Supplementary Material

related to “Silveira PSP, Rocha FT, Vieira JE, Siqueira JO. From Blue November to Broader Diagnosis: The Youden Index to Evaluate the Performance of Any Diagnostic Tests.

prepared by Paulo Sergio Panse Silveira
and Jose Oliveira Siqueira

Departmento de Patologia da Faculdade de Medicina da Universidade de São Paulo, SP, Brazil

Contacts:
Paulo Sergio Panse Silveira <silveira@usp.br>
Jose Oliveira Siqueira <silveira@usp.br>
Av. Dr. Arnaldo 455 sala 1103
01246-903 - Sao Paulo, SP, Brazil

May 17, 2025

Detailed Methodological Description and Examples, complementary to

From Blue November to Broader Diagnosis: The Youden Index to Evaluate the Performance of Any Diagnostic Tests

Paulo Sergio Panse Silveira[1],*, Flavio Trigo Rocha[2,3], Joaquim Edson Vieira[4,5], and Jose Oliveira Siqueira[1]

[1] Departmento de Patologia da Faculdade de Medicina da Universidade de São Paulo, SP, Brazil
[2] Disciplina de Urologia, Departmento de Cirurgia, Faculdade de Medicina da Universidade de São Paulo, SP, Brazil
[3] Chief Urologist of Voiding dysfunction section, Hospital Sirio Libanes, Sao Paulo, SP, Brazil
[4] Disciplina de Anestesiologia, Departmento de Cirurgia, Faculdade de Medicina da Universidade de São Paulo, SP, Brazil
[5] Faculdade Israelita de Ciências da Saúde Albert Einstein, Sao Paulo, SP, Brazil

Contents

1 Statistical methods **ii**
1.1 Background ii
1.2 Youden's index v
1.3 One sample test of the Youden index viii
1.4 Independent samples test of Youden indices (between groups) x
1.5 Dependent samples test of Youden indices (within-group) . xi

2 Computational procedures **xii**
2.1 Packages to install xii
2.2 Fig1Createdata.R and Fig1Prevalence.R xii
2.3 Replication of Chen et. al (2015) results xvii
2.4 Chen2015between.R xvii
2.5 Chen2015within.R xx
2.5.1 output xxiii
2.6 ErdoganvsReckerOneSample.R xxvi
2.6.1 output xxix
2.7 ErdoganvsRecker.R xxxiii
2.7.1 output xliii

3 Replication of Manuscript Tables **lx**

4 Replication of Manuscript Figures **lxviii**

List of Tables

Y-1 D_+ and D_- represent the presence or absence of a disease or patient condition, respectively. T_+ and T_- denote the occurrence of a positive or negative result from any symptom, signal, or test. Along the main diagonal, *a* and *D* indicate the counts or proportions of agreement between *D* and *T*, while along the secondary diagonal, *b* and *c* indicate the counts or proportions of disagreement between *D* and *T*. ii

Y-2 General 2x2 table rearrangement proposed by Chen et al. (2015) for within-group comparison of diagnostic exams. This test is based on the separation between diseased and healthy (control) individuals, obtained from raw data or by authors providing the values for a_{D_+}, b_{D_+}, c_{D_+}, d_{D_+}, a_{D_-}, b_{D_-}, c_{D_-}, or d_{D_-}. If these values are not provided, all arrangements that satisfy the marginal conditions of the bottom tables should be used to verify the agreement between the exams for the two patient groups.* All procedures are implemented by the R function *twosample.Youden* for testing the null hypothesis (H_0: $J_2 - J_1 = 0$). xxi

1 D_+ and D_- represent the presence or absence of a disease or patient condition, respectively. T_+ and T_- denote the occurrence of a positive or negative result from any symptom, signal, or test. Along the main diagonal, *a* and *D* indicate the counts or proportions of agreement between *D* and *T*, while along the secondary diagonal, *b* and *c* indicate the counts or proportions of disagreement between *D* and *T*. lx

2 Reproduction of Chen et al. (2015) data for the estimation of exams for bladder carcinoma: immunoCyt™ and Cytology. For each test, sensitivity (*se*), specificity (*sp*), Youden index (*J*), and p value were computed using the R function *onesample.Youden* testing $H_0 : J \leq 0$. For the difference between tests, 95% confidence interval was computed by the R function *twosample.Youden* testing $H_0 : J_2 - J_1 = 0$. . lxi

3 Reproduction of Chen et al. (2015) data for the estimation of exams for tuberculosis detection in a within-group design: all patients were tested with ELISA and ELISPOT methods. For each test, sensitivity (*se*), specificity (*sp*), Youden index (*J*), and p value were computed using the R function *onesample.Youden* testing H_0: $J < 0$. For the difference between tests, confidence interval was computed by the R function *twosample.Youden* testing H_0: J_2-$J_1 = 0$.* lxii

4 Analysis of Erdogan et al. (2020) data for the estimation of prostate volume (PV) as a predictor of prostate cancer (PCa). For each test, sensitivity (*se*), specificity (*sp*), Youden index (*J*), and p value were computed using the R function *onesample.Youden* testing H_0: $J \leq 0$. For the difference between tests, confidence interval was computed by the R function *twosample.Youden* testing H_0: J_2-$J_1 = 0$. . lxiii

5 Reproduction of Recker et al. (1998) data for the estimation of Prostate-Specific Antigen (PSA) tests: total PSA (tPSA), free to total PSA (f/tPSA020) with a cut-off point at 0.20, and free to total PSA with a cut-off point at 0.15 (f/tPSA015). Patients were diagnosed with prostate cancer (D_+) or benign prostatic hyperplasia (D_-), and the PSA tests could result in positive (T_+) or negative (T_-) outcomes. Sensitivity (*se*), specificity (*sp*), Youden index (*J*), and p value were computed using the R function *onesample.Youden* testing H_0: $J \leq 0$. lxiv

6 Performance difference (within-group design) with data reproduced from Recker et al. (1998) evaluated by the Youden index (*J*) of three Prostate-Specific Antigen (PSA) tests: total PSA (tPSA), free to total PSA (f/tPSA020) with a cut-off point at 0.20, and free to total PSA with a cut-off point at 0.15 (f/tPSA015). 95% confidence intervals and *p* values were computed using the R function *twosample.Youden* testing H_0: J_2-$J_1 = 0$. GD_+ and GD_- are agreement estimates obtained from Gwet's AC_1 (see text for explanation). lxv

7 Performance difference (between-group design) with data reproduced from Erdogan et al. (2020) and Recker et al. (1998) evaluated by the Youden index (*J*). Erdogan proposed the prostate volume (PV) as a predictor of cancer in two groups of patients with two different cut-off points (PSA 2.5-10 ng/ml with cutoff of PV = 43.5 ml; PSA 10.1-30.0 ng/ml with cutoff of PV = 63.5 ml). Recker applied three Prostate-Specific Antigen (PSA) tests: total PSA (tPSA), free to total PSA (f/tPSA 0.20) with a cut-off point at 0.20, and free to total PSA with a cut-off point at 0.15 (f/tPSA 0.15). 95% confidence intervals of *J* and *p* values were computed using the R function *twosample.Youden* testing H_0: J_2-J_1 = 0. lxvi

8 Comparison of Recker et al.'s total PSA (reference) with proposed biomarkers. Statistical significance is highlighted with a gray background. When a proposed biomarker qualifies as a diagnostic test (95% confidence interval for J is entirely above zero), advance is considered if the difference from the reference (*Diff. J*) is also significantly greater than zero. lxvii

List of Figures

Y-1 Figures obtained with `Chen2015between.R` (not included in the manuscript). xix
Y-2 Figures obtained with `Chen2015between.R` (not included in the manuscript). xxv
Y-3 Figure obtained with `ErdoganvsReckerOneSample.R`, not included in the manuscript (see section "One sample test of the Youden index" in the manuscript related to the test utility). xxxii
Y-4 Exam comparisons. First row: comparison of prostate volume (PV) with cutoff values at 43.5 ml and 61.5 ml in patients with serum PSA levels of 2.5–10.0 ng/ml and 10.1–30.0 ng/ml (Erdogan et al., 2020). Second row: comparison of versus total PSA with a cutoff of 4 ng/ml, and free/total PSA ratios with cutoff values of 0.20 and 0.15 (Recker et al., 1998). Third and forth rows: comparison of two conditions of Erdogan et al (2020) and three conditions of Recker et al. (1998). The plots depict the probability of disease (PPV and 1-NPV) in function of prevalence or pre-test probability. The shadows correspond to overlaping confidence intervals of PPV and 1-NPV, which corresponds to the lack of evidence of difference between all attempts to improve prostate cancer diagnosis. lvii
Y-5 The figure was obtained using `ErdoganvsReckerComparison.R` (see section "Independent samples test of Youden indices (between groups)" in the manuscript). The left and right columns represent the proposition by Erdogan et al. (2020)[1], where prostate volume is used as a diagnostic tool with different cut-off points for patients with varying PSA levels. The lines represent the approach by Recker et al. (1998)[2], using total PSA or two cut-off points for free-to-total PSA. It can be observed that at least one hypothetical line for PPV and $1 - NPV$ can be drawn across the common area shared by both tests. This graphical representation indicates the absence of a significant difference between the tests. . . . lviii

1 Simulation of 500,000 2×2 tables ($n = 500$) with the Youden index ranging from $0 < J < 1$ in intervals of 10% (light to dark gray). The figure shows that sensitivity (se) and specificity (sp) do not depend on prevalence. However, the probability of disease is affected by prevalence, increasing more with positive test results (positive predictive values, PPV) and decreasing more with negative test results (complement of the negative predictive values, $1 - NPV$) as the value of J increases. lxviii

2 Examples of diagnostic exams with varying sensitivity and specificity values. It is illustrated with updated probabilities of disease (pre-test) with an initial suspicion of 50.0% after a positive test result (PPV) or negative test result ($1 - NPV$). Shadows are 95% confidence bands ($n = 500$). **A** and **B**: hypothetical exams with equal values of Youden index (J) switching sensitivity (se) and specificity(sp); **C**: estimated values for ELISA for HIV detection (França et al., 2018); **D**, **E** and **F**: prostate detection assessed by seric PSA varying the cutoff points, respectivelly 4.1, 1.1, and 10.1 ng/ml (Thompson & Ankerst, 2007). lxix

3 Decision-making algorithm integrating Youden index (J) with prevalence-adjusted positive (PPV) and negative (NPV) predictive values. lxx

1 Statistical methods

1.1 Background

The following explains when a test has diagnostic value, converging into a practical, grounded rule: the Youden index. The notation adopted in this text (Table Y-1) summarizes all possible interactions between an available exam result, such as an observable symptom, a signal detected by physical examination, the result of a laboratory test or an imaging diagnosis (collectively referred to as the 'test', T), and any patient status, such as the presence of a disease, the occurrence of death, or the existence of a morbid condition (referred to as the 'disease', D). Any test may provide correct results, represented by the counts in the a or d cells, but it is also subject to errors, represented by counts in the b or c cells.

Table Y-1: D_+ and D_- represent the presence or absence of a disease or patient condition, respectively. T_+ and T_- denote the occurrence of a positive or negative result from any symptom, signal, or test. Along the main diagonal, *a* and *D* indicate the counts or proportions of agreement between *D* and *T*, while along the secondary diagonal, *b* and *c* indicate the counts or proportions of disagreement between *D* and *T*.

	D_+	D_-	total
T_+	*a* true positive	*b* false positive	*a+b*
T_-	*c* false negative	*D* true negative	*c+d*
total	*a+c*	*b+d*	*a+b+c+d*

The probability of a positive test to be correct is therefore the counting of positive tests given that the patient has a disease, expressed by

$$P(T_+|D_+) = \frac{a}{a+c} = se \tag{1}$$

Literally read as "the probability, P, of a positive test, T_+, provided that '|' the patient is diseased, D_+, is equal to the counting of positive tests, a, among the patients that are known to be diseased, $a + c$", which is the proportion of true positives and the definition of test sensitivity (se).

Similarly, the probability of a negative test being correct is defined by

$$P(T_-|D_-) = \frac{d}{b+d} = sp \tag{2}$$

which is the proportion of true negatives known as test specificity (sp). The complementary probabilities to se and sp are the probabilities of a false negative computed by $P(T_+|D_-) = \frac{c}{a+c}$ and a false positive by $P(T_-|D_+) = \frac{d}{b+d}$.

Physicians are primarily interested in the reverse probabilities $P(D_*|T_*)$, which indicate the likelihood of having a disease given a test result, rather than $P(T_*|D_*)$. This distinction is crucial, as many mistakenly equate sensitivity with the probability of disease when a test is positive. Even a highly sensitive and specific test may not yield a high probability of disease, known as positive predictive value (PPV = $P(D_+|T_+)$), which differs from $P(T_+|D_+)$. Similarly, the negative predictive value (NPV = $P(D_-|T_-)$) reflects the reverse order of specificity ($P(T_-|D_-)$).

In a laboratory setting, sensitivity and specificity are determined by testing patients with known disease statuses, reflecting the intrinsic quality of the test. In contrast, medical practice involves interpreting test results where the presence of disease is unknown. Here, positive and negative predictive values depend on disease prevalence, which affects the quality of diagnosis[3]. Prevalence can be expressed as $P(D_+)$ or simply as p.

$$\begin{aligned} p = P(D_+) &= \frac{a+c}{a+b+c+d} \\ 1-p = 1-P(D_+) = P(D_-) &= \frac{b+d}{a+b+c+d} \end{aligned} \tag{3}$$

where $P(D_-)$ is the prevalence of the absence of disease.

It is not immediately obvious why the estimators of sensitivity (se) and specificity (sp) are not correlated with prevalence, while the positive predictive value (PPV) and the negative predictive value (NPV) are profoundly affected by prevalence[4]. Referring to the main manuscript Table Y-1, we can understand that the operator '|' ('provided that') represents a conditional probability that segments the table by columns or rows. For sensitivity, the condition that "there is disease" limits the probability to the first column. Here, the division between a and c depends solely on $P(T_+)$ and $P(T_-)$, resulting in $se = \frac{a}{a+c}$, independent of how many diseased subjects ($a+c$) are included in the table. The same logic applies to specificity, which is confined to the second column, giving $sp = \frac{d}{b+d}$. In contrast, for PPV (when the test result is positive), the condition applies to the first row, and for NPV (when the test result is negative), it applies to

the second row, making the division dependent on $P(D_+)$. These values, based on the slice of these rows, compute:

$$\begin{aligned} PPV &= P(D_+|T_+) = \frac{a}{a+b} \\ NPV &= P(D_-|T_-) = \frac{d}{c+d} \end{aligned} \tag{4}$$

Consequently, in situations where the researcher is interested in measuring sensitivity and specificity, any number of healthy and diseased subjects (gold standard) can be recruited to compose this contingency table. However, determining PPV or NPV applying the rule of equation 4 is valid only for a contingency table that reflects the prevalence of the population from which the patients come. Otherwise, we should not calculate PPV or NPV. In these cases, the correct way to obtain them is through Bayes' theorem, using the prevalence of the disease (p) in the population from which the patient of interest originates, through the following expressions:

$$\begin{aligned} PPV &= P(D_+|T_+) = \frac{se \cdot p}{se \cdot p + (1-sp) \cdot (1-p)} \\ NPV &= P(D_-|T_-) = \frac{sp \cdot (1-p)}{sp \cdot (1-p) + (1-se) \cdot p} \\ 1 - NPV &= P(D_+|T_-) \end{aligned} \tag{5}$$

Furthermore, it is sometimes confusing to interpret NPV, as it represents the probability of absence of disease provided that the test result is negative (a double negative). Therefore, throughout this text and to keep the focus on the disease, we prefer to present $P(D_+|T_-) = 1 - NPV$, which indicates the probability of still having the disease when the test result is negative.

Another popular coefficient is the likelihood ratio (LR). Since positive and negative LR are relationships between se and sp, they are also

measures of test quality (not diagnosis quality), computed by

$$LR_+ = \frac{P(T_+|D_+)}{P(T_+|D_-)} = \frac{se}{1-sp} = \frac{\frac{a}{a+c}}{\frac{b}{b+d}}$$

$$LR_- = \frac{P(T_-|D_+)}{P(T_-|D_-)} = \frac{1-se}{sp} = \frac{\frac{c}{a+c}}{\frac{d}{b+d}} \quad (6)$$

In other words, LR_+ can be interpreted as the odds between true and false positives or how many times a test has sensitivity compared to nonspecificity. LR_- is the odds between false and true negatives or how many times a test has unsensitivity compared to specificity. In a test of good quality (high se and sp), LR_+ can assume values from 1 to infinity (true positive $>$ false positive, therefore, the higher its value, the better the test performance when the result is positive) and LR_- can assume values from 0 to 1 (false negative $<$ true negative, thus the lower its value, the better the test performance when the result is negative). The odds ratio is equal to the ratio of the likelihood ratios, $OR = \frac{LR_+}{LR_-} = \frac{ad}{bc}$.

1.2 Youden's index

Tests must have high sensitivity and specificity. Although in practice a test with 100% sensitivity (se) and 100% specificity (sp) does not exist, theoretically such a test would result in positive or negative predictive values 100%, irrespective of prevalence. This hypothetical test would be a perfect oracle. Technological advances aim to approach this ideal by producing tests with ever-increasing sensitivity and specificity, although never achieving 100% performance. Therefore, clinical assessment (prevalence or pre-test medical belief) remains necessary for accurate diagnosis.

Medical researchers require an index to determine whether a new test significantly outperforms its predecessor, providing physicians with greater confidence and certainty in making diagnoses. This statistical measure, known as the Youden index (J), was originally presented by William John

Youden in 1950[5]. It is based on the concept of test success, which is measured by the rate of misclassifications. Thus, for diseased patients, a test correctly classifies a proportion $\frac{a}{a+c}$ (true positives) and incorrectly classifies a proportion $\frac{c}{a+c}$ (false negatives). The success rate is thus $\frac{a}{a+c} - \frac{c}{a+c} = \frac{a-c}{a+c}$. Similarly, for healthy patients, the difference between $\frac{d}{b+d}$ (true negative) and $\frac{b}{b+d}$ (false positive) proportions results in a success rate of $\frac{d-b}{d+b}$. Youden also highlighted that false negatives are particularly problematic when delayed treatment can alter the course of a disease. In contrast, false positives can lead to the misuse of resources that should be available for genuinely diseased patients. Deciding which type of diagnostic error carries more weight is not a statistical matter but a clinical one.

Youden, then, assuming that false negative and false positive are equally important, computes the index as the average success, obtaining

$$J = \frac{\frac{a-c}{a+c} + \frac{d-b}{d+b}}{2} = \frac{ad-bc}{(a+c)(b+d)} = \frac{a}{a+c} - \frac{b}{b+d} \tag{7}$$

According to Tadeusz & Tadeusz, 2020[6] alternative formulas of J can reinforce its different aspects:

- Formula for J reflecting the excess in the proportion of a positive result between patients with the disease versus patients without the disease (i.e., positive minus false positive):

$$J = se - (1 - sp) \tag{8}$$

- Formula for J reflecting the excess in the proportion of a negative result between patients without the disease versus patients with the disease: (i.e., negative minus false negative)

$$J = sp - (1 - se) \tag{9}$$

- Formula for J reflecting the excess in balanced detection accuracy (i.e., average of true positive and negative tests) versus balanced detection error (i.e., average of false positive and negative tests):

$$J = \frac{se + sp}{2} - \frac{(1 - se) + (1 - sp)}{2} \tag{10}$$

- Formula for J reflecting the excess in the joint probabilities of correct detection (positive and negative) versus the joint probabilities of incorrect detection (positive and negative):

$$J = se \cdot sp - (1 - se) \cdot (1 - sp) \tag{11}$$

All these equations are equivalent to:

$$J = se + sp - 1 \tag{12}$$

which is the easiest way to remember and apply it in practice as the "rule of thumb" to quickly assess the quality of a test, that is $se + sp > 1$.

Equation 7 to 12 in many equivalent forms reveals a series of convenient properties of Youden index:

- By construction, J is a direct measure of correct classifications; therefore, independently of the table configuration, two different tests with equal J produce $1 - J$ misclassifications, although they may show different proportions of false positives or negatives.

- The numerator $ad - bc$ shows a tension between diagonals, with ad representing the test correctness and bc its false results; in that sense, J has similarities with the odds ratio. In this context, only $ad > bc$ or the lower limit $J > 0$ makes sense by definition: no one has application to a test that produces more false than true results.

- $ad - bc$ is the determinant of the 2 $\times$ 2 table presented in Table Y-1. According to Tadeusz & Tadeusz, 2020[6] and Rogan & Gladen[3], this determinant is positive, i.e., $ad - bc > 0$ if and only if $J > 0$ or $LR_+ > 1$ or $OR > 1$ or $PPV > p$ or $PPV > NPV$, therefore being a necessary condition for an exam to be a diagnostic test.

- Any test able to produce more true than false results (i.e., $J > 0$) will indicate $PPV > p$ when the test result is positive (increasing the pre-test medical belief, p) and $1 - NPV < p$ when the test result is negative (decreasing the medical belief in a given diagnoses).
- The upper limit of J is 1, which occurs when there are no false results; $b = 0$ and $c = 0$ would represent the oracle test that does not exist in practice. Since there is always some kind of false result, due to the denominator $(a + c)(b + d)$, it results that $0 < J < 1$.
- Mathematically, J can be less than zero ($-1 \leq J \leq 1$), but negative values ($J < 0$) should not occur in practice. If a diagnostic test had $se + sp < 1$, it would function inversely, reducing the probability of diagnosis when positive and increasing it when negative.
- J is a direct measure of test quality, which is obtained from $se+sp-1$. Like se and sp, J does not depend on prevalence. Thus, observing $se + sp > 1$ is another practical approach to perceive if a test is of good quality.

According to Haimes, 2008, Chap. 12[7], PPV and NPV can be expressed in terms of the Youden index (J), prevalence (p), sensitivity (se), and specificity (sp):

$$\begin{aligned} PPV &= \frac{se\ p}{1 - sp + pJ} \\ NPV &= \frac{sp(1-p)}{1 - se + (1-p)J} \end{aligned} \tag{13}$$

1.3 One sample test of the Youden index

All computations above are point estimates of populational values based on sample contingency tables. Therefore, in order to achieve conclusions of enough quality of a test to present diagnostic performance, inferential statistics are necessary. We have discussed that $J > 0$ is a minimum requirement for a test to present more true than false results. Inferential tests must require that the confidence interval of J be above zero, i.e., one-sided statistical analysis.

$$\begin{aligned} H_0 &: J = 0 \\ H_1 &: J > 0 \end{aligned} \tag{14}$$

According to Youden[5], estimators of sensitivity (among diseased subjects) and specificity (among healthy subjects) are proportions with binomial distributions, thus

$$\begin{aligned} s_{se} &= \sqrt{\frac{se(1-se)}{np}} = \sqrt{\frac{\frac{a}{a+c}\left(1-\frac{a}{a+c}\right)}{a+c}} = \sqrt{\frac{ac}{(a+c)^3}} \\ s_{sp} &= \sqrt{\frac{sp(1-sp)}{n(1-p)}} = \sqrt{\frac{\frac{d}{b+d}\left(1-\frac{d}{b+d}\right)}{b+d}} = \sqrt{\frac{bd}{(b+d)^3}} \end{aligned} \tag{15}$$

Since these two distributions are independent from each other, the global standard error is the standard error of the Youden index (since J is the proportion of global correctness):

$$s_J = \sqrt{s_{se}^2 + s_{sp}^2} = \sqrt{\frac{ac}{(a+c)^3} + \frac{bd}{(b+d)^3}} \tag{16}$$

The test statistic is

$$z = \frac{J}{s_J} \tag{17}$$

Consequently, the 95% confidence interval of J is

$$CI^{95\%}(J) = [J \pm 1.64 \cdot s_J] \tag{18}$$

which means that there is a minimum value of J for a test to be assumed as a diagnostic test, corresponding to the rejection of the null hypothesis (Equation 14) when $J - 1.64 \cdot s_J > 0$

This standard error was recently improved by Chen et al.[8], taking into account that sensitivity and specificity may not be truly independent and can covariate, which was also implemented in our R scripts:

$$\begin{aligned} s'_J &= \sqrt{s_{se}^2 + s_{sp}^2 + 2 \cdot \mathrm{cov}(se, sp)} \\ s'_J &= \sqrt{\left(\frac{d}{b+d}\right)^2 \frac{ac}{(a+c)^3} + \left(\frac{a}{a+c}\right)^2 \frac{bd}{(b+d)^3}} \end{aligned} \tag{19}$$

1.4 Independent samples test of Youden indices (between groups)

In order to compare diagnostic performance with a modified and supposedly improved exam, an experiment can be conducted with two independent groups of subjects. Inferential statistics has the null hypothesis

$$\begin{aligned} H_0 &: J_1 = J_2 \\ H_1 &: J_1 \neq J_2 \end{aligned} \tag{20}$$

Based on Chen et al.[8], the squared standard error for the test statistic of H_0 is

$$\begin{aligned} \left(s^i_{J_1-J_2}\right)^2 = \; & \left(\frac{d_1}{b_1+d_1}\right)^2 \frac{a_1c_1}{(a_1+c_1)^3} + \left(\frac{a_1}{a_1+c_1}\right)^2 \frac{b_1d_1}{(b_1+d_1)^3} + \\ & \left(\frac{d_2}{b_2+d_2}\right)^2 \frac{a_2c_2}{(a_2+c_2)^3} + \left(\frac{a_2}{a_2+c_2}\right)^2 \frac{b_2d_2}{(b_2+d_2)^3} \end{aligned} \tag{21}$$

The test statistic is

$$z = \frac{J_1 - J_2}{s^i_{J_1-J_2}} \tag{22}$$

and the 95% confidence interval is

$$CI^{95\%}\left(J_1 - J_2\right) = \left[(J_1 - J_2) \pm 1.96 \cdot s^i_{J_1-J_2}\right] \tag{23}$$

1.5 Dependent samples test of Youden indices (within-group)

A more powerful possibility is to apply both the original and modified diagnostic exam to the same subjects, obtaining

$$\begin{aligned}\left(s^d_{J_1-J_2}\right)^2 = \quad & s_{J_1} + s_{J_2} - 2G_{D_+}\sqrt{\frac{a_1c_1}{(a_1+c_1)^3}\frac{a_2c_2}{(a_2+c_2)^3}} \\ & -2G_{D_-}\sqrt{\frac{b_1d_1}{(b_1+d_1)^3}\frac{b_2d_2}{(b_2+d_2)^3}}\end{aligned} \tag{24}$$

where G_{D_+} and G_{D_-} are the agreement of both tests in diseased and non-diseased subjects, respectively. It is a modification of Chen et al.[8], who applied Cohen's kappa as agreement measure. We rather apply Gwet's $AC1$ or Holley and Guilford's G, more suitable estimators of agreement[9].

The test statistic is

$$z = \frac{J_1 - J_2}{s^d_{J_1-J_2}} \tag{25}$$

and the 95% confidence interval is

$$CI^{95\%}\left(J_1 - J_2\right) = \left[(J_1 - J_2) \pm 1.96 \cdot s^d_{J_1-J_2}\right] \tag{26}$$

2 Computational procedures

The codes below replicate all the results presented in the manuscript.

2.1 Packages to install

To run them, it is necessary to install the following packages (in that order), available on the Harvard Dataverse:

- `eirasdata`: https://doi.org/10.7910/DVN/DLQTPH
- `eiras`: https://doi.org/10.7910/DVN/TBBAVU
- `eiras2x2`: https://doi.org/10.7910/DVN/5QTMBW

2.2 Fig1Createdata.R and Fig1Prevalence.R

Simulations offer an alternative way to illustrate the relationship among sensitiviy, specificity, predictive values and Youden index without requiring an in-depth quantitative background. Through simulations, we can effectively demonstrate how prevalence influences the reliability of diagnostic tests, making it easier to grasp the implications of these diagnostic metrics in clinical practice.

The proposed algorithm visualizes how these key diagnostic parameters vary across different prevalences. This approach reinforces the mathematical argument that sensitivity and specificity, which reflect the quality of the test, are independent of prevalence, while PPV and NPV, which reflect the quality of the diagnosis, are affected by prevalence.

The manuscript Figure Fig. 1 is provided by `Fig1Createdata.R` (it takes time to generate all necessary data, producing `simulation.xlsx`) and `Fig1Prevalence.R` (it reads this data to create the figure).

Fig1Createdata.R:

```
levels.prev <- seq(0.01, 0.99, 0.01)
levels.J <- seq(0.0, 1.0, 0.1)
levels.J[1] <- 0.01
levels.J[length(levels.J)] <- 0.99
```

```
5
6  n <- 500
7  # for the paper we applied 5e5 (it may take a long time)
8  attempts <- 5e4
9
10 namesdata <- c("a","b","c","d","n","prev","se","sp","J","
       ppv","1-npv")
11 dt_data <- data.frame(matrix(nrow=attempts, ncol=length(
       namesdata)))
12 names(dt_data) <- namesdata
13
14 cat("\n")
15 for (s in 1:attempts)
16 {
17   if((s%%100)==0)
18   {
19     cat(".")
20     if((s%%10000)==0)
21     {
22       cat(s,"/",attempts,"\n")
23     }
24   }
25   a <- b <- c <- d <- 0
26   while(a<1 | b<1 | c<1 | d<1)
27   {
28     # random level of prevalence
29     v.aux <- sample(x=1:length(levels.prev),size=1)
30     sp <- 0
31     while(sp<=0 | sp>=1)
32     {
33       # random level of Youden
34       j.aux <- sample(x=1:length(levels.J),size=1)
35       # random se
36       se <- round(runif(1,0,1),3)
37       # consequent specificity has to be possible
38       sp <- round(levels.J[j.aux]+1-se,3)
39     }
40     # value of a
41     ac <- levels.prev[v.aux]*n
```

```
    # a <- levels.se[s.aux]*ac
    a <- se*ac
    se <- a/ac
    c <- ac-a
    # value of d
    d <- sp*(n-ac)
    sp <- d/(n-ac)
    b <- (n-ac)-d
    n <- a+b+c+d
  }

  prev <- ac/n
  J <- se+sp-1
  ppv <- a/(a+b)
  npvm1 <- c/(c+d)

  dt_data$a[s] <- a
  dt_data$b[s] <- b
  dt_data$c[s] <- c
  dt_data$d[s] <- d
  dt_data$n[s] <- n
  dt_data$prev[s] <- prev
  dt_data$se[s] <- se
  dt_data$sp[s] <- sp
  dt_data$J[s] <- J
  dt_data$ppv[s] <- ppv
  dt_data$`1-npv`[s] <- npvm1
}

openxlsx::write.xlsx(dt_data,"simulation.xlsx")
cat("\n")
```

Fig1Prevalence.R:

```
levels.prev <- seq(0.01, 0.99, 0.01)
levels.J <- seq(0.0, 1.0, 0.1)
levels.J[1] <- 0.01
levels.J[length(levels.J)] <- 0.99
```

```
5
6  n <- 500
7  # for the paper we applied 5e5 (it may take a long time)
8  attempts <- 5e4
9
10 namesdata <- c("a","b","c","d","n","prev","se","sp","J","
       ppv","1-npv")
11 dt_data <- data.frame(matrix(nrow=attempts, ncol=length(
       namesdata)))
12 names(dt_data) <- namesdata
13
14 cat("\n")
15 for (s in 1:attempts)
16 {
17   if((s%%100)==0)
18   {
19     cat(".")
20     if((s%%10000)==0)
21     {
22       cat(s,"/",attempts,"\n")
23     }
24   }
25   a <- b <- c <- d <- 0
26   while(a<1 | b<1 | c<1 | d<1)
27   {
28     # random level of prevalence
29     v.aux <- sample(x=1:length(levels.prev),size=1)
30     sp <- 0
31     while(sp<=0 | sp>=1)
32     {
33       # random level of Youden
34       j.aux <- sample(x=1:length(levels.J),size=1)
35       # random se
36       se <- round(runif(1,0,1),3)
37       # consequent specificity has to be possible
38       sp <- round(levels.J[j.aux]+1-se,3)
39     }
40     # value of a
41     ac <- levels.prev[v.aux]*n
```

```
    # a <- levels.se[s.aux]*ac
    a <- se*ac
    se <- a/ac
    c <- ac-a
    # value of d
    d <- sp*(n-ac)
    sp <- d/(n-ac)
    b <- (n-ac)-d
    n <- a+b+c+d
  }

  prev <- ac/n
  J <- se+sp-1
  ppv <- a/(a+b)
  npvm1 <- c/(c+d)

  dt_data$a[s] <- a
  dt_data$b[s] <- b
  dt_data$c[s] <- c
  dt_data$d[s] <- d
  dt_data$n[s] <- n
  dt_data$prev[s] <- prev
  dt_data$se[s] <- se
  dt_data$sp[s] <- sp
  dt_data$J[s] <- J
  dt_data$ppv[s] <- ppv
  dt_data$`1-npv`[s] <- npvm1
}

openxlsx::write.xlsx(dt_data,"simulation.xlsx")
cat("\n")
```

2.3 Replication of Chen et. al (2015) results

As a first exercise, we reproduced the results of Chen et al.[8], who proposed some improvements to the statistical tests originally proposed by Youden[5]. Following their ideas, all equations were written in terms of Table Y-1, using the quantities a, b, c, d. Our contributions are (1) to rewrite the equations to show more explicitly the origin of standard deviations of J from se and sp for the estimation of the statistics and confidence interval (equations 15 to 19, 21 to 23, and 24 to 26), (2) to declare the null hypotheses (equations 14 and 20), and (3) to implement all procedures in the R functions.

Two functions in R were developed. The first (`onesample.Youden`) tests the null hypothesis expressed in equation 14. The second (`twosample.Youden`) compares two diagnostic tests considering the equation 20 and implements both between-groups and within-group situations.

2.4 Chen2015between.R

Replication of data used in manuscript Table 2. This script also creates Figure Y-1, which was not included in the manuscript.

```
# Chen 2015
#    D+  D-
# T+  a   b
# T-  c   d

t <- a  <- b  <- c  <- d  <- rep(NA,2)
t[1] <- "immunoCyt(TM)"
a[1] <- 85
b[1] <- 51
c[1] <- 16
d[1] <- 182
t[2] <- "Cytology"
a[2] <- 41
b[2] <- 12
c[2] <- 13
d[2] <- 24
```

```
17
18 # immunoCyt(TM)
19 m <- list()
20 for (aux in 1:2)
21 {
22   m[[aux]] <- eiras2x2::abcd2matrix(a=a[aux],b=b[aux],c=c[
       aux],d=d[aux],
23              rownames=c("T+","T-"),
24              colnames=c("D+","D-"))$m
25   print(res <- eiras2x2::onesample.Youden(m[[aux]],main=t[
       aux]))
26   # PPV and NPV
27   plot(res, col="forestgreen", lty=1, lwd=2, plot.npv =
       FALSE)
28   plot(res, col="tomato", lty=2, lwd=2, plot.ppv = FALSE,
       add=TRUE,
29        legend.txt=c("PPV","1-NPV"),
30        legend.col=c("forestgreen","tomato"),
31        legend.lwd=c(2,2),
32        legend.lty=c(1,2))
33   # PPV and NPV and hypothetical pre-test value
34   plot(res, col="royalblue",
35        lty=1, lwd=2, curve.lab.at = rep(0.9,2),
36        mcex=1,
37        tp="12", # less intense polygon shadow (use "00" for
              none),
38        digits=2, # default is 3
39        main=paste0("Performance of a hypothetical\nresult
             for ",t[aux]),
40        xlab="Probability of disease before testing",
41        ylab="Probability of disease after test result",
42        prevalence=0.63
43   )
44 }
45
46 # Comparison
47 res.between <- eiras2x2::twosample.Youden(m1=m[[1]],m2=m
     [[2]],
48        main1=t[1],
```

```
        main2=t[2])
print(res.between)
```

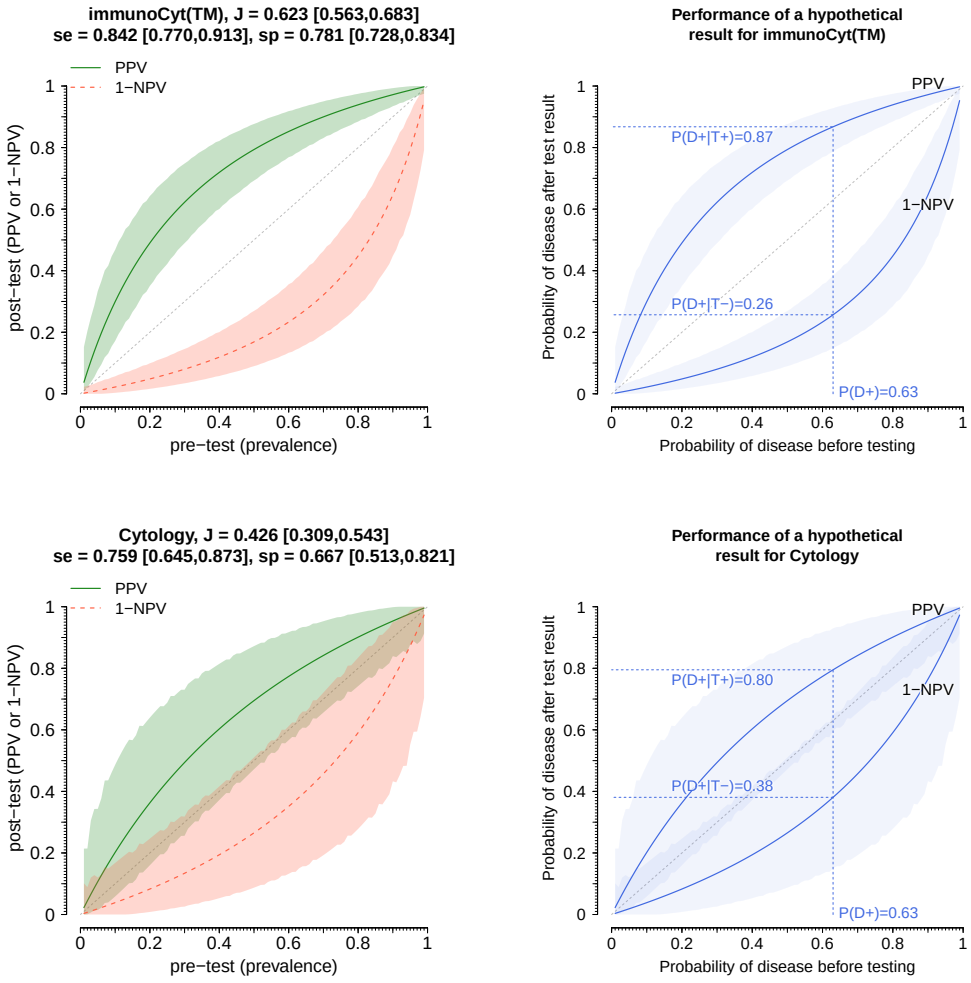

Figure Y-1: Figures obtained with `Chen2015between.R` (not included in the manuscript).

2.5 Chen2015within.R

This script replicates data shown in manuscript Table 3.

For within-group comparison of diagnostic tests, we provide a general schema for constructing these tables based on the number of patients with both tests yielding positive or negative results (manuscript Table Y-2). In doing so, we also identified a minor error in the header of Chen's original Table XVI, which we corrected in manuscript Table 3.

Table Y-2: General 2x2 table rearrangement proposed by Chen et al. (2015) for within-group comparison of diagnostic exams. This test is based on the separation between diseased and healthy (control) individuals, obtained from raw data or by authors providing the values for a_{D+}, b_{D+}, c_{D+}, d_{D+}, a_{D-}, b_{D-}, c_{D-}, or d_{D-}. If these values are not provided, all arrangements that satisfy the marginal conditions of the bottom tables should be used to verify the agreement between the exams for the two patient groups.* All procedures are implemented by the R function *twosample.Youden* for testing the null hypothesis (H_0: $J_2 - J_1 = 0$).

Exam 1

	D_{1+}	D_{1-}
T_{1+}	a_1	b_1
T_{1-}	c_1	d_1

Exam 2

	D_{2+}	D_{2-}
T_{2+}	a_2	b_2
T_{2-}	c_2	d_2

Diseased group:

Exam 1

Exam 2		T_{1+}	T_{1-}	
	T_{2+}	a_{D+}	b_{D+}	$P(T_{2+}\|D_{2+}) = a_2$
	T_{2-}	c_{D+}	d_{D+}	$P(T_{2-}\|D_{2+}) = c_2$
		$P(T_{1+}\|D_{1+}) = a_1$	$P(T_{1-}\|D_{1+}) = c_1$	

Control group:

Exam 1

Exam 2		T_{1-}	T_{1+}	
	T_{2-}	a_{D-}	b_{D-}	$P(T_{2-}\|D_{2-}) = d_2$
	T_{2+}	c_{D-}	d_{D-}	$P(T_{2+}\|D_{2-}) = b_2$
		$P(T_{1-}\|D_{1-}) = d_1$	$P(T_{1+}\|D_{1-}) = b_1$	

* Marginals to the bottom tables are obtained from the upper 'Exam' 2x2 tables. Since 2x2 tables have only one degree of freedom, it suffices to provide for *twosample.Youden* the values of a_{D+} (the number of patients for whom both tests were positive) and a_{D-} (the number of patients for whom both tests were negative) to determine the partition between the diseased and control groups.

This script also creates Figure Y-2, which was not included in the manuscript:

```
# Chen 2015
#    D+  D-
# T+  a   b
# T-  c   d
```

```
5
6  t <- a  <- b  <- c  <- d  <- rep(NA,2)
7  t[1] <- "immunoCyt(TM)"
8  a[1] <- 85
9  b[1] <- 51
10 c[1] <- 16
11 d[1] <- 182
12 t[2] <- "Cytology"
13 a[2] <- 41
14 b[2] <- 12
15 c[2] <- 13
16 d[2] <- 24
17
18 # immunoCyt(TM)
19 m <- list()
20 for (aux in 1:2)
21 {
22   m[[aux]] <- eiras2x2::abcd2matrix(a=a[aux],b=b[aux],c=c[
       aux],d=d[aux],
23            rownames=c("T+","T-"),
24            colnames=c("D+","D-"))$m
25   print(res <- eiras2x2::onesample.Youden(m[[aux]],main=t[
       aux]))
26   # PPV and NPV
27   plot(res, col="forestgreen", lty=1, lwd=2, plot.npv =
       FALSE)
28   plot(res, col="tomato", lty=2, lwd=2, plot.ppv = FALSE,
       add=TRUE,
29     legend.txt=c("PPV","1-NPV"),
30     legend.col=c("forestgreen","tomato"),
31     legend.lwd=c(2,2),
32     legend.lty=c(1,2))
33   # PPV and NPV and hypothetical pre-test value
34   plot(res, col="royalblue",
35     lty=1, lwd=2, curve.lab.at = rep(0.9,2),
36     mcex=1,
37     tp="12", # less intense polygon shadow (use "00" for
           none),
38     digits=2, # default is 3
```

```
39        main=paste0("Performance of a hypothetical\nresult
             for ",t[aux]),
40        xlab="Probability of disease before testing",
41        ylab="Probability of disease after test result",
42        prevalence=0.63
43    )
44  }
45
46  # Comparison
47  res.between <- eiras2x2::twosample.Youden(m1=m[[1]],m2=m
      [[2]],
48           main1=t[1],
49           main2=t[2])
50  print(res.between)
```

2.5.1 output

```
Separated exams:

-----
ELISA
-----

One condition test of Youden index

Data:
       D+  D- Total
T+    182  47   229
T-     45 211   256
Total 227 258   485

Youden (1950)
Youden index = 0.6196
p-value 1.30e-67
95 percent confidence interval:
 0.5608 0.6784

Chen et al. (2015)
Youden index = 0.6196
```

```
p-value 9.38e-102
95 percent confidence interval:
 0.5719 0.6673
-------
ELISPOT
-------

One condition test of Youden index

Data:
       D+  D- Total
T+    192  45   237
T-     35 213   248
Total 227 258   485

Youden (1950)
Youden index = 0.6714
p-value 7.55e-89
95 percent confidence interval:
 0.6160 0.7268

Chen et al. (2015)
Youden index = 0.6714
p-value 2.84e-126
95 percent confidence interval:
 0.6251 0.7177

-----------------------------
Comparison: ELISA vs. ELISPOT
-----------------------------


Within-group test of Youden index

Data:
       D+  D-  Total       D+  D-  Total
 T+    182 47  229   T+    192 45  237
 T-    45  211 256   T-    35  213 248
 Total 227 258 485   Total 227 258 485
```

```
Differences:

* Agreement by Cohen's kappa:
       Condition AgrD+  AgrD-  Diff. J lower   upper  z      p-value
       Kappa     0.8185 0.8677 0.0518  -0.0234 0.1270 1.3507 0.1767948

* Agreement by Holley & Guilford's G:
       Condition AgrD+  AgrD-  Diff. J lower   upper  z      p-value
       G         0.8943 0.9225 0.0518  -0.0231 0.1267 1.3563 0.1750082

* Agreement by Gwet's AC1:
       Condition AgrD+  AgrD-  Diff. J lower   upper  z      p-value
       AC1       0.9255 0.9452 0.0518  -0.0229 0.1265 1.3586 0.1742696
```

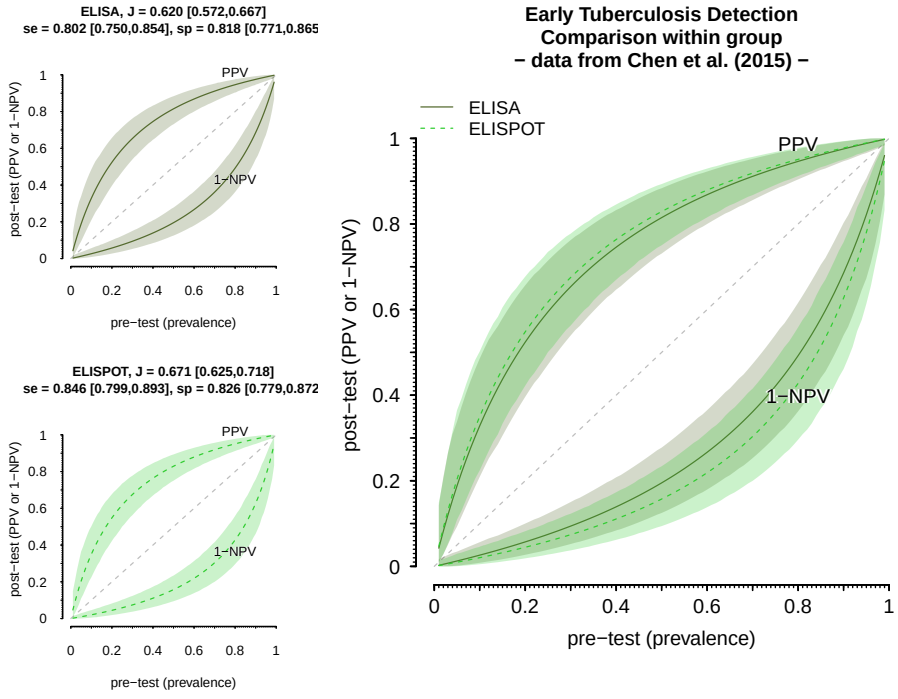

Figure Y-2: Figures obtained with `Chen2015between.R` (not included in the manuscript).

2.6 ErdoganvsReckerOneSample.R

Using the function `eiras2x2::onesample.Youden` to demonstrate that the tests proposed by Erdogan et al. (2020)[1] and Recker et al. (1998)[2] qualify as diagnostic tests, albeit with relatively mediocre performance. It is the source of data for manuscript Tables 4 and 5. This script also creates Figure Y-3, which is not included in the manuscript.

```
# Erdogan 2020 vs Recker 1998
#    D+  D-
# T+  a   b
# T-  c   d

o.par <- par(no.readonly = TRUE)
o.mai <- par()$mai

# provide filenames if you want the output stored in disk
figname1 <- NA # "Fig_ErdoganRecker1.pdf" # isolated curves
usecolor <- TRUE # FALSE for graphs in shades of gray

Dnames1 <- Tnames1 <- Dnames2 <- Tnames2 <- c()
t <- a <- b <- c <- d <- c()

# colors
if(usecolor)
{
  color <- c("gold3","tan3","violetred3","steelblue3","
     tomato1")
} else
{
  gray_palette <- colorRampPalette(c("gray64","gray12"))
  color <- gray_palette(5)
}

# Erdogan, 2020
# lower PSA, cutoff by prostate volume at 43.5 ml
Dnames1[1] <- "PCa+"; Dnames2[1] <- "PCa-"
Tnames1[1] <- "PV < 43.5 ml"; Tnames2[1] <- "PV > 43.5 ml"
t[1] <- "PV cutoff 43.5 ml (PSA 2.5-10.0 ng/ml)"
a[1] <- 29
```

```
b[1] <- 13
c[1] <- 19
d[1] <- 83
# higher PSA, cutoff by prostate volume at 61.5 ml
Dnames1[2] <- "PCa+"; Dnames2[2] <- "PCa+"
Tnames1[2] <- "PV < 61.5ml"; Tnames2[2] <- "PV > 61.5 ml"
t[2] <- "PV cutoff 61.5 ml (PSA 10.1-30.0 ng/ml)"
a[2] <- 21
b[2] <- 10
c[2] <- 5
d[2] <- 31

# Recker, 1998
# total PSA with traditional cutoff
Dnames1[3] <- "PCa+"; Dnames2[3] <- "PCa+"
Tnames1[3] <- "PSA > 4 ng/ml"; Tnames2[3] <- "PSA < 4 ng/ml
    "
t[3] <- "Total PSA cutoff 4 ng/ml"
ac <- 69 # a+c
bd <- 219 # b+d
n <- ac+bd
se <- 0.88
sp <- 0.57
a[3] <- 61
d[3] <- 126
c[3] <- ac-a[3]
b[3] <- bd-d[3]
# free / total PSA, cutoff at 0.20
Dnames1[4] <- "PCa+"; Dnames2[4] <- "PCa+"
Tnames1[4] <- "f/tPSA > 0.20"; Tnames2[4] <- "f/tPSA < 0.20
    "
t[4] <- "f/tPSA cutoff=0.20"
sp2 <- 0.66
d[4] <- round(bd*sp2)
b[4] <- bd-d[4]
a[4] <- a[3]
c[4] <- c[3]
# free / total PSA, cutoff at 0.15
Dnames1[5] <- "PCa+"; Tnames2[5] <- "PCa+"
```

```
Tnames1[5] <- "f/tPSA > 0.15"; Dnames2[5] <- "f/tPSA < 0.15
   "
t[5] <- "f/tPSA cutoff=0.15"
se3 <- 0.70
sp3 <- 0.82
d[5] <- round(bd*sp3)
b[5] <- bd-d[5]
a[5] <- round(ac*se3)
c[5] <- ac-a[5]

if(!is.na(figname1))
{
  pdf(figname1, width=10, height=8)
  mcex <- 1
} else
{
  mcex <- 0.8
}

m <- matrix(data=c(0,1,1,2,2,0,3,3,4,4,5,5),
            nrow=2,ncol=6,byrow=TRUE)
layout(m)

res <- list()
for (i in 1:5)
{
  cat("\n\n",t[i],"\n")
  res[[i]] <- eiras2x2::onesample.Youden(a=a[i],b=b[i],c=c[
     i],d=d[i],
                                rownames = c(Dnames1[i],
                                   Dnames2[i]),
                                colnames = c(Tnames1[i],
                                   Tnames2[i]))
  print(res[[i]])
  # cat("\n"); print(res[[i]]$estimatives_human)
  par(mai=c(o.mai[1],o.mai[2]*1.2,o.mai[3],o.mai[4]))
  plot(res[[i]],main=paste0(t[i],"\n",
                            "J = ",res[[i]]$estimatives_
                               human[12,3],
```

```
                              " [",res[[i]]$estimatives_human
                                 [12,4],", ",
                              res[[i]]$estimatives_human
                                 [12,5],"]"),
       col=color[i], lwd=2, curve.lab.at=c(0.6,0.8), mcex=
          mcex
  )
}
par(o.par)
if(!is.na(figname1)) {dev.off()}
```

2.6.1 output

```
 PV cutoff 43.5 ml (PSA 2.5-10.0 ng/ml)

One condition test of Youden index

Data:
             PCa+ PCa- Total
PV < 43.5 ml   29   13    42
PV > 43.5 ml   19   83   102
Total          48   96   144

Youden (1950)
Youden index = 0.4688
p-value 1.32e-09
95 percent confidence interval:
 0.3392 0.5983

Chen et al. (2015)
Youden index = 0.4688
p-value 1.94e-13
95 percent confidence interval:
 0.3625 0.5750

 PV cutoff 61.5 ml (PSA 10.1-30.0 ng/ml)

One condition test of Youden index
```

```
Data:
             PCa+ PCa+ Total
PV < 61.5ml    21   10    31
PV > 61.5 ml    5   31    36
Total          26   41    67

Youden (1950)
Youden index = 0.5638
p-value 1.80e-08
95 percent confidence interval:
 0.3955 0.7321

Chen et al. (2015)
Youden index = 0.5638
p-value 7.46e-13
95 percent confidence interval:
 0.4327 0.6949

 Total PSA cutoff 4 ng/ml

One condition test of Youden index

Data:
              PCa+ PCa+ Total
PSA > 4 ng/ml   61   93   154
PSA < 4 ng/ml    8  126   134
Total           69  219   288

Youden (1950)
Youden index = 0.4594
p-value 1.05e-19
95 percent confidence interval:
 0.3755 0.5433

Chen et al. (2015)
Youden index = 0.4594
p-value 7.87e-36
95 percent confidence interval:
 0.3987 0.5201
```

```
 f/tPSA cutoff=0.20

One condition test of Youden index

Data:
              PCa+ PCa+ Total
f/tPSA > 0.20   61   74   135
f/tPSA < 0.20    8  145   153
Total           69  219   288

Youden (1950)
Youden index = 0.5462
p-value 5.29e-28
95 percent confidence interval:
 0.4638 0.6285

Chen et al. (2015)
Youden index = 0.5462
p-value 5.75e-47
95 percent confidence interval:
 0.4835 0.6088

 f/tPSA cutoff=0.15

One condition test of Youden index

Data:
              PCa+ f/tPSA < 0.15 Total
f/tPSA > 0.15   48            39    87
PCa+            21           180   201
Total           69           219   288

Youden (1950)
Youden index = 0.5176
p-value 1.26e-17
95 percent confidence interval:
 0.4170 0.6181

Chen et al. (2015)
```

```
Youden index = 0.5176
p-value 1.99e-26
95 percent confidence interval:
 0.4371 0.5981
```

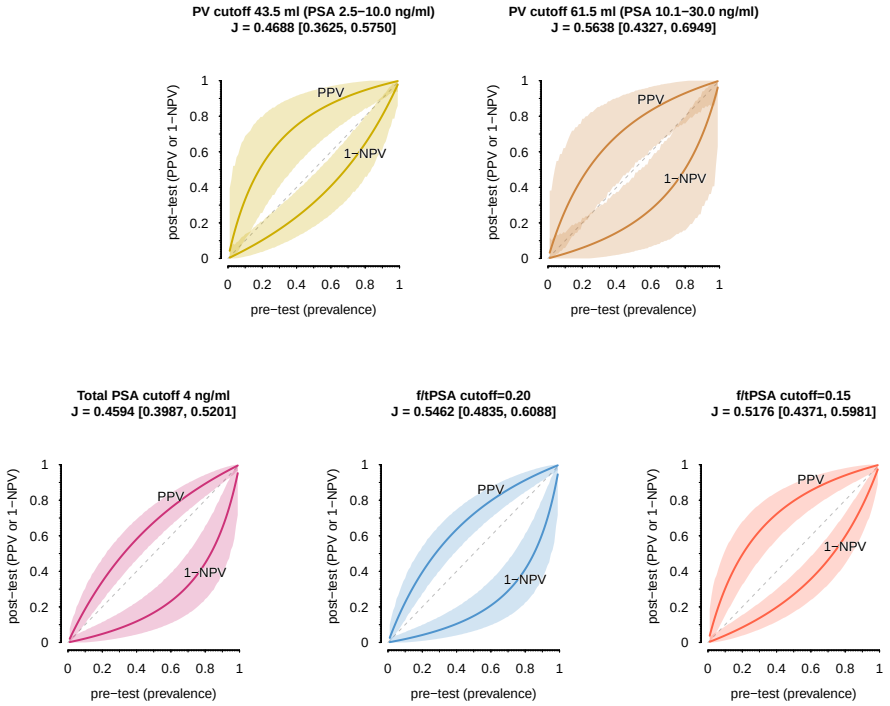

Figure Y-3: Figure obtained with `ErdoganvsReckerOneSample.R`, not included in the manuscript (see section "One sample test of the Youden index" in the manuscript related to the test utility).

2.7 ErdoganvsRecker.R

Figure Y-4 is generated by with `eiras2x2::onesample.Youden` and `eiras2x2::twosample.Youden` functions. It is not presented in the main manuscript, but it is related to manuscript Tables 4, 5, 6, and 7, which compare the results of Erdogan et al. (2020)[1] and Recker et al. (1998)[2].

```
# Erdogan 2020 vs Recker 1998
#    D+  D-
# T+  a   b
# T-  c   d

o.par <- par(no.readonly = TRUE)
o.mai <- par()$mai

figname <- "Fig_ErdoganRecker.pdf"
usecolor <- TRUE # TRUE for color graphs

Dnames1 <- Tnames1 <- Dnames2 <- Tnames2 <- c()
t <- a <- b <- c <- d <- c()

# colors
if(usecolor)
{
  color <- c("gold3","tan3","violetred3","steelblue3","
     tomato1")
} else
{
  gray_palette <- colorRampPalette(c("gray64","gray12"))
  color <- gray_palette(5)
}

# Erdogan, 2019
Dnames1[1] <- "PCa+"; Dnames2[1] <- "PCa-"
Tnames1[1] <- "PV < 43.5 ml"; Tnames2[1] <- "PV > 43.5 ml"
t[1] <- "PV cutoff 43.5 ml (PSA 2.5-10.0 ng/ml)"
a[1] <- 29
b[1] <- 13
c[1] <- 19
d[1] <- 83
```

```
33
34 Dnames1[2] <- "PCa+"; Dnames2[2] <- "PCa+"
35 Tnames1[2] <- "PV < 61.5ml"; Tnames2[2] <- "PV > 61.5 ml"
36 t[2] <- "PV cutoff 61.5 ml (PSA 10.1-30.0 ng/ml)"
37 a[2] <- 21
38 b[2] <- 10
39 c[2] <- 5
40 d[2] <- 31
41
42 # Recker, 1998
43 Dnames1[3] <- "PCa+"; Dnames2[3] <- "PCa+"
44 Tnames1[3] <- "PSA > 4 ng/ml"; Tnames2[3] <- "PSA < 4 ng/ml
      "
45 t[3] <- "Total PSA cutoff 4 ng/ml"
46 ac <- 69 # a+c
47 bd <- 219 # b+d
48 n <- ac+bd
49 se <- 0.88
50 sp <- 0.57
51 a[3] <- 61
52 d[3] <- 126
53 c[3] <- ac-a[3]
54 b[3] <- bd-d[3]
55
56 # free / total PSA
57 Dnames1[4] <- "PCa+"; Dnames2[4] <- "PCa+"
58 Tnames1[4] <- "f/tPSA > 0.20"; Tnames2[4] <- "f/tPSA < 0.20
      "
59 t[4] <- "f/tPSA cutoff=0.20"
60 sp2 <- 0.66
61 d[4] <- round(bd*sp2)
62 b[4] <- bd-d[4]
63 a[4] <- a[3]
64 c[4] <- c[3]
65
66 Dnames1[5] <- "PCa+"; Tnames2[5] <- "PCa+"
67 Tnames1[5] <- "f/tPSA > 0.15"; Dnames2[5] <- "f/tPSA < 0.15
      "
68 t[5] <- "f/tPSA cutoff=0.15"
```

```
se3 <- 0.70
sp3 <- 0.82
d[5] <- round(bd*sp3)
b[5] <- bd-d[5]
a[5] <- round(ac*se3)
c[5] <- ac-a[5]

if(!is.na(figname))
{
  pdf(figname, width=18, height=18)
  mcex <- 1.5
} else
{
  mcex <- 0.6
}

m <- matrix(data=c(2,6,16,16,
                   3,7,8,9,
                   rep(1,4),
                   4,10,11,12,
                   5,13,14,15),
            nrow=5,ncol=4,byrow=TRUE)
layout(m, widths=c(0.7,1.1,1,1), heights = c
   (1.1,1.1,0.1,1.1,1.4))
par(mai=c(0,0,0,0))
plot (NA,xlim=c(0,10),ylim=c(0,3),axes=FALSE)
text (5,1.5,"Erdogan: Prostate volume (PV) vs. Recker: PSA
   or free/total PSA",cex=1.7*mcex)
plot (NA,xlim=c(0,10),ylim=c(0,10), axes=FALSE)
text (5,5,paste0("Erdogan et al. (2020)\n",
                 "Prostate volume\n",
                 "(cutoff at 43.5 ml and 61.5 ml)\n",
                 "in two groups of patients\n",
                 "with serum levels of\n",
                 "2.5-10.0 ng/ml and 10.1-30.0 ng/ml"),
      cex=mcex*1.3)
plot (NA,xlim=c(0,10),ylim=c(0,10), axes=FALSE)
text (5,5,paste0("Recker et al. (1998)\n",
                 "Total PSA with a\n",
```

```
                  "cutoff of 4 ng/ml\n",
                  "and free/total PSA\n",
                  "with cutoffs of\n",
                  "0.20 and 0.15"),
      cex=mcex*1.3)
plot (NA,xlim=c(0,10),ylim=c(0,10), axes=FALSE)
text (5,5,paste0("Erdogan et al. (2020)\n",
                  "Prostate volume (cutoff at 43.5 ml)\n",
                  "in patients with 2.5-10.0 ng/ml\n",
                  "vs.\n",
                  "Recker et al. (1998)\n",
                  "Total PSA (with cutoff of (4 ng/ml)\n",
                  "and free/total PSA with cutoffs\n",
                  "points at 0.20 and 0.15"),
      cex=mcex*1.3)
plot (NA,xlim=c(0,10),ylim=c(0,10), axes=FALSE)
text (5,5,paste0("Erdogan et al. (2020)\n",
                  "Prostate volume (cutoff at 61.5 ml)\n",
                  "in patients with 10.1-30.0.0 ng/ml\n",
                  "vs.\n",
                  "Recker et al. (1998)\n",
                  "Total PSA (with cutoff of (4 ng/ml)\n",
                  "and free/total PSA with cutoffs\n",
                  "points at 0.20 and 0.15"),
      cex=mcex*1.3)
par(mai=o.mai)

res <- list()
for (i in 1:5)
{
  cat("\n\n",t[i],"\n")
  res[[i]] <- eiras2x2::onesample.Youden(a=a[i],b=b[i],c=c[
     i],d=d[i],
        rownames = c(Dnames1[i],Dnames2[i]),
        colnames = c(Tnames1[i],Tnames2[i]))
  print(res[[i]])
}

cat("\n------------\n")
```

```
cat(  "Comparisons:")
cat("\n------------\n")

res.between <- list()
pair <- c(
            1,2, # y
            3,4, # y
            3,5,
            4,5,
            1,3, # y
            1,4,
            1,5,
            2,3, # y
            2,4,
            2,5
)
for (p in seq(1,length(pair),by=2))
{
  label <- paste0(t[pair[p]]," vs. ",t[pair[p+1]])
  cat("\n
     ------------------------------------------------------\
     n")
  cat("\n",t[pair[p]],"x",t[pair[p+1]],"\n")
  # recover 2x2 tables
  m1 <- res[[pair[p]]]$table2x2
  m1 <- m1[1:2,1:2]
  m2 <- res[[pair[p+1]]]$table2x2
  m2 <- m2[1:2,1:2]
  res.between[[label]] <- eiras2x2::twosample.Youden(m1=m1,
     m2=m2,
      main1=t[pair[p]],
      main2=t[pair[p+1]],
      colnames1 = c(Dnames1[pair[p]],Dnames2[pair[p]]),
      rownames1 = c(Tnames1[pair[p]],Tnames2[pair[p]]),
      colnames2 = c(Dnames1[pair[p+1]],Dnames2[pair[p+1]]),
      rownames = c(Tnames1[pair[p+1]],Tnames2[pair[p+1]])
                                            )
  print(res.between[[label]])
  mai <- c(0.8,0.6,0.1,0.1)
```

```
180   if(p==1 | p==3 | p==9 | p==15)
181   {
182     ylab <- "Prob. of Disease"
183     y.axis <- TRUE
184   } else
185   {
186     ylab <- ""
187     y.axis <- FALSE
188     mai[2] <- 0
189   }
190   if(p>=15)
191   {
192     if(p==17)
193     {xlab <- "Prevalence or Pre-test probability"} else {
          xlab<-""}
194     x.axis <- TRUE
195     mai[3] <- 0.3
196   } else
197   {
198     xlab=""
199     x.axis <- FALSE
200     mai[1] <- 0
201   }
202   par(mai=mai)
203
204   plot(res[[pair[p]]], col=color[pair[p]],
205        xlab=xlab, x.axis=x.axis,
206        ylab=ylab, y.axis=y.axis,
207        lty=1, lwd=3, main="",mcex=mcex)
208   plot(res[[pair[p+1]]], col=color[pair[p+1]], lty=2, lwd
        =3,
209        add=TRUE, mcex=mcex,
210        legend.txt = c(t[pair[p]],t[pair[p+1]]),
211        legend.lty = c(1,2),
212        legend.lwd = c(3,3),
213        legend.col = c(color[pair[p]],color[pair[p+1]]),
214        curve.lab.at=c(0.6,0.8))
215 }
216
```

```
par(o.par)
if(!is.na(figname)) {dev.off()}
```

Another alternative using the function `eiras2x2::twosample.Youden` to compare the proposals of Erdogan et al. (2020)[1] and Recker et al. (1998)[2], showing that there are no significant differences between the tests. This script also creates Figure Y-5, which was not included in the manuscript.

```
# Erdogan 2020 vs Recker 1998
#    D+  D-
# T+  a  b
# T-  c  d

o.par <- par(no.readonly = TRUE)
o.mai <- par()$mai

# provide filenames if you want the output stored in disk
figname2 <- NA # "Fig_ErdoganRecker2.pdf" # curves in pair
usecolor <- TRUE # TRUE for color graphs

Dnames1 <- Tnames1 <- Dnames2 <- Tnames2 <- c()
t <- a <- b <- c <- d <- c()

# colors
if(usecolor)
{
  color <- c("gold3","tan3","violetred3","steelblue3","
     tomato1")
} else
{
  gray_palette <- colorRampPalette(c("gray64","gray12"))
  color <- gray_palette(5)
}

# Erdogan, 2019
# lower PSA, cutoff by prostate volume at 43.5 ml
Dnames1[1] <- "PCa+"; Dnames2[1] <- "PCa-"
Tnames1[1] <- "PV < 43.5 ml"; Tnames2[1] <- "PV > 43.5 ml"
```

```
t[1] <- "PV cutoff 43.5 ml (PSA 2.5-10.0 ng/ml)"
a[1] <- 29
b[1] <- 13
c[1] <- 19
d[1] <- 83
# higher PSA, cutoff by prostate volume at 61.5 ml
Dnames1[2] <- "PCa+"; Dnames2[2] <- "PCa+"
Tnames1[2] <- "PV < 61.5ml"; Tnames2[2] <- "PV > 61.5 ml"
t[2] <- "PV cutoff 61.5 ml (PSA 10.1-30.0 ng/ml)"
a[2] <- 21
b[2] <- 10
c[2] <- 5
d[2] <- 31

# Recker, 1998
# total PSA with traditional cutoff
Dnames1[3] <- "PCa+"; Dnames2[3] <- "PCa+"
Tnames1[3] <- "PSA > 4 ng/ml"; Tnames2[3] <- "PSA < 4 ng/ml
    "
t[3] <- "Total PSA cutoff 4 ng/ml"
ac <- 69 # a+c
bd <- 219 # b+d
n <- ac+bd
se <- 0.88
sp <- 0.57
a[3] <- 61
d[3] <- 126
c[3] <- ac-a[3]
b[3] <- bd-d[3]
# free / total PSA, cutoff at 0.20
Dnames1[4] <- "PCa+"; Dnames2[4] <- "PCa+"
Tnames1[4] <- "f/tPSA > 0.20"; Tnames2[4] <- "f/tPSA < 0.20
    "
t[4] <- "f/tPSA cutoff=0.20"
sp2 <- 0.66
d[4] <- round(bd*sp2)
b[4] <- bd-d[4]
a[4] <- a[3]
c[4] <- c[3]
```

```
67  # free / total PSA, cutoff at 0.15
68  Dnames1[5] <- "PCa+"; Tnames2[5] <- "PCa+"
69  Tnames1[5] <- "f/tPSA > 0.15"; Dnames2[5] <- "f/tPSA < 0.15
        "
70  t[5] <- "f/tPSA cutoff=0.15"
71  se3 <- 0.70
72  sp3 <- 0.82
73  d[5] <- round(bd*sp3)
74  b[5] <- bd-d[5]
75  a[5] <- round(ac*se3)
76  c[5] <- ac-a[5]
77
78  cat("\n------------\n")
79  cat(  "Comparisons:")
80  cat("\n------------\n")
81
82  if(!is.na(figname2))
83  {
84    pdf(figname2, width=10.4, height=14.7)
85    mcex <- 1.3
86  } else
87  {
88    mcex <- 0.8
89  }
90
91  # testing onesample.Youden
92  res <- list()
93  for (i in 1:5)
94  {
95    res[[i]] <- eiras2x2::onesample.Youden(a=a[i],b=b[i],c=c[
          i],d=d[i],
96      rownames = c(Dnames1[i],Dnames2[i]),
97      colnames = c(Tnames1[i],Tnames2[i]))
98  }
99
100
101 m <- matrix(data=c(1,1,2:7),
102             nrow=4,ncol=2,byrow=TRUE)
103 layout(m, heights = c(0.1,1,1,1))
```

```
104 par(mai=c(0,0,0,0))
105 plot (NA,xlim=c(0,10),ylim=c(0,1.5),axes=FALSE)
106 text (5,0.5,"Prostate volume (PV) vs. PSA or free/total PSA
       ",cex=2*mcex)
107 par(mai=o.mai)
108 res.between <- list()
109 pair <- c(
110   1,3,
111   2,3,
112   1,4,
113   2,4,
114   1,5,
115   2,5
116 )
117 for (p in seq(1,length(pair),by=2))
118 {
119   label <- paste0(t[pair[p]]," vs. ",t[pair[p+1]])
120   cat("\n
         ------------------------------------------------------\
         n")
121   cat("\n",t[pair[p]],"x",t[pair[p+1]],"\n")
122   # recover 2x2 tables
123   m1 <- res[[pair[p]]]$table2x2
124   m1 <- m1[1:2,1:2]
125   m2 <- res[[pair[p+1]]]$table2x2
126   m2 <- m2[1:2,1:2]
127   res.between[[label]] <- eiras2x2::twosample.Youden(m1=m1,
         m2=m2,
128     main1=t[pair[p]],
129     main2=t[pair[p+1]],
130     colnames1 = c(Dnames1[pair[p]],Dnames2[pair[p]]),
131     rownames1 = c(Tnames1[pair[p]],Tnames2[pair[p]]),
132     colnames2 = c(Dnames1[pair[p+1]],Dnames2[pair[p+1]]),
133     rownames = c(Tnames1[pair[p+1]],Tnames2[pair[p+1]])
134   )
135   print(res.between[[label]])
136   par(mai=c(o.mai[1],o.mai[2]*1.2,0,o.mai[4]))
137   plot(res[[pair[p]]], col=color[pair[p]], lty=1, lwd=2,
         main="",mcex=mcex)
```

```
138   plot(res[[pair[p+1]]], col=color[pair[p+1]], lty=2, lwd
         =2,
139         add=TRUE, mcex=mcex,
140         legend.txt = c(t[pair[p]],t[pair[p+1]]),
141         legend.lty = c(1,2),
142         legend.lwd = c(2,2),
143         legend.col = c(color[pair[p]],color[pair[p+1]]),
144         curve.lab.at=c(0.6,0.8))
145 }
146
147 par(o.par)
148 if(!is.na(figname2)) {dev.off()}
```

2.7.1 output

```
------------
Comparisons:
------------

-----------------------------------------------------

 PV cutoff 43.5 ml (PSA 2.5-10.0 ng/ml) x Total PSA cutoff 4 ng/ml

Separated exams:

--------------------------------------
PV cutoff 43.5 ml (PSA 2.5-10.0 ng/ml)
--------------------------------------

One condition test of Youden index

Data:
             PCa+ PCa- Total
PV < 43.5 ml   29   13    42
PV > 43.5 ml   19   83   102
Total          48   96   144

Youden (1950)
```

```
Youden index = 0.4688
p-value 1.32e-09
95 percent confidence interval:
 0.3392 0.5983

Chen et al. (2015)
Youden index = 0.4688
p-value 1.94e-13
95 percent confidence interval:
 0.3625 0.5750
------------------------
Total PSA cutoff 4 ng/ml
------------------------

One condition test of Youden index

Data:
              PCa+ PCa+ Total
PSA > 4 ng/ml   61   93   154
PSA < 4 ng/ml    8  126   134
Total           69  219   288

Youden (1950)
Youden index = 0.4594
p-value 1.05e-19
95 percent confidence interval:
 0.3755 0.5433

Chen et al. (2015)
Youden index = 0.4594
p-value 7.87e-36
95 percent confidence interval:
 0.3987 0.5201

--------------------------------------------------------------------------------
Comparison: PV cutoff 43.5 ml (PSA 2.5-10.0 ng/ml) vs. Total PSA cutoff 4 ng/ml
--------------------------------------------------------------------------------
```

```
Between-groups test of Youden index

Data:
              PCa+ PCa- Total               PCa+ PCa+ Total
 PV < 43.5 ml 29   13   42    PSA > 4 ng/ml 61   93   154
 PV > 43.5 ml 19   83   102   PSA < 4 ng/ml 8    126  134
 Total        48   96   144   Total         69   219  288


Differences:

Youden (1950)

J1 = 0.4688
J2 = 0.4594
Differ (J2-J1) = -0.0093
z = -0.0996, p-value = 0.9206
95 percent confidence interval (difference):
 -0.1932 0.1745

Chen et al. (2015)

J1 = 0.4688
J2 = 0.4594
Differ (J2-J1) = -0.0093
z = -0.1257, p-value = 0.9000
95 percent confidence interval (difference):
 -0.1551 0.1364
-------------------------------------------------------

 PV cutoff 61.5 ml (PSA 10.1-30.0 ng/ml) x Total PSA cutoff 4 ng/ml

Separated exams:

---------------------------------------
PV cutoff 61.5 ml (PSA 10.1-30.0 ng/ml)
---------------------------------------

One condition test of Youden index
```

```
Data:
             PCa+ PCa+ Total
PV < 61.5ml    21   10    31
PV > 61.5 ml    5   31    36
Total          26   41    67

Youden (1950)
Youden index = 0.5638
p-value 1.80e-08
95 percent confidence interval:
 0.3955 0.7321

Chen et al. (2015)
Youden index = 0.5638
p-value 7.46e-13
95 percent confidence interval:
 0.4327 0.6949
------------------------
Total PSA cutoff 4 ng/ml
------------------------

One condition test of Youden index

Data:
              PCa+ PCa+ Total
PSA > 4 ng/ml   61   93   154
PSA < 4 ng/ml    8  126   134
Total           69  219   288

Youden (1950)
Youden index = 0.4594
p-value 1.05e-19
95 percent confidence interval:
 0.3755 0.5433

Chen et al. (2015)
Youden index = 0.4594
p-value 7.87e-36
95 percent confidence interval:
```

```
 0.3987 0.5201

------------------------------------------------------------------------------
Comparison: PV cutoff 61.5 ml (PSA 10.1-30.0 ng/ml) vs. Total PSA cutoff 4 ng/ml
------------------------------------------------------------------------------


Between-groups test of Youden index

Data:
              PCa+ PCa+ Total               PCa+ PCa+ Total
 PV < 61.5ml  21   10   31    PSA > 4 ng/ml 61   93   154
 PV > 61.5 ml 5    31   36    PSA < 4 ng/ml 8    126  134
 Total        26   41   67    Total         69   219  288


Differences:

Youden (1950)

J1 = 0.5638
J2 = 0.4594
Differ (J2-J1) = -0.1044
z = -0.9130, p-value = 0.3612
95 percent confidence interval (difference):
 -0.3285 0.1197

Chen et al. (2015)

J1 = 0.5638
J2 = 0.4594
Differ (J2-J1) = -0.1044
z = -1.1886, p-value = 0.2346
95 percent confidence interval (difference):
 -0.2765 0.0677
------------------------------------------------------

 PV cutoff 43.5 ml (PSA 2.5-10.0 ng/ml) x f/tPSA cutoff=0.20
```

```
Separated exams:

--------------------------------------
PV cutoff 43.5 ml (PSA 2.5-10.0 ng/ml)
--------------------------------------

One condition test of Youden index

Data:
             PCa+ PCa- Total
PV < 43.5 ml   29   13    42
PV > 43.5 ml   19   83   102
Total          48   96   144

Youden (1950)
Youden index = 0.4688
p-value 1.32e-09
95 percent confidence interval:
 0.3392 0.5983

Chen et al. (2015)
Youden index = 0.4688
p-value 1.94e-13
95 percent confidence interval:
 0.3625 0.5750
------------------
f/tPSA cutoff=0.20
------------------

One condition test of Youden index

Data:
              PCa+ PCa+ Total
f/tPSA > 0.20   61   74   135
f/tPSA < 0.20    8  145   153
Total           69  219   288

Youden (1950)
Youden index = 0.5462
```

```
p-value 5.29e-28
95 percent confidence interval:
 0.4638 0.6285

Chen et al. (2015)
Youden index = 0.5462
p-value 5.75e-47
95 percent confidence interval:
 0.4835 0.6088

------------------------------------------------------------------------
Comparison: PV cutoff 43.5 ml (PSA 2.5-10.0 ng/ml) vs. f/tPSA cutoff=0.20
------------------------------------------------------------------------


Between-groups test of Youden index

Data:
              PCa+ PCa- Total               PCa+ PCa+ Total
 PV < 43.5 ml 29   13   42    f/tPSA > 0.20 61   74   135
 PV > 43.5 ml 19   83   102   f/tPSA < 0.20 8    145  153
 Total        48   96   144   Total         69   219  288


Differences:

Youden (1950)

J1 = 0.4688
J2 = 0.5462
Differ (J2-J1) = 0.0774
z = 0.8295, p-value = 0.4068
95 percent confidence interval (difference):
 -0.1055 0.2603

Chen et al. (2015)

J1 = 0.4688
J2 = 0.5462
```

```
Differ (J2-J1) = 0.0774
z = 1.0327, p-value = 0.3018
95 percent confidence interval (difference):
 -0.0695 0.2243
-------------------------------------------------------

 PV cutoff 61.5 ml (PSA 10.1-30.0 ng/ml) x f/tPSA cutoff=0.20

Separated exams:

---------------------------------------
PV cutoff 61.5 ml (PSA 10.1-30.0 ng/ml)
---------------------------------------

One condition test of Youden index

Data:
             PCa+ PCa+ Total
PV < 61.5ml    21   10    31
PV > 61.5 ml    5   31    36
Total          26   41    67

Youden (1950)
Youden index = 0.5638
p-value 1.80e-08
95 percent confidence interval:
 0.3955 0.7321

Chen et al. (2015)
Youden index = 0.5638
p-value 7.46e-13
95 percent confidence interval:
 0.4327 0.6949
------------------
f/tPSA cutoff=0.20
------------------

One condition test of Youden index
```

```
Data:
              PCa+ PCa+ Total
f/tPSA > 0.20   61   74   135
f/tPSA < 0.20    8  145   153
Total           69  219   288

Youden (1950)
Youden index = 0.5462
p-value 5.29e-28
95 percent confidence interval:
 0.4638 0.6285

Chen et al. (2015)
Youden index = 0.5462
p-value 5.75e-47
95 percent confidence interval:
 0.4835 0.6088

------------------------------------------------------------------------------
Comparison: PV cutoff 61.5 ml (PSA 10.1-30.0 ng/ml) vs. f/tPSA cutoff=0.20
------------------------------------------------------------------------------


Between-groups test of Youden index

Data:
               PCa+ PCa+ Total                PCa+ PCa+ Total
 PV < 61.5ml   21   10   31    f/tPSA > 0.20 61   74   135
 PV > 61.5 ml  5    31   36    f/tPSA < 0.20 8    145  153
 Total         26   41   67    Total         69   219  288


Differences:

Youden (1950)

J1 = 0.5638
J2 = 0.5462
Differ (J2-J1) = -0.0176
```

```
z = -0.1548, p-value = 0.8770
95 percent confidence interval (difference):
 -0.2409 0.2057

Chen et al. (2015)

J1 = 0.5638
J2 = 0.5462
Differ (J2-J1) = -0.0176
z = -0.1996, p-value = 0.8418
95 percent confidence interval (difference):
 -0.1907 0.1555
-------------------------------------------------------

 PV cutoff 43.5 ml (PSA 2.5-10.0 ng/ml) x f/tPSA cutoff=0.15

Separated exams:

---------------------------------------
PV cutoff 43.5 ml (PSA 2.5-10.0 ng/ml)
---------------------------------------

One condition test of Youden index

Data:
             PCa+ PCa- Total
PV < 43.5 ml   29   13    42
PV > 43.5 ml   19   83   102
Total          48   96   144

Youden (1950)
Youden index = 0.4688
p-value 1.32e-09
95 percent confidence interval:
 0.3392 0.5983

Chen et al. (2015)
Youden index = 0.4688
p-value 1.94e-13
```

```
95 percent confidence interval:
 0.3625 0.5750
------------------
f/tPSA cutoff=0.15
------------------

One condition test of Youden index

Data:
              PCa+ f/tPSA < 0.15 Total
f/tPSA > 0.15   48            39    87
PCa+            21           180   201
Total           69           219   288

Youden (1950)
Youden index = 0.5176
p-value 1.26e-17
95 percent confidence interval:
 0.4170 0.6181

Chen et al. (2015)
Youden index = 0.5176
p-value 1.99e-26
95 percent confidence interval:
 0.4371 0.5981

------------------------------------------------------------------------
Comparison: PV cutoff 43.5 ml (PSA 2.5-10.0 ng/ml) vs. f/tPSA cutoff=0.15
------------------------------------------------------------------------


Between-groups test of Youden index

Data:
              PCa+ PCa- Total               PCa+ f/tPSA < 0.15 Total
 PV < 43.5 ml 29   13   42    f/tPSA > 0.15 48   39            87
 PV > 43.5 ml 19   83   102   PCa+          21   180           201
 Total        48   96   144   Total         69   219           288
```

```
Differences:

Youden (1950)

J1 = 0.4688
J2 = 0.5176
Differ (J2-J1) = 0.0488
z = 0.4897, p-value = 0.6243
95 percent confidence interval (difference):
 -0.1466 0.2442

Chen et al. (2015)

J1 = 0.4688
J2 = 0.5176
Differ (J2-J1) = 0.0488
z = 0.6025, p-value = 0.5468
95 percent confidence interval (difference):
 -0.1100 0.2076
-------------------------------------------------------

 PV cutoff 61.5 ml (PSA 10.1-30.0 ng/ml) x f/tPSA cutoff=0.15

Separated exams:

---------------------------------------
PV cutoff 61.5 ml (PSA 10.1-30.0 ng/ml)
---------------------------------------

One condition test of Youden index

Data:
             PCa+ PCa+ Total
PV < 61.5ml    21   10    31
PV > 61.5 ml    5   31    36
Total          26   41    67

Youden (1950)
Youden index = 0.5638
```

```
p-value 1.80e-08
95 percent confidence interval:
 0.3955 0.7321

Chen et al. (2015)
Youden index = 0.5638
p-value 7.46e-13
95 percent confidence interval:
 0.4327 0.6949
------------------
f/tPSA cutoff=0.15
------------------

One condition test of Youden index

Data:
              PCa+ f/tPSA < 0.15 Total
f/tPSA > 0.15   48            39    87
PCa+            21           180   201
Total           69           219   288

Youden (1950)
Youden index = 0.5176
p-value 1.26e-17
95 percent confidence interval:
 0.4170 0.6181

Chen et al. (2015)
Youden index = 0.5176
p-value 1.99e-26
95 percent confidence interval:
 0.4371 0.5981

------------------------------------------------------------------------
Comparison: PV cutoff 61.5 ml (PSA 10.1-30.0 ng/ml) vs. f/tPSA cutoff=0.15
------------------------------------------------------------------------


Between-groups test of Youden index
```

```
Data:
              PCa+ PCa+ Total               PCa+ f/tPSA < 0.15 Total
 PV < 61.5ml  21   10   31    f/tPSA > 0.15 48   39            87
 PV > 61.5 ml 5    31   36    PCa+          21   180           201
 Total        26   41   67    Total         69   219           288


Differences:

Youden (1950)

J1 = 0.5638
J2 = 0.5176
Differ (J2-J1) = -0.0462
z = -0.3877, p-value = 0.6982
95 percent confidence interval (difference):
 -0.2798 0.1874

Chen et al. (2015)

J1 = 0.5638
J2 = 0.5176
Differ (J2-J1) = -0.0462
z = -0.4942, p-value = 0.6211
95 percent confidence interval (difference):
 -0.2295 0.1371
```

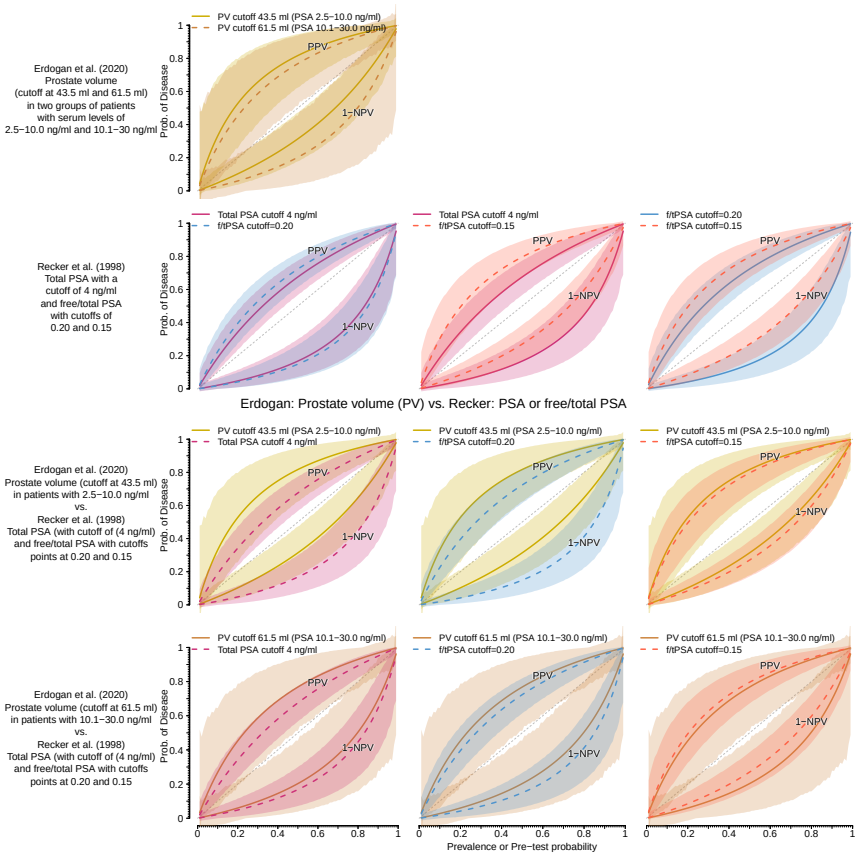

Figure Y-4: Exam comparisons. First row: comparison of prostate volume (PV) with cutoff values at 43.5 ml and 61.5 ml in patients with serum PSA levels of 2.5–10.0 ng/ml and 10.1–30.0 ng/ml (Erdogan et al., 2020). Second row: comparison of versus total PSA with a cutoff of 4 ng/ml, and free/total PSA ratios with cutoff values of 0.20 and 0.15 (Recker et al., 1998). Third and forth rows: comparison of two conditions of Erdogan et al (2020) and three conditions of Recker et al. (1998). The plots depict the probability of disease (PPV and 1-NPV) in function of prevalence or pre-test probability. The shadows correspond to overlaping confidence intervals of PPV and 1-NPV, which corresponds to the lack of evidence of difference between all attempts to improve prostate cancer diagnosis.

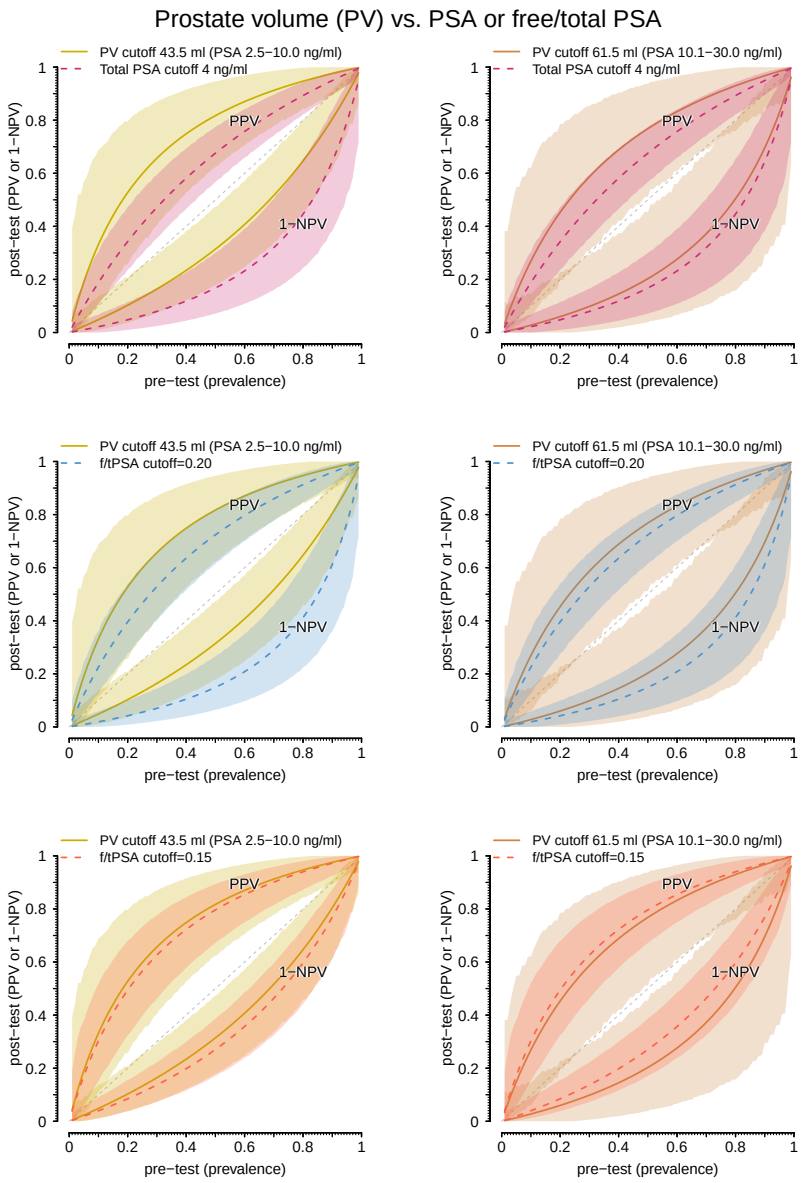

Figure Y-5: The figure was obtained using `ErdoganvsReckerComparison.R` (see section "Independent samples test of Youden indices (between groups)" in the manuscript). The left and right columns represent the proposition by Erdogan et al. (2020)[1], where prostate volume is used as a diagnostic tool with different cut-off points for patients with varying PSA levels. The lines represent the approach by Recker et al. (1998)[2], using total PSA or two cut-off points for free-to-total PSA. It can be observed that at least one hypothetical line for PPV and $1 - NPV$ can be drawn across the common area shared by both tests. This graphical representation indicates the absence of a significant difference between the tests.

References

[1] Erdogan A, Polat S, Keskin E, Turan A. Is prostate volume better than PSA density and free/total PSA ratio in predicting prostate cancer in patients with PSA 2.5–10 ng/mL and 10.1–30 ng/mL? Aging Male. 2020;23(1).

[2] Recker F, Kwiatkowski MK, Piironen T, Pettersson K, Goepel M, Tscholl R. Free-to-total prostate-specific antigen (PSA) ratio improves the specificity for detecting prostate cancer in patients with prostatism and intermediate PSA levels. British Journal of Urology. 1998;81(4).

[3] Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. American Journal of Epidemiology. 1978;107(1).

[4] Li J, Fine JP. Assessing the dependence of sensitivity and specificity on prevalence in meta-analysis. Biostatistics. 2011;12(4).

[5] Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1).

[6] Tadeusz R O, Tadeusz O. The Basic Four Measures and their Derivates in Dichotomous Diagnostic Tests. International Journal of Clinical Biostatistics and Biometrics. 2020;6(1).

[7] Haimes YY. Risk Modeling, Assessment, and Management, Third Edition; 2008.

[8] Chen F, Xue Y, Tan MT, Chen P. Efficient statistical tests to compare Youden index: Accounting for contingency correlation. Statistics in Medicine. 2015;34(9).

[9] Silveira PSP, Siqueira JO. Better to be in agreement than in bad company: A critical analysis of many kappa-like tests. Behavior Research Methods. 2023;55(7).

3 Replication of Manuscript Tables

Tables

Table 1: D_+ and D_- represent the presence or absence of a disease or patient condition, respectively. T_+ and T_- denote the occurrence of a positive or negative result from any symptom, signal, or test. Along the main diagonal, *a* and *D* indicate the counts or proportions of agreement between *D* and *T*, while along the secondary diagonal, *b* and *c* indicate the counts or proportions of disagreement between *D* and *T*.

	D_+	D_-	total
T_+	*a* true positive	*b* false positive	*a+b*
T_-	*c* false negative	*D* true negative	*c+d*
total	*a+c*	*b+d*	*a+b+c+d*

Table 2: Reproduction of Chen et al. (2015) data for the estimation of exams for bladder carcinoma: immunoCyt™ and Cytology. For each test, sensitivity (*se*), specificity (*sp*), Youden index (*J*), and p value were computed using the R function *onesample.Youden* testing $H_0 : J \leq 0$. For the difference between tests, 95% confidence interval was computed by the R function *twosample.Youden* testing $H_0 : J_2 - J_1 = 0$.

	immunoCyt™		**Cytology**	
	D_+	D_-	D_+	D_-
T_+	85	51	41	12
T_-	16	182	13	24
se	84.16 [77.04, 91.28]		75.93 [64.52, 87.33]	
sp	78.11 [72.80, 83.42]		66.67 [51.27, 82.07]	
J (Youden 1950)	0.6227 [0.5482, 0.6972]		0.4259 [0.2651, 0.5867]	
	$p = 2.91 \cdot 10^{-43}$		$p = 6.61 \cdot 10^{-6}$	
J (Chen 2015)	0.6227 [0.5628, 0.6826]		0.4259 [0.3089, 0.5430]	
	$p = 6.67 \cdot 10^{-66}$		$p = 1.08 \cdot 10^{-9}$	
J_2-J_1 (Youden 1950)	-0.1968 [-0.4080, 0.0144]		$p = 0.06784$	
J_2-J_1 (Chen 2015)	-0.1968 [-0.3534, -0.0401]		$p = 0.01382$	

Table 3: Reproduction of Chen et al. (2015) data for the estimation of exams for tuberculosis detection in a within-group design: all patients were tested with ELISA and ELISPOT methods. For each test, sensitivity (*se*), specificity (*sp*), Youden index (*J*), and p value were computed using the R function *onesample.Youden* testing H_0: $J < 0$. For the difference between tests, confidence interval was computed by the R function *twosample.Youden* testing H_0: $J_2 - J_1 = 0$.*

	ELISA			**ELISPOT**		
	D_{1+}	D_{1-}		D_{2+}	D_{2-}	
T_{1+}	182	47	T_{2+}	192	45	
T_{1-}	45	211	T_{2-}	35	213	
Diseased group:			**Control group:**			
	ELISA			**ELISA**		
ELISPOT	T_{1+}	T_{1-}	Total	T_{1-}	T_{1+}	Total
T_{2+}	181	11	192	T_{2-}	207	6	213
T_{2-}	1	34	35	T_{2+}	4	41	45
Total	182	45	227	Total	211	47	258
se	80.18 [74.99, 85.36]			84.58 [79.88, 89.28]		
sp	81.78 [77.07, 86.49]			82.56 [77.93, 87.19]		
J	0.6196 [0.5719, 0.6673]			0.6714 [0.6251, 0.7177]		
	$p = 9.38 \cdot 10^{-102}$			$p = 2.84 \cdot 10^{-126}$		
Agreement (AC_1)	0.926			0.945		
Diff. *J*	0.05180 [-0.02293, 0.12654]					
	$p = 0.1743$					

* Boxes are one option that can provide enough information to recover the partition between diseased and control groups: marginals are transported from the original tables. The number of patients with positive or negative results in both methods allows the reconstruction of 2x2 tables.

Table 4: Analysis of Erdogan et al. (2020) data for the estimation of prostate volume (PV) as a predictor of prostate cancer (PCa). For each test, sensitivity (*se*), specificity (*sp*), Youden index (*J*), and p value were computed using the R function *onesample.Youden* testing H_0: $J \leq 0$. For the difference between tests, confidence interval was computed by the R function *twosample.Youden* testing H_0: J_2-$J_1 = 0$.

	PSA 2.5-10.0 ng/ml			**PSA 10.1-30.0 ng/ml**	
	PCa_+	PCa_-		PCa_+	PCa_-
PV < 43.5 ml	29	13	*PV < 61.5 ml*	21	10
PV > 43.5 ml	19	83	*PV > 61.5 ml*	5	31
se	60.42 [46.58, 74.25]			80.77 [65.62, 95.92]	
sp	86.46 [79.61, 93.30]			75.61 [62.46, 88.75]	
J	0.4688 [0.3625, 0.5750]			0.5638 [0.4327, 0.6949]	
	$p = 1.94 \cdot 10^{-13}$			$p = 7.46 \cdot 10^{-13}$	
	J_2-J_1 0.0950 [-0.1060, 0.2961], $p = 0.3541$				

Table 5: Reproduction of Recker et al. (1998) data for the estimation of Prostate-Specific Antigen (PSA) tests: total PSA (tPSA), free to total PSA (f/tPSA020) with a cut-off point at 0.20, and free to total PSA with a cut-off point at 0.15 (f/tPSA015). Patients were diagnosed with prostate cancer (D_+) or benign prostatic hyperplasia (D_-), and the PSA tests could result in positive (T_+) or negative (T_-) outcomes. Sensitivity (*se*), specificity (*sp*), Youden index (J), and p value were computed using the R function *onesample.Youden* testing H_0: $J \leq 0$.

	tPSA		**f/tPSA020**		**f/tPSA015**	
	D_+	D_-	D_+	D_-	D_+	D_-
T_+	61	93	61	74	48	39
T_-	8	126	8	145	21	180
se	88.41 [80.85, 95.96]		88.41 [80.85, 95.96]		69.57 [58.71, 80.42]	
sp	57.53 [50.99, 64.08]		66.21 [59.95, 72.47]		82.19 [77.12, 87.26]	
J	0.4594 [0.3987, 0.5201]		0.5462 [0.4835, 0.6088]		0.5176 [0.4371, 0.5981]	
	$p = 7.87 \cdot 10^{-36}$		$p = 5.75 \cdot 10^{-47}$		$p = 1.99 \cdot 10^{-26}$	

Table 6: Performance difference (within-group design) with data reproduced from Recker et al. (1998) evaluated by the Youden index (*J*) of three Prostate-Specific Antigen (PSA) tests: total PSA (tPSA), free to total PSA (f/tPSA020) with a cut-off point at 0.20, and free to total PSA with a cut-off point at 0.15 (f/tPSA015). 95% confidence intervals and *p* values were computed using the R function *twosample.Youden* testing H_0: J_2-J_1 = 0. GD_+ and GD_- are agreement estimates obtained from Gwet's AC_1 (see text for explanation).

			tPSA *J* = 0.45940	**f/tPSA 0.20** *J* = 0.54616
f/tPSA 0.20	*J* = 0.54616	*min*(GD_+)	0.7080	
		min(GD_-)	-0.4440	
		Diff. *J*	0.087 [-0.015, 0.188]	
			p = 0.0945	
		min(GD_+)	0.7080	
		max(GD_-)	0.8360	
		Diff. *J*	0.087 [-0.010, 0.183]	
			p = 0.0775	
		max(GD_+)	1.0000	
		min(GD_-)	-0.4440	
		Diff. *J*	0.087 [-0.013, 0.187]	
			p = 0.0890	
		max(GD_+)	1.0000	
		max(GD_-)	0.8360	
		Diff. *J*	0.087 [-0.008, 0.181]	
			p = 0.0724	
f/tPSA 0.15	*J* = 0.51757	*min*(GD_+)	0.3710	0.3710
		min(GD_-)	-0.0410	0.1640
		Diff. *J*	0.058 [-0.059, 0.175]	-0.029 [-0.147, 0.089]
			p = 0.3307	*p* = 0.6346
		min(GD_+)	0.3710	0.3710
		max(GD_-)	0.5740	0.7410
		Diff. *J*	0.058 [-0.057, 0.173]	-0.029 [-0.145, 0.087]
			p = 0.3221	*p* = 0.6292
		max(GD_+)	0.7180	0.7180
		min(GD_-)	-0.0410	0.1640
		Diff. *J*	0.058 [-0.056, 0.172]	-0.029 [-0.144, 0.086]
			p = 0.3186	*p* = 0.6261
		max(GD_+)	0.7180	0.7180
		max(GD_-)	0.5740	0.7410
		Diff. *J*	0.058 [-0.054, 0.170]	-0.029 [-0.142, 0.085]
			p = 0.3097	*p* = 0.6204

Table 7: Performance difference (between-group design) with data reproduced from Erdogan et al. (2020) and Recker et al. (1998) evaluated by the Youden index (*J*). Erdogan proposed the prostate volume (PV) as a predictor of cancer in two groups of patients with two different cut-off points (PSA 2.5-10 ng/ml with cutoff of PV = 43.5 ml; PSA 10.1-30.0 ng/ml with cutoff of PV = 63.5 ml). Recker applied three Prostate-Specific Antigen (PSA) tests: total PSA (tPSA), free to total PSA (f/tPSA 0.20) with a cut-off point at 0.20, and free to total PSA with a cut-off point at 0.15 (f/tPSA 0.15). 95% confidence intervals of *J* and *p* values were computed using the R function *twosample.Youden* testing H_0: J_2-J_1 = 0.

			PV 43.5 ml J_1 = 0.46875	**PV 61.5 ml** J_1 = 0.56379
tPSA	J_2 = 0.45940	J_2 - J_1	-0.00935 [-0.15514, 0.13644]	-0.10439 [-0.27652, 0.06774]
			p = 0.9000	p = 0.2346
f/tPSA 0.20	J_2 = 0.54616	J_2 - J_1	0.07741 [-0.06951, 0.22433]	-0.01763 [-0.19072, 0.15546]
			p = 0.3018	p = 0.8418
f/tPSA 0.15	J_2 = 0.51757	J_2 - J_1	0.04882 [-0.10999, 0.20763]	-0.04622 [-0.22951, 0.13707]
			p = 0.5468	p = 0.6211

Table 8: Comparison of Recker et al.'s total PSA (reference) with proposed biomarkers. Statistical significance is highlighted with a gray background. When a proposed biomarker qualifies as a diagnostic test (95% confidence interval for J is entirely above zero), advance is considered if the difference from the reference (*Diff. J*) is also significantly greater than zero.

a	b	c	d	*J*		*Diff.* $J = J_{\text{ref}} - J$	
Recker et al., 1998, tPSA				**(J_{ref})**			
61	93	8	126	0.4594 [0.3987,0.5201]	$p = 7.87 \cdot 10^{-36}$	**reference**	
Deras et al., 2008, PCA3 overall							
112	92	96	262	0.2786 [0.2317,0.3254]	$p = 7.14 \cdot 10^{-23}$	-0.1808 [-0.2722,-0.0894]	$p = 1.06 \cdot 10^{-04}$
PCA3 with PSA $<$ 4							
17	22	17	75	0.2732 [0.1587,0.3877]	$p = 4.36 \cdot 10^{-05}$	-0.1862 [-0.3407,-0.0317]	$p = 0.0181$
PCA3 with 4 $<$ PSA $<$ 10							
69	62	62	153	0.2383 [0.1807,0.2960]	$p = 5.22 \cdot 10^{-12}$	-0.2211 [-0.3208,-0.1213]	$p = 1.41 \cdot 10^{-05}$
PCA3 with PSA $>$ 10							
25	9	16	36	0.4098 [0.2930,0.5265]	$p = 3.87 \cdot 10^{-09}$	-0.0496 [-0.2064,0.1072]	$p = 0.5349$
Ramos et al. 2013, tPSA							
19	23	4	6	0.0330 [-0.0727,0.1387]	$p = 0.3039$	-0.4264 [-0.5717,-0.2812]	meaningless
PCA3							
12	4	11	25	0.3838 [0.2262,0.5414]	$p = 3.09 \cdot 10^{-05}$	-0.0756 [-0.2768,0.1256]	$p = 0.4616$
tPSA with previous negative biopsy							
8	5	0	2	0.2857 [0.0049,0.5666]	$p = 0.0471$	-0.1737 [-0.5161,0.1687]	$p = 0.3201$
PCA3 with previous negative biopsy							
3	1	5	6	0.2321 [-0.0226,0.4869]	$p = 0.0669$	-0.2273 [-0.5393,0.0848]	meaningless
tPSA without previous biopsy							
19	18	5	6	0.0417 [-0.0784,0.1617]	$p = 0.2840$	-0.4177 [-0.5780,-0.2574]	meaningless
PCA3 without previous biopsy							
14	3	10	21	0.4583 [0.2997,0.6170]	$p = 1.01 \cdot 10^{-06}$	-0.0011 [-0.2035,0.2014]	$p = 0.9918$
Thompson et al., 2014, Scenario 1: (more strict grade only, Gleason score $\geq 4+3$)							
70	37	5	38	0.4400 [0.3482,0.5318]	$p = 1.61 \cdot 10^{-15}$	-0.0194 [-0.1506,0.1118]	$p = 0.7719$
Scenario 2: (less strict grade only, Gleason score $\geq 3+4$)							
72	40	3	35	0.4267 [0.3341,0.5193]	$p = 1.75 \cdot 10^{-14}$	-0.0327 [-0.1647,0.0992]	$p = 0.6269$
Scenario 3: (more strict grade + volume, Gleason score $\geq 4+3$ or > 50% core involvement)							
70	35	4	40	0.4793 [0.3867,0.5718]	$p = 8.12 \cdot 10^{-18}$	0.0199 [-0.1120,0.1518]	$p = 0.7677$
Scenario 4: (less strict grade + volume, Gleason score $\geq 3+4$ or > 33% core involvement)							
72	37	3	38	0.4667 [0.3736,0.5598]	$p = 8.20 \cdot 10^{-17}$	0.0073 [-0.1252,0.1397]	$p = 0.9144$
Saidi et al. 2017, tPSA							
22	80	6	28	0.0450 [-0.0188,0.1087]	$p = 0.1229$	-0.4144 [-0.5193,-0.3095]	meaningless
PHI							
23	21	5	87	0.6270 [0.5181,0.7358]	$p = 1.32 \cdot 10^{-21}$	0.1676 [0.0191,0.3161]	$p = 0.0270$
%p2PSA							
18	19	10	89	0.4669 [0.3382,0.5956]	$p = 1.21 \cdot 10^{-09}$	0.0075 [-0.1621,0.1771]	$p = 0.9306$
Van Neste et al., 2016, PCA3							
448	309	44	77	0.1101 [0.0793,0.1408]	$p = 1.98 \cdot 10^{-09}$	-0.3493 [-0.4305,-0.2682]	$p \ll 0.0001$
TDRD1							
443	343	49	43	0.0118 [-0.0120,0.0357]	$p = 0.2077$	-0.4476 [-0.5253,-0.3698]	meaningless
DLX1							
408	324	84	62	-0.0101 [-0.0360,0.0158]	$p = 0.2603$	-0.4695 [-0.5482,-0.3908]	meaningless
HOXC4							
448	301	44	85	0.1308 [0.0988,0.1627]	$p = 8.12 \cdot 10^{-12}$	-0.3286 [-0.4104,-0.2469]	$p = 3.33 \cdot 10^{-15}$
HOXC6							
448	259	44	127	0.2396 [0.2031,0.2761]	$p = 1.72 \cdot 10^{-27}$	-0.2198 [-0.3042,-0.1354]	$p = 3.35 \cdot 10^{-07}$
HOXC4 and DLX1							
448	267	44	119	0.2189 [0.1831,0.2547]	$p = 4.38 \cdot 10^{-24}$	-0.2405 [-0.3246,-0.1565]	$p = 2.01 \cdot 10^{-08}$
HOXC4 and TDRD1							
448	270	44	116	0.2111 [0.1756,0.2466]	$p = 7.32 \cdot 10^{-23}$	-0.2483 [-0.3322,-0.1645]	$p = 6.46 \cdot 10^{-09}$
HOXC4, DLX1, and TDRD1							
448	267	44	119	0.2189 [0.1831,0.2547]	$p = 4.38 \cdot 10^{-24}$	-0.2405 [-0.3246,-0.1565]	$p = 2.01 \cdot 10^{-08}$
HOXC6 and DLX1							
448	247	44	139	0.2707 [0.2333,0.3081]	$p = 5.20 \cdot 10^{-33}$	-0.1887 [-0.2737,-0.1037]	$p = 1.35 \cdot 10^{-05}$
HOXC6 and TDRD1							
448	251	44	135	0.2603 [0.2232,0.2974]	$p = 4.11 \cdot 10^{-31}$	-0.1991 [-0.2839,-0.1143]	$p = 4.21 \cdot 10^{-06}$
HOXC6, DLX1, and TDRD1							
448	254	44	132	0.2525 [0.2157,0.2894]	$p = 9.96 \cdot 10^{-30}$	-0.2069 [-0.2915,-0.1222]	$p = 1.68 \cdot 10^{-06}$
HOXC6, HOXC4, DLX1, and TDRD1							
448	259	44	127	0.2396 [0.2031,0.2761]	$p = 1.72 \cdot 10^{-27}$	-0.2198 [-0.3042,-0.1354]	$p = 3.35 \cdot 10^{-07}$
Wu et al., 2023, SelectMDx							
941	681	220	737	0.3303 [0.3100,0.3505]	$p = 5.15 \cdot 10^{-159}$	-0.1291 [-0.2054,-0.0529]	$p = 0.0009$
Parekh et al., 2021, 4KScore							
207	371	24	410	0.4211 [0.3895,0.4526]	$p = 3.08 \cdot 10^{-107}$	-0.0383 [–0.1199,0.0432]	$p = 0.3569$
Derderian et al., 2022, Liquid Biopsy, risk classification from 14-gene panel							
17	0	3	48	0.8500 [0.7187,0.9813]	$p = 9.12 \cdot 10^{-27}$	0.3906 [0.21820,0.5630]	$p = 8.99 \cdot 10^{-6}$

4 Replication of Manuscript Figures

Figure captions

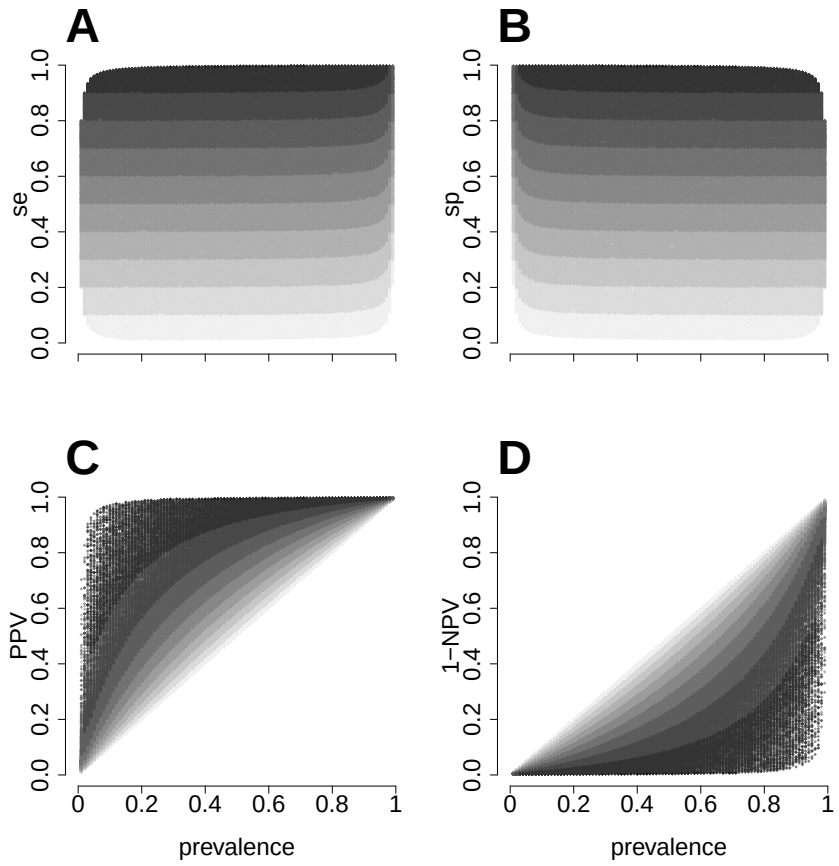

Figure 1: Simulation of 500,000 2×2 tables ($n = 500$) with the Youden index ranging from $0 < J < 1$ in intervals of 10% (light to dark gray). The figure shows that sensitivity (se) and specificity (sp) do not depend on prevalence. However, the probability of disease is affected by prevalence, increasing more with positive test results (positive predictive values, PPV) and decreasing more with negative test results (complement of the negative predictive values, $1 - NPV$) as the value of J increases.

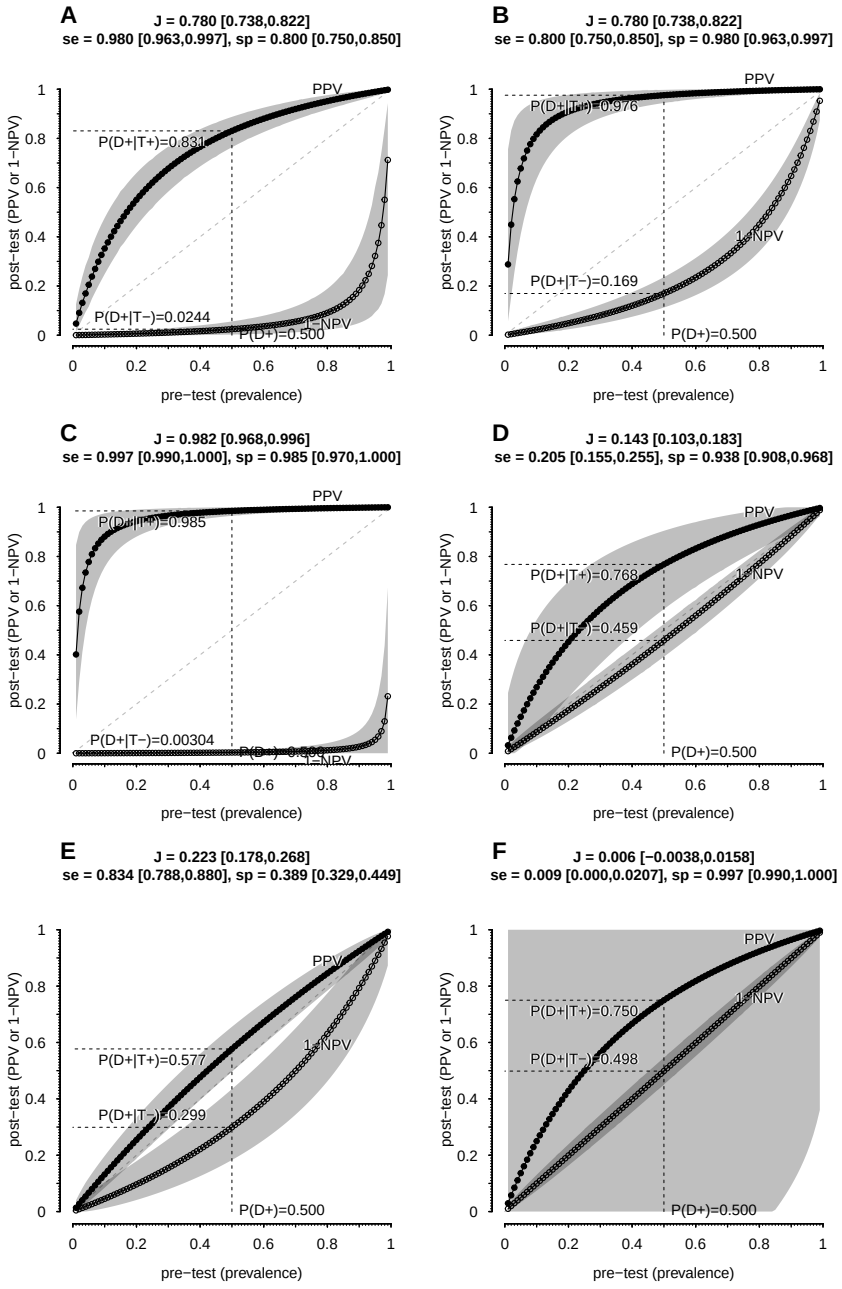

Figure 2: Examples of diagnostic exams with varying sensitivity and specificity values. It is illustrated with updated probabilities of disease (pre-test) with an initial suspicion of 50.0% after a positive test result (PPV) or negative test result ($1 - NPV$). Shadows are 95% confidence bands ($n = 500$). **A** and **B**: hypothetical exams with equal values of Youden index (J) switching sensitivity (se) and specificity(sp); **C**: estimated values for ELISA for HIV detection (França et al., 2018); **D**, **E** and **F**: prostate detection assessed by seric PSA varying the cutoff points, respectivelly 4.1, 1.1, and 10.1 ng/ml (Thompson & Ankerst, 2007).

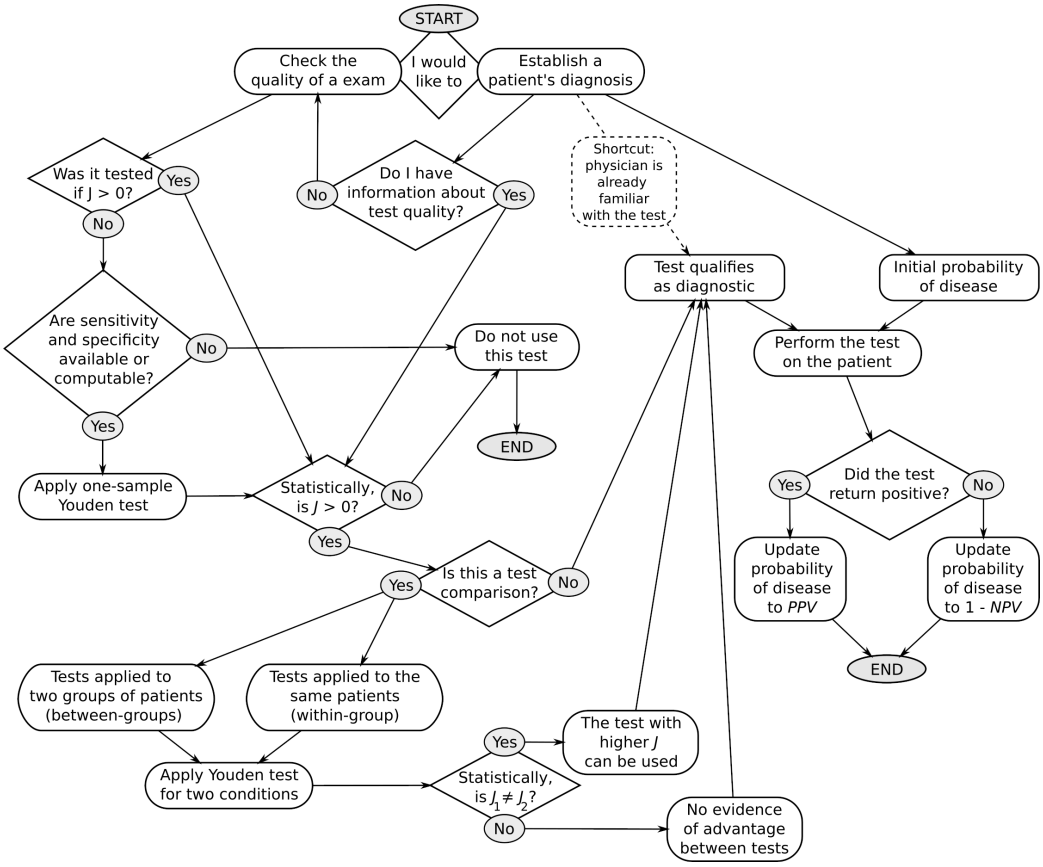

Figure 3: Decision-making algorithm integrating Youden index (J) with prevalence-adjusted positive (PPV) and negative (NPV) predictive values.