PART A

1a) C.

Explanation: ARG is a better predictor of stillbirth than OGTT because the area under the ROC curve (AUROC) for it as seen from Fig. D is considerably more than that corresponding to OGTT for stillbirth (Fig. B). A higher AUROC establishes higher discriminatory power(sensitivity and specificity) of the ARG over OGTT for stillbirth diagnosis.

1b) D

Explanation: The ROC values to the right of the diagonal indicate lower discriminatory power of the diagnosis test. The straight line itself indicates average discriminatory power because the area is 0.5 when the ROC curve overlaps with the line. As the ROC curve further falls below the diagonal on the right side, the ability of the diagnostic test to discriminate cases from non-cases , and vice-versa worsens because the area under the curve decreases below 0.5. This indicates that it is doing worse that average.

1c) E

Explanation: As we can see from parts on ROC in Fig. B and D that for a range of values of Sensitivity, Specificity is constant (a vertical bar in steps) even though the cut-off point is changing. Similarly, for a range of specificity values, sensitivity is constant (horizontal bar in steps) in parts of the ROC curve despite change in cut-off values. This indicates that Sensitivity and Specificity are relatively stable or they do not change together very frequently with change in cut-off values. In contrast, for A and C, the pair of (Sensitivity, Specificity) changes at different cut-off values quite rapidly indicating less stability between the values.

2a) D

Explanation: From the overall population plot (A) we see that the differences are distributed fairly randomly around the mean difference line (y=4.5) , no pattern is visible among the data points, hence the possibility of systematic bias appears implausible. On looking at these plots segregated by population type (septic vs. non-septic) we see no notable differences or violations of random distribution of differences. However, we observe that for smaller magnitude of true values a higher proportion of positive differences exist. Whereas for higher true values, these differences lie close to the mean difference (mainly below it in B and roughly homogeneously around mean difference in C). Moreover, for non-septic shock patients the methods perform relatively better than for septic shock patients. Thus, it appears that co-oximetry may be systematically giving higher values for lower ranges of true values; especially among non-septic shock patients.

2b) C

Explanation: Agreement is worse for Septic shock patients because their limits of agreement are wider (-6.7 to 17.8 , range = 24.5) as compared to that for non-septic shock patients (-7.2 to 14.2, range=21.4). This indicates that the variation in the differences is much higher for the former group than the latter group indicating lower agreement between the two methods. This is also visible from the amount of dispersion of differences around the mean. For septic shock patients, the differences are more spread out around the mean difference whereas this is not observed for the non-septic shock patients.

2c) C

Explanation: There is a clear indication of bias in the total population. For the smaller true values a higher proportion of differences lie above the mean differences than below it. An absence of bias demands homogeneous distribution of residuals (differences), but in this case it is seen only among higher true values (x-axis). On deeper look, it is clear that these differences exist because of lower agreement between methods in septic shock patients.

3a) B

Explanation: The primary outcome of measurement is categorical: presence and absence - 2 categories. The number of raters is 5. Thus, this would mean that each subject is measured 5 times, once by each rater. This gives us a contingency table (140 X 5) dimensions. To measure agreement among raters we will use Cohen's kappa because it will help us in measuring degree of agreement among the raters for each subject over and above agreement by chance. The other options are not suitable for this kind of scenario.

3b) E

Explanation: The primary outcome of measurement is lateral acromial angle (LAA) in degrees. It is a continuous variable. The number of radiologists is 5. Each radiologist measures the patient to find the value of LAA. To estimate the agreement between radiologists (inter-radiologist agreement) we will use ICC. Specifically we will use ICC2(A,1) which will measure agreement between raters when patients are measured once; A indicates ICC Agreement. This is because ICC makes use of total variability in the sample (in this case due to patients, raters, interaction of raters and patients(heterogeneity), random error) and using the formula ICC Agreement=Subject Variability / (Subject Variability + Systematic Bias + Random Error) we can estimate the degree of agreement.

3c) E

Explanation: When the assessment is made by the same radiologist, we are interested in how consistent the measurements taken by him/her are. This consistency, an indicator of reliability, is measured using ICC when the measurement is continuous in nature, as is the case here. Thus, we will use Intra-class correlation to measure the reliability in this case. This ICC is different in that, it will use the formula
ICC=(Subject Variability/ (Subject Variability + Measurement Error)). This ICC is often called ICC_consistency, an indicator of reliability.

PART B

Answer 4

ICC: Intra-class correlation coefficient is used to measure reliability of measurement in different situations. It may be used to measure reliability of measurements taken by a single observer (intra-rater reliability also called consistency) or to measurement reliability of measurements taken by different observers for different subjects (inter-rater reliability, also called agreement). Both these kinds of reliability: consistency and agreement, are provided by the intra-class correlation coefficient.

Context: In the context of this work, the researchers intended to assess how reliable the MMT8 and HHD were in measuring the muscular strength among patients for single muscle groups and also total muscle weakness. They set to assess this reliability in two ways: intra-rater reliability , when the rater measured consistency in observation for measurement M1 and M2 (one week apart). In the second case, they measured agreement using two raters (inter-rater agreement) for measurement M2 and M3 taken on the same day, one hour apart. Thus, in context of the research ICC indicates the reliability of MMT8 and HHD used to measure muscle strength among myositis patients.

SEM: Standard Error of Measurement, is a statistic used to estimate the measurement error in an instrument when multiple measurements are made on the same subject. It indicates the amount of variation in the true measurement which may be attributed to random chance/measurement error. Thus, it helps in assessing the reliability of an instrument by measuring the consistency of measurement across multiple measurements for a given subject(s) by the same observer.

Context: In the context of this work, it is used to indicate the measurement error in the muscle strength when measured using MMT8 and HHD. Since the goal is to identify which instrument is better, the researchers set to compare the SEM of MMT and HHD for different muscle groups.

SDC: Smallest Detectable Change, is the smallest change in the measurement which can be detected over and above the measurement error (SEM). Numerically, it is 1.96 X sqrt(2) X SEM and also called repeatability coefficient. We want it to be small so that we may detect changes across the patients and separate them from each other on measurement scales. Small SDC promises high reliability of the instrument because it can detect smaller differences.

Context: In the context of this work, SDC is used to understand the reliability of MMT8 and HHD for different muscle groups. A lower value of SDC would mean higher reliability or that the repeated observations of the same subject do not deviate from each other grossly.

Elbow Flexion and Ankle Extension:

Table 1 provides Inter-rater and intra-rater reliability for Peak force measured with a hand held dynamometer in Elbow and Ankle.

Muscle Group	Intra ICC (95% CI)	Intra SEM(%)	Inter ICC (95% CI)	Inter SEM(%)
Elbow Flexion	0.83 (0.83 - 0.91)	22 (17)	0.90 (0.83 - 0.95)	15 (12)
Ankle Extension	0.82 (0.70 - 0.90)	21 (14)	0.45 (0.19 - 0.65)	36 (27)

From the above Table we can see that:

1. Intra ICC is almost comparable for Elbow Flexion (0.83) and Ankle Extension (0.82), however the reported confidence intervals for Elbow Flexion appear problematic - the lower CI overlaps with the estimate. For both muscle groups HHD shows excellent consistency in measurements taken by Rater 1, hence excellent reliability. Thus, HHD is able to differentiate across subjects differing in their muscle strengths based on peak force.
2. The magnitude of SEM is quite less (22 for elbow flexion, 21 for ankle extension) as compared to the mean observations (131 for Elbow and 144 for Ankle) which tells us that measurement error is quite small and hence not indicative of major handicap in measurement. Both high Intra ICC and low Intra SEM for these two muscle groups indicates higher consistency of HHD in measuring strength in each of these muscle groups.
3. The Inter-rater reliability for Elbow Flexion is excellent at 0.90 CI (0.83 - 0.95) whereas it is fair for the Ankle Extension (0.45 (0.19 - 0.65). This indicates that when two raters were used no major differences were found in the agreement of their individual assessment of the muscle strength for the elbow flexion muscle. However, in the case of Ankle Extension, the raters' measures differed quite substantially, indicating 55% chances of error in measurement which is quite high and compromises agreement between raters, hence the reliability of the instrument. Interestingly, this ICC is higher than the Intra-rater ICC for elbow flexion which further indicates a higher agreement between raters.
4. The SEM for Elbow (15) is much less than that for Ankle extension muscle (36), thus the latter suffers from high measurement error and low agreement. For Ankle Extension, poor inter-rater reliability indicates limited use of HHD for measuring ankle muscle strength. Its SEM is also quite higher than that observed in the Intra-rater case (36 vs 21).

Overall, HHD appears to be a reliable instrument for measuring strength in Elbow but does not perform as well for Ankle. This is because for overall reliability of an instrument, both consistency and agreement should be good.

Answer 5

Definition: Cohen's kappa is a statistic used to measure agreement between two observations at different times or measurements by two raters at a time, when the measurement is of categorical (ordinal or nominal) nature across different subjects. In this study, for all 8 muscle groups an ordinal scale is used to assess the muscle strength (4 scales: Severe, moderate, mild, no detectable weakness). Thus, for each muscle group , kappa is computed both for one rater(two time periods separated by one week) and two raters (same day, an hour apart).

In the context of this study, the kappa values indicate how reliable the assessment of muscle weakness by MMT8 is for different muscle groups, both across the raters (inter- rater kappa) and across the two measurements taken by a single rater (intra- rater kappa). A higher value of kappa will reflect higher agreement between the different assessments of the strength for a given muscle, hinting at the possibility of using MMT8 widely to measure muscle weakness.

A weighted version of kappa has been used here because the degree of meaningful difference between categorical pairs cannot be assumed to be the same. For eg, classifying severe as mild is more undesirable than classifying severe as moderate. Thus there is a need to penalize the incorrect categorizations based on the degree of discordance between the ordinal categories. This will ensure that the final kappa value takes into account the ordinal nature of the muscle weakness assessment rather than treating it as nominal. If weighted kappa is not used, it would mean that intrinsic differences between these 4 scales have not been taken into account and are assumed to be the same. This may under-estimate or overestimate the kappa value, leading to an inaccurate estimate of reliability.

Muscle	**Intra-tester Kappa**	**Inter-tester Kappa**
Elbow Flexion	0.66	0.30
Ankle Extension	0.35	0.20

Table 2: Cohen Kappa values for Elbow and ankle.

According to the table above, for elbow muscle a substantial reliability (0.66) is observed in the intra-tester scenario whereas for the ankle extension it is much less and achieves only a fair level of reliability. This means that a higher amount of agreement was observed between the two observations by the single rater on 2 different occasions for elbow muscle weakness assessment than for the ankle. Thus, MMT8 appears to be more reliable in measuring muscle strength of the elbow than that of the ankle.

On the other hand, both muscle groups show a much lower level of Inter-rater kappa as compared to the intra-rater kappa, indicating that the agreement between the two raters for muscle weakness assessment for these muscle groups was fair. Thus, the inter-rater reliability is quite low. These figures indicate agreement over and above that expected due to random chance. Overall, it appears that MMT8 may be a reliable instrument to measure muscle strength in the elbow because but may not so for the ankle extension.

Answer 6

Bland-Altman plots are plots which are used to assess the agreement between two methods of measurement measuring the same thing across different subjects. The plots are used to analyse the relationship between the true value of the estimate (obtained from the mean value of the measurements from the two observations) and the difference between the two measurements.

By analysing the patterns in the differences with mean, Bland-Altman plots intend to comment on the presence of systematic difference between the two methods of measurement. If the difference of measurement indicates some systematic patterns with mean values, such as funneling or over-difference, under-difference, then it may be said that the two methods under consideration provide different values for the entity under measurement. This would prohibit using the two methods interchangeably or substitute one for the other entirely.

Fig. 3 and 4 show the BA plots for HHD Peak force.

Elbow Flexion: In case of the BA plot for Intra-rater agreement we can observe that the differences between measurements M1 and M2 taken by the same rater lie close to the mean differences (0.9) which is quite close to the line of zero difference (y=0). At higher values of true estimate of elbow muscle strength(mean of M1 and M2), these differences tend to fly off especially in the positive direction showing some possibility of systematically higher measurement in M1 by MMT8. This could be attributed to the difference of one week between these two measurements. LOA : -61 <=Mean <=63

For measurements M2 and M3, taken on the same day by different raters, we see a much tighter envelope of differences around the mean difference, indicating lower level of dispersion, or higher agreement between the two assessments by two raters. However, the mean difference is 5, higher than that observed for intra-rater. LOA : -37 <=Mean<=47.

From the limits of agreement for both cases we can see that agreement is stronger for the second case (inter-rater) as compared to the first case (intra-rater). Moreover, no pattern is observed for inter-rater, thus, plausibility of systematic bias is almost non-existent. Therefore, the MMT8 method seems to match across the two raters.

Ankle Extension: In the case of ankle extension, the results are opposite to that observed for elbow flexion. For intra-rater, the limits of agreement (-55, 61, Mean difference = 2.8) are much more tighter than that for inter-rater (-86, 115, Mean difference=14). For the former, the mean difference (2.8) lies very close to the mean difference of 0 given the range of difference between M1 and M2. Thus, it seems that intra-rater agreement is higher for ankle extension muscle strength assessment than the inter-rater agreement. This could hint at the possibility of some difference in the measurement method by the two raters to measure ankle strength. In both the cases, no systematic pattern is observed among differences across the true/mean values. The inter-rater agreement is low because the differences are widely dispersed as compared to the intra-rater.

Thus, while for elbow flexion, a higher inter-rater reliability is observed, for ankle extension higher intra-rater reliability is observed. These differences could be attributed to the change in time in the case of the former and difference in measurement technique between the two raters in the case of the latter.

PART C

Answer 7

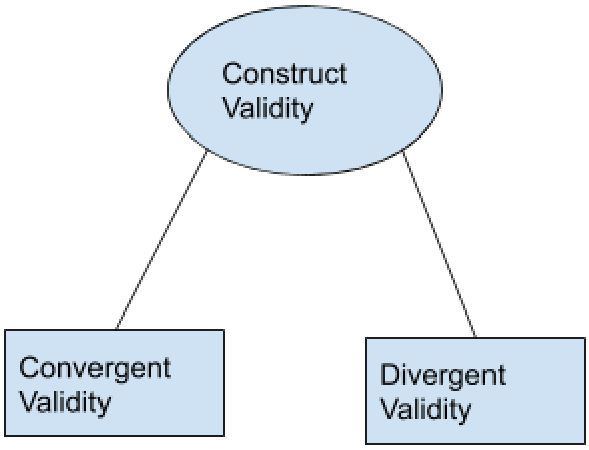

Figure 1: Constituent elements of Construct Validity.

Convergent Validity: It is one of the ways to measure the construct validity of a measurement instrument or scale. Convergent validity intends to assess whether the two measurement scales given by two different methods converge together or correlate together (one increases, other also increases, one decreases, the other also decreases). A higher value of convergence (high correlation) is required to ascertain higher convergent validity. Normally, two measurements that measure the same concept are expected to show high convergent validity.

Divergent Validity: It is used to measure the degree to which the method of measurement/scale under consideration measures only the desired construct and thus, is not related to some other different or unrelated construct. In this case, a divergence is desired (low correlation) between two constructs measuring the same concept/construct.

For a scale to have good construct validity, it should not only show good convergent validity but also good divergent validity.

In the context of the study, the researchers intended to measure the construct validity of PY-BOCS II. For convergent validity they used COI scale and for divergent validity they used BDI and STAI (State and Trait) scales. This was because COI assesses the obsessive and compulsive symptoms and they wanted their scale to be highly convergent with COI scale values for different patients. On the other hand, they did not want PY-BOCS II to measure anything other than Obsession and Compulsion, such anxiety of depression, so they used scales which measure the different constructs (BDI - Depression, STAI (State and trait) - Anxiety) to assess the divergent validity of PY-BOCS II.

Table 3 in the paper shows the outcome of convergent and divergent validities. These values are obtained from the correlation coefficients between different values of

PY-BOCS II total, obsession and compulsion and COI total, BDI total and STAI (Trait and State) total.

For each patient, these scores are computed and correlated to one another using Pearson product moment correlation.

All	**Y-BOCSII obsession**	**Y-BOCS compulsion**	**YBOCS-total**	**Type of Validity**
COI	High	High	High	**Convergent Validity**
BDI	Low	Low	Low	**Divergent validity**
STAI state	Low	Low	Low	
STAI Trait	Low	Low	Low	

Table 3: **Desirable Validity Values Table.** High could indicate more than 0.7, low could indicate less than 0.3 (Hinkel et. al, 2003).

In the ideal case for PY-BOCS II to show high construct validity, we would expect it to have a high convergent validity (high correlation between COI and PY-BOCS II) and high divergent validity (low correlation coefficient between BDI and PY-BOCS II and STAI (State, Trait) and PY-BOCSII).

The authors mention that they were able to achieve high convergent validity but the divergent validity was only acceptable for Y-BOCS II compulsion scale. In my opinion, although the convergent validity is acceptable (0.67 overall) but the divergent validity of Y-BOCS II is poor. It is 0.68 with STAI-trait total and 0.57 with BDI total. These fairly indicate a high level of correlation where low level is required to distinguish OCD people from those with other mental disorders. This demonstrates that the scale although measures obsession, compulsion but the possibility of it tapping into depression and anxiety cannot be fully negated. The construct validity of the YBOCS-II scale is questionable in the Portuguese population.

Answer 8

Factor Analysis, also known as Principal Axis Factoring, is a technique which is used to identify factors/hypothetical constructs in the data which may be used to explain the variation between items in a measurement scale. In general, it is an attempt to reduce the dimensionality of the data and to explain the observed variations using these meaningful hypothetical constructs made from variables under consideration.

It was conducted in this study, because the authors wanted to explore whether the two factors identified for Y-BOCS II in previous studies, namely obsession and compulsion, also exist when using this scale in the Portuguese population. Furthermore, due to contention over the existence of a third factor, “resistance to obsession”, they carried out exploratory factor analysis rather than a confirmatory factor analysis.

The authors decided on the number of factors based on the eigenvalues of the 10 different components provided by Principal Component Analysis (PCA). As per PCA, two components had eigenvalues of more than 1 and thus, the authors chose the number of factors as 2. Yes, I agree with their choice because the third highest eigenvalue was less than 1 (As seen from Fig. 1 Scree plot), which means that it explains even less variation than its baseline level of variation explained (which is normally considered to be 1). The first two eigenvalues (5.05 and 1.47) are higher than 1 and no major contribution is made by components beyond these two components. Thus, two factor structure is a correct choice.

Item 1: Time spent on obsession. (Factor 1 loading : 0.06 , Factor 2 loading: 0.76)

Item 3: Distress associated with obsession. (Factor 1 loading : 0.36 , Factor 2 loading: 0.39)

Item 4: Resistance to obsession. (Factor 1 loading : -0.27 , Factor 2 loading: 0.98)

These 3 items relate to questions on obsession, form a part of obsession subscale, hence contribute to the factor 2 which stands for obsession. Looking from the factor loadings we can see that all three items load more on Factor 2 than on Factor 1, although, for Item 3 factor loading magnitudes differ very less (0.36 vs 0.39). For this, higher factor loading is for Factor 2, but this item is factorially complex. For item 3, factorial complexity may be associated with the inclusion of word “distress” which may be quite prominent in compulsion also, hence the loadings do not vary widely. Nevertheless, it is used to determine Factor 2. For item 1, it is clear that it is highly correlated with Factor 1 because its magnitude of factor loading is high (0.76). In case of Item 4, while the absolute value of factor loading is high for Factor 2, a negative factor loading is observed for Factor 2 indicative negative correlation. This is not worrisome because while creating factors one looks for high value of factor loadings each item. For this case, it is quite tidy. Thus, all these three items are highly related to Factor II, which is the obsession factor. These values are consistent with the nature of questions asked through these items, as shown above. These relate to obsession.

Answer 9

Cronbach's alpha is a measure of internal consistency which is used to assess the reliability of questionnaire scales. It allows us to know the degree of correlation between all the questions in a questionnaire. Ideally, if the questions are measuring the same thing/concept, they must be highly correlated in their responses, and the Cronbach's alpha value would be high. In general, a high value (more than 0.7) represents high reliability and is considered acceptable, indicating that the scale is reliable in measuring the desired concept and may be used in the population. A very high value, 0.95, may instead indicate redundancy and hint at superficiality of some items. Please note that Cronbach's alpha value is not universal, it may differ from population to population and indicates reliability of a scale in a given population.

In the context of this study, the researchers used Cronbach's alpha to measure the internal consistency of the questionnaire for all the participants and for OCD sub-sample (obsessive and compulsive scales) separately. They wanted to assess the reliability/internal consistency of PY-BOCS II both for the overall population and for the population with OCD.

For all participants, they found a Cronbach alpha value of 0.96 whereas for the OCD sample, they found that it was 0.94 for both obsessive and compulsive scales. This indicates a drop in reliability of the scale among OCD subset which is actually not strange because, this subset consists of people only with OCD and PY-BOCSII had failed to provide good divergent validity, therefore, a decrease in reliability is expected. However, such high values are indicative of the presence of some redundancy of items in both the situations - all participants and OCD subsample.

For total values, authors report no gross change in the Cronbach's alpha when items are removed one at a time, 0.96 is fairly maintained indicating no fluctuations in internal consistency of the scale for the total population. However, this is not observed for the OCD sample subset. We can see that when each item is deleted, the alpha value decreases indicating that the deleted item was an important component contributing to the overall reliability of the scale for the OCD sample. Item 4 (resistance to obsession) and Item 7 (degree of control over compulsion) show the least drop (alpha=0.88) whereas for others it is 0.87. This means that the loss of items other than these is much more costly in terms of reliability than of these items.

The drop in alpha values among OCD subset is an expected behaviour because each item is crucial for the overall OCD measurement, 5 items each make up the obsessive and compulsive scale, hence, their inclusion is instrumental for a higher degree of internal reliability of the scale. However, absence of this drop for the overall population actually reflects nothing major because the total population consists of people with and without OCD (healthy, those with other mental health disorders) , thus, a higher reliability in that case could be attributed to measurement of something other than OCD. Perhaps, trait/anxiety because these showed low divergent validity with PY-BOCS II.

Answer 10

Sensitivity and Specificity are two criteria used to assess the diagnostic ability of a test. Sensitivity refers to the ability of a test to correctly detect true positive cases (those with the condition) from the total number of true cases (total number diagnosed with that condition).

Confusion matrix		Gold Standard	
		Yes	No
Test	Yes	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

Table 4: Confusion Matrix

As per the table above Sensitivity = TP/(TP+FN)

Specificity refers to the ability of a diagnostic test to correctly detect true negative cases (those without the condition) from the total number of true cases without that condition.

Specificity = TN/(TN+FP)

The total number of true cases with or without diseases are usually obtained from a gold standard and the result of the diagnostic test is compared to results obtained from it.

In the context of this work (Table 4 in paper), sensitivity and specificity indicate the diagnostic robustness of the test at different cut-off values used for detecting OCD among all the participants. In order for a test to be useful, high values of sensitivity and specificity are required, considering the trade-off between them at different cut-offs. A cut-off is chosen which optimises sensitivity and specificity in context of the research objective while considering the trade-off between the two. This cut-off is chosen using the ROC plot which is produced from Table 4 values; this point is generally close to the (0,1) point on the ROC curve. In this case, the authors intended to maximise specificity while maintaining a substantial sensitivity. Cut off of 13 seemed to be achieving this, hence they chose this point. Moreover, beyond this cut-off (both above and below) the trade-off between sensitivity and specificity increased, compromising the discriminatory power of the scale.

These additional analyses were conducted because the main objective of the study was to assess the criterion validity of the PY-BOCS II scale. To be confident in saying that PY-BOCS successfully discriminates those with OCD from those without in the Portuguese population, authors demonstrated its discrimination power for three groups of samples, each with different characteristics. The authors wanted to be sure that that the scale correctly differentiated people with and without OCD

1. in over all population (Plot 1 Fig 2),
2. among age-, gender-, education-matched controls (Plot 2)
3. From those with mood and anxiety disorders (Plot 3)

These analyses are helpful in testing and demonstrating the validity of PY-BOCS II because they take care of confounding and error due to non-blinded assessment. Based on their results, they indicate that PY-BOCS II may be used for diagnosing OCD among patients who visit hospitals and indicate obsession-like ideas or compulsive-like behaviours.

According to their main analysis, AUROC=96% and at cutoff = 13, Sensitivity = 85% and Specificity=97%. The test is thus, excellent in segregating those without OCD in the overall population.

In the matched population, for the same cut-off of 13 Sensitivity=85% and Specificity=96%. These are not grossly different from the overall population further validating the scale. Adjusting for confounding, authors show that PY-BOCS is robust in discriminating OCD from non-OCD. For this case, I expected the results similar from main analysis because the stratification of the data would uphold main results if the populations are not starkly different.

For a heterogeneous sample having people with depression, anxiety and OCD they used a blinded assessment and achieved Sensitivity= 90% and Specificity = 94%. For this case, I expected lower Specificity because the divergent validity of PY-BOCS II was not good. It seemed to be related to even anxiety/depression. Even though, specificity has decreased - the test no longer cleanly tells who does not have OCD and confuses its absence with its presence. Increase in sensitivity is expected because the population consists of mentally sick people.

References

Hinkle DE, Wiersma W, Jurs SG. Applied Statistics for the Behavioral Sciences. 5th ed. Boston: Houghton Mifflin; 2003. [Google Scholar]