A generalized meta-analysis model for binary diagnostic test performance

Ben A. Dwamena, MD

The University of Michigan & VA Medical Centers, Ann Arbor

FNASUG - November 13, 2008
DIAGNOSTIC TEST
Any measurement aiming to identify individuals who could potentially benefit from preventative or therapeutic intervention. This includes:

1. Elements of medical history
Any measurement aiming to identify individuals who could potentially benefit from preventative or therapeutic intervention. This includes:

1. Elements of medical history
2. Physical examination
DIAGNOSTIC TEST
Any measurement aiming to identify individuals who could potentially benefit from preventative or therapeutic intervention. This includes:

1. Elements of medical history
2. Physical examination
3. Imaging procedures
Diagnostic Test Evaluation

**DIAGNOSTIC TEST**
Any measurement aiming to identify individuals who could potentially benefit from preventative or therapeutic intervention. This includes:

1. Elements of medical history
2. Physical examination
3. Imaging procedures
4. Laboratory investigations
DIAGNOSTIC TEST

Any measurement aiming to identify individuals who could potentially benefit from preventative or therapeutic intervention. This includes:

1. Elements of medical history
2. Physical examination
3. Imaging procedures
4. Laboratory investigations
5. Clinical prediction rules
The performance of a diagnostic test assessed by comparison of index and reference test results on a group of subjects.

**Binary test data often reported as $2 \times 2$ matrix**

<table>
<thead>
<tr>
<th></th>
<th>Reference Test Positive</th>
<th>Reference Test Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Positive</td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td>Test Negative</td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
</tbody>
</table>
The performance of a diagnostic test assessed by comparison of index and reference test results on a group of subjects.

Ideally these should be patients suspected of the target condition that the test is designed to detect.

**Binary test data often reported as $2 \times 2$ matrix**

<table>
<thead>
<tr>
<th></th>
<th>Reference Positive</th>
<th>Test Positive</th>
<th>Reference Negative</th>
<th>Test Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Positive</td>
<td>True Positive</td>
<td></td>
<td>False Positive</td>
<td></td>
</tr>
<tr>
<td>Test Negative</td>
<td>False Negative</td>
<td></td>
<td>True Negative</td>
<td></td>
</tr>
<tr>
<td>Measure</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity (true positive rate)</td>
<td>The proportion of people with disease who are correctly identified as such by test</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specificity (true negative rate)</td>
<td>The proportion of people without disease who are correctly identified as such by test</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive predictive value</td>
<td>The proportion of test positive people who truly have disease</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative predictive value</td>
<td>The proportion of test negative people who truly do not have disease</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Measures of Diagnostic Performance

Likelihood ratios (LR)  The ratio of the probability of a positive (or negative) test result in the patients with disease to the probability of the same test result in the patients without the disease.

Diagnostic odds ratio  The ratio of the odds of a positive test result in patients with disease compared to the odds of the same test result in patients without disease.

ROC Curve  Plot of all pairs of (1-specificity, sensitivity) as positivity threshold varies.
Rationale

1. Evaluation of the quality and scope of available primary studies
Meta-analysis of Diagnostic Performance

Rationale

1. Evaluation of the quality and scope of available primary studies

2. Determination of the proper and efficacious use of diagnostic and screening tests in the clinical setting in order to guide patient treatment
Meta-analysis of Diagnostic Performance

Rationale

1. Evaluation of the quality and scope of available primary studies

2. Determination of the proper and efficacious use of diagnostic and screening tests in the clinical setting in order to guide patient treatment

3. Decision making about health care policy and financing
Meta-analysis of Diagnostic Performance

Rationale

1. Evaluation of the quality and scope of available primary studies

2. Determination of the proper and efficacious use of diagnostic and screening tests in the clinical setting in order to guide patient treatment

3. Decision making about health care policy and financing

4. Identification of areas for further research, development, and evaluation
Meta-analysis of Diagnostic Performance

Major steps

1. Framing objectives of the review
Meta-analysis of Diagnostic Performance

Major steps

1. Framing objectives of the review
2. Identifying the relevant literature
Meta-analysis of Diagnostic Performance

Major steps

1. Framing objectives of the review
2. Identifying the relevant literature
3. Assessment of methodological quality and applicability to the clinical problem at hand
Meta-analysis of Diagnostic Performance

Major steps

1. Framing objectives of the review
2. Identifying the relevant literature
3. Assessment of methodological quality and applicability to the clinical problem at hand
4. Summarizing the evidence qualitatively and if appropriate, quantitatively (meta-analysis)
Meta-analysis of Diagnostic Performance

Major steps

1. Framing objectives of the review
2. Identifying the relevant literature
3. Assessment of methodological quality and applicability to the clinical problem at hand
4. Summarizing the evidence qualitatively and if appropriate, quantitatively (meta-analysis)
5. Interpretation of findings and development of recommendations
Validity of Meta-analysis of Diagnostic Test Accuracy

Depends on presence, extent and sources of variability due to:

1. Methodological quality bias
Validity of Meta-analysis of Diagnostic Test Accuracy

Depends on presence, extent and sources of variability due to:

1. Methodological quality bias
2. Covariate Heterogeneity
Validity of Meta-analysis of Diagnostic Test Accuracy

Depends on presence, extent and sources of variability due to:

1. Methodological quality bias
2. Covariate Heterogeneity
3. Publication and other sample size-related bias
Validity of Meta-analysis of Diagnostic Test Accuracy

Depends on presence, extent and sources of variability due to:

1. Methodological quality bias
2. Covariate Heterogeneity
3. Publication and other sample size-related bias
4. Threshold Effects
Validity of Meta-analysis of Diagnostic Test Accuracy

Depends on presence, extent and sources of variability due to:

1. Methodological quality bias
2. Covariate Heterogeneity
3. Publication and other sample size-related bias
4. Threshold Effects
5. Unobserved heterogeneity
Extent of Heterogeneity

1 Assessed statistically using the quantity $I^2$ described by Higgins and Colleagues (2002).

$I^2$ is calculated as:

$$I^2 = \left( \frac{Q - df}{Q} \right) \times 100.$$ (1)

$Q$ is Cochran's heterogeneity statistic; $df$ equals degrees of freedom.

$I^2$ lies between 0% and 100%: 0% indicates no observed heterogeneity, greater than 50% considered substantial heterogeneity.

Advantage of $I^2$: does not inherently depend on the number of the studies.
Extent of Heterogeneity

1. Assessed statistically using the quantity $I^2$ described by Higgins and Colleagues (2002).

2. Defined as percentage of total variation across studies attributable to heterogeneity rather than chance.
Extent of Heterogeneity

1. Assessed statistically using the quantity $I^2$ described by Higgins and Colleagues (2002).

2. Defined as percentage of total variation across studies attributable to heterogeneity rather than chance.

3. $I^2$ is calculated as:

$$I^2 = \left(\frac{Q - df}{Q}\right) \times 100. \quad (1)$$

$Q$ is Cochran’s heterogeneity statistic; $df$ equals degrees of freedom.
Extent of Heterogeneity

1. Assessed statistically using the quantity $I^2$ described by Higgins and Colleagues (2002).

2. Defined as percentage of total variation across studies attributable to heterogeneity rather than chance.

3. $I^2$ is calculated as:

$$I^2 = \left(\frac{Q - df}{Q}\right) \times 100. \quad (1)$$

$Q$ is Cochran’s heterogeneity statistic; $df$ equals degrees of freedom.

4. $I^2$ lies between 0% and 100%: 0% indicates no observed heterogeneity, greater than 50% considered substantial heterogeneity.
Extent of Heterogeneity

1. Assessed statistically using the quantity $I^2$ described by Higgins and Colleagues (2002).

2. Defined as percentage of total variation across studies attributable to heterogeneity rather than chance.

3. $I^2$ is calculated as:

$$I^2 = \left( \frac{Q - df}{Q} \right) \times 100. \quad (1)$$

$Q$ is Cochran’s heterogeneity statistic; $df$ equals degrees of freedom.

4. $I^2$ lies between 0% and 100%: 0% indicates no observed heterogeneity, greater than 50% considered substantial heterogeneity.

5. Advantage of $I^2$: does not inherently depend on the number of the studies.
There are different sources of heterogeneity in meta-analysis: characteristics of the study population, variations in the study design (type of design, selection procedures, sources of information, how the information is collected), different statistical methods, and different covariates adjusted for (if relevant).
Sources of Heterogeneity: Meta-regression

1. There are different sources of heterogeneity in meta-analysis: characteristics of the study population, variations in the study design (type of design, selection procedures, sources of information, how the information is collected), different statistical methods, and different covariates adjusted for (if relevant).

2. Formal investigation of sources of heterogeneity is performed by meta-regression, a collection of statistical procedures (weighted/unweighted linear, logistic regression) in which the study effect size is regressed on one or several covariates.
Methodological Quality

The assessment of quality has to consider details of study design and execution such as:

1. Cogency of the research question and clinical context
Methodological Quality

The assessment of quality has to consider details of study design and execution such as:

1. Cogency of the research question and clinical context
2. Appropriateness of patient population
The assessment of quality has to consider details of study design and execution such as:

1. Cogency of the research question and clinical context
2. Appropriateness of patient population
3. Sufficient description and well-defined interpretation of index diagnostic technique(s)
Methodological Quality

The assessment of quality has to consider details of study design and execution such as:

1. Cogency of the research question and clinical context
2. Appropriateness of patient population
3. Sufficient description and well-defined interpretation of index diagnostic technique(s)
4. Appropriateness and sufficient description of reference standard information
Methodological Quality

The assessment of quality has to consider details of study design and execution such as:

1. Cogency of the research question and clinical context
2. Appropriateness of patient population
3. Sufficient description and well-defined interpretation of index diagnostic technique(s)
4. Appropriateness and sufficient description of reference standard information
5. Other factors that can affect the integrity of the study and the generalizability of the results
Methods of quality assessment may focus on:

1. Absence or presence of key qualities in the study report (checklist approach)
Methods of quality assessment may focus on:

1. Absence or presence of key qualities in the study report (checklist approach)

2. Scores developed for this purpose (scale approach)
Methods of quality assessment may focus on:

1. Absence or presence of key qualities in the study report (checklist approach)

2. Scores developed for this purpose (scale approach)

3. Levels-of-evidence methods by which a level or grade is assigned to studies fulfilling a predefined set of criteria
Threshold effects

1. Most diagnostic tests have multiple or continuous outcomes
1. Most diagnostic tests have multiple or continuous outcomes.

2. Dichotomization or application of cutoff value used to classify results into positive or negative.
Threshold effects

1. Most diagnostic tests have multiple or continuous outcomes

2. Dichotomization or application of cutoff value used to classify results into positive or negative

3. **Implicit positivity threshold**: based on interpretation/judgement/machine calibration e.g. radiologists classifying images as normal or abnormal
Threshold effects

1. Most diagnostic tests have multiple or continuous outcomes

2. Dichotomization or application of cutoff value used to classify results into positive or negative

3. **Implicit positivity threshold**: based on interpretation/judgement/machine calibration e.g. radiologists classifying images as normal or abnormal

4. **Explicit positivity threshold**: based on a numerical threshold e.g. blood glucose level above which patient may be said to have diabetes
The chosen threshold may vary between studies of the same test due to inter-laboratory or inter-observer variation.
1. The chosen threshold may vary between studies of the same test due to inter-laboratory or inter-observer variation

2. The higher the cut-off value, the higher the specificity and the lower the sensitivity
Threshold effects

1. The chosen threshold may vary between studies of the same test due to inter-laboratory or inter-observer variation.

2. The higher the cut-off value, the higher the specificity and the lower the sensitivity.

3. Threshold-based interdependence between sensitivity and specificity tested \textit{a priori} using a rank correlation test such as Spearman’s rho after logit transformation.
Publication bias Tendency for investigators, reviewers, and editors to submit or accept manuscripts for publication based on the direction or strength of the study findings.

Funnel plot Exploratory tool for investigating publication bias, plotting a measure of effect size versus a measure of study precision

1 Funnel plot should appear symmetric if no bias is present
Publication bias: Tendency for investigators, reviewers, and editors to submit or accept manuscripts for publication based on the direction or strength of the study findings.

Funnel plot: Exploratory tool for investigating publication bias, plotting a measure of effect size versus a measure of study precision.

1. Funnel plot should appear symmetric if no bias is present.
2. Assessment of such a plot is very subjective.
Publication and Other Precision-related Biases

**Publication bias**  
Tendency for investigators, reviewers, and editors to submit or accept manuscripts for publication based on the direction or strength of the study findings.

**Funnel plot**  
Exploratory tool for investigating publication bias, plotting a measure of effect size versus a measure of study precision.

1. Funnel plot should appear symmetric if no bias is present.
2. Assessment of such a plot is very subjective.
3. Non-parametric and linear regression methods used to formally test funnel plot asymmetry.
Examples of Tests For Funnel Plot Asymmetry

(Begg 1994) Rank correlation between standardized effect and its standard error

(Egger 1997) Linear regression of intervention effect against its standard error weighted by inverse of the variance of intervention effect estimate

(Macaskill 2001) Linear regression of intervention effect on sample size

(Harbord 2006) Modified version of (Egger 1997) based on ”score” and ”score variance” of the log odds ratio

(Peters 2006) Linear regression of intervention effect on inverse of sample size
Problems with sample size and standard error

1. The asymptotic standard error is a biased estimate of the true standard error, with larger bias for smaller cell sizes, as occurs with larger DORs and smaller studies.
Problems with sample size and standard error

1. The asymptotic standard error is a biased estimate of the true standard error, with larger bias for smaller cell sizes, as occurs with larger DORs and smaller studies.

2. Diagnostic studies have unequal sample sizes in diseased and non-diseased groups which reduces the precision of an estimate of test accuracy for a given sample size.
Problems with sample size and standard error

1. The asymptotic standard error is a biased estimate of the true standard error, with larger bias for smaller cell sizes, as occurs with larger DORs and smaller studies.

2. Diagnostic studies have unequal sample sizes in diseased and non-diseased groups which reduces the precision of an estimate of test accuracy for a given sample size.

3. The standard error of the logDOR depends on proportion testing positive. However, individual studies often differ in positivity threshold leading to variability in proportion testing positive.
The most commonly used and easy to implement method

1 Linear regression analysis of the relationship
\[ D = a + bS \]
where:
\[ D = (\text{logit TPR}) - (\text{logit FPR}) = \ln \text{DOR} \]
\[ S = (\text{logit TPR}) + (\text{logit FPR}) = \text{proxy for the threshold} \]

The most commonly used and easy to implement method

1. Linear regression analysis of the relationship
   \[ D = a + bS \]
   where:
   \[ D = (\logit \ TPR) - (\logit \ FPR) = \ln \ DOR \]
   \[ S = (\logit \ TPR) + (\logit \ FPR) = \text{proxy for the threshold} \]

2. \( a \) and \( b \) may be estimated by weighted or unweighted least squares or robust regression, back-transformed and plotted in ROC space

The most commonly used and easy to implement method

1. Linear regression analysis of the relationship
   \[ D = a + bS \]
   where:
   \[ D = (\text{logit TPR}) - (\text{logit FPR}) = \ln \text{DOR} \]
   \[ S = (\text{logit TPR}) + (\text{logit FPR}) = \text{proxy for the threshold} \]

2. \( a \) and \( b \) may be estimated by weighted or unweighted least squares or robust regression, back-transformed and plotted in ROC space.

3. Differences between tests or subgroups may examined by adding covariates to model.

1 Assumes variability in test performance due only to threshold effect and within-study variability
Assumes variability in test performance due only to threshold effect and within-study variability

Does not provide average estimates of sensitivity and specificity
Assumes variability in test performance due only to threshold effect and within-study variability.

Does not provide average estimates of sensitivity and specificity.

Continuity correction may introduce non-negligible downward bias to the estimated SROC curve.
1. Assumes variability in test performance due only to threshold effect and within-study variability

2. Does not provide average estimates of sensitivity and specificity

3. Continuity correction may introduce non-negligible downward bias to the estimated SROC curve

4. Does not account for measurement error in S
Summary ROC Meta-analysis of Diagnostic Test Accuracy

1. Assumes variability in test performance due only to threshold effect and within-study variability
2. Does not provide average estimates of sensitivity and specificity
3. Continuity correction may introduce non-negligible downward bias to the estimated SROC curve
4. Does not account for measurement error in S
5. Ignores potential correlation between D and S
Summary ROC Meta-analysis of Diagnostic Test Accuracy

1. Assumes variability in test performance due only to threshold effect and within-study variability
2. Does not provide average estimates of sensitivity and specificity
3. Continuity correction may introduce non-negligible downward bias to the estimated SROC curve
4. Does not account for measurement error in S
5. Ignores potential correlation between D and S
6. Confidence intervals and p-values are likely to be inaccurate
Publication Bias test for Diagnostic Meta-analysis

1. linear regression of log odds ratio on inverse square root of effective sample size

Bivariate Mixed Effects Models

Arends et al. Med Decis Making. Published online June 30, 2008
Publication Bias test for Diagnostic Meta-analysis

1. linear regression of log odds ratio on inverse square root of effective sample size
2. Uses the effective sample size as weight

Bivariate Mixed Effects Models

Arends et al. Med Decis Making. Published online June 30, 2008
Publication Bias test for Diagnostic Meta-analysis

1. linear regression of log odds ratio on inverse square root of effective sample size
2. Uses the effective sample size as weight
3. Effective sample size = \(4 \times \frac{\text{ndis} \times \text{nndis}}{\text{sample size}}\)

Bivariate Mixed Effects Models

Arends et al. Med Decis Making. Published online June 30, 2008
Recent Developments

Publication Bias test for Diagnostic Meta-analysis

1. Linear regression of log odds ratio on inverse square root of effective sample size
2. Uses the effective sample size as weight
3. Effective sample size = $4*(n_{dis}*n_{ndis})/\text{sample size}$

Bivariate Mixed Effects Models

1. Focused on inferences about sensitivity and specificity but SROC curve(s) can be derived from the model parameters

Arends et al. Med Decis Making. Published online June 30, 2008
Recent Developments

Publication Bias test for Diagnostic Meta-analysis

1. linear regression of log odds ratio on inverse square root of effective sample size
2. Uses the effective sample size as weight
3. Effective sample size = \(4 \times (n_{dis} \times n_{ndis}) / \text{sample size}\)

Bivariate Mixed Effects Models

1. Focused on inferences about sensitivity and specificity but SROC curve(s) can be derived from the model parameters
2. Generalization of the commonly used DerSimonian and Laird random effects model

Arends et al. Med Decis Making. Published online June 30, 2008
Publication Bias test for Diagnostic Meta-analysis

Deeks’ Funnel Plot Asymmetry Test
p-value = 0.89
Bivariate Linear Mixed Model

Level 1: Within-study variability

\[
\begin{pmatrix}
\text{logit}(p_{Ai}) \\
\text{logit}(p_{Bi})
\end{pmatrix}
\sim N\left(\begin{pmatrix}
\mu_{Ai} \\
\mu_{Bi}
\end{pmatrix}, C_i\right)
\]

\[
C_i = \begin{pmatrix}
s^2_{Ai} & 0 \\
0 & s^2_{Bi}
\end{pmatrix}
\]

\(p_{Ai}\) and \(p_{Bi}\) Sensitivity and specificity of the \(i\)th study

\(\mu_{Ai}\) and \(\mu_{Bi}\) Logit-transforms of sensitivity and specificity of the \(i\)th study

\(C_i\) Within-study variance matrix

\(s^2_{Ai}\) and \(s^2_{Bi}\) variances of logit-transforms of sensitivity and specificity

Bivariate Linear Mixed Model

Level 2: Between-study variability

\[
\begin{pmatrix}
\mu_{Ai} \\
\mu_{Bi}
\end{pmatrix} \sim N\left(\begin{pmatrix}
M_A \\
M_B
\end{pmatrix}, \Sigma_{AB}\right)
\]

\[
\Sigma_{AB} = \begin{pmatrix}
\sigma_A^2 & \sigma_{AB} \\
\sigma_{AB} & \sigma_B^2
\end{pmatrix}
\]

\(\mu_{Ai}\) and \(\mu_{Bi}\) Logit-transforms of sensitivity and specificity of the \(i\)th study

\(M_A\) and \(M_B\) Means of the normally distributed logit-transforms

\(\Sigma_{AB}\) Between-study variances and covariance matrix

Bivariate Binomial Mixed Model

Level 1: Within-study variability

\[ y_{Ai} \sim Bin(n_{Ai}, p_{Ai}) \]
\[ y_{Bi} \sim Bin(n_{Bi}, p_{Bi}) \]

\( n_{Ai} \) and \( n_{Bi} \) Number of diseased and non-diseased

\( y_{Ai} \) and \( y_{Bi} \) Number of diseased and non-diseased with true test results

\( p_{Ai} \) and \( p_{Bi} \) Sensitivity and specificity of the \( i \)th study

Bivariate Binomial Mixed Model

Level 2: Between-study variability

\[
\begin{pmatrix}
\mu_{Ai} \\
\mu_{Bi}
\end{pmatrix} \sim N\left(\begin{pmatrix}
M_A \\
M_B
\end{pmatrix}, \Sigma_{AB}\right)
\]

\[
\Sigma_{AB} = \begin{pmatrix}
\sigma^2_A & \sigma_{AB} \\
\sigma_{AB} & \sigma^2_B
\end{pmatrix}
\]

\(\mu_{Ai}\) and \(\mu_{Bi}\) Logit-transforms of sensitivity and specificity of the \(i\)th study

\(M_A\) and \(M_B\) Means of the normally distributed logit-transforms

\(\Sigma_{AB}\) Between-study variances and covariance matrix

1 Exact binomial approach preferred especially for small sample data and for avoiding continuity correction

The relation between logit-transformed sensitivity and specificity is given by

\[ \logit(A_i) = a + b \times \logit(B_i) \]

with slope \( b = \frac{\sigma_{AB}}{\sigma_A^2} \) and intercept \( a = M_A - b \times M_B \).

SROC may be obtained after anti-logit transformation of the regression line.
1 Exact binomial approach preferred especially for small sample data and for avoiding continuity correction

2 The relation between logit-transformed sensitivity and specificity is given by $\mu_{Ai} = a + b \times \mu_{Bi}$ with slope $b = \frac{\sigma_{AB}}{\sigma_{A}^2}$ and intercept $a = M_A - b \times M_B$
1. Exact binomial approach preferred especially for small sample data and for avoiding continuity correction.

2. The relation between logit-transformed sensitivity and specificity is given by $\mu_{Ai} = a + b \times \mu_{Bi}$ with slope $b = \frac{\sigma_{AB}}{\sigma_A^2}$ and intercept $a = M_A - b \times M_B$.

3. SROC may be obtained after anti-logit transformation of the regression line.
Propose a generalized framework for diagnostic meta-analysis based on a modification of the bivariate Dale model:

1. Univariate random-effects logistic models for sensitivity and specificity are associated through a log-linear model of odds ratios with effective sample size as independent variable.
Propose a generalized framework for diagnostic meta-analysis based on a modification of the bivariate Dale model:

1. Univariate random-effects logistic models for sensitivity and specificity are associated through a log-linear model of odds ratios with effective sample size as independent variable.

2. This unifies the estimation of summary test performance and assessment of the presence, extent, and sources of variability.
Methodological Framework

Discuss specification, estimation, diagnostics, and prediction of model:

1. Using a motivating dataset of 43 studies investigating FDG-PET for staging the axilla in patients with newly diagnosed breast cancer
Discuss specification, estimation, diagnostics, and prediction of model:

1. Using a motivating dataset of 43 studies investigating FDG-PET for staging the axilla in patients with newly diagnosed breast cancer

2. Taking advantage of the ability of gllamm to model a mixture of discrete and continuous outcomes
Bivariate Dale Model (Correlated Binary Responses)

1. Joint probabilities decomposed into two marginal distributions for the main effects

2. One log-cross-ratio for the association between two responses

\[ h_1(p_1+(x)) = B_1 x; \]
\[ h_2(p_1(x)) = B_2 x; \]
\[ h\left(\frac{p_{11}(x) \cdot p_{22}(x)}{p_{12}(x) \cdot p_{21}(x)}\right) = B_3 x \]

1. \( h_1, h_2, h_3 \) are link functions in the GLM terminology

2. \( p_1+ \) and \( p_1+ \) are the marginal probabilities for \( \text{response1}=1 \) and \( \text{response2}=1 \) respectively

3. Most popular choice for \( h_1=h_2 \) is the logit function

4. Commonly used link function for \( h_3 \) is the natural logarithm:

\[ \ln(\text{cross-ratio}) = \ln\left\{\frac{p_{11}(x) \cdot p_{22}(x)}{p_{12}(x) \cdot p_{21}(x)}\right\} \]
Within-study variability

\[ y_{Ai} \sim Bin(n_{Ai}, p_{Ai}) \]

\[ y_{Bi} \sim Bin(n_{Bi}, p_{Bi}) \]

\( n_{Ai} \) and \( n_{Bi} \) Number of diseased and non-diseased

\( y_{Ai} \) and \( y_{Bi} \) Number of diseased and non-diseased with true test results

\( p_{Ai} \) and \( p_{Bi} \) Sensitivity and specificity of the \( i \)th study
Between-study variability

\[
\begin{pmatrix}
\mu_{Ai} \\
\mu_{Bi}
\end{pmatrix}
\sim N\left(\begin{pmatrix}
M_A \\
M_B
\end{pmatrix}, \Sigma_{AB}\right)
\]

\[
\Sigma_{AB} = \begin{pmatrix}
\sigma_A^2 & 0 \\
0 & \sigma_B^2
\end{pmatrix}
\]

\(\mu_{Ai}\) and \(\mu_{Bi}\) Logit-transforms of sensitivity and specificity of the \(i\)th study

\(M_A\) and \(M_B\) Means of the normally distributed logit-transforms

\(\Sigma_{AB}\) Between-study variances
Association Model

Associates the univariate random-effects logistic models for sensitivity and specificity in the form a log-linear model:

\[ \log \text{DOR}_i = a + b \times \text{ESS}_i \]

intercept \( a = \) adjusted odds ratio

and slope \( b = \) bias coefficient
PET or Positron Emission Tomography uses radiolabeled glucose analog to evaluate tumor metabolism.
PET or Positron Emission Tomography uses radiolabeled glucose analog to evaluate tumor metabolism.

This radiological test may be used to stage and/or examine the extent of breast cancer.
PET or Positron Emission Tomography uses radiolabeled glucose analog to evaluate tumor metabolism.

This radiological test may be used to stage and/or examine the extent of breast cancer.

The accuracy of axillary PET has been studied by many researchers.
PET or Positron Emission Tomography uses radiolabeled glucose analog to evaluate tumor metabolism.

This radiological test may be used to stage and/or examine the extent of breast cancer.

The accuracy of axillary PET has been studied by many researchers.

We obtained, by searching PUBMED, 43 studies published between 1990 and 2008.
Table: Dataset

<table>
<thead>
<tr>
<th>Idnum</th>
<th>Author</th>
<th>Year</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
<th>SIZE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tse</td>
<td>1992</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>Adler1</td>
<td>1993</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>Hoh</td>
<td>1993</td>
<td>6</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>Crowe</td>
<td>1994</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>Avril</td>
<td>1996</td>
<td>19</td>
<td>1</td>
<td>5</td>
<td>26</td>
<td>51</td>
</tr>
<tr>
<td>6</td>
<td>Bassa</td>
<td>1996</td>
<td>10</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>7</td>
<td>Scheidhauer</td>
<td>1996</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>8</td>
<td>18</td>
</tr>
<tr>
<td>8</td>
<td>Utech</td>
<td>1996</td>
<td>44</td>
<td>20</td>
<td>0</td>
<td>60</td>
<td>124</td>
</tr>
<tr>
<td>9</td>
<td>Adler2</td>
<td>1997</td>
<td>19</td>
<td>11</td>
<td>0</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>10</td>
<td>Palmedo</td>
<td>1997</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>14</td>
<td>20</td>
</tr>
<tr>
<td>11</td>
<td>Noh</td>
<td>1998</td>
<td>12</td>
<td>0</td>
<td>1</td>
<td>11</td>
<td>24</td>
</tr>
<tr>
<td>12</td>
<td>Smith</td>
<td>1998</td>
<td>19</td>
<td>1</td>
<td>2</td>
<td>28</td>
<td>50</td>
</tr>
<tr>
<td>13</td>
<td>Rostom</td>
<td>1999</td>
<td>42</td>
<td>0</td>
<td>6</td>
<td>26</td>
<td>74</td>
</tr>
<tr>
<td>14</td>
<td>Yutani1</td>
<td>1999</td>
<td>8</td>
<td>0</td>
<td>2</td>
<td>16</td>
<td>26</td>
</tr>
<tr>
<td>15</td>
<td>Hubner</td>
<td>2000</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>22</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>32</td>
<td>Wahl</td>
<td>2004</td>
<td>66</td>
<td>40</td>
<td>43</td>
<td>159</td>
<td>308</td>
</tr>
<tr>
<td>33</td>
<td>Zornoza</td>
<td>2004</td>
<td>90</td>
<td>2</td>
<td>17</td>
<td>91</td>
<td>200</td>
</tr>
<tr>
<td>34</td>
<td>Weir</td>
<td>2005</td>
<td>5</td>
<td>3</td>
<td>13</td>
<td>19</td>
<td>40</td>
</tr>
<tr>
<td>35</td>
<td>Gil-Rendo</td>
<td>2006</td>
<td>120</td>
<td>2</td>
<td>22</td>
<td>131</td>
<td>275</td>
</tr>
<tr>
<td>36</td>
<td>Kumar</td>
<td>2006</td>
<td>16</td>
<td>2</td>
<td>20</td>
<td>40</td>
<td>80</td>
</tr>
<tr>
<td>37</td>
<td>Stadnik</td>
<td>2006</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>38</td>
<td>Chung</td>
<td>2006</td>
<td>25</td>
<td>0</td>
<td>17</td>
<td>18</td>
<td>51</td>
</tr>
<tr>
<td>39</td>
<td>Veronesi</td>
<td>2006</td>
<td>38</td>
<td>5</td>
<td>65</td>
<td>128</td>
<td>236</td>
</tr>
<tr>
<td>40</td>
<td>Cermik</td>
<td>2008</td>
<td>40</td>
<td>15</td>
<td>39</td>
<td>125</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>Ueda</td>
<td>2008</td>
<td>34</td>
<td>6</td>
<td>25</td>
<td>118</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>Fuster</td>
<td>2008</td>
<td>14</td>
<td>0</td>
<td>6</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>Heuser</td>
<td>2008</td>
<td>8</td>
<td>0</td>
<td>2</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>
Recode Data for gllamm

```
gen dor = (tp*tn)/(fp*fn)
gen ldor = ln(dor)
gen ldorvar = (1/fn)+(1/tn)+(1/fp)+(1/tp)
gen ldorse = sqrt((1/fn)+(1/tn)+(1/fp)+(1/tp))
tempvar n1 n2 ESS zero thetai sethetai
gen ‘n1’ = tp + fn
gen ‘n2 ’= tn + fp
gen ‘ESS’ = (4 * ‘n1’ * ‘n2’)/('n1’ + ‘n2’)
gen ‘thetai’=(tp * tn)/(fp * fn)
replace ‘thetai’=log(‘thetai’)
gen ‘sethetai’=sqrt(‘ESS’)
gen size = 1/‘sethetai’
```
Recode Data for gllamm

gen ttruth1 = tn /* number truly disease-free */
gen ttruth2 = tp /* number truly diseased */
gen ttruth3 = 'thetai'
gen num1 = tn+fp /* total disease-free */
gen num2 = tp+fn /* total diseased */
gen num3 = 1
reshape long num ttruth, i(study) j(dtruth) string
qui tabulate dtruth, generate(disgrp)
eq disgrp1: disgrp1
eq disgrp2: disgrp2
eq disgrp3: disgrp3
gen gvar = .
replace gvar = 1 if dtruth == "1"
replace gvar = 2 if dtruth == "2"
replace gvar = 3 if dtruth == "3"
forvalues i=1/3 {
    g size_‘i’ = disgrp‘i’* size
}

Bivariate Binomial Mixed Model

gllamm ttruth disgrp1 disgrp2 if dtruth !="3", nocons ///
i(study) nrf(2) eqs(disgrp1 disgrp2) ///
f(bin) l(logit) denom(num) ip(m) adapt

Table: Estimation results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient (Std. Err.)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed Effects</strong></td>
<td></td>
</tr>
<tr>
<td>logitsen</td>
<td>3.084 (0.260)</td>
</tr>
<tr>
<td>logitspe</td>
<td>0.925 (0.197)</td>
</tr>
<tr>
<td><strong>Random-Effects</strong></td>
<td></td>
</tr>
<tr>
<td>logitsen</td>
<td>1.144 (0.232)</td>
</tr>
<tr>
<td>logitspe</td>
<td>1.109 (0.174)</td>
</tr>
<tr>
<td>Correlation</td>
<td>-0.319 (0.256)</td>
</tr>
</tbody>
</table>
Bivariate Binomial Mixed Model

Table: Summary estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>(Std. Err.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sens</td>
<td>0.716</td>
<td>(0.040)</td>
</tr>
<tr>
<td>spec</td>
<td>0.956</td>
<td>(0.011)</td>
</tr>
<tr>
<td>ldor</td>
<td>4.009</td>
<td>(0.305)</td>
</tr>
<tr>
<td>lrp</td>
<td>16.362</td>
<td>(4.047)</td>
</tr>
<tr>
<td>lrn</td>
<td>0.297</td>
<td>(0.042)</td>
</tr>
</tbody>
</table>
SROC Curve

- Observed Data
- Summary Operating Point: SENS = 0.72 [0.63 – 0.79], SPEC = 0.96 [0.93 – 0.97]
- AUC = 0.94 [0.92 – 0.96]
- 95% Confidence Contour
- 95% Prediction Contour
gllamm ttruth disgrp1 disgrp2 disgrp3, nocons nocor ///
i(study) nrf(2) eqs(disgrp1 disgrp2) f(bin bin gauss) ///
l(logit logit id) denom(num) ip(m) adapt fv(gvar) lv(gvar)

**Table:** Estimation results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>(Std. Err.)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed effects</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>logitsen</td>
<td>3.119</td>
<td>(0.265)</td>
</tr>
<tr>
<td>logitspe</td>
<td>0.921</td>
<td>(0.193)</td>
</tr>
<tr>
<td>logdor</td>
<td>3.694</td>
<td>(0.211)</td>
</tr>
<tr>
<td><strong>Random effects</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>logitsen</td>
<td>1.196</td>
<td>(0.246)</td>
</tr>
<tr>
<td>logitspe</td>
<td>1.143</td>
<td>(0.173)</td>
</tr>
</tbody>
</table>
No bias Uncorrelated Random-Effects

Table: Summary estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>(Std. Err.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sens</td>
<td>0.715</td>
<td>(0.039)</td>
</tr>
<tr>
<td>spec</td>
<td>0.958</td>
<td>(0.011)</td>
</tr>
<tr>
<td>ldor</td>
<td>3.694</td>
<td>(0.211)</td>
</tr>
<tr>
<td>lrp</td>
<td>16.888</td>
<td>(4.384)</td>
</tr>
<tr>
<td>ln</td>
<td>0.297</td>
<td>(0.041)</td>
</tr>
</tbody>
</table>
Bias Correlated Random-Effects

gllamm ttruth disgrp1 disgrp2 disgrp3 size_3, nocons ///
i(study) nrf(2) eqs(disgrp1 disgrp2) f(bin bin gauss) ///
l(logit logit id) denom(num) ip(m) adapt fv(gvar) lv(gvar)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>(Std. Err.)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed Effects</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>logitsen</td>
<td>3.084</td>
<td>(0.260)</td>
</tr>
<tr>
<td>logitspe</td>
<td>0.925</td>
<td>(0.197)</td>
</tr>
<tr>
<td>logdor</td>
<td>4.324</td>
<td>(0.543)</td>
</tr>
<tr>
<td>bias</td>
<td>-3.801</td>
<td>(3.032)</td>
</tr>
<tr>
<td><strong>Random-effects</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>logitsens</td>
<td>1.144</td>
<td>(0.232)</td>
</tr>
<tr>
<td>logitspe</td>
<td>1.109</td>
<td>(0.174)</td>
</tr>
<tr>
<td>Correlation</td>
<td>-0.319</td>
<td>(0.256)</td>
</tr>
</tbody>
</table>
### Table: Summary estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>(Std. Err.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sens</td>
<td>0.716</td>
<td>(0.040)</td>
</tr>
<tr>
<td>spec</td>
<td>0.956</td>
<td>(0.011)</td>
</tr>
<tr>
<td>ldo</td>
<td>4.324</td>
<td>(0.543)</td>
</tr>
<tr>
<td>lrp</td>
<td>16.362</td>
<td>(4.046)</td>
</tr>
<tr>
<td>lrn</td>
<td>0.297</td>
<td>(0.042)</td>
</tr>
</tbody>
</table>
Bias Uncorrelated Random-Effects

gllamm ttruth disparp1 disparp2 disparp3 size_3, nocons nocor ///
i(study) nrf(2) eqs(disgrp1 disparp2) f(bin bin gauss) ///
l(logit logit id) denom(num) ip(m) adapt fv(gvar) lv(gvar)

Table: Estimation results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>(Std. Err.)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed effects</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>logitsen</td>
<td>3.119</td>
<td>(0.265)</td>
</tr>
<tr>
<td>logitspe</td>
<td>0.921</td>
<td>(0.193)</td>
</tr>
<tr>
<td>logdor</td>
<td>4.324</td>
<td>(0.543)</td>
</tr>
<tr>
<td>bias</td>
<td>-3.801</td>
<td>(3.032)</td>
</tr>
<tr>
<td><strong>Random effects</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>logitsen</td>
<td>1.196</td>
<td>(0.246)</td>
</tr>
<tr>
<td>logitspe</td>
<td>1.144</td>
<td>(0.173)</td>
</tr>
</tbody>
</table>
Table: Summary estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>(Std. Err.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sens</td>
<td>0.715</td>
<td>(0.039)</td>
</tr>
<tr>
<td>spec</td>
<td>0.958</td>
<td>(0.011)</td>
</tr>
<tr>
<td>ldor</td>
<td>4.324</td>
<td>(0.543)</td>
</tr>
<tr>
<td>lrp</td>
<td>16.888</td>
<td>(4.384)</td>
</tr>
<tr>
<td>ln</td>
<td>0.297</td>
<td>(0.041)</td>
</tr>
</tbody>
</table>
## Comparative Results

### Table: Fit and Complexity Measures

<table>
<thead>
<tr>
<th>Model</th>
<th>nparm</th>
<th>Deviance</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Bias</td>
<td>7</td>
<td>548.42</td>
<td>582.44</td>
</tr>
<tr>
<td>Bias Correlated Random-effects</td>
<td>8</td>
<td>548.42</td>
<td>587.30</td>
</tr>
<tr>
<td>Bias Uncorrelated Random-effects</td>
<td>7</td>
<td>548.37</td>
<td>582.39</td>
</tr>
</tbody>
</table>

### Table: Sensitivity and Specificity

<table>
<thead>
<tr>
<th>Model</th>
<th>Sens</th>
<th>Spec</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Bias</td>
<td>0.716 (0.638 - 0.795)</td>
<td>0.956 (0.935 - 0.978)</td>
</tr>
<tr>
<td>Bias Correlated RE</td>
<td>0.716 (0.638 - 0.795)</td>
<td>0.956 (0.935 - 0.978)</td>
</tr>
<tr>
<td>Bias Uncorrelated RE</td>
<td>0.715 (0.638 - 0.792)</td>
<td>0.958 (0.937 - 0.979)</td>
</tr>
</tbody>
</table>
Prediction and Diagnostics

May use **gllapred** for empirical bayes predictions, residual analysis, influence analysis, normality testing etc

![Model Diagnostic Plots](image)

(a) Goodness–Of–Fit

(b) Bivariate Normality

(c) Influence Analysis

(d) Outlier Detection
Conclusions

1. The preferred model is the Bias Uncorrelated Random-effects Model.
Conclusions

1 The preferred model is the **Bias Uncorrelated Random-effects Model**

2 If interest is in diagnostic performance only, then the **Bivariate binomial mixed** and **modified bivariate Dale** models are equivalent.
Conclusions

1. The preferred model is the **Bias Uncorrelated Random-effects Model**

2. If interest is in diagnostic performance only, then the **Bivariate binomial mixed** and **modified bivariate Dale** models are equivalent.

3. The **modified bivariate Dale** models may be extended further to include study-level covariates to assess impact on summary test performance jointly or separately.