Clinical Biostatistics: Evaluating diagnostic tests

Assoc. Prof. Cameron Hurst
cphurst@gmail.com

Faculty of Medicine,
Chulalongkorn University

7th December, 2559

Get notes at: research.md.chula.ac.th

Preamble

- **Diagnostic tests** represent a tool whereby patients are tested for the presence or absence of a disease
- They can be used in a range of applications ranging from screening, surveillance, or confirmation. Typically:
  - A **diagnostic test** is used to confirm the absence or presence of a disease where we suspect they have the disease (symptomatic)
  - Whereas, **screening tests** are for detecting the disease in asymptomatic individuals
- The relative importance of *getting it wrong* might change across these settings
- Today, we will consider methods to evaluate new diagnostics against an existing gold standard.

What we will cover....

Diagnostic tests

- We will consider the situation where we want to compare a new method for diagnosing disease (Present Vs Absent) against some existing 'gold standard'
- At first glance, this would appear to be the same as considering a binary instrument for two raters (a reliability problem)
- BUT, there is a very important difference:

Important point:

Unlike inter-rater agreement problem, the source of error in assessing diagnostic tests comes **ONLY** from the new test. The existing method (**gold standard**) is assumed to be 100% correct.

Diagnostic test accuracy

Let's start by looking at the elements of a diagnostic test assessment:

		diagnostic test(new)	
		T^+	T^-
Actual disease status	D^+	**True Pos**(TP)	**False Neg**(FN)
(gold standard)	D^-	**False Pos**(FP)	**True Neg**(TN)

Where:

- D^+ is the event that the patient is **actually** diseased
- D^- is the event that the patient is **actually** non-diseased
- T^+ is the event that the patient has a positive test result
- T^- is the event that the patient has a negative test result

Note: **Green** is where we get it **right**, and **Red** is where we get it **wrong**

Test accuracy, Sensitivity and specificity

After we have compared our new test with the gold standard we will have numbers to populate the above table.

The (**green**) numbers on the main diagonal represents the patients classified correctly. The probabilities of these main diagonals jointly represent the **test accuracy**:

$$\text{Test accuracy} = \frac{TP+TN}{n} = \frac{TP+TN}{TP+TN+FP+FN}$$

Sensitivity and Specificity

There are two main quantities that we are interested in when we first evaluate a new diagnostic test: The **Sensitivity** and the **Specificity**

- **Sensitivity** is the probability that the test **correctly** classifies a **diseased** patient, $P(T^+|D^+)$
- **Specificity** is the probability that the test **correctly** classifies a **non-diseased** patient, $P(T^-|D^-)$

Sensitivity and **Specificity** tell us how well a diagnostic test discriminates between patients **with** and **without** a disease

Notation:

The notation $P(A|B)$ means the event A given (or conditional on) event B. So $P(T^+|D^+)$ implies a positive test result (T^+), given you have the disease (D^+).

Example: Sensitivity and Specificity

We have 100 patients **known** to have a disease and 1000 known to be disease-free (based on the gold standard). These patients are tested using a new diagnostic procedure

		diagnostic test(new)		
		T^+	T^-	**Total**
Actual disease status	D^+	90(TP)	10(FN)	100
(gold standard)	D^-	200(FP)	800(TN)	1000
	Total	290	810	1100

Now:

- $Sensitivity = P(T^+|D^+) = \frac{90}{100} = 0.9$
- $Specificity = P(T^-|D^-) = \frac{800}{1000} = 0.8$

Is accuracy everything

- We see from our example we have a **Sensitivity** of **0.9**, and a **Specificity** of **0.8**. This seems OK.
- Let's dig a little deeper:
  - For every thousand diseased patients, we will miss 100 (let them go home)
  - For every 1000 non-diseased patients, we will accidentally classify 200 with the disease (which might be followed up by dangerous or invasive procedures)
- We ALSO need to consider if we are likely to encounter just as many diseased patients as diseased patients. **What is the prevalence of the disease?**

Why is prevalence important?

Let's see why prevalance matters. Three sceneries (using our example diagnostc tests with **sensitivity=0.9** and **specificity = 0.8**). All three cases consider 10,000 patients

1. **Hypertension in elderly males (prevalence = 50%)** Here, of the 5000 men who have HTN, 4500 will be correctly classified (500 will be missed), and of the 5000 without hypertension, 1000 will be falsely diagnosed as hypertensive.
2. **Recurrent breast cancer (prevalence = 10%)**. Of the 1000 women who have recurrent breast cancer, 100 will be missed. Of the 9000 who don't, 1800 women will be told (falsely) that their cancer has recurred.
3. **Screening for Colo-rectal cancer (prevalence = 1%)**. Of the 100 people who have CRC, 10 are missed. Of 9900 who don't have CRC, 1980 will be told that they have CRC.

Predictive value of a test

This means sensitivity and specificity (while indicating accuracy) doesn't tell us much about the **utility of the test** (its clinical usefulness)

- A set of quantities that give an idea about the **usefulness** of a new diagnostic tests are the **predictive values of the test**
- Theses values have a positive test version (PV^+) and a negative test version (PV^-) and are often called the *posterior probabilities* (more later)

1. PV^+ represents the probability of having the disease GIVEN (conditional on) a positive test result: $P(D^+|T^+)$
2. PV^- represents the probability of being disease-free GIVEN (conditional on) a negative test result: $P(D^-|T^-)$

That is, of those that tested **positive**, how many did we get right, and of those that tested **negative**, how many did we get right?

Calculation of (PV^+) and (PV^-)

Using Bayes' theorem:

$$PV^+ = P(D^+|T^+) = \frac{P(D^+)P(T^+|D^+)}{P(D^+)P(T^+|D^+) + P(D^-)P(T^+|D^-)}$$

or simply,

$$PV^+ = P(D^+|T^+) = \frac{TP}{TP + FP} = \frac{\text{True Postives}}{\text{All Positives}} \quad (1)$$

$$PV^- = P(D^-|T^-) = \frac{P(D^-)P(T^-|D^-)}{P(D^-)P(T^-|D^-) + P(D^+)P(T^-|D^+)}$$

or simply,

$$PV^- = P(D^-|T^-) = \frac{TN}{TN + FN} = \frac{\text{True Negatives}}{\text{All Negatives}} \quad (2)$$

Prevalence and study design

- Perusal of the above equations show that sensitivity and specificity are directly related to (PV^+) and (PV^-)
- Also, disease prevalence, $P(D^+)$ (and $P(D^-) = 1 - P(D^+)$) are in the calculations. This implies:
  1. If we want to use our sample to estimate $P(D^+)$, using $\frac{D^+}{n}$, it should be representative of the clinical population.
  2. If this isn't the case (e.g. case-control sample), equations (1) and (2) (PV^+ and PV^-) aren't valid for this population.
  3. Finally, this also implies that PV^+ and PV^- are **susceptible to levels of disease prevalence**

Prevalence (pre-test probability) and PV^+ and PV^-

Unlike Sensitivity ans Specificity, the prevalence of the disease directly effects PV^+ and PV^-. In other words, PV^+ and PV^- give us a good idea of the actual usefulness of the test.

Example: (PV^+) and (PV^-)

Recall our example:

		diagnostic test(new)		
		T^+	T^-	**Total**
Actual disease status	D^+	90(TP)	10(FN)	100
(gold standard)	D^-	200(FP)	800(TN)	1000
	Total	290	810	1100

Now assuming our sample is representative of the population:

$$PV^+ = P(D^+|T^+) = \frac{TP}{TP+FP} = \frac{90}{90+200} = 0.310$$

$$PV^- = P(D^-|T^-) = \frac{TN}{TN+FN} = \frac{800}{800+10} = 0.988$$

Also note: $Prevalence = \widehat{P(D^+)} = \frac{100}{1100} = 0.09091$

Bringing it together: Interpretation of results

1. The test was quite sensitive ($P(T^+|D^+) = 0.9$) with 90% of individuals with the disease being identified as positive
2. The test was also quite specific ($P(T^-|D^-) = 0.8$) with 80% of disease-free individuals correctly identified
3. The positive predictive value was not very high ($PV^+ = 0.31$). The chance a patient with a positive test really having the disease is only 0.31 (i.e. Only about 3 out of 10 who test positive have the disease)
4. And 98.8% testing negative are truly disease-free ($PV^- = 0.988$)

Note:

In this case: The major difference between $PV^+ = 0.31$ and $PV^- = 0.988$ relate to the quite low prevalence ($P(D^+) = 0.09091$) of this condition. As we have seen, higher prevalence (holding sensitivity and specificity constant) would result in higher PV^+ but lower PV^-

Disease likelihood

- When we conduct a diagnostic test, we are trying to **rule in** (positive result) or **rule out** (negative result) our patient
- This involves an initial assessment of the likelihood our patient has the disease (**pre-test probability** = disease prevalence),
- We then conduct the diagnostic test (using **patient-specific** information) to shift (one way or the other) our level of confidence about the presence (or absence) of the disease (**post-test probability**)
- **Likelihood ratios tell us how much we should shift our suspicions for a particular test result**: Are we more certain a patient has the disease (after a test result), or more certain a patient does not have the disease?

Positive and negative likelihood ratios

- As diagnostic tests can give us two possible results (positive or negative), there are two likelihood ratios for any test:
  - The **Positive likelihood ratio** (LR_+) indicates how much we should increase the probability of the disease (from the pre-test probability of the disease) if the test is positive; and
  - The **Negative likelihood ratio** (LR_-) tells how much we should decrease the probability (from the pre-test probability) if the test is negative;
- **Rem:** pre-test probability is based **purely** on the population (e.g. prevalence), whereas post-test probabilities ALSO includes info from the diagnostic test (patient-specific information)

Likelihood ratios

LRs uses the extra information (on top of pre-test probability) that a test gives us for a positive (LR_+) and negative (LR_-) diagnosis

Positive Likelihood ratio, LR_+

The formula for the positive likelihood ratio is:

$$LR_+ = \frac{\text{probability individual } \textbf{WITH} \text{ the disease tests } \textbf{POSITIVE}}{\text{probability individual } \textbf{WITHOUT} \text{ the disease tests } \textbf{POSITIVE}}$$

or,

$$LR_+ = \frac{\text{Sensitivity}}{\text{1-Specificity}}$$

Negative Likelihood ratio, LR_-

And the formula for the negative likelihood ratio is:

$$LR_- = \frac{\text{probability individual } \textbf{WITH} \text{ the disease tests } \textbf{NEGATIVE}}{\text{probability individual } \textbf{WITHOUT} \text{ the disease tests } \textbf{NEGATIVE}}$$

or,

$$LR_- = \frac{\text{1-Sensivity}}{\text{Specificity}}$$

Example: Calculating LR_+ and LR_-

From our example

		diagnostic test(new)		
		T^+	T^-	**Total**
Actual disease status	D^+	90(TP)	10(FN)	100
(gold standard)	D^-	200(FP)	800(TN)	1000
	Total	290	810	1100

Recall

$$\text{Sensitivity} = \frac{90}{100} = 0.9 \text{ and Specificity} = \frac{800}{1000} = 0.8$$

Now,

$$LR_+ = \frac{\text{Sensitivity}}{\text{1-Specificity}} = \frac{0.9}{1-0.8} = 4.5$$

And:

$$LR_- = \frac{\text{1-Sensitivity}}{\text{Specificity}} = \frac{1-0.9}{0.8} = 0.13$$

Clinical application of LR_+ and LR_-

So we have $LR_+ = 4.5$ and $LR_- = 0.13$. What do these values tell us?

- LRs represent a sliding scale of disease probability and are also stable. Unlike the PV^+ and PV^-, LRs are independent of prevalence (LRs only depend on *Sens* and *Spec*)
- LRs also more interpretable to clinicians as thresholds can be set for **poor**, **acceptable** and **good** values of the LRs (see below)
- We can interpret the LRs as:
  - LR_+ represents the **ruling in** of disease (based on a positive result); and
  - LR_- represents the **ruling out** of disease (based on a negative result)

Interpreting our LRs

These cutpoints are widely used in interpreting the LRs:

LR	Interpretation
>10	Large (and conclusive) increase in the likelihood of disease
5-10	Moderate increase in the likelihood of disease
2-5	Small increase in the likelihood of disease
1-2	Minimal increase in the likelihood of disease
1	No change in the likelihood of disease
0.5-1	Minimal decrease in the likelihood of disease
0.2-0.5	Small decrease in the likelihood of disease
0.1-0.2	Moderate decrease in the likelihood of disease
<0.1	Large (and conclusive) decrease in the likelihood of disease

So given our values of $LR_+ = 4.5$ and $LR_- = 0.13$. $LR_+ = 4.5$ suggest a positive result for **our** diagnostic test provides a **small increase** (upper end) in the likelihood of the disease, and $LR_- = 0.13$ suggests a negative result of our test indicates a **moderate decrease** in the probability of disease. **So? Is our test useful?**

Back to the 2 x 2 table

Often many new diagnostic tests are based on a continuous scale, and we don't know which cut-off will give us the best diagnostic accuracy (and utility). Let's go back to our 2 x 2 table.

		Diagnostic test(new)	
		'High' test result(T^+)	'Low' test result(T^-)
Disease status	D^+	True Pos(TP)	False Neg(FN)
(gold standard)	D^-	False Pos(FP)	True Neg(TN)

We can imagine that:

- If we lower the threshold (value of cut-off), we might capture some more true positives (↑ sensitivity) but we are likely to let in some false positive (↓ specificity)
- Conversely, if we increase the threshold, we would be able more accurately exclude more people without the disease (↑ specificity), but may also miss more diseased individuals (↓

Receiver-Operator Characteristic (ROC) curves

- We need a way to assess this balance of sensitivity and specificity
- Receiver-Operator Characteristic (ROC) curves do exactly this
- The horrible (and seemingly irrelevant) name of the ROC comes from their first area of application (early radio technology)

Note:

All ROC curves do is explore what happens to sensitivity (fraction of true positives) and specificity (usually via 1-specificity = fraction of false positives) when we move the clinical measure threshold (cut-off) up or down

Example: Fever in children

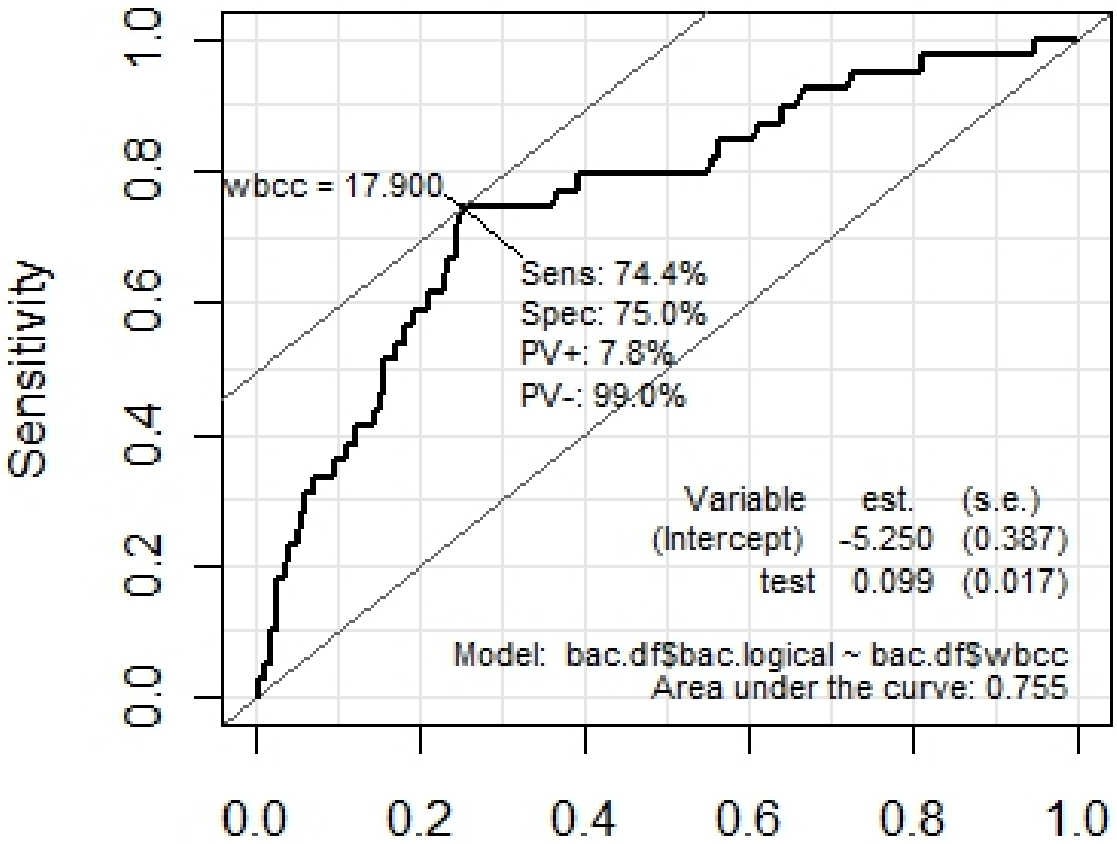
Fever is children in a majority of cases is a result of self-limiting viral infections, but some of these children will be bacteraemic which can progress to quite serious conditions

Despite many studies having been done, the management of febrile children still presents problems to treating physicians. For example:

1. The consequences of not treating a child with bacteraemia are potentially life-threatening (false negatives are bad); also
2. Over-use of antibiotics also presents problems (false positives are bad)

A number of clinical measures were trialled to distinguish between 'bacteraemic' and 'non-bacteraemic' febrile children. We will consider two: **(1) White blood cell count**; and **(2) Temperature**

Our first ROC curve: White blood cell count



White blood cell count

1. wbcc: 17.9
2. Sensitivity: 74.4%
3. Specificity: 75%
4. PV^+: 7.8
5. PV^-: 99
6. LR_+: 2.98
7. LR_-: 0.34
8. AUC: 0.755

Interpretation

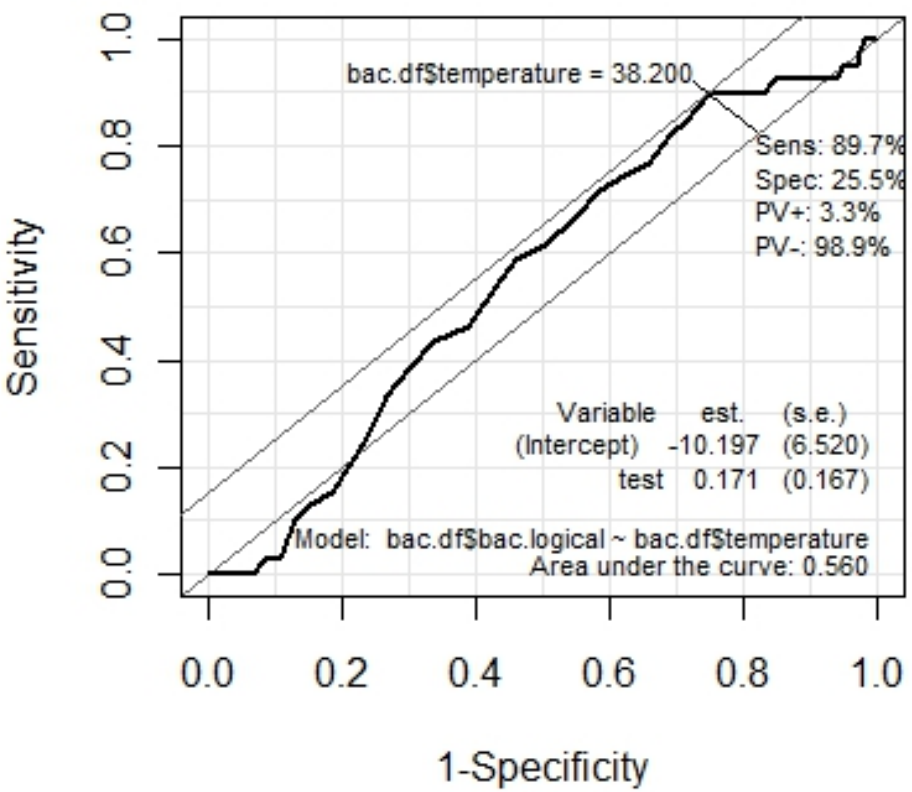
- If we use a white blood cell cut-off of 17.9 $10^9 L^{-1}$ (1:wbcc) we would expect to (correctly) identify 74.4% of bacteraemia children (2:sensitivity) and correctly exclude (negative test) 75% of non-bacteraemic children (3: specificity)
- The probability of a child testing positive and really being bacteraemic is only 0.078 (4: PV^+) AND the probability of a child testing negative and not having the condition is 0.99 (5: PV^-)
- $LR_+ = 2.98$ (6) suggests a positive test only slightly increases the chance of having the disease (see table of guidelines), and $LR_- = 0.34$ (7) suggests negative result only slightly changes the likelihood of no disease.
- The Area under the ROC curve (8: **AUC**) of 0.755 shows WBC count (as a tool) is quite effective: discriminates between those with and without bacteraemia

AUC: Area under (ROC) curve

- As indicated above, AUC tells us about a clinical measure's ability (independent of the cut-off) to discriminate between those with a disease and those disease free.
- So AUC represents an 'overall' measure of instrument accuracy
- **The closer an AUC is to 1, the better the instrument**
  - An instrument whose ROC curve has a pronounced convex bulge towards the upper left, will have a higher AUC
- **The closer the AUC to 0.5, the worse the instrument**
  - Test's without the pronounced bulge (a straight diagonal line) will have an AUC close to 0.5, indicating low instrument accuracy $\Rightarrow$ Flip a coin

Temperature

Now let's look at a not so successful diagnostic test:



Temperature

1. Temp: 38.2
2. Sensitivity: 89.7%
3. Specificity: 25.5%
4. PV^+: 3.3
5. PV^-: 98.9
6. LR_+: 1.2
7. $LR\_$: 0.4
8. AUC: 0.56

Interpretation

- Straight away from the shape of the ROC curve we can see temperature is not a good diagnostic tool. This is supported by the very low AUC (6:AUC=0.56)
- Using a cut-off of 38.2^oC (1:temp), we can see the test is quite sensitive (2:sensitivity) where 89.7% of bacteraemic children testing postive
- In sharp contrast, only 25.5% of bacteraemia-free children tested negative (3:specificity). Lot of false positives.
- Of those testing positive, only 3.3% are bacteraemic (4:PV^+)
- but of those testing negative $PV^- = 98.9\%$ were disease free
- A positive test doesn't really change the chance of having the disease ($LR_+ = 1.2$), and nor does testing negative really increase (much) the chance of being disease free ($LR_- = 0.4$)

Question for you to consider

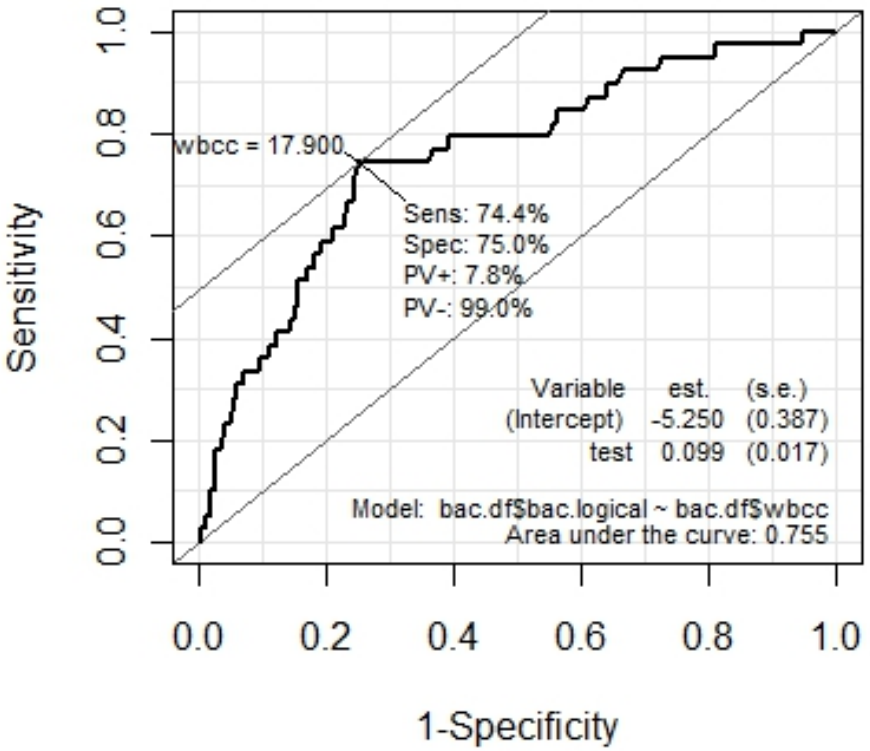
1. For both the **White blood cell count** and **Temperature** diagnostic tests, PV^+ was very low, and PV^- very high, what might explain this?
2. If you take WBCC to be the better test, are you happy with the 'optimal' threshold ($17.9\ 10^9 L^{-1}$)?
   - When answering this question: think about the two major considerations in developing this tool:(1) identifying febrile children at risk of the serious conditions that arise from unreated bacteraemia; and (2) Overuse of antibiotics?
   - What type of test(s) best satisfies these two objectives

Finding an appropriate cut-off

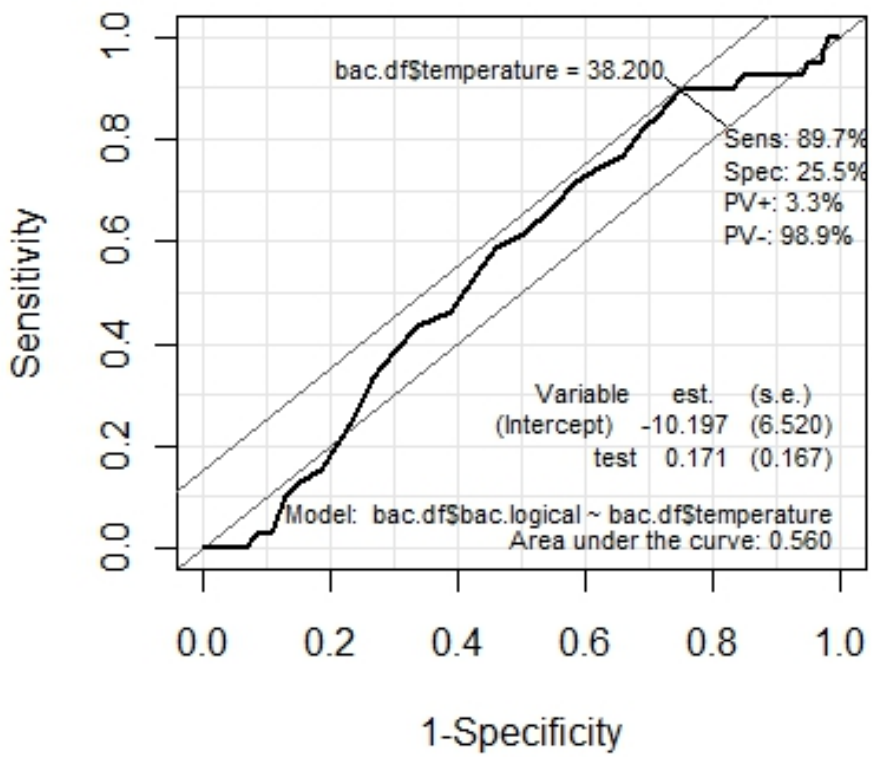
- If we consider both sensitivity and specificity as equally important, the point (on the ROC) closest to the top left corner, usually represents the optimal threshold.
- But if we have a stronger desire for higher sensitivity (or higher specificity) we may choose another cut-off
- This is why the SHAPE of the ROC curve is important
  1. A strongly convex ROC curve (one with a high AUC that bulges towards the top left hand corner) is one where a gain in sensitivity can be expected without too much cost in terms of specificity (or conversely, we can gain in specificity, without losing too much sensitivity)
  2. An ROC curve with AUC close to 0.5 (a straight line with no convexity), more or less has an equivalent cost in specificity (sensitivity) for an equal gain in sensitivity (specificity)

Shape of ROC curve and AUC

White blood cell count



Temperature



What would you do???

Reporting results: Summarizing a diagnostic test evaluation

As a general convention, **seven** numbers should provided in any paper evaluating a diagnostic test:

- (1) Sensitivity and (2) specificity;
- (3) PV^+ and (4) PV^-;
- (5) LR_+ and (6) LR_-;
- (7) Area under the ROC curve (AUC).

Finally, the above assumes that we already possess a optimal test cut-point for the disease. If our test is on a continuous scale, we may need to identify this cut-point first(using the ROC curve).

Sample size calculations for diagnostic tests

- A large majority of sample size calculations I do in diagnostic test studies are for either Sensitivity, or Specificity (although it is possible to power a study on one of the other statistics)
- The first thing to note is that **both sensitivity and specificity are proportions**. Therefore, we can employ the standard sample size calculations for proportions
- **BUT WITH ONE SLIGHT BUT IMPORTANT MODIFICATION (see next)**
- The most common sample size calculations are for:
  1. **Precision of an estimate**: e.g. Estimate sensitivity with a particular level of precision
  2. **To compare two (new) diagnostic tests**: e.g. Is diagnostic test A more sensitive (or specific) than diagnostic test B?

Sample size for proportions:

To calculate the sample size to get a specific precision of proportion (i.e. with margin of error, E), we would use:

$$n = \frac{(Z_{1-\alpha/2})^2[p(1-p)]}{E^2}$$

The sample size required to compare two (independent sample) proportions is,

$$N_{pergroup} = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2(p_1(1-p_1) + p_2(1-p_2))}{MCD^2}$$

Note. The above formula assumes the two diagnostic tests are run of two different set of patients.

Now defining all the terms:

1. α (Significance level): How sure do we want to be to not falsely conclude a difference $\Rightarrow$ prob(Type I error) = prob(False positive)
2. β (for power): How sure do we want to be we can detect a difference (that is truly there). Note: Power = 1 - β where β = prob(Type II error) = prob(False negative)
3. The $p(1-p)$ in formula 1, and the $p_1(1-p_1)+p_2(1-p_2)$ in formula 2 represent the variance of the esitmate, σ^2.
4. The E in formula 1 (precision) represents how precise we want to be. Do we want our estimate to be within 1%, 2.5%, 5% etc?
5. MCD in formula 2 (only) represents the Minimal **clinical** difference: What would represent a **scientifically** meaningful improvement. Often represented by Δ or $p_1 - p_2$ (but this confuses people)

Modifying the 'proportions' formulae

There is one little problem with the above formulae. Let's look a bit closer (and we will adapt for sensitivity):

$$n = \frac{(Z_{1-\alpha/2})^2[p(1-p)]}{E^2} = \frac{(Z_{1-\alpha/2})^2[Sens(1-Sens)]}{E^2}$$

CAN YOU SEE THE PROBLEM???

Hint: Who is included in the caculation of sensitivity? All patients??

Sample size for diagnostic tests

- Only DISEASED patients are used in the calculation of sensitivity, and only NON-DISEASED patients in the calculations specificity
- So the formula on the previous page only tells us the number of CASES we need (not the number of PATIENTS)
- We need to modify this formula a bit to account for this.

So,

$$n_{cases} = \frac{(Z_{1-\alpha/2})^2[Sens(1-Sens)]}{E^2}$$

Is modified to:

$$n_{patients} = \frac{(Z_{1-\alpha/2})^2[Sens(1-Sens)]}{E^2 Prev}$$

The same for specificity

Similalry, we can modify the calculation for specificity:

$$n_{controls} = \frac{(Z_{1-\alpha/2})^2[Spec(1-Spec)]}{E^2}$$

Becomes:

$$n_{patients} = \frac{(Z_{1-\alpha/2})^2[Spec(1-Spec)]}{E^2(1-Prev)}$$

Now for a worked example....

Sample size example: Precision of sensitivity

We have a new diagnostic test and we would like to estimate the sensitivity to within 3% based on the 95% confidence interval. We believe the sensitvity of the test will be about 0.8, and we believe the prevalence of the disaese (in our clinical sample to be about 20%). Now we have everything we need.

1. Significance level: $\alpha = 0.05$ so we will use $Z_{1-\alpha/2} = Z_{0.975} = 1.96$ (as always).
2. We believe or sensitivity will be about $Sens = 0.8$
3. We believe the disease prevalence to be about $Prev = 0.2$
4. We want want our to be within 3% ($E = 0.03$)

Note 1: $Z_{0.975} = 1.96$ comes from the standard normal distribution.

Note 2: $E = 0.03$ implies our confidence interval will be 6% wide (i.e. $2E$)

Sample size: Putting it all together

Now:

$$n_{patients} = \frac{(Z_{1-\alpha/2})^2[Sens(1-Sens)]}{E^2 Prev}$$

$$n_{patients} = \frac{1.96^2[0.8(1-0.8)]}{0.03^2 0.2} = 3415$$

ARGH! We don't have 3415 patients. Let's try 5% precision (a 95%CI 10% wide):

$$n_{patients} = \frac{1.96^2[0.8(1-0.8)]}{0.05^2 0.2} = 1229$$

But this means that if our sensitivity is 75% the the 95%CI will be 70%, 80%. Not that precise.

Bringing it on home...

The last thing I would like to talk about is where diagnostic test evaluations sit in terms of 'scientific evidence'. If you were at my last lecture, I talked about three types of studies:

1. Exploratory studies;
2. Hyopthesis testing studies; and
3. Predictive studies

Where do you think Diagnostic test evaluations sit????

Clinical evidence and diagnostic test studies

- Diagnostic test attempt to **predict** the disease membership of **individual patients**, based on their clinical characteristics
- In this respect, p-values aren't enough
  - BMI is a well established risk factor of type 2 diabetes ($p < 0.05$), but we would hardly use BMI as a diagnostic test for T2DM
- For predictive studies we have MUCH higher standards. We want to show that our diagnostic test gets it right a (large) majority of the time (sensitivity and specificity).
- Routinely, we do this in a training sample (standard approach)
- But ideally, we would also want to establish this in a hold out ('test') sample to establish external validity.
- Some areas already require this (e.g. 'omics studies), and it is likely that in the future other disciplines will require this, too

THANK-YOU

Questions?????