Clinical biostatistics: Assessing agreement and diagnostic test evaluation

Dr Cameron Hurst
cphurst@gmail.com

DAMASAC and CEU, Khon Kaen University

26th September, 2557



What we will cover....

1. Introduction
   - Biostatistics vs. clinical biostatistics
   - Methods specific to clinical biostatistics
2. Measures of agreement: continuous
   - Continuous measures: Bland-Altman plots
   - Continuous measures: Intra-class correlation
3. Measures of agreement: categorical
   - Nominal outcomes
   - Ordinal measures
4. Assessing diagnostic tests
   - Sensitivity, specificity and predictive values
   - Likelihood ratio tests
   - Reciever Operator Characteristic (ROC) curves

Biostatistics vs. clinical biostatistics

- In this session and the next, I will cover the biostatistical methods that are used solely in clinical epidemiology
- The main reason for this, is that they don't really have a purpose outside the clinical setting (at least in the health sciences)
- Theoretically, they also don't really tie in that strongly with the 'modelling' approaches we will have covered in this lecture series

Clinical Biostatistical Methods

The **Clinical Biostatistical** methods we will cover today fall under the two closely related areas:

1. Assessing reliability \ agreement (this session) for
   - Continuous measures
   - Categorical measures
2. Assessing diagnostic tests (next session)

Assessing agreement

In clinical research, we often want to answer questions like:

- Is a new method (of measurement) good enough to replace an old one?
- Do measurements made by clinician A agree with those made by clinician B?

Analysis involving these **agreement issues** is called **reliability**, **repeatability**, **reproducibility** and/or **consistency** (depending on the exact setting in which the questions is asked)

Assessing agreement

The most common situation in which we want to assess agreement are as follows:

1. A few raters measure the same characteristic (on a group of subjects) with each rater measuring all subjects. In this case we want to see how much the different raters agree in terms of their measurements.
2. One rater successively measures a characteristic while the subjects' characteristic stay constant (Known as **intra-rater reliability**, **repeatability** or **reproducibility**)

Aside:

We should note that **rater** or **method** can refer to a person, a machine or a technical method

Assessing agreement

The are two main features we need to consider in any of the above situations:

1. **Bias: The degree to which two methods systematic disagree (i.e. consistently over- or under-estimation**
2. **Variation: How much 'random' noise there is**

We will consider methods to assess continuous, nominal and ordinal measures.

Methods for agreement of continuous measures

First we will consider measures of reliability on continuous measures. We will focus on two different approaches:

1. The graphical method: Bland-Altman plots; and
2. The statistic: The Intra-class correlation (ICC) coefficient

There is a lot of disagreement among users of these methods about which is the better approach. However, both methods have their advantages and limitations.

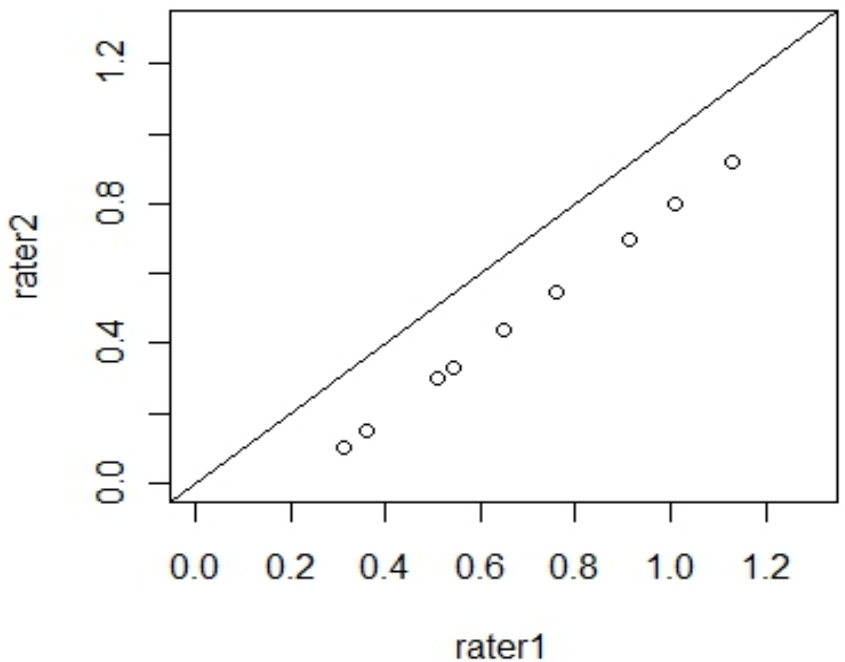
Pearson's correlation coefficient: A traditional but inappropriate measure of agreement

- In the past many studies used Pearson correlation coefficient, r, to assess the agreement between two raters
- The problem with this approach is two raters can be highly correlated, but have low agreement
- This happens when one rater **consistently** over- or underestimates the other
- That is, when there is **bias**

Problems with Pearson's correlation coefficient

Measures from 2 raters with perfect correlation (r=1)

We can see that Rater 1 consistently underestimates the measurements taken by Rater 2. This is example of **Bias**



Bland-Altman plots: 2 raters only

1. Calculate the difference between each rater's measure for each subject
2. Calculate the average scores (of 2 raters) for each subject
3. Generate plot of differences(3) against average score(2)
4. Include line representing the **overall average difference**
5. Plot the **limits of agreement** around the average difference

Interpretation:

- Values within **limits of agreement** $\Rightarrow$ 'agreement'
- Values outside $\Rightarrow$ lack of agreement
- A **overall average difference** substantially different from zero indicate **bias**
- Wide **limits of agreement** $\Rightarrow$ high **noise** (random variation)

Bland-Altman plots: Calculating the limits of agreement

The limits of agreement is very similar to a 95% confidence interval, except the **Standard deviation of the differences** rather than the standard error is used:

$$\bar{\delta} \pm 1.96 S_{\delta}$$

where:
$\bar{\delta}$ is the average difference between the raters' scores
S_{δ} is the standard deviation of the difference between the raters' scores

Example or Bland-Altman plot: Heart rate

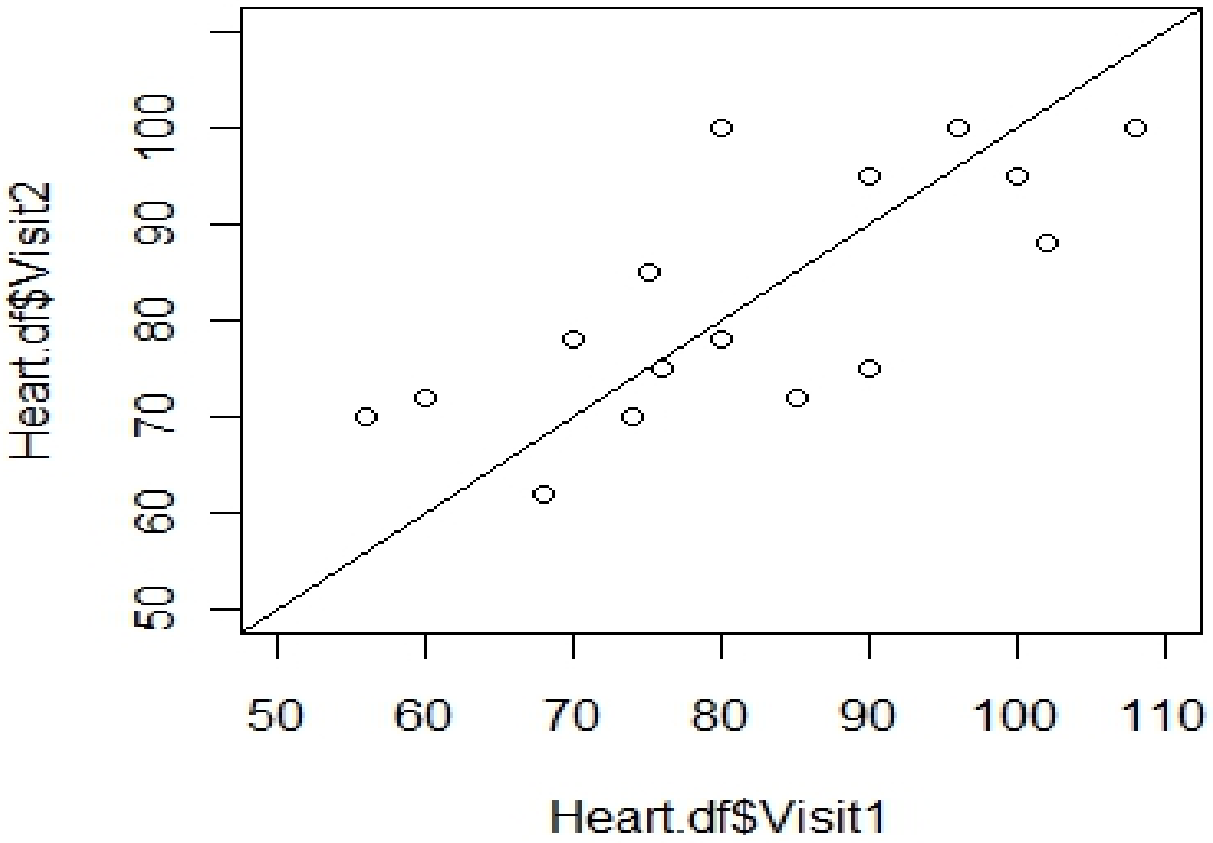
Here we will consider the measurement of heart rates of 16 patients taken by the same clinician, on two successive visits. Since we have a single clinician taking measurements on two occasions this is a **reproducibility** (or repeatability,or intra-rater agreement) study:

Patient ID	Visit 1	Visit 2	Patient ID	Visit 1	Visit 2
1	90	75	9	85	72
2	100	95	10	108	100
3	80	72	11	75	85
4	56	70	12	74	70
5	76	75	13	70	78
6	80	100	14	80	78
7	90	95	15	68	62
8	96	100	16	102	68

We will assume that there hasn't been any disease progression (or other relevant patient changes) between visits

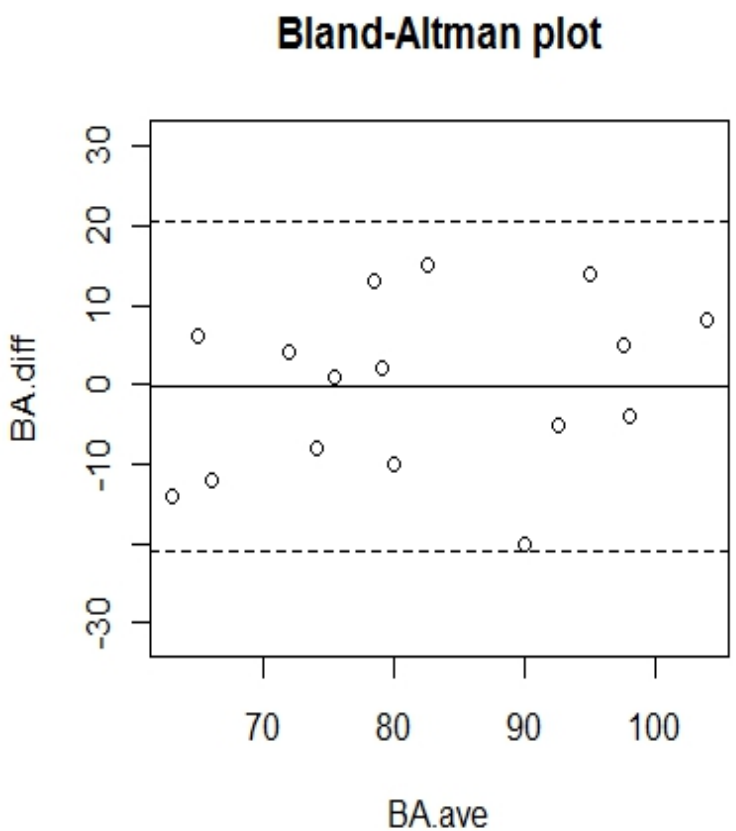
A preliminary perusal

Scatter plot of rater measures for two different visits of 16



patients

The Bland-Altman plot



Interpretation:

- Average difference close to zero suggests low bias
- Level of disagreement independent of level
- All values fall within limits of agreement (suggest low variation/noise)
- No outliers (to investigate)

Conclusion:

This rater demonstrated a high level of reproducibility

Pros and Cons of the Bland-Altman approach

BA-plots have some major advantages, but this method has also come under a lot of criticism.

- Strengths:
  - BA-plot easy and intuitive to understand
  - Able to peruse all observations (not just get a measure of the 'average level of agreement')
  - Can visualize both **Bias** and **Noise** from the plot
- Limitations:
  - As a graph, we don't have any 'magic cutoff boundry' as to what constitutes agreement and disagreement
    - i.e. Subjective: What one researcher might think of as agreement, might represent disagreement for another
  - Does not provide a single 'quotable' statistic (measure)

R code for Bland-Altman plot

None of the R libraries generates a Bland-Altman plot. I have written the following function to generate one (provided in R script file for this session)

```
BA.plot<-function(rater1, rater2)
{
BA.diff<-rater1-rater2
BA.ave<-(rater1 + rater2)/2
ave.diff<-mean(BA.diff)
sd.diff<-sd(BA.diff)
upper.lim<-ave.diff + 1.96* sd.diff
lower.lim<-ave.diff - 1.96* sd.diff
y.lim<-c((1.5*lower.lim), (1.5*upper.lim))
plot(BA.ave, BA.diff, main="Bland-Altman plot", ylim=y.l
abline(a=ave.diff, b=0)
abline(a=upper.lim, b=0, lty=2)
abline(a=lower.lim, b=0, lty=2)
}
```

R output

Now running it on our data, we generate the Bland-Altman plot from a few slides back:

```
#Read in data
setwd("c:\\myworkingdirectory")
Heart.df<-read.csv("HeartRateAgreement.csv")

#Use BA.plot function on the Heart rate data
BA.plot(Heart.df$Visit1, Heart.df$Visit2)
```

Intra-Class Correlation (ICC)

An alternative way of assessing agreement is the Intra-class correlation coefficient (ICC).

- Statistic representing the 'average' level of agreement
- Flexible: allows for greater than 2 raters and other reliability study designs
- However, this flexibility comes with a price: we have to think about and choose the right model
- ICC is not as intuitive as Bland-Altman plots
- There are a number of different models for ICC. EG:
  1. Simple m raters (random) problem
  2. m raters (random) by p methods (random)
  3. m raters (random) by p methods (fixed)
- We will only cover the first (simplest) case. See *Armitage and Berry*(1994) and *Chinn*(1990) for other models

The Intra-Class Coefficient (ICC)

- ICCs are essentially a statistic representing the proportion of variation of an observation due to subject-to-subject variability in error free scores
- For the above reason, ICCs can be calculated using various Analysis of Variance (ANOVA) models.
- We will consider the first case from the previous slide (m raters) which can be calculated using a *Components of Variance* model; A one way ANOVA model with a **Random** effect (Raters are considered a random sample from the *population of* raters)

Aside:

ICCs are also used in a totally different area of biostatistics: To represent (or measure) the level of within cluster association in multi-centre studies.

Analysis of components model for m-rater ICC

n subjects are rated by m raters (i.e. $n \times m$ observations). The Analysis of components (aka One-way random effects model) is given by:

$$Y_{ij} = \mu + s_i + \epsilon_{ij}$$

where $i = 1, 2, \cdots, n$ (number of patients); $j = 1, 2, \cdots, m$ (number of raters), μ is an unknown constant (mean of all observations), $s_i \sim N(0, \sigma_s^2)$ (random variation due to the subject) and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$

The components of variance ANOVA table:

Source of variation	df	MS	E(MS)
Between subjects	$n-1$	M_s	$\sigma_\epsilon^2 + m\sigma_s^2$
Residual	$n(m-1)$	M_r	σ_ϵ^2
Total	$nm-1$		

Analysis of components model for m-rater ICC

The quantity we are trying the estimate:

$$\rho_{ICC} = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_\epsilon^2}$$

Using the quantities given in the ANOVA table:

$$ICC = \hat{\rho}_{ICC} = \frac{M_s - M_r}{M_s + (m-1)M_r}$$

For two raters (i.e. $m = 2$)

$$ICC = \hat{\rho}_{ICC} = \frac{M_s - M_r}{M_s + M_r}$$

R code for ICC (nice and simple)

```
library(psychometric) #Contains ICC function
#Have to stack data in long format
#Combine values from visit1 and visit2 into a single col
heart.rate<-c(Heart.df$Visit1, Heart.df$Visit3)
#Generate a patient ID variable
patient <- rep(c(1:16), times=2)
#Combine into a data frame
hold.df<-data.frame(heart.rate, patient )
#Caluclate one way random effects ICC
ICC1.CI(dv=heart.rate, iv=patient, data=hold.df, level =
```

R Output

This results in:

```
  heart.rate, patient
       LCL       ICC1       UCL
1 0.2507445 0.6462212 0.8591235
```

Interpretation

With an ICC of 0.65 we would conclude that the ICC (and therefore reproducibility) is moderately strong. Generally:

- $ICC < 0.4 \Rightarrow$ weak
- $0.4 \leq ICC < 0.7 \Rightarrow$ moderate
- $ICC \geq 0.7 \Rightarrow$ strong

The 95 % confidence intervals $[0.25, 0.85]$ give us an idea of the certainty of our estimate. Narrower confidence intervals give a higher degree of certainty.

However, to test the hypothesis $H_0 : \rho_{ICC} = 0$ is not really meaningful. Although a 95%CI containing zero would give us cause for alarm.

Bland-Altman or ICC???

- As I mentioned previously, there is a lot of contention about whether the Bland-Altman or ICC approach is best
- My advice is to use both, only presenting the full Bland-Altman plot if it adds something to the story (i.e. Demonstrates the nature of bias or noise)

In terms of writing up the results I would use something like:

> *Both the Bland-Altman plot and Intraclass correlation coefficient(ICC) were used to evaluate the reliability of the XXX measure. The ICC showed moderately strong agreement between the measurements (ICC=0.65, 95%CI: 0.25, 0.86). The Bland-Altman plot supported this by showing a low degree of bias (average difference=-0.312) and no values falling outside the 95% lower and upper limits of agreement (-21.05, 20.43)...*

Assessing agreement for categorical measures

- Now let's consider how to assess the reliability of categorical measures
- We will consider two cases
  - Nominal measures (**no** basis for ordering responses)
  - Ordinal measures (**is** a basis for ordering responses)
- Fortunately there is a standard approach for doing this: Cohens' Kappa (κ).
- There are two variants of Cohens' Kappa. One to deal with the **nominal** case (**Cohens' Kappa**) and one for the **ordinal** case (**Cohens' Weighted Kappa**)

However, before we use these methods for assessing agreement for categorical measures, let's consider a naive (and inappropriate) approach for this purpose: the χ^2 test of independence

χ^2 test of independence based on cross-tabulation

Consider two situations (Table 1 and Table 2) where two raters are asked to diagnose a disease:

Table 1:

		Rater B	
		+	-
Rater A	+	75	25
	-	25	75

Table 2:

		Rater B	
		+	-
Rater A	+	50	50
	-	50	50

- Here there seems to be reasonable agreement. If we (inadvisedly) used χ^2 tes we might conclude that there is some agreement($\chi^2 = 48.02$, $df = 1$, $p < 0.0001$)
- Problem: *Isn't it possible that a large proportion of the classifications could have been correct just by chance??*

If we just got the two raters to 'flip of the coin' (as in **Table 2**) we would expect 50% agreement **just by chance**

Agreement by chance

This problem is exacerbated when we have diseases of very low (or high) prevalence. Consider a disease that has a prevalence of 0.01:

Table 3:

		Rater B	
		+	-
Rater A	+	0	1
	-	1	98

In this case, both raters failed to agree on any patients with the disease, but they still managed to agree on 98/100 cases.

We need a method that accounts for the **agreement just by chance**

Cohens' Kappa: Binary case

Consider a general table (similar to those above) representing the agreements and disagreements between raters on a two-point (Binary) scale:

		Rater B		
		+	-	
Rater A	+	a	b	$a+b$
	-	c	d	$c+d$
		$a+c$	$b+d$	n

One way of representing agreement would be to calculate:

$$I_o = \frac{a+d}{n}$$

I_o would be 0.75, 0.5 and 0.98 in Tables 1, 2 and 3 respectively. But as we have seen already, we need to account for the agreement by chance

Cohens' Kappa

Cohens' kappa allows us to calculate the agreement we would expect just by chance. As before **observed** agreement is given by:

$$I_o = \frac{a+d}{n}$$

and **chance** agreement can be calculated:

$$I_e = \frac{E(a)+E(d)}{n}$$

where $E(a) = \frac{(a+c)(a+b)}{n}$ and $E(d) = \frac{(c+d)(b+d)}{n}$ (as in a χ^2 test of independence: $\frac{row_{total} \times col_{total}}{overall_{total}}$)

Note:

Only expected frequencies of *agreement* cells are used to calculate I_e

Cohens' Kappa

Now Cohens' Kappa represents **the difference between the observed and expected frequencies as a fraction of the maximal difference**. This maximal difference also accounts for the agreement just by chance.

$$\kappa = \frac{\mathbf{I_o} - \mathbf{I_e}}{\mathbf{1} - \mathbf{I_e}}$$

Example: *tardive dyskinesia*

Two raters are asked to administer a new test to diagnose *tardive dyskinesia*, with the following results:

		Rater B		
		+	-	
Rater A	+	123	10	133
	-	6	29	35
		129	39	168

Now:

$$I_o = \frac{152}{168} = 0.905$$

$$I_e = \frac{\frac{(129)(133)}{168} + \frac{(39)(35)}{168}}{168} = 0.656$$

$$\kappa = \frac{0.905 - 0.656}{1 - 0.656} = 0.72$$

How high should Cohens' κ be?

A value of 0.72 suggests a reasonably high degree of reliability. We would conclude that the raters generally agree.

So what represents a high 'enough' level of Cohens' κ?

Fleiss(1999) suggests the following guidelines:

- $\kappa \leq 0.4 \Rightarrow$ *Poor*
- $0.4 < \kappa \leq 0.75 \Rightarrow$ *Fair to Good*
- $\kappa > 0.75 \Rightarrow$ *Excellent*

Cohens' Kappa for nominal outcomes

It is quite simple to extend the binary form of Cohens' Kappa to the nominal (>2 class) problem

		Rater B		
		absent	typeR	typeQ
Rater A	absent	a	b	c
	typeR	d	e	f
	typeQ	h	i	j

Gives: $I_o = \frac{a+e+j}{n}$ and $I_e = \frac{E(a)+E(e)+E(j)}{n}$

and as before: $\kappa = \frac{I_o - I_e}{1 - I_e}$

Note:

Again note that only the frequencies of the 'agreement cells' are used in both I_o and I_e

Ordinal outcomes and Cohens' weighted Kappa

- Often we are presented with the case where our measurement scale is ordinal:
  - *Mild, Moderate, Severe*
  - *Absent, Benign, Suspect, Cancer*
- Cohens' Kappa can be simply extended to account for this ordering
- Basic idea: Increase the amount of penalty with higher levels of disagreement
  - For example, we might penalize disagreements in diagnoses two categories apart (e.g. Mild vs Severe) twice as highly as those adjacent (Mild vs Moderate and Moderate vs Severe) which we might term **partial agreement**
- This idea is implemented in the method: **Cohens' Weighted Kappa**

Calculation of Cohens' Weighted Kappa

- I wont go into too much details about how the statistic is calculated (those interested are directed to Armitage and Berry, 1994)
- The main innovation in this method is that both the observed and expected agreements are calculated using weights that reflect the level of agreement (i.e. full agreement, partial agreement and full disagreement)
- Higher levels of disagreement are penalized higher (weighed lower) than lower levels of disagreement (partial agreement)
- The main decision to make is how much to penalize for different levels of (dis)agreement
- A number of ways to do this. A common approach used is the **equally spaced** penalty (see below)

Agreement, partial agreement and disagreement

I will illustrate how the Weighted κ statistic works using a number of examples. Consider a 2 rater by 3 point ordinal scale of disease severity:

		Rater B		
		mild	intermediate	severe
Rater A	mild	a	b	c
	intermediate	d	e	f
	severe	h	i	j

The 'standard' Cohens' Kappa

We might decide to use a 'standard' κ statistic (and assume all disagreement is equally bad). That is:

Weight table: unweighted κ

		Rater B		
		mild	intermediate	severe
Rater A	mild	1	0	0
	intermediate	0	1	0
	severe	0	0	1

That is, in this '**weight table**' (**the unweighted case**) there is either 'full agreement' (with weight of 1), or 'disagreement' (weight of 0).

Equally-spaced Cohens' weighted kappa

Or we might use an 'equally spaced' approach where closer disagreements (partial agreement) is weighted more highly (penalized less) than full disagreements:

Weight table: equally spaced κ for 3 x 3 problem

		Rater B		
		mild	intermediate	severe
Rater A	mild	1	0.5	0
	intermediate	0.5	1	0.5
	severe	0	0.5	1

Aside:

Weights can be worked out for the k point ordinal scale problem using: $w_i = 1 - \frac{i}{k-1}$

Asymmetric weight tables

- We may also want to penalize more highly particular types of disagreements
- For example, consider a situation where patients classified as moderate or severe are triaged to further testing, whereas those classified as mild aren't
- In this case we would want to **highly** penalize a disagreement involving a mild classification:

Weight table: Example of asymmetric weighting

		Rater B		
		mild	intermediate	severe
Rater A	mild	1	0.3	0
	intermediate	0.3	1	0.8
	severe	0	0.8	1

Example: Patient vs Nurse rating of cholesterol levels

In this example, 40 patients self-rated themselves as having **low** (chol < 3.8), **high** (chol $\epsilon(3.8, 42]$ or **very high** ((chol ≥ 4.2) cholesterol levels. Nurses were then asked to classify these patients cholesterol levels (using the same method) with the following results: **Patient vs nurse cholesterol**

		Patients			
		low	high	very high	
Nurses	low	17	0	0	**Eyeball:**
	high	4	6	1	
	very high	1	7	4	

Summarize what you think is happening here

Nurses are overestimating (and\or patients are underestimating) cholesterol levels.

Now let's use R to calculate:

1. The standard (unweighted) κ
2. The equally-spaced Weighted κ
3. An example of a asymmetric weighted κ

I will use the *epicalc* library developed by Prof Virasakdi Chongsuvivatwong at PSU

Unweighted (nominal) case

```
#Assume data set with ratings read in
library(epicalc)

#Cross tabulate patient and nurse ratings
my.tab<-table(Chol.df$nurse.rate,
+ Chol.df$patient.rate, dnn=c("Nurses","Patients"))

#Run unweighted Kappa
kap(my.tab)
```

Results (Unweighted Kappa):

```
   Patients
Nurses  0  1  2
     0 17  0  0
     1  4  6  1
     2  1  7  4

Observed agreement = 67.5 %
Expected agreement = 36.06 %
Kappa = 0.492
Standard error=0.109, Z=4.523, P value<0.001
```

- We would say the agreement was on the lower end to moderate
- Note that here all types of disagreement are just as bad
- The hypothesis test ($H_0 : \kappa = 0$) is rather meaningless

Equally-spaced Kappa:

R code:

```
#wttable="w" means use equally-spaced weight table
kap(my.tab, wttable = "w")
```

Results:

```
      Patients
Nurses  0  1  2
     0 17  0  0
     1  4  6  1
     2  1  7  4

Observed agreement = 82.5 %
Expected agreement = 57.12 %
Kappa = 0.592
Standard error = 0.117, Z = 5.05, P value < 0.001
```

- By accounting for ordinality, we can see an improvement ($\kappa = 0.592$), which comfortably falls in the 'good range'
- In this case disagreements in close proximity are penalized less than those further apart

Asymmetric weighted Kappa:

```
my.wts <-as.table(rbind(c(1,0.8,0), c(0.3,1,0.8), c(0, 0
my.wts
    A   B   C
A 1.0 0.8 0.0
B 0.3 1.0 0.8
C 0.0 0.3 1.0
kap(my.tab, wttable = my.wts)
      Patients
Nurses  0  1  2
     0 17  0  0
     1  4  6  1
     2  1  7  4
Observed agreement = 77.75 %
Expected agreement = 57.32 %
Kappa = 0.479
Standard error = 0.115, Z = 4.178, P value < 0.001
```

- Disagreements involving mild levels are penalized more highly

- Not as successful, but maybe a better measure of agreement

A last word on Cohens' κ

- The choice of, and if and how we should weight our disagreements, is one that should be governed by 'design'
- We should **not** just choose the result that sells our idea better
- Cohens' Kappa is a widely used and standard measure of gauging agreement between raters using categorical instruments
- It is quite robust, but we should note that it is susceptible to the very low (very high) prevalence problem (that I mentioned before)
- Some modifications have been proposed to the Kappa the circumvent this problem (e.g See Byrt et al, 1993)

Diagnostic tests

- Now 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 agreement of a binary instrument for two raters
- BUT, there is a very important difference:

Key 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 (**infallible**).

Diagnostic test accuracy

Let's start by looking at the elements of a new 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

Test accuracy, Sensitivity and pecificity

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

The numbers on the main diagonal represents the patients classified correctly. Similar to agreement (in reliability studies) the probabilities of these main diagonals jointly represent the **test accuracy**:

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

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 the **without** a disease

Example: Sensitivity and Specificity

We have 100 patients **known** to have a disease and 1000 to be disease-free(from 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:

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

Predictive value of a test

- Another set of quantities can also be calculated in this situation: 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*

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: $P(D^-|T^-)$

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

Using Bayes' theorem:

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

$$PV^+ = P(D^+|T^+) = \frac{TP}{TP + FP} \quad (1)$$

and

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

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

Prevalence and study design

- Perusal of the above equations show that sensitivity and specificity are directly related to (PV^+) and (PV^-)
- We should also note that the disease prevalence, $P(D^+)$ (and indirectly via $P(D^-) = 1 - P(D^+)$) is present in the calculations. This leads to a number of implications:
  1. If we want to use our sample to estimate $P(D^+)$ using $\frac{D^+}{n}$ our sample should be representative (i.e. cross-sectional) of the target population.
  2. If this isn't the case, equations (1) and (2) (the quick way of calculating PV^+ and PV^-) aren't valid.
  3. If so, we need to estimate prevalence from other sources and use the long form to calculate PV^+ and PV^-
  4. Finally, this also implies that PV^+ and PV^- (as good estimates) are **susceptible to levels of disease prevalence**

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:

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

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 ($P(T^-|D^-) = 0.8$). 80% of individuals without the disease were correctly identified
3. The positive predictive value was not very high ($PV^+ = 0.31$) meaning that the probability of a patient with a positive test actually having the disease is only 0.31
4. The probability of an individual with a negative test not having the disease ($PV^- = 0.988$) was very high.

Note:

The major difference between $PV^+ = 0.31$ and $PV^- = 0.988$ relates to the low prevalence ($P(D^+) = 0.09091$). A smaller prevalence (holding sensitivity and specificity constant) would have resulted in even a more pronounced difference

Likelihood ratios

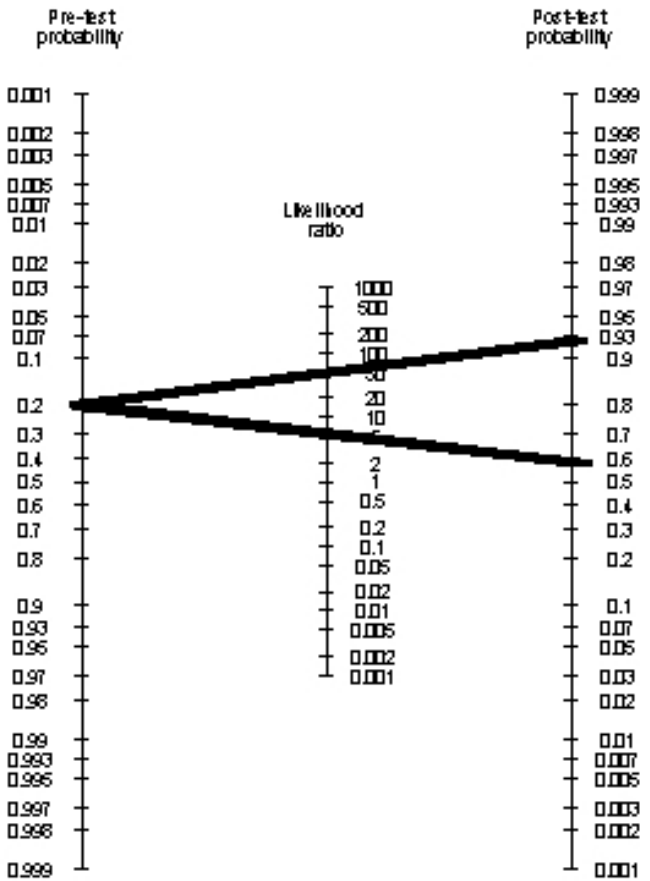
- Likelihood ratios are an alternative way of gauging the success (or otherwise) of a diagnostic tests
- Like the predictive values (PV_+ and PV_-) they have a positive and negativeversion(LR_+ and LR_-)
- Basic idea: Use the **pre-test odds** and the **likelihood ratio** to get the **post-test odds**
- Anadvantage of the likelihood ratios is that we can use it for different 'cut-offs'

Why use Likelihood ratios?

- Why master the concept of likelihood ratios when they are more complicated than prevlaence, sensitivity, specificity and the predicitve values?
- ANS: They go beyond the 'less informative' measures of sensitivity and specificity to account for the 'acutal value' rather than just 'above' or 'below' the cutoff
  - For example, somebody with a VERY abnormal value is more likely to have a disease than someone JUST ABOVE the cutoff
- So with the LR is is possible to summarize information for test results at different values

Using likelihood ratios

One of the best ways of present the information contained in a LR is through the **nomogram**



1. Given a pre-test probability (prevalence) of 0.2 and $LR_+ = 5$, the pro-test probablility would be 0.6
2. Given a pre-test probability (prevalence) of 0.2 and $LR_+ = 50$, the pro-test probablility would be 0.93

Relative importance of Sensitivity and Specificity

- Often the clinical measure(s) used for diagnostic tests are measured on a continuous (or close to continuous) scale
- A threshold can then be chosen where we decide a subject is diseased or disease free
- Usually choices of threshold value can be weighed in how we choose the cut off noting:
  - (Usually) increasing test sensitivity implies a cost in terms of test specificity
  - (Usually) increasing test specificity occurs at the expense of sensitivity
- How we balance these two aspects should depend (mainly) on our research objectives. For example:
  - For population surveillance of serious infectious diseases we may want higher sensitivity at the cost of lower specificity

Back to the 2 x 2 table

So let's go back to our 2 x 2 table. We have a continuous clinical measure and we want to choose a cut-off, that best suits our research objective

		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 ($\uparrow$ sensitivity) but we are likely to let in some false positive ($\downarrow$ specificity)
- Conversely, if we increase the threshold, we would be able exclude people without the disease ($\uparrow$ specificity), but may also miss some diseased individuals ($\downarrow$ sensitivity)

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 in children in a majority of cases is a result of self-limiting viral infections, but a minority 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.

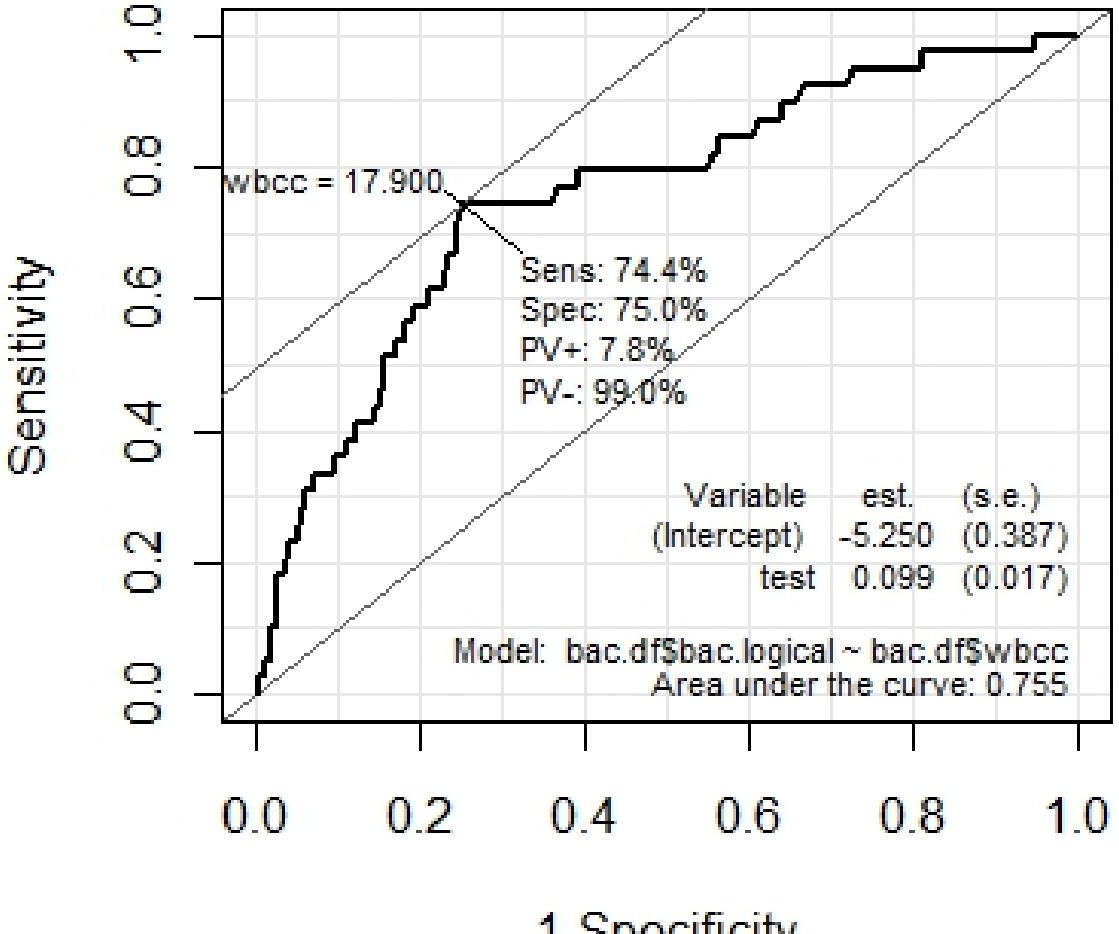
For example:

1. The consequences of not treating a child with bacteraemia are potentially life-threatening; but also
2. Over-use of antibiotics also presents a problem

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**

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. AUC: 0.755

Interpretation

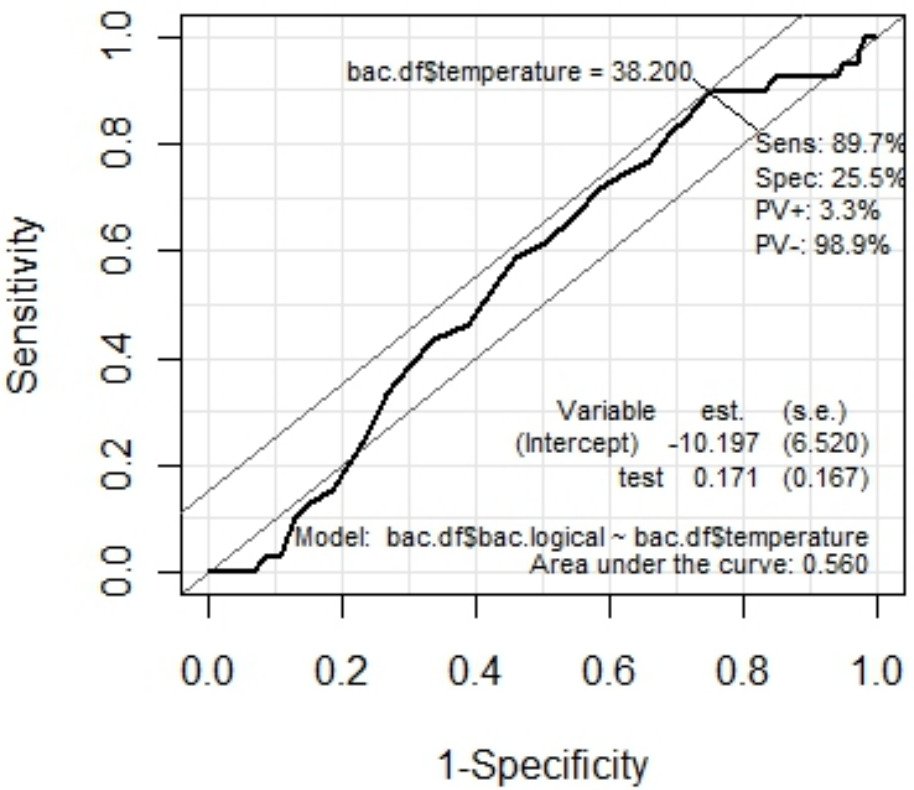
- If we use a white blood cell cut-off of 17.9 $10^9 L^{-1}$ (1:wbcc) we would expect to (correctly) identify 74.4% of bacteraemia children (2:sensitivity) and correctly exclude (negative test) 75% of non-bacteraemic children (3: specificity)
- The probability of a child testing positive for bacteraemia and really being bacteraemic is only 0.078 (4: PV^+)
- and the probability of a child testing negative for bacteraemia actually not having the condition is 0.99 (5: PV^+)
- The Area under the ROC curve (6: **AUC**) of 0.755 shows that white blood cell count is quite an effective measure to discriminate between children with and without bacteraemia
- Also note the **convex shape** of the ROC (bulge towards upper left quadrant) suggesting a good diagnostic test.

AUC: Area under (ROC) curve

- As indicated above, AUC tells us about a clinical measure's ability (independent of cut-off) to discriminate between those with a disease and those disease free.
- So AUC represents an 'overall' measure of test accuracy
- **The closer an AUC is to 1, the better the test**
  - A diagnostic test whose ROC curve has a pronounced convex bulge towards the upper right, will have a higher AUC
- **The closer the AUC to 0.5, the worse the test**
  - Test's without the pronounced bulge will have an AUC close to 0.5, and will have low test accuracy

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. AUC: 0.56

Interpretation

- Straight away from the shape of the ROC curve we can see temperature is not a good diagnostic test. 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) with 89.7% of bacteraemic children testing postive
- In sharp contrast, only 25.5% of bacteraemia-free children tested negative (3:specificity)
- Of those that tested positive, 3.3% were actually bacteraemic (4:PV^-)
- But of those that tested negative (98.9%) were not bacteraemic

Question for you to consider

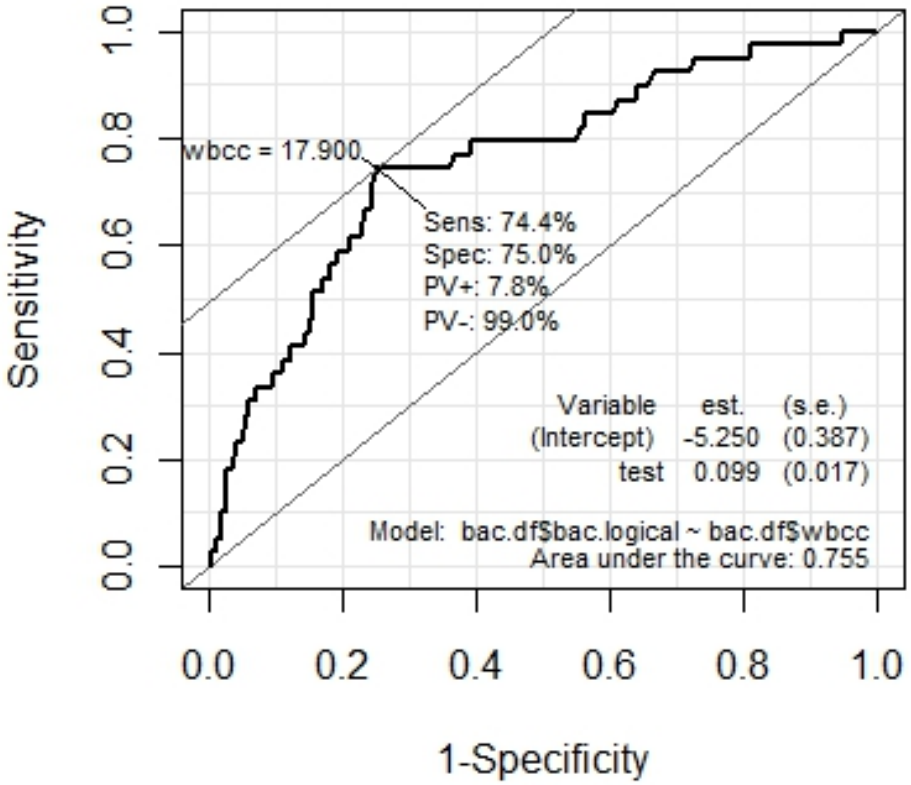
1. For both the **White blood cell count** and **Temperature** diagnostic tests, PV^+ was very low, and PV^- very high, what might explain this?
2. If you take WBCC to be the better test, are you happy with the 'optimal' threshold ($17.9\ 10^9 L^{-1}$)?
   - When answering this question: think about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

Finding an appropriate cut-off

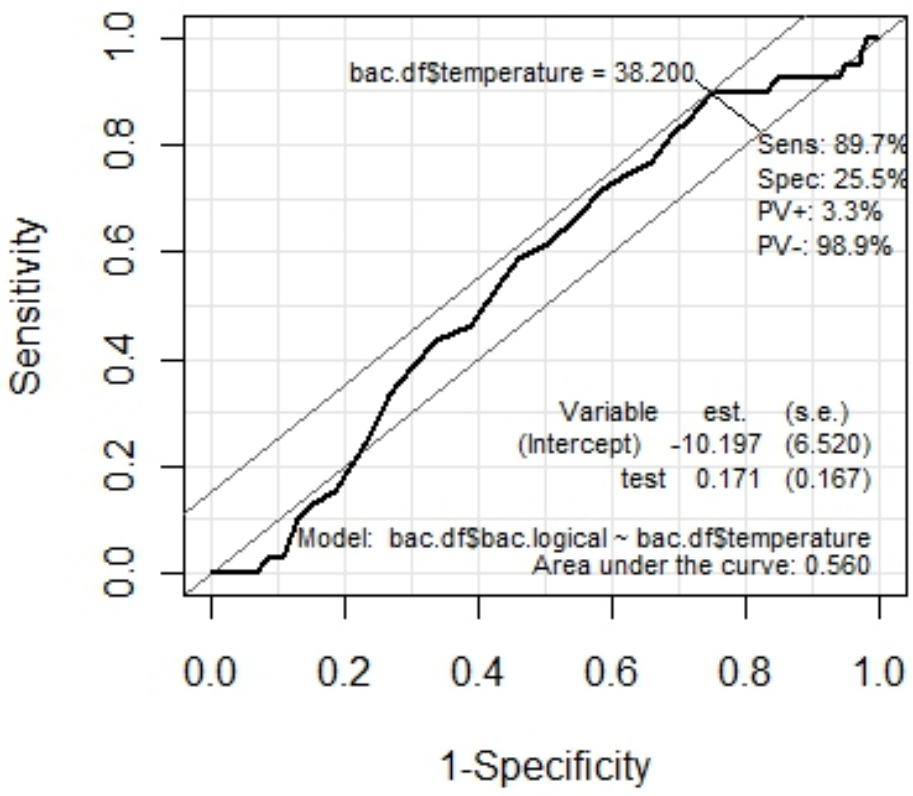
- If we consider both sensitivity and specificity as equally important, the point (on the ROC) closest to the top left corner, usually represents the optimal threshold.
- But if we have a stronger desire for higher sensitivity (or higher specificity) we may choose another cut-off
- This is where the 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 (and has 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???

R code for generating ROC curves:

```
library(Epi)

setwd("D:\\mydirectory")
bac.df<-read.csv("bacteremia.csv")

#Define disease status
bac.df$bac.logical<-(bac.df$bacteremia==1)

#White blood cell count
ROC.wbcc<-ROC(test=bac.df$wbcc, stat=bac.df$bac.logical, plot="ROC")
ROC.wbcc

#Temperature
ROC.temp<-ROC(test=bac.df$temp, stat=bac.df$bac.logical, plot="ROC")
ROC.temp
```

References and resources

Armitage, P. and Berry, G. (1994) *Statistical Methods in Medical Research (3ed)*

Byrt, T., Bishop, J., and Carlin, J. (1993) Bias, Prevalence, and Kappa. *Journal of Clinical Epidemiology* **46(5)**, 423-429

Chinn, S.(1990) The assessment of methods of measurement. *Statistics in Medicine* **9**, 351-362

Fleiss, J. L. (1999) *The Design and Analysis of Clinical Experiments*, John Wiley & Sons, Inc., Hoboken, NJ, USA.

THANK-YOU

Questions?????