Incorporation of meta-analyses of diagnostic test accuracy studies into a clinical/economic decision analytic framework

Thesis submitted for the degree of

PhD

at the University of Leicester

by

Nicola Novielli

April 2011

Incorporation of meta-analyses of diagnostic test accuracy studies into a clinical/economic decision analytic framework

Nicola Novielli, BSc, MSc

Abstract

An accurate diagnosis is a crucial part of an effective treatment. Diagnostic errors cause unwanted side effects for healthy individuals and witheld treatments for diseased patients. Meta-analysis techniques allow the accuracy of diagnostic tests to be estimated using all the available sources of evidence. The most common measures of diagnostic accuracy are sensitivity (true positive rate) and specificity (true negative rate).

As part of this thesis, current methods developed for synthesising data from diagnostic test studies are reviewed and critiqued, and then applied to estimate the accuracy of the Ddimer test for diagnosing Deep Vein Thrombosis (DVT). The fit of the different models is assessed via the Deviance Information Criterion and the Residual Deviance and the most complex synthesis models are found to provide the best fit to the data. When covariates are added to these models, only the incorporation of *study setting* sensitivity is found to improve the fit of the model.

Diagnostic tests are rarely used in isolation and consideration of multiple tests in combination may also require evaluation. In this thesis, a multiple equations with shared parameters approach is proposed which estimated the accuracy of a combination of tests in two stages: i) estimate the conditional accuracy of the tests; and ii) estimate the accuracy of possible combinations of tests as functions of the conditional accuracies. Such a modeling approach allows the inclusion of different sources of evidence to be used simultaneously. The final part of the thesis evaluated the cost-effectiveness of different strategies for diagnosing DVT by incorporating the results from the aforementioned evidence synthesis models into an economic decision analytic model.

In conclusion, the assumption of conditional independence can affect the analyses of the effectiveness and the cost-effectiveness of combinations of diagnostic tests, thus leading to potentially wrong decisions if the dependence is not explicitly modelled.

Acknowledgments

I am deeply grateful to Dr Nicola J. Cooper, Professor Alexander J. Sutton and Professor Keith R. Abrams for their support and enthusiasm.

I would like to thank Professor Tracey Roberts, and conference delegates, for their interesting and useful discussion of a previous version of Chapter 6 (and relative publication) presented at the Health Economists' Study Group meeting in Sheffield, UK July 2009. Thanks also goes to the anonymous reviewers of the papers published as a result of the work of this PhD. I am also grateful to Professor Steve Goodacre for the useful comments on Chapter 6 and for allowing the use of the data on the accuracy of Ddimer used in chapter 5. As I am to Dr Matthew Stevenson allowing access to the cost-effectiveness analysis model he originally developed for Deep Vein Thrombosis.

Finally, thanks to my colleagues and friends, parents and relatives, and everyone who shared the emotions of this great adventure.

Contents

Contents of the CD-ROM.

The CD-ROM folders (underlined) and files (italics) contained in the CD-ROM are:

- Chapter 2 - GERD 1 study
    - *Accuracy measures -1 study - GERD.odc*
- Chapter 4 - GERD meta-analysis
    - *GERDbugs.odc*
- Chapter 5 - meta-analysis of DD for DVT
    - *asymmetric ROC FI data.txt*
    - *asymmetric ROC FI init.txt*
    - *asymmetric ROC FI model.txt*
    - *asymmetric ROC RI data.txt*
    - *asymmetric ROC RI init.txt*
    - *asymmetric ROC RI model.txt*
    - *bivariate data.txt*
    - *bivariate init.txt*
    - *bivariate model.txt*
    - *Covariates.txt*
    - *hsROC data.txt*
    - *hsROC init.txt*
    - *hsROC model.txt*
    - *independent estimates FE data.txt*

- *independent estimates FE init.txt*
- *independent estimates FE model.txt*
- *independent estimates RE data.txt*
- *independent estimates RE init.txt*
- *independent estimates RE model.txt*
- *symmetric ROC FE data.txt*
- *symmetric ROC FE init.txt*
- *symmetric ROC FE model.txt*
- *symmetric ROC RE data.txt*
- *symmetric ROC RE init.txt*
- *symmetric ROC RE model.txt*

- Chapter 8 - combinations of WS and DD for DVT
  - *model of data A B C D E - conditional accuracy and sequences.odc*
  - *model that assumes independence.odc*
  - *decision model - assume independence.odc*
  - *decision model - conditional accuracy - predictions.odc*
  - *decision model - conditional accuracy.odc*

For a description of the contents of the CD-ROM, see section 1.6.

List of abbreviations

AIC: “Akaike Information Criterion”

AUC: “Area Under the ROC Curve”

BN: “Believe the Negative”

BP: “Believe the Positive”

CEA: “Cost-Effectiveness Analysis”

CEAC: “Cost-Effectiveness Acceptability Curve”

DD: “Ddimer test”

DIC: “Deviance Information Criterion”

DOR: “Diagnostic Odds Ratio”

DVT: “Deep Vein Thrombosis”

GERD: “Gastro Esophageal Reflux Disease”

FNR: “False Negative Rate”

FPR: “False Positive Rate”

HsROC: “Hierarchical summary Receiver Operating Characteristic”

HTA: “Health Technology Assessment”

ICER: “Incremental Cost-Effectiveness Ratio”

IPD: “Individual Patient Data”

LOR: “Likelihood Odds Ratio”

LR(+/-):(positive/negative) “Likelihood ratio”

MCMC: “Markov Chain Monte Carlo”

NE: “North-East”

NICE: “National Institute for Health and Clinical Excellence”

NIHR: “National Institute for Health Research”

NMB: “Net Monetary Benefit”

NPV: “Negative Predictive Value”

NW: “North-West”

PE: “Pulmonary Embolism”

PPI: “Proto Pump Inhibitors”

PPV: “Positive Predictive Value”

QALY: “Quality Adjusted Life Year”

QoL: “Quality of Life”

ROC: “Receiver Operating Characteristic”

SE: “South-East”

sROC: “summary Receiver Operating Characteristic”

SW: “South-West”

TNR: “True Negative Rate”

TPR: “True Positive Rate”

UK: “United Kingdom”

VTE: “Venous Thromboembolism”

WS: “Wells score”

Chapter 1. Introduction

1.1 Background

Early diagnosis can lead to diseases being treated more successfully than if treatment were delayed. Therefore, the evaluation of test performance is crucial to improve patient outcomes and treatments' effectiveness. Test performance is a multifaceted idea, where correct diagnoses are opposed to misdiagnoses. Correct diagnoses occur when diseased patients are positive to the test, and also when healthy patients are negative to the test. On the other hand, misdiagnosis occurs when *i*) a diseased patient is diagnosed as negative by a test (false negative) leading the appropriate care to not be delivered and the disease potentially evolving, or *ii*) a healthy patient is diagnosed positive by a test (false positive) and is exposed to potential side effects as a result of unneeded treatment. Unfortunately, misdiagnoses are unavoidable where the perfect test cannot be applied because it is invasive, not available, or more often does not exist. The effect of misdiagnosis can be observed directly on clinical outcomes. For example, Halfon et al (Halfon, Eggli et al. 2002) show that a proportion (5% in a hospital in Switzerland) of unforeseen (thus avoidable) hospital readmissions after 1 month from hospital discharge is due to "missing or erroneous diagnosis or inappropriate treatment".

Evaluation of the performance of diagnostic tests through systematic review and meta-analysis is less established than for interventions but is increasing rapidly. Presently, reviews are characterised by poor reporting and poor quality (only 56% of reviews for the accuracy of tests in cancer research reported sensitivity, specificity and sample size until 2006 (Mallett, Deeks et al. 2006)). Methods for synthesis of diagnostic test studies are more complicated than for healthcare intervention studies due to additional issues relating to threshold levels and dependence between sensitivity and specificity (until 2006, only 61% of reviews attempted to formally synthesise diagnostic accuracy data in cancer research (Mallett, Deeks et al. 2006)). To date, at least five different synthesis methods have been developed ranging from the simplistic (i.e. assume independence of point estimates of sensitivity and specificity) to the most sophisticated (i.e. express test performance as an asymmetric summary Receiver Operating Characteristic (sROC) curve).

Evaluating the performance of a diagnostic test is only the first step along the pathway to establishing its role in clinical practice. The impact of misdiagnosis in terms of patient outcomes needs to be considered as an important part of the evaluation. Therefore, of more direct relevance is establishing whether a test is beneficial in terms of clinical and economic outcomes. To date, there has been little work conducted integrating the synthesis models with the economic decision models to address policy questions such as, "At which threshold value is the test

most beneficial/cost-effective?" Furthermore, diagnostic tests are rarely used in isolation and consideration of other tests as alternatives or in combination may also require evaluation. Since it is prohibitively expensive and time-consuming to set up individual trials to answer complex questions of policy relevance related to diagnostic tests, often decision models are developed instead.

1.2 Aims of the thesis

The main objectives of this PhD have been to:

i) Review and critique methods currently developed for synthesising data from diagnostic test studies and assess the fit of the different synthesis models when applied to an example dataset (i.e. deep vein thrombosis (DVT) diagnosis and subsequent treatment for (suspected) pulmonary embolism);

ii) Develop methodology for the incorporation of meta-analyses of diagnostic test accuracy studies into an economic decision modeling framework when evaluating the cost-effectiveness of either individual tests or combination of tests

The modeling approaches used throughout this thesis do not aim to tune the threshold of a test to the best test accuracy although some of these approaches will account for variability into the threshold.

1.3 Example I: The diagnosis of Gastroesophageal Reflux Disease

As a first example dataset, the diagnosis of GastroEsophageal Reflux Disease (GERD) will be considered. In the case of uncomplicated GERD, lifestyle modifications and acid suppressive medications are given. Short term treatment is often required given its symptoms (i.e."heartburn"), and fast diagnosis techniques are preferred. The most common tests available are 24-hours pH monitoring, endoscopy, and structured symptom scoring system. The first two are very accurate, however endoscopy is quite invasive and 24-hours pH monitoring may be too expensive. Structured symptom scoring systems may not be very accurate to use as reference standards but are non-invasive and cheap to administer. In common practice, Proton-Pump Inhibitors (PPI) are often used as first therapy for GERD given they are very effective and considered as validation of a first symptomatic diagnosis. However, they are not officially validated as a test, and need diagnostic accuracy evaluations.

This example will be used in Chapter 3 and Chapter 4. In Chapter 3 a single study of the accuracy of GERD will be used to explain the ideas of usefulness of a diagnostic test and to calculate the main measures of diagnostic accuracy. This study is part of an existing meta-analysis; such a meta-analysis dataset is used in Chapter 4 where the meta-analysis techniques presented are applied, assumptions behind each statistical models are discussed, and for comparison between statistical models.

The sensitivity of GERD is estimated at 68% (95% CrI 51% to 82%), and the specificity 57% (95% CrI 39% to 74%) form Bate et al (1999) (repeated in section 3.3.8).

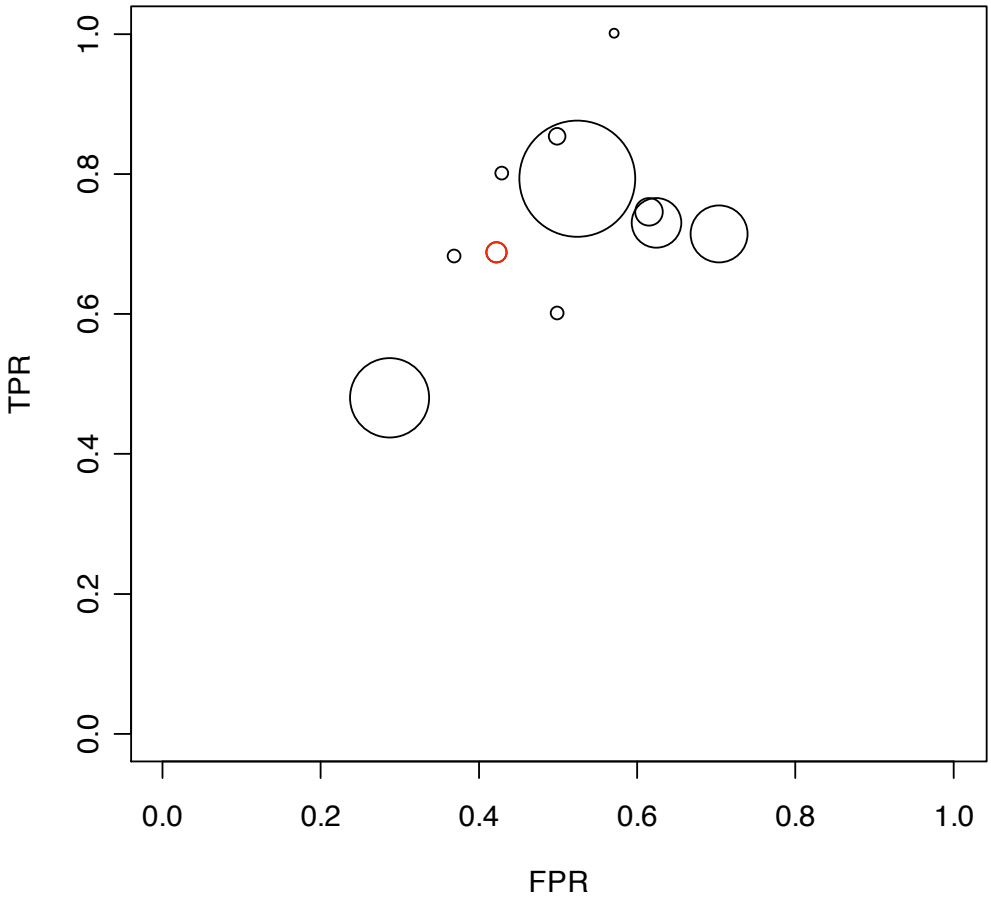


Figure 1-1 Sensitivity and specificity for PPI test from studies identified in a previous meta-analysis. The width of the circles is proportional to the sample size of each study. The red circle corresponds to the study of Bate et al (1999) used in Chapter 3.

The GERD example has been selected because the meta-analysis dataset presented some important differences from DVT example, described in section 1.4: firstly, the GERD dataset for meta-analysis is a smaller dataset than DVT, and secondly there is not much evidence of a diagnostic threshold effect and the implications of this characteristic will be discussed throughout section 4.5 where the different statistical models will be applied.

1.4 Example II: The diagnosis of Deep Vein Thrombosis

The principal example used throughout this thesis is based on the accuracy of tests for the diagnosis of DVT. The medical term Venous Thromboembolism (VTE) is used to identify either the presence of DVT or Pulmonary Embolism (PE) or both. DVT is a blood clot in a deep vein (lower limb) that is usually treated with anticoagulants. It is well known that PE is very likely to originate by a non treated DVT in the lower limbs (i.e. in 90% of cases as reported by Hull (Hull, Raskob et al. 1986)). However, anticoagulants may have serious side effects (i.e. intracranial bleeding). The correct diagnosis of DVT is crucial to lower the mortality due to VTE related adverse events and to lower the impact of side effects from anticoagulant treatment given to healthy patients. A recent Health Technology Assessment (HTA) publication evaluated the clinical effectiveness and cost-effectiveness of diagnostic tests for DVT when used singularly and in combination (Goodacre, Wailoo et al. 2006). They compared the accuracy and the cost effectiveness of a wide range of medical tests for DVT and evaluated the accuracy of 31 combinations of tests (diagnostic algorithms). Some reference tests are Ultrasound or Venography; however, several other tests exist that are less accurate but cheaper, quicker and less invasive, such as Ddimer test (DD) and Wells score (WS). DD measures the concentration of an enzyme in the blood, the higher the measurement the more likely DVT. WS is a checklist of symptoms and clinical history of the patients (Wells, Hirsh et al. 1995; Wells, Anderson et al. 1997). A simplified and widely used version of WS categorises patients into low (score <1), moderate (score 1 or 2) and high (score >2) probability of DVT.

Goodacre et al found that *i)* DD and WS score were not accurate enough as stand-alone diagnostic tools (see Table 1-1) and *ii)* there was evidence that algorithms containing WS and DD performed better. Figure 1-2 (a) and (b) represent the accuracy data for DD (at the threshold used within the original publications) and the accuracy of WS(for the two possible thresholds: low vs moderate-high; low-moderate vs high) respectively.

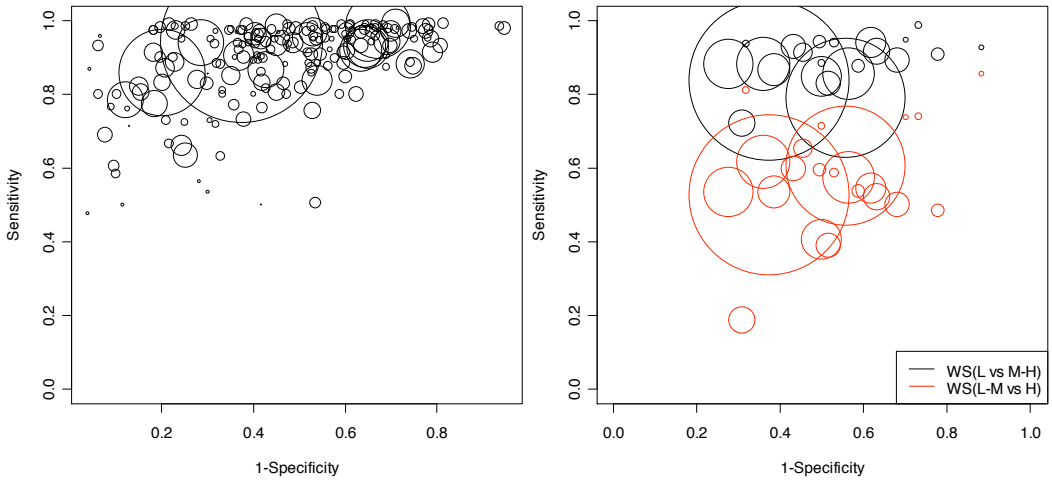
	sensitivity	specificity
DD	90.5% (95% CI 90.0% to 91.1%)	54.7% (95% CI 54.0% to 55.4%)
WS - low vs moderate-high	89.0% (95% CI 86.0% to 92.0%)	48.0% (95% CI 40.0% to 56.0%)
WS - low-moderate vs high	57.0% (95% CI 51.0% to 63.0%)	89.0% (95% CI 85.0% to 92.0%)

Table 1-1 Values of sensitivity and specificity of DD and WS as a result of the analysis performed by Goodacre et al. (2005).

More information on DD dataset (i.e. covariates) will be given in section 5.2, and Chapter 8 will describe the dataset of the accuracy of DD and WS used in combination.

The major limitation of the Goodacre at al. analysis of WS and DD in combination was the assumption of independence of the tests within the combinations, although potential correlation between DD and WS was acknowledged (the accuracy of DD given WS is analysed in Chapter 8).

Further details of the DVT dataset specific for our meta-analysis example will be given in section 5.2.



a) Ddimer b) Wells score

Figure 1-2 Sensitivity and specificity for WS and DD test form studies identified in a previous meta-analysis. The width of the circles is proportional to the sample size of each study.

1.5 Outline of the thesis

This thesis can be divided in two parts; each pursues one of the two main objectives of this PhD as listed above. The first part (Chapters 2, 3, 4, 5) is an introduction to the accuracy of diagnostic tests and an exploration of the methodology used for the synthesis of evidence of the accuracy of dichotomous diagnostic tests. Chapter 2 introduces and justifies the methodological framework used for the statistical analyses: Bayesian modeling. Chapter 3 describes the main characteristics of diagnostic tests and explores the most common measures of performance of diagnostic tests. Chapters 4 and 5 respectively explore and apply the currently proposed techniques used for the meta-analysis of diagnostic test data, to the DVT example. The choice of the correct approach is an important issue, therefore the usefulness of model choice statistics is also explored.

The second part of the thesis (Chapters 6, 7, 8) explores the inclusion of evidence synthesis results of the accuracy of tests used individually or in combination into economic evaluations. Chapter 6 presents a systematic review of Health Technology Assessment (HTA) reports to identify which evidence synthesis methods have been used for diagnostic test accuracy and how the results from these analyses have been used to inform the economic decision model. Chapter 6 also highlights the lack of evidence synthesis methods for the accuracy of combinations of diagnostic tests. Chapter 7 explores the main characteristics of combinations of diagnostic tests and their accuracy. Chapter 8 then describes the modeling framework that is proposed in this thesis for the meta-analysis of the

accuracy of combinations of diagnostic tests. Such a framework is also applied to the DVT example. In addition to the HTA data on the diagnostic accuracy of WS and DD used individually described in section 1.4, conditional data of DD given WS is also identified via a systematic reiew of the literature.

Finally, Chapter 9 concludes the thesis describing the major contributions and the limitations that characterised this research project. Some implications of our findings in the area of diagnostic test accuracy and directions for further work are also presented.

1.6 Overview of the content of the CD-ROM

The WinBUGS code to implement the modeling approaches used throughout this thesis is given in the CD-ROM attached to this thesis. The content of the CD-ROM has been listed at the end of the "Contents" section, under the heading "Contents of the CD-ROM".

The folder Chapter 2 - GERD 1 study contains a Bayesian model for the estimates of the accuracy parameters of PPI test for GERD using data from one single study and presented throughout chapter 2 (the GERD example is presented in section 1.3). The folder Chapter 4 - GERD meta-analysis contains one single WinBUGS file with all the meta-analysis models for the accuracy of PPI test for GERD (illustrative example) and used to produce the results presented throughout Chatper 4. The folder Chapter 5 - meta-analysis of DD for DVT contains all the meta-analysis models for the accuracy of DD test for DVT presented in Chatper 5. Each model is given in three *.txt* files: one for the model, one for the data and one for the initial values. Also a file containing the list of covariates used for the model fitting exercise presented in section 5.3.2 is given. These files can be directly opened in WinBUGS or, alternatively, can be used to run WinBUGS through other softwares (i.e. R). Such files are also available on request for the publication attached in Appendix E. Finally, the folder Chapter 8 - combinations of WS and DD for DVT contains the models for the meta-analysis of the accuracy of DD and WS used in combination, assuming conditional dependence and conditional independence, presented in Chapter 8. Also the code for the decision

models is given in three files, one that contains the meta-analysis model that assumes conditional independence, one that contains the meta-analysis model that assumes conditional dependence and uses credible intervals to represent parameter uncertainty and one that contains the meta-analysis model that assumes conditional dependence and uses predictive intervals to represent parameter uncertainty

Chapter 2. Introduction to the Bayesian approach for statistical modeling

2.1 Chapter overview

The Bayesian approach to statistics is known for the explicit use of external information. For this reason, often objective and subjective statistics have been used as synonyms of Classical (Frequentist) and Bayesian statistics. For example, Blyth (Blyth 1972) comments on the use of Bayesian prior opinions saying "*The publication of Bayesian prior and posterior probabilities would be the antithesis of scientific method*". However, during the last 20 years the Bayesian approach to statistics has gained in popularity, and the objectivity of these methods has been reviewed under the methodological perspective rather than on the mere use of external information in the statistical analysis (Lilford and Braunholtz 1996).

This section aims to give a general introduction to the Bayes theorem and Bayesian modeling, which will form the basis of the analytical approaches to meta-analysis of diagnostic tests used throughout this thesis, and in particular in Chapter 4, Chapter 5 and Chapter 8. This chapter is divided into three main parts: Part 1 (section 2.2) introduces Bayes theorem for simple events and for simple problems of inference; Part 2 (section 2.3) describes the Bayesian approach to statistical modeling for more complex problems of inference, that require the use of Markov Chain Monte Carlo (MCMC) simulation methods; and Part 3 (section 2.4) focuses on Bayesian model selection criteria implemented via MCMC.

2.2 Bayes theorem

In 1763, a new theorem on probability was discovered by a Presbyterian minister named Thomas Bayes. In a letter published *post mortem*, Bayes described the law for reversing a conditional probability by means of marginal probabilities. Bayes had formulated the principles of a new approach to statistical inference. Since then, Bayes theorem has been used to update earlier understanding (beliefs) of a phenomenon by using data from current experiments (see section 2.3). One of the main innovations of the Bayesian approach is the use of external evidence in the form of prior distribution. Prior distributions can be also non informative, also called flat prior distributions. However, it is interesting to observe how results vary under different assumptions on the prior knowledge, representing alternative ideas, in a sensitivity analysis context.

2.2.1 Formula for simple events

Let A and B be two events, and their complements to be $\bar{A}$ and $\bar{B}$, so that the probability $P(A\ or\ \bar{A}) = P(A) + P(\bar{A}) = 1$, similarly for B. Also, let $P(B|A)$ be the conditional probability, for example, of the event B conditional to A; that is the probability of B accurring when A has been observed. Then the conditional probability $P(A|B)$ can be calculated using Bayes theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})}$$

Equation 2-1

This can be interpreted as the probability of an event A, posterior to the knowledge of an event B. $P(A)$ is equal to the (prior) knowledge of the sole event A, the denominator at the right side of Equation 2-1 is $P(B)$ calculated as a function of the conditional probabilities $P(B|A)$, $P(B|\bar{A})$ and the marginal probability $P(A)$ (i.e. $P(\bar{A}) = 1 - P(A)$).

2.2.2 Bayes theorem for inference

The theorem as in Equation 2-1 is also called Bayes theorem for complementary events. Equation 2-2 represents the application of Bayes theorem for inference of continuous or discrete data (Spiegelhalter DJ, Abrams et al. 2004):

$$f\{\theta|Y\} = \frac{L\{Y|\theta\}g\{\theta\}}{\int L\{Y|\theta\}g\{\theta\}d\theta}$$

Equation 2-2

Where *f* is called the posterior density function, *L* is the likelihood and *g* is the prior distribution. The formula above is valid for continuous parameters and either continuous or discrete data. For discrete parameters the integration becomes a summation.

In Equation 2-2, when the prior distribution $g\{\theta\}$ and the posterior distribution $f\{\theta|Y\}$ belong to the same distributional family (Ntzoufras 2010) the analysis is a conjugate analysis. In this case the computation of the formula above is algebraically possible; the integral on the denominator has an algebraic solution; and the posterior distribution also belongs to the same distributional family. A conjugate analysis is possible the functional form of the prior distribution is proportional to the functional form of the likelihood (Spiegelhalter, Abrams et al. 2004). Some common cases of conjugate prior distributions are: Normal-Normal, Gamma-Poisson, Beta-Binomial.

For example, a Normal-Normal conjugate model of a continuous measurement X can be described defining X_n as a sample of dimension n where $X_n \sim N(\mu, \sigma^2/n)$ is the likelihood function associated to this sample. μ is the true mean of X that needs to be estimated, and σ^2 is the known variance. If there is some prior knowledge about the parameter of interest μ (i.e. from a pilot experiment or an expert belief) this can be quantified using the assumption of normality with mean μ_0 and variance σ^2, and the model can be written as:

$$X_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\mu \sim N(\mu_0, \sigma^2)$$

Equation 2-3

The prior distribution is conjugate to the likelihood because it has a similar algebraic form and the posterior distribution of μ given the data can be calculated in closed form:

$$\mu|X \sim N\left(\frac{\mu_0 + \sum_{i=1}^{n} X_i}{1+n}, \frac{\sigma^2}{1+n}\right)$$

Equation 2-4

2.3 More complex modeling with Bayesian statistics

2.3.1 The Markov Chain Monte Carlo and the Gibbs sampler

The general problem of Bayesian inference can be represented by the estimation of the mean parameter using its posterior distribution $f\{\theta|X_n\}$ as in the equation below:

$$E(\theta|X_n) = \int \theta f\{\theta|X_n\} d\theta$$

Equation 2-5

Where θ is the parameter of interest and X_n is the data.

The calculation of the integral in Equation 2-5 is not always algebraically possible in closed form, for example when likelihood and prior distribution are not conjugate. The techniques to estimate the parameters of interest for non-conjugate models are briefly explored in this section.

When the likelihood is not conjugated to the prior distribution, the solution to the integral in Equation 2-5 above can be obtained by numerical techniques but when the dimension of the parameter θ exceeds 4, then standard quadrature-based methods are ruled out (Press 2002). Until a few years ago this was a big limitation. The development of powerful computers and *ad-hoc* software made possible the application of specific numerical algorithms to solve such integrals. Currently, the main and most developed method is the MCMC algorithm (Gilks, Richardson et al. 1996).

The MCMC algorithm transforms the problem of solving the integral in Equation 2-5 by sampling adequately a number of M draws $\theta^{(1)}, \theta^{(2)}, \theta^{(3)}$ ... $\theta^{(m)}$ directly from the posterior density function of θ. By the definition of a Markov process, the conditional density of any $\theta^{(j)}$ depends only on the conditional density of $\theta^{(j-1)}$. This conditional density, also called transition density, can be denoted as $T(\theta^{(j-1)} \mid Y)$. In order to draw the first sample $\theta^{(1)}$, an initial value $\theta^{(0)}$ arbitrarily chosen is needed. The choice of the initial value $\theta^{(0)}$ must not affect the convergence of T(.) to f(.); however, it may have a delaying effect on the convergence. Thus, the first N draws are discarded as a burn-in period, and the remaining M minus N draws are considered as draws from the posterior density(Gilks, Richardson et al. 1996). The length of N and M depends on the MCMC algorithm.

An MCMC chain can be constructed in different ways. The most common is the Metropolis-Hasting algorithm, and a special case of this is the Gibbs sampler. This algorithm simplifies the sampling by considering all the full conditional distributions of $\theta^{(j)} = (\theta_1^{(j)}, \theta_2^{(j)}, \dots, \theta_p^{(j)})$, i.e.

$$T\left\{\theta_1^{(j)} \middle| Y, \theta_2^{(j)}, \dots, \theta_p^{(j)}\right\}$$

$$\dots\dots\dots\dots$$

$$T\left\{\theta_p^{(j)} \middle| Y, \theta_1^{(j)}, \dots, \theta_{p-1}^{(j)}\right\}.$$

Equation 2-6

The steps of the Gibbs algorithm are as follows (Gilks, Richardson et al. 1996):

1. To specify a initial value for θ, i.e. $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)})$
2. For j=1,2 ..,N,…M; generate $\theta_i^{(j)}$ from $T\left\{\theta_i^{(j)} \middle| Y, \theta_{-i}^{(j)}\right\}$
3. Return values of $\theta^{(\forall j: N<j\leq M)}$

Where $\theta_{-i}^{(j)}$ is the vector of parameters $\theta^{(j)}$ excluding the element i, N is the length of the burn-in period and M−N is the number of samples used to build the posterior density.

2.3.2 The software WinBUGS

MCMC algorithm can be implemented using statistical and/or programming software. This requires that code specific to the model has to be developed and complex programming skills are required. The software WinBUGS simplifies this by offering a powerful platform for MCMC simulations (Lunn, Thomas et al. 2000). WinBUGS allows the model to be implemented by the specification of the distributional relationship (stochastic nodes) and of the functional relationship (deterministic nodes) between parameters. When functional relationships between parameters are explicated by equations the uncertainty around the stochastic nodes propagates into the deterministic nodes, thus allowing different sources of evidence to be integrated within the same modeling framework (Ades and Sutton

2006). This property will be used in chapter 8, when a model for the accuracy of combinations of diagnostic tests will be presented.

2.3.3 The choice of prior distributions

The choice of prior distributions is a crucial and multifaceted aspect of Bayesian statistics. In fact, it is related both to the functional form of the prior distributions and to the information that is included in the analysis. A brief overview of prior distributions with respect to the type of information that is represented follows in this section. Finally, the approach to the choice of prior distributions adopted in this thesis and the strategy to sensitivity analyses will be presented.

Prior distribution and prior information

The prior information that is used in the analysis can be either substantial or weak.

In the case of substantial information in favour of some parameter values, those values are assumed to have a higher probability a priori than others to be the true values, therefore the prior information will favour this values when combined with the data by means of the likelihood. Substantial information can derive from existing (e.g. published, such as from randomised clinical trials) evidence (Spiegelhalter, Abrams et al. 2004), can be elicited from domain experts who do not have a (psychological or material) personal stake from the results of the analysis (Khalil 2010) by means of ad-hoc techniques (O'Hagan, Buck et al.

2006), or ad-hoc prior distributions can be constructed to test the impact of *a priori* assumptions (e.g. scepticism or optimism on the effect of a new treatment compared to the standard care) on the posterior distribution (Spiegelhalter, Abrams et al. 2004). In some cases (e.g. evaluation of the effectiveness of interventions at the development stage), prior elicited information is the only information available to inform decision analysis (Cosh, Girling et al. 2007; McAteer, Cosh et al. 2007).

Weak information does not greatly favour any parameter values and it is expressed by means of reference prior distributions. For these prior distributions, the meaning of very similar terminologies is still subject of debate, for example between "vague", "non-informative" or "weakly informative" distributions (Kass and Wasserman 1996). The main reason why this debate has been active for a long time is that reference prior distributions do contain some information, and therefore may have an (unexpected/undesired) impact on posterior distributions. A simulation study has assessed that the use of vague prior distributions is not enough to assure the non-influence of the prior distributions themselves on the final posterior parameter estimates (Lambert, Sutton et al. 2005) particularily when there is little data in the context of hyper-parameters in hierarchical models. Although location parameters were always correctly estimated, the precision parameters were often badly estimated, especially when a few data points were available, consequently leading to bad predictions. More evidence on the unexpected impact of vague distributions on posterior distributions is also

available elsewhere (Gelman 2006). In section 3.3.8 an example of the impact of vague distribution on posterior distributions will be given.

What is the best approach between using informative and reference prior distributions? Including the existing evidence in the analysis of the data is one of the main advantage of Bayesian statistics and reference prior distributions are useful benchmarks to assess the impact of informative prior distributions (Spiegelhalter, Abrams et al. 2004). Technical problems such as how the information (either vague or substantial) is implemented analytically should not represent barriers but stimuli to improve the existing techniques.

For simplicity vague and non-informative prior distributions (that is, distributions that do not contribute with substantial information to the analysis) will be considered having equal meaning (Gelman, Carlin et al. 2003). Sensitivity analises are recommended in either case.

Throughout this thesis non-informative prior distributions will be used. However, nothing would limit the use of informative prior distributions in the models that will be presented in the following chapters. The next section presents the strategy adopted in this thesis to check for sensitivity of the results to prior distributions.

Sensitivity analysis to prior distributions

As already mentioned above, vague prior distributions may unexpectedly have an impact on the posterior distribution. Therefore, sensitivity analyses should always be performed to assess the robustness of posterior distributions to the coice of

different prior distributions (Ntzoufras 2010). This can be done under two different perspectives: to assess the sensitivity to the information contained in the prior distribution (in this case all distributions are meant to be non informative, therefore sceptical and optimistic priors will not be included in the analyses)(Spiegelhalter, Abrams et al. 2004); to assess the sensitivity to the distributional form of the prior distribution (i.e. the same information can be represented with different distributions, either on the same parameter or on transformations of the parameter).

Generally, the prior distribution of the means (i.e. logit-sensitivity) was normal with a very low precision, and sensitivity against different prior means and standard deviations was checked. This type of prior is also called weakly informative (Gelman 2009) because it gives slightly more relevance to plausible values of the parameters (i.e. it is locally vague on the interval of plausible values).

Different prior distributions for heterogeneity parameters were used with respect to the shape of the distribution and to different parameterizations of the parameters (i.e. prior distributions on either precision or variance).

The models that will be presented in this thesis have been tested against sensitivity on prior distributions and major agreement problems were not observed. Instead, models with more complex parameterizations presented issues with the choice of initial values (i.e. to get the model run), in these cases plausible

set of initial values have been suggested. This approach has already been used for meta-analysis models of diagnostic test data (Rutter and Gatsonis 2001).

2.3.4 Assessment of convergence and length of chains

Convergence is a crucial property of MCMC algorithms for Bayesian modeling. If an MCMC algorithm does not converge, then the algorithm is not sampling from the posterior distributions of the parameters which has obvious and serious implications. A"burn-in" period of the MCMC algorithm is required to ensure convergence of the sampler, but a key challenge is that the length of "burn-in" required is model specific and no methods exist for determining the length a priori. Due to this, a number of convergence diagnostics have been developed and a brief overview is provided below. Convergence diagnostics of two types exist, those based on: 1) statistical tests; and 2) more advanced procedures based on checking the characteristics of the sampled chain (i.e. history plots, autocorrelation plots, MCMC error).

The most common tests for convergence are:

1. The Geweke test (Geweke 1992) is based on the comparison of the averages estimated from two different sub-samples of the MCMC chain for an individual parameters. Simply, if the algorithm converges (and the burn-in period is long enough) then the two averages should not be statistically different from each other.

2. The Gelman and Rubin test (Gelman and Rubin 1992) compares the running means of two (or more) parallel samples (i.e. different chains characterised by different initial values).
3. The Raftery and Lewis test (Raftery and Lewis 1992; Raftery and Lewis 1995) focuses on the accuracy of specific quantiles of the sampled chain. Differently from the first to approaches, this is not based on the sampled means.
4. The Heidelberger and Welch test (Heidelberger and Welch 1983) is used on univariate observations and requires one chain to be run. This test controls that stationarity of the MCMC chain (i.e. the chain is sampling from one posterior distribution) is achieved. If stationarity of the chain is not achieved, then the burn-in period is increased with the first 10% of the candidate posterior iterations. The hypothesis of stationarity, and therefore convergence, is rejected when more than 50% of candidate posterior iterations are rejectet.

The main shortcoming of these approaches is that they have to be repeated for all relevant parameters in case of multidimensional parameter space. Only the Gelman and Rubin statistic has been adapted for multidimensional parameter space (Brooks and Gelman 1997). Moreover, a personal opinion is that these statistics may be very useful as complementary tools in prespecified strategies but they try to give a simple answer to very complex problems when used individually. The risk is that the complexity of the problem is not appreciated. In fact, some authors suggest that these statistics all represent different aspects of the

problem of convergence and should be used simultaneously (Ntzoufras 2010). Also, they suggest that more advanced users should assess convergence by observing directly the characteristics of the MCMC sample. However, a standard procedure based on the characteristics of the chain does not exist. Below I describe the procedure used in this thesis to assess convergence step by step:

1 Convergence and the length of the burn-in period was initially assessed by initalising the chain in different points of the space of parameters (i.e. different sets of plausible values of the parameters) (Racine-Poon and Wakefield 1996). History plots for multiple chains where used to check when the chains were converging by setting the burn-in period to zero (for every model presented in this thesis the algorithm started samping within the same range of values below 1000 iterations). Therefore, the burn-in period (N) was set between 4000 and 5000 iterations.

2 The length of the chain after the burn in period (M-N) that is used to build the posterior distribution was determined when the following two criteria where both met: 1) if *s* is a relatively small number of further iterations that can be run after the first M terations, then the posterior distributions (i.e. posterior means and standard deviations) considering did not change when considering (M+*s)*-N iterations; 2) the MC error was lower than 10E-4, where the MC error is a measure of variability in the estimates due to the sampling algorithm; the higher the number of iterations the lower the MC error (Ntzoufras 2010).

3 The length of MCMC chains also depends on the assumption of independence of an iteration at time t_m from the iterations at time t_j, for $j < m - 1$. Therefore, autocorrelation was checked via the autocorrelation plots available in WinBUGS. Where autocorrelation was observed the length of the chain was increased proportionally to the extent of autocorrelation (i.e. if the maximum lag was l on the autocorrelation plot, then the length of the chain was set to $M * l$).

2.3.5 The representation of uncertainty in Bayesian modeling compared to Classical methods

Parameter estimates are supposed to be robust estimates of the parameters true values. The quantification of uncertainty around these estimates depends on the data available, and on the type of prior distribution used. Measures of uncertainty depend on the precision of the estimates (i.e. the more the observations the more precise the estimate).

In classical statistics, parameter estimates are followed by confidence intervals. Their interpretation is quite complex and needs the understanding of the terms confidence and significance. In statistical testing, significance is used to indicate the probability to reject a true null hypothesis (false positive result). Therefore, confidence is the probability to accept a true null hypothesis (true negative result). For the construction of confidence intervals, if an indefinite number of samples could be drawn each corresponding to a confidence interval, in 95% of such samples the interval would cover the true value of the parameter(Snedecor and Cochran 1967). Different approaches can be used to construct confidence intervals (e.g. deviance based, likelihood based, profile likelihood based, quasi-likelihood based, score finction based (Hinkley, Reid et al. 1991)). In Bayesian statistics, the representation of uncertainty via Credible intervals simplifies their interpretation. Credible intervals are true probability statement on the variability of the parameter; that is the true parameter has 95% probability to lie within its credible interval (given the prior information). The interpretation seems to be simpler than in Classical analysis, and it is simpler and more interpretable by non

statistician. However, the description of prior distribution and their influence on the posterior distribution is important for a correct understanding of those intervals.

2.4 Model selection

Some model selection techniques are shared and others are exclusively used in either the Classical or the Bayesian approach to statistics. For example, section 2.4.1 describes a model selection statistic that is only valid when MCMC-Bayesian models are used, and section 2.4.2 describes a model selection statistic that can be calculated for either approach. The application of a comparison between these statistics will be presented in Chapter 5, where an example of the choice between different models and the inclusion of covariates will be presented.

2.4.1 Deviance information criterion as model choice criterion

The DIC (Spiegelhalter, Best et al. 2002) is a compound measure of goodness of fit and complexity of the model, and can be used as a basis on which to choose between competing models. It is defined as

$$DIC = \overline{D} + p_D$$
$$p_D = \overline{D} - \hat{D}$$

Equation 2-7

Where $\overline{D}$ is the posterior mean of the deviance taken as a Bayesian measure of fit(Spiegelhalter, Best et al. 2002), penalised by a measure of complexity p_D (the 'effective number of parameters'), where $\hat{D}$ is the deviance evaluated at the posterior mean of the model parameters.

The posterior mean of the deviance can be obtained in WinBUGS by monitoring the node "deviance", or alternatively the posterior distribution of the deviance can be computed manually by sampling within the model the sum of the individual contribution to the deviance (i.e. $\overline{D_j} = \sum \overline{D_{ij}}$, for observation i, j=1 to M indicates the MCMC iteration, iterations between N (burn-in period) and M will be used to build the posterior distribution of the eviance, see section 2.3.1).

$\overline{D_{ij}} = -2 * log\left(\binom{N_i}{y_i} {p_{ij}}^{y_i} * (1 - p_{ij})^{N_i - y_i}\right)$ for binomial likelihood, where N_i may be the number of diseased (healthy) patients and y_i may be the number of true positive (negative) patients, and therefore p_{ij} indicates the j^{th} MCMC sample of sensitivity (specificity) of the diagnostic test (true positives and true negatives are counts data, sensitivity and specificity are proportions, a detailed description of these quantities will be presented in chapter 3). p_{ij} is the estimated rate at iteration j of the MCMC algorithm.

The DIC is a natural generalisation of the Akaike Information Criterion (AIC) (Akaike 1973). The DIC was developed to solve the problem of determining the 'effective' number of parameters (p_D) in complex non-nested hierarchical models and is implemented in WinBUGS (i.e. penalisation for random effects is not available via the AIC which motivates, in part, the Bayesian approach taken throughout this paper). The lower the value of DIC the better the model fits the data. It has been recommended that differences of 5 or more in the DIC between competing models may be considered as substantial (The BUGS project). Note

that the DIC is only comparable across models with exactly the same observed data. Moreover, the version of DIC presented in this section is not ideal in the case of missing data (Celeux, Forbes et al. 2006).

2.4.2 Residual deviance

An alternative statistic for assessing model fit is inspired on the theory for generalised linear model and is based on the likelihood ratio statistic. Using a slightly different notation, the deviance can be represented as $-2l(y_i; \bar{p}_i)$ (for example where $\bar{p}_i$ is a likelihood based estimate for the parameter p_i and $l(.;.)$ represent the log-likelihood function). The posterior distribution of the deviance was defined in section 2.4.1 for binomial data by sampling via the MCMC algorithm $-2l(y_i; p_{ij})$, where $l(.;.)$ indicates the (natural) logarithm of the likelihood function (log-likelihood) and p_{ij} is the j^{th} sample of the MCMC algorithm for the i^{th} observation (samples between N and M-N are used to build the posterior distribution of the deviance).

The residual deviance statistic can be defined as the difference between the deviance for the model currently being fitted and the deviance for the saturated model (i.e. a model that perfectly fits the data because it has as many parameters as there are values to be fitted) where the deviance measures the fit of the model to the data points using the likelihood function (Hinkley, Reid et al. 1991). The term "residual deviance" is due to an analogy with normal-theory models where the likelihood ratio statistic for comparing two models reduces to the difference

between the the respective residual sums of squares after fitting the models (Hinkley, Reid et al. 1991).

The formula for residual deviance is:

$$residual\ deviance \quad = \sum_i \left[2 \left(y_i \log\left(\frac{y_i}{N_i * \bar{\mathrm{p}}_\mathrm{i}}\right) + (N_i - y_i) * \log\left(\frac{(N_i - y_i)}{(N_i - N_i * \bar{\mathrm{p}}_\mathrm{i})}\right)\right)\right]$$

Equation 2-8

As for section 2.4.1, if N_i is the number diseased patients and y_i is the number of true positive patients, then $\bar{p}_i$ indicates the posterior mean of sensitivity; similarly for specificity. The residual deviance is inversely proportional to the likelihood, the highest value of the likelihood function corresponds to the lowest value of the residual deviance, therefore the lower the residual deviance the better the model fits the data.

If the deviance of the satureated model is defined as $-2l(y_i; y_i) = -2 * log\left(\binom{N_i}{y_i}\left(\frac{y_i}{N_i}\right)^{y_i} * \left(1 - \frac{y_i}{N_i}\right)^{N_i - y_i}\right)$, the residual deviance (Equation 2-8) can be obtained as the difference between $-2l(y_i; y_i)$ and $-2l(y_i; \bar{p}_i)$. In this thesis the posterior distribution of the residual deviance will be calculated by using the j^{th} MCMC sample for the parameter p_i instead of $\bar{p}_i$ (i.e. $-2l(y_i; p_{ij})$ instead of $-2l(y_i; \bar{p}_i)$) and the posterior mean of the residual deviance will be used as model choice statistics.

Under the null hypothesis that the model provides an adequate fit to the data, it is expected that residual eviance would have a mean approximately equal to the number of unconstrained data points (this is exact if the data have a Normal likelihood (Dempster 1997)). For the models presented in this thesis, the residual deviance can be calculated by summing the residual deviance for sensitivity and the residual deviance for specificity. However, more research is needed into the properties of the posterior mean of the residual deviance, especially with respect to the shape of the posterior distribution of the parameter (e.g. asymmetric posterior distribution), and to the implementation of missing data models.

2.4.3 Some considerations on the relationship between DIC and residual deviance

Some considerations on the relationship between the residual deviance and the DIC can be helpful to interpret both statistics.

The DIC has been defined as the sum of the posterior estimate of the deviance $-2l(y_i; p_{ij})$ and the effective number of parameters p_D. Residual deviance has been defined as the difference between the posterior estimate of the deviance $-2l(y_i; p_{ij})$ and the deviance of the saturated model $-2l(y_i; y_i)$:

$$DIC = E_j\left(-2l(y_i; p_{ij})\right) + p_D$$

$$Residual\ deviance = E_j\left(-2l(y_i; p_{ij})\right) + 2l(y_i; y_i)$$

Where $E_j\left(-2l(y_i; p_{ij})\right)$ indicates the expected value of the statistic calculated from the respective posterior distribution. It is evident that the DIC will be higher than the posterior estimate of the deviance (p_D should assume positive values otherwise the model fit can be considered as poor (Spiegelhalter, Best et al. 2002)). Similarly, the residual deviance will be lower than the posterior estimate of the deviance ($2l(y_i; y_i)$ assumes negative values). Moreover, $2l(y_i; y_i)$ depends only on the data (i.e. $2l(y_i; y_i)$ will be constant for all the models and for every MCMC iteration). Therefore, the lower the value of p_D compared to the posterior mean of the deviance, the more similar will be the comparison between models based on the two statistics. This will be the case of the model choice exercise presented in chapter 5.

The reason why the DIC and the residual deviance were compared to each other relies in their interpretation. As for the frequentist AIC, also the DIC provides a relative measure of fit (Spiegelhalter, Best et al. 2002). Therefore it is useful to compare models but it does not provide the fit of the model in absolute terms. Alternatively, the residual deviance provides an absolute measure of fit. In chapter 5 these two measures will be applied to an existing dataset for the selection between models for the meta-analysis of diagnostic test data.

2.5 Summary

Bayesian models will be described in detail for the meta-analysis of diagnostic tests data in Chapter 4 and for the meta-analysis of combinations of diagnostic tests in Chapter 8. Applications and examples will be given in Chapter 4, Chapter 5 and Chapter 8, where the results of a series of Bayesian models with shared equations for the meta-analysis of combinations of tests will be given. The use of Bayesian techniques will also allow economic evaluations that incorporate parameter uncertainty (see Chapter 4 for description of comprehensive decision modeling and Chapter 8 for an application to combinations of tests). The Bayesian approach to statistics will be at the basis of the statistical modeling throughout this thesis.

For this reason, this chapter has been the first methodological chapter of the thesis, where Bayes theorem for inference, the most common techniques for parameter estimation and model selection, and some pros (interpretation of uncertainty, complexity of models) and cons (convergence, sensitivity of the model) of such an approach have been discussed.

This chapter constitutes an important part of the technical background for the statistic techniques used in this thesis. Before the description of methodologies and examples specific to the accuracy of diagnostic tests, which would constitute the main body of the work behind this thesis, a description of the theory behind diagnostic tests needs to be explored as another important part of the technical background.

Chapter 3. Introduction to the accuracy of diagnostic tests

3.1 Chapter overview

This chapter aims to provide an overview of the relevant theory behind diagnostic tests and their accuracy, and is split into 3 parts. The first part contains a general description of the motivations behind diagnosis and explores the characteristics and roles of diagnostic tests (section 3.2). It includes a description of *i*) the types of tests according to their role in the diagnostic pathway (3.2.2) and the different types of data on accuracy that may be collected (3.2.3); and *ii*) the usefulness of a test (3.2.4). Finally, these ideas will be discussed and applied to an example dataset selected from the literature (3.2.5), which will also be used throughout Chapter 4**.**

The second part of this chapter describes and discusses the most common measures for the accuracy of medical tests and the relations between these measures (section 3.3).

3.2 Diagnosis and accuracy of medical tests

3.2.1 Motivations of diagnostic tests

Diagnostic tests have clinical utility as they are used to detect diseases. However, different types of test can be used to explore the same set of symptoms which may be at the basis of a set of different diseases. This is often simplified saying that (a number of) diagnostic tests can be used to detect the presence or the absence of a disease.

Measuring how well a diagnostic test performs has statistical and clinical relevance and diagnostic tests often need to be compared to each other. Sometimes, diagnostic tests can be used in conjunction with other tests and so these evaluations are even more complex. Zweig et al (Zweig and Campbell 1993) classified the motivations for the evaluation of diagnostic test accuracy into four categories:

1. TEST VALUES: The case when one simply wants to know the test values in relation to the presence of the disease.

2. REPLACEMENT: This case is to compare two tests to consider whether it is worth replacing the currently used test with a new one.

3. COMPLETING: Adding a new test can make the clinical diagnosis of a disease (or a set of diseases) more accurate. When the symptoms relate to different possible diseases, the diagnostician may aim to exclude a disease rather than detect one.

4. ELIMINATION: Sometimes a test can perform so badly that it can be eliminated. Obviously, an evaluation of its accuracy is needed even without any comparison.

3.2.2 Types of diagnostic tests

The accuracy of a diagnostic test is usually obtained by comparison with a reference standard, regardless of the motivation of the evaluation as mentioned in section 3.2.1. A reference standard is a test that does (or it is supposed to) discriminate perfectly between the populations of the diseased and the healthy individuals. This standard, despite being perfect, is sometimes not used in current practice for different reasons (invasiveness, very expensive, availability, etc).

In order to evaluate the accuracy of a new test it is very important to define its role. Beyond the explorative diagnostic test (i.e. one may want to classify individuals according to the value of a single test), three roles have been identified by Bossuyt (Bossuyt, Irwig et al. 2006). Identifying the role of a new test may help either to *i*) design a new study, *ii*) interpret the results of an existing one or *iii*) interpret meta-analytic results. A further test could be used also to screen for the false positives or the false negatives, which are the main drivers for the costs resulting from the treatment (Altman and Bland 1994).

The following types of diagnostic tests can be identified according to their role in the diagnostic pathway: first, a new test can replace the existing one and it is called *Replacement test*; second, a new test can follow the existing test being

added-on to the existing pathway and it is called *Add-on test*; third, a new test can be positioned at the beginning of the pathway, and a certain result (positive or negative) can be the condition to continue with the existing test, this is called *Triage test*. The awareness of the possible roles of a test can help to reduce the waste of resources resulting from wrong decisions (i.e. money, time, human lives, etc). Thus, when a new test has to be evaluated, it is important to define whether it will replace the new test or not, and if not, how it will be used jointly in the diagnostic pathway. A complete exploration of the types of combinations of diagnostic test and their properties will be given in Chapter 7 and Chapter 8.

3.2.3 Type of data

A diagnostic test involves the measurements of a characteristic of the patient which is believed to be associated with the presence of the disease. These measurements can produce different types of data:

1. Dichotomous data - positive or negative test results: For example in the case of a qualitative test which involves the observation of the change in the colour of a patch after the blood is mixed with an enzyme (usually called biomarkers), or the presence/absence of a symptom;
2. Multiple discrete data: For example clinical scores can classify patients into more than two categories (i.e. low risk, moderate risk or high risk of having the disease), or imaging tests may lead to unclear results, thus the doctor who reads an x-ray may spot the condition (positive test), may

exclude the presence of the condition (negative test), or may consider the output of the x-ray unclear for a definitive classification;

3. Continuous test data: For example, the level of a certain biomarker in the blood can be measured after a laboratory analysis (Glas, Lijmer et al. 2003).

While for the first class of results the classification of positive against negative results is obvious, for the multiple (ordinal) and continuous test results it may be necessary to choose a threshold (also called cut-off value) that makes a test value either positive or negative before taking the definitive decision of treating/discharging or further testing. Then, given the threshold, all the test results can be viewed as dichotomous. Such a threshold causes variability that will affect the accuracy of diagnostic test, and that will need to be accounted for in the statistical analyses. The statistical approaches presented in Chapter 4 and Chapter 5 will refer to dichotomous or dichotomised test results, and some of these account for variability in threshold.

3.2.4 Useful tests

A number of different terms have been used to express the clinical performance of a diagnostic test; for example, efficiency, accuracy, utility, value, worth, effectiveness, usefulness, efficacy and diagnostic accuracy. All these terms may mean different things to different people or they may sometimes be used as

synonyms (Zweig and Campbell 1993). Throughout this thesis, diagnostic accuracy will be used to identify the ability of the test to discriminate between diseased and healthy individuals. Efficiency and usefulness will be used to identify the value of the test under an economic/decision perspective. Zweig and Campbell (1993) identified two criteria that aim to identify useful tests: the quality of information and the practical/clinical value of information. According to the *quality of information*, the accuracy of a generic test is its ability to distinguish between two different health states; the more accurate a test the better the quality of the information provided by the test. However, a good accuracy does not always correspond to a useful test. According to the *practical clinical-value of the information* criteria, possible reasons for non-useful tests may be *i*) that they elevate costs (i.e. economic usefulness), *ii*) the intolerability to false results (i.e. clinical usefulness), and/or *iii*) limited availability of the test (scarce technical resources). Moreover, it may be so invasive or uncomfortable that it is unacceptable to patients. Thus, usefulness relates not only to the test itself, but it relates to environmental characteristics.

Evaluating the performance of a diagnostic test is only the first step along the pathway to establishing its role in clinical practice. Of more direct relevance is establishing whether a test is beneficial/useful in terms of clinical and economic outcomes. To evaluate the whole diagnostic-to-treatment pathway (including treatment strategies and possible consequences which may derive from a certain classification; for example, false positives may be given expensive treatments and unwanted side effects may occur), decision analytic models can be developed.

Such methods are briefly described in Chapter 4 and applied to assess the cost-effectiveness of different strategies of diagnosing DVT in Chapter 8.

3.2.5 GERD example – description of the study

The general framework of the GERD example has been presented in section 1.3. This section presents a single study of the accuracy of PPI for GERD that is used throughout this chapter to represent the measures of accuracy described in section 3.3.

Bate, Riley *et al* (1999) conducted a study in the United Kingdom on 58 patients (55.1% men whose age was on average 47.4 years) to evaluate the clinical and economic effectiveness of PPI's. 24h pH monitoring (pH <4 during at least 4% of monitoring time) was used as the gold standard. PPI's therapy was considered successful (PPI test positive) when complete relief or at least 50% reduction in symptoms was achieved. This is a rare case where the diagnostic test is also a therapy and its clinical usefulness is given by the fact that it is also a therapy. However, given the possible side effects of PPI, if people without the disease are wrongly diagnosed as having GERD (false positives), the usefulness of PPI as a test may not be fully expressed.

This dataset has been chosen here because it is small and simple to represent. A larger dataset (i.e. diagnosing DVT outlined in section 1.4), will be used in subsequent Chapter 5.

3.3 Measures of Diagnostic Accuracy

3.3.1 The accuracy of diagnostic tests is not a univariate measure

Diagnostic accuracy has been defined as the ability of the test to classify patients into one of two health states. Unfortunately, the possible alternative results of a test are not simply diseased or healthy (in which case the test would be perfectly accurate) but positive or negative with some uncertainty on the true disease status which is not known. Table 3-1 shows how patients may be classified in one of the following four categories given the disease status is known, which is a condition to measure the accuracy of a diagnostic test:

1. positive and diseased (True Positive, *TP*)
2. positive and healthy (False Positive, *FP*)
3. negative and healthy (True Negative, *TN*)
4. negative and diseased (False Negative, *FN*)

where the total number of patients with the condition (D^+) is equal to $TP+FN$, and the number of patients without the condition (D^-) is equal to $TN+FP$.

Figure 3-1 shows the assumptions behind the problem of diagnostic test accuracy. The populations of diseased and healthy are assumed to be normally distributed (Figure 1-1 a) over the test value. As the threshold varies, the probability of being true or false positive or negative changes. The conventional assumption of normality can be replaced by other distributional assumptions (Figure 1-1 b).

		Reference test / disease status	
		Diseased	Healthy
Experimental test	Positive	TP	FP
	Negative	FN	TN
		D^+	D^-

Table 3-1 Dichotomous classification of the results of a diagnostic test.

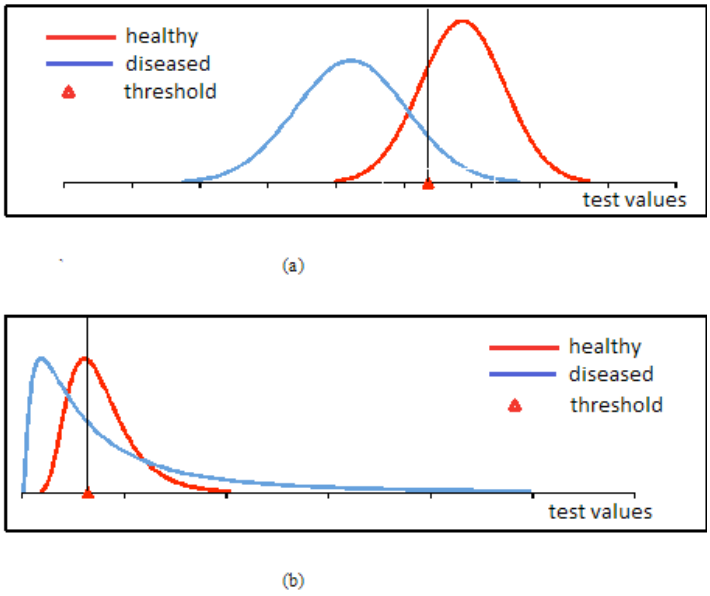


Figure 3-1 Distribution of healthy and diseased patients over the test results under the assumption of (a) normality and (b) under an alternative distributional assumption, i.e. log-normal.

3.3.2 Sensitivity and Specificity

By definition, s*ensitivity* is the proportion of individuals classified as positive by the test, among those who are diseased. Conversely, *specificity* is the proportion of individuals classified as negative by the test, among those who are not diseased.

$$sensitivity = P(T+|D+) = \frac{TP}{TP+FN}$$

$$specificity = P(T-|D-) = \frac{TN}{TN+FP}$$

Equation 3-1

Specificity is also called the *true negative rate*, and its complement (1-Specificity) is called the *false positive rate*. This estimates the proportion of people without the condition who have a positive test result. Similarly, the *false negative rate* is equal to (1-Sensitivity), where Sensitivity is called the *true positive rate*.

A measure of uncertainty and confidence intervals can be calculated using, for example, asymptotic methods (i.e. by assuming normality of the sampling distribution(Newcombe 1998)). Newcombe (1998) presents a comparison of methods for building confidence intervals for single proportions. Alternatively, the posterior distribution of sensitivity and specificity along with standard deviation and credible intervals can be calculated using a beta-binomial conjugate Bayesian model (Ntzoufras 2010).

Sensitivity and specificity can be represented as probabilities conditional to the disease status, for example the probability that the test is positive (negative) given the condition is present (absent) (Altman and Bland 1994) (see Equation 3-1). These measures are usually assumed to be independent to the prevalence of disease, although this is the subject of several controversies (Brenner and Gefeller 1997). Here it is presented a personal interpretation of this variability that needs to be explored further. Both sensitivity and specificity and prevalence of disease may vary with the severity of disease; for example fewer people are affected by severe occurrence of the disease (i.e. lower prevalence for more severe occurrence of disease) and the test may be positively correlated with the severity of disease (i.e. the test is more sensitive when the disease is more severe). Therefore, sensitivity and specificity appear to vary with prevalence.

3.3.3 Positive and Negative predictive values

Analogously to sensitivity and specificity, Positive Predictive Value (PPV) is the proportion of patients with positive test results who are correctly diagnosed, and Negative Predictive Value (NPV) is the proportion of patients with negative results who are correctly diagnosed (Altman and Bland 1994).

$$PPV = p(D+|T+) = \frac{P(D+,T+)}{P(T+)} = \frac{P(T+|D+) * P(D+)}{P(T+|D+) * P(D+) + P(T+|D-) * P(D-)}$$

$$NPV = p(D- \mid T-) = \frac{P(D-, T-)}{P(T-)} = \frac{P(T- \mid D-) * P(D-)}{P(T- \mid D-) * P(D-) + P(T- \mid D+) * P(D+)}$$

Equation 3-2

Equation 3-2 represents the predictive values in terms of probabilities and gives a better understanding of these numbers in terms of probabilities. It can be seen how the predictive values can be expressed as functions of sensitivity, specificity, and the prevalence of disease. Therefore, it is clear that the prevalence of disease directly affects the predictive values. Figure 3-2 shows how the PPV varies as the prevalence changes, given fixed values of sensitivity and specificity. It can also be seen that, even when sensitivity and specificity are very high, the PPV will be quite low if the prevalence is low (Altman and Bland 1994). Similarly for specificity (see Figure 3-3).

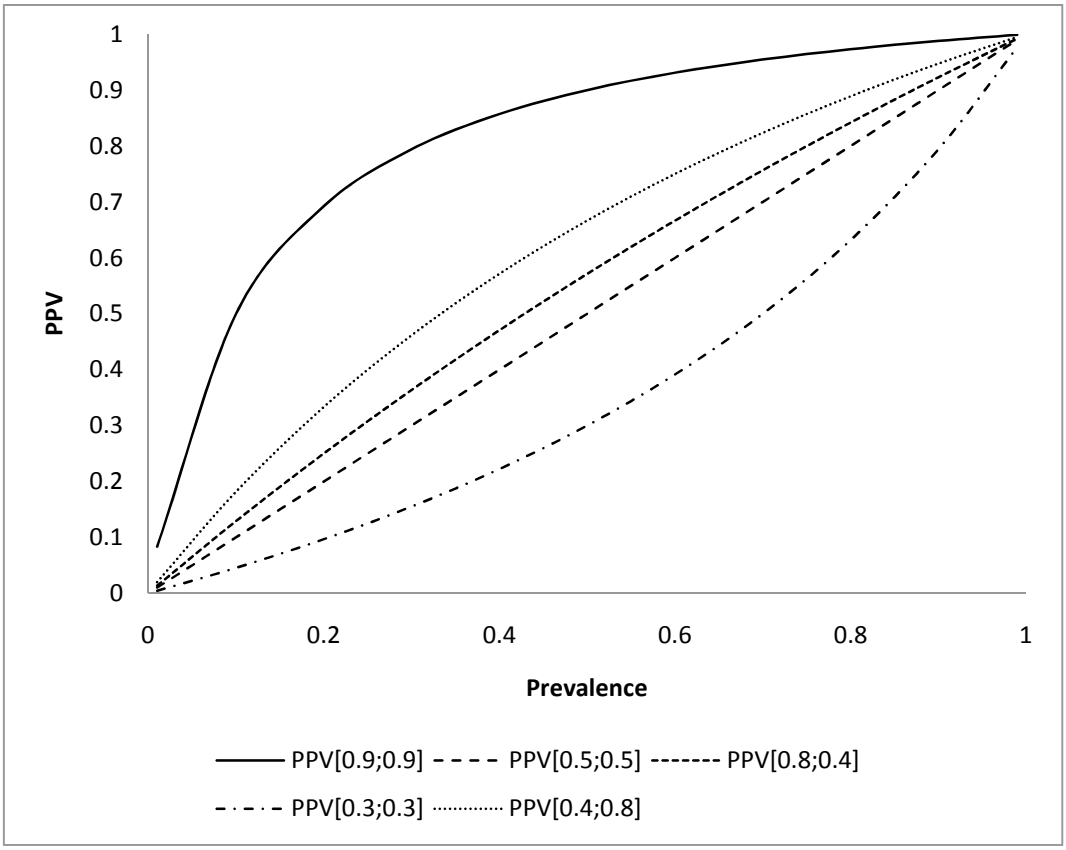


Figure 3-2 Graphical representation of the dependence between PPV and the prevalence of disease, for given levels of sensitivity and specificity (i.e. PPV[sensitivity;specificity]).

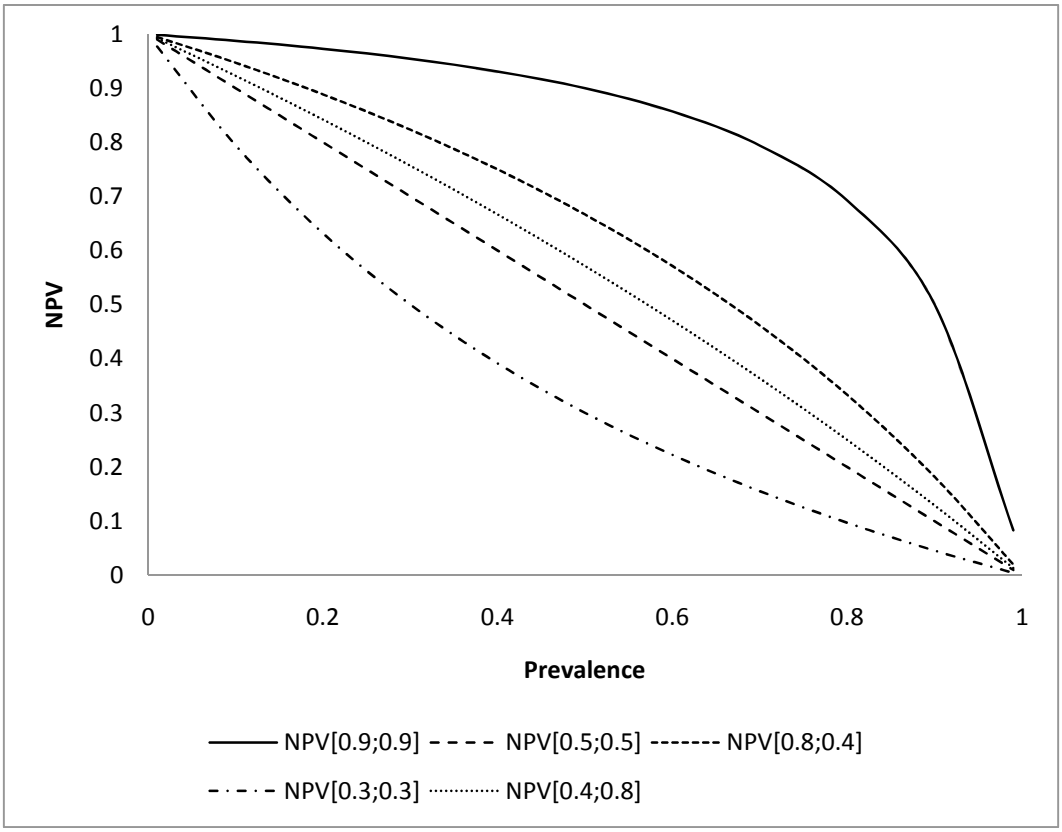


Figure 3-3 Graphical representation of the dependence between NPV and the prevalence of disease, for given levels of sensitivity and specificity (i.e. NPV[sensitivity;specificity]).

Prevalence may be interpreted as the probability to observe the condition prior to the test; consequently, the PPV is the probability to observe the condition posterior to the test (i.e. P(D+|T+)), when the test is positive. Similarly, 1-NPV is the prevalence of disease posterior to the test, when the test is negative. The difference between the prior prevalence and posterior prevalence (i.e. predictive

value) has been indicated as a possible measure of the clinical usefulness of the test (Altman and Bland 1994).

Equation 3-3 shows how to calculate predictive values form the data in a contingency table as Table 3-1.

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

Equation 3-3

Similarly to sensitivity and specificity, predictive values are proportions and confidence or credible intervals can be calculated using analogous techniques.

3.3.4 Likelihood ratios

Other measures can be used to evaluate diagnostic accuracy. They are Likelihood Ratios (LR) and can be calculated in terms of sensitivity and specificity (see Equation 3-4). Deeks and Altman (2004) interpret these measures using the Bayes theorem. Given the interpretation of the prevalence as the pre-test probability of disease as mentioned in section 3.3.3 , it can be easily demonstrated that the post-test odds of having disease is equal to the prior odds [prevalence/(1-prevalence)] multiplied by the positive likelihood ratio (LR+) [sensitivity/(1-specificity)]. Thus, when LR of a test is known, it can be used to transform the prevalence into predictive values for the positives and the negatives results to the test. If LR+>1 the test result is associated with the presence of the disease. If LR+<1 the test

result is associated with the absence of the disease (Deeks and Altman 2004). The further away LR+ is from 1, the stronger the evidence of presence [>10] or absence [<0.01] (Deeks and Altman 2004).

$$LR+= \frac{sensitivity}{1-specificity}$$

$$LR-= \frac{specificity}{1-sensitivity}$$

Equation 3-4

where LR- is the negative likelihood ratio, LR+ is the positive likelihood ratio. Similarly as for relative risks, confidence intervals can be calculated assuming that the log(LR) is normally distributed (Deeks and Altman 2004). Credible intervals can be obtained fitting a conjugate beta-binomial model to the accuracy data to estimate sensitivity and specificity and then these can be used to calculate LR. Either model can be implemented using WinBUGS software for Bayesian statistics or analogous, LR can be expressed as functions of the estimated sensitivities and specificities and uncertainty will be propagated into the estimated LR.

3.3.5 Diagnostic Odds Ratio

Although diagnostic accuracy is naturally bivariate, different attempts exist to use univariate measures. These would simplify the representation of the accuracy of diagnostic tests but, at the same time, may not be able to represent the whole information. The most common univariate measure is the Diagnostic Odd Ratio (DOR) (see Equation 3-5).

$$DOR = \frac{TP\ x\ TN}{FP\ x\ FN} = \frac{sens/(1-sens)}{(1-spec)/spec} = \frac{LR+}{LR-}$$

Equation 3-5

DOR may vary between 0 and plus infinity. Similar to odds ratios, Glas, Lijmer et al (2003) give DOR two possible interpretations:

- how many times is a test more likely to find a positive result in diseased rather than in non diseased,
- how many times is a test more likely to find diseased in those tested positive rather than in those tested negative

As for sensitivity and specificity, DOR does not depend much on the prevalence of disease (Glas, Lijmer et al. 2003) but it is likely to depend on the spectrum

(severity) of disease (Glas, Lijmer et al. 2003). It cannot be used to evaluate a test error rate, at a particular prevalence.

The main problem with this measure is that two tests with the same DOR can have very different sensitivities and specificities (Glas, Lijmer et al. 2003). Therefore, DOR can be useful but it is not very good for comparisons between tests, especially if the impact of false positive results is very different from the impact of false negative results.

Confidence intervals can be calculated using the techniques used for Odds Ratios (Bland and Altman 2000), while credible intervals can be calculated using similar techniques to those introduced for likelihood ratios above.

3.3.6 Receiver Operating Characteristic curves

When the test gives positive/negative results (i.e. it is a dichotomous or dichotomised test), it is easy to use the measures of accuracy described above. But when a test gives results on a continuous or ordinal scale, there is not a clear cut-off point (threshold), and the exploration of diagnostic accuracy over a range of thresholds is needed; Receiver Operating Characteristic (ROC) curves can be used.

A ROC curve is a plot of all the sensitivities and the specificities as the diagnostic threshold varies (Zweig and Campbell 1993). Because sensitivity and specificity derive from two different populations which can be assumed independent (see

Figure 3-1), analogously, ROC curves are independent of the prevalence of disease.

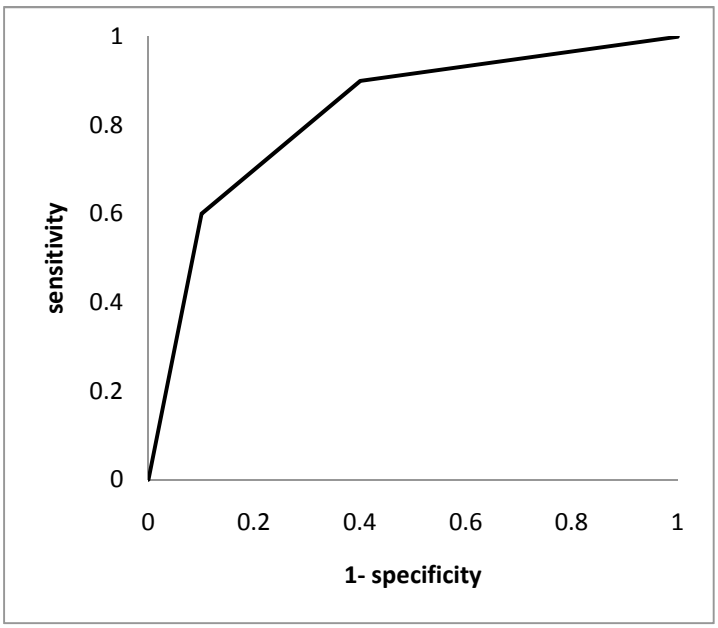


Figure 3-4 Example of ROC curve for a test with 4 possible thresholds.

ROC curves can be used to compare two or more different tests. Only if the ROC curve of a test lies completely above the ROC curve of a second test, it can be said that the first test is better than the second. The best threshold at which the test should operate is called the Q point, which is the point on the ROC curve that lies on the diagonal where sensitivity equals specificity. This may be the best solution for an ROC curve derived from single study data. In Chapter 4 it will be discussed that this choice is more complex when a meta-analytical ROC curve is derived, and in Chapter 7 that the best threshold really depends on the aim of the test (i.e. a triage test would aim to exclude safely healthy patients).

Confidence or credible bands can be obtained based on the fact that an ROC curve is calculated in terms of sensitivity expressed as a function of specificity (or vice versa).

3.3.7 Area under the ROC curve

The ROC curve is a graphical statistic for diagnostic test accuracy. This curve can be used to generate another measure of accuracy: the Area Under the ROC Curve (AUC). This has different interpretations. Glas, Lijmer et al (2003) describes it as the probability that a test correctly ranks 2 individuals (one diseased and one non-diseased). Alternatively, it is the average sensitivity across all possible specificities (Glas, Lijmer et al. 2003) or vice versa. Another interpretation given for diagnostic markers but that can be generalised to the other tests explains the AUC as the probability that the value of the test in the diseased group will be higher than the value of the test in the healthy group (Bamber 1975).

AUC is strictly related to the DOR. In fact, when the DOR is constant for all possible thresholds values, the ROC is symmetric, and the
AUC can be calculated as a function of the DOR (Glas, Lijmer et al. 2003):

$$AUC = \int_0^1 \frac{1}{1 + \frac{1}{DOR(\frac{x}{1-x})}} dx$$

Equation 3-6

where x=1-specificity

It can be seen that DOR and AUC are proportional to each other. Moreover, AUC and DOR have the same limitation: two tests with the same AUC may have different sensitivities and specificities.

The discriminating ability of a test is not always proportional to AUC. Usually, "the greater the AUC the better the test" is valid when tests work on continuous scales, assuming a distribution for the populations of diseased and healthy patients (i.e. normal, log-normal, etc), and a decision rule similar to the following:

- if test >=threshold then T+ (test positive),
- else, T- (test negative).

In this case AUC=0.5 gives a test with very bad discriminating ability, and AUC =1 perfect discrimination. Zhou at al (Zhou, McClish et al. 2002) gave an example of a perfectly discriminating test when AUC =0.5. Suppose the distribution of the 'well' supports values between 80 and 120 (i.e. uniform, truncated normal), and the distribution of the 'sick' are halved symmetrically for test values below 80 and above 120, the resulting ROC curve will have AUC=0.5 despite perfect discrimination ability of the test. However, throughout this thesis the general assumption that higher values of a test are associated to the disease will be made.

If the AUC is used, it should be kept in mind that there is not a functional relationship between AUC and a measure of accuracy. The function

$$f: AUC \rightarrow Accuracy$$

does not exist because the inverse relationship is not monotone (i.e. Accuracy can be expressed in terms of sensitivity and specificity). For example, 2 different asymmetric ROC curves can give the same AUC.

A simple way to calculate confidence or credible intervals for the AUC is by calculating the area under the bands of the ROC curve as introduced in section 3.3.6.

3.3.8 GERD example - accuracy measures

The accuracy measures stated above will now be applied to the GERD dataset introduced in section 3.2.5. A ROC curve and relative AUC will not be given because this dataset provides data for one threshold. The accuracy measures have been obtained by modeling the accuracy data in WinBUGS by means of binomial likelihoods and uniform prior (i.e. conjugate model when the prior is espressed as a beta distribution with both parameters equal to 1, see section 2.2.2 for conjugate models), the code is available in the folder "Chapter 2 – GERD 1 study" contained in the CD-ROM attached to this thesis.

Table 3-2 represents data from a study (Bate, Riley et al. 1999) of the accuracy of PPI therapy used for the diagnosis of GERD (see section 1.3 and 3.2.5 for more details on PPI and GERD).

		Reference test / disease status	
		Diseased	Healthy
Experimental test	Positive	22	11
	Negative	10	15
		32	26

Table 3-2 GERD example single study data.

Table 3-3 presents the parameter estimates for the accuracy of GERD dataset. The prevalence of GERD according to the gold standard is 0.55 (95% CrI 0.42 to 0.68). Only 68% (95% CrI 51% to 82%) of individuals with GERD are correctly diagnosed (estimated sensitivity). Conversely, 57% (95% CrI 39% to 74%) of individuals who are not diseased are correctly diagnosed (estimated specificity). According to this study, the probability of suffering of GERD given a consistent symptomatic reduction (>50% reduction) is 66% (95% CrI 49% to 80%) (estimated Positive Predictive Value). The probability of not having the disease given a not consistent symptomatic reduction (<50% reduction) is 59% (95% CrI 41% to 76%). Predictive values seem to be very informative for patients, which can quantify how much they can rely on their diagnosis. However, for statistical purposes, sensitivity and specificity are independent to disease prevalence.

The odds of having GERD given a consistent symptomatic reduction is 1.64 (95% CrI 1.00 to 2.68) (estimated positive likelihood ratio). The odds of having GERD given a non-consistent symptomatic reduction is 1.875 (95% CrI 1.004 to 3.324) (estimated negative likelihood ratio). Multiplying these two odds together the DOR can be calculated as 3.30 (95% CrI 1.01 to 8.03). The data available were not enough to plot a ROC curve. However, since no test comparisons were required and the threshold was well defined, there was no need to plot any ROC curve. Only in the case where one wants to explore the possibility of other thresholds, a ROC curve would also be plotted [i.e. 40% symptomatic reduction in favour of experimental test].

GERD - parameter estimates	
prevalence	0.55 (95% CrI 0.42 to 0.68)
sensitivity	68% (95% CrI 51% to 82%)
specificity	57% (95% CrI 39% to 74%)
PPV	66% (95% CrI 49% to 80%)
NPV	59% (95% CrI 41% to 76%)
LR+	1.64 (95% CrI 1.00 to 2.68)
LR-	1.875 (95% CrI 1.004 to 3.324)
DOR	3.30 (95% CrI 1.01 to 8.03)

Table 3-3 Parameter estimated for the diagnostic accuracy of GERD dataset.

It may be noted that the estimates of sensitivity and specificity calculated via the Bayesian conjugate beta-binomial model slightly differ from the likelihood based estimates of sensitivity (0.69, standard error 0.08) and specificity (0.58, standard error 0.10) reported by Bate et al. Although this small difference (one percentage point in both cases) may be neither clinically significant (i.e. it would not change the decision on whether to give the test or not) nor economically significant (i.e. the trade-off between false negatives and false positives may not be affected by this difference), it was not expected after the use of a vague prior distribution. As already discussed by Lambert et al. (2005), the vagueness of prior distributions is relative to the amount of information that derives by the data. For the example presented in this section, the calculation of sensitivity and specificity is based on 32 and 26 observed patients respectively; therefore, the prior distribution has a stronger influence on the estimated posterior rates than if more patients were observed.

3.4 Discussion: what is the most suitable measure to represent Diagnostic Accuracy?

Which of these measures should be used is difficult to say. In general, this may depend on the type of test and on the individual who is going to use these measures. The individuals that may be interested to know about the accuracy of a diagnostic test are:

- *Patients*, in order to interpret correctly a statement about their health status. In fact, this information can be helpful to patients to decide between different treatments where possible. This may involve some better representations (i.e. less mathematical) of diagnostic accuracy than the ones that have been mentioned in this chapter. Furthermore, often diagnosis is done by patients themselves (*self-diagnosis*), and they are asked to give a first interpretation or make a first inference based on a test result. For example, pregnancy tests are available in pharmacies.
- *Clinical staff or doctors*. A clinician uses the results of a test to make choices that may seriously affect the conditions of a patient and his quality of life. Thus it is crucial to have in mind how likely the test is to be wrong, and what are the associated implications.
- *Policy makers*. Policy makers not only take decisions on which tests must not be used, but often indicate the best way to use a test, or a pattern that maximizes its accuracy.

- *Statisticians*. Although policy makers are often statistical experts, in this category all the epidemiologists can be included that conduct their experiments often ignoring the effect of diagnostic accuracy. The appraisal of the efficacy of a treatment should not ignore the effect of wrong diagnosis.

For each of these individuals a different way to represent diagnostic accuracy may be needed. And each of the accuracy measures presented above can be helpful to better understand the role of a test, to compare tests or to consider its usefulness or efficiency.

3.5 Summary

In this chapter, characteristics of diagnostic tests and the most common measures used for diagnostic accuracy were described. Finally, a brief discussion tries to identify the most suitable diagnostic accuracy measure and concludes that different measures can be meaningful according to the purpose such measures are going to be used for. The most common accuracy measures and their graphical representation were illustrated using the results of a single study of the diagnosis of GERD (see section 1.3 for details on GERD example). The estimates of such accuracy measures were obtained via Bayesian modeling; the model is given in the CD-ROM within the folder Chapter 2 - GERD 1 study. In conclusion, the measures that have better statistical properties are sensitivities and specificities because they are not strongly affected by the prevalence of disease. Also, the other measures can be obtained in terms of sensitivity and specificity.

Chapter 4. Meta-Analysis of diagnostic test accuracy measures and principles of cost-effectiveness analysis

4.1 Chapter overview

This chapter describes the main parametric statistical techniques, and classifies and quantifies heterogeneity for meta-analysis of diagnostic test data. It also describes the methodology for economic decision modeling in brief.

Meta-analysis for diagnostic test data can be considered a two stage process: firstly summary measures are derived for each trial/study as a result of a systematic review, and then pooled estimates are calculated (Egger, Smith et al. 2001). An introduction to meta-analysis is given in section 4.2.

As discussed in Chapter 3, diagnostic test data are naturally bivariate, and the diagnostic threshold represents a further source of heterogeneity along with differences between studies (i.e. design) and populations (i.e. baseline prevalence of disease). The fact that studies are conducted in randomized populations, a single number can still be inappropriate to summarize the study results in a meta-analysis. How to quantify and interpret heterogeneity will be presented throughout the following section 4.4.

Section 4.5 introduces the most common statistical model for the meta-analysis of dichotomised diagnostic test data, shows how to implement these in a Bayesian framework and how to explore heterogeneity via the inclusion of covariates.

Throughout section 4.5, the GERD example introduced in section 3.2.5 will be extended to the context of a meta-analysis (Numans, Lau et al. 2004) and be used to show the results and discuss some issues relative to each technique (see Table 4-3 for GERD data). The code to implement these models in WinBUGS are available in the folder "Chapter 4 – GERD meta-analysis" contained in the CD-ROM that is attached to this thesis.

Although the models presented in this chapter are based on different assumptions, they are all attempts to synthesise the same type of data (i.e. dichotomous or dichotomised diagnostic test accuracy data from a number of studies). Section 4.3 presents the data for meta-analysis of diagnostic test accuracy in tabular form. Section 4.7 describes the relationships between the different syntheses models described in section 4.5.

Formulae to plot sROC curves are given throughout section 4.5; however, section 4.6 gives a brief review to the approaches to summary ROC curves (Novielli, Cooper et al. 2010) and exploits the differences between these using similar formulae (sROC, i.e. as a result of a meta-analysis).

Finally, an overview of the theory behind and methods used for economic evaluations and comprehensive cost-effectiveness analysis is given in section 4.8.

4.2 Introduction to meta-analysis

Systematic reviews and, consequently, meta-analyses play the crucial role of key-sources of evidence for medical research (Lau, Schmid et al. 1995; Mosteller and Colditz 1996; Wallace, Schmid et al. 2009). Meta-analysis is used to combine and summarise quantitative evidence from a number of articles. Some argue that a decision cannot be based on a single study any longer (Lau, Ioannidis et al. 1998; George, Oscar et al. 2009) and meta-analysis needs to be considered a high level statistical exercise rather than a mere procedure.

Meta-analysis increases the statistical power of the single studies (Wallace, Schmid et al. 2009), thus allowing the detection of small effects for which single studies may be underpowered. Moreover, study specific characteristics can determine differences between study estimates, which cause problems in the choice of one of those studies either for decision making or for inference. Therefore, meta-analysis allows for the exploration of such systematic differences also called heterogeneity (Higgins, Thompson et al. 2009), and separate these from aleatoric uncertainty which is due by chance. Similarly, it allows for the detection and corrections of some forms of biases (i.e. publication bias).

4.3 Diagnostic data from multiple studies

Diagnostic data from one study were represented in Table 3-1. Similarly, Table 4-1 presents diagnostic data from one study that will be used in a meta-analysis, that is where the fact that the data come from a specific study is recorded by mean of the index i. TP_i is the number of True Positives for the i^{th} study. Similarly, FP_i is the number of False Positives, TN_i is True Negative and FN_i is False Negative. Diagnostic measures can be easily calculated for every study as shown in Chapter 3. Usually true positive rates (i.e. sensitivity) and true negative rates (i.e. specificity) are considered to have better statistical properties than others (i.e. predictive values are strongly variable with prevalence, see 3.3.3 for details).

i^{th} study		Reference test / disease status	
		Diseased	Healthy
Experimental test	Positive	TP_i	FP_i
	Negative	FN_i	TN_i
		D_i^+	D_i^-

Table 4-1. Aggregate study data, single study, i in 1 to N.

Table 4-2 represents the data that can be collected as part of a systematic review, where X_i is the value of the study level covariates for study i. These covariates can be used to explore the possible sources of heterogeneity (presented throughout section 4.5 for different approaches to meta-analysis of diagnostic test data).

Stuy ID (i)	TP	FP	FN	TN	Covariate
1	TP_1	FP_1	FN_1	TN_1	X_1
2	TP_2	FP_2	FN_2	TN_2	X_2
..	..	..	..	..	..
I	TP_i	FP_i	FN_i	TN_i	X_i
..	..	..	..	..	..
N	TP_N	FP_N	FN_N	TN_N	X_N

Table 4-2 Example of meta-analysis data for diagnostic test.

Sometimes, for rare diseases or small studies, zero count data may be extracted. Classical analysis usually uses continuity corrections in order to calculate diagnostic accuracy statistics (i.e. if $TP = 0$ the DOR can not be calculated, therefore a small quantity δ can be added to each cell of the 2 by 2 table, usually $\delta = 0.5$). However, the results may be sensitive to the correction depending on the number of patients observed and on the value of δ (Sweeting, Sutton et al. 2004). In this thesis, the use of Bayesian models as specified in Chapter 2 and implemented in the software for Bayesian meta-analysis WinBUGS (Lunn, Thomas et al. 2000) will allow the likelihoods for count data to be modelled directly using binomial distributions and continuity corrections will not be needed.

4.4 Classification and qualification of heterogeneity for meta-analyses of diagnostic test data

4.4.1 Statistical, clinical and diagnostic heterogeneity

Egger et al. (2001) give a very clear explanation of the different motivations and purposes behind individual studies and meta-analysis studies. Individual studies test the effect of a treatment on a clinical situation (i.e. risk of death) given test regimen/protocol (i.e. duration) and a selected population (i.e. eligibility criteria). Meta-analysis studies estimate the extent to which a treatment (represented by a variety of study averages) influences the clinical situation, bringing a gain in objectivity, applicability of results, and precision from all available evidence. Similarly, for meta-analysis of diagnostic data, the subject is a diagnostic test, its effect is the classification in positives or negatives of a sample of the population of symptomatic patients, and the clinical effect can be measured in terms of accuracy of the diagnosis. Moreover, meta-analysis results can be more precise for the possibility to evaluate between studies variability (Statistical heterogeneity). Statistical heterogeneity is quantifiable (see next section) and the exploration of its sources gives the possibility to describe it qualitatively. Egger et al. (2001) described statistical heterogeneity as a result of:

1. Clinical heterogeneity. Due to qualitative differences between studies (i.e. protocol, eligibility criteria, etc)
2. Methodological heterogeneity. Due to different analytical strategies (i.e. study design: cohort study or case control study)

3. Residual heterogeneity. Due to unknown/unrecorded trial characteristics, which other authors simply call uncertainty (not explicable) (Briggs, Claxton et al. 2006)

The following form of heterogeneity is specific to diagnostic data and is due to those factors that make the diagnostic threshold vary between studies:

Diagnostic heterogeneity: Variability between studies can be due to three main factors (Littenberg and Moses 1993; Egger, Smith et al. 2001):

1. By chance (as part of uncertainty which cannot be explained)
2. By changes in diagnostic threshold. These variations can be either explicit or implicit (due to difference between observers, measurement techniques, laboratories protocols, precision of instruments, even the prevalence of disease can influence the choice of a different threshold).
3. Other factors (different reference tests, different types of patients, different populations with different prevalence, different reference tests, different study methodologies, patient selection criteria/method, study design, etc), sometimes referred to as implicit threshold (Sterne 18 November 2009)

When the change in threshold causes heterogeneity, the sensitivities and specificities plotted on the ROC plane lie along an underlying ROC curve,

deviations from this curve are caused by methodological/clinical heterogeneity, and uncertainty.

4.4.2 Sources of variation and bias

The classification of the sources of heterogeneity given above is essential for diagnostic data accuracy, and highlights the importance of investigating such sources.

While diagnostic heterogeneity is quite a clear concept (since it is generated by different experimental tests, different standard test, different technologies, different observers, etc), methodological and clinical heterogeneity may be quite vague terms since the range of sources is very broad. Here it is distinguished between two main types of statistical (both diagnostic and methodological and clinical) heterogeneity that can be found in model fitting: variation and bias. Whiting et al performed a literature search in order to explore the main sources of variation and bias in diagnostic accuracy studies (Whiting, Rutjes et al. 2004). They identified many areas of bias and variation:

Population: Differences in populations and demographic features affect diagnostic accuracy measures to different extents and directions. Also disease severity and prevalence were considered population characteristics able to affect accuracy;

Test protocol: Differences were found according to the degree of expertise required to perform the test;

Reference standard: Strong evidence was found on the influence of verification procedures on accuracy results; inappropriate or inaccurate reference test may strongly bias the analysis, inflating the sensitivity (less false positive may be detected);

Interpretation: Reading processes are related to interpretation; they were found to affect sensitivity. Also different observers may be element of bias/variation;

Analysis: A few studies investigated the effect of different analysis strategies. This analysis may be carried out via sensitivity analyses, that are aimed to study the sensitivity of results to different approaches (i.e. or use of different model strategies or use of different prior distributions or multiple chains for Bayesian models);

Study design (Lijmer, Mol et al. 1999): Study design related bias strongly affect study results. For example *selection bias* occurs when not all the people that meet the inclusion criteria are selected or this selection is not properly randomized. *Verification bias* and *partial verification bias* (or workup bias) occur when not all the studies were verified with a reference test (i.e. when it is too invasive, it may not be performed or it may be performed indirectly by follow up). Others are: *Inappropriate blinding, Publication bias* (encouraging results may give more chance of publishing).

The idea of bias is strictly linked to the idea of variation since the same source of variation (i.e. observer) may also be interpreted as a source of bias. Some of these sources of variation may also be study specific, and different phenomena are called by the same name or vice versa, making difficult any accurate qualitative or quantitative description of sources of variation. Nowadays, it is not always clear whether the word variation is more correct than the word bias in studies of diagnostic accuracy, it is difficult to generalize and study specific considerations are needed.

Although diagnostic heterogeneity is qualitatively clear, it is difficult to distinguish diagnostic heterogeneity from other sources of heterogeneity when they are quantified. Similarly, the exploration of bias is very difficult (i.e. publication bias methods for meta-analysis of diagnostic accuracy are not established).

4.4.3 The quantification of heterogeneity in meta-analysis

Quantification of heterogeneity may be a mere calculation of an index or a more complex exploration of the causes of such variability. Since in diagnostic modeling the bivariate nature of the data reflects in multiple sources of heterogeneity, exploration of its sources becomes essential. Three main approaches to quantification and exploration of heterogeneity can be identified from the literature: the use of random effects; the use of *ad-hoc* indexes (based on random effects models); and the use of covariates or subgroup analysis.

Assumption of heterogeneity via random effect modeling

When a meta-analysis is performed it is possible to estimate the precision of parameter estimates. Let one say LOR (Log Odds Ratio) is the estimated quantity, its precision ($^{1}/_{var(LOR)}$) is the inverse of the sum of a within study component (σ_i^2) and a between study component (τ^2). If total consistency of single study estimates with the grand mean from the meta-analysis (i.e. homogeneity) is assumed, it is consequently assumed that $\tau^2 = 0$ (Egger, Smith et al. 2001) which corresponds to the calculation of fixed effects estimates. Where it is assumed τ^2 not necessarily equal to zero, random effect estimates are being calcualted. Thus, the inclusion of random effects such as τ^2 in hierarchical models allows for the quantification of residual heterogeneity (Sutton, Abrams et al. 2000).

In the case of the estimation of LOR, it is assumed that study LORs are normally distributed with known variances. Sparse data may affect this assumption, thus a Bayesian modeling approach may be helpful as explained above. In this case, it would also be possible to consider precision in the estimate of τ^2 via credible intervals.

Some shortcomings of τ^2 include (Higgins, Thompson et al. 2009):

1. Aggregation bias (i.e. considering in the same estimate measures based on different popupations or different study designs)

2. It depends on the scale the effect is measured on (i.e. LOR on log scale, difficult to interpret, difficult to compare two effects on two different scales)
3. Depends on the number of studies in the meta-analysis

Q-test for diagnostic test data

The issue of quantifying heterogeneity in meta-analysis is well known, and needs to be considered in diagnostic accuracy studies since they are subject to many sources of between study variability. This can be achieved by describing the between study variance via random effect models. Thus they are required to follow properties well described by Higgins and Thompson (2002). They need to be independent of the extent of heterogeneity (i.e. monotone as the heterogeneity varies), scale invariant and size invariant (i.e. independent of the number of studies). However, the measures they elaborate (H^2, I^2, R^2) still depend on the between study variance term $\sigma_i^2 = var(\theta_i)$, where θ_i is the study specific estimated effect in the meta-analysis. Moreover, those measures fit well univariate meta-analyses, but cannot be easily generalised in case of multivariate meta-analysis and existing methods need to be adapted and tested for diagnostic studies (Jackson, White et al. 2010).

Exploration of sources of heterogeneity and measuring the extent of heterogeneity: regression modeling and sub-group analysis

The indexes presented above do not have enough power to detect small evidence of heterogeneity (Egger, Smith et al. 2001). Moreover, all the studies included in a meta-analysis can be considered heterogeneous according to clinical and methodological differences. Therefore, further exploration of heterogeneity, that is independent to the results of such tests, is necessary in order to find out its possible sources. Two approaches are possible at this stage: *i*) to explore the effect of covariates, and *ii*) to perform sub-groups analysis.

Covariates. Two types of covariates exist according to whether they vary between or within studies. In both cases they can be included in a meta-regression. In the first case they will participate directly to the explanation of heterogeneity while within studies covariates can be implemented only if Individual Patient Data (IPD) is available. When variables are strongly related, interpretation of results becomes more difficult (Egger, Smith et al. 2001). Thus, explorative analysis in this sense should be preformed.

Subgroups. The availability of IPD, when either continuous or ordering variables are considered, makes possible an accurate subgroup analysis since subgroups difference across studies can be replicated. Even when there are no IPD available

it is possible to carry out a subgroup analysis in which a number of studies are considered in each group.

4.4.4 Heterogeneity in GERD dataset

Here some brief examples are given of the ideas described in section 4.4 applied to GERD dataset presented in Chapter 3. The GERD dataset for the meta-analysis is represented in Table 4-3 (Numans, Lau et al. 2004) and includes the study by Bate et al already used in Chapter 3 to illustrate examples (Bate, Riley et al. 1999). Note that study number 3 will need a continuity correction if classical approaches to meta-analysis are used as specified in section 4.3. Some possible sources of heterogeneity are introduced (i.e. study setting affects the methods for data collection, hence their precision). However, study settings may also produce diagnostic heterogeneity if a threshold is not well specified or the clinicians are not well trained to use the test. Another source of heterogeneity is the reference test. In this dataset the reference test was not unique. Either a symptomatic check (i.e. presence of esophagitis) or a pH measurement approach is used. This may introduce clinical heterogeneity, since the threshold does only depend on the experimental test.

Random effects may be calculated to quantify heterogeneity. However, a random effect does depend on the model (i.e. model parameter, structure). In section 4.5

different approaches will be presented to the meta-analysis of diagnostic test data and applied then to this dataset.

ID	Author	Year	TP	FP	FN	TN	Reference test	Study setting
1	Bate	1999	22	11	10	15	24h-PH	Secondary/Specialist care
2	Fass	1999	28	3	7	4	24h-PH	Secondary/Specialist care
3	Fass	2000	21	8	0	6	24h-PH	Secondary/Specialist care
4	Juul-Hansen	2001	29	11	5	11	24h-PH	Secondary/Specialist care
5	Schenk	1997	15	7	7	12	24h-PH	Secondary/Specialist care
6	Carlsson	1998	66	25	72	62	Esophagitis	Primary
7	Galmiche	1997	27	65	10	39	Esophagitis	Secondary/Specialist care
8	Hatlebakk	1999	55	59	22	25	Esophagitis	Primary
9	Schenk	1997	9	13	6	13	Esophagitis	Secondary/Specialist care
10	Johnsson	1998	50	8	17	5	Esophagitis	Secondary/Specialist care
11	Venables	1998	80	120	21	109	Esophagitis	Primary

Table 4-3 GERD meta-analysis example dataset.

4.5 Meta-analysis techniques for diagnostic accuracy studies

Sensitivities and specificities are not the only accuracy measures for diagnostic data but they are recommended as the best measures for meta-analysis compared to predictive values and likelihood ratios. Predictive values are not usually meta-analysed because they have been shown to be strongly dependent on the prevalence of disease (see Chapter 3). The meta-analysis of the likelihood ratio statistics can give impossible results of sensitivities and specificities when back-transformed (Zwinderman and Bossuyt 2008). It has been recommended by other authors that sensitivities and specificities be used for meta-analysis (Egger, Smith et al. 2001), and then use these to obtain likelihood ratios, which could then be used to calculate posterior disease probabilities (predictive values). On the contrary, direct meta-analysis of likelihood ratios would not allow calculation of other accuracy measures.

Recently, methods for the meta-analysis of diagnostic test data that aimed to select the optimal threshold have been published (Rucker and Schumacher 2010); however, the focus of this thesis is on methods for the meta-analysis of diagnostic accuracy data for tests evaluated at their operative threshold (i.e. the threshold used in practice such that recommended by the test producers). The most common methods for the meta-analysis of diagnostic test data for dichotomised tests evaluated at their operative threshold can be divided in three groups:

1. **Independent estimates of diagnostic rates.** The main assumption is independence between rates, which corresponds to no heterogeneity due to differences in diagnostic threshold (see section 4.5.1).
2. **Summary ROC (sROC).** These are attempts to relax the assumption of independence between rates and capture variability due to differences in diagnostic threshold. Symmetric and asymmetric sROC curve approaches to the meta-analysis of diagnostic test data are considered in section 4.5.2.
3. **Bivariate estimates of diagnostic rates** are considered in section 4.5.3. This allows direct estimation of correlation between diagnostic rates as a measure of heterogeneity due to variability in threshold. In the same section, Hierarchical ROC (HsROC) modeling is presented as an almost always equivalent modeling approach.

Each technique will be presented in its classical form for general consideration. Also the Bayesian modeling to be implemented in WinBUGS will be given as published by Novielli, Cooper et al (2010). Models that are fit on sensitivity and specificity are equivalent to those that are fit on sensitivity and 1-specificity. The choice of one or the other will be made to facilitate the comparison between these approaches in section 4.7.

4.5.1 Meta-analysis approaches where independence between rates is assumed

General framework

A first step to the meta-analysis of diagnostic accuracy is to consider independently sensitivity and specificity. The main assumption is that the diagnostic threshold is the same among the studies, and this means assuming no correlation between sensitivities and specificities. If this assumption is not true then *i*. sensitivities and specificities appear correlated on the ROC plane (i.e. they do not look randomly scattered but lie along an underlying ROC curve, that is with a negative correlation); *ii*. the model can be improved to quantify diagnostic heterogeneity (i.e. due to variability in threshold, see section 4.4.1). This assumption can be verified either by graphical methods (i.e. see point *i* above) or by statistical methods (i.e. the chi-squared test can be used to check for the heterogeneity if there is enough data). The Spearman Rho can be used to check for independency between sensitivities and specificities. In the case of sparse data, chi-squared or Rho may not be used and it still remains an undetected variability that needs to be explored. If this assumption holds or if a first naive measure of diagnostic accuracy needs to be obtained, pooled estimates can be calculated with an approximation of the inverse variance approach (Egger, Smith et al. 2001):

$$Pooled\ sensitivity = sens^* = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} D_i^+}$$

$$Pooled\ specificity = spec^* = \frac{\sum_{i=1}^{N} TN_i}{\sum_{i=1}^{N} D_i^-}$$

Equation 4-1

Standard errors can be calculated when the study sample sizes are large (i.e. >30). If they are not, bootstrapping methods can be used. In general, methods for proportions are well suited to summarize such data.

$$se(sens^*) = \sqrt{\frac{sens^*(1 - sens^*)}{\sum_{i=1}^{N} D_i^+}}$$

$$se(spec^*) = \sqrt{\frac{spec^*(1 - spec^*)}{\sum_{i=1}^{N} D_i^-}}$$

Equation 4-2

Bayesian model for independent sensitivities and specificities

Bayesian modeling based on the MCMC algorithm offers a way to calculate pooled estimates. Models implemented in WinBUGS do not need continuity corrections in case of zero counts. Also, within study variability can be modelled directly using binomial distributions (Rutter and Gatsonis 1995; Hamza, van Houwelingen et al. 2008) and this structure will be common to all models.

$$TP_i \sim binomial(p_{1i}, D_i^+)$$

$$FP_i \sim binomial(p_{2i}, D_i^-)$$

$$logit(p_{1i}) = \mu_{1i}$$

$$logit(p_{2i}) = \mu_{2i}$$

Fixed effect $\begin{cases} \mu_{1i} = \theta_1 \\ \mu_{2i} = \theta_2 \end{cases}$

Random effect $\begin{cases} \mu_{1i} \sim N(\theta_1, \sigma_1^2) \\ \mu_{2i} \sim N(\theta_2, \sigma_2^2) \end{cases}$

Prior distribution on $\theta_.$ and $\sigma_.^2$

Model 4-1 Bayesian model for independent estimates of sensitivity and specificity, fixed and random effect, no covariate included.

Model 4-1 shows both fixed effect and random effect models for independent estimates. p_{1i} (p_{2i}) are study specific logit(sensitivity) (logit(1-specificities)) estimated directly by the model. Study specific logit transformations are calculated and assumed to be distributed normally (μ_{1i} , μ_{2i}). In the case of the fixed effect model, all studies are assumed to have the same effect for sensitivity (θ_1) and for specificity (θ_2). In the case of random effect modeling, all the studies are considered exchangeable (Bernardo and Smith 1994). That is, study specific effect estimates are realizations from the same underlying distribution (i.e. normal) with mean θ_1 (θ_2) and variance σ_1^2 (σ_2^2). Assuming the estimates to be independent, Model 4-1 is actually the synthesis of four different models that

could be fitted separately (i.e. different pieces of code): two for sensitivities (fixed and random effect) and two for 1-specificities (fixed and random effect).

The inclusion of covariates. Since the inclusion of covariates may complicate the models, for clarity, their inclusion will be described separately for every model.

$$TP_i \sim binomial(p_{1i}, D_i^+)$$

$$FP_i \sim binomial(p_{2i}, D_i^-)$$

$$logit(p_{1i}) = \mu_{1i}$$

$$logit(p_{2i}) = \mu_{2i}$$

$$\text{Fixed effect} \quad \begin{cases} \mu_{1i} = \alpha_1 + \beta_1 X_i \\ \mu_{2i} = \alpha_2 + \beta_2 X_i \end{cases}$$

Prior distribution on $\alpha_.$ and $\beta_.$

$$\text{Random effect} \quad \begin{cases} \mu_{1i} \sim N(\theta_{1i}, \sigma_1^2) \\ \mu_{2i} \sim N(\theta_{2i}, \sigma_2^2) \end{cases}$$

$$\theta_{ji} = \alpha_j + \beta_j X_i; j = 1,2$$

Prior distribution on $\alpha_.$, $\beta_.$ and $\sigma_.^2$

Model 4-2 Bayesian model for independent estimates of sensitivity and specificity, fixed and random effect, covariate included.

Model 4-1 are based on well known logistic regression models for meta-analysis (Ntzoufras 2010). In case of fixed effect, the study specific effects (μ_{1i} , μ_{2i}) are associated to a linear predictor that is a regression equation rather than to a single

parameter. Thus, β_j is a vector of parameters describing the effect of the vector of covariates X. α_j is the intercept of the linear meta-regression line. In case of random effect modeling, study specific estimates are assumed as realizations θ_{ji} of the same (normal) distributions, with between study variability term σ_j^2 , which measures the amount of heterogeneity remained unexplained (Higgins, Thompson et al. 2009). In both fixed and random effect estimate, normal priors can be put on the regression parameters.

GERD example- independent estimates

GERD data in Table 4-3 can be summarized in terms of single independent estimates in order to have a first estimation of diagnostic accuracy rates. Sensitivity is estimated at 0.694 (95%CI 0.657 to 0.732) and false positive rate is estimated at 0.523 (95%CI 0.484 to 0.562). Specificity can be obtained by subtracting the false positive rate from 1. In this case, I did not apply any continuity correction. A continuity correction would have lowered slightly this estimate. However there is no evidence on what is considered the right approach in this case of sparse data (Sweeting, Sutton et al. 2004).

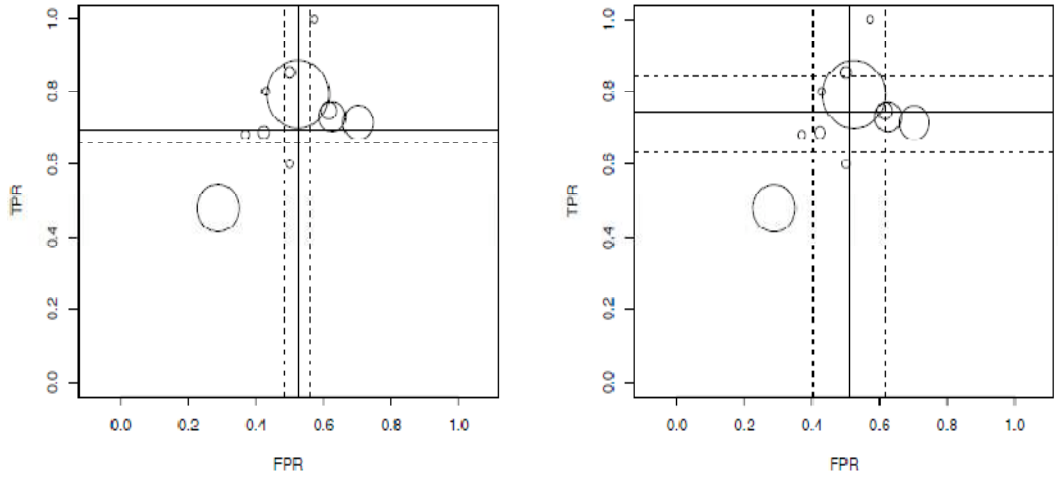


Figure 4-1 Independent estimates model: (a) fixed effect, (b) random effect.

When models are implemented in a Bayesian framework, estimated fixed effect sensitivity (0.695, 95%CrI 0.657 to 0.731) and false positive rate (FPR) (0.523, 95%CrI 0.483 to 0.561) are very similar to those obtained by the classical approach (i.e. using non informative priors). Figure 4-1(a) shows pooled rates and confidence intervals on the ROC plane: two single and uncorrelated estimates crossing at the point on the ROC plane that represents the overall accuracy of the test. When random effects are considered as in Figure 4-1 (b), the estimate of sensitivity is slightly higher (0.744, 95%CrI 0.634 to 0.842), similarly specificity (1-specificity is 0.511, 95%CrI 0.403 to 0. 617). The estimated between-study standard deviations were σ_1 =0.567 (95%CrI 0.280 to 1.03) and σ_2 = 0.770 (95%CrI 0.368 to 1.55) respectively; these indicate an amount of unexplained

heterogeneity that needs to be investigated Since the heterogeneity detected above has to be explored when possible, the effect of study settings on accuracy estimates will be analysed. A meta-regression equation can be put on either the fixed effect or the random effect model. According to the fixed effect model, little difference in study settings on specificity is observed ($FPR_{primary}$ 0.473 95%CrI 0.472 to 0.592) and ($FPR_{secondary}$ 0.510 95%CrI 0.461 to 0.559). Differences in settings seem to affect more the sensitivities: ($TPR_{primary}$ 0.481, 95%CrI 0.349 to 0.618) and ($TPR_{secondary}$ 0.635 95%CrI 0.580 to 0.688).

When random effect modeling is used, little difference are detected between specificities: ($FPR_{primary}$ 0.4502, 95%CrI 0.1116 to 0.7878) and ($FPR_{secondary}$ 0.4829, 95%CrI 0.2832 to 0.6598); while higher is the effect of setting on the sensitivity: ($TPR_{primary}$ 0.5514, 95%CrI 0.0973 to 0.9032) and ($TPR_{secondary}$ 0.6695, 95%CrI 0.4087 to 0.8552). Still estimated between-study standard errors may be observed: (σ_1 1 0.789, 95%CrI 0.344 to 1.655) and (σ_2 2 0.620, 95%CrI 0.230 to 1.204). The inclusion of study setting does not resolve all the heterogeneity. This means that other sources could be explored.

There was no evidence of sensitivity to prior distributions for this model. The prior distributions used for the logit rates parameters and the regression parameters where normal with mean close to zero (i.e. 0.5 or 1) and very low precision (i.e. 1.0E-6 in WinBUGS code corresponds to a standard deviation of 1000). The heterogeneity parameter (i.e. the random effect) was given a prior

uniform on the range of its plausible values; for example σ was assumed a priori uniform between 0 and 10.

4.5.2 Meta-analysis approaches that pool summary ROC curves

Meta-analysis of sensitivity and specificity assuming independence does not consider variability in threshold. This assumption of independence between the rates can be relaxed via the calculation of sROC curves, which, by definition, represent pairs of sensitivities and specificities at different threshold levels. ROC curves can be pooled considering the DOR (see Chapter 3). When DOR is considered *i*. it represents a unique measure of accuracy (although different couples of diagnostic rates can lead to the same DOR, see Chapter 3), and *ii*. sROC curves can be calculated. DOR can be assumed constant across studies and produce symmetric sROC curves. When this assumption is relaxed asymmetric sROC are produced. In both cases, a random effect approach for the quantification of the unexplained clinical and methodological heterogeneity is possible.

Symmetric summary ROC curves

A sROC curve symmetric around the line sensitivity=specificity on the ROC plane can be fitted if the DOR is assumed constant across studies (Leeflang, Deeks et al. 2008). This approach does not allow for a proper exploration of variability in threshold (i.e. quantification of diagnostic heterogeneity) but relaxes the assumption of independence between rates.

After a pooled DOR is calculated using standard meta-analytic methods, the symmetric sROC curve in terms of sensitivity as function of specificity (or vice versa) can be calculated using Equation 4-3.

$$sens_j = \frac{1}{1 + \frac{1}{DOR\frac{(1 - spec_j)}{spec_j}}}$$

Equation 4-3

Where, for example, $spec_j$ is a given set of values of specificity between 0 and 1 (0.1,0.2,0.3,...,0.8,0.9). The more precise the plot of the sROC curve the less the space between these values. Thus, Area Under the sROC Curve (AUC) can be calculated by integrating the curve between zero and one (see Equation 4-4).

$$AUC = \int_0^1 \frac{1}{1 + \frac{1}{DOR\frac{(1 - x)}{x}}} dx$$

Equation 4-4

Bayesian model for symmetric ROC curves

The implementation that is proposed for this model is based on the formulation of the log DOR as the difference between the two logit rates. In the case of the fixed

effect model, study specific log-DOR are assumed to be the same value, while in the case of the random effect model, study specific log-DOR are assumed normally distributed with mean Θ_{RE} and between study variance σ^2.

$$TP_i \sim binomial(p_{1i}, D_i^+)$$

$$FP_i \sim binomial(p_{2i}, D_i^-)$$

$$logit(p_{1i}) = \mu_i + d_i$$

$$logit(p_{2i}) = \mu_i$$

$$\text{Fixed effect} \qquad \{d_i = \Theta_{\mathrm{FE}}$$

Prior distributions on μ_i and Θ_{FE}

$$\text{Random effect} \qquad \{d_i \sim N(\Theta_{\mathrm{RE}}, \sigma^2)$$

Prior distributions on μ_i, Θ_{RE} and σ^2.

Model 4-3 Symmetric sROC model, without covariates.

Some discussion is required on the interpretation of random effect models for the DOR. If there is heterogeneity in the pooled log-DOR it means either *i*. a fixed effect is not enough to estimate the constant DOR (inclusion of random effect or covariates is recommended) or *ii*. DOR cannot really be considered constant across studies (i.e. definition of source of heterogeneity as between studies differences). When the random effect model is used, different log-DOR for every study are calculated and assumed exchangeable. Graphically on the ROC plane it is similar to plot a family of parallel symmetric ROC curves.

The exploration of heterogeneity by adding covariates is straightforward and requires a little modification to the linear predictor of the model:

$$TP_i \sim binomial(p_{1i}, D_i^+)$$

$$FP_i \sim binomial(p_{2i}, D_i^-)$$

$$logit(p_{1i}) = \mu_i + d_i$$

$$logit(p_{2i}) = \mu_i$$

$$\text{Fixed effect} \qquad \{d_i = \alpha + \beta X$$

Prior distributions on α, β and μ_i

$$\text{Random effect} \qquad \begin{cases} d_i = \alpha_i + \beta X \\ \alpha_i \sim N(A, \sigma^2) \end{cases}$$

Prior distributions on μ_i, A, σ^2 and β.

Model 4-4 Symmetric sROC model, with covariates.

where α and β are the intercept (baseline log-DOR) and the slope (incremental log-DOR). In case of random intercept model, a family of ROC curves (parallel lines on the logit scale) are estimated, with different slopes according to the value of the covariate.

Asymmetric summary ROC curves

In the previous section, a set of fixed effect and random effect methods are presented to estimate a common DOR between studies, assumed constant between studies. Littenberg and Moses (1993) relaxed this assumption of constant DOR allowing for a better account of diagnostic heterogeneity. This model accounts for

differences in the variances of the distributions of the diseased and the healthy patients over the test values.

This method was conceived as a fixed effect model although an improved version of this model exists that allows for the inclusion of covariates, but still does not allow for a quantification of methodological and clinical heterogeneity.

This modeling approach can be represented by the following three steps:

1. To transform the vertical and horizontal axes of the ROC plane in S and D, where

$$S_i = logit(sens_i) - logit(spec_i)$$
$$D_i = logit(sens_i) + logit(spec_i)$$

Equation 4-5

2. To estimate the slope and intercept of the line

$$D_i = a + bS_i$$

Equation 4-6

3. To reverse the transformation in order to find the corresponding sROC curve

S can be interpreted as a measure of the threshold effect (i.e. *S* decreases as the threshold increases). *S* is related to how often the test is positive. In fact, *P*(*T*+)

rises as the threshold goes down. *S* may be interpreted as the attitude of the test in preferring positive or negative results. Reitsma, Glas et al (2005) suggest the relation between *S* and the diagnostic rates:

- if TPR = 1− FPR, then S = 0
- if TPR < 1 − FPR, then S < 0
- if TPR > 1 − FPR, then S > 0

D is the logarithm of the DOR, that is a measure of how well the test discriminates the diseased from the healthy patients (i.e. depends especially on how distant the distributions of the healthy and diseased are).

The regression parameter *b* is a measure of the variation of the diagnostic performance:

1. if *b* equals 0, no variation of the diagnostic performance with the threshold. In this case the ROC curve is symmetric.
2. If *b* significantly different from 0, a significant variation of the diagnostic performance with the threshold is detected. Studies are likely to be based on different thresholds (i.e. variability due to observers). More analyses are needed.

If difference in variances between the sick and the well populations is high, b tends to be different to 0 and the summary ROC is distorted (Asymmetrical - (Littenberg and Moses 1993)). Littenberg and Moses (1993) also suggest the following rule:

- If $b \in (-0.5, +0.5)$, sROC is symmetrical
- If $b \notin (-0.5, +0.5)$, sROC is asymmetrical

The AUC can be calculated by the following equation:

$$AUC = \int_0^1 \frac{1}{1 + \frac{1}{e^{\frac{a}{1-b}} \left[\frac{(1-x)}{x}\right]^{\frac{1+b}{1-b}}}} dx$$

Equation 4-7

The bigger a, the closer to the left-upper corner the sROC curve is. It mainly represents the intercept of the transformed line and it measures the ability of the test to discriminate between healthy and diseased patients. The greater the distance between the average of the populations of the diseased and the healthy the greater is a.

The sROC curve can be plotted by back-transforming the straight line on the ROC plane by the formula (Egger, Smith et al. 2001).

$$sens_j = \int_0^1 \frac{1}{1 + \frac{1}{e^{\frac{a}{1-b}} \left[\frac{(1 - spec_j)}{spec_j}\right]^{\frac{1+b}{1-b}}}} dx$$

Equation 4-8

The differences between an sROC and a single ROC curve can help to understand better what an sROC represents. A ROC curve describes how TPR and FPR vary as the threshold varies all else being constant. An sROC describes how TPR and FPR vary as the threshold varies and all the rest not being constant (what varies is not stated). It follows that an sROC comes from a set of independent populations and a ROC curve comes from a single population.

Bayesian model for asymmetric ROC curves

A limitation of this approach is that the uncertainty in the predictor variable (i.e. the sum of the logit) is ignored. The implementation of the asymmetric modeling approach in a Bayesian framework solves this problem and allows for random effects to be implemented (Novielli, Cooper et al. 2010). The third row of Model 4-5 derives directly from equation Equation 4-6 and Equation 4-5.

$$TP_i \sim binomial(p_{1i}, D_i^+)$$

$$FP_i \sim binomial(p_{2i}, D_i^-)$$

$$logit(p_{1i}) = \mu_{1i}$$

$$logit(p_{2i}) = \mu_{2i}$$

$$\text{Fixed effect} \quad \begin{cases} \mu_{1i} = \frac{\mu_{2i}(1+b)+a}{1-b} \end{cases}$$

Prior distributions on μ_{2i}, a and b

Model 4-5 Asymmetric FE sROC model, without covariates.

Model 4-6 includes covariates using the same parameterization proposed in Model 4-5. The result is a series of sROC curves, each defined by a different level of the covariate, which are parallel on the logit space.

$$TP_i \sim binomial(p_{1i}, D_i^+)$$

$$FP_i \sim binomial(p_{2i}, D_i^-)$$

$$logit(p_{1i}) = \mu_{1i}$$

$$logit(p_{2i}) = \mu_{2i}$$

$$\text{Fixed effect} \quad \begin{cases} \mu_{1i} = \frac{\mu_{2i}(1+b)+a_i}{1-b} \end{cases}$$

$$a_i = \alpha + \beta^T X$$

Prior distributions on μ_{2i}, a and b

Model 4-6 Asymmetric FE sROC model, with covariates.

Bayesian random intercept model for asymmetric sROC curves

In this section, the random effect version of this model will be presented based on the same parameterisation of Model 4-5 (Hamza, Reitsma et al. 2008). This can be obtained by substituting Equation 4-6 with the following equation:

$$D_i = \alpha_i + \beta S_i \quad \text{with} \quad \alpha_i \sim N(A, \tau_\alpha^2)$$

Equation 4-9

The following parameterization can be considered to implement such a model in a Bayesian framework (i.e. WinBUGS software for Bayesian data analysis).

$$TP_i \sim binomial(p_{1i}, D_i^+)$$

$$FP_i \sim binomial(p_{2i}, D_i^-)$$

$$logit(p_{1i}) = \mu_{1i}$$

$$logit(p_{2i}) = \mu_{2i}$$

$$\text{Fixed effect} \quad \begin{cases} \mu_{1i} = \frac{\mu_{2i}(1+b)+a_i}{1-b} \end{cases}$$

$$a_i \sim N(a_0, \sigma_a^2)$$

Prior distributions on μ_{2i}, a_0 and b

Model 4-7 Asymmetric random intercept sROC model, without covariates.

This allows a quantification of heterogeneity via the estimation of the parameter σ_a^2. This variance parameter represents between-study variability and identifies a

family of sROC that are parallel on the logit scale. The random intercept model can also be adapted to include covariates (see Model 4-8).

$$TP_i \sim binomial(p_{1i}, D_i^+)$$

$$FP_i \sim binomial(p_{2i}, D_i^-)$$

$$logit(p_{1i}) = \mu_{1i}$$

$$logit(p_{2i}) = \mu_{2i}$$

$$\text{Fixed effect} \quad \begin{cases} \mu_{1i} = \frac{\mu_{2i}(1+b)+a_i}{1-b} \end{cases}$$

$$a_i = a_{0i} + \beta^T X$$

$$a_{0i} \sim N(a_0, \sigma_a^2)$$

Prior distributions on μ_{2i}, a_0 and b

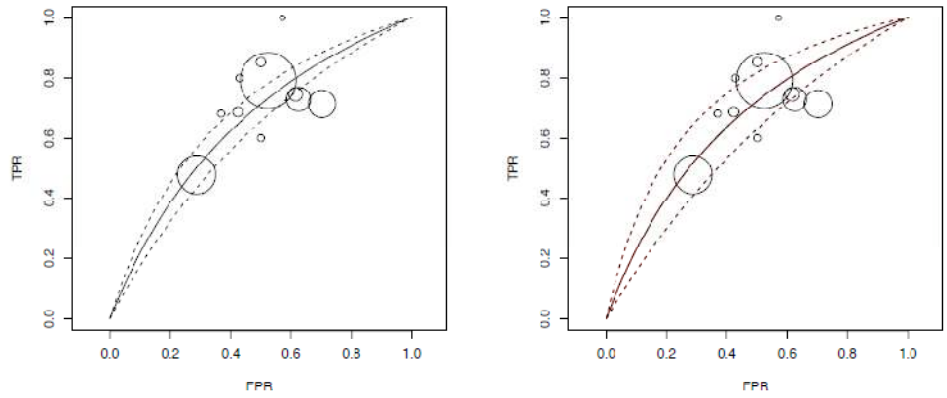
Model 4-8 Asymmetric RI sROC model, with covariates.

Recently, a Bayesian random effect version of this model for meta-analysis and meta-regression has been modelled by putting a bivariate normal distribution directly on pairs of (D_i, S_i) (Verde 2010); see Equation 4-5 for a definition of D_i and S_i. However, D_i and S_i are two different functions of the same pair of parameters, which means that part of their correlation will be due to the fact that the same quantities are used to calculate D_i and S_i (i.e. the dependence between D_i and S_i is not due completely to variability in the threshold, this is also discussed in the section below "GERD example – summary ROC estimation")

and this may be reflected in the variance and covariance matrix specified for this structure.

GERD example - summary ROC estimation

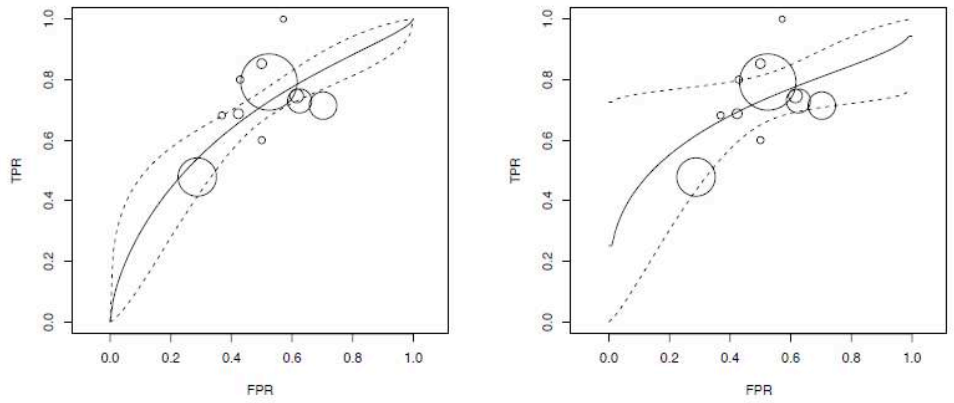
The estimated sROC curves in Figure 4-2 are symmetric sROC curves based on the estimation of the DOR for GERD dataset. Not all the points on the sROC represent the accuracy of PPI. In this case both a Q point can be extrapolated (i.e. point on the curve that crosses the line sensitivity=specificity) or an estimate of the area under the curve (AUC 0.65 95%CrI 0.61 to 0.69). This model assumes the same value of the DOR for every study of 2.53 (95%CrI 1.91 to 3.27). A little heterogeneity remains unexplained and needs to be explored (σ^2 0.4898 95%CrI 0.04063 to 1.218). This model allows the natural tension between rates to be considered although it does not allow for diagnostic heterogeneity to be fully considered in terms of asymmetries in the sROC curve.



(a) Fixed Effect model (b) Random Effect model

Figure 4-2 Symmetric sROC estimates: (a) fixed effect, (b) random effect.

Asymmetric curves overcome this characteristic of symmetric curves. In fact it estimates a threshold parameter (*b* -0.109, 95%CrI -0.3708 to 0.1383) and accuracy parameter (*a* 1.005, 95%CrI 0.6751 to 1.35). In this case the accuracy parameter does not tell us anything if it is not used for model or group comparison. There is no evidence of diagnostic heterogeneity (i.e. threshold parameter not significantly different from zero). Credible intervals are more precise where observations are recorded.



(a) Fixed Effect model (b) Random Intercept model

Figure 4-3 Asymmetric sROC estimates: (a) fixed effect, (b) random intercept.

The random intercept model has been shown to be equivalent to the bivariate approach (see section 4.7.4). However, in a fully Bayesian framework, there is the suspect that this approach may have some shortcomings with small datasets like GERD. This is related to the values of *b*. In fact, this is the threshold parameter (also shape parameter), and it affects both the symmetry of the curve and its slope. The relation between *b* and the symmetry of the curve has already been shown above in this section; this is how the slope of the ROC curve varies as *b* varies in equation 5.9:

- $b < -1$: negative slope
- $b = -1$: null slope, horizontal curve
- $b > -1$: positive slope.

In small datasets such as the GERD dataset, failure in the convergence of *b* has been observed when vague prior distributions are used. This can be overcome using prior distributions on *b* restricted to its plausible values. Since it is known than there is negative tension between sensitivity and specificity due to the threshold effect and the type of test (i.e. values of the test higher for diseased than for healthy), then it has been used a uniform prior distribution between -1 and 100. Also negative correlation in the chains has been measured between the values of *b* and *a* (i.e. -0.75). This may be due to the fact that *S* and *D* are affected by a sort of regression dilution bias (i.e. they are not independent to each other since they both are transformations of the same two parameters sensitivity and specificity).

It should also be noted that the sROC curve in Figure 4-3 shows unusual values at the boundaries of the curve; for example, uncertainty is very large when the curve reaches the point (0,0) and (1,1) on the ROC plane. This may be due to the low number of studies and to the random intercept at the logit level, which both result in larger uncertainty around the sROC curve especially where there is not any observation. For this curve, the sROC curve has still sensitivity=0 where specificity=1 (and sensitivity=1 where specificity=0) in theory, but the curve has not been plotted all the way to the extremities (i.e. it has been plotted the interval between the observed points). Some authors suggest plotting sROC curves only in the range of the data (Leeflang, Deeks et al. 2008).

For these models also non informative prior distributions were used. Logit rates and regression parameters were given normal prior distributions with mean equal to zero and very low precision. For the asymmetric model, either fixed effect or random intercept version, the estimated proportion of false positives (*fp*) was given a uniform prior between 0 and 1. The standard deviation (heterogeneity) parameter for the symmetric random effect model was given a uniform distribution between 0 and 10. The precision (heterogeneity) parameter for the asymmetric model random intercept was given a gamma prior with parameters both equal to 0.0001.

4.5.3 Bivariate estimates of sensitivity and specificity

The modeling approaches described in this section are based either on the estimation of sensitivities and specificities (bivariate approach) or on the estimation of an accuracy and a threshold parameter (Hierarchical summary ROC - HsROC). These relax many of the assumptions that characterised the other simpler models and account explicitly for diagnostic heterogeneity. Although different software have been suggested to estimate this model in either classical or Bayesian framework (Paul, Riebler et al. 2009), in this thesis WinBUGS software for Bayesian modeling (Lunn, Thomas et al. 2000) has been used also to estimate the bivariate models, as discussed in Chapter 3.

Bivariate model

The most complex model proposed to date to meta-analyse individual estimates of sensitivity and specificity from multiple studies includes bivariate random effects (Van Houwelingen, Zwinderman et al. 1993) to allow for the correlation between sensitivity and specificity (Reitsma, Glas et al. 2005). Following the notation used previously (Harbord, Deeks et al. 2007), μ_{Ai} is defined as the logit(sensitivity) in study $i = 1. \ldots k$, and μ_{Bi} as the logit(specificity). The model is then written as

$$TP_i \sim binomial(\pi_{Ai}, (TP_i + FN_i))$$

$$TN_i \sim binomial(\pi_{Bi}, (FP_i + TN_i))$$

$$\mu_{Ai} = logit(\pi_{Ai})$$

$$\mu_{Bi} = logit(\pi_{Bi})$$

$$\begin{pmatrix} \mu_{Ai} \\ \mu_{Bi} \end{pmatrix} \sim MVN\left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \Sigma_{\mathrm{AB}}\right) \quad \text{with} \quad \Sigma_{AB} = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix}$$

Model 4-9 Bayesian specification of Bivariate RE sROC model, without covariates.

where, TP_i, FP_i, TN_i and FN_i are the number of individuals who are true positive, false positive, true negative and false negative respectively in the i^{th} study. μ_A and

μ_B are the mean logit-transformed sensitivity and logit-transformed specificity respectively and Σ_{AB} is the associated variance-covariance matrix, with components σ_A^2 , σ_B^2 and σ_{AB} for the between-study variance in sensitivity, specificity and covariance, respectively. Using these estimates from the bivariate model, sROC curves and confidence and prediction regions (within which the results of a future study may be expected) can be constructed around the pooled sensitivity and specificity and plotted in ROC space (Reitsma, Glas et al. 2005). Full technical details on the construction of such intervals is available elsewhere (Reitsma, Glas et al. 2005; Harbord, Deeks et al. 2007).

Briefly, the following formulae define the region:

$$\mu_A = \hat{\mu}_A + \hat{s}_A \, c \, cos(t)$$
$$\mu_B = \hat{\mu}_B + \hat{s}_B \, c \, cos(t + \arccos(\hat{r}))$$

Equation 4-10

where $\hat{\mu}_A$ and $\hat{\mu}_B$ are the posterior estimates of μ_A and μ_B as given in equation 1, $\hat{s}_A$and $\hat{s}_B$ are the standard errors of the posterior distributions of μ_A and μ_B, and $\hat{r}$ is an estimate of the correlation between $\hat{\mu}_A$ and $\hat{\mu}_B$. The latter is estimated by calculating the correlation between the sampled values of $\hat{\mu}_A$ and those for $\hat{\mu}_B$ across all iterations of the MCMC sampler. c is the boundary constant and is

calculated as $c = \sqrt{\chi^2_{2,.\alpha}}$, where n is the number of studies and $1 - \alpha$ is the level of credibility of the region, and $\chi^2_{2,.\alpha}$ refers to the Chi-squared probability density function with 2 degrees of freedom. t takes values between 0 and 2π and the higher the number of data points calculated across this range, the higher the definition of the confidence region. The inverse logit of the pairs of values of μ_A and μ_B can then be plotted on ROC space. The choice of c is not unique; Alexandersson (Alexandersson 2004) reviews some of the possible alternatives. Further, it is possible to plot the highest posterior density regions for the joint region for (logit) sensitivity and specificity. For example, this can be achieved by using the package *hdrcde* available in R (R Foundation for Statistical Computing 2005) to plot non-parametric regions using the MCMC samples for the relevant posterior distributions directly.

Predictive regions may be calculated in a similar manner to those above by substituting $\hat{s}_A + \hat{\sigma}_A$ and $\hat{s}_B + \hat{\sigma}_B$ for $\hat{s}_B$ and $\hat{s}_A$ respectively, where $\hat{\sigma}_A$ and $\hat{\sigma}_B$ are the random effect variances estimated within the bivariate model, and the correlation coefficient $\hat{r}$ is substituted by

$$\hat{r}^{pred} = (\hat{s}_{AB} + \hat{\sigma}_{AB}) \Big/ [(\hat{s}_A + \hat{\sigma}_A)((\hat{s}_B + \hat{\sigma}_B)]\cdot$$

Equation 4-11

Covariates can be added to the bivariate model to explore residual heterogeneity (Hamza, van Houwelingen et al. 2009). A version of this model to be implemented in a Bayesian framework is described in Model 4-10.

$$TP_i \sim binomial(\pi_{Ai}, (TP_i + FN_i))$$

$$TN_i \sim binomial(\pi_{Bi}, (FP_i + TN_i))$$

$$\mu_{Ai} = logit(\pi_{Ai})$$

$$\mu_{Bi} = logit(\pi_{Bi})$$

$$\begin{pmatrix} \mu_{Ai} \\ \mu_{Bi} \end{pmatrix} = MVN\left[M_i = \begin{pmatrix} \mu_{Ai} \\ \mu_{Bi} \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix} \right]$$

$$M_i = \begin{pmatrix} \alpha_A \\ \alpha_B \end{pmatrix} + \begin{pmatrix} \sum_{j=1}^{p} \beta_{jA} x_{ji} \\ \sum_{j=1}^{p} \beta_{jB} x_{ji} \end{pmatrix}$$

Prior distributions on α and β_j's.

Model 4-10 Bayesian specification of Bivariate RE sROC model, with covariates.

where $\alpha = \begin{pmatrix} \alpha_A \\ \alpha_B \end{pmatrix}$ is the vector of the intercepts; $\beta_j = \begin{pmatrix} \beta_{jA} \\ \beta_{jB} \end{pmatrix}$ is the jth couple of regression parameters; $X_j = x_{ji}$ is the j^{th} covariate replicated for both rates (assume symmetric regression, same covariates on both parameters), where i indicates the study in the meta-analysis.

The bivariate model has also been adapted to explicitly adjust for dependence of sensitivity and specificity with the prevalence of disease (Chu, Nie et al. 2009) by mean of a trivariate random effect model. However, this model is more complex (no publications have been found that used this model) and its relationships to the other model approaches (in particular tot he HsROC model) have not been investigated.

Hierarchical Summary ROC analysis (HsROC)

The Hierarchical summary ROC (HsROC) modeling approach is based on a Bayesian hierarchical model with three levels (Rutter and Gatsonis 1995; Rutter and Gatsonis 2001).

As for the other models presented before, within study variability is described via binomial distributions. Then, logit(sensitivity) and logit(specificity) are expressed as functions of an accuracy parameter (α_i), a threshold parameter (θ_i), and a shape parameter (β) which is assumed constant. $X_{i,\cdot}$ indicates the presence ($X_{i,+}$) or the absence ($X_{i,-}$) of disease. This may take arbitrary values (i.e. 1 and 0). The authors suggest 0.5 and -0.5.

α_i and θ_i are assumed conditionally independent which is an underlying assumption of the ROC analysis (i.e. positivity threshold and accuracy are independent, and together impose tension between accuracy rates).

Covariates are easily included in the model through regression equations. γ (λ) models the systematic differences in positivity criteria (accuracy) across studies due to the covariate Z_i. This model is quite sensitive to prior distributions and the authors suggest a set of prior distributions which allow the parameters to be samples among plausible values and were set to be locally vague.

$$TP_i \sim binomial(p_{i1}, D_i^+)$$

$$TN_i \sim binomial(1 - p_{i2}, D_i^-)$$

$$logit(p_{i1}) = \left(\theta_i + \alpha_i X_{i,+}\right) \exp\left(-\beta X_{i,+}\right) = \mu_{1i}$$

$$logit(p_{i2}) = \left(\theta_i + \alpha_i X_{i,-}\right) \exp\left(-\beta X_{i,-}\right) = \mu_{2i}$$

$$\theta_i | \Theta, \gamma, Z_i, \sigma_\theta^2 \sim N(\Theta + \gamma Z_i, \sigma_\theta^2)$$

$$\alpha_i | \Lambda, \lambda, Z_i, \sigma_\alpha^2 \sim N(\Lambda + \lambda Z_i, \sigma_\alpha^2)$$

$\Theta \sim UNIF(\mu_{\theta 1}, \mu_{\theta 2})$; $\gamma \sim UNIF(\mu_{\gamma 1}, \mu_{\gamma 2})$; $\sigma_{2\theta} \sim \Gamma^{-1}(\psi_{\theta 1}, \psi_{\theta 2})$

$\Lambda \sim UNIF(\mu_{\alpha 1}, \mu_{\alpha 2})$; $\lambda \sim UNIF(\mu_{\lambda 1}, \mu_{\lambda 2})$; $\sigma_{2\alpha} \sim \Gamma^{-1}(\psi_{\alpha 1}, \psi_{\alpha 2})$

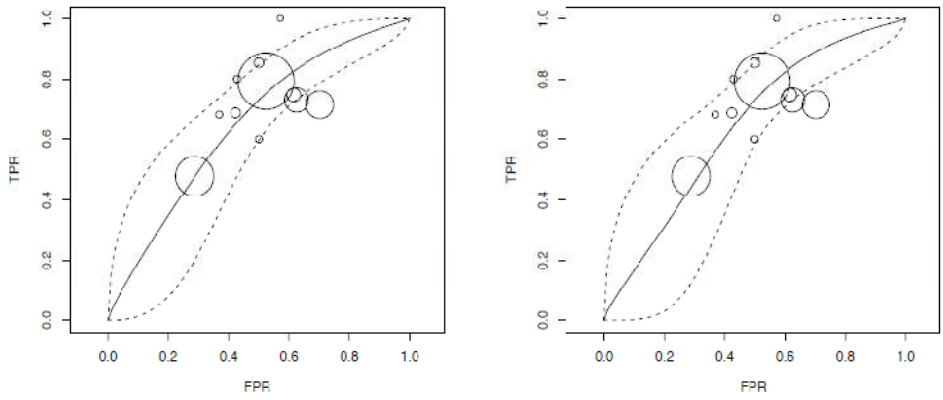
$\beta \sim UNIF(\mu_{\beta 1}, \mu_{\beta 2})$

Model 4-11 Hierarichical summmary ROC (HsROC) model

GERD example - correlated estimates.

The sROC curves produced by applying the bivariate approach and the HsROC approach to GERD data (see Figure 4-4) are not different from each other. The

equivalence of these approaches will be presented in section 4.6. Small differences may be due to different prior distributions and may be more evident for small dataset such as GERD.



(a) Bivariate (b) Hierarchical SROC

Figure 4-4 Correlated estimates of diagnostic rates : (a) bivariate model, (b) HsROC model.

The bivariate approach allows the estimation of a correlation parameter between sensitivities and specificities which quantified the diagnostic heterogeneity (ρ -0.34, 95%CrI -0.81 to 0.34). The fact that it is not significantly different from zero confirms what the results on threshold parameter from the asymmetric fixed effect model that there is weak evidence from the data for diagnostic heterogeneity. Standard errors (squared root of heterogeneity parameters) give

evidence of unexplained clinical and methodological heterogeneity that affects the data: (σ_1 0.74, 95%CrI 0.4189 to 1.293) and (σ_1 0.63, 95%CrI 0.38 to 1.04).

HsROC approach does not estimate directly the correlation factor because it already incorporates a threshold parameter. Similarly to the asymmetric model, it estimates a threshold parameter and an accuracy parameter but it is not based on the DOR. Differently to the asymmetric random intercept model, it includes variability in both the accuracy and the threshold parameter through random effects. In this case, the threshold parameter is close to zero although still significantly positive (Θ 0.53, 95%CrI 0.05 to 1.07) and fair amount of unexplained heterogeneity is indicated by its random effects parameters (σ_θ 0.69, 95%CrI 0.46 to 1.06). This tells that this threshold parameter may be either i. more precise of other indicators of diagnostic heterogeneity, or ii. not correctly interpreted. Like in the asymmetric models, the accuracy parameter (Λ 0.91, 95%CrI 0.24 to 1.638) is useful for model comparison. However, estimated between studies standard error (σ_α 0.75, 95%CrI 0.47 to 1.22) indicated some clinical-methodological heterogeneity to be explored (i.e. through variables inclusion). This model also estimates a scale parameter (β -0.26, 95%CrI -1.14 to 0.57).

Prior distributions for the logit rates of the bivariate model where normal with mean equal to 0 and very low precision. The matrix of variances and covariances

(heterogeneity parameters) was given a Wishart prior distribution with parameters $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

4.6 Comparison and interpretation of summary ROC curves

4.6.1 Approaches to construction of sROC curves

In this section a summary for the techniques for the construction of sROC curves using the output of the different models described above is given (Novielli, Cooper et al. 2010). Note, it is not possible to plot sROC curves where independence between sensitivity and specificity is assumed. In this situation, the pooled sensitivity and specificity can be plotted on the ROC plane with their associated uncertainty.

A symmetric sROC curve based on a pooled DOR is given by the equation below (Egger, Smith et al. 2001)

$$sensitivity = \frac{1}{1 + \frac{1}{DOR * \left(\frac{1 - specificity}{specificity}\right)}}$$

Equation 4-12

Graphical representations of sROC curves based on the output of the bivariate model have been presented previously (Reitsma, Glas et al. 2005; Arends, Hamza et al. 2008; Hamza, Arends et al. 2009)**.** One approach is to estimate logit(sensitivities) for different values of logit(specificities) (Reitsma, Glas et al. 2005)

$$\boldsymbol{\mu_A \mid \mu_B = \hat{\mu}_A + \frac{\sigma_{AB}}{\sigma_B^2}(\mu_B - \hat{\mu}_B)}$$

Equation 4-13

Where $\mu_A \mid \mu_B$ is the logit-sensitivity calculated as a function of a given set of values of logit-specificity, $\hat{\mu}_A$ and $\hat{\mu}_B$ are the *estimated* logit(sensitivities) and logit(specificities) respectively from the bivariate model, and σ_B^2, σ_{AB} are estimates of the variances and covariances matrix described in Model 4-9.

Alternatively, using the same notation conventions as above, one can also calculate specificities for different values of sensitivities (Arends, Hamza et al. 2008)

$$\mu_B \mid \mu_A = \hat{\mu}_B + \frac{\sigma_{AB}}{\sigma_A^2}(\mu_A - \hat{\mu}_A)$$

Equation 4-14

It is important to remember that sensitivity and specificity are correlated because of their common threshold within each study, and hence defining one as the outcome and the other as the predictor (measured with error) is inappropriate. Hence, the below equations can be considered more appropriate (and these will fall within the (extreme) bounds defined by Equation 4-13 and Equation 4-14).

Considering the regression modeling approach for asymmetric sROC curves described in section 4.5.2, another sROC curve can be expressed using parameter estimates from Model 4-9 as follows (Arends, Hamza et al. 2008):

$$\mu_A \mid \mu_B = \hat{\mu}_A + \frac{\sigma_A^2 + \sigma_{AB}}{\sigma_B^2 + \sigma_{AB}}(\mu_B - \hat{\mu}_B)$$

Equation 4-15

This curve can be viewed as a sort of compromise between the regression of μ_A on μ_B and the regression of μ_B on μ_A, described above, since its slope lies in between these slopes. Rutter and Gatsonis (Rutter and Gatsonis 2001; Arends, Hamza et al. 2008) suggest another approach to the calculation of the slope of a sROC curve:

$$\mu_A \mid \mu_B = \hat{\mu}_A + \frac{\sigma_A}{\sigma_B}(\mu_B - \hat{\mu}_B)$$

Equation 4-16

This sROC curve can also be viewed as a compromise between

Equation 4-13 and Equation 4-14, since its slope is the geometric mean of the estimates from these approaches (Arends, Hamza et al. 2008).

This list is not exhaustive, and further possibilities have been described elsewhere (Arends, Hamza et al. 2008). There is no overriding reason to select one of these alternative curves over the others as they all represent the accuracy of the test for different characterizations of the bivariate model (Arends, Hamza et al. 2008).

4.6.2 Interpretation of sROC curves

In Chapter 3, ROC curves were described as means to represent diagnostic accuracy rates of a test with varying threshold form a single study. For meta-analysis of diagnostic data, sROC curves are considered attempts to represent diagnostic rates of a test from multiple studies accounting for correlation between them. It is tempting to give sROC curves the same interpretation that it is given to ROC curves; however sROC curves are not pooled ROC curves. Although sROC curves are meant to be similar to ROC curves, they are different tools designed for different problems. The modeling approaches that can be used to produce sROC curves when dichotomous test data are available were presented in section 4.5 and a summary of the methods for plotting sROC curves has been presented in section 4.6.1. In this section the differences between sROC and ROC curves will be discussed in relation to the different approaches to plot sROC curves.

When diagnostic rates are assumed independent it is not possible to plot sROC curves. This does not necessarily mean that the same threshold has been used in the studies included in the meta-analysis, nor that individual studies had not reported ROC curves. This is the case when dichotomised data are considered for meta-analysis and correlation between study rates is assumed to be zero.

Two groups of approaches have been presented to account for correlation between rates. These were introduced in the first part of section 4.5 and described as "summary ROC curves" and "Bivariate estimates of diagnostic rates". It must be noted that correlation between rates may not be entirely due to variability in

diagnostic threshold, for example factors that may cause heterogeneity (i.e. differences between populations) may be a source of negative correlation between sensitivities and specificities included in the meta-analysis.

The methods belonging to the first group are based on the meta-analysis of the DOR. In the simplest of these methods an ROC curve can be obtained by equation 4-3. This equation represents the sensitivity as a function of specificity (the opposite is also possible) and the DOR, where specificity is fixed and the DOR is the output of the meta-analysis. In fact, these methods for meta-analyses do not pool directly ROC curves but use the pooled DOR to represent graphically a sROC curve. sROC curves are then plotted using the output of these models by using ad-hoc transformations of the equivalence sensitivity=sensitivity (i.e. in order to include in the equation specificity and the pooled DOR). Moreover, as already mentioned above, the correlation between sensitivity and specificity may be due to causes other than variability in threshold. Finally, since the pooled DOR can be considered as an average between study specific DOR, the sROC curve has been interpreted as an average between ROC curves (Deeks 2001). However, this is true only if study specific DOR are consistent with each other (Deeks 2001). Unlike ROC curves, some authors believe that sROC curves from these methods have no clinical relevance(Deeks 2001).

More complex approaches build a regression line between study specific S (measure of accuracy, independent variable) and D (DOR, dependent variable). These are defined in section 4.5.2. The regression line is then back-transformed in the ROC space and a sROC curve is obtained. As it is also acknowledged by Littenber and Moses (1993), the output of these methods can be used to make considerations on the DOR (i.e. if slope is zero, then DOR is constant across studies; how the DOR varies with certain covariates). As these methods are based on regression approaches, the line that is fitted on the logistic space is the line that best fits the pairs the sensitivities and specificities obtained form a number of individual studies, where parameters can be estimated with ad-hoc techniques (i.e. least squares estimates). As before, these sROC curves are not pooled ROC curves, although for a decade they have been considered as the best means available to represent diagnostic accuracy form multiple studies. They are attempts to summarise diagnostic data accounting for correlation between sensitivities and specificities, however, such correlation may have multiple sources. The main drawback of these curves is that, in cases like the one represented in Figure 4-3(b) the sROC curve and its boundaries may not evidently join the points (0,0) and (1,1) of the ROC plane as it is expected from ROC curves. For this reason, the authors suggested to restrict the interval of a priori specificities on the interval of plausible values (i.e. observed form the data)(Moses, Shapiro et al. 1993). It must be also acknowledged that the problem

in Figure 4-3(b) may be simply due to computational issues; more research is needed.

Recently, as mentioned in section 4.5.3, bivariate estimates of sensitivity and specificity is considered the ideal approach to pool diagnostic data from sROC curves. It has been shown in section 4.5.3 that sROC curves can be also obtained from these methods, although the natural graphical output of these methods is a pooled estimate of sensitivity and specificity and a confidence/credible region. Therefore, also in this case, sROC curves cannot be interpreted as ROC curves and have not clinical relevance.

In conclusion, sROC curves are very different from ROC curves when dichotomised data are available. They can be interpreted as mean sensitivities for given values of specificities, and different approaches can be used to obtain them. The methods used to plot the sROC curves presented in section 4.5 and 4.6.1 are the most common approaches used in practice. The bivariate models allow the estimation of pooled rates and variability can be represented by regions, which are clinically more relevant and thus avoiding misleading interpretations of sROC curves as ROC curves. Methods for pooling ROC curves are available when studies report data from multiple thresholds (Hamza, Arends et al. 2009; Putter, Fiocco et al. 2009); these methods may be able to produce sROC curves that can be considered pooled ROC curves, therefore with more similar properties and interpretation.

4.7 Relationships between models

4.7.1 Introduction to the section

Harbord et al recently published an empirical comparison of methods for the meta-analysis of diagnostic data, including methods for pooling likelihood ratios and predictive values; they applied the methods to a number of published statistical datasets and concluded that bivariate and hierarchical summary ROC methods where ideal solutions for a thorough exploration of the diagnostic heterogeneity and to obtain robust results (Harbord, Whiting et al. 2008). In this section, the different meta-analysis of diagnostic test accuracy data approaches described above are theoretically revisited (section 4.5) and the mathematical relationships between them are shown by expressing all models in terms of logit(sensitivity) and logit(specificity), as first presented in our publications (Novielli, Cooper et al. 2010). This is an attempt to show that all the proposed methods can be viewed as either simplifications or special cases of the bivariate model (section 4.5.3). Considering the models in this manner provides insight into the assessment of goodness of fit of, and selection between, competing models of varying complexity. These aspects are explored in Chapter 5 where the different meta-analysis models are applied to the complete DVT dataset (introduced in section 1.4). Since this dataset is unusually large for a meta-analysis of diagnostic test data, the meta-analysis models have also been applied to a subset of the DVT dataset for comparison.

Throughout, it is assumed that the data available from each study to be meta-analysed is a cross-tabulation of test result (positive or negative) and disease status (diseased or non-diseased) from which sensitivity and specificity can be derived (Egger, Smith et al. 2001) (as represented in Table 4-1). This assumption is also at the basis of the modeling approaches presented in section 4.5.

4.7.2 Independent estimates of sensitivity and specificity

Fixed effect

As seen in section 4.5.1, the simplest meta-analysis model for diagnostic test accuracy data assumes there is no heterogeneity between study estimates, and that sensitivity and specificity are independent of one another (Egger, Smith et al. 2001). This can be fitted by setting all terms of the variance-covariance matrix (σ_A^2 , σ_B^2 and σ_{AB}) to 0 in the bivariate model given in Model 4-9. This in turn implies that all the study specific estimates are identical (i.e. $\mu_{A1} = \ldots . = \mu_{Ak} = \theta_1$ & . $\mu_{B1} = \ldots . = \mu_{Bk} = \theta_2$) returning us to Model 4-1 in section 4.5.1. This model is only appropriate if all studies use the same test threshold and there is no heterogeneity in study results (Egger, Smith et al. 2001).

Random effects

Independent random effects estimates of sensitivity and specificity can be obtained if the covariance between them (σ_{AB}) in Model 4-9 is set to 0. As with the fixed effect model above, this model assumes no correlation between sensitivity and specificity and hence is only appropriate if all studies use the same test threshold, although heterogeneity between sensitivity and specificity estimates is incorporated.

4.7.3 Combining diagnostic odds ratios

These models assume the diagnostic odds ratio (DOR), as defined below, is constant across studies (fixed effect approach) or exchangeable between studies (random effect approach) (Egger, Smith et al. 2001), see section 4.5.2. When the DORs are assumed exchangeable between studies, the study specific DORs are assumed to be drawn from the same (random effect) distribution for every study.

$$\boldsymbol{DOR} = \frac{\left(\frac{\boldsymbol{sensitivity}}{\boldsymbol{1-sensitivity}}\right)}{\left(\frac{\boldsymbol{1-specificity}}{\boldsymbol{specificity}}\right)}$$

Equation 4-17

A constant DOR results in a combination of sensitivities and specificities which trace out an sROC (Leeflang, Deeks et al. 2008) which is symmetric around the line sensitivity = specificity.

Fixed effect

The DOR model may also be specified as a special case of the bivariate model. For the fixed effect model, μ_{Bi} is constrained to be equal to $\mu_{Ai} + d$, where d = underlying mean DOR, and all terms in the variance-covariance matrix, Σ_{AB}, are set to zero.

Random effects

The random effect DOR model can be derived from the bivariate model (Model 4-9) by constraining μ_{Bi} to equal $\mu_{Ai} + d_i$, where $d_i \sim N(D, \sigma_D^2)$ and D = underlying mean DOR and σ_D^2 its variance. Thus,

$$\begin{pmatrix} \mu_{Ai} \\ \mu_{Ai} + d_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_A \\ \mu_A + D \end{pmatrix}, \Sigma_{\mathrm{AB}}\right) \quad \text{with} \quad \Sigma_{AB} = \begin{pmatrix} 0 & 0 \\ 0 & \sigma_d^2 \end{pmatrix}$$

Equation 4-18

Note that the logit(sensitivity) from each study (the μ_{Ai}'*s*) are not assumed to be related to each other in this model for either the fixed or random effects formulations.

4.7.4 Asymmetric sROC models

Allowing the DOR to change with diagnostic threshold implies that the sROC can be asymmetrical. The original model described to do this was proposed by

Littenberg and Moses (Littenberg and Moses 1993), and is based on a simple regression of the difference (D_i) of logit(sensitivity) and logit(1-specificity) on the sum (S_i) of the two for each study (see section 4.5.2 for details).

Fixed effect

Using the notation defined previously, noting

$$D_i = \mu_{Ai} + \mu_{Bi}$$

$$S_i = \mu_{Ai} - \mu_{Bi},$$

the regression in Equation 4-6 is equivalent to

$$\mu_{Ai} = \frac{\alpha - \mu_{Bi}(1+\beta)}{(1-\beta)}$$

Equation 4-19

Hence, this can be viewed as a special case of the bivariate model in Model 4-9 in which Σ_{AB} is set to zero and μ_{Ai} and μ_{Bi} are dependent on each other via the relationship outlined in Equation 4-19.

Random intercept

While the original approach of Littenberg and Moses was a fixed effect model, its extension to include a random effect term on the intercept has recently been considered by others (Arends, Hamza et al. 2008), viz

$$D_i = \alpha_i + \beta S_i \quad \text{with} \quad \alpha_i \sim N(A, \tau_\alpha^2)$$

Equation 4-20

Re-expressing Equation 4-20 to obtain a model in terms of logit(sensitivity) and logit(specificity) results in

$$\mu_{Ai} \sim N\left(\frac{\alpha_i - \mu_{Bi}(1+\beta)}{(1-\beta)}, \frac{\tau_\alpha^2}{(1-\beta)^2}\right)$$

Equation 4-21

4.7.5 Full bivariate random effect

For completeness, it can be noted that if a full bivariate normal structure is placed on the data (as in section 4.5.3), then it is possible to derive estimates for the regression parameters for Equation 4-6 from Model 4-9 as described in detail elsewhere (Arends, Hamza et al. 2008) and given below using our formulation modeling logit(sensitivity) and logit(specificity).

$$\beta\ (slope) = \frac{\sigma_A^2 - \sigma_B^2}{\sigma_B^2 + \sigma_A^2 - 2\sigma_{AB}}$$

$$\alpha\ (intercept) = \ (\mu_B + \mu_A) - \beta(\mu_A - \mu_B)$$

$$\sigma_{D|S}^2(residual\ variance) = \left(\sigma_B^2 + \sigma_A^2 + 2\sigma_{AB}\right) - \frac{(\sigma_A^2 - \sigma_B^2)^2}{\sigma_B^2 + \sigma_A^2 - 2\sigma_{AB}}$$

Equation 4-22

4.7.6 Hierarchical sROC (HsROC) model

In a similar vein to the model considered in 4.7.4 above, Rutter and Gatsonis (2001) specify a hierarchical sROC model incorporating two random effects to accommodate between-study heterogeneity. They define this in terms of π_{ij}, where π_{i1} is the sensitivity and π_{i0} is 1 – specificity in study *i*:

$$logit(\pi_{ij}) = (\theta_i + \alpha_i X_{ij}) \exp(-\beta X_{ij})$$

$$\theta_i \sim N(\Theta, \sigma_\theta^{\ 2})$$
$$\alpha_i \sim N(\Lambda, \sigma_\alpha^{\ 2})$$

Equation 4-23

where X_{ij} denotes the true disease status for a patient in study *i* with disease status *j*, θ_i is the 'positivity criteria' parameter which measures the trade-off between sensitivity and specificity in the i^{th} study (assumed to be sampled from a normal distribution with mean Θ and variance $\sigma_\theta^{\ 2}$) with sensitivity and 1-specificity

increasing with the value of the parameter. α_i is the 'accuracy parameter' in the *i*th study (again, assumed to be sampled from a normal distribution with mean Λ and variance σ_α^2) where a higher value indicates increased accuracy of the test. β is the 'scale parameter' or 'shape parameter' which models possible asymmetry in the ROC curve (in a similar way to the models described in section 4.7.4) .

Harbord et al (Harbord, Deeks et al. 2007) have previously shown that this model is equivalent to the bivariate model (Model 4-9) with:

$$\mu_A = b^{-1}\left(\Theta + \frac{1}{2}\Lambda\right) \qquad \mu_B = -b\left(\Theta - \frac{1}{2}\Lambda\right)$$

$$\sigma_A^2 = b^{-2}\left(\sigma_\theta^2 + \frac{1}{4}\sigma_\alpha^2\right) \qquad \sigma_B^2 = b^2\left(\sigma_\theta^2 + \frac{1}{4}\sigma_\alpha^2\right) \qquad \sigma_{AB} = -\left(\sigma_\theta^2 - \frac{1}{4}\sigma_\alpha^2\right)$$

where b=exp(β/2).

4.8 Economic decision modeling and cost-effectiveness analysis

4.8.1 Introduction

Economic perspective

Economy originates from the scarcity of resources that need to be managed for the best satisfaction of community needs, and money is a tool at the service of economy to quantify and compare the values of resources. An economic evaluation does not account merely for the values of the resources used but also accounts for the implications derived by every action or event (NICE 2010).

Formal decisions are commonly made using decision models. The simplest analysis aims to minimise costs (i.e. cost-minimisation), more sophisticated techniques also consider measure of quality of life to quantify the benefit related to each choice (i.e. cost-utility analysis).

Decision making organizations in UK for Health

NHS (National Health Service) and NICE (National Institute for Clinical Excellence), the major decisional organizations in UK for Health, use cost-effectiveness analyses (CEA) as *modus operandi* for the evaluation of new technologies. In these evaluations, often decision analytic models are developed to

assess the cost-effectiveness of new technologies. Such models integrate information on the effectiveness of the new technologies with information on natural history of the disease in question, adverse events, quality of life, costs and resource use, and the results are usually expressed as cost-effectiveness acceptability curves (see section 4.8.4). (http://www.nice.org.uk/aboutnice/whatwedo/niceandthenhs/nice_and_the_nhs.jsj).

4.8.2 Cost-effectiveness analysis

Measuring health outcomes

CEA can be assessed alongside trials or via decision analytic models. Decision models provide a framework to combine data on effectiveness, resource use, costs, natural history and health outcomes (usually expressed as Quality Adjusted Life Years (QALYs)).

QALY is a measure of the quality of life as a consequence of a certain health condition, weighted by the time lived in that condition. The quality of life can be elicited using different techniques. For example, EuroQol (EQ-5D) is one of the most common set of questionnaires that are used to elicit the quality of life (QoL). The QoL is assumed to be zero in case of death, and one in case of full health. For

certain gruesome conditions, the QoL can be assumed to be negative (i.e. from EQ-5D classification - health condition 33332; description: Confined to bed; unable to wash or dress self; unable to perform usual; QoL: -0.429 (Phillips and Thompson 1998))

For example, a certain condition A can lead to a QoL of 0.8 for 3 years, therefore involving 2.4 QALY for an individual experiencing A. Let suppose that a drug is developed to reduce the time in condition A to 2 years, and that the same drug also reliefs condition A to a mild state (i.e. QoL 0.9). In this case, the patient would experience 2.8 QALY (assuming the third year is lived at full health). Thus, the use of the drug leads to a gain of 0.4 QALY for each patient. In a population of 10,000 patients, with condition A and assuming the same drug has the same effect on all patients, there will be 400 QALY gained as a consequence of the drug.

The cost-effectiveness threshold

In order to compare costs and effects, it is necessary to bring these to the same scale. For example, it is necessary to express the effect in monetary terms, to allow the comparison with the costs or vice versa. This involves attributing a value to each unity of the effect (i.e. 1 QALY) and is commonly called cost-effectiveness threshold. In other words, this represents the amount of money that

the system is willing to pay for a unity increase in health (i.e. 1 QALY). In UK this value is between 20,000£ and 30,000£.

Cost-effectiveness analysis

CEA aims to compare different interventions in terms of costs and effects via the definition of a cost-effectiveness threshold, as defined above. At this purpose, the Incremental Cost-effectiveness Ratio (ICER) can be defined as the cost of each individual unit of effect (usually QALY) gained as a result of the intervention (Briggs, Claxton et al. 2006). Therefore, ICER can be calculated as the ratio between the difference in costs ($\Delta cost$) and the difference in effect ($\Delta effect$) as may derive form the implementation of the intervention (see Equation 4-24)

$$ICER = \frac{\Delta cost}{\Delta effect}$$

Equation 4-24

As costs and effects vary for each patient i, then different values of ICER can be calculated as $ICER_i = \Delta c_i / \Delta e_i$, where it is assumed that an intervention A (i.e. new intervention) is compared with an intervention B (i.e. current standard), then

$\Delta c_i = cost\ of\ intervention\ A - cost\ of\ intervention\ B$; and

$\Delta e_i = effect\ of\ intervention\ A - effect\ of\ intervention\ B$.

The geometrical interpretation of the ICER needs the definition of the cost-effectiveness plane, which is the plane were individual patient differences in effects (x-axis) are plotted against differences in costs (y-axis) (Briggs, Claxton et al. 2006). As represented in Figure 4-5, the cost-effectiveness plane is divided in 4 areas using, for example, the four main directions as defined on a compass:

- NW (Nord West), where the intervention B is said dominant (less expensive and more effective)
- NE (Nord East), in this case none of the intervention is dominant, a trade off between costs and effects is needed, for example, using the threshold as explained above.
- SE (South East), where the intervention A is said dominant (less expensive and more effective)
- SW (South West), same as NE.

Thus, when none of the intervention is dominant (i.e. when none of the interventions is both cheaper and more effective than the other), the cost for a unity increase in health is compared to the cost-effectiveness threshold. This is the ICER as defined above, and on the cost-effectiveness plane it is represented, for each patient i, by the slope of the line passing through the origin of the axis and the point $Q_i = (\Delta c_\mathrm{i}; \Delta e_\mathrm{i})$.

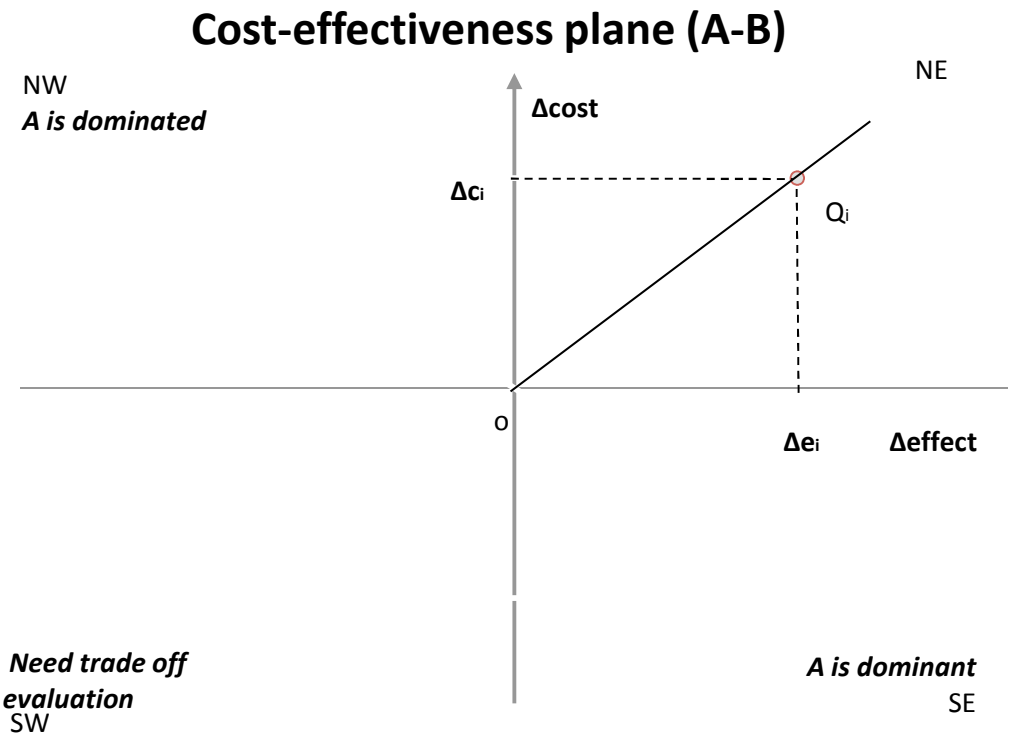


Figure 4-5 The cost-effectiveness plane.

Let's suppose that the trade off between costs and effects for an intervention is expressed as the following: intervention A is cost-effective if $ICER > \lambda$, where λ represents the willingness to pay of a funding body for a unit increase in health (i.e. the cost-effectiveness threshold as defined above). This is a comparison between slopes of lines in the NE quadrant of the cost-effectiveness plane; if the observed line (i.e. with slope equal to ICER) lies below the threshold line (i.e. with slope equal to λ), then treatment A is cost-effectiveness ($ICER < \lambda$), otherwise intervention B is cost-effective ($ICER > \lambda$). Formally, the intervention

A is cost-effective if $\Delta c \ / \Delta e \ < \lambda$, that is if $\lambda \Delta e - \Delta c > 0$, where $\lambda \Delta e - \Delta c$ is also called Net Monetary Benefit (NMB)(Briggs, Claxton et al. 2006). In other words, intervention A is cost effective if the associated Net Monetary Benefit, as calculated in Equation 4-25, is positive (Drummond, Sculpher et al. 2005).

$$\mathrm{NMB} = \lambda \Delta e - \Delta c$$

Equation 4-25

Unfortunately, rarely one intervention will be dominant (i.e. the implementation of a new intervention usually involves more costs and has a better effect than the standard intervention); and, when a trade off between costs and effects needs to be made, individual patients ICERs rarely will all lie above or below the line with slope λ (CE threshold). Thus, a better expression of economic decision models is probabilistic because it attempts to capture such variability between patients. In this case, cost-effectiveness acceptability curves (CEAC, see section 4.8.4), which are derived by the NMB when the decision model is probabilistic.

The decision tree

Decision trees are used to compare strategies (decisions) against each other. An example of decision tree to evaluate sequences of diagnostic tests is presented in Figure 4-6. In a decision tree lines connect different nodes. Some nodes are

deterministic (i.e. they represent decisions, indicated by squares), others are stochastic and therefore subject to uncertainty (i.e. they represent uncertain events, indicated by circles). Red triangles represent the end of the causal chain that needs to be evaluated in terms of costs and benefits. Uncertain events are associated to their probability to occur. The problem of combinations/sequences of tests will be explored in Chapter 7 and Chapter 8.

When a decision is made between tests, Briggs, Claxton et at (2006) suggest to include two extreme options:

- No test, treat all. This is the case when nobody is tested and all patients are directly treated, and answers the question “is it worth treating at all?”.
- No teat, no treat. This is the case when no one is neither treated nor tested.

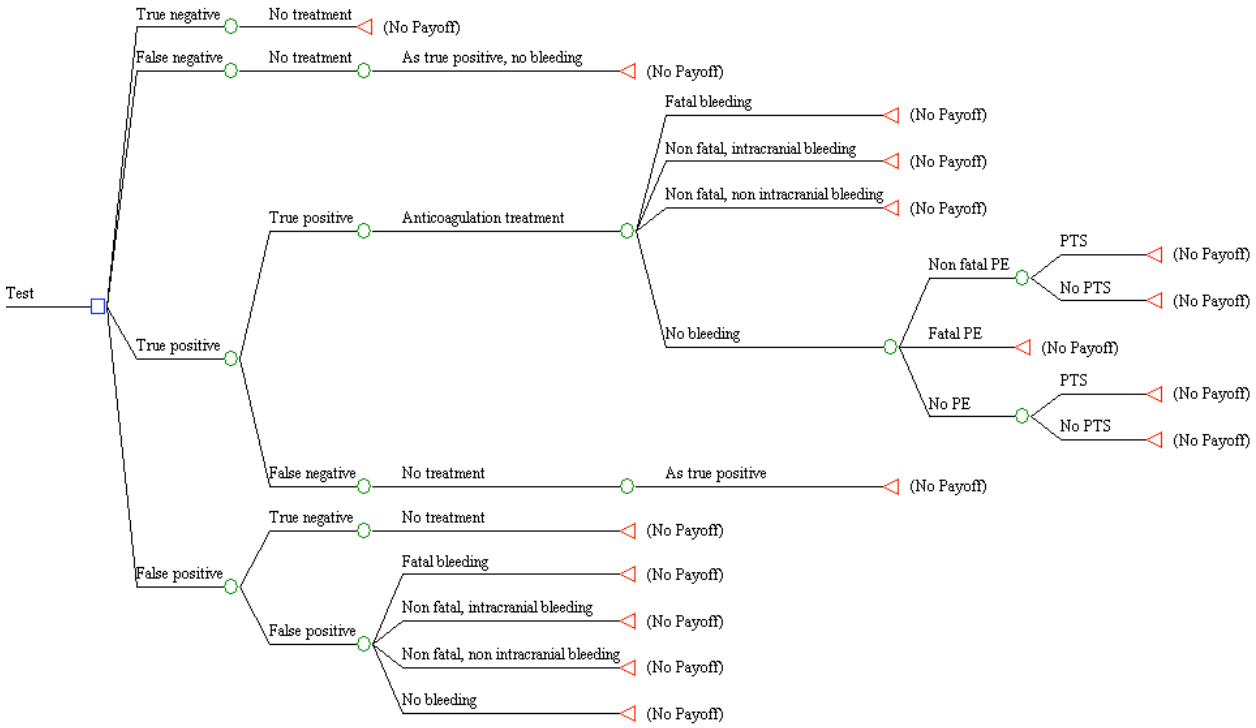


Figure 4-6 Example of decision tree for the accuracy of a sequence of tests (Made by NJ Cooper, 2010 Department of health sciences, University of Leicester).

4.8.3 Probabilistic and Comprehensive decision modelling

Motivations to probabilistic decision modeling

Decision models aid the decision making process by transforming certain input parameters into a set of output parameters, which, subsequently, are associated with some decision rules. This transformation, for example $g()$, may not be a linear transformation, therefore it may be that $E(g(\quad)) \neq g(E(\quad))$ (Drummond, Sculpher et al. 2005). Consequently, the distribution of the output parameters obtained by applying probabilistic decision modeling techniques is useful to calculate correctly the expectation of the output parameter directly from its distribution and to capture parameter uncertainty (Briggs, Claxton et al. 2006). In fact, NICE recommends probabilistic decision models (Briggs, Claxton et al. 2006).

Probabilistic decision models can be implemented using simulation techniques. WinBUGS (software for Bayesian modeling via MCMC simulations, see Chapter 2) allows the model of the data to be implemented along with the model of the decision in a comprehensive decision models. This is presented in the following section.

Comprehensive decision modeling

Bayesian MCMC simulations (see Chapter 2) have recently been used to include all the available sources of evidence, the model of the data and the model of the decision in a "single coherent model" (Cooper, Sutton et al. 2004).

Modeling in WinBUGS allows the same piece of code to *i*. produce an estimate of the parameters from the model of the data, and *ii*. use these estimates to evaluate the decision model and the uncertainty around these estimates is reflected into the decision model parameters. This approach is called comprehensive decision modeling and consists of 4 steps:

1. Develop the decision model. The development of the decision model as first step avoids that this may depend on the information collected at the following stages.
2. Systematic review of the relevant data and its meta-analysis.
3. Estimation of all inputs parameters:
    a. Effectiveness.
    b. Transition probabilities
    c. Costs
4. Evaluation of the model and sensitivity analyses for model, data specification, prior distributions (type of and initial values).

4.8.4 Cost-effectiveness Acceptability Curves

Comprehensive decision modeling is a generalization of probabilistic decision modeling where parameter uncertainty derives directly from the available sources of evidence. In this case, the analysis of the data (meta-analysis) and the decision are evaluated in the same coherent model. If the costs and effects of the new intervention (i.e. intervention A) and the control intervention (i.e. intervention B) are known and do not vary, then the rules explained above can be applied to determine which intervention is cost-effective (i.e. NMB>0). However, costs and effects are characterised by a certain amount of uncertainty as explained above, which is quantified using probabilistic decision modeling. This involves the possibility of simulating a number of times the difference between costs and effects of the two interventions and represent them on the cost-effectiveness plane. For example, Figure 4-5 represents 100 simulations of the costs and effects of two hypothetical interventions. The straight line is associated to a cost-effectiveness threshold of λ=30,000£ and all the simulations can be compared to this line: the simulation below the line are in favour of the intervention A when in the quadrant NE, in this case the NMB>0. The proportion of simulations that follow this rule represents an estimate of the probability of the intervention A being CE.

However, the choice of the cost-effectiveness threshold presents some arbitraries. Thus, the probability of the new intervention A to be cost effective at different cost-effectiveness thresholds can be estimated by calculating the NMB at different

values of λ, and counting the number of simulations that respect the cost-effectiveness rule as presented above (i.e. NMB>0). Those probabilities can be plotted against the respective thresholds to form the cost effectiveness acceptability curve (see Figure 4-8)

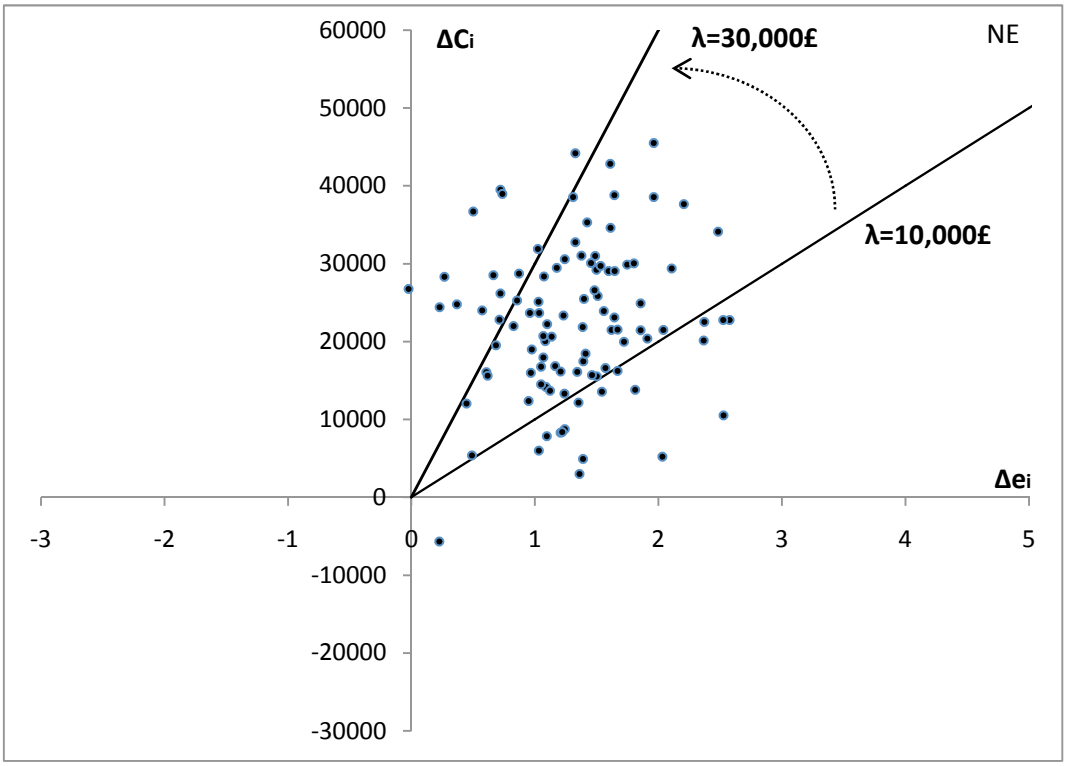


Figure 4-7 Cost-effectiveness plane representing 100 simulations of differences in costs and effects between two interventions from a hypothetical probabilistic decision model.

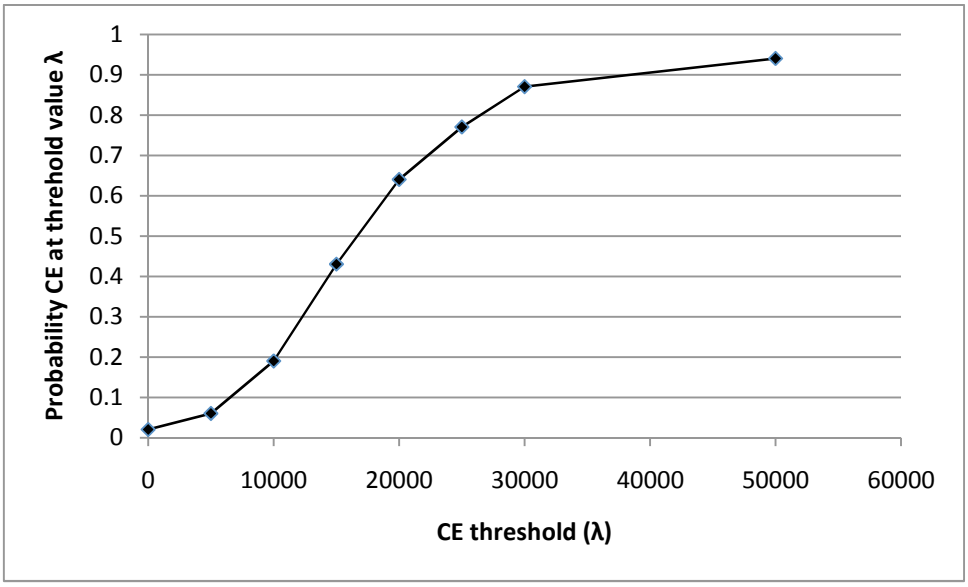


Figure 4-8 Example of Cost-Effectiveness Acceptability Curve.

4.9 Summary

Meta-analysis of diagnostic data is a complex statistical exercise and needs to be performed under the awareness of the main available techniques and the issues that can potentially affect its results. Before listing the existing techniques for the meta-analysis of dichotomous diagnostic test data, it has been necessary to describe the main types of heterogeneity, and distinguish clinical and methodological heterogeneity from diagnostic heterogeneity, where the last is due to variability in diagnostic threshold. A number of techniques exist for the Meta-Analysis of dichotomous diagnostic test data, which have been classified in three groups: *i)* those assuming independence between rates; *ii)* those which attempt to relax this assumption by pooling ROC curves by means of the DOR, and implicitly consider correlation between sensitivities and specificities; and *iii)* those which explicitly consider correlation between sensitivities and specificities. The description of these techniques took a Bayesian perspective, which allowed some of these methods to be improved; especially those based on pooling asymmetric ROC curves. The main advantages of using Bayesian modeling techniques concern the interpretation of the parameter estimates (the uncertainty in posterior estimates can be used to build probability statement) and the possibility of fitting more complex models (i.e. the Bivariate model can be fit by using covariates either on sensitivity or specificity, or by fitting a better version of the asymmetric sROC curve model). Finally, via a unification of the literature on the topic, it has been possible to present all models expressed in terms of the

bivariate model parameters and, where necessary, constraining some of the bivariate model parameters. In particular, formulae to convert the bivariate model parameters into the asymmetric RI model parameters have been given. However, although these formulae attest the mathematical equivalence of these two models, the complexity of the models and the use of different prior distributions may result in slightly different parameter values after the transformation. Also the inclusion of covariates does not assure that these relationships are maintained. In conclusion to the presentation of these techniques, the existing approaches to the construction of sROC curves were presented as from a review of the literature.

While the relationships between models have been treated theoretically, all the other topics were applied to a small meta-analytical dataset GERD, for which, the clinical problem was introduced in Chapter 3. The decision of using this example throughout this chapter is justified because: *i)* it is a small dataset, quite typical for diagnostic tests; *ii)* there was small evidence of diagnostic heterogeneity, and this has been helpful to explore the modeling techniques when the assumption of heterogeneity was weakly met; *iii)* this dataset permitted to discover possible problems of convergence for the slope parameter of the asymmetric RI modeling approach. In contrast, for the rest of the thesis a completely different diagnostic problem will be considered (introduce in section 1.4), for which systematic reviews would retrieve a large number of articles and a stronger evidence of diagnostic heterogeneity is detected. Finally, when the Bayesian bivariate model

has been presented, the procedure to plot credible and predictive regions as adaptation of the existing formulae was presented.

Behind the choice of one of these techniques (see Chapter 5 for a practical example of model choice in a Bayesian framework) there is the need of taking a decision on the best model for the diagnosis of a certain condition. A decision can be taken under the clinical perspective simply considering the summary results from these models (i.e. comparing the sROC curves via the AUC), or it can be taken considering the costs and the benefits of each diagnostic test (economic perspective), for example via a cost-effective analysis.

In conclusion, this chapter was a summary of the methodology for the meta-analysis of diagnostic test presented along with some improvements that were achieved within the Bayesian modeling framework and a summary of principles for cost-effectiveness analysis. Methods for the meta-analysis of diagnostic tests well be applied in the next chapter for the estimate of the accuracy of DD for DVT.

Chapter 5. Application of methods for the meta-analysis of diagnostic accuracy data: the diagnosis of Deep Vein Thrombosis using Ddimer test

5.1 Chapter overview

A number of statistical models for the Meta-Analysis of diagnostic test accuracy data have been explored in Chapter 4. Such analyses are more complex than for studies of therapeutic interventions due to additional issues relating to threshold levels, dependence between sensitivity and specificity, and substantial between-study heterogeneity (Egger, Smith et al. 2001). Throughout Chapter 4, the example of Proto Pump Inhibitor for the diagnosis of Gastroesophageal Reflux Disease (GERD dataset) was used to describe and interpret model results (see section 1.3 for details on GERD example, also sections 3.2.5 and 3.3.8 for a single study of the accuracy of PPI therapy for GERD, also 4.4.4 for the heterogeneity in GERD dataset , and also 4.5.1, 4.5.2, and 4.5.3 for meta-analysis models of the accuracy of PPI therapy for GERD). GERD dataset was used because it was a small dataset without evidence of threshold effect. Very different from GERD dataset is Deep Vein Thrombosis (DVT) dataset, which will be used throughout this chapter. The DVT dataset, differently from GERD, is composed by a large number of publications as already explained in section 1.4. As a consequence, a large amount of unexplained heterogeneity is potentially detected by each method

and can be investigated by means of covariates which will be described in section 5.2, or via the inclusion of random effects as described throughout section 4.5.

In this chapter, the alternative meta-analysis models described in Chapter 4 are applied to the DVT dataset (section 5.3.1) and their results compared. The use of the Deviance Information Criteria (DIC) as well as residual deviance, to decide which model is the most appropriate, is considered, including the consideration of covariates, for a given dataset (Section5.3.2). Different assays are available for Ddimer (DD) test for DVT, which can be all performed using a small blood sample. Thus, often studies evaluate more than one assay. The bivariate model is adapted to adjust for those factors within a study that evaluate multiple assays, and the DIC is used to compare such adaptation to the other models (see section 5.3.3). Models are then fitted to a subset of the data to examine the performance of the model selection approaches to a reduced dataset (see section 5.3.4).

The code to implement the models used in this chapter are available in the folder "Chapter 5 - meta-analysis of DD for DVT" contained in the CD-ROM attached to this thesis. Files are in .txt format, and the model, initials and data are given in separate files; these files are the same that are made available on request for the publication presented in Appendix E.

5.2 The Accuracy of Ddimer for the diagnosis of Deep Vein Thrombosis

As part of an assessment into the most cost-effective approach of diagnosing DVT, Goodacre et al (Goodacre, Sutton et al. 2005) performed a meta-analysis to evaluate the accuracy of DD as a test for DVT. The clinical problem of DVT and the meta-analysis dataset has already been described in section 1.4. This dataset consisted of diagnostic test performance information on 198 assays extracted from 97 publications. In this chapter 196 of the 198 assays were selected (i.e. 2 of the 198 assays were excluded) due to missing covariate data on two assays, which would have not allowed the inclusion of covariates as in section 5.3.2. Figure 5-1 graphically presents the data in ROC space as specified below. (It also includes the results from the bivariate model which will be discussed in the section 5.3.1).

In the original meta-analysis, meta-regression, assuming sensitivity and specificity as independent, was performed to explore the considerable between-study heterogeneity. They identified the study level covariates *study setting* and *type of reference test* to be statistically significant. Some of the most relevant covariates are described below.

Study settings were retrieved from the original dataset in terms of place of measurement and rearranged in 4 levels: 1. *in-patient only* measurement (**IPo**), 15 assays; 2. *emergency department only* (**EDo**), 28 assays; 3. *in-clinic only*

measurement (**Co**), 61 assays; 5. *Mixed/Unclear* (**Mixed**), 92 assays. This last included studies with either multiple or not recorded settings.

Type of Experimental test was considered relevant for the exploration of diagnostic heterogeneity (i.e. variability due to changes in threshold). This variable was missing data for 2 assays. Such records were wholly excluded from the analysis resulting in the 196 assays. There were three main test types: ***ELISA*** (91 assays); ***LATEX*** (76 assays) and *whole blood agglutination* (**WBA**, 29 assays).

There were also data about past history of the patients recruited. In particular, the covariates that were considered were *past history of DVT* (**ExDVT**, binary covariate:Yes/No, 32 studies excluded patients with past history of DVT) and *past history of anti-coagulants* (**ExAC**, binary covariate:Yes/No, 71 studies excluded patients with past history of Anticoagulant). Table 5-1 shows that these two variables may be correlated because patients with past history of DVT (those included in studies where ExDVT=NO), have very likely been treated with anticoagulants (thus ExAC=NO).

Year of publication (**yop**) was considered as the only continuous variable. This may be able to explain differences that occur over time due to improvement in technologies, in medical competencies, better understanding of the experimental test and disease.

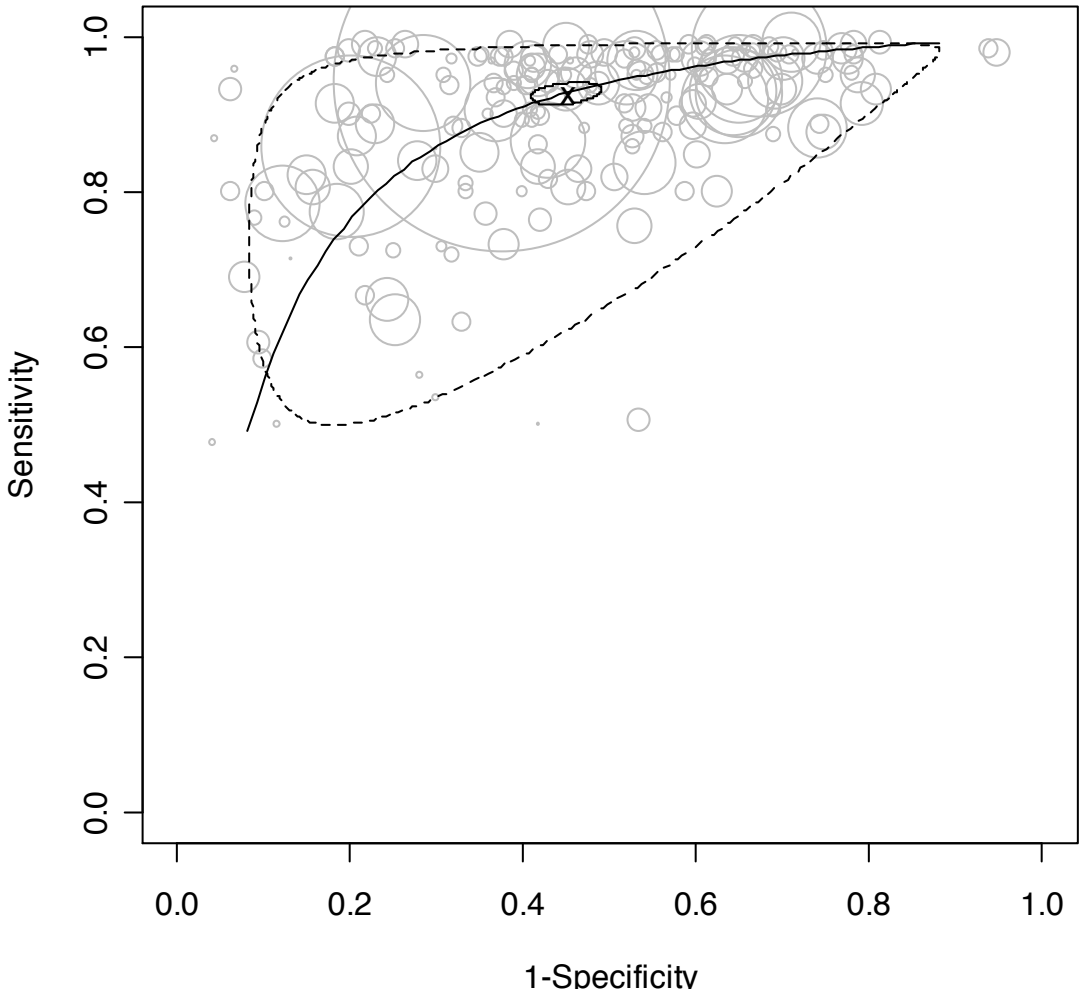


Figure 5-1: Results from the bivariate model plotted on the ROC plane together with credible region (solid line) and predictive region (dashed line). Circles represent the diagnostic test accuracy data for DD from each of the 196 assays included in the meta-analysis (The size of the circles are proportional to total number of patients in the study (i.e. true positives + true negatives + false positives + false negatives)).

		Excluded patients who previously used Anti-Coagulant		
		No	Yes	Total
Excluded patients with history of DVT	No	116	48	164
	Yes	9	23	32
	Total	125	71	196

Table 5-1 Ddimer Assays classified according to the variables "excluded patients with history of DVT (ExDVT)" and "excluded patients who previously used Anti-Coagulant (ExAC)".

5.3 Application to Ddimer test for Deep Vein Thrombosis

All models are evaluated within a Bayesian framework using MCMC simulation and fitted in WinBUGS (Lunn, Thomas et al. 2000). For the purposes of these analyses, all prior distributions are intended to be vague. The MCMC chains were run for 20,000 iterations after a 'burn in' of 5,000 iterations in order to ensure convergence of the MCMC sampler (these initial values were discarded)(Gilks, Richardson et al. 1996).

For all models, posterior distributions of parameter were not sensitive to the type of prior distributions and initial values. A graphical check for convergence did not reveal convergence issues.

5.3.1 Results of Bayesian meta-analysis models

Table 5-2 presents the DICs and residual deviances for all the meta-analysis models for diagnostic test data when applied to the DD dataset. Based on the DIC, it can be observed that, in general, the more complex random-effects meta-analysis models fit the data best, indicated by lower DIC values, with the independent sensitivity and specificity model providing the worst fit to the data (DIC = 5541.05).

	Deviance information criteria		Residual deviance		
	DIC	*pD*	*Sensitivity*	*Specificity*	*Total*
Independent sensitivity and specificity – Fixed effect	5541.05	1.97	576	2410	2986
Independent sensitivity and specificity – Random effects	2129.20	318.81	112	161	273
Diagnostic Odds Ratio - Fixed effect	2412.31	197.31	278	251	529
Diagnostic Odds Ratio - Random effects	2122.40	315.00	113	157	270
Asymmetric sROC – Fixed effect	2388.68	189.72	289	206	495
Asymmetric sROC - Random effects intercept	2110.76	304.60	115	133	249
Bivariate / Hierarchical sROC	2112.88	297.68	116	147	263
Bivariate with covariates	2100.00	288.00	116	146	262
Hierarchical sROC with covariates	2099.65	288.00	116	145	261

Table 5-2. Model fit criteria results applied to the DD dataset.

Overall, the model with the lowest DIC is the asymmetric sROC random effects intercept (2110.76). However, the Bivariate / HsROC model fits almost as well (DIC=2112.88). As guidance suggests that a difference in DIC of more than 5 is important (http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/dicpage.shtml#q9), there would appear to be little to choose between these two models. I have chosen to further explore the Bivariate / HsROC model due to the recent interest in these models and the appealingly direct interpretation of the Bivariate model parameters. It should be acknowledged that it could be considered further exploration of the asymmetric sROC random effect intercept model to be equally valid although potential problems with the convergence of such a model with small datasets heve been observed in section 4.5.2, which is a reason in favour to the choice of the bivariate model.

Parameters		Mean	95% Highest Posterior Density Interval	
			lower	upper
Logit-sensitivity	μ_A	2.572	2.400	2.750
Logit-specificity	μ_B	0.203	0.070	0.330
Residual heterogeneity on sensitivity	σ_A^2	1.035	0.890	1.190
Residual heterogeneity on specificity	σ_B^2	0.897	0.791	0.997
Covariance between logit-sensitivity and logit-specificity	σ_{AB}	-0.551	-0.739	-0.378
Accuracy rates				
Sensitivity	$e^{\mu_A}/(1+e^{\mu_A})$	0.929	0.917	0.940
Specificity	$e^{\mu_B}/(1+e^{\mu_B})$	0.550	0.517	0.582

Table 5-3: Model parameters and accuracy rates estimated using the bivariate model applied to DD dataset.

The parameter estimates relating to the bivariate formulation of this model are presented in Table 5-3 and the corresponding graphical representation is presented in Figure 5-1. The ROC curve was derived following the original formulation by Reitsma et al (Reitsma, Glas et al. 2005) as described in Equation 4-13 in Chapter 4. In this figure, it can be seen that while the mean sensitivity and specificity are estimated precisely, there is considerable between-study heterogeneity which is reflected by the large predictive region.

Based on the residual deviance (Table 5-2) all models containing random effects provide an adequate fit to the data compared to 392 unconstrained data points. In the majority of instances, the residual deviance is higher for specificity than sensitivity. This can potentially be attributed to the observed greater variability in specificity values between studies clearly evident in Figure 5-1.

5.3.2 Inclusion of covariates

In an attempt to further improve the fit of the model to the DD data, and reduce between study heterogeneity, consideration is given to study level covariates. It is natural to consider their inclusion in the best fitting model without covariates – i.e. the bivariate / HsROC model. Recall from Chapter 4 that it is possible to include covariates to influence either or both dimensions being modelled and that the bivariate and HsROC model are only equivalent when covariates are included in both dimensions. Therefore, I have fit covariates to both formulations of the model and compare the results in the following section.

			95% Highest Posterior Density Interval	
Parameters		**Mean**	**Lower**	**Upper**
Logit-sensitivity for In Patient only	$\beta_{1,0}$	2.264	1.746	2.683
Change in logit-sensitivity for Mixed or Unclear setting *vs* In-patient only	$\beta_{1,1}$	0.003	-0.480	0.531
Change in logit-sensitivity for Emergency department only *vs* In-Patient only	$\beta_{1,2}$	0.160	-0.400	0.765
Change in logit-sensitivity for Clinic only *vs* In-Patient only	$\beta_{1,3}$	0.934	0.413	1.470
Overall logit-specificity	$\beta_{2,0}$	0.205	0.083	0.349
Residual heterogeneity on sensitivity	σ_A^2	1.026	0.728	1.317
Residual heterogeneity on specificity	σ_B^2	0.807	0.633	1.001
Covariance between Logit sensitivity and Logit Specificity	σ_{AB}	-0.622	-0.808	-0.442
Accuracy rates				
Sensitivity of In Patient only	$e^{\beta_{1,0}}/(1+e^{\beta_{1,0}})$	0.903	0.860	0.942
Sensitivity of Mixed or Unclear setting	$e^{\beta_{1,0}+\beta_{1,1}}/(1+e^{\beta_{1,0}+\beta_{1,1}})$	0.905	0.886	0.924
Sensitivity of Emergency department only	$e^{\beta_{1,0}+\beta_{1,2}}/(1+e^{\beta_{1,0}+\beta_{1,2}})$	0.917	0.888	0.944
Sensitivity of Clinic only	$e^{\beta_{1,0}+\beta_{1,3}}/(1+e^{\beta_{1,0}+\beta_{1,3}})$	0.960	0.949	0.971
Overall Specificity	$e^{\beta_{2,0}}/(1+e^{\beta_{2,0}})$	0.551	0.528	0.586

Table 5-4: Model parameters and accuracy rates estimated using the bivariate model with inclusion of the study setting covariate for sensitivity applied to the DD dataset.

Initially covariates were added individually to each dimension in each underlying model (i.e. bivariate or HsROC) to explore their effect and contribution to model fit (using the DIC statistic). Following this, a "best model" for each underlying model was constructed by sequentially considering covariates for either dimension, starting with the null model and then choosing those that improved model fit greatest when added stepwise to the model. A covariate was retained in the model if it reduced the DIC by more than 5 (The BUGS project).

Using this approach for the bivariate model, only *study setting* improved the fit of the model when added to sensitivity and no covariates improved the fit of the model when added to specificity (DIC 2100). Thus, the specification of the final model is as follows:

$$\boldsymbol{logit}(\boldsymbol{sens_i}) = \boldsymbol{\beta_{1,0}} + \boldsymbol{\beta_{1,1}} * \boldsymbol{Mixed_i} + \boldsymbol{\beta_{1,2}} * \boldsymbol{EDo_i} + \boldsymbol{\beta_{1,3}} * \boldsymbol{Co_i}$$

$$\boldsymbol{logit}(\boldsymbol{spec_i}) = \boldsymbol{\beta_{2,0}}$$

Table 5-4 presents the parameters and accuracy rates estimated by the final bivariate model for DD data including this covariate. Also included are the heterogeneity parameters for sensitivity and specificity as well as the associated covariance term. It would appear that the inclusion of this covariate has had minimum impact on the (residual) between-study heterogeneity and thus the majority of variability remains unexplained. The sROC curves corresponding to the different values of the study setting covariate are presented in Figure 5-2

according to the formulation given by Equation 4-13 in Chapter 4. It can be concluded that the diagnostic accuracy of DD is greatest when the test is carried out in a *clinic* setting.

Parameters	**Mean**	**95% Highest Posterior Density Interval**	
		Lower	**Upper**
Accuracy parameter for In Patient only $\beta_{1,0}$	2.20	1.70	2.70
Change in accuracy parameter for Mixed or Unclear setting *vs* In-patient only $\beta_{1,1}$	0.07	-0.44	0.58
Change in accuracy parameter for Emergency department only *vs* In-Patient only $\beta_{1,2}$	0.44	-0.18	1.02
Change in accuracy parameter for Clinic only *vs* In-Patient only $\beta_{1,3}$	1.08	0.54	0.61
Overall threshold parameter $\beta_{2,0}$	1.09	0.94	1.25
Residual heterogeneity on the accuracy parameter σ_{Λ}^2	0.75	0.63	0.88
Residual heterogeneity on the threshold parameter σ_{Θ}^2	0.86	0.76	0.96
Accuracy rates			
Sensitivity of In Patient only $(\beta_{2,0}+\beta_{1,0}/2)e^{-\beta/2}/[1+(\beta_{2,0}+\beta_{1,0}/2)e^{-\beta/2}]$	0.912	0.889	0.935
Sensitivity of Mixed or Unclear setting $[\beta_{2,0}+(\beta_{1,0}+\beta_{1,1})/2]e^{-\beta/2}/\{1+[\beta_{2,0}+(\beta_{1,0}+\beta_{1,1})/2]e^{-\beta/2}\}$	0.925	0.900	0.929
Sensitivity of Emergency department $[\beta_{2,0}+(\beta_{1,0}+\beta_{1,2})/2]e^{-\beta/2}/\{1+[\beta_{2,0}+(\beta_{1,0}+\beta_{1,2})/2]e^{-\beta/2}\}$	0.929	0.912	0.945
Sensitivity of Clinic only $[\beta_{2,0}+(\beta_{1,0}+\beta_{1,3})/2]e^{-\beta/2}/\{1+[\beta_{2,0}+(\beta_{1,0}+\beta_{1,3})/2]e^{-\beta/2}\}$	0.949	0.938	0.960
Specificity of In Patient only $(\beta_{2,0}-\beta_{1,0}/2)e^{\beta/2}/[1+(\beta_{2,0}-\beta_{1,0}/2)e^{\beta/2}]$	0.504	0.445	0.566
Specificity of Mixed or Unclear setting $[\beta_{2,0}-(\beta_{1,0}+\beta_{1,1})/2]e^{\beta/2}/\{1+[\beta_{2,0}-(\beta_{1,0}+\beta_{1,1})/2]e^{\beta/2}\}$	0.509	0.474	0.547
Specificity of Emergency department only $[\beta_{2,0}-(\beta_{1,0}+\beta_{1,2})/2]e^{\beta/2}/\{1+[\beta_{2,0}-(\beta_{1,0}+\beta_{1,2})/2]e^{\beta/2}\}$	0.552	0.501	0.601
Specificity of Clinic only $[\beta_{2,0}-(\beta_{1,0}+\beta_{1,3})/2]e^{\beta/2}/\{1+[\beta_{2,0}-(\beta_{1,0}+\beta_{1,3})/2]e^{\beta/2}\}$	0.626	0.586	0.666

Table 5-5: Model parameters and accuracy rates estimated using the HsROC model with the inclusion of the study setting covariate for the accuracy parameter applied to the DD dataset.

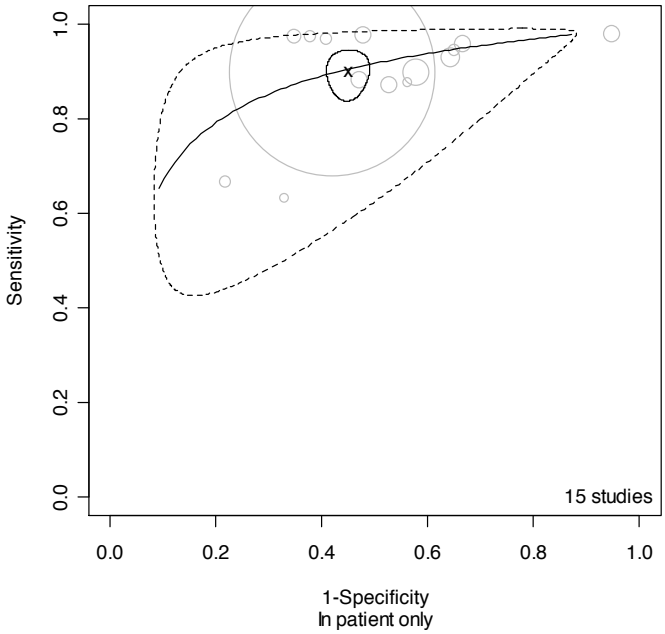
Using the same "step-forward" approach, but considering their effect on accuracy and threshold parameters form the HsROC model, only *study settings* improved the fit of the model when added to the accuracy parameter, and no covariates improved the fit when included on the threshold parameter. The final model for HsROC parameters is

$$\boldsymbol{logit}(\boldsymbol{\Lambda_i}) = \boldsymbol{\beta_{1,0}} + \boldsymbol{\beta_{1,1}} * \boldsymbol{Mixed_i} + \boldsymbol{\beta_{1,2}} * \boldsymbol{EDo_i} + \boldsymbol{\beta_{1,3}} * \boldsymbol{Co_i}$$

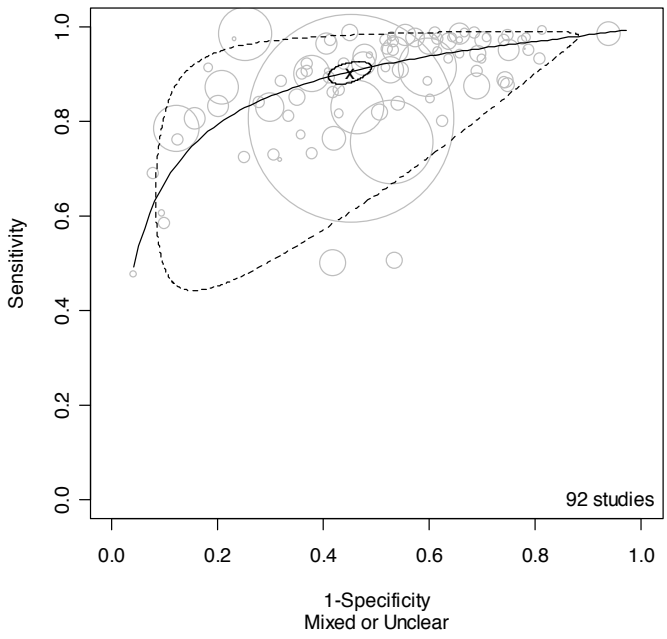
$$\boldsymbol{logit}(\boldsymbol{\Theta_i}) = \boldsymbol{\beta_{2,0}}$$

Table 5-5 presents the parameters and accuracy rates estimated by the final HsROC model for DD data including the covariate. It can be seen that the effect of the covariate on the accuracy parameter (Λ_i) affects the estimates of both sensitivity and specificity and thus results in different estimates of test performance for the four patient subgroups as indicated by the accuracy rates in Table 5-4 and Table 5-5. The DIC for the 'final' bivariate (2100.00 pD 288) and HsROC (2099.65, pD 288) models including covariates are very similar suggesting the goodness of fit of both models is approximately equal.

(a)

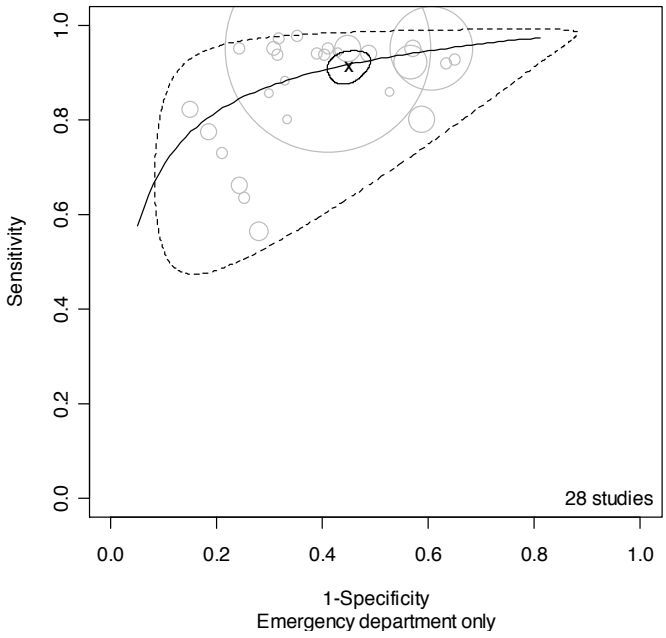

Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
15 studies
1-Specificity
In patient only

(b)


Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
92 studies
1-Specificity
Mixed or Unclear

[continued]

(c)



(d)

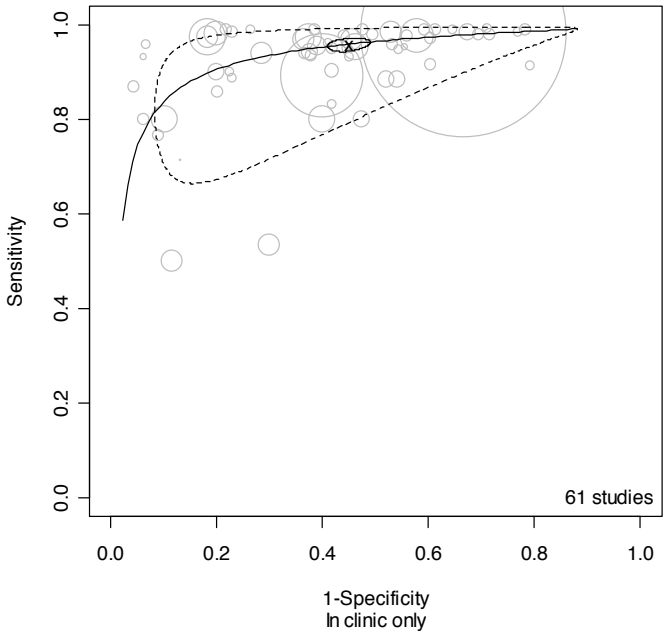


Figure 5-2: Summary ROC curves (solid line), and 95% credible (solid eclipse) and predictive regions (dashed eclipse) for the best fitting model for the (full) D-dimer data. Settings categories are a) In Patient only, b) Mixed/Unclear c) Emergency department only, and d) Clinic only.

5.3.3 Random effect modeling for studies reporting results from multiple assays

In the case of meta-analysis, studies may report the accuracy of a number of different assays of the same test. For our example dataset on the accuracy of DD for DVT, two different assays are either assays which are technically different (i.e. one is based on a qualitative assessment of the colour of a patch and the other requires the quantitative measurement of an enzyme in the blood) or technically similar but produced by different companies. Thus, the same study may report the accuracy of ELISA DD produced by A, ELISA DD produced by B, LATEX DD produced by C and so on, where A, B and C are three different companies/producers.

Amongst the 96 publications included in our review, 57 evaluated the accuracy of a single assay and the remaining 39 evaluated the accuracy of 2 or more assays (2 assays analysed in each of 18 publications, 3 assays in each of 9 publications, 4 assays in each of 5 publications, 5 assays in 1 publication, 6 assays in each of 4 publications, 7 assays in each of 2 publications and 13 assays in 1 publication). Accounting for multiple assays studies may be crucial when heterogeneity is explored. In fact, it may be assumed that the accuracy of the assays from the same studies are similar to each other because they are affected by some characteristics of the study itself (i.e. assays applied to the same population, similar inclusion criteria, same group of assessors, study design, etcetera).

Using the bivariate random effect model as in Model 4-9 in Chapter 4 it is possible to estimate the mean logit sensitivity and specificity, the standard deviations of their posterior distributions, and 3 heterogeneity parameters σ_A^2, σ_B^2 and σ_{AB}^2. In a dataset where each study contributes with one record to the data, these would quantify the unexplained heterogeneity due to unobserved characteristics (clinical or methodological differences between studies) and variability due to differences in implicit and explicit threshold between studies. However, in our dataset, each study contributes to the data with one or more records (assay). Thus, a random effect to account for unexplained heterogeneity due to differences between studies (similarly, due to similarities between assays within the same study) has been added to the bivariate model following the formulation in Equation 5-1.

$$TP_{ij} \sim binomial(\pi_{Aij}, (TP_{ij} + FN_{ij}))$$

$$TN_{ij} \sim binomial(\pi_{Bij}, (FP_{ij} + TN_{ij}))$$

$$\mu_{Aij} = logit(\pi_{Aij}) + \upsilon_{Aj}$$

$$\mu_{Bij} = logit(\pi_{Bij}) + \upsilon_{Bj}$$

$$\upsilon_{Aj} \sim N(0, \sigma_{\upsilon A}^2)$$

$$\upsilon_{Bj} \sim N(0, \sigma_{\upsilon B}^2)$$

$$\begin{pmatrix} \mu_{Aij} \\ \mu_{Bij} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \Sigma_{\mathrm{AB}}\right) \quad \text{with} \quad \Sigma_{AB} = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix}$$

Equation 5-1

Where μ_{Aij} and μ_{Bij} are the logit sensitivities and specificities for assay i in study j and Σ_{AB} is the variances and covariances as defined for equation 1. υ_{Aj} and υ_{Aj} are study specific random effects.

According to the DIC this model fits the data better than the bivariate model (DIC 2092.16). It is true that the more complex the model, the better one should expect the fit of the model. However, the DIC already incorporates a measure of complexity of the model (*pD* the estimated number of parameters). There is no doubt there is still a big amount of heterogeneity still to be explained in this dataset.

The model fitting exercise in section 5.3.2 has been repeated with this new model. This means that the study random-effects first and then covariates are used to explore the residual heterogeneity. None of the covariates had a significant impact on the DIC. Apparently, after the random-effect had been set, there were not significant differences between settings (DIC 2092).

However, Higgins et al (Higgins, Thompson et al. 2009) suggests that random effects should be used to explore heterogeneity that cannot be explained using covariates. In this case, it can be noted that the bivariate approach with study setting as covariate fitted in section 5.3.2 (DIC 2112) is fitting worse than the new random effect model when settings is still considered as explanatory of part of the heterogeneity on sensitivity (DIC 2092).

Thus, should study setting be considered as useful to explain part of the heterogeneity? The answer may be easy if it is considered that study setting is a study specific covariate. That means that all the assays from the same study have the same study setting. In this case it may be still very useful to consider this covariate into the model along with the study random effect.

5.3.4 Analysis of a smaller dataset

In patients only (15 studies)	Deviance information criteria		Residual deviance		
	DIC	*pD*	*Sensitivity*	*Specificity*	*Total*
Independent sensitivity and specificity – Fixed effect	360.30	1.99	41.74	120.83	162.57
Independent sensitivity and specificity – Random effects	169.34	25.96	8.35	11.41	19.76
Diagnostic Odds Ratio - Fixed effect	175.79	16.04	15.98	13.65	29.63
Diagnostic Odds Ratio - Random effects	171.18	24.42	10.21	11.57	21.7
Asymmetric sROC – Fixed effect	176.93	16.36	16.81	13.55	30.36
Asymmetric sROC - Random effects intercept	171.00	25.33	9.85	11.29	21.14
Bivariate / Hierarchical sROC	169.14	24.31	9.19	11.93	21.15

Table 5-6: Model fit criteria results applied to the sub-sample of 15 hospital in-patient only DD studies.

The dataset presented above is rather atypical in that many meta-analyses will include a considerably smaller number of studies. I have repeated the analysis for the sub-set of 15 studies relating to hospital inpatient-only patients (Figure 5-3) to assess how well the DIC works for discriminating between models when less data is available. The model that accounts for studies reporting multiple assays was not

evaluated in this case. The differences between the DIC values for the different models (Table 5-6) is now much less pronounced, with only the fixed effect independent estimates of sensitivity and specificity standing out as being markedly worse than the other models. This highlights that although comparison between DIC values provides a helpful framework for choosing between models, its discriminatory ability is limited by the amount of data available.

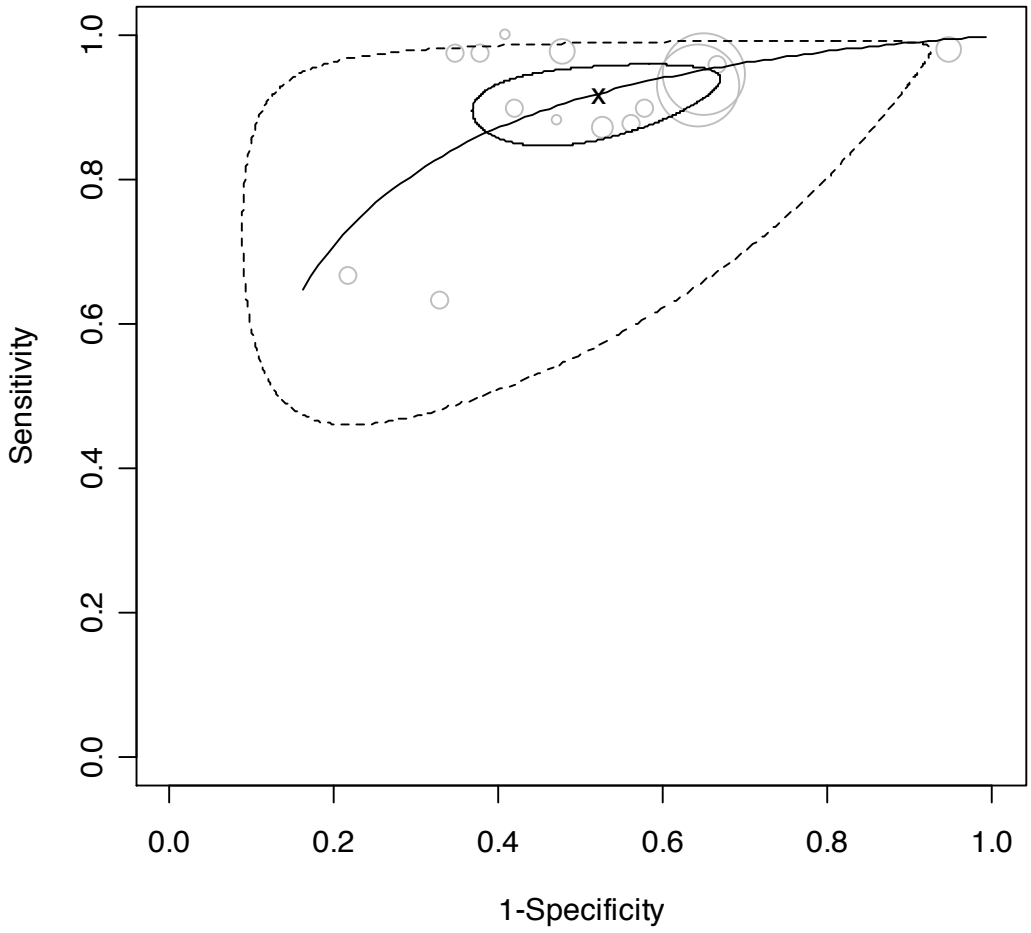


Figure 5-3: Observed accuracies and results from the bivariate model plotted on the ROC plane together with credible region (solid line) and predictive region (dashed line) for the sub-sample of 15 hospital in-patient only DD accuracy studies.

5.4 Summary

Methods for meta-analysis of diagnostic accuracy data have evolved rapidly over recent years. The use of the DIC statistic has been considered for choosing between the models. Model selection methods have been somewhat overlooked in meta-analysis. This is probably as a result of a combination of factors including: i) the lack of measures to compare between (non-nested) hierarchical models until recently; ii) the fact that the majority of people who conduct meta-analyses are not statistical experts and view meta-analysis more as a "procedure" than a statistical model fitting exercise. However, this is not the first time the DIC and residual deviance statistics have been used to inform model choice in evidence synthesis, although their use has largely been reserved to applications using modeling approaches beyond "standard" meta-analysis models (Welton, Cooper et al. 2008; Cooper, Sutton et al. 2009). While such statistics could be used to compare the simplest meta-analysis models (e.g. deciding between the usual fixed and random effect models commonly applied to randomised controlled trials) they are perhaps at their most useful in situations where many model specification options exist, such as is the case for diagnostic accuracy data.

A recurring issue is that in modeling of the type of data presented in this paper, there can be many possible candidate models which mean assessing the fit of each one will often be impractical. What is required is a modeling strategy to overcome this. Here a two-stage approach has been used to select the best underlying model for the data and explore the effect of covariates afterwards. This seems sensible,

but, there is perhaps no guarantee that, if covariates had been added to an alternative underlying model, a better fitting model could not be identified.

A further challenge in meta-analysis of diagnostic test data is a clinically meaningful presentation of the data with the construction of confidence regions and (multiple alternative) sROC curves possible from the same statistical model and thus model fit statistics do not inform such presentational issues. A recent paper (Chappell, Raab et al. 2009) presented an alternative strategy for deciding the most appropriate meta-analysis model for diagnostic test data and the most appropriate way to present the results of the resulting analysis. A specific aspect of this strategy was deciding when presenting a summary ROC curve is appropriate. It was suggested that this should be based on an assessment of the degree of heterogeneity for the true and false positive rates and the correlation between them. A study comparing the use of DIC with this alternative approach to model selection would be worthwhile. Further, there would be nothing to stop presentational decisions following the use of DIC to choose between models being informed by the ideas presented in this alternative approach.

An increasingly important use of results of meta-analysis of diagnostic test accuracy data is to inform economic decision models (Sutton, Cooper et al. 2008). While differences in point estimates of sensitivity and specificity did not appear to change considerably between models in the application presented, the associated uncertainty and variability around these estimates did. Correct quantification of these is important when comparing alternative tests in a decision making context.

Such a framework also highlights the limitations in comparing tests using area under the sROC curve since relative test performance will change with test thresholds.

I have focused on the situation where only one datapoint in ROC space is used from each study. Methods are emerging which relax this and allow for multiple data-points per study, and these should be utilised where possible to maximise the information used in the analysis (Dukic and Gatsonis 2003; Hamza, Arends et al. 2009). Further approaches exist for dealing with diagnostic data which categorises individuals into more than two categories (Bipat, Zwinderman et al. 2007), or is based on the underlying distributions on the test scores (Hellmich, Abrams et al. 1999). It is important to note that data on test threshold are not included in any of the meta-analysis models and are not always presented in primary publications. This would seem an important limitation of all approaches considered here and an issue which is in need of further work.

The next chapter investigates the use of the meta-analysis approaches presented in Chapter 4 and applied in chapter 4 (to the GERD dataset) and in this chapter (to the DVT dataset) with respect to the estimation of the diagnostic accuracy and to their use in economic decision models.

Chapter 6. A review of the evidence synthesis methods used to inform economic evaluations in NIHR - Health Technology Assessment publications

6.1 Chapter overview

The creation of structures in the UK (i.e. National Institute for Health and Clinical Excellence) and elsewhere to facilitate evidence-based health policy decision-making has highlighted the role that systematic reviews including, where appropriate, meta-analysis, and economic evaluations have to play in the decision-making process. These methodologies provide answers to fundamental questions such as: Does the technology work, for whom, at what cost, and how does it compare with alternatives (NICE 2008)? In the area of diagnostic test performance, such evidence-based evaluations are crucial to the decision making process as early diagnosis can lead to diseases being treated more successfully than if treatment were delayed.

Previous chapters have explored all the different meta-analysis techniques developed for the accuracy of diagnostic test data and shown how the format of the results produced by each of the different meta-analysis models differs considerably (see Chapter 4 and Chapter 5). The challenge when evaluating the cost-effectiveness of diagnostic tests is how best to synthesise the available

evidence and then appropriately incorporate the results of the synthesis into an economic decision model. In this chapter, it is investigated how evidence on test accuracy is used to inform decision models developed to evaluate the cost-effectiveness of diagnostic tests. In particular, the focus is on diagnostic tests evaluated as part of the NHS Research and Development Health Technology Assessment (HTA) programme since 1997 and investigate how the evidence on diagnostic test accuracy identified as part of the systematic review is used to inform the diagnostic test accuracy parameter(s) of the economic decision model. Where evidence synthesis methods have been applied to combine test accuracy data from a number of studies, the review focuses on the specific meta-analysis models adopted and how these pooled results are used in the economic evaluation, if at all.

6.2 Methods

All NHS Research & Development Health Technology Assessment (HTA) Programme reports listed on their website (http://www.ncchta.org/project/htapubs.asp) as published between 1997 and May 2009 inclusively were reviewed by myself with the aim of identifying reports that evaluated the performance of diagnostic tests. First the HTA reports were categorised, based on their title, as: (i) Methodology, (ii) Treatments alone, or (iii)

Testing. Where classification was unclear from the title, abstracts followed by executive summary and then introduction were reviewed as necessary.

The second step was to sub-divide those HTA reports classified as Testing into one of the following subgroups: i) Diagnosis, ii) Screening, iii) Prognosis and iv) Monitoring. Occasionally, a report could be classified into more than one subgroup. If a report contained diagnosis and prognosis, screening or monitoring then the report was classified as diagnosis. For all other combinations the report was classified according to its main objective established by reading the main text of the report. Where the purpose(s) of the testing was unclear, categorisation was established via consensus forming discussions with two of my supervisors (NJ Cooper & AJ Sutton, Department of Health Science, University of Leicester, Leicester).

All reports evaluating diagnostic tests were reviewed to identify whether an economic decision model had been developed as part of the HTA. Those reports where economic models had been developed were examined further to establish whether they contained meta-analyses of diagnostic accuracy data in the clinical review section of the report. Those reports that had defined our sample of interest were scrutinised further. Specifically, data were extracted on:

I. All meta-analysis methods used in the clinical review.

II. Whether any of the meta-analysis methods were used to derive estimates of test performance for the economic model. If yes, which method used. If no, the alternative method used to estimate diagnostic test accuracy parameters specifically for the economic model.

III. Whether the economic model had considered pathways involving multiple test combinations, and if so, how test performance had been estimated for the combinations of tests.

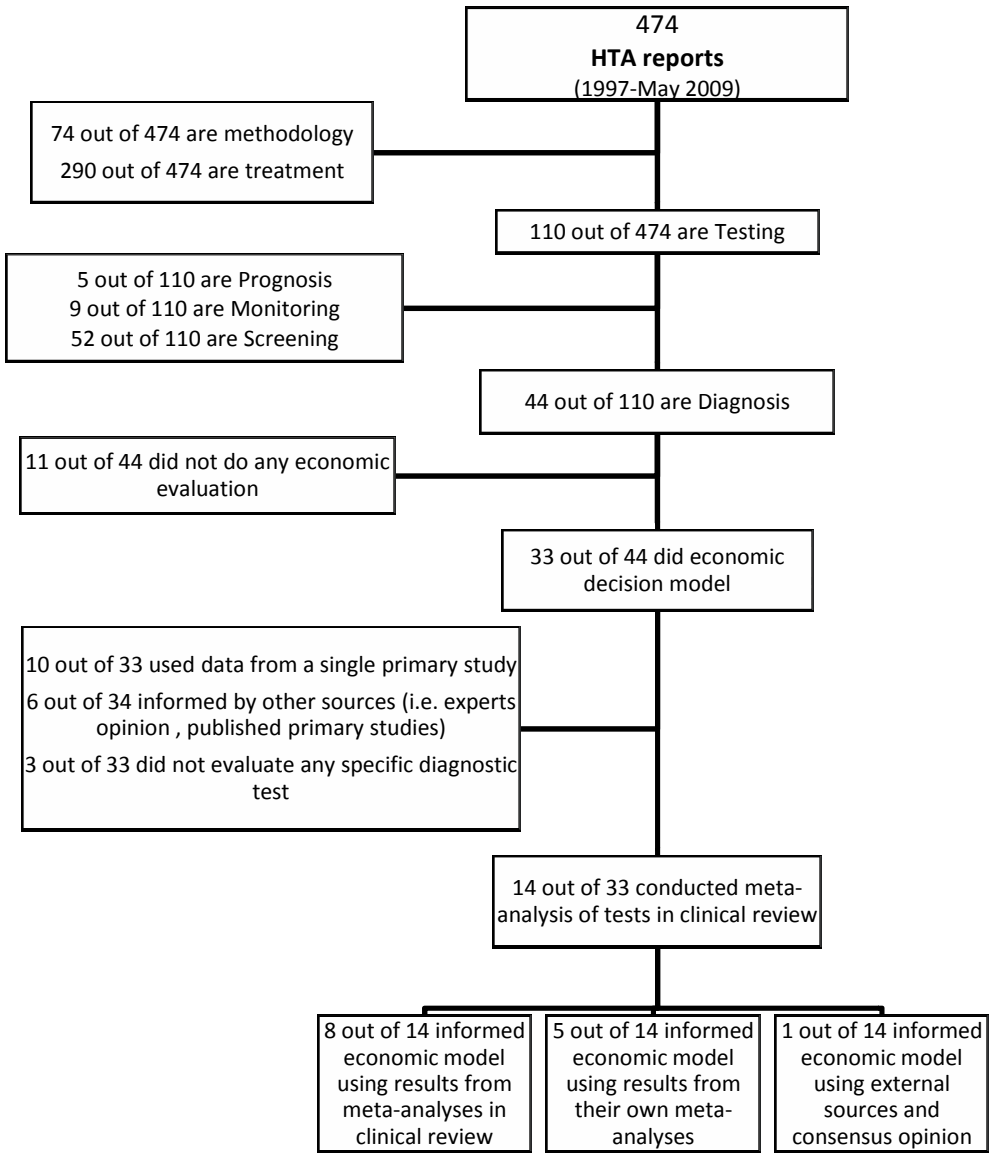


Figure 6-1: Flowchart of excluded and included studies.

6.3 Results

Figure 6-1 shows our classifications of the 474 HTA reports published between 1997 and May 2009 inclusively. 110 out of the 474 reports (23%) were classified as 'Testing' with 44 (40%) of these focusing on 'Diagnosis'. Thirty-three out of the 44 'Diagnosis' reports (75%) included an economic evaluation. Of these 33, 14 (42%) included meta-analysis of diagnostic test accuracy in the clinical review section of the report and these 14 reports defined our sample of interest (A numbered reference list (S1-14) for this sample is provided in Table 6-2, and online at: http://www.ispor.org/Publications/value/ViHsupplementary/ViH13i8_Cooper.asp).

In Table 6-1 the 14 reports that performed meta-analysis as part of the clinical review are listed chronologically together with the meta-analysis method(s) used (denoted by the letter R in the table). The methods are listed broadly in order of complexity and it can be observed that most reports used more than one meta-analysis method. All of the reports except one (S9), included an independent meta-analyses on specificity and sensitivity thus assuming the two measures to be independent. One of these reports used individual participant data in their meta-analysis rather than summary data (S13)). Two reviews adopted a strategy based on heterogeneity; that is, where evidence of heterogeneity existed the Littenberg and Moses regression approach was adopted otherwise independent pooled estimates of sensitivity and specificity were obtained. (S1, S6) The most

sophisticated methods of bivariate and hierarchical summary receiver operating characteristic curve were only applied by 2 of the reviews. (S9, S10) Five of the reports considered study-level covariates in their analyses (S4, S5, S9, S14, S15).

	HTA Report Volume/Number													
	2002	2004			2006					2007		2008		2009
Meta-analytic methods used to evaluate diagnostic accuracy	S2	S7	S6	S11	S8	S5	S12	S13	S14	S3[*]	S1	S9	S10	S4[†]
Independent sensitivity and specificity	M	R, M	R, M[‡]	R, M	R, M	R, M	M[‖]	R, M	R	R	M		R, M	R
Likelihood ratio		R	R			R	R		R	R			R	
Diagnostic odds ratio					R			R		R	R		R	
Littenberg and Moses regression approach	R						R		R					
Littenberg and Moses if heterogeneity, if not independent			R								R			
Bivariate model									M[§]			R, M[¶]		
Hierarchical Summary Receiver Operating Characteristic curve													R	

Table 6-1: Meta-analysis methods applied in the systematic review of diagnostic test accuracy (R) and results used as input parameter in the economic decision model (M).

[*]Used data from systematic review to obtain negative predictive values (number of true negatives divided by total number of negatives) and ratio of test positives to test negatives

[continued]

†Used data from external sources and consensus opinion

‡Used median sensitivity and specificity

‖Expert opinion used where no studies identified in the systematic review

§Performed a series of regression analyses to establish the relationship between sensitivity and specificity

¶Unclear how the Bivariate data is dealt with in the probabilistic decision model

S1 Abubakar, I., L. Irvine, C. F. Aldus, G. M. Wyatt, R. Fordham, S. Schelenz, L. Shepstone, A. Howe, M. Peck, and P. R. Hunter. 2007. A systematic review of the clinical, public health and cost-effectiveness of rapid diagnostic tests for the detection and identification of bacterial intestinal pathogens in faeces and food. *Health Technology Assessment* 11 (36):iii-110.

S2 Berry, E., S. Kelly, M. E. Westwood, L. M. Davies, M. J. Gough, J. M. Bamford, J. F. M. Meaney, C. M. Airey, J. Cullingworth, M. Barbieri, A. Jackson, and M. A. Smith. 2002. The cost-effectiveness of magnetic resonance angiography for carotid artery stenosis and peripheral vascular disease: A systematic review. *Health Technology Assessment* 6 (7).

S3 Collins, R., G. Cranny, J. Burch, R. Aguiar-Ibanez, D. Craig, K. Wright, E. Berry, M. Gough, J. Kleijnen, and M. Westwood. 2007. A systematic review of duplex ultrasound, magnetic resonance angiography and computed tomography angiography for the diagnosis and assessment of symptomatic, lower limb peripheral arterial disease. *Health Technology Assessment* 11 (20):iii-120.

S4 Fortnum, H., C. O'Neill, R. Taylor, R. Lenthall, T. Nikolopoulos, G. Lightfoot, G. O'Donoghue, S. Mason, D. Baguley, H. Jones, and C. Mulvaney. 2009. The role of magnetic resonance imaging in the identification of suspected acoustic neuroma: A systematic review of clinical and cost-effectiveness and natural history. *Health Technology Assessment* 13 (18):iii-106.

S5 Goodacre, S., F. Sampson, M. Stevenson, A. Wailoo, A. Sutton, S. Thomas, T. Locker, and A. Ryan. 2006. Measurement of the clinical and cost-effectiveness of non-invasive diagnostic testing strategies for deep vein thrombosis. *Health Technology Assessment* 10 (15):iii-99.

S6 Kaltenthaler, E., Y. B. Vergel, J. Chilcott, S. Thomas, T. Blakeborough, S. J. Walters, and H. Bouchier. 2004. A systematic review and economic evaluation of magnetic resonance cholangiopancreatography compared with diagnostic endoscopic retrograde cholangiopancreatography. *Health Technology Assessment* 8 (10):iii, 1-89.

S7 Mant, J., R. J. McManus, R. A. L. Oakes, B. C. Delaney, P. M. Barton, J. J. Deeks, L. Hammersley, R. C. Davies, M. K. Davies, and F. D. R. Hobbs. 2004. Systematic review and modeling of the investigation of acute and chronic chest pain presenting in primary care. *Health Technology Assessment* 8 (2):iii-78.

S8 Martin, J. L., K. S. Williams, K. R. Abrams, D. A. Turner, A. J. Sutton, C. Chapple, R. P. Assassa, C. Shaw, and F. Cheater. 2006. Systematic review and evaluation of methods of assessing urinary incontinence. *Health Technology Assessment* 10 (6):iii-87.

S9 Meads, C. A., J. S. Cnossen, S. Meher, A. Juarez-Garcia, G. Ter Riet, L. Duley, T. E. Roberts, B. W. Mol, J. A. Van der Post, M. M. Leeflang, P. M. Barton, C. J. Hyde, J. K. Gupta, and K. S. Khan. 2008. Methods of prediction and prevention of pre-eclampsia: Systematic reviews of accuracy and effectiveness literature with economic modeling. *Health Technology Assessment* 12 (6):1-249.

S10 Mowatt, G., E. Cummins, N. Waugh, S. Walker, J. Cook, X. Jia, G. S. Hillis, and C. Fraser. 2008. Systematic review of the clinical effectiveness and cost-effectiveness of 64-slice or higher computed tomography angiography as an alternative to invasive coronary angiography in the investigation of coronary artery disease. *Health Technology Assessment* 12 (17):iii-143.

S11 Mowatt, G., L. Vale, M. Brazzelli, R. Hernandez, A. Murray, N. Scott, C. Fraser, L. McKenzie, H. Gemmell, G. Hillis, and M. Metcalfe. 2004. Systematic review of the effectiveness and cost-effectiveness, and economic evaluation, of myocardial perfusion scintigraphy for the diagnosis and management of angina and myocardial infarction. *Health Technology Assessment* 8 (30):iii-89.

S12 Rodgers, M., J. Nixon, S. Hempel, T. Aho, J. Kelly, D. Neal, S. Duffy, G. Ritchie, J. Kleijnen, and M. Westwood. 2006. Diagnostic tests and algorithms used in the investigation of haematuria: Systematic reviews and economic evaluation. *Health Technology Assessment* 10 (18).

S13 Wardlaw, J. M., F. M. Chappell, M. Stevenson, E. De Nigris, S. Thomas, J. Gillard, E. Berry, G. Young, P. Rothwell, G. Roditi, M. Gough, A. Brennan, J. Bamford, and J. Best. 2006. Accurate, practical and cost-effective assessment of carotid stenosis in the UK. *Health Technology Assessment* 10 (30):iii-128.

[continued]

S14 Whiting, P., M. Westwood, L. Bojke, S. Palmer, G. Richardson, J. Cooper, I. Watt, J. Glanville, M. Sculpher, and J. Kleijnen. 2006. Clinical effectiveness and cost-effectiveness of tests for the diagnosis and investigation of urinary tract infection in children: a systematic review and economic model. *Health Technology Assessment* 10 (36):iii-iv, xi-xiii, 1-154.

Table 6-2 Studies selected as a result of the systematic review.

Table 6-1 also highlights which meta-analysis method (if any) is used to inform the test accuracy parameters in the economic decision model (denoted by the letter M in the table). Where the letters R and M appear in the same cell of the table, this indicates that one of the meta-analysis approaches used in the clinical review was also used to inform the economic decision model. Where the letter M appears in a cell on its own, this indicates that a different meta-analysis method was used specifically to inform the decision model.

Eight out of 14 reports (57%) used independent pooled estimates of sensitivity and specificity obtained from meta-analyses performed in the clinical review as inputs into the decision model, 5 (36%) used study data identified by the clinical review but performed their own meta-analyses (3 out of 5 reports did independent meta-analyses on sensitivity and specificity, 1 out of 5 report did a bivariate meta-analysis model, and 1 out of 5 report obtained negative predictive values and ratio of test positive to test negative), and 1 report used sources external to the clinical review plus consensus opinion. Overall, the majority of reports (10 out of 14 (71%)) used pooled estimates of sensitivity and specificity obtained from the

simplest meta-analysis method, that assumes the two measures are independent of one another, as inputs into the economic decision model. Only two economic decision models used estimates of sensitivity and specificity from meta-analyses that allowed for the correlation between the two quantities attributed to test thresholds varying between studies (i.e. a bivariate model). None of the models used a meta-analysis method that derives an sROC (i.e. Diagnostic odds ratio, Littenberg and Moses regression method, HsROC curve). Ten out of the 14 models reviewed (77%) incorporated the uncertainty associated with pooled estimates to perform a probabilistic cost-effectiveness evaluation.

6.3.1 Evaluation of a combination of diagnostic tests

Six out of the 14 (43%) reports listed in Table 6-1 considered a combination of diagnostic tests in the economic decision modeling. Two of these (S8, S13) assumed the tests to perform independently of one another and thus input the pooled estimates of sensitivity and specificity obtained for each test direct from the meta-analyses. Two reports (S3, S14) assumed the second test to have 100% sensitivity and 100% specificity (i.e. a perfect test). Only one report (S5) clearly stated that the specificity of a second test (d-dimer) depended on the result obtained from the first test (Wells criteria). This was possible due to the data available; that is, a number of studies reported the sensitivity and specificity of the d-dimer stratified by the Wells score (Test performance was assumed independent for all other test combinations evaluated in this report (S5)). The

remaining report (S12) provided no details about how the combination of tests was evaluated.

6.4 Discussion

The focus of this review has been to assess how evidence on test accuracy is synthesised and used to inform economic decision models evaluating diagnostic pathways. The 14 HTA reports reviewed here were all published in the last 7 of the 12 year period considered suggesting that economic evaluation of diagnostic tests via decision models is in its infancy. Due to this it is perhaps not surprising that little has been written on the associated methodology (Sutton, Cooper et al. 2008).

Many of the reports used a range of different meta-analysis methods to synthesise the test performance data. This in itself can be problematic since virtually all the methods make different assumptions, and therefore, theoretically cannot simultaneously be appropriate for a given dataset. Ideally, authors should assess how well each of the proposed models fits the data to identify the 'best' fitting model and thus facilitate interpretation regarding the most appropriate summary of test performance (Novielli, Cooper et al. 2010). Multiple methods were used in many of the clinical reviews but, despite this, the majority of the reports applied the simple meta-analytic approach of assuming sensitivity and specificity to be independent for informing the decision model. This goes against recent guidance from the Cochrane Diagnostic Test Accuracy Group which advises reviewers to

use the hierarchical sROC or the bivariate model for synthesising diagnostic test data as both of these methods overcome the limitations of symmetric and asymmetric sROC curve methods. The Group discredits the Meta-Analysis of independent sensitivities and specificities because such an approach may identify a summary point that is not representative of the paired sensitivity and specificity data (that is, a point that does not lie on the sROC curve). Deeks et al (Egger, Smith et al. 2001) established that when the independent model is used inappropriately (i.e. the primary studies evaluate tests at different thresholds) the resulting point estimate underestimates true test performance (i.e. it lies below the sROC curve that would be produced by an analysis that takes threshold into account). Additionally, if a probabilistic modeling approach is used, this approach will estimate the uncertainty incorrectly.

Although half the reports calculated pooled likelihood ratios for test performance, none went on to use these estimates to inform the decision model. This is understandable since it is not as straightforward to use likelihood ratios compared to estimates of sensitivity and specificity to estimate the number of true positives, true negatives, false positives and false negatives required by the typical parameterisation of decision models evaluating diagnostic tests. Similarly, although methods that estimate an sROC curve (i.e. diagnostic odds ratios & the regression method of Littenberg and Moses) were conducted quite frequently, the output from these analyses was never used to inform the decision model. Again,

this may well be because it is not obvious how to parameterise output in the form of a sROC curve in the decision model. Indeed, one report (S13) stated that meta-analyses were performed on sensitivity and specificity separately, rather than calculating an sROC curve, to obtain the parameters needed for the economic decision model. An sROC curve describes how test performance varies with changing test threshold, therefore it would be possible to consider the cost effectiveness of a diagnostic strategy as a function of test threshold. This could be achieved most simply by running a series of decision models using estimates of sensitivity and specificity for the test(s) at different locations on the sROC curve. In this way, it is possible to identify the optimum threshold – in terms of cost-effectiveness – to use a test at (although it should be acknowledged that, in practice, specifying an exact threshold may or may not be achievable). To our knowledge, this approach has only been attempted once in the published literature (Sutton, Cooper et al. 2008).

A bivariate model, which accounts for the correlation between sensitivity and specificity, was used in two of the reports. There would appear to be growing consensus in the statistical literature that this is the most appropriate model for meta-analysing test performance data (Harbord, Deeks et al. 2007; Arends, Hamza et al. 2008). Therefore, this finding could be interpreted as disappointing. However, it is important to remember that this approach to meta-analysis of diagnostic test data was only described in the literature in 2005 (Reitsma, Glas et

al. 2005) with custom software appearing even more recently (e.g. a macro for Stata (Harbord and Whitting 2009)). It is important to appreciate that it is likely that the research for the HTA reports reviewed here was undertaken prior to the publication of this key paper (Reitsma, Glas et al. 2005) in the majority cases.

Even once the parameter estimates for the bivariate model have been obtained, for a probabilistic decision model, it will be necessary to specify a multivariate normal distribution or a re-parameterisation or approximation to it that is non-trivial (i.e. one of the papers stated using Cholesky Decomposition for this (S14)). Alternatively, it is possible to use a one-stage comprehensive approach to the decision modeling where the meta-analyses are carried out simultaneously in the same computer program that evaluates the decision model. This has been described elsewhere (Sutton, Cooper et al. 2008) using the WinBUGS software (Spiegelhalter, Thomas et al. 2003) which implements MCMC simulation methods, and perhaps provides the most elegant approach available to date.

Despite the above, it is important to note that the bivariate approach should not be used uncritically for the following reason. The method estimates a 95% confidence region for the average sensitivities and specificities observed in the primary studies. Therefore, it is implicit that all the studies are representative of how the test will be used in routine practice. If for example, particular studies use test thresholds which are not representative of routine practice / a particular

threshold being considered, then such an analysis would seem inappropriate. In such cases, exploring cost-effectiveness as a function of an sROC curve, or at one particular point on the curve, would seem more appropriate (although, study level data relating to test threshold is not routinely included in the meta-analysis models and therefore it is not obvious which point on an sROC curve relates to a particular threshold). Given this, further research is required to establish the optimal approach in different situations and this is ongoing.

To add further confusion to this already complex area, it was recently established that the bivariate model and the hierarchical sROC approach are actually re-parameterisations of the same model (Harbord, Deeks et al. 2007) although the two parameterisations lead naturally to different model summaries (i.e. a confidence region in ROC space and an sROC curve respectively, see section 4.7.6 for a discussion of this relationship). Thus, owing to this re-parameterisation, it is possible to obtain an sROC curve from the bivariate analysis and therefore the discussion relating to sROC curves above is also pertinent for this model leading to even more possibilities of how diagnostic test data may be used to inform decision models.

How the application of the different synthesis methods would affect the conclusions in any particular decision problem is difficult to predict since multiple tests may be compared in an economic decision model, and the synthesis

estimates of test performance may be deficient in similar ways (i.e. due to the problems highlighted above). In a previous paper (Sutton, Cooper et al. 2008) the application of the different synthesis methods to a particular decision problem (which incidentally is reference S5 in Table 6-2 and included in the review) is explored. Here the initial HTA assumed independent fixed estimates of sensitivity and specificity but alternative approaches were compared to this. In this example, only relatively small changes in the cost-effectiveness acceptability curves (CEACs) were observed and the decision would not change for most willingness to pay thresholds, but the impact may be considerably greater in other contexts; for example, where the accuracy (and costs) of the competing test strategies are more similar than Ddimer and ultrasound are in this example.

Six of the models reviewed considered diagnostic pathways using multiple tests in combination. The use of combinations of tests is common in clinical practice, e.g. a cheap or non-invasive test may initially be used which has poor specificity and those diagnosed as diseased may go on to receive a more expensive / more invasive test with superior test performance. The main concern is that estimation of accuracy of test combinations was dealt with too simplistically in these reviews (i.e. assuming tests to be independent or the second test to be perfect). Crucially, this is perhaps a limitation of the available data as much as the modeling per se as many primary studies estimating test performance only consider a single test so results of tests conditional on the results of other tests are rarely available. A

concern is that if the strong assumption of test independence is violated, this could lead to misleading conclusions. Further work is needed to establish ways of estimating such correlations. The following Chapter 7 and Chapter 8 respectively describe the characteristics of combinations of tests and their correlation, and develop a modeling approach for the meta-analysis of the accuracy of combinations of tests. Even if they are estimated with considerable uncertainty, including them in the modeling allows the possibility of using value of information methods (Claxton 1999; Ades, Lu et al. 2004) to demonstrate the importance of conducting primary studies to estimate them more accurately.

In conclusion, meta-analytic methods for diagnostic test accuracy data have developed rapidly in recent years. Decision modellers need to be aware of the recent developments in this area and appreciate the limitations of simplistic approaches used commonly in the past. However, more research is needed to refine and develop synthesis methods in this context for the purpose of decision modeling.

Chapter 7. Introduction to the accuracy of combinations of diagnostic tests

7.1 Chapter overview

As discussed in Chapter 3, Chapter 4 and Chapter 5, medical tests are used in routine practice to diagnose patients for the presence or absence of a disease and the measures used to quantify the performance of a dichotomous diagnostic test were also explored. Although the reference test is, if not perfect, more accurate than the index test, it may still be preferable not to perform it as a first choice in clinical practice. For example, it may be invasive, or very expensive, or simply rare or unavailable. For this reason the assessment of the accuracy of an index test is still of great interest. In fact, the index test may have either *i)* a very high sensitivity, and/or *ii)* a very high specificity, or *iii)* neither high sensitivity nor high specificity. In the first case it may be very helpful to exclude healthy patients. In the second case, the positives will be very likely to be diseased, thus can be safely treated. In the third case, it may not be very helpful if used alone. However, it can still be part of a combination (i.e. combined with another, or several other, tests) testing strategy, and the combination itself may be as in case (*i*) (highly sensitive) or case (*ii*) (highly specific) above and thus of diagnostic value.

Several meta-analytic approaches to the evaluation of dichotomized diagnostic tests have been developed in the last decades. A review and application of

methods is presented in Chapter 4 and Chapter 5 respectively (Novielli, Cooper et al. 2010). They have also been integrated into a comprehensive decision modeling framework (Sutton, Cooper et al. 2008). However, while at an individual level diagnosis may be eventually based on a single test, at a population level there is always a proportion of patients whose diagnosis is based on a combination of test results (Zweig and Campbell 1993). For example, individuals who undertake only one test may be those who exit the diagnostic pathway at the first stage, whilst other individuals continue to have more tests. The accuracy of combinations of tests needs to be correctly evaluated before any diagnostic strategy is implemented on a large scale. A challenge in doing this is that tests may not be independent which complicates the evaluation. Section 7.2 presents an overview of statistical approaches to conditional dependence of tests (dependence conditional to the disease status); this includes approaches to testing for conditional independence and approaches to determine and evaluate the best combination of tests.

The exploration of the accuracy of combinations of diagnostic tests presented in section 7.3 will be at the basis of the modeling approach for meta-analysis proposed in Chapter 8.

7.2 Narrative methodological review of the literature for combinations of medical tests

Approaches to the estimation of the accuracy of combinations of tests are not new, although they are rare in the literature. In this section a brief overview of the different approaches will be given.

7.2.1 Methods relating to or inspired by the case of imperfect gold standard

Conditional covariances

Some authors have tried to adjust for the absence of a gold standard by accounting for the dependence between the index test and the reference tests (Hui and Walter 1980; Vacek 1985; Enoe, Georgiadis et al. 2000; Dendukuri and Lawrence 2001). However, their estimates of the probability that the two tests are both positive for a diseased patient, and of the probability that the two tests are both negative for a healthy patient, are calculated as the product of the probabilities of the two events (as if these were independent) plus a term which represent the covariance between the tests (i.e. if this term was zero then the tests would have been independent). For discordant test results conditional to the disease status, covariances had negative sign. Such covariances represent the differences between the expected and the observed proportions that populate the 2 by 2 tables of the joint results of

the tests, conditional to the disease status. A similar model has been used for the case where the disease status is known; that is, in the field of veterinary science to test the hypothesis that tests are conditionally dependent by means of the methods used for hypothesis testing on null hypothesis on Odds Ratios (Gardner, Stryhn et al. 2000). Other authors have developed a test for the hypothesis of conditional independence (null hypothesis) that is based on the calculation of the correlation between two tests (Shen, Wu et al. 2001). Such correlation was calculated as a component of the covariance between the tests as defined above, and there was no evidence of dependence between the tests if the correlation coefficient was not significantly different from zero.

The case where a gold standard is not available (or similarly unknown false positive/negative rates) is not our focus here. However, these techniques are sometimes based on the assumption of conditional dependence between tests and make use of the conditional covariances. The latent variable model is reviewed below.

Latent variable models

Some authors have used latent variable models for the case where a perfect gold standard is absent; such methods have been adapted to allow for a number of tests to be used simultaneously in order to adjust for the dependence between test

results (Joseph, Gyorkos et al. 1995; Dendukuri, Hadgu et al. 2009; Principato, Vullo et al. 2010). However, this model is based on the assumption that the values of a number of tests are available for a number of patients and can be used when the true disease status is not available. If the disease status were available then a latent class model would not be required, but conditional covariances or conditional accuracy rates could be directly estimated. Moreover, such models are based on quantitative measurements and therefore imaging or dichotomous tests would be excluded form this methods.

Moreover, such latent class models model the accuracy of multiple tests based on the assumption of conditional independence of the test accuracy a priori (Joseph, Gyorkos et al. 1995). The effect of this assumption has not been explored.

7.2.2 Methods to build or evaluate the best combination

Combination schemes and conditional accuracy

Initially, the problem of conditional dependence has been explored under the perspective of the use of the Bayes theorem for transforming sensitivities and specificities into predictive values. In particular, the effect of conditional dependence on such application of the Bayes theorem has been explored (Fryback 1978), and the need to consider conditional accuracy rates has been highlighted. The author concludes that assuming conditional independence brings more "degradation of the model" as more "variables" (tests) are considered.

Once the role of conditional accuracy was made clear, the ways of combining tests in sequences must be clear. Thompson (Thompson 2003) describes the two possible schemes for couples of diagnostic tests: *i)* believe the negative and *ii)* believe the positive result (see sections below). The latter scheme, also described as a between tests positivity criteria (see section 7.2.2) means that in the sequence only patients negative to the first test will be further tested. Zou et al. (Zou, Bhagwat et al. 2006) then generalises these criteria to a number of tests (>2) ,and identifies a third combination scheme called Majority (i.e. if the majority of tests are positive then the combination is positive) which is more indicated if the same test is repeated a number of times in a period of time.

Linear discriminant procedures

Various procedures are based on linear discriminant procedures, which select the best combination of tests according to some maximising function (Su and Liu 1993; Liu, Schisterman et al. 2005; Qin and Zhang 2010). However, these are based on assumptions that sometimes can be implausible or difficult to test. For example, the bivariate normality of the test populations of diseased and healthy over the test values, restrictions on the variance-covariance matrices, and the sensitivity must be the highest for every level of specificity. This would be the case of selecting the combination with the ROC curve above all the others (i.e. maximise the AUC) but in section 2 it has been discussed that ROC curves often are asymmetrical and the assumption above is not met. Moreover, this class of

methods can be used only for tests giving numerical measurements (i.e. biomarkers), therefore they cannot be used for imaging techniques and other qualitative tests (i.e. simply red ELISA Ddimer test, as presented in Chapter 5). Similar approaches based on distribution free statistics are also available (Su and Liu 1993; Pepe and Thompson 2000; Huang, Qin et al. 2010). For the implementation of linear discriminant procedures to determine the best combinations of tests, these were compared for the cases where likelihood ratio functions and logistic regression functions were used alternatively (McIntosh and Pepe 2002; Jin and Lu 2008).

The problem of choosing the best combination of tests by comparing different alternatives via the measurement of the trade-off between true and false results – for either diseased in case of believe the positive combination scheme, or healthy for believe the negative combination schemes- has also been considered (Macaskill, Walter et al. 2002).

Probability modifying plot

Another approach that aims to build the best sequence of tests is that based on the probability modifying plot, which is a plot that represents each sequence as a decision tree where tests are applied at each stage (i.e. non scaled horizontal axis), and at each stage the prevalence of disease after each test is represented (i.e. vertical axis) (Severens, de Vries Robbé et al. 1999). These methods used a treatment threshold to decide which sequence of tests is best. Such a threshold is

compared with the prevalence of disease posterior to the test (i.e. negative or positive predictive values). For example, if the positive predictive value after a test is higher than such a threshold (say 80%), then it is worth considering such test in the sequence. Such methods required that tests are applied to the same population of patients and have not been adapted to other cases (i.e. multiple studies).

Meta-analysis techniques

The meta-analysis of the accuracy of combinations of tests has been considered under a few different perspectives. The first, and most common approach, assumes conditional independence between tests (as mentioned in Chapter 6) even when there is not evidence of independence between the tests; for example, when two tests are applied on the same population of patients (i.e. test values reported on the same paper, tests are likely to be correlated). A first modeling attempt to model test accuracy data allowing for conditional dependence between tests was based on repeated measures modeling (Siadaty, Philbrick et al. 2004). This modeling approach was a generalization of the asymmetric sROC fixed effect model proposed by Littenberg and Moses (1993) (see Chapter 4 for asymmetric sROC model) and therefore was based on the meta-analysis of the DOR (see Chapter 3 for Diagnostic Odds Ratio). However, such modeling approach aimed to produce accuracy estimates that were adjusted for conditional dependence rather than to measure the accuracy of combinations of such tests. Moreover, this

approach was mainly fixed effects; random effects versions have been presented for some types of data (i.e. "multilayer cluster structure" of the data). Also, it seemed that the structure of the data, indexed by paper ID (p) and test ID (t), used the test ID as a proxy for the time when the measure was repeated, therefore assuming that the test sequence followed the order of the index t, which is usually not true. They have not investigated the robustness of the results of their modeling approaches for different permutations of the index test where possible (combinations of data reported rather than sequences).

Economic evaluations

Some authors have used economic criteria to select the best combination of tests. The rationale of using costs to compare combinations of tests was that of minimization of costs in order to reduce the waste of resources that may occur if all tests were given to all patients (Rhea, DeLuca et al. 1982). However, the total cost of a combination of tests needs to consider the cost of the tests, and the expected costs of the consequences of the diagnosis (i.e. as a result of false diagnosis); in this case costs can be used to compare very different quantities (adverse events, treatment effects, cost of testing) using the same measure and a cost minimization analysis can be used to choose the best combination (Henschke and Whalen 1994).

Also, in the review of HTA publications presented in Chapter 6, when economic decisions models have been used to compare combinations of tests, the accuracy

of each combination was calculated based on the assumption of independence between tests. If this assumption was not true, then both the meta-analysis (for the estimates of the accuracy rates) and the economic evaluation (informed by the meta-analysis results) are likely to give misleading conclusions.

7.2.3 Summary of the methodological review

According to the aim of this thesis stated in section 1.2, the methods reviewed in this section are not ideal for incorporation of meta-analysis techniques into economic decision framework. In fact, the majority of these techniques are not meta-analyses.

Moreover, the assumptions of the methods based on the estimation of the covariance term have already been discussed in section 7.2. These do not produce results suitable for economic decision modeling. For example, the covariace term indicates the difference between the accuracy of the combination when conditional dependence is assumed instead of independence. In fact, it does not measure how many patients are being tested by the second test (i.e. conditional accuracy rate). At this purpose, the estimation of such conditional rates needs to consider the combination schemes described in section 7.2.2 ("*Combination schemes and conditional accuracy*").

Also, techniques based on the estimation of the covariance term and methods based on linear discriminant analysis are more suitable for tests based on

quantitative measurements. Imaging and more generally qualitative tests are excluded from this class of methods.

The method that can be based on the direct estimation of conditional accuracies is the probability modifying plot. However, it is not clear what the post-test prevalence threshold should be to choose between combinations and it offers only a methods to compare combinations rather than to inform economic decision models. However, it is not excluded that the probability modifying plot can be used to represent graphically the effect of adding a new test on the accuracy of the combination.

Also, the meta-analysis methods described are either based on the assumption of conditional independence of tests, or they produce estimates of the accuracy of the tests adjusted for the conditional independence. On the contrary, for our aim it would be more useful to have the conditional accuracy rates based on the assumption of dependence instead of accuracy estimates adjusted for conditional dependence.

When economic techniques were used, a simplistic cost minimization analysis was used to select the best combination. This only considers the cost implications of testing and ignores the effect of tests on the quality of life of patients. Cost-effectiveness analysis would consider these effects but were informed by meta-analysis that assumed conditional independence.

The next section will describe the conditional accuracy of two dichotomised tests with respect to the combination schemes presented above. In Chapter 8, a cost-effectiveness analysis informed by a meta-analysis model that accounts for dependence between tests will be proposed as an alternative approach to the evaluation/choice of the best combination of tests.

7.3 Combination of diagnostic tests

In this section the words combination and sequence have been used with the same meaning. Other expressions have been used almost as synonyms. The expressions *diagnostic algorithm* or *diagnostic strategy* have been used interchangeably to indicate either combinations or sequences of tests. However, it must be noted that these words are very generic, and an algorithm/strategy may easily be defined by one single test, and not necessarily involve more than one test unless specified. For clarity, in the next section the difference between a combination and a sequence will be explained, and these words will be used with these precise meanings throughout the chapter. Diagnostic strategy will be used generically to define one or more tests combined for diagnosis.

7.3.1 From combinations to sequences of two dichotomised diagnostic tests

Defining a combination of a number of dichotomous tests is not a simple task and it becomes more complex as the number of tests increases. I will focus and describe combinations of two dichotomous tests and their properties.

The simplest combination involves only two tests which are performed on a given set of symptomatic patients. For simplicity, let's refer to two generic tests T1 and T2, where each test gives either a positive or a negative answer to the question "does the subject present the condition of interest?" As discussed in Chapter 3, even in case of continuous test results these can be dichotomised by means of a threshold and represented as in Table 3-1.

Initially, the two tests will be assumed to be given simultaneously to each patient. This is equivalent to the use of both tests independently to each other on the same patient (i.e. the result of one test does not influence whether the second test is undertaken), and then interpret their results simultaneously. However, if a diagnostician wants to use both tests, he clearly needs to choose a positivity criterion. When a single test is used, the positivity criterion is built into the test as described by the diagnostic threshold. The diagnostic threshold is a within test positivity criterion because it defines the positivity of the test (see Chapter 3, Chapter 4 and Chapter 5). In the case of two (or more) tests used simultaneously, another positivity criterion is needed to combine test results; hence, a between test positivity criterion is required. Combination schemes have already been mentioned in section 7.2.2; a detailed representation will be given in this section.

It may usually be decided that patients found to be positive to both tests must be classified as positive, and patients found to be negative to both tests must be classified as negative. The challenge comes for discordant tests results. Table 7-1 shows all the possible combination of results using T1 and T2. The logic operator for combining single test results is always *and;* the problem is how to classify a patient that was positive to T1 *and* negative to T2, or *vice versa*.

Thus, in case of discordant tests results, the diagnostician has the following options:

1. Not taking any decision for discordant test results. This is not a sensible and ethical choice. Moreover, delaying the decision corresponds to classifying everyone as negative (test ignored, no treat), or as positive (test ignored and treat all);
2. Believe one of the two tests as in combination 1 in Table 7-1, this strategy actually makes worthless the execution of the other test. As a matter of fact, the overall accuracy of the combination will be exactly the accuracy of the believed test, but sometimes that will be T1 and sometimes that will be T2 so that combinations still will have defined characteristics. This choice will break down the strategy into a single test diagnosis;
3. Finally, a diagnostician may either *believe the positive* (combination 2) or *believe the negative* (combination 3) test results from either of the two tests (see Table 7-1 and Table 7-2) (Thompson 2003). As it will be shown later in this chapter, this choice depends on the objective of the strategy.

An alternative is to believe a third test, which may be a gold standard or another index test. However, this would still require the estimation of the conditional accuracy of the third test. At the moment only combinations of two tests will be considered.

It may be noted that the between test positivity criteria in the third point of the list above are the only ones that make sense in this case. *Believe the negative* criterion states that if one of the two tests is negative also the strategy is negative. *Believe the positive* (BP) criterion states that if one of the two tests is positive then the strategy is positive.

In case of *believe the negative* criterion, every time a patient is negative to one test, the diagnostician may ignore the other test result. Therefore, if the two tests are not performed simultaneously but one after the other, he may give the second test only to the proportion of patients that were positive to the first test (T1). This will lower the overall cost of the combination strategy since a lower number of tests will be given, and its accuracy will remain unvaried. In case of uncomfortable tests (i.e. invasive tests, tests that involve travelling to a hospital, time consuming tests, etc) there may also be advantage in the quality of life impact and safety of the testing procedure. A similar argument exists for the believe the positives strategy.

In the following section, all the possible combinations/sequences of two dichotomised diagnostic tests and issues related to the calculation of their accuracy (i.e. dependence between tests) will be explored.

T1	+	+	-	-
Logic operator for the combination	and	and	and	and
T2	+	-	+	-
Combination 1: Believe test 1	+	+	-	-
Combination 2: Believe the positive	+	+	+	-
Combination 3: Believe the negative	+	-	-	-

Table 7-1 Possible combinations of test results from two dichotomised diagnostic tests.

	Between tests positivity Criterion	
Diagnostic strategy output	**Believe the Negative**	**Believe the positive**
Positive	T1+ *and* T2+	T1+ *or* T2+
Negative	T1- *or* T2-	T1- *and* T2-

Table 7-2 How to combine test results when one of the two positivity criteria is used.

7.3.2 Sequences of two diagnostic tests

As already discussed in the previous section, a diagnostic strategy of two tests is a combination of tests where one is given after the other according to a between tests positivity criterion, for example *believe the negative*. Thus, the words *sequence* or *combination* of two tests may be used interchangeably if a between test positivity criterion is specified to complete the description of the combination and therefore the definition of the sequence.

Given a between test positivity criteria, the overall accuracy of a sequence remains unchanged regardless of which test is used first (Thompson 2003). Thus, the order of a sequence of tests should only involve considerations about economics and quality if life. Hence, in this section T1 will usually be referred to as the first test and T2 the second test in the strategy. Table 7-3 shows the two

sequences using two generic tests T_1 and T_2 with *believe the negative* between tests negativity criterion. The following notation will be used: $(T1 \text{ and } T2)_{BN}$ to identify the sequence where T_1 is given first and T_2 after, and are combined believing the negative test result (Table 7-3 A). The accuracy of this strategy is equivalent to the accuracy of $(T2 \text{ and } T1)_{BN}$ (Table 7-3 B). In both cases, the sensitivity of the sequence will be equal to $sensitivity = a/(a + c + e) = a/(a + g + i)$. Since the number of diseased and non diseased people per study does not change, $c + e = g + i$ and, therefore, $d + f = h + j$. Similarly, the specificity of the sequence will be $specificity = (d + f)/(b + d + f) = (h + j)/(b + h + j)$. Table 7-3 shows that, although the overall accuracy does not change, part of the conditional accuracy of the second test in the sequence does change. For example,

$$d/(b + d) = spec_{T2|T1+} = P(T2 - |T1+, non\ diseased)$$

$$\neq$$

$$h/(h + b) = spec_{T1|T2+} = P(T1 - |T2+, non\ diseased)$$

These conditional probabilities will be at the basis of Chapter 8 for the economic evaluations of combinations/sequences of tests; for example to choose which of the two equivalent strategies is better. When a between test positivity criteria is used to combine two tests in a sequence where T1 is applied after T2, the sequence obtained by applying T2 after T1 has equivalent accuracy.

A	Description of the sequence $(T1 \text{ and } T2)_{BN}$			Disease status	
				+	-
Permutation 1	T1	+	T2 +	a	b
			T2 -	c	d
		-		e	f

B	Description of the sequence $(T2 \text{ and } T1)_{BN}$			Disease status	
				+	-
Permutation 2	T2	+	T1 +	a	b
			T1 -	g	h
		-		i	j

Table 7-3 Classification of patients by two sequences with equivalent accuracy of T1 and T2 when the negatives are believed.

Sensitivity and specificity of either of two strategies with equivalent accuracy (i.e. $(T1 \text{ and } T2)_{BN}$) is further described in the rest of this section. Similar results may be derived for the case of *believe the positives* positivity criterion.

The following formulae define the sensitivity (specificity) as a function of the sensitivity (specificity) of a test and the conditional sensitivity (specificity) of the other test given the first test (Thompson 2003):

$$sens(T1\ and\ T2)_{BN}$$

$$=$$

$$P(T1+\ and\ T2+|diseased)$$

$$=$$

$$P(T1+|diseased) * P(T2+|T1+\ and\ diseased)$$

;

$$spec(T1\ and\ T2)_{BN}$$

$$=$$

$$P(T1-\ or\ T2-|non\ diseased)$$

$$=$$

$$1 - P(T1+|non\ diseased) * P(T2+|T1+\ and\ non\ diseased)$$

Equations 7-1

For simplicity, specificity was obtained through the calculation of its complement, the false positive rate (1- specificity). Since probabilities are, by definition, numbers between zero and one, the multiplication between a pair of probabilities will be a number smaller than each factor. Consequently, the sensitivity of the

sequence will be lower that the conditional sensitivities of either test used in Equations 7-1 (i.e. the sensitivity of T_1 and the sensitivity of $T_2|T_1$ are both higher than the sensitivity of $(T1\ and\ T2)_{BN}$), conversely, the specificity will be higher (i.e. the specificity of T_1 and $T_2|T_1$ will be both lower than the specificity of $(T1\ and\ T2)_{BN}$). It can be said that under the assumptions presented in chapter 3 (diseased and healthy are normally distributed over the test results and diseased patients tend to have higher values of the tests), one can believe the negatives to increase the sensitivity (and decrease the specificity), or can believe the positives to increase the specificity (and decrease the sensitivity) (Pepe 2003). As generally happens with diagnostic individual tests (considered in Chapter 3), for sequences of tests, an increase in one accuracy parameter (i.e. sensitivity) will result in a penalization of the other one (i.e. specificity).

It has been mentioned above that the conditional accuracies of the tests involved in the sequence are important for the (economic) evaluation of the strategy. Let's consider the sensitivity of T_2 conditional to T_1 as in Equations 7-1: $P(T2+|T1+\ and\ diseased)$. If T_1 and T_2 have different costs, as it is likely in a real evaluation, this conditional accuracy would be useful for the evaluation of the overall cost of T_2 given after T_1 when applied to a population, that will differ from the cost of T_1 given after T_2. This is due to *i)* the cost and qualitative implication of the single tests may be different, and *ii)* the number of patients who take the second test.

Equation 7-2 gives an application of the Bayes theorem to calculate the conditional sensitivity of T2 conditional to T1 being positive given the sensitivity of T1 conditional to T2 being positive and the sensitivities of T1 and T2 individually. In the case where T1 is cheaper and less invasive than T2, then the strategy where T1 is performed first is dominant compared to the strategy where T1 is performed after. If this is not the case, the sensitivity calculated in Equation 7-2 facilitates the economic evaluation of the sequence where T2 is given first in the case where this was not directly evaluated (i.e. there is not data to estimate the accuracy of T1|T2+). Similar equations can be obtained for specificity, or conditional to the second test being negative.

$$sens_{T1|T2+} = sens_{T2|T1+} * \frac{sens_{T1}}{sens_{T2}}$$

Equation 7-2

In terms of probabilities, Equation 7-2 can be rewritten as:

$$P(T2 + |T1 + and\ diseased)$$

$$=$$

$$\frac{P(T1 + |T2 + and\ diseased)P(T2 +|diseased)}{P(T1 +|diseased)}$$

$$=$$

$$\frac{P(T1 + and\ T2 +|\ diseased)}{P(T1 +|diseased)}$$

As a consequence, applying some rule of probabilities, the relation between unconditional and conditional accuracies is described by the following formula:

$$spec_{T2}$$

$$=$$

$$P(T2-|non\ diseased)$$

$$=$$

$$P(T2-and\ T1+|non\ diseased)+P(T2-and\ T1-|non\ \ diseased)$$

$$=$$

$$P(T2-|\ T1+,non\ \ diseased)P(T1+|non\ diseased)$$

$$+$$

$$P(T2-|T1-,non\ diseased)P(T1-|non\ diseased)$$

$$=$$

$$spec_{T2|T1+}*(1-spec_{T1})+spec_{T2|T1-}*spec_{T1}$$

Equation 7-3

The formula above shows that the accuracy of the second test may be calculated as the sum of the accuracies conditional to the first test results, weighted by the accuracy of the first test. This formula will be adapted to the clinical problem presented in Chapter 8 and will be crucial for the development of the model, to allow the inclusion of different types of data into the same modeling framework.

It should be clear that the dependence of the tests is conditional to the disease status (conditional dependence) as indicated in section 7.2.

7.3.3 Clinical Scores

A clinical score is a combination of tests (often individually of weak discriminating value) often derived using regression analysis from datasets including individual characteristics. For example, every predictor may be given a score of 1 if present (positive) and 0 if absent (negative). In this case a test may also be the presence or absence of a condition (symptom). It may also happen that different values are given to different test/questions (i.e. if an alternative diagnosis is possible, the score may be penalised). In this case the score by each test adds up to an overall score, and the overall combination is considered positive if the overall score is above a certain threshold. This is the case where the tests are combined in a combination, not in a sequence. This is the case of very weak tests (i.e. those that are part of the score), which may be observing the presence/absence of a symptom. One of these is the Wells score (Wells, Owen et al. 2006). This test has already been described in section 1.4 and will be described in more detail in next section.

Finally, it is possible to combine clinical scores with other tests although clinical scores can already be thought of as combinations of a number of tests where the accuracy of each component is not directly measured. For example, in a clinical score the individual tests are combined according to the scheme Majority (see

section 7.2.2 under "*Combination schemes and conditional accuracy*" for details on this combination scheme). For example, Wells score is a clinical score where the Majority scheme is applied (if WS>2 then positive).

7.4 Summary

This chapter has explored the characteristics of combinations of dichotomised diagnostic tests and the basic algebra behind the accuracy of diagnostic sequences. This exploration has been very useful to understand the issues behind combinations of diagnostic tools and has produced some algebra that will be used in the next chapter when a synthesis model to estimate the accuracy of combinations/sequences of tests will be developed.

When two dichotomous tests are considered, it is possible to build only 2 pairs of clinically relevant diagnostic strategies in which both tests are used simultaneously regardless of which test is used first. Given the tests to combine and the between tests positivity criterion, the issue of which test is used first concerns more the economic evaluation of the diagnosis, since the two permutations of the tests into the strategy are characterised by the same overall accuracy of the strategy. When three or more tests are used the evaluation becomes much more complex because of the number and types of combinations that can be obtained.

It is assumed that test results are dichotomised into positive and negative in case of continuous tests, which are usually reported as such and data for meta-analysis are often in this format. The case of naturally dichotomous tests is very rare, but it is very common to report test results at a given threshold. In fact, although a true threshold does not exist, variation in the implicit threshold still causes correlation between accuracy rates (Whiting 2008). The possibility of considering changes in

the threshold in either test exists but it is not considered here; tests are considered at their operative thresholds (i.e. threshold suggested by the producer) and the aim is not to find the best threshold but to evaluate tests at their operative threshold.

In conclusion, considerations about test strategies with more than two tests are not much different than considering two-test strategies, although the extent of difficulty is much higher as the number of tests increases. According to the objective of the strategy (i.e. to have the best sensitivity, or specificity, to have a non expensive and non invasive specific test, etc), the first step would be to choose the first test according to its characteristics, then add a second test in order to maximise the sensitivity (or specificity) of the strategy. After the best second test has been placed into the strategy, eventually, a best third test may be placed into the strategy in the best position. For this reason, sequencing should be regarded as the best way of combining tests (Pepe 2003) with the exception of building clinical scores (i.e. very low accuracy tests which borrow strength from each other under the combination scheme Majority). The next chapter will present the methodological development and practical application of meta-analysis techniques to the accuracy of combinations of Wells score and Ddimer for DVT.

Chapter 8. Meta-analysis and cost effectiveness analysis of the diagnostic accuracy of sequences of tests accounting for dependency between tests

8.1 Chapter overview

The review of HTA reports presented in Chapter 6 (Novielli, Cooper et al. 2010), identified that few reports evaluated combinations of diagnostic tests and, those that did, assumed independence between the tests. This review provided the motivation for this chapter, to develop statistical models to correctly evaluate the accuracy of a combination of tests (i.e. allowing for the correlation between the tests). Chapter 7 introduced some important features of combinations of diagnostic tests and of their accuracy and provided the basis for the statistical modeling approach developed in this chapter. This approach was developed within a Bayesian framework and incorporated multiple components (Ades and Cliffe 2002) to evaluate the accuracy of combinations of tests. The approach is applied to estimate the accuracy of Ddimer (DD) and Wells score (WS) in combination for Deep Vein Thrombosis (DVT – see section 1.4 for an introduction to this example) and the cost-effectiveness assessed by incorporating the results into an existing comprehensive decision analytic model (Sutton, Cooper et al. 2008). The results of this analysis (i.e. allowing for correlations between the tests) are compared with those obtained when independence between tests is assumed.

This chapter starts with a systematic review to identify studies reporting the accuracy of DD and WS for DVT in combination (sections 8.2, 8.3 and 8.4). These studies are then classified according to the types of data extractable from the studies (section 8.4). The data also include the accuracy of WS and DD used individually, which have been taken from two recent systematic reviews (section 8.4). The modeling approach used to meta-analyse these data, and the relation between the parameters estimated directly from the data and those representing the accuracy of different combinations of the two tests are presented (section 8.7). Finally, the results and the implications of such a modeling approach are described (section 8.8).

8.2 Systematic review of the conditional accuracy of DD given WS

The estimation of the accuracy of a combination of tests is based on the estimation of the conditional accuracy of either test given the other. Therefore, a systematic review of the literature must be conducted to identify those publications reporting such accuracy data. In this chapter, the focus is on the accuracy of DD and WS used in combinations. The data for the accuracy of DD and WS used individually were also used and were available from two recent systematic reviews; data are presented in section 8.4. First, relevant publications were identified from the Goodacre et al HTA report (Goodacre, Sampson et al. 2005; Goodacre, Sutton et al. 2005); that is, publications that provided data for both DD and WS were scrutinised for relevant data on the accuracy of combinations of DD and WS. These publications were then used to develop, test and improve the search strategy described below. One of the main problems when developing the search strategy was how to define WS which has been described using a number of very generic terms (i.e. clinical probability score, clinical assessment, etc) which may be attributed to a number of other meanings.

The final search strategy used for the accuracy of DD and WS used in combinations is provided in details in Appendix A. The search strategy had to be organised around the intersection of four different domains (see Figure 8-2):

- The disease: Deep vein thrombosis
- The first test: WS

- The second test: DD
- The objective of the study: Evaluation of the accuracy of diagnostic tests

One-thousand and twenty articles were identified as potentially relevant to the review via the search engine OVID XP, excluding review papers and before applying the inclusion and exclusion criteria.

The data of the accuracy of DD and WS used individually were also included into the analysis. Normally, the systematic review should have been set to retrieve the publications reporting also the accuracy of DD and WS used individually. Because the focus of this chapter is more methodological than clinical, such accuracy data were extracted from two existing systematic reviews for WS (Goodacre, Sutton et al. 2005) (updated with study T33 in Appendix B) and DD (Goodacre, Sampson et al. 2005) respectively. More details are given in section 8.4 and Figure 8-1, and references are given in Appendix B.

8.3 Reasons for exclusion and inclusion criteria of studies

Eight hundred and sixty-nine out of 1020 articles were excluded after a review of their titles/abstract because they were irrelevant to our analysis (e.g. mainly because they were presenting evaluations of the efficacy of clinical interventions, or the disease of interest was not DVT). Thus, the remaining 152 articles focused on the diagnosis of DVT (Figure 8-1). A further 125 articles were also excluded based on titles and abstracts for the following reasons:

- **39** were reviews not excluded by the filter
- **14** were guidelines or methodological publications
- **25** were not considering either WS score or DD test in their analysis.
- **26** were impossible to extract data
- Others as listed below Figure 8-1.

A complete account of the reasons for exclusions is given in Figure 8-1. This left only 26 studies evaluating the accuracy of DD given WS. Five of these studies dichotomised WS into:

- likely if WS>=2
- Unlikely if WS <=1.

As this is an uncommon representation of WS, these 5 studies were excluded from the review leaving a final sample of 21 studies.

After the exclusion of non relevant studies a total of 21 were potentially included in the analysis. These remaining 21 studies all considered a threefold categorization of WS:

- low pre test probability of DVT if WS<1
- moderate pre test probability of DVT if WS = 1 or 2
- high pre test probability of DVT if WS>2

The final inclusion criteria were: i) the publications reported the accuracy of DD and WS when used simultaneously on the same population of symptomatic patients; ii) the accuracy data was reported either in terms of count data (number of true positives, false negatives and/or number of true negatives and false positives) or in terms of proportions and a measure of statistical error (i.e. standard error or variance); iii) the aim of the publication was not to maximize the accuracy of DD by determining the best diagnostic threshold (i.e. the aim was to measure the accuracy of DD at the operative threshold as suggested by the manufacturer).

One study was excluded because the reporting of the data presented some incoherence (i.e. the data reported in the result section did not correspond to the data reported in the tables; it was classified as type I - see Appendix C). Two studies were also excluded because they reported the accuracy of DD at the threshold maximising the sensitivity; classified as type F, G and H in Appendix C (section 8.4 for further details).

A total of 18 studies were finally included in the meta-analysis (they are described in detail in section 8.4 and are classified as type A, B and C). The meta-analysis included also data on the accuracy of WS (classified as type D) and DD (classified as type E) used individually which were taken from systematic reviews recently conducted from other authors (see section 4.8 for more details).

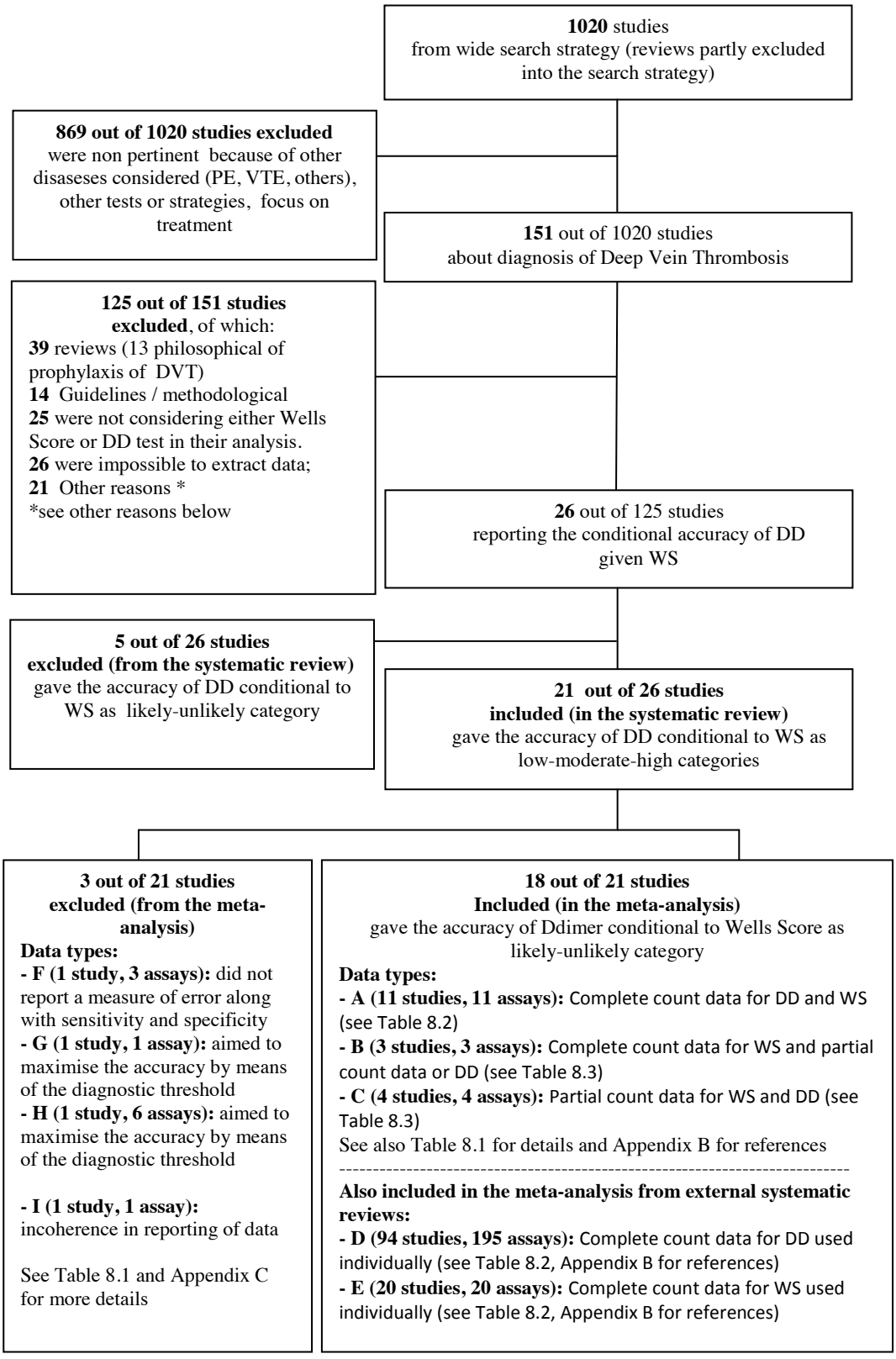


Figure 8-1 Flow chart of the studies excluded/included in the systematic review and meta-analysis.

[continued]

(*Other reasons:

3 studies analysed the agreement between DD assays; **1** study was based on a single observation; **3** were commentaries to other studies; **5** were a duplicate already retrieved, One of these was actually using a subset of data used in another article (Anderson, Wells et al. 2000); **2** based on different populations than symptomatic patients (colorectal cancer patients, long air-travellers);**1** unclear threshold for WS (if WS<2 then negative; if WS>=3 then positive; if WS=2 then ?); **1** reanalysed data from another study; **1** the WS was part of the GS; **1** multiple threshold, difficult to chose one; **1** analysed 398 legs of 343 patients, our analysis is based on the number of patients diagnosed, rather than on the number of legs examined; **1** performed a health economic evaluation and the accuracy data derived from another study; **1** analysed the accuracy of WS given DD negative.)

8.4 Description of the data available from the systematic review and data on the accuracy of WS and DD used alone

Different types of data were extracted from the systematic review, each informing a different set of parameters among those described in sections 8.5 and 8.6. The modeling approach that will be developed is a multi-component model. The main advantage of using a multi-component model is the use of all data through the definition of different likelihoods (components). The data can be represented in three different categories:

- Data on the accuracy of WS and DD when used together (from the systematic review presented in section 8.2 and 8.3).
- Data on the accuracy of WS when used individually (Update of an existing review (Goodacre, Sutton et al. 2005))
- Data on the accuracy of DD when used individually (An existing review(Goodacre, Sampson et al. 2005))

The types of data collected by the systematic review are described in Table 8-1. This data could be divided into nine different groups according to the type (counts or proportions), the aim (analyse the accuracy of the test at the usual threshold or maximizing the accuracy via an optimum threshold), level of missing information. Data A (Table 8-2), B and C (Table 8-3) measure the accuracy of WS and DD|WS and have been included in the analysis; these are 18 of the 21 potentially included studies indicated in Figure 8-1, that also met the three final inclusion criteria specified in section 8.3. Type D and E (Appendix B for list of references) measure

the accuracy of WS alone (T19 to T36 in Appendix B) and DD alone (T37 to T133 in Appendix B) respectively. Type F, G and H were excluded from the systematic review because did not meet the three final inclusion criteria specified in section 8.3. One study (type I) was excluded for poor quality of reporting of the data (i.e. data reported in the main body of the article were inconsistent with data reported in the table). The description of the included data form the systematic review (type A, B and C), of the data on the tests used alone (type D and E) and the reasons for exclusion of data F, G, H and I are explained in detail below (references for the latter are available in Appendix C, and accuracy data are presented in Table E1 Table E2 in Appendix C).

Type of data	Description	Included / Excluded	Aim of the study
A	Complete count data for DD and WS	Included	To measure the accuracy of DD after WS
B	Complete count data for WS and partial count data or DD	Included	To measure the accuracy of DD after WS
C	Partial count data for WS and DD	Included	To measure the accuracy of DD after WS
D	Complete count data for WS	Included	To measure the accuracy of WS
E	Complete count data for DD	Included	To measure the accuracy of DD
F	Proportions for DD (**no standard errors**) and complete count data for WS	Excluded	To measure the accuracy of DD after WS
G	Proportions for DD (no standard errors) and complete count data for WS	Excluded	**To maximize the accuracy of DD via optimum threshold**
H	Proportions for DD (with confidence intervals) and complete count data for WS	Excluded	**To maximize the accuracy of DD via optimum threshold**
I	Partial count data for WS and DD (**poor data reporting**)	Excluded	To measure the accuracy of DD after WS

Table 8-1 Classification and description of the types of data, and inclusion exclusion criteria (in bold). Main reasons for exclusion are indicated in bold.

Type A: *Complete count data of the conditional accuracy of DD given WS* (i.e. $TP_{ij}, FP_{ij}, FN_{ij}, TN_{ij}$ from study i and WS category j), *and complete count data for all categories of WS* (i.e. number of diseased d_{ij} and healthy h_{ij} patients in each category, where for example $d_{ij} = TP_{ij} + FN_{ij}$ and $h_{ij} = FP_{ij} + TN_{ij}$). This data is the best data that could be extracted: It is complete (no missing bits of information) and gives full information on the accuracy of WS (intermediate parameters p_{dj} and p_{hj} as will be specified in section 8.6 and 8.7), of the accuracy of DD given a category of WS (intermediate parameters $sens_{DD|.}$ and $spec_{DD|.}$) and therefore of DD alone. Table 8-2 lists this data type and full references are available in Appendix B.

Type B: *Complete count data of the conditional accuracy of DD given WS for one or two categories of WS* ($TP_{ij}, FP_{ij}, FN_{ij}, TN_{ij}$ for some j), *and complete count data for all categories of WS* (number of diseased $d_{ij}$and healthy h_{ij} patients for some categories of WS j). This data contributes to the estimation of the accuracy of DD given WS but not for all categories (i.e. $sens_{DD|WS\ low}$ and $spec_{DD|WS\ low}$ cannot be estimated if data are available only for WS high and moderate), and to the estimation of the accuracy of WS in terms of the proportion of diseased/healthy per some WS category (p_{dj} and p_{hj} for all j). Table 8-3 lists data type B and full references are available in Appendix B

Type C: *Complete count data of the conditional accuracy of DD given WS for one or two categories of WS* ($TP_{ij}, FP_{ij}, FN_{ij}, TN_{ij}$ for some j), *and, for the same categories, count data for the proportion of diseased and healthy patients*

(number of diseased $d_{ij}$and healthy h_{ij} patients for the same j's). This data contributes to the estimation of $sens_{DD|.}$ and $spec_{DD|.}$ and p_{dj} and p_{hj}. for some j's . Table 8-3 also lists data type C and references are available in Appendix B.

Type D: *Complete data for the accuracy of WS alone*. Since this project is methodological and aims to create a modeling framework that achieves the inclusion of all data available, a systematic review on the accuracy WS has not been performed directly, instead data from a recent existing systematic reviews has been used (Goodacre, Sutton et al. 2005) and an updated version of the dataset was obtained from the authors (Appendix B, study T33 has been added to the original set of articles). Twenty-three publications were included in our dataset in order to inform the accuracy of WS (i.e. contribute to the estimation of p_{dj} and p_{hj} for all j). Among these 23 studies, 3 have been already considered in the systematic review of the conditional accuracy of DD given WS presented in section 8.2 (i.e. are already considered among data type A, B). Therefore, they have been excluded from this dataset in order not to avoid duplication of the data.

Type E: *Complete data for the unconditional accuracy of DD*. The data presented and already analysed in Chapter 5 were included (Goodacre, Sampson et al. 2005), however this time all 198 assays have been used since covariate effect is not considered in this chapter. This data would concur indirectly to the estimation of the accuracy of DD conditional on WS level. Via the definition of the functional relation between the unconditional accuracy of DD and the accuracies of WS and DD|WS, the information contained in this data should contribute to the

estimation of the final parameters. For the same reason as for WS data type D, three studies have been excluded because they contain type A data. Thus, 195 assays extracted from 94 studies giving data on the accuracy of DD alone were included in our analysis.

Type F: *conditional sensitivity and specificity of DD given WS for one or two categories of WS, and, for the same categories, complete count data for the accuracy of WS*. This data come from one single study that evaluated, in particular, the accuracy of DD given moderate WS. This have not been included into the analysis because: i) When trying to calculate $TP_{ij}, FP_{ij}, FN_{ij}, TN_{ij}$ for this study the results were not exact (i.e. 20.45, the rounding was not clear), ii) they did not report the standard error of the estimated sensitivity and specificity.

Type G: *Same as type F, but the accuracy of DD|WS was maximised by choosing an ad hoc threshold* rather than using the threshold suggested by the manufacturer. This study was excluded for the reasons (i) and (ii) explained above and also because the accuracy given an ad-hoc threshold would have inflated one accuracy rate and deflated the other (iii). The main focus was on operative thresholds (i.e. used in real practice as those recommended by the manufacturer).

Type H: *Conditional sensitivity and specificity and confidence intervals of DD (at the best threshold to maximise sensitivity) given WS for all categories of WS, and number or people in each category of Wells per diseased and healthy*. Potentially, this study could be included by calculating the number of $TP_{ij}, FP_{ij}, FN_{ij}, TN_{ij}$ for each assay, however it was excluded for the same reason (*iii*) explained above

(DD assays in this study are evaluated on the basis of an ad-hoc threshold to maximise sensitivity).

Type I: *Complete count data of the conditional accuracy of DD given WS for one or two categories of WS* ($TP_{ij}, FP_{ij}, FN_{ij}, TN_{ij}$ for some j), *and, for the same categories, count data for the proportion of diseased and healthy patients* (number of diseased $d_{ij}$and healthy h_{ij} patients for the same j's). This data was relatifve to one study (T136, reference and data in Table E2 available in appendic C) and was excluded for poor and incoherent reporting. Moreover, it was not clear why diseased patients were not reported. This data was considered poor and unclear. This data would have been categorised as type C in the case it was not excluded.

First Author	Ddimer assay	Ddimer Accuracy data				Wells score	
		TP	FP	FN	TN	WS level	(Disased/ Total)
Type A studies							
T1. Shields 2002	SimpliRED	1	8	0	32	low	1/41
		6	18	0	20	moderate	6/44
		8	2	2	5	High	10/17
T2. Lennox 1999	SimpliRED	3	8	1	76	low	4/45
		9	12	3	43	moderate	12/67
		30	8	0	7	High	30/88
T3. Kearon 2001	SimpliRED	4	25	1	176	low	5/49
		17	51	7	113	moderate	24/188
		33	8	2	6	High	35/206
T4. Ruiz-Gimenez 2004	Rapid ELISA	15	49	1	70	low	16/145
		31	51	0	54	moderate	30/144
		54	36	1	21	High	50/112
T5. Yamaki 2005	ELISA	1	20	0	17	low	1/38
		22	23	0	19	moderate	22/64
		35	9	0	12	High	35/56
T6. Anderson 2000	SimpliRED	4	17	0	97	low	21/118
		6	9	3	48	moderate	15/66
		13	2	2	13	High	15/30
T7. Anderson 2002	SimpliRED	17	113	3	313	low	20/446
		61	93	15	23	moderate	76/192
		79	55	15	50	High	94/199
T8. Bates 2003	LATEX	18	85	0	193	low	18/296
		16	83	1	89	moderate	17/189
		21	30	0	20	High	21/71
T9. Rio Sola 1999	LATEX	23	1	5	3	low	28/32
		37	6	7	5	moderate	44/55
		9	3	0	2	high	9/14
T10. Williams 2005	Rapid ELISA	6	42	0	41	low	6/89
		15	59	3	46	moderate	18/123
		10	16	1	4	high	11/31
T11. Yamaki 2009	LATEX	28	233	1	243	low	29/508
		117	104	0	16	moderate	118/237
		109	29	0	3	high	109/141

Table 8-2 Data extracted by the systematic review of WS and DD used in combination; type A (complete data for both tests).

Study, Author and year of publication	Ddimer Accuracy data					Wells score	
	Ddimer Assay	TP	FP	FN	TN	WS level	(Disased/ Total)
Type B studies							
T12. Borg 1997	Rap. ELISA	NA	NA	NA	NA	low	02/32
		NA	NA	NA	NA	moderate	04/15
		25	2	1	1	high	26/29
T13. Dewar 2008	LATEX	9	70	0	87	low	9/166
		NA	NA	NA	NA	moderate	17/161
		NA	NA	NA	NA	high	30/108
T14. Elf 2008	**Unclear**	12	37	1	109	low	14/159
		NA	NA	NA	NA	moderate	37/141
		NA	NA	NA	NA	high	33/57
Type C studies							
T15. Aguilar-Franco 2002a	LATEX	2	76	0	71	low	2/149
		NA	NA	NA	NA	moderate	NA/NA
		NA	NA	NA	NA	high	NA/NA
T16. Walsh 2009	Rap. ELISA	4	23	0	22	low	04/49
		NA	NA	NA	NA	moderate	NA/NA
		NA	NA	NA	NA	high	NA/NA
T17. Aguilar-Franco 2002b	LATEX	NA	NA	NA	NA	low	NA/NA
		26	73	0	35	moderate	26/134
		NA	NA	NA	NA	high	NA/NA
T18. Bucek 2002	**Unclear**	2	43	0	48	low	02/93
		NA	NA	NA	NA	moderate	NA/NA
		NA	NA	NA	NA	high	NA/NA

Table 8-3 Data extracted by the systematic review of WS and DD used in combination; type B (partial data for DD), type C (partial data for DD and WS).

In total, 233 assays extracted from 132 studies were included in the analysis. It is quite a large number of studies for a meta-analysis. Studies maximising accuracy via *ad hoc* threshold selection, if included in the analysis, could inflate the sensitivity (or specificity) of DD.

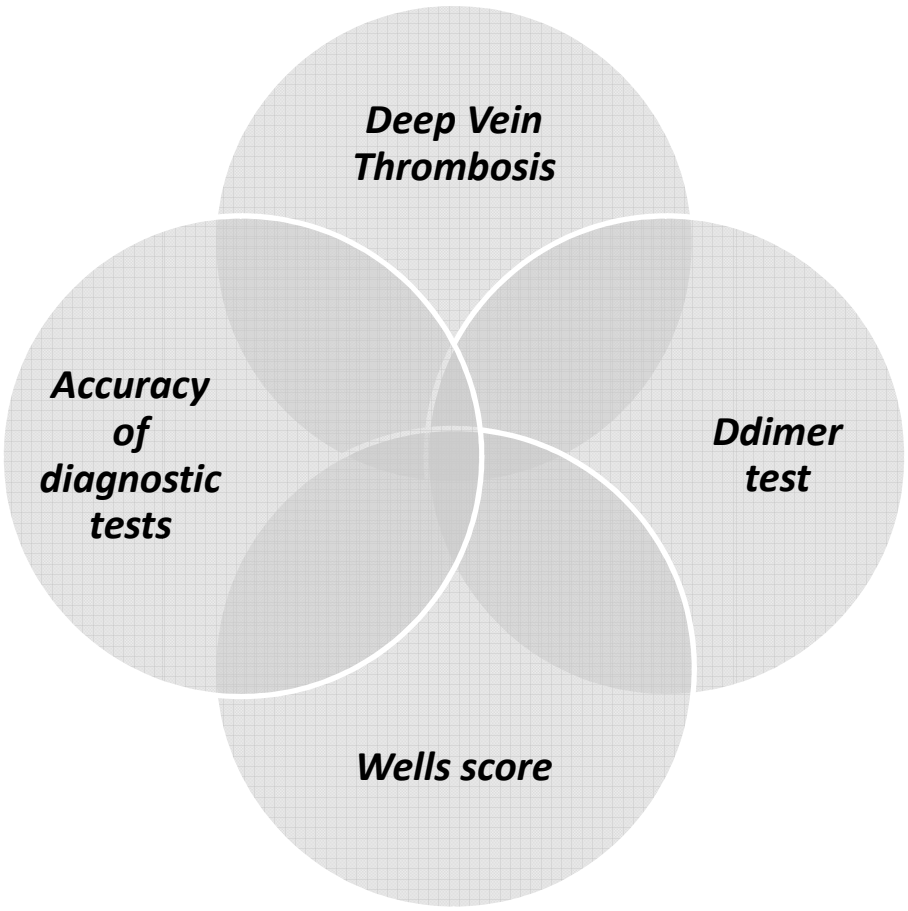
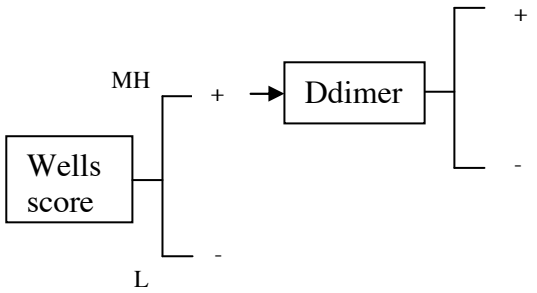


Figure 8-2. The search strategy to detect the studies for the accuracy of DD and WS in combination consisted of four overlapping domains.

8.5 Diagnostic strategies under evaluation

The review highlighted that WS has always been used as a pre-test probability score. WS can be considered as a "proper"diagnostic test (i.e. used to diagnose rather than to stratify patients into groups with more homogeneous prevalence of disease) with two thresholds: i) low vs moderate/high (i.e. a patient is negative if low WS, positive if moderate or high), and ii) low/moderate vs high (i.e. a patient is negative if low or moderate WS, positive if high). Although some diagnostic algorithms described above use WS as a "proper" test, these strategies have never been implemented in practice. Thus, the final parameters which are the focus of the analysis are sensitivity and specificity of the possible diagnostic strategies including WS and DD. All 8 options are described below.

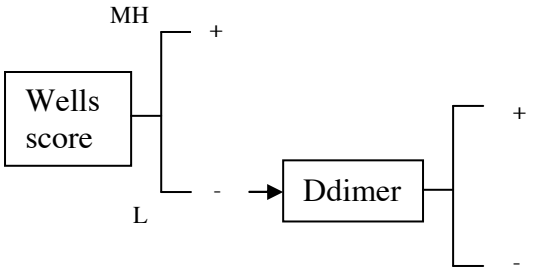
- Strategy 1: WS at threshold 1 (low vs moderate/high), believe the negatives



$$sens_{(WS_{t1}\,and\,DD)_{BN}} = P(WS\ moderate\ or\ high\ and\ DD\ +|diseased)$$

$$spec_{(WS_{t1}and\ DD)_{BN}} = 1 - P(WS\ moderate\ or\ high\ and\ DD\ +|healthy)$$

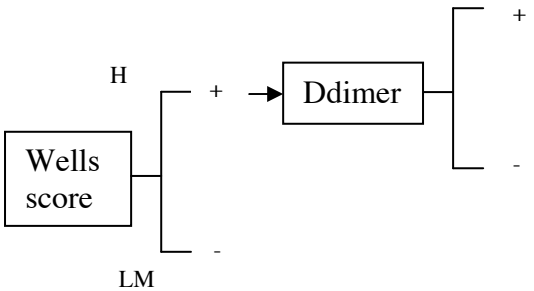
- Strategy 2:WS at threshold 1 (low vs moderate/high), believe the positives



$$sens_{(WS_{t1}and\ DD)_{BP}} = 1 - P(WS\ low\ and\ DD\ -|diseased)$$

$$spec_{(WS_{t1}and\ DD)_{BP}} = P(WS\ low\ and\ DD\ -|healthy)$$

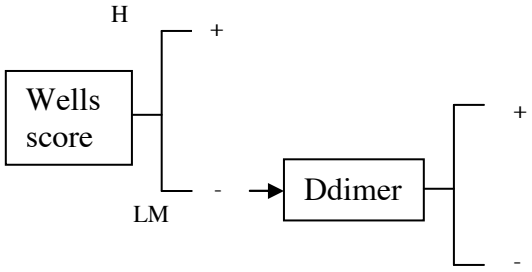
- Strategy 3:WS at threshold 2 (low/moderate vs high), believe the negatives



$$sens_{(WS_{t2}and\ DD)_{BN}} = P(WS\ high\ and\ DD\ +|diseased)$$

$$spec_{(WS_{t2}and\ DD)_{BN}} = 1 - P(WS\ high\ and\ DD\ +|healthy)$$

- Strategy 4:WS at threshold 2 (low/moderate vs high), believe the positives

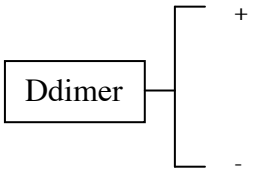


$$sens_{(WS_{t2}and\ DD)_{BP}} = 1 - P(WS\ low\ or\ moderate\ and\ DD - |diseased)$$

$$spec_{(WS_{t2}and\ DD)_{BP}} = P(WS\ low\ or\ moderate\ and\ DD - |healthy)$$

These need then to be compared to the tests when used alone, they are:

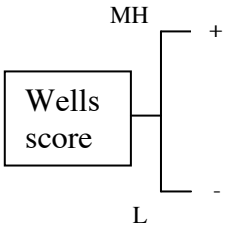
- Strategy 5:DD when used alone



$$sens_{DD} = P(DD\ + |diseased)$$

$$spec_{DD} = P(DD - |healthy)$$

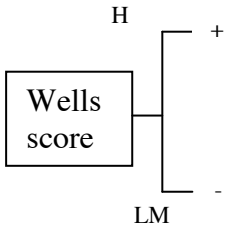
- Strategy 6: WS used alone at the first threshold (low vs moderate/high)



$$sens_{WS_{t1}} = P(WS\ moderate\ or\ high \mid diseased)$$

$$spec_{WS_{t1}} = P(WS\ low \mid healthy)$$

- Strategy 7: WS used alone at the second threshold (low/moderate vs high)

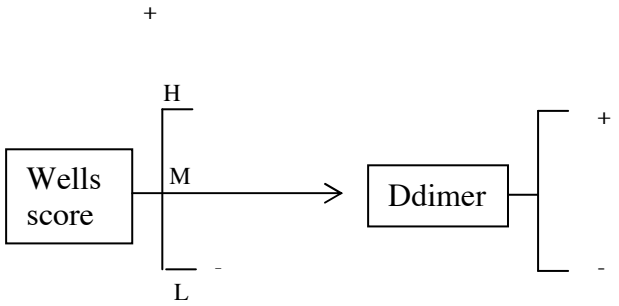


$$sens_{WS_{t2}} = P(WS\ high \mid diseased)$$

$$spec_{WS_{t2}} = P(WS\ low\ or\ moderate \mid healthy)$$

- Strategy 8: WS to every patient, if high then treat, if low then discharge, if moderate then further test with DD. This is not the only further option to combine the two tests, others can be identified. However, the sequences presented above represent coherent and sufficient choices to maximize either sensitivity or specificity (Pepe 2003) and this combination, included

as an example, is not expected to be more accurate in both parameters nor economically better (see section 8.9).



$$sens_{if\ WS\ mod \rightarrow DD}$$

$$= P(WS\ high | diseased)$$

$$+ P(WS\ moderate\ and\ DD + | diseased)$$

$$spec_{if\ WS\ mod \rightarrow DD}$$

$$= P(WS\ low \mid healthy)$$

$$+\ P(WS\ moderate\ and\ DD - \mid healthy)$$

8.6 Parameters to be estimated by the models

The accuracy of each strategy described above can be calculated as a function of the accuracy of WS and the conditional accuracy of DD|WS. Thus, the parameters of interest are:

i. the accuracy rates of the four sequences (i.e. $sens_{(WS_{t1}and\ DD)_{BN}}$, $sens_{(WS_{t1}and\ DD)_{BP}}$, $sens_{(WS_{t2}and\ DD)_{BN}}$, $sens_{(WS_{t2}and\ DD)_{BP}}$ and $spec_{(WS_{t1}and\ DD)_{BN}}$, $spec_{(WS_{t1}and\ DD)_{BP}}$, $spec_{(WS_{t2}and\ DD)_{BN}}$, $spec_{(WS_{t2}and\ DD)_{BP}}$),

ii. the accuracy rates of the tests used alone ($sens_{DD}$, $sens_{WS_{t1}}$, $sens_{WS_{t2}}$ and $spec_{DD}$, $spec_{WS_{t1}}$, $spec_{WS_{t2}}$,)

iii. the accuracy of the combination described as strategy 8 in section 8.5above ($sens_{if\ WS\ mod \rightarrow DD}$ and $spec_{if\ WS\ mod \rightarrow DD}$).

According to Equation 7-3 in Chapter 7, these will be calculated as a function of WS accuracy parameters (p_{Dk} proportion of diseased patients in category (k) 1=low, 2=moderate, 3=high; and p_{Hk} proportion of healthy patients in WS category (k) 1=low, 2=moderate, 3=high) and DD conditional on WS accuracy parameters ($sens_1$, $sens_2$, $sens_3$, and $spec_1$, $spec_2$, and $spec_3$ as defined in Figure 8-3).

The process used to describe these relations has been used in the past by Ades *et al* (Ades and Cliffe 2002) for the assessment of the effectiveness of an intervention for HIV. This approach is also called a shared component model as it

contains equations with shared parameters, where two models (one for the accuracy of WS data and the other for the accuracy of DD|WS data) are linked by using separate equations but using parameters in common (Knorr-Held and Best 1999). Figure 8-3 shows this process for our example. The data are collected in order to estimate the values of the parameters of interest (final parameters). However the model does not estimate them directly, but these are expressed as functions of intermediate parameters. These intermediate parameters are directly estimated by the model, and the uncertainty in their estimates is propagated into the final parameters. Thus, the relationship between the data and the final parameters is explained in two stages, first the relation between the data and the intermediate parameters is presented via the description of the likelihoods, then, the relation between the intermediate parameters and the final parameters is given via the description of the formulae used to transform intermediate parameters into final parameters.

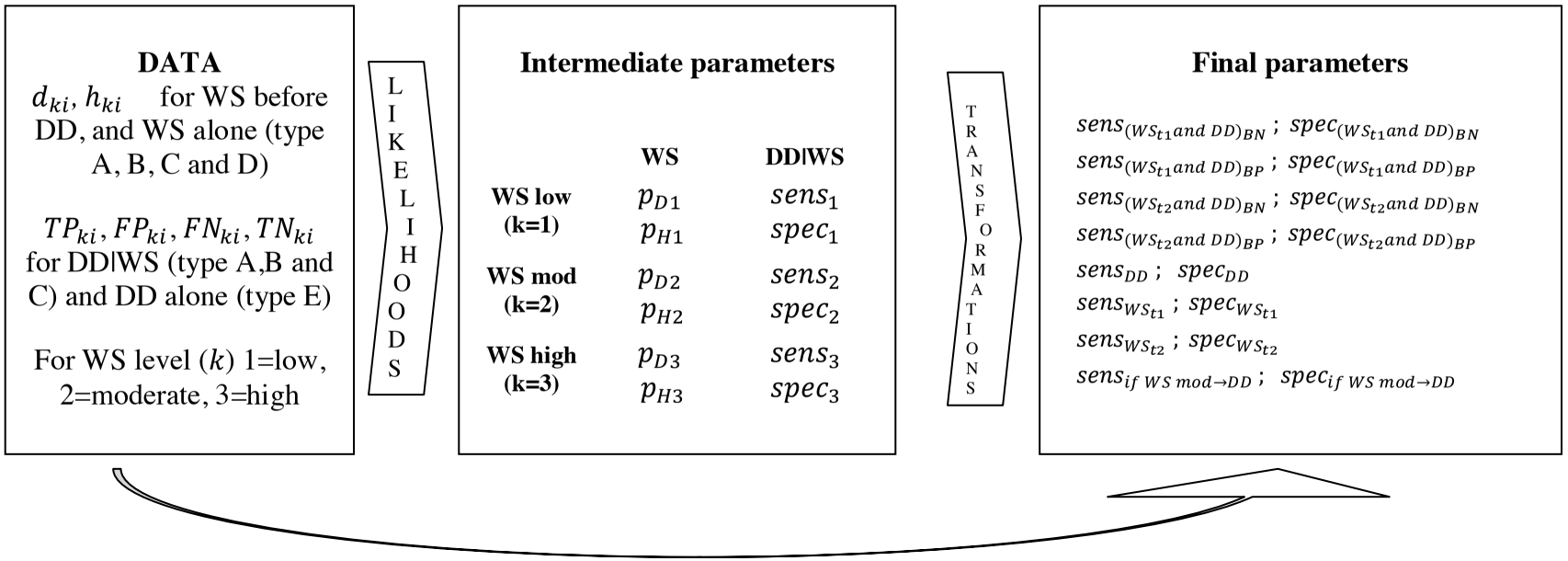


Figure 8-3 Process of description of the data, intermediate and final parameters.

8.7 Description of the model

In this section, the WS and DD relative components of the multi-component model will be described starting from the likelihoods used for each type of included data (type A, B, C, D and E). For WS, a multinomial random effect logit-model for the estimation of the proportion of diseased and healthy per each category has been used. For DD, the model component is based on the bivariate random effect approach (Model 4.9) presented in Chapter 4 and Chapter 5. These two models are linked together via equations that allow the missing bits of information on either test to be estimated and all types of data to be used simultaneously.

The proposed modeling approach has two types of component that interact together, one for the intermediate accuracy parameters of WS (see Table 8-4) and one of the accuracy intermediate parameters of DD|WS (see Table 8-5). The next section 8.7.1 describes how the data are linked to the intermediate parameters via the specification of the likelihoods. The model as implemented in WinBUGS is available in the folder "Chapter 8 - combinations of WS and DD for DVT", contained in the CD-ROM attached to this thesis, in the WinBUGS file "*model of data A B C D E - conditional accuracy and sequences.odc*". In the same folder the code for the implementation of the model that assumes independence between tests is given in the file "*model that assumes independence.odc*".

8.7.1 Definition of the first part of the model for the estimate of the intermediate parameters

A clear representation of the data by means of tables

Table 8-4 and Table 8-5 represent a generic complete data that can hypothetically be extracted for WS and DD respectively. Symbols are used instead of numbers in order to establish the relationships between the data in the two tables. The same symbols will be used throughout this chapter to describe the formulae that define the model (following in this section 8.7).

Study i	LOW	MOD	HIGH	TOTAL
Diseased	d_{1i}	d_{2i}	d_{3i}	N_{Di}
Healthy	h_{1i}	h_{2i}	h_{3i}	N_{Hi}

Table 8-4 Complete data that would be extracted for the accuracy of Wells score.

	WS levels			Overall accuracy of DDimer
Study i	LOW (k=1)	MOD (k=2)	HIGH (k=3)	
Sensitivity	$\frac{r_{D1i}}{d_{1i}}$	$\frac{r_{D2i}}{d_{2i}}$	$\frac{r_{D3i}}{d_{3i}}$	$\frac{\sum_k r_{Dki}}{N_{Di}}$
Specificity	$\frac{r_{H1i}}{h_{1i}}$	$\frac{r_{H2i}}{h_{2i}}$	$\frac{r_{H3i}}{h_{3i}}$	$\frac{\sum_k r_{Hik}}{N_{Hi}}$
$r_{Dki} = TP_{ki}$; $d_{1i} - r_{Dki} = FN_{ki}$; $r_{Hki} = TN_{ki}$; $h_{1i} - r_{Hki} = FP_{ki}$ for WS level (k) $1 = LOW, 2 = MOD, 3 = HIGH$, for study i				

Table 8-5 Complete data that would be extracted for the conditional accuracy of Ddimer.

In Table 8-4, d_{ki} and h_{ki} indicate the number of diseased and non diseased respectively, within WS level k (1=low, 2=moderate and 3=high) for study i; N_{Di} and N_{Hi} are the total number of diseased and non diseased for study i. Table 8-5 represents the data for the conditional and overall accuracy of DD: r_{Dki} is the number of diseased patients that are correctly classified as positive by DD, conditional to the k^{th} level of WS and for study i (true positive); r_{Hki} is the number of non diseased patients that are correctly classified as negative by DD, conditional to the k^{th} level of WS and for study i (true negative); d_{ki} and h_{ki} are already defined in Table 8-4 and represent the denominator of the fractions used to define sensitivity and specificity; the last column of Table 8-5 represents the overall sensitivity and specificity of DD as a function of the data on the left, it is clear that the overall accuracy of DD and WS is the sum of the numerators divided by the sum of the denominators of the conditional accuracies. This last relation will then lead to the formulation of the overall accuracy of DD as the sum of the accuracies of DD conditional to WS, weighted by the proportion of diseased/non diseased patients into WS categories (see Equation 8-10).

The different data types A, B, C, D and E can all be expressed by the symbols presented in the tables above. The number of studies will be indicated as n_A, n_B, n_C, n_D, n_E for data kinds A, B, C, D and E respectively.

Multinomial random effect logistic model for the meta-analysis of WS data

Given the multinomial nature of WS data, the multinomial logistic model should allow the threshold effect to be implicitly considered because the accuracy of WS is estimated for each of its two possible thresholds. However, meta-analyses are often characterised by a residual amount of unexplained heterogeneity, which is usually accounted for either using random effects or covariates where possible (Higgins, Thompson et al. 2009). Therefore, I have adapted the Bayesian fixed effect multinomial logistic model presented by Ntzoufras (Ntzoufras 2010) and obtained a Bayesian random effect multinomial logistic model.

For WS types A, B and D data, the likelihood is specified via multinomial distributions (see Equation 8.1) with parameters p_{Dki} (for diseased) and p_{Hki} (for healthy patients) for study i and WS level (k) 1=low, 2=moderate and 3=high. The order of the multinomial is the sum of diseased/healthy ($N_{Di} = \sum_{k=1}^{3} d_{ki}$ and $N_{Hi} = \sum_{k=1}^{3} h_{ki}$ as defined in Table 8-4).

$$(d_{1i}, d_{2i}, d_{3i}) \sim multinom((p_{D1i}, p_{D2i}, p_{D3i}); N_{Di})$$

$$(h_{1i}, h_{2i}, h_{3i}) \sim multinom\big((p_{H1i}, p_{H2i}, p_{H3i}); N_{Hi}\big)$$

$$for\ i\ from\ 1 \quad to \quad n_A + n_B\ (type\ A, B)$$

$$and\ for\ i\ from\ n_A + n_B + n_C + 1 \quad to \quad n_A + n_B + n_C + n_D\ (type\ D)$$

Equation 8-1

Type C data is incomplete for WS; thus, multinomial likelihoods cannot be used because the order of the multinomial would not be available (i.e. studies only reported the number of patients classified in some but not all levels of WS). This data does not contain information to estimate the parameters p_{Dki} or p_{Hki} , because the total number of diseased and non diseased is not available. However, type C studies give important information on the accuracy of DD given the reported levels of WS. Thus, WS category specific data can be included into the modeling approach by using a combination of binomial likelihoods as substitutive of the multinomial likelihoods by means of a constraint on the parameters (see Equations 8-2), and the assumption of exchangeability between studies (Bernardo and Smith 1994) which allows the model itself to inform the estimate of the missing data via the indirect estimation of the parameter for which there is not information.

For example, let's consider the case where only the number of diseased/healthy patients for low or moderate WS is available (d_{1i}, d_{2i} and h_{1i}, h_{2i}, see Equations 8-2(a)). The proportion of patients in such categories will be estimated by the model by means of Equation 8-1 based on the assumption of exchangeability between studies. This estimation will be based on the total number of diseased and healthy patients for type C studies that can be also estimated from the model via the assumption of exchangeability, for example, $\widehat{N}_{Di} = d_{ki}/p_{Dki}$ and $\widehat{N}_{Di} = h_{ki}/p_{Hki}$. Using these formulae, all the missing information for WS will be

estimated by the model and type C data will influence neither the estimate of the intermediate parameters nor the estimate of the final parameters for WS, but type C data for the conditional accuracy of DD will still be considered.

WS high (*k*=3) is missing	**WS moderate(*k*=2) is missing**	**WS low (*k*=1) is missing**
(a)	**(b)**	**(c)**
$d_{1i} \sim binomial(p_{D1i}, \widehat{N}_{Di})$	$d_{1i} \sim binomial(p_{D1i}, \widehat{N}_{Di})$	$d_{2i} \sim binomial(p_{D2i}, \widehat{N}_{Di})$
$h_{1i} \sim binomial(p_{H1i}, \widehat{N}_{Hi})$	$h_{1i} \sim binomial(p_{H1i}, \widehat{N}_{Hi})$	$h_{2i} \sim binomial(p_{H2i}, \widehat{N}_{Hi})$
$d_{2i} \sim binomial(p_{D2i}, \widehat{N}_{Di})$	$d_{3i} \sim binomial(p_{D3i}, \widehat{N}_{Di})$	$d_{3i} \sim binomial(p_{D3i}, \widehat{N}_{Di})$
$h_{2i} \sim binomial(p_{H2i}, \widehat{N}_{Hi})$	$h_{3i} \sim binomial(p_{H3i}, \widehat{N}_{Hi})$	$h_{3i} \sim binomial(p_{H3i}, \widehat{N}_{Hi})$
$p_{D3i} = 1 - p_{D1i} - p_{D2i}$	$p_{D2i} = 1 - p_{D1i} - p_{D3i}$	$p_{D1i} = 1 - p_{D2i} - p_{D3i}$
$p_{H3i} = 1 - p_{H1i} - p_{H2i}$	$p_{H2i} = 1 - p_{H1i} - p_{H3i}$	$p_{H1i} = 1 - p_{H2i} - p_{H3i}$

Equations 8-2

(The binomial likelihoods model with constraint on one proportion presented in Equation 8-2, as used for type C data, can actually be used on type A, B and D data, with the difference that for these the total number of diseased and non diseased is known, however resulting in more lines of code when implemented in WinBUGS.)

The multinomial logistic model is then completed by the specification of the between study variability structure and the logit transformations as specified in Equation 8-3.

$$p_{Dki} = \frac{\eta_{Dki}}{\sum_{k=1}^{3} \eta_{Dki}} \quad , \quad \xi_{Dki} = \ln(\eta_{Dki})$$

$$p_{Hki} = \frac{\eta_{Hki}}{\sum_{k=1}^{3} \eta_{Hki}} \quad , \quad \xi_{Hki} = \ln(\eta_{Hki})$$

$$\xi_{Dki} \sim norm(\xi_{Dk}, \sigma_D^2)$$

$$\xi_{Hki} \sim norm(\xi_{Hk}, \sigma_H^2)$$

$$for\ i\ from\ \ 1\ \ to\ \ n_A + n_B + n_C + n_D\ (type\ A, B, C\ and\ D)$$

$$for\ WS\ level\ (k)\ \ 1 = low,\ \ 2 = moderate\ \ and\ \ 3 = high$$

Equation 8-3

The parameters estimated by this meta-analytical model are the overall proportion of patients in each WS category for diseased and healthy patients, expressed in a multinomial logit scale in Equation 8-3 as ξ_{Dk} and ξ_{Hk}, for WS level (k) 1=low, 2=moderate and 3=high. These need to be back-transformed to obtain the overall proportions using Equation 8-4:

$$p_{D1}^{pooled} = \exp(\xi_{D1}) \Big/ \sum_{k=1}^{3} \exp(\xi_{Dk})$$

$$p_{D2}^{pooled} = \exp(\xi_{D2}) \Big/ \sum_{k=1}^{3} \exp(\xi_{Dk})$$

$$p_{D3}^{pooled} = \exp(\xi_{D3}) \Big/ \sum_{k=1}^{3} \exp(\xi_{Dk})$$

$$p_{H1}^{pooled} = \exp(\xi_{H1}) \Big/ \sum_{k=1}^{3} \exp(\xi_{Hk})$$

$$p_{H2}^{pooled} = \exp(\xi_{H2}) \Big/ \sum_{k=1}^{3} \exp(\xi_{Hk})$$

$$p_{H3}^{pooled} = \exp(\xi_{H3}) \Big/ \sum_{k=1}^{3} \exp(\xi_{Hk})$$

$$for\ i\ from\ \ 1\ \ \ to\ \ \ n_A + n_B + n_C + n_D\ (type\ A, B, C\ and\ D)$$

Equation 8-4

Where $\left(p_{D1}^{pooled}, p_{D2}^{pooled}, p_{D3}^{pooled}\right)$ is the vector of the pooled proportions (across studies) of diseased patients for WS equal to 1=low, 2=moderate and 3=high, where the study specific proportions that have been pooled have been estimated in Equation 8-1and Equations 8-2. Similarly $\left(p_{H1}^{pooled}, p_{H2}^{pooled}, p_{H3}^{pooled}\right)$ represents the vector of pooled proportions of healthy patients.

Such parameters refer to the generic data represented in Table 8-4 which represents study specific data. A similar table can also be used to represent such estimates (see Table 8-6).

	LOW	MOD	HIGH	TOTAL
Diseased	p_{D1}^{pooled}	p_{D2}^{pooled}	p_{D3}^{pooled}	1
Healthy	p_{H1}^{pooled}	p_{H2}^{pooled}	p_{H3}^{pooled}	1

Table 8-6 Generic pooled estimates of the proportions of diseased and health patients for WS categories.

Bivariate random effect logistic models for the meta-analysis of DD|WS data

Data on the conditional accuracy of DD given WS has been identified as type A, B and C. These also contain information on WS and have been meta-analysed in the last section. Type C has been included in the meta-analytic model of WS data although it does not contribute to the accuracy of WS but because of its potential contribution to the accuracy of DD.

Unconditional DD data have already been analysed in Chapter 5 by using the bivariate model. This model can be indicated as the best one given the interpretability and the explicit account for the correlation of sensitivity and specificity (as discussed throughout Chapter 4 and Chapter 5). Therefore, the bivariate approach will be used also for the conditional accuracy of DD.

Type A (DD) data is easily added to the model through the implementation of three "independent" bivariate random effect models (the models are independent, although the data clearly refer to the accuracy of the same test and they are correlated through the thresholds of the tests). For type B data, the missing

information on the accuracy of DD is estimated by the model based on the assumption of exchangeability.

$$TP_{ki} \sim binomial(sens_{ki}, d_{ki})$$

$$TN_{ki} \sim binomial(spec_{ki}, h_{ki})$$

$$logit(sens_{ki}) = \mu_{1ki}$$

$$logit(spec_{ki}) = \mu_{2ki}$$

$$\begin{pmatrix}\mu_{1ki}\\ \mu_{2ki}\end{pmatrix} = MVN\left[M = \begin{pmatrix}\mu_{1k}\\ \mu_{2k}\end{pmatrix}, \Sigma = \begin{pmatrix}\sigma_{1k}^2 & \sigma_{12k}\\ \sigma_{21k} & \sigma_{2k}^2\end{pmatrix}\right]$$

$$for\ i\ in\ 1\ \ to\ \ n_A + n_B + n_C\ \ (type\ A, B\ and\ C)$$

$$for\ WS\ level\ (k)\ \ 1 = low,\ \ 2 = moderate\ \ and\ \ 3 = high$$

Equation 8-5

Where the correlation for sensitivity and specificity within WS level k is $\rho_k = \frac{\sigma_{12k}}{\sigma_{2k} * \sigma_{2k}}$.

For type C data, the missing total number of diseased (d_{ki}) and healthy (h_{ki}) is estimated by the model for the meta-analysis of WS data specified above in this section. The accuracy of DD conditional to WS, where not reported in the data, is estimated by the model as specified in Equation 8-5 based on the assumption of exchangeability (i.e. if for study i^* TP_{ki^*} and TN_{ki^*} are not available, then these are estimated by the model by mean of μ_{1k} and μ_{2k}).

It needs to be noted at this point that type D data only contributes to the estimate of the accuracy parameters of WS, and therefore no model exists for the accuracy of DD conditional to WS. Rather, if the model in Equation 8-5 is applied to type

D data (i.e. $for\ i\ in\ 1\ \ to\ \ n_A + n_B + n_C + n_D\ \ (type\ A, B\ , C\ and\ D)$), then based on the assumption of exchangeability that characterise all the meta-analytical modeling framework, a predictive estimate of the sensitivity (q_{1ki}) and specificity (q_{2ki}) for type D studies ($i\ in\ n_A + n_B + n_C + 1\ \ to\ \ n_A + n_B + n_C + n_D$) will be sampled by the model (i.e. what would the sensitivity and specificity of DD conditional to WS likely to be if DD was performed in these studies?).

The pooled estimates of sensitivity and specificity of DD conditional to WS can be calculated from the parameters estimated in Equation 8-5 and are presented in Table 8-7, which is similar to Table 8-5 that was used to describe the conditional DD data above in this section.

WS levels →	LOW (k=1)	MOD (k=2)	HIGH (k=3)
$\boldsymbol{sens_k^{pooled}}$	$\frac{\exp(\mu_{11})}{1+\exp(\mu_{11})}$	$\frac{\exp(\mu_{12})}{1+\exp(\mu_{12})}$	$\frac{\exp(\mu_{13})}{1+\exp(\mu_{13})}$
$\boldsymbol{spec_k^{pooled}}$	$\frac{\exp(\mu_{21})}{1+\exp(\mu_{21})}$	$\frac{\exp(\mu_{22})}{1+\exp(\mu_{22})}$	$\frac{\exp(\mu_{23})}{1+\exp(\mu_{23})}$

Table 8-7 Formulae to calculate the pooled estimates of the conditional accuracy of DD given WS levels.

Inclusion of data type E, count data for unconditional accuracy of DD

The last data that can be added to the model refers to the overall accuracy of DD (type E). The assumption at the basis of the inclusion of this data is that the overall accuracy of DD can be expressed as a function of the proportion of diseased and healthy in each WS category and the accuracy of DD conditional to WS. In studies reporting data of type E, the proportion of diseased/healthy patients per WS category and the conditional accuracy data of DD are missing but their sums across WS levels is reported. The model is fit to the data based on this relationship, on the data structure proposed in Table 8-4 and Table 8-5 and analysed via the multinomial logit model for WS and the bivariate models for conditional DD, and using the overall DD accuracy data to constrain the parameters.

Step 1: The first step is to set up the bivariate logit model for the meta-analysis of the unconditional accuracy of DD:

$$TP_i \sim binomial(sens_i, N_{Di})$$

$$TN_i \sim binomial(spec_i, N_{Hi})$$

$$logit(sens_i) = \mu_{1i}$$

$$logit(spec_i) = \mu_{2i}$$

$$\begin{pmatrix}\mu_{1i}\\ \mu_{2i}\end{pmatrix} = MVN\left[M = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \Sigma = \begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{21} & \sigma_2^2\end{pmatrix}\right]$$

$$for\ i\ in\ n_A + n_B + n_C + n_D + 1 \quad to \quad n_A + n_B + n_C + n_D + n_E \ \ (type\ E)$$

Equation 8-6

Where TP_i =number of true positive, TN_i =number of true negative, N_{Di} =number of diseased and N_{Hi} =number of healthy.

Step 2: The second step is to replicate the conditional structure of the data also for type E data using the multinomial logit model and the three bivariate models for the conditional accuracy of DD to estimate, under the assumption of exchangeability, the proportion of diseased and healthy in every WS category (Equation 8-7), and the sensitivity and specificity for DD|WS (Equation 8-8)

$$\xi^{new}{}_{Dki} \sim norm(\xi_{Dk}, \sigma_D^2)$$
$$\xi^{new}{}_{Hki} \sim norm(\xi_{Dk}, \sigma_H^2)$$
$$p^{new}{}_{Dki} = \frac{\exp(\xi^{new}{}_{Dki})}{\sum_{k=1}^{3} \exp(\xi^{new}{}_{Dki})}$$
$$p^{new}{}_{Hki} = \frac{\exp(\xi^{new}{}_{Dki})}{\sum_{k=1}^{3} \exp(\xi^{new}{}_{Dki})}$$
$$for\ i\ in\ n_A + n_B + n_C + n_D + 1 \quad to \quad n_A + n_B + n_C + n_D + n_E \ \ (type\ E)$$
$$for\ WS\ level\ (k) \quad 1 = low, \quad 2 = moderate \quad and \quad 3 = high$$

Equation 8-7

$$\begin{pmatrix} \boldsymbol{\mu^{new}}_{\mathbf{1ki}} \\ \boldsymbol{\mu^{new}}_{\mathbf{2ki}} \end{pmatrix} \sim \boldsymbol{MVN}\left[\boldsymbol{M} = \begin{pmatrix} \boldsymbol{\mu_{1k}} \\ \boldsymbol{\mu_{2k}} \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\sigma^2_{1k}} & \boldsymbol{\sigma_{12k}} \\ \boldsymbol{\sigma_{21k}} & \boldsymbol{\sigma^2_{2k}} \end{pmatrix}\right]$$

$$sens_{ki} = \frac{\exp\left(\mu^{new}{}_{1ki}\right)}{1 + \exp\left(\mu^{new}{}_{1ki}\right)}$$

$$spec_{ki} = \frac{\exp\left(\mu^{new}{}_{2ki}\right)}{1 + \exp\left(\mu^{new}{}_{2ki}\right)}$$

$$for\ i\ in\ n_A + n_B + n_C + n_D + 1 \quad to \quad n_A + n_B + n_C + n_D + n_E \ \ (type\ A, B\ and\ C)$$

$$for\ WS\ level\ (k) \quad 1 = low, \quad 2 = moderate \quad and \quad 3 = high$$

Equation 8-8

Step 3: The third step is to link the parameters estimated in step 1 with the parameters simulated in step 2 using a generalisation of Equation 7-3 in Chapter 7 where the first test is not dichotomous but has three categories (Equation 8-9). Here, the unconditional accuracy parameters of DD (left side of Equation 8-9) will influence the estimate of the parameters for the conditional accuracy of WS and DD|WS (right side of Equation 8-9) which are estimated from the data types A, B, C and D. Based on the assumption of exchangeability, the parameters at the right side of Equation 8-9 will affect the overall estimates of the accuracy of WS and WS|DD when these are allowed to be part of Equation 8-3 and Equation 8-5.

$$sens_i = \sum_{k=1}^{3} sens_{ki} * p^{new}{}_{Dki}.$$

$$spec_i = \sum_{k=1}^{3} spec_{ki} * p^{new}{}_{Hki}.$$

$$for\ i\ in\ n_A + n_B + n_C + n_D + 1 \quad to \quad n_A + n_B + n_C + n_D + n_E \ \ (type\ A, B\ and\ C)$$

Equation 8-9

Where $sens_i$ and $spec_i$ are the unconditional sensitivity and specificity of DD, and these are expressed as functions of the conditional sensitivity and specificity of DD given WS - $sens_{ki}$ and $spec_{ki}$ respectively – and the accuracy parameters for WS -$p^{new}{}_{Dki}.$ and $p^{new}{}_{Hki}.$ respectively-.

The demonstration of the relationships above is obtained from the formulation of the sensitivity (and specificity) expressed in the last column of Table 8-5:

$$\frac{\sum_k r_{Dki}}{N_{Di}} = \frac{r_{D1i} + r_{D2i} + r_{D3i}}{N_{Di}} = \frac{d_{1i}}{d_{1i}} * \frac{r_{D1i}}{N_{Di}} + \frac{d_{2i}}{d_{2i}} * \frac{r_{D2i}}{N_{Di}} + \frac{d_{3i}}{d_{3i}} * \frac{r_{D3i}}{N_{Di}} =$$

$$\frac{d_{1i}}{N_{Di}} * \frac{r_{D1i}}{d_{1i}} + \frac{d_{2i}}{N_{Di}} * \frac{r_{D2i}}{d_{2i}} + \frac{d_{3i}}{N_{Di}} * \frac{r_{D3i}}{d_{3i}} = \sum_{k=1}^{3} sens_{ki} * p^{new}{}_{Hki}$$

Where $\frac{d_{ki}}{N_{Di}} = sens_{ki}$ and $\frac{r_{Dki}}{d_{ki}} = p^{new}{}_{Dki}$. Similarly for specificity.

8.7.2 Linking of intermediate to final parameters (transformations)

Accuracy of DD for intersections of WS categories

This section will discuss how to calculate the accuracy of DD for patients classified as either moderate or high WS (similarly for either low or moderate, or for either low or high). This will be very useful to simplify the formulae for the accuracy of the strategies in the next section.

In the modeling approach described in the last section, variability in threshold for WS is considered by estimating the proportion of diseased/healthy patients for each WS category. Then, the (conditional) accuracy of DD is estimated given each WS category. Now the problem is, given these estimates, can the accuracy of DD for patients classified low or high be estimated?

The parameters estimated in the multinomial logit model are, for WS low and diseased: $p_{D1i} = \frac{d_{1i}}{d_{1i}+d_{2i}+d_{3i}}$, similarly for WS moderate or high, and healthy.

The overall sensitivity of DD is

$$\frac{\sum_k r_{Dki}}{N_{Di}} = \sum_k \frac{r_{Dki}}{d_{1i}+d_{2i}+d_{3i}} = \sum_k \left(\frac{r_{Dki}}{dki} * \frac{d_{ki}}{d_{1i}+d_{2i}+d_{3i}}\right) = \sum_k \left(sens_{ki} * p_{k1i}\right).$$

Similarly, the overall specificity of DD is

$$\frac{\sum_k r_{Hki}}{N_{Hi}} = \sum_k \frac{r_{Hki}}{h_{1i} + h_{2i} + h_{3i}} = \sum_k \left(\frac{r_{Hki}}{h_{ki}} * \frac{h_{ki}}{h_{1i} + h_{2i} + h_{3i}} \right) = \sum_k (spec_{ki} * p_{Hki})$$

Equation 8-10

Where p_{Dki} and p_{Hki} are the weights of the weighted average of the conditional sensitivities and specificities of DD given WS level. The formulae above are a generalization of Equation 7-3 in the case when the first test is not dichotomised but categorical. Similarly, if one wants to calculate the sensitivity and specificity of DD only for patients categorised either moderate or high WS, similar formulae can be derived:

$$\begin{aligned} sens_{k=2,3\,;\,i} &= \frac{r_{\mathrm{D2i}} + r_{\mathrm{D3i}}}{d_{2i} + d_{3i}} = \frac{r_{\mathrm{D2i}}}{d_{2i} + d_{3i}} + \frac{r_{\mathrm{D3i}}}{d_{2i} + d_{3i}} \\ &= \frac{r_{\mathrm{D2i}}}{d_{2i}} * \frac{d_{2i}}{d_{2i} + d_{3i}} + \frac{r_{\mathrm{D3i}}}{d_{3i}} * \frac{d_{3i}}{d_{2i} + d_{3i}} \\ &= sens_{2i} * w_{D1\mathrm{i}} + sens_{3i} * w_{D2\mathrm{i}} \end{aligned}$$

$$\begin{aligned} spec_{k=2,3\,;\,i} &= \frac{r_{\mathrm{H2i}} + r_{\mathrm{H3i}}}{h_{2i} + h_{3i}} = \frac{r_{\mathrm{H2i}}}{h_{2i} + h_{3i}} + \frac{r_{\mathrm{H3i}}}{h_{2i} + h_{3i}} \\ &= \frac{r_{\mathrm{H2i}}}{h_{2i}} * \frac{h_{2i}}{h_{2i} + h_{3i}} + \frac{r_{\mathrm{H3i}}}{h_{3i}} * \frac{h_{3i}}{h_{2i} + h_{3i}} \\ &= spec_{2i} * w_{H1i} + spec_{3i} * w_{H2i} \end{aligned}$$

Equation 8-11

Where $w_{D1\mathrm{i}} = \frac{d_{2i}}{d_{2i}+d_{3i}}$ and $w_{D2\mathrm{i}} = \frac{d_{3i}}{d_{2i}+d_{3i}}$, $w_{H1\mathrm{i}} = \frac{h_{2i}}{h_{2i}+h_{3i}}$ and $w_{H2\mathrm{i}} = \frac{h_{3i}}{h_{2i}+h_{i3}}$.

Multiplying numerator and denominator of the weights for diseased by the total number of diseased N_{Di}, and the numerator and denominator of the weights for healthy by N_{Hi}, the weights can be expressed in terms of proportions:

$w_{D1\mathrm{i}} = \frac{p_{\mathrm{D2}i}}{p_{\mathrm{D2i}}+p_{\mathrm{D3i}}}$ and $w_{D2\mathrm{i}} = \frac{p_{\mathrm{D3i}}}{p_{\mathrm{D2i}}+p_{\mathrm{D3i}}}$, $w_{H1\mathrm{i}} = \frac{p_{\mathrm{H2i}}}{p_{\mathrm{H2i}}+p_{\mathrm{H3i}}}$ and $w_{H2\mathrm{i}} = \frac{p_{\mathrm{H3}i}}{p_{\mathrm{H2}i}+p_{H3i}}$

Accuracy of the diagnostic strategies

The unconditional accuracy of DD (5th strategy in section 8.5) was not obtained from the multi-component model. Instead, data A for DD were aggregated over WS levels, and merged to type E data, and then analysed using a bivariate random effect logit model as presented in Chapter 4 and Chapter 5 (Model 4-9).

The accuracy of WS considered as a single test with one of two possible thresholds can be derived using parameters estimates presented in Equation 8-4. In section 8.5, where all the relevant strategies including WS and DD are presented, WS with the first threshold (i.e. Low vs Moderate/High) is the 6th strategy and its accuracy is:

$$sens_{WS_{t1}} = p_{D2} + p_{D3}$$

$$spec_{WS_{t1}} = p_{H1}$$

Equation 8-12

WS with the second threshold (i.e. Low/Moderate vs High) is the 7th strategy and its accuracy is:

$$sens_{WS_{t2}} = p_{D3}$$

$$spec_{WS_{t2}} = p_{H1} + p_{H2}$$

Equation 8-13

The accuracy of the four strategies characterised by sequences of WS and DD (strategies 1 to 4 in section 8.5) can be estimated considering the following equations:

- Strategy 1: WS at threshold 1 (low vs moderate/high), believe the negatives

$$sens_{(WS_{t1} and\ DD)_{BN}} = sens_{WS_{t1}} * sens_{k=2,3}$$

$$spec_{(WS_{t1} and\ DD)_{BN}} = 1 - [(1 - spec_{WS_{t1}}) * (1 - spec_{k=23})]$$

Equation 8-14

Where $sens_{k=2,3}$ is the sensitivity of DD for patients classified either moderate or high and can be calculated as a weighted average of the sensitivities of DD for WS moderate and high $sens_{k=2,3} = w_{D1} * sens_2 + w_{D2} * sens_3$, and the weights can be demonstrated to be equal to $w_{D1} = \frac{p_{D2}}{p_{D2}+p_{D3}}$ and $w_{D2} = 1 - w_{D1} = \frac{p_{D3}}{p_{D2}+p_{D3}}$ (see first subsection above). Similarly,

$spec_{k=2,3} = 1 - \{[(1 - spec_2)w_{H1}] * [(spec_3)w_{H2}]\}$.

Where $spec_2$ is directly estimated by the model and does not need to be calculated.

- Strategy 2:WS at threshold 1 (low vs moderate/high), believe the positives

$$sens_{(WS_{t1} and\ DD)_{BP}} = 1 - [(1 - sens_{WS_{t1}}) * (1 - sens_1)]$$

$$spec_{(WS_{t1} and\ DD)_{BP}} = spec_{WS_{t1}} * spec_1$$

Equation 8-15

- Strategy 3:WS at threshold 2 (low/moderate vs high), believe the negatives

$$sens_{(WS_{t2} and\ DD)_{BN}} = sens_{WS_{t2}} * sens_3$$

$$spec_{(WS_{t2} and\ DD)_{BN}} = 1 - [(1 - spec_{WS_{t2}}) * (1 - spec_3)]$$

Equation 8-16

- Strategy 4:WS at threshold 2 (low/moderate vs high), believe the positives

$$sens_{(WS_{t2} and\ DD)_{BP}} = 1 - [(1 - sens_{WS_{t2}}) * (1 - sens_{k=1,2})]$$

$$spec_{(WS_{t2} and\ DD)_{BP}} = spec_{WS_{t2}} * spec_{k=1,2}$$

Equation 8-17

Where, similarly to the 1st strategy, $sens_{k=1,2}$ and $spec_{k=1,2}$ are the sensitivity and specificity of DD given patients have been classified either WS low or moderate , which can be calculated as weighted averages of the sensitivities and specificities of DD given low and moderate WS as discussed in the last section.

The 8th strategy has been included although it is expected that only sequences as the 1st to 4th make sense since they represent better ways of increasing either sensitivity or specificity, which is the main reason to combine tests together (Pepe 2003). The accuracy of this strategy is expected to be in between the sensitivity maximising strategy (i.e. the 2nd) and the specificity maximising strategy (i.e. the 3rd). However, beyond the clinical assessment of the strategy (i.e. in terms of solely sensitivity or specificity maximization) this strategy can be included in a cost effectiveness analysis to assess the overall performance (i.e. including costs and quality data) compared to the other options.

The sensitivity of this combination is the probability that either of the two following events occur: 1. a diseased patient who scored moderate to WS is also positive to DD; or 2. a diseased patient scores high to WS, that is equal to $P(WS\ high|Diseased) + P(DD+, WS\ moderate|Diseased)$.

Similarly, specificity is the probability that either 1. a diseased patient scores low to WS, or 2. a diseased patients who scores moderate to WS also scores negative to DD, that is equal to

$P(WS\ low|Non\ Diseased) + P(DD-, WS\ moderate|Non\ Diseased)$.

Therefore the accuracy of the 8th strategy is represented in Equation 8-18

$$sens_{if\ WS\ mod \rightarrow DD} = sens_{WS_{t2}} + sens_2 * p_{D2}$$

$$spec_{if\ WS\ mod \rightarrow DD} = spec_{WS_{t1}} + spec_2 * p_{H2}$$

Equation 8-18

Since this modeling framework aims to account for dependence between tests and the accuracy of DD alone is not affected by this issue, the accuracy of DD alone was calculated by a separate bivariate model, using data type E and data type A aggregated across WS levels (i.e. $data\ for\ i\ form\ 1\ \ to\ \ n_A$: $TP_i = \sum_{k=1}^{3} TP_{ki}$; $TN_i = \sum_{k=1}^{3} TN_{ki}$; $data\ for\ i\ in\ \ n_A + 1\ \ to\ \ n_E\ directly\ available$).

8.8 Results of the data analysis

This section presents the results obtained from the modeling approach developed throughout this chapter where all the available data types are used, and either independence or dependence between the two different tests in the strategy were assumed for comparison. When independence is assumed, the accuracy strategies were calculated *i*) running models for WS and DD separately and *ii*) using the formulae presented in section 8.7.2 (from Equation 8-12 to Equation 8-18) but substituting the unconditional accuracy of DD to the conditional accuracy rates for DD.

sROC curves and credible or predictive ellipses were calculated adapting the formulae presented in Chapter 4 (Equation 4-10).

8.8.1 MCMC Diagnostics

The model developed in this chapter needs to be checked against the problems that may potentially affect MCMC chains: length of the burn-in period, convergence of the chains, sensitivity to initial priors and sensitivity to prior distributions.

The length of the burn-in period has been safely set at 5000 iterations. This has been graphically assessed by setting two different chains and using the history tool available in WinBUGS, which plots the sampled values for both chains for

every parameter. All couples of chains (for every parameter) overlap before 1000 iterations.

The convergence and sensitivity to initial values of the algorithm have been assessed, as before, by visualising the two chains via the history tool after the burn-in period. Using a sample of 20,000 iterations, the chains overlap after the burn-in period, although the chains sometimes appear waving rather than completely over-impose. However, such waves do not correspond to poor convergence but they correspond to a high degree of autocorrelation, confirmed by the autocorrelation plots available via WinBUGS. For this reason, the length of the sample has been set to 100,000. The parameter estimates do not vary if the length of the sample is set to 20,000 iterations, indicating that the model generates parameter estimated that are robust to autocorrelation problems (i.e. autocorrelation was not too high).

Although there was no evidence that the model results were sensitive to the initial values, such initial values needed to be chosen carefully to get the model run (i.e. perform successfully the first simulation) with respect to the prior distributions of the multinomial random effect logit model. Since a non linear transformation links the proportions estimated at the likelihood level and the logit transformations, it is quite easy to initialise these parameters with implausible values. For our model such parameters are initialised to -4, which ensure that the model updates. For complex models, suggesting initial values and restricting prior distributions to

plausible range of the parameter values has been done before (i.e. for the HsROC model (Rutter and Gatsonis 2001)).

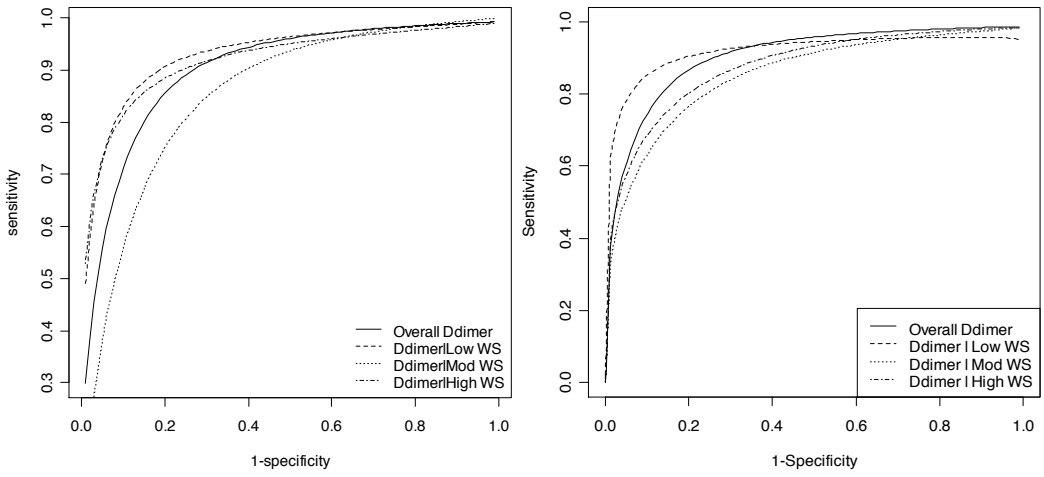
Position parameters were usually associated with normal prior distributions with mean 0 and very low precision (non informative prior distributions). Heterogeneity parameters for the multinomial logit model (i.e. standard deviations) were given a half normal (i.e. on the positive side, the mean of the normal distribution was set to 0) prior distribution with low precision. The heterogeneity parameters of the bivariate model for the conditional accuracy of DD (i.e. the matrix of variances and covariances) were given a Wishart prior distribution with parameters $\begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$. The model results were not sensitive to different values of the parameters of the Wishart prior distribution $\begin{pmatrix} 3 & -5 \\ -2 & 2 \end{pmatrix}$; i.e. only small differences were observed in the mean values of the scale parameters and the boundaries of the credible intervals, which were mainly overlapping.

8.8.2 Estimates of the intermediate parameters

The modeling approach developed in the last section accounts for correlation between tests by estimating the conditional accuracies of the tests. Figure 8-4 (a) represents the unconditional accuracy of DD compared with the conditional accuracy of DD given WS when all data types are included into the modeling approach. It is clear that if the assumption of independence between WS and DD was true then the four sROC curves would be overlaid on top of each other. Demonstrating this rule is not difficult: first, consider that sROC curves are couples of sensitivities and specificities; second, Equation 8-19 shows that the conditional and unconditional sensitivity are equal if tests are independent; third, the same rule applies for specificities; finally, if Equation 8-19 is applied for all sensitivities that compose the sROC curve (and similarly for all specificities), then the sROC curves for conditional and unconditional accuracy would overlap if the tests are conditionally independent. For example, type A data can be considered which report full data on the overall and conditional accuracy of DD; Figure 8-4 (b) compares the accuracy of DD and DD given WS for this data. It is important to note that when all data are considered, the only worst fitting curve for DD is related to WS moderate, while when type A data are considered, the sROC curves relating to DD given WS is high also indicates low accuracy.

$$P(T2+ \mid diseased) = P(T2+ \mid T1+, diseased)$$

Equation 8-19



a) b)

Figure 8-4 a)Roc curves for the accuracy of DD and DD|WS based on data types A,B,C,D and E. b) sROC curves for the accuracy of DD and DD|WS based on data type A.

Table 8-8 presents and compares the estimates of the intermediate parameters for the model when each data type is added to the model (sequentially A, B, C, D, E) and credible intervals at a credibility level of 95%. Also, the estimates of the same parameters are reported when independence is assumed for comparison; in this case the data types that could be used to inform the multinomial logistic regression for WS and the bivariate mode for DD are type A, B and D (for WS) and type A and E (for DD) respectively. It can be seen that parameter estimates change as different data types are reported, and, equally important, they are different from the estimates obtained by assuming independence. Therefore, the impact of the assumption of independence and the inclusion of part of the available evidence directly reflect on these parameter estimates. In fact, the estimated proportions of diseased patients per WS category does not change much

when the assumption of independence is relaxed and all data are used, however, the conditional accuracy of DD given WS is very different: sensitivity 0.930 (0.919 to 0.941) and specificity 0.552 (0.522 to 0.583) when independence is assumed; when dependence is accounted for then sensitivity varies between 0.930 (0.863 to 0.981) for WS low, and 0.960 (0.933 to 0.982) for WS high, while specificity is estimated at 0.390 (0.212 to 0.561) for WS moderate and at 0.699 (0.598 to 0.797) for WS high. When assuming independence, an amount of unexplained heterogeneity is estimated for both WS and DD (see estimates of σ_1, σ_2 in Table 8-8). However, when dependence is accounted for, such unexplained heterogeneity (see estimates of $\sigma_{1\mathrm{k}}$, $\sigma_{2\mathrm{k}}$ for the different data type) *i*) is larger than assuming independence *ii*) it is much larger when data type E is included into the model (which is not surprising since data type E involves the estimates of the conditional accuracy parameters via predicting them for the 198 record of data type E), and *iii*) it is very large for DD given moderate WS, confirming that considering the strategy number 8 in section 8.6 may be a sensible choice. However, this suggests that such heterogeneity may be explored, for example by adding covariate, which has not been the focus of this chapter but is an issue to explore in the future. Finally, the assumption of independence would be misleading in this case leading to wrong conclusions and badly supported clinical and economic decisions.

Wells score -->			Low (k=1)	Moderate (k=2)	High (k=3)
IND	WS	p_{Dk}^{pooled}	0.121 (0.096 to 0.148)	0.345 (0.295 to 0.398)	0.534 (0.479 to 0.591)
		p_{Hk}^{pooled}	0.497 (0.416 to 0.576)	0.385 (0.312 to 0.462)	0.118 (0.087 to 0.153)
	Type A, B,D	σ_D	<---------- *0.223 (0.124 to 0.368)* ---------->		
		σ_H	<---------- *0.665 (0.263 to 5.710)* ---------->		
	DD|WS	$sens_k^{pooled}$	<---------- *0.930 (0.919 to 0.941)* ---------->		
		$spec_k^{pooled}$	<---------- *0.552 (0.522 to 0.583)* ---------->		
	Type A, E	σ_1	<---------- *0.223 (0.124 to 0.368)* ---------->		
		σ_2	<---------- *0.665 (0.263 to 5.710)* ---------->		
		σ_{12}	<---------- -0.544 (-0.731 to -0.382) ---------->		
		ρ	<---------- -0.586 (-0.697 to -0.459) ---------->		
DEP Type A	WS	p_{Dk}^{pooled}	0.136 (0.085 to 0.198)	0.397 (0.289 to 0.515)	0.467 (0.352 to 0.584)
		p_{Hk}^{pooled}	0.522 (0.420 to 0.617)	0.368 (0.280 to 0.467)	0.110 (0.076 to 0.154)
		σ_D	<---------- *0.328 (0.121 to 0.755)* ---------->		
		σ_H	<---------- *0.245 (0.114 to 0.480)* ---------->		
	DD|WS	$sens_k^{pooled}$	0.921 (0.840 to 0.974)	0.913 (0.800 to 0.983)	0.968 (0.919 to 0.995)
		$spec_k^{pooled}$	0.721 (0.596 to 0.829)	0.502 (0.336 to 0.670)	0.431 (0.277 to 0.599)
		σ_{1k}	1.041 (0.163 to 3.831)	3.250 (0.657 to 10.970)	2.956 (0.513 to 10.250)
		σ_{2k}	0.876 (0.315 to 2.204)	1.243 (0.470 to 3.046)	1.015 (0.273 to 2.777)
		σ_{12k}	-0.432 (-1.782 to 0.372)	-1.042 (-3.395 to 0.102)	-0.935 (-3.260 to 0.211)
		ρ_k	-0.426 (-0.903 to 0.508)	-0.528 (-0.880 to 0.071)	-0.548 (-0.914 to 0.176)
DEP Type AB	WS	p_{Dk}^{pooled}	0.135 (0.092 to 0.188)	0.372 (0.283 to 0.468)	0.492 (0.396 to 0.590)
		p_{Hk}^{pooled}	0.527 (0.444 to 0.605)	0.362 (0.290 to 0.442)	0.111 (0.082 to 0.146)
		σ_D	<---------- *0.282 (0.122 to 0.569)* ---------->		
		σ_H	<---------- *0.208 (0.103 to 0.394)* ---------->		
	DD|WS	$sens_k^{pooled}$	0.929 (0.862 to 0.976)	0.913 (0.800 to 0.983)	0.968 (0.923 to 0.993)
		$spec_k^{pooled}$	0.711 (0.604 to 0.807)	0.502 (0.335 to 0.666)	0.427 (0.278 to 0.588)
		σ_{1k}	1.025 (0.168 to 3.619)	3.240 (0.656 to 10.950)	2.443 (0.480 to 8.052)
		σ_{2k}	0.749 (0.297 to 1.771)	1.235 (0.467 to 3.027)	0.955 (0.260 to 2.624)
		σ_{12k}	-0.430 (-1.585 to 0.255)	-1.034 (-3.377 to 0.100)	-0.823 (-2.760 to 0.156)
		ρ_k	-0.466 (-0.900 to 0.415)	-0.527 (-0.877 to 0.065)	-0.545 (-0.907 to 0.147)

[continued]

DEP Type ABC	WS	p_{Dk}^{pooled}	0.130 (0.086 to 0.179)	0.363 (0.277 to 0.457)	0.508 (0.413 to 0.604)
		p_{Hk}^{pooled}	0.571 (0.472 to 0.665)	0.329 (0.244 to 0.422)	0.101 (0.068 to 0.140)
		σ_D	<---------- *0.302 (0.132 to 0.613)* ---------->		
		σ_H	<---------- *0.351 (0.186 to 0.654)* ---------->		
	DD|WS	$sens_k^{pooled}$	0.940 (0.882 to 0.982)	0.916 (0.812 to 0.982)	0.973 (0.935 to 0.995)
		$spec_k^{pooled}$	0.673 (0.575 to 0.765)	0.501 (0.358 to 0.651)	0.437 (0.296 to 0.592)
		σ_{1k}	1.114 (0.173 to 4.028)	3.487 (0.801 to 11.080)	2.628 (0.512 to 8.679)
		σ_{2k}	0.704 (0.309 to 1.498)	1.108 (0.447 to 2.518)	0.901 (0.253 to 2.439)
		σ_{12k}	-0.457 (-1.548 to 0.208)	-1.107 (-3.196 to -0.085)	-0.723 (-2.546 to 0.287)
		ρ_k	-0.499 (-0.899 to 0.373)	-0.577 (-0.878 to -0.065)	-0.483 (-0.888 to 0.240)
DEP Type A,B, C,D	WS	p_{Dk}^{pooled}	0.124 (0.101 to 0.149)	0.343 (0.298 to 0.393)	0.533 (0.483 to 0.583)
		p_{Hk}^{pooled}	0.511 (0.447 to 0.576)	0.377 (0.317 to 0.440)	0.112 (0.088 to 0.139)
		σ_D	<---------- *0.177 (0.099 to 0.289)* ---------->		
		σ_H	<---------- *0.343 (0.232 to 0.499)* ---------->		
	DD|WS	$sens_k^{pooled}$	0.939 (0.881 to 0.980)	0.917 (0.815 to 0.982)	0.973 (0.935 to 0.995)
		$spec_k^{pooled}$	0.675 (0.577 to 0.766)	0.500 (0.354 to 0.644)	0.439 (0.296 to 0.589)
		σ_{1k}	1.051 (0.172 to 3.651)	3.435 (0.788 to 10.600)	2.659 (0.518 to 8.908)
		σ_{2k}	0.710 (0.307 to 1.538)	1.106 (0.447 to 2.544)	0.897 (0.251 to 2.403)
		σ_{12k}	-0.443 (-1.522 to 0.203)	-1.102 (-3.145 to -0.087)	-0.733 (-2.571 to 0.282)
		ρ_k	-0.492 (-0.900 to 0.357)	-0.578 (-0.878 to -0.067)	-0.485 (-0.893 to 0.247)
DEP Type AB C DE	WS	p_{Dk}^{pooled}	0.127 (0.105 to 0.152)	0.347 (0.301 to 0.394)	0.526 (0.476 to 0.577)
		p_{Hk}^{pooled}	0.482 (0.422 to 0.540)	0.406 (0.351 to 0.464)	0.112 (0.089 to 0.138)
		σ_D	<---------- 0.170 (0.094 to 0.282) ---------->		
		σ_H	<---------- 0.298 (0.199 to 0.437) ---------->		
	DD|WS	$sens_k^{pooled}$	0.930 (0.863 to 0.981)	0.957 (0.893 to 0.995)	0. 960 (0.933 to 0.982)
		$spec_k^{pooled}$	0.699 (0.598 to 0.797)	0.390 (0.212 to 0.561)	0.433 (0.300 to 0.566)
		σ_{1k}	1.979 (0.277 to 6.768)	7.957 (2.813 to 19.960)	1.093 (0.426 to 2.296)
		σ_{2k}	1.597 (0.664 to 3.082)	3.934 (1.393 to 8.029)	0.905 (0.213 to 2.538)
		σ_{12k}	-1.358 (-3.633 to -0.103)	-4.633 (-9.844 to -1.711)	-0.623 (-1.673 to -0.016)
		ρ_k	-0.747 (-0.977 to -0.007)	-0.845 (-0.953 to -0.675)	-0.632 (-0.922 to -0.026)

Table 8-8 Estimates of intermediate parameters (and 95% credible intervals in brackets) when assuming independence between tests (IND) and when assuming dependence between tests (DEP) for different data types included in the model amongst the available A, B, C, D and E.

The use of different sources of data is very important where these data can add value to the estimate with their contribution to the overall information. Whether the contribution to the information is worth or not the effort of the inclusion of each data into the model can be evaluated only after the data itself is included. For example, the inclusion of type D data to the model influences the estimates of WS accuracy parameters to some extent. Whether this contribution is significant or not (i.e. overlapping credible intervals across data types) may not be the right point of view, since a number of small changes in the parameter estimated can result in a bigger change when these data are then combined together. Moreover, this model can be used or adapted for other diagnostic problems, where the contribution of each data type can be more evident.

This issue is even more evident when data type E is included into the model. The contribution of type E data is very small because it is an indirect contribution (i.e. the sum of three terms influences the single terms of the sum). However, some parameters can be more influenced than others from this type of data (i.e. the estimate of $spec_2^{pooled}$). However, the inclusion of such data influences majorly the estimates of the heterogeneity parameters, which consequently will have a different impact when predictions are calculated.

8.8.3 Estimates of the final parameters

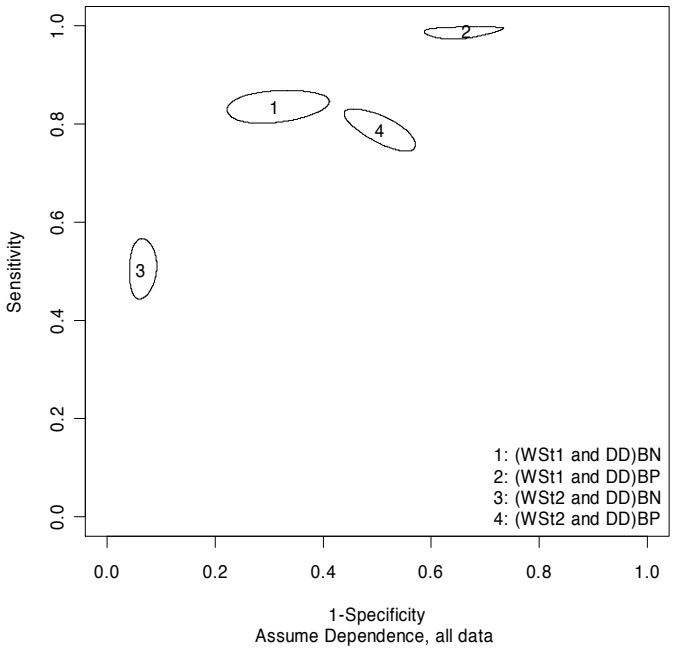
Graphical comparison of combinations

Formulae used to combine intermediate parameters in order to obtain the accuracy of the possible combinations of DD and WS have been given in section 8.7.2. The same formulae can be used to calculate the accuracy of combinations of these two tests if the unconditional accuracy of DD is used instead of the conditional accuracy of DD given WS. Table 8-9 presents the results of the strategies of WS and DD; for example, these are the final parameters that can be used in a cost effectiveness analysis (section 8.9) that aims to compare combinations of WS and DD.

Figure 8-5 shows the confidence ellipses when either independence (dashed lines) or dependence (plain lines) between tests is assumed. When independence is assumed, the credible regions for the 4 sequences seem to lie on an underlying ROC curve where the threshold variability is due to differences in between tests positivity criteria (i.e. when believe the negative the accuracy corresponds to a higher specificity and a lower sensitivity) and WS threshold (given the between tests positivity criteria, where WS is dichotomised according to the first threshold, the sensitivity is higher and specificity is lower than sequences where WS is dichotomised according to the second threshold). Credible regions already account for threshold variability for DD. The reason because credible regions are

used instead of sROC curve is that these allow the variability in thresholds to be disentangles as described above. However, when the assumption of independence is relaxed, the credible regions look different. That is, solid regions look wider although more studies have been used (type C studies could not be included into the model when independence was assumed), especially the region relative to the sequence $(WS_{t2} and\ DD)_{BP}$. It needs to be noted that the assumption of either dependence or independence also influences the data that will be included into the model; for example, when dependence is assumed the proposed modeling approach allows the inclusion of all data types, when independence is assumed than data types A, B and D are used for WS and data types A and E are included for DD.

(a)



(b)

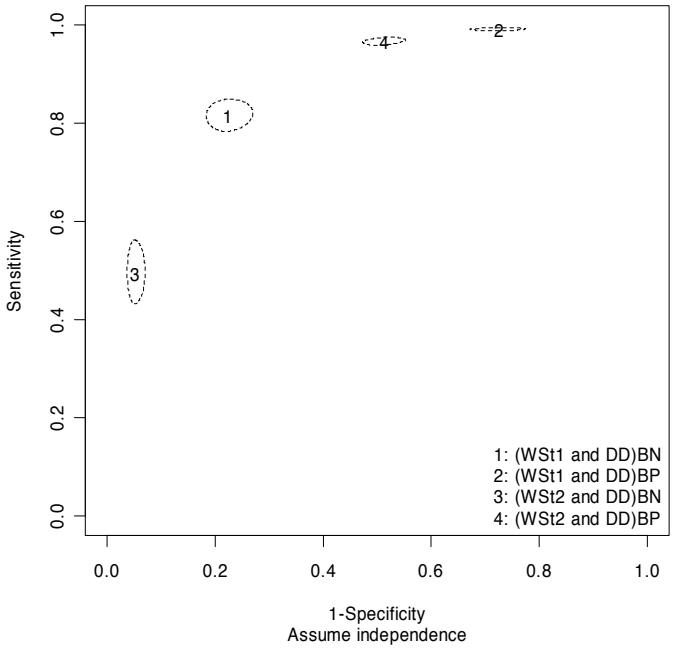


Figure 8-5 Credible ellipses for the four sequences of DD and WS assuming independence and dependence compared on the same ROC curve (plain line for dependence; dashed line for independence).

Confidence regions can be plotted for all the diagnostic strategies evaluated into the model. Figure 8-6 compares the credible regions of all the diagnostic strategies including WS and DD. It is evident that DD does not add much to the accuracy of WS when this is dichotomised according to the second threshold and negatives are believed (region 3 and 7 on Figure 8-6). Sequence $(WS_{t1} and\ DD)_{BP}$ has the highest sensitivity and the lowest specificity (region 2 on Figure 8-6), and sequence $(WS_{t2} and\ DD)_{BN}$ has the highest specificity and the lowest sensitivity (region 3 on Figure 8-6).

The choice of one of these is not straightforward. According to a common rule used for sROC curves and that is also applied to regions (Egger, Smith et al. 2001) the best combination of sensitivity and specificity could be that corresponding to the region closer to the upper left corner of the plot (the point with sensitivity and specificity both equal to 1). In this case, either DD alone or $(WS_{t1} and\ DD)_{BN}$ seem to be candidate to the best combination. Sensitivity and specificity of the sequence $(WS_{t1} and\ DD)_{BN}$ are both around 80% and may not be good compared to other tests (i.e. in a cost effectiveness analysis). Other criteria can be used to choose the best combination as already mentioned in Chapter 7. For example, one may be looking for a triage test to exclude safely as many healthy patients as possible, given that the other tests have side effects (i.e. Venography) or given the treatments side effects (i.e. anticoagulants may cause intracranial bleeding). Under this perspective, sequence $(WS_{t1} and\ DD)_{BP}$ may be the best, because it corresponds to the highest sensitivity (i.e. highest negative predictive value). Oppositely, if one wants to capture as many diseased patients as

possible (i.e. need to treat quickly because the disease evolves quickly), sequence $(WS_{t2} and\ DD)_{BN}$ maximizes the positive predictive values and may result the best choice. The trade off in terms of costs and effects between false positive and false negative results will be considered in the next section 8.9. Combinations will be evaluated as final decision gates rather than triage strategies where, for examples, more tests can be used after. This implies that positives to the combination are treated and negatives are discharged.

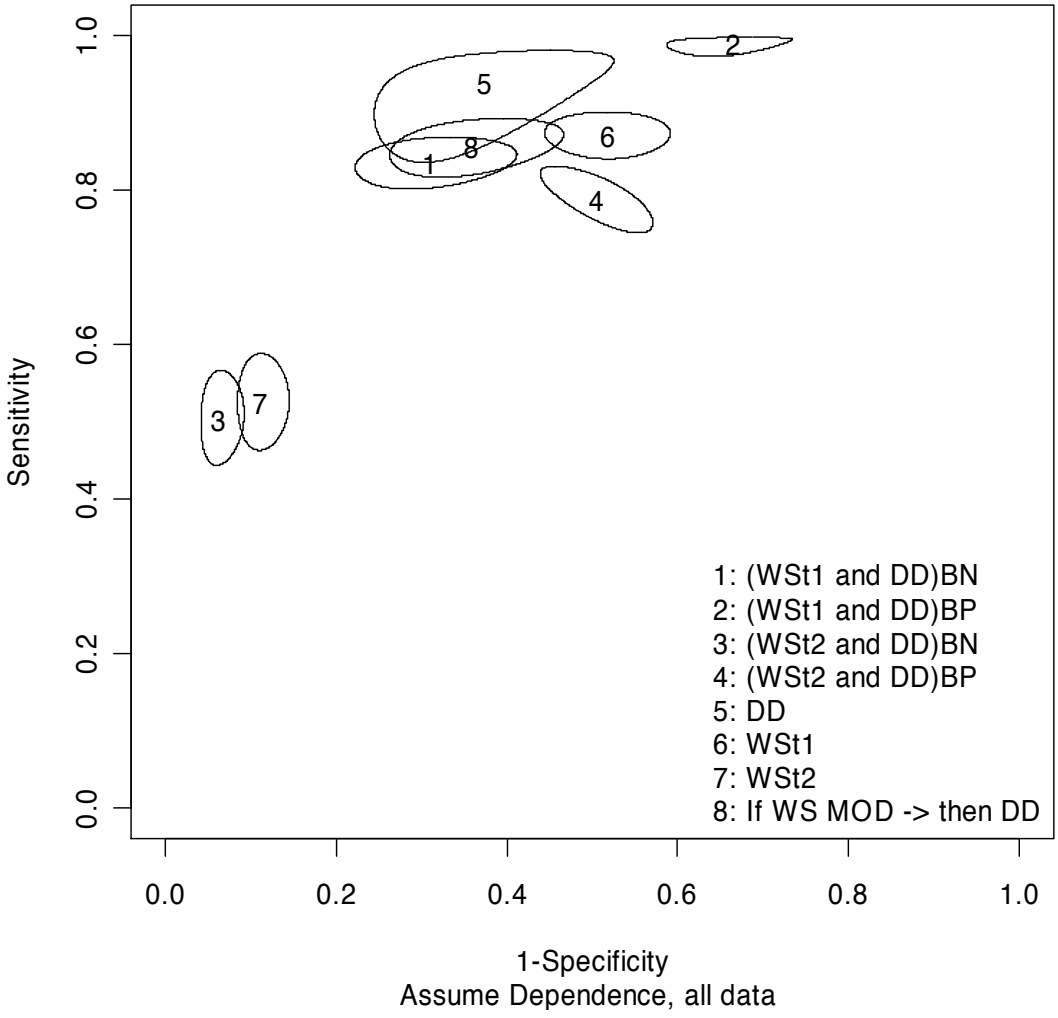
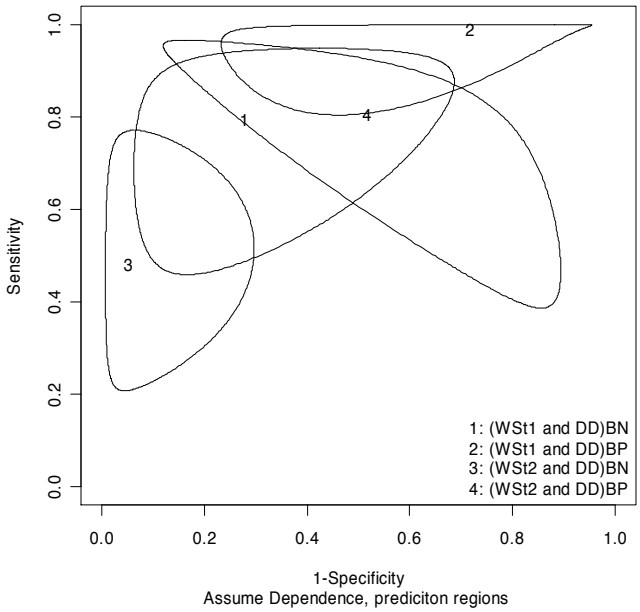


Figure 8-6 Credible ellipses for all the considered strategies of DD and WS assuming dependence between tests.

Predictive variability

Predictions allow the unexplained heterogeneity to be considered as part of the uncertainty in parameter estimates; for example it can be included into a decision analysis (Higgins, Thompson et al. 2009). WinBUGS allows the calculation of the predictive variability of the intermediate parameters to propagate into the predictive distribution of the final parameters; therefore, although no distribution is directly available for the final parameters still their predictive distributions can be obtained. Figure 8-7 shows such predictive regions in the case of independence (plot b) and dependence (plot a). When independence is assumed, the prediction areas overlap to a much greater extent.

(a)



(b)

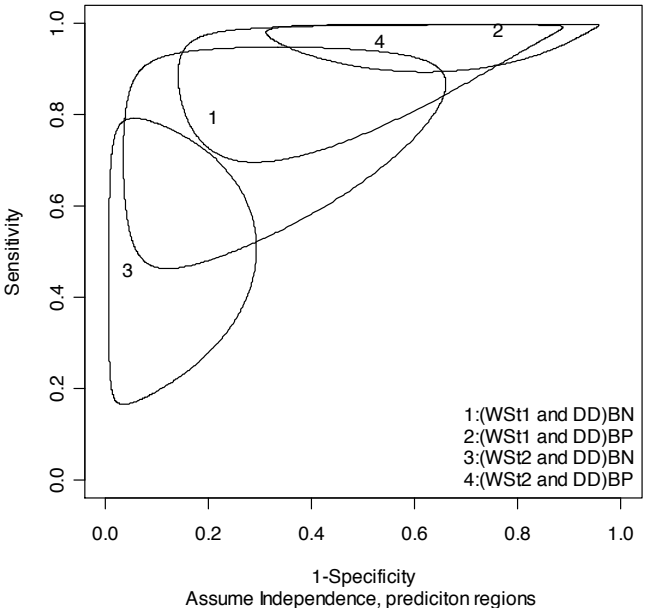


Figure 8-7 Predictive regions for the 4 sequences of DD and WS assuming either independence (b) or dependence (a) between tests.

Sensitivities and specificities of combinations

Table 8-9 presents the estimates of sensitivity and specificity for the final combinations of WS and DD described in section 8.6 and listed in the right box of Figure 8-3. It is difficult to compare the strategies on the basis of the sole numbers presented in this table, especially because of the negative correlation between sensitivity and specificity (when a higher sensitivity will be achieved, a lower specificity is likely to be estimated). Graphical methods have been proposed above (credible regions), and a cost-effectiveness analysis will be proposed in the next section 8.9. However, it can be observed how the accuracy varies when *i*) the assumption of dependence is used instead of the assumption of independence, and *ii*) when more data (from type A to type E) are included into the model.

When dependence is accounted for, the estimates of sensitivity and specificity are not much different for the tests used on their own (i.e. DD, WS_{t1} and WS_{t2}); however, the differences may be larger for the other combinations (i.e. for $(WS_{t2} and\ DD)_{BP}$) the larger difference is observed, sensitivity when independence is assumed is 0.968 (0.961 to 0.974), while when dependence is assumed sensitivity is 0.790 (0.754 to 0.823)).

Generally, smaller credible intervals are obtained as more data are included into the model. Also the estimated sensitivities and specificities vary, although the inclusion of data type E does not lead to large differences compared to the rest of data (A, B, C and D).

		$Strategy\ (ST) = \ DD$		
	Type of data	$sens_{ST}$	$spec_{ST}$	**n**
Assume independence		0.930 (0.919 to 0.941)	0.552 (0.522 to 0.583)	233
Assume dependence	**A**	0.940 (0.889 to 0.974)	0.609 (0.516 to 0.699)	11
	AB	0.942 (0.893 to 0.975)	0.603 (0.517 to 0.686)	14
	ABC	0.948 (0.905 to 0.977)	0.593 (0.516 to 0.669)	19
	ABCD	0.949 (0.910 to 0.977)	0.582 (0.505 to 0.657)	39
	ABCDE	0.936 (0.871 to 0.977)	0.624 (0.502 to 0.733)	234
		$ST = \ WS_{t1}$		
	Type of data	$sens_{ST}$	$spec_{ST}$	
Assume independence		0.879 (0.852 to 0.904)	0.497 (0.416 to 0.576)	
Assume dependence	**A**	0.864 (0.802 to 0.915)	0.522 (0.420 to 0.617)	
	AB	0.865 (0.812 to 0.908)	0.527 (0.444 to 0.605)	
	ABC	0.871 (0.821 to 0.914)	0.571 (0.472 to 0.665)	
	ABCD	0.876 (0.851 to 0.900)	0.511 (0.447 to 0.576)	
	ABCDE	0.873 (0.848 to 0.895)	0.482 (0.422 to 0.540)	
		$ST = \ WS_{t2}$		
	Type of data	$sens_{ST}$	$spec_{ST}$	
Assume independence		0.534 (0.479 to 0.591)	0.882 (0.847 to 0.913)	
Assume dependence	**A**	0.467 (0.352 to 0.584)	0.890 (0.846 to 0.924)	
	AB	0.492 (0.396 to 0.590)	0.889 (0.854 to 0.918)	
	ABC	0.508 (0.413 to 0.604)	0.899 (0.860 to 0.932)	
	ABCD	0.533 (0.483 to 0.583)	0.889 (0.861 to 0.912)	
	ABCDE	0.526 (0.476 to 0.577)	0.888 (0.863 to 0.911)	
		ST =**WS low-> neg; WS mod -> DD; WS high-> pos**		
	Type of data	$sens_{ST}$	$spec_{ST}$	
Assume independence		0.855 (0.828 to 0.881)	0.709 (0.661 to 0.755)	
Assume dependence	**A**	0.829 (0.755 to 0.890)	0.707 (0.618 to 0.787)	
	AB	0.832 (0.768 to 0.885)	0.708 (0.628 to 0.781)	
	ABC	0.840 (0.778 to 0.892)	0.735 (0.656 to 0.805)	
	ABCD	0.847 (0.805 to 0.881)	0.700 (0.631 to 0.764)	
	ABCDE	0.858 (0.825 to 0.886)	0.641 (0.557 to 0.720)	

[continued]

	Type of data	$sens_{ST}$	$spec_{ST}$
		$ST = (WS_{t1} \text{ and } DD)_{BN}$	
Assume independence		0.818 (0.790 to 0.844)	0.775 (0.735 to 0.813)
Assume dependence	A	0.815 (0.740 to 0.876)	0.754 (0.669 to 0.830)
	AB	0.816 (0.750 to 0.871)	0.756 (0.678 to 0.827)
	ABC	0.826 (0.764 to 0.879)	0.779 (0.705 to 0.845)
	ABCD	0.833 (0.788 to 0.870)	0.749 (0.680 to 0.813)
	ABCDE	0.837 (0.709 to 0.863)	0.689 (0.608 to 0.761)
		$ST = (WS_{t1} \text{ and } DD)_{BP}$	
	Type of data	$sens_{ST}$	$spec_{ST}$
Assume independence		0.992 (0.989 to 0.994)	0.275 (0.227 to 0.321)
Assume dependence	A	0.989 (0.976 to 0.997)	0.376 (0.282 to 0.469)
	AB	0.990 (0.980 to 0.997)	0.374 (0.297 to 0.453)
	ABC	0.992 (0.984 to 0.998)	0.384 (0.301 to 0.470)
	ABCD	0.993 (0.985 to 0.998)	0.345 (0.280 to 0.411)
	ABCDE	0.991 (0.983 to 0.998)	0.337 (0.280 to 0.397)
		$ST = (WS_{t2} \text{ and } DD)_{BN}$	
	Type of data	$sens_{ST}$	$spec_{ST}$
Assume independence		0.497 (0.445 to 0.550)	0.947 (0.931 to 0.961)
Assume dependence	A	0.452 (0.339 to 0.568)	0.937 (0.906 to 0.963)
	AB	0.476 (0.381 to 0.572)	0.936 (0.909 to 0.959)
	ABC	0.494 (0.400 to 0.590)	0.943 (0.915 to 0.965)
	ABCD	0.518 (0.466 to 0.570)	0.937 (0.914 to 0.958)
	ABCDE	0.505 (0.456 to 0.555)	0.936 (0.915 to 0.954)
		$ST = (WS_{t2} \text{ and } DD)_{BP}$	
	Type of data	$sens_{ST}$	$spec_{ST}$
Assume independence		0.968 (0.961 to 0.974)	0.487 (0.453 to 0.520)
Assume dependence	A	0.803 (0.729 to 0.868)	0.561 (0.467 to 0.653)
	AB	0.810 (0.748 to 0.865)	0.556 (0.469 to 0.639)
	ABC	0.808 (0.750 to 0.860)	0.549 (0.470 to 0.627)
	ABCD	0.813 (0.767 to 0.855)	0.533 (0.458 to 0.609)
	ABCDE	0.790 (0.754 to 0.823)	0.495 (0.440 to 0.545)

Table 8-9 Estimates of the accuracy (sensitivity and specificity) and number of studies (n) for the final combinations of DD and WS.

8.8.4 The best combination of DD and WS: the clinical perspective

As discussed in section 8.8.3, sensitivities and specificities were not sufficient to choose between the available diagnostic strategies even when represented on an ROC plane with credible regions or sROC curves. An alternative approach to an effective choice under the clinical perspective is to change the general objective of the strategy as already mentioned in Chapter 7. This means that the objective of a diagnostic strategy is no more to correctly diagnose the presence or absence of a condition, which does not consider the bivariate nature of diagnosis: in the majority of cases a better sensitivity does not correspond to a better specificity. Two possible objectives that better would address the role of diagnostic tests are (Pepe 2003):

- To identify and exclude as many healthy patients as possible in order to avoid the consequences of a non treated and diseased patient. This can be achieved by using a highly sensitive test, that corresponds to a high negative predictive value for a given level of the prevalence (see chapter 3 for relation between predictive values and prevalence).
- To identify and treat as many diseased as possible in order to avoid the effect potentially harmful treatments. This corresponds to having a high specificity (i.e. high positive predictive value).

Finally, under a clinical perspective, none of the diagnostic strategies that can be identified are likely to be clinically dominant, where a dominant strategy has both diagnostic rates higher than any of the other strategies for all the plausible levels of prevalence, especially because the consequences of testing are not considered

(i.e. invasiveness, eventual side effects). Moreover, none of these choices fully considers the trade-off between false negatives and false positives in terms of future consequences of a misdiagnosis. Then the inclusion of more information like the effects and costs as consequences of each diagnostic strategy can be used to choose the best strategy.

8.9 Cost-effectiveness analysis of combinations of Ddimer and Wells score for DVT

The meta-analytical model developed and applied in this chapter allows for the comparison of combinations of tests and accounts for dependence between the tests in each combination. However, none of the combinations seems to be sensibly better than the others (i.e. both sensitivity and specificity very high) considering that the aim proposed to combine tests is to increase either sensitivity (to exclude safely negative patients) or specificity (to treat early as many diseased as possible). Combinations that maximize either sensitivity (region 2 on Figure 8-6) or specificity (region 3 on Figure 8-6) have been found. However, such methodology does not consider the trade off between false positives and false negatives in terms of costs of the diagnostic procedures and consequent treatment regimes, the costs and the effects derived by the occurrence of clinical events (i.e. bleeding). Such meta-analytical approach can be used to inform an economic decision framework that considers such factors and leads to a more informed decision. Under this perspective, the best strategy after the cost-effectiveness analysis is likely to be very different form the strategy that it could be selected according to the clinical perspective.

8.9.1 Structure and parameters of the decision model

This section presents an application of the modeling techniques developed in the previous chapter to inform a comprehensive decision model. Such model is an

adaptation of an existing cost effectiveness analysis, presented by Sutton at al (Sutton, Cooper et al. 2008), for the accuracy of diagnostic strategies for the diagnosis of DVT. The diagnostic strategies in the original article were composed by single tests or by no tests at all: Ddimer, Ultrasound, discharge without test, treat without test. The last two represents the extreme diagnostic choices, that is assuming that every patient is negative (i..e discharge without test) or that every patient is positive (treat without test). Such decision modeling framework is adapted in order to compare the diagnostic strategies described in section 8.6.

Such economic evaluation considers a simplified diagnosis-to-treatment pathway for DVT. It assumes that positive patients are treated with anticoagulants, which potentially may cause armful side effects such as bleeding at different intensities (i.e. false and true positive patients may be subject to non fatal bleeding, fatal intracranial bleeding, non fatal intracranial bleeding or no bleeding when treated with anticoagulant). Of course, true positive patients which do not suffer any bleeding may still develop pulmonary embolism as a consequence of the thrombosis.

The structure or the model is a decision tree described in Sutton et al (Sutton, Cooper et al. 2008) and adapted for combinations of tests (Figure 8-8). Accuracy parameters are informed by the meta-analysis model developed and presented above in this chapter, parameters other than those for the diagnostic accuracy (i.e. prevalence of DVT, risk of pulmonary embolism, quality of life adjusted life years

per each possible health status, etc) have been described in Goodacre et al (Goodacre, Wailoo et al. 2006).

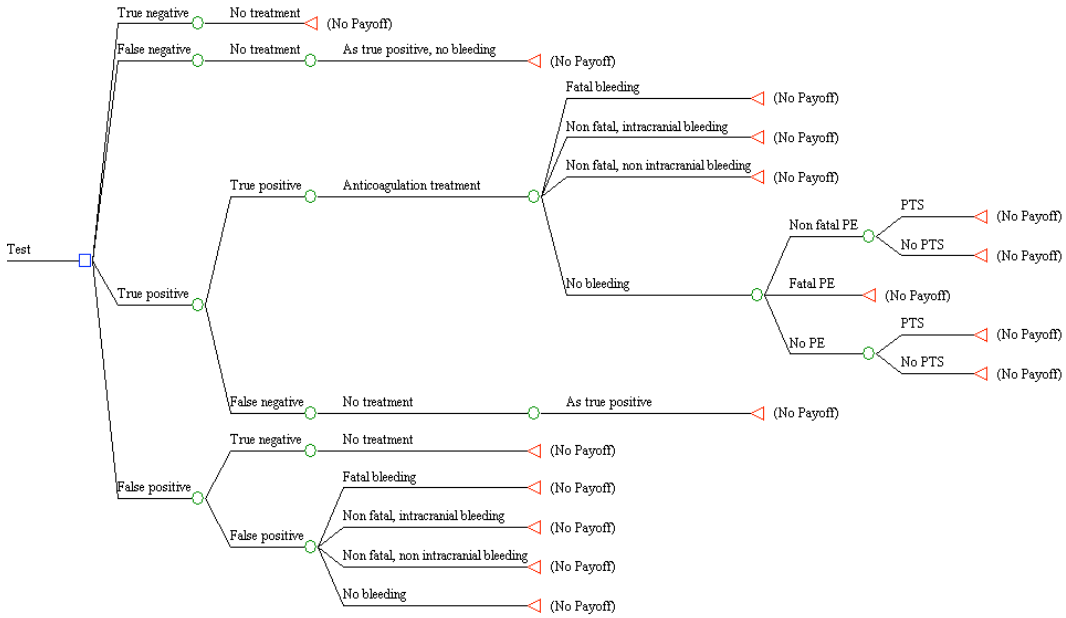


Figure 8-8 Decision tree for the economic decision model, adapted for the inclusion of a combination of a couple of diagnostic tests (i.e. in this case the *"believe the negatives*" combination scheme is used, similarly the tree can be adapted for the other schemes).

8.9.2 The best combination of DD and WS: the cost-effectiveness analysis

As mentioned in Chapter 4, comprehensive decision modeling consists of 4 steps:

1. *Develop the decision model*. A framework for comprehensive cost-effectiveness analysis for the accuracy of diagnostic tests has been recently presented by Sutton at al. (Sutton, Cooper et al. 2008).
2. *Systematic review of the relevant data and its meta-analysis*. This stage has been exhaustively described in this chapter for the accuracy of DD and WS for the diagnosis of DVT.
3. Estimation of all others inputs parameters:
    a. Effectiveness of the treatments considered in the evaluation;
    b. Transition probabilities from a health status to another;
    c. Costs and quality of life relative to different health stata.

    Measures of the effects (QALY), costs and transition probabilities as a consequence of the different diagnostic outputs are already given in the reference paper (Goodacre, Wailoo et al. 2006).
4. Evaluation of the (comprehensive decision) model in one coherent piece of code implemented in the software WinBUGS and sensitivity analyses.

Figure 8-9 represents the Cost Effectiveness Acceptability Curves (CEAC) for each diagnostic strategy identified in section 8.5 when their accuracy is estimated

assuming dependence between tests, independence between tests, and uncertainty around the accuracy estimates is explained by predictive intervals.

Many strategies under different assumptions do not result cost-effective for any values of the cost-effectiveness threshold (i.e. their CEAC are mostly overlaid on top of the the x-axis – Figure 8-9). When independence is assumed (lower-left CEAC plot in Figure 8-9), strategy 1 $(WS_{t1} and\ DD)_{BN}$ seems to be likely to be the most cost-effective at a cost-effectiveness threshold between 20,000£ and 30,000£, with a probability of circa 80%. However, when the model accounts for dependence (upper-left CEAC plot in Figure 8-9), the best strategy seems to be DD alone (strategy 5), with a probability of about 40% which highlights that there is much more uncertainty around this decision. This result leads to a much different decision compared to the assumption of independence, which shows the importance of our new modeling approach. Given the great amount of uncertainty when dependence is allowed for, an alternative decision may be to invest more resources in producing more information to inform this decision, for example running a larger trial for the accuracy of DD and WS in combination.

The use of predictions allows the unexplained heterogeneity to be included as part of the uncertainty into parameter estimation, this means that such variability propagates into the decision models and is reflected into the CEACs. In fact, when predictions are considered, the CEAC appear in the same order of importance, but the probability associated to each strategy results deflated as a consequence of the greater uncertainty that characterises the parameter estimates.

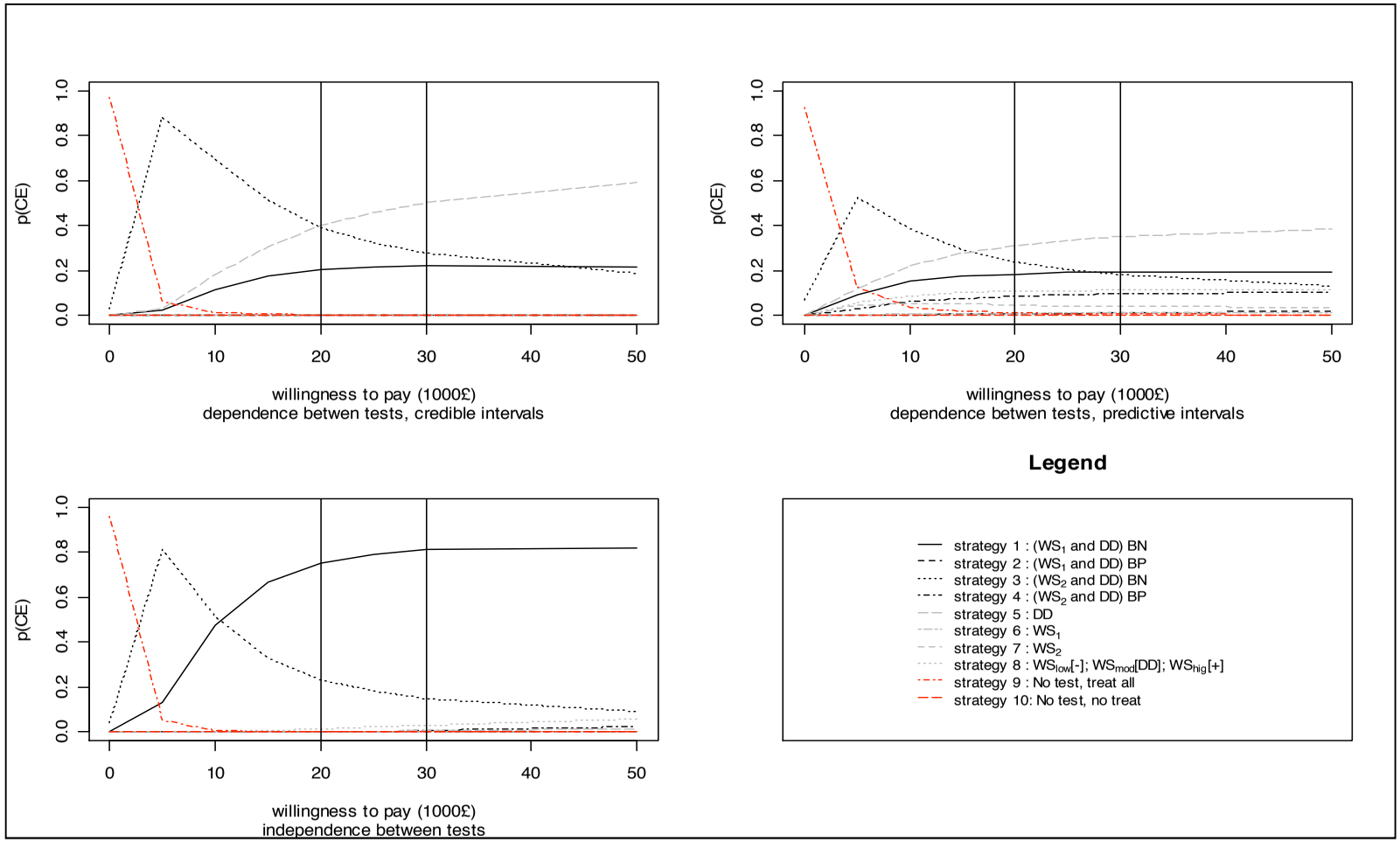


Figure 8-9 Cost effectiveness acceptability curves under different assumptions: dependence between tests using credible regions, independence between tests using credible regions, dependence between tests using predictive regions.

8.10 Discussion of model results

The Bayesian modeling framework described in section 8.7 produces results at 2 levels. First, the estimates of the intermediate parameters and their random effect parameters relatively, which refer to the conditional accuracy of WS and DD. Second, the accuracy of the combinations of WS and DD; these are calculated as functions of the intermediate parameters and uncertainty propagates to the final parameters. Random effect parameters are not directly estimated for the final parameters but they can be included in the final parameter uncertainty if predictions are calculated.

The inclusion of more data sometimes has changed the parameters estimates (in both cases of intermediate and final parameters). This is a very important characteristic of this modeling framework.

The best strategy is not always possible to be identified as the most accurate. In fact, when tests are combined, higher sensitivity corresponds to lower specificity. Therefore, the inclusion of economic information into the economic decision model can be helpful to consider the trade-off between false positives and false negatives. The CEAC plot points to DD alone as the best strategy, however with only 40% chance of being the best strategy, indicating that *i*) further information may be considered into the decision model, and *ii*) more single studies are needed to investigate the conditional accuracy of DD and WS.

The assumption of independence has produced very different cost-effectiveness results, which lead to the question whether it is worth assuming independence at all, even when conditional accuracy data are not available. In this case, perhaps a study for the evaluation of the conditional accuracy may lead to more correct conclusions.

As this thesis does not aim to find the best solution to the problem of the diagnosis of DVT, but offers a methodological basis for the choice of sequences of tests accounting for the dependence of the tests in the combination, there is one assumption that needs to be clarified. It is assumed that the list of strategies is exhaustive (I am not considering all the possible and/or plausible combinations of the existing tests) and the strategies are complete (positives are treated and negatives are discharged). In fact, the strategies considered are not necessarily complete. They may be better considered as triage strategies, that is further testing can be considered according to the objective as explained above (i.e. a strategy which aims to exclude as many healthy patients as possible produces a low positive predictive value; thus, positive patients can be further tested with a highly specific test to improve the overall accuracy of the sequence). Overall, this strategy represents the best trade-off between the consequences of being a false positive or a false negative.

A remark on the quality of the studies included in the analysis needs to be done.

Only a crude quality assessment on the quality of reporting of data has been done.

8.11 Summary

The problem of a correct estimate of the accuracy of sequences of diagnostic tests is widely undervalued in the evidence synthesis context. Independence is usually assumed between tests but this assumption can be heavily misleading. This chapter represents one of the first attempts to allow for dependence between tests in meta-analysis. Systematic reviews on the use of a number of tests simultaneously are not available, thus this chapter initiated with an *ad-hoc* systematic review for the assumption of DD and WS for DVT.

An important distinction has been specified for the parameter directly informed by the data (intermediate) and the parameter of interest (final). The intermediate parameters have been estimated by fitting to the data a multiple component model with shared parameters. The result produced within this modeling framework lead to different conclusions than those that can be produced by assuming independence between tests.

In the majority of cases, it is impossible to identify the clinically dominant strategy. Therefore, the trade-off between sensitivity and specificity (or equivalently between predictive values) needs to be evaluated considering all the chain of clinical and potentially economic consequences that derives from the choice of a particular diagnostic strategy. This can be achieved via decision modelling techniques. Recently, comprehensive cost-effectiveness analysis has been proposed as the ideal tool for modelling a decision in these circumstances.

The same decision model has been adapted to the choice of the best combination of DD and WS.

The code for the meta-analysys models and for the cost-effectiveness analysis, assuming either independence or dependence between tests, is included in the folder "Chapter 8 - combinations of WS and DD for DVT" contained in the CD-ROM attached to this thesis.

Chapter 9. Discussion and directions for further development

9.1 Overview of the thesis

This thesis reviews and where appropriate improves methods for the meta-analysis of diagnostic accuracy for dichotomised test results, and develops a framework for the meta-analysis of accuracy data from combinations of tests for their inclusion into a comprehensive decision model framework.

All models where implemented in WinBUGS (Lunn, Thomas et al. 2000) software for Bayesian statistical analysis and, where needed, checked for convergence and sensitivity to prior assumptions which may bias the results of MCMC based models.

The focus of this thesis is mainly methodological and an overview of the methodology is given in section 9.1.1.

9.1.1 Overview of the methodology

The most used models for the accuracy of dichotomised diagnostic test data have been presented in the first part of this thesis and were compared when implemented in a Bayesian data analysis framework. Although every model is based on a different set of assumptions, they have been related to each other in

Chapter 4 and all resulted as either equivalent or special cases of the bivariate model. Which model should be used cannot be established a priori: the assumption behind the model must be met by the data. However, a systematic review of these methods used for meta-analysis and to inform decision models in Chapter 6 found that simple methods are more likely to be used. To date, testing the assumptions behind the model is not always straightforward for the meta-analysis of diagnostic tests for the bivariate nature of the data; moreover, models are fitted on different parameterizations of the data, which complicates the comparison between them, especially when covariates are used to explore residual heterogeneity. Therefore, Chapter 5 has proposed the use of the Bayesian model choice statistics DIC (described in Chapter 2) as compared to residual deviance to chose the best fitting model and to the inclusion of covariates.

A further finding of the systematic review presented in Chapter 6 concerns the assumptions behind the estimates of the accuracy of combinations of tests: rarely combinations are considered, and, when they are considered, independence between tests is assumed in a systematic review meta-analysis context (i.e. individual studies may have considered dependence between tests). Therefore, Chapter 7 and Chapter 8 are dedicated to the exploration of the accuracy of combinations of two diagnostic tests and to the development of a meta-analytic framework for their accuracy. Such a modelling framework is based firstly on the estimates of the conditional accuracies that characterise the combination of the

test (intermediate parameters), and secondly on the estimates of the accuracy of different types of combinations expressed as a functions of the intermediate parameters. Such modelling framework is based on multiple equations (for different kinds of data and for the different tests at different levels in the combination) with shared parameters.

9.2 Contributions to knowledge

This thesis contains a number of contributions to knowledge, some of which directly resulted from the direction taken for the main investigation, and some that were results of the application of the methodologies explored herein.

Firstly, the direction of research was to explore the meta-analytical approaches to dichotomised diagnostic data. Although some of these methods were initially built in a Bayesian framework, their implementation in such a framework has been explored (WinBUGS code made available) for the first time with respect to the relationships between such models. Also, for the first time the use of the DIC was proposed to choose between meta-analytical models of the accuracy of diagnostic tests. The asymmetric sROC model initially developed by Littenberg and Moses (see section 4.5.2) has been criticised for not considering uncertainty in both parameters S and D; a parameterization that overcomes this inconvenience

is proposed. It is also described how to plot credible and predictive regions for when such models are implemented in WinBUGS.

The second direction for investigation was not clear from the beginning, and resulted from the systematic review of HTA reports of diagnostic accuracy presented in Chapter 6. Briefly, the major finding of this systematic review was that simplistic methods to the meta-analysis of diagnostic data are used more often to inform economic evaluations. Another finding that was secondary to the chatper but important to this thesis, was that tests are rarely considered in combination, and in the case they were considered in combinations the assumption of independence between tests was usually made.

Consequently, the second direction for research was to develop a modelling framework to the meta-analysis of combination of diagnostic tests. Probably for the first time it has been presented a systematic review that specifically concerns the accuracy of a combination of two tests. The meta-analysis of Wells score was performed using a multinomial logistic model as presented in Chapter 8.

Finally, the major contribution to the knowledge resulting form this thesis is the modelling framework developed, applied and presented in Chapter 8. Also, it has been shown how to make predictions and how to apply this model into a comprehensive decision model.

9.3 Discussion and limitations

The accuracy of diagnostic tests is crucial to maximize the efficacy of treatments. Meta-analysis techniques allow differences between different settings to be accounted for; however, a number of meta-analytical approaches exist that are based on a number of assumptions. The most of these methods allow sROC curves to be plotted on the ROC plane even where pairs of sensitivities and specificities have not been recorded. Graphical representations of such sROC curves can be constrained to the range of variation of the accuracy data. Alternatively, credible regions may be used to represent graphically the accuracy of diagnostic strategies; credible regions account for correlation between diagnostic rates due to variability in the diagnostic threshold and also represent the residual heterogeneity in the meta-analysis dataset.

When tests are combined, the assumption of conditional independence between the tests may not hold, therefore meta-analysis methods need to allow for dependence between tests. The approach that is proposed also estimated the conditional accuracy of the tests. However, the evaluation may become very complicated and data may not be available. This highlights the need of studies that consider a number of the possible tests on the same population so that the assumption of independence can be relaxed without the need of a number of strong assumptions based on real data, especially if the tests being treated are

cheap and non invasive and their combination can be used as a triage diagnostic strategy (see section 3.2.2 for a definition of triage diagnostic strategy).

Finally, the issue of imperfect reference test has not been investigated in this thesis where the reference test was assumed to be perfect for all examples. A simulation exercise may be helpful to investigate the impact of such an assumption.

An important limitation that needs to be mentioned for the meta-analysis methods listed in section 4.5. Some of these approaches are based on assumptions which may not be necessarily true. For example, models based on an estimate of the pooled DOR assume that the log(DOR) is normally distributed; however, there is some evidence that the log(DOR) is not necessarily normally distributed but its distribution may be asymmetrical when its values are further away from zero due to small study effect (Sterne, Gavaghan et al. 2000). The fixed DOR is usually very distant form zero, being sensitivity usually much higher then 1-specificity. The Bayesian methods presented in Chapter 4 are flexible and other distributions may be used for the symmetric approach.

9.4 Directions for further development and Conclusions

Individual Patient data (IPD) data for the accuracy of diagnostic tests has not been considered in the modelling approach developed in Chapter 8. Where IPD is available, this could be included into the modelling approach developed in Chapter 8 by means of Bernoulli likelihoods (Riley, Dodd et al. 2008). Reporting of IPD should be encouraged by means of guidelines.

Diagnostic tests have clinical utility as they are used to detect diseases. However, different types of test can be used to explore the same set of symptoms which may be at the basis of a set of different diseases. Therefore, the methods to investigate meta-analytically the accuracy of a (single or combination of) test could be adapted to consider the range of diseases that that test may indicate.

An assumption that is usually made in the field of diagnostic test is that presented in Chapter 3: the distributions of the test measurements for the diseased and the healthy patients are normally distributed and diseased patients usually are characterised by higher values of the test. This assumption is difficult to test unless individual patient test measurements are available (i.e. very difficult for qualitative or imaging tests). This assumption is currently impossible to test in case of qualitative diagnostic tests. The impact of this characteristic of the data on the meta-analysis modelling approaches should be investigated, for example via a

simulation study (i.e. what is the impact on the meta-analysis modelling approaches if patients are not normally distributed (for example, measurements of an enzyme in the blood may be right skewed) or if diseased patients are characterised by lower values of the test? Such simulation should consider a quantitative test, however results could than be generalised for imaging tests for which the populations of diseased and healthy are assumed to be normally distributed over the underlying threshold too.

The meta-analytical model for the accuracy of combinations of tests presented in Chapter 8 could be generalised for more than two tests, although there may be very little, if any, data. Also, such approach can be generalised for publications reporting the accuracy of tests with a continuous thresholds at different levels of such threshold. Moreover, the use of covariates for the exploration of heterogeneity should be explored.

In conclusion, the awareness that meta-analyses are important tools for the robust estimation of the diagnostic accuracy is increasing (Jones and Athanasiou 2009) and, more complex methods are indicated as more appropriate to capture the complexity behind the estimation of diagnostic accuracy. In fact, the estimation of accuracy of diagnostic strategies including more than one test via meta-analytic techniques needs to consider correlation between the tests included in the strategy. In the main example (DVT) used in this thesis, correlation between the tests WS and DD was evident and the unexplained heterogeneity large. Ignoring such

correlation would lead to biased accuracy estimates and wrong decisions based on economic evaluations. The use of random effect modelling allows for the quantification of the residual heterogeneity. The model that is developed in this thesis accounts for conditional dependence between the tests. The Bayesian framework gives the amount of flexibility needed to develop such types of modelling approaches. More work is needed to fully explore the potential and the properties of such modelling approach.

Appendix A. Search strategy of the systematic review of the accuracy of DDimer and Well score used in combination (Chapter 8)

Database: EMBASE, Ovid MEDLINE(R)

Period: since inception to 03/2011

Sub-Search Strategy for accuracy of diagnostic test studies:

1 exp "Sensitivity and Specificity"
2 exp diagnostic errors
3 reference values.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, nm, ui]
4 reproducibility of results.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, nm, ui]
5 likelihood functions.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, nm, ui]
6 1 or 2 or 3 or 4 or 5
7 specificity.af.
8 sensitivity.af.
9 false negative$.af.
10 false positive$.af.
11 true positive$.af.
12 true negative$.af.
13 predictive value$.af.
14 reproducibility.af.
15 ROC curve.af.
16 diagnos$.ti.
17 reference value$.af.
18 likelihood function$.af.
19 likelihood ratio$.af.
20 11 or 7 or 9 or 17 or 12 or 15 or 14 or 8 or 18 or 19 or 16 or 10 or 13

Sub-Search Strategy for Deep Vein Thrombosis:

21 exp venous thrombosis
22 exp deep vein thrombosis
23 exp phlebothrombosis
24 venous thrombosis.af.
25 venous thromboembolism.af.
26 deep venous thrombosis.af.
27 DVT.af.
28 26 or 24 or 21 or 27 or 23 or 25 or 22

29 6 or 28 or 20

Sub-Search Strategy for diagnostic algorithms, combinations of test:

30 algorithm$.af.
31 clinical protocol$1.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, nm, ui]
32 algorithm$1.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, nm, ui]
33 protocol$.af.
34 diagnostic strateg$3.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, nm, ui]
35 diagnostic combination$1.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, nm, ui]
36 combination$1 of test$1.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, nm, ui]
37 sequence$1 of test$1.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, nm, ui]
38 diagnostic sequence$1.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, nm, ui]
39 sequence$1 of diagnostic test$1.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, nm, ui]
40 combination$1 of diagnostic test$1.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, nm, ui]
41 sequence$3.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, nm, ui]
42 combination$1.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, nm, ui]
43 algorithm$1.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, nm, ui]
44 management.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, nm, ui]
45 35 or 33 or 32 or 39 or 40 or 36 or 41 or 42 or 38 or 34 or 30 or 37 or 43 or 44 or 31

Sub-Search Strategy for Ddimer test:

46 d-dimer$.af.
47 ddimer$.af.
48 enzyme-linked immunosorbent assay$1.af.
49 simplired.af.
50 ELISA.af.
51 LATEX.af.
52 whole blood agglutination.af.
53 vidas.af.
54 vidas.af.
55 turbidimeter.af.
56 turbidimetric.af.
57 50 or 53 or 51 or 48 or 47 or 52 or 56 or 46 or 49 or 55 or 54

Sub-Search Strategy for Wells score:

58 WELLS score.af.
59 clinical probability.af.
60 pre test probability.af.
61 clinical probability model.af.
62 pre test clinical probability.af.
63 "WELLS test".mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, nm, ui]

64 "clinical score".mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, nm, ui]
65 clinical assessment.af.
66 clinical scoring system.af.
67 clinical probability assessment.af.
68 clinical assessment.af.
69 standardized model.af.
70 pretest probability.af.
71 clinical model.af.
72 clinical probability score.af.
73 clinical evaluation$1.af. (
74 67 or 63 or 71 or 70 or 68 or 72 or 65 or 64 or 61 or 58 or 59 or 69 or 60 or 66 or 73 or 62

Final combination of results from the previous sub-search strategies:

75 74 and 57 and 29
76 remove duplicates from 75

Appendix B- References to the studies included in the meta-analysis, (Chapter 8 table 3)

Type A

T1. Shields, P., S. Turnipseed, E. Panacek, N. Melnikoff, R. Gosselin, and R. White. (2002). "Validation of the canadian clinical probability model for acute venous thrombosis." Acad Emerg Med **9**(6):561-566.

T2. Lennox, A., K. Delis, S. S, Z. Zarka, S. Daskalopoulou, and A. Nicolaides. (1999). "Combination of a clinical risk assessment score and rapid whole blood d-dimer testing in the diagnosis of deep vein thrombosis in symptomatic patients." Journal of vascular surgery **30**:794-804.

T3. Kearon, C., J. Ginsberg, J. Douketis, M. Crowther, P. Bill-Edwards, J. Wietz, and J. Hirsh (2001). "Management of suspected deep vein thrombosis in poutpatients by using clinical assessment and ddimer testing." Annals of internals medicine **135** (2):108-111.

T4. Ruiz-Gimenez, N., A. Friera, P. Artieda, P. Caballero, P. Sanchez Molini, M. Morales, and C. Suarez (2004). "Rapid d-dimertest combined a clinical model for Deep Vein Thrombosis." Thrombosis and Haemostasis **91**:1237-1246.

T5. Yamaki, T., M. Nozaki, H. Sakurai, M. Takeuchi, K. Soejima, and T. Kono (2005). "Prospective evaluation of a screening protocol to exclude deep vein thrombosis on the basis of a combination of quantitative d-dimer testing and pre-test clinical probability score." journal of American college of surgeons **201**:701-709.

T6. Anderson, D., P. Wells, I. Stiell, B. MacLeod, M. Simms, L. Gray, K. Robinson, J. Bormanis, M. Mitchell, L. Bernard, and G. Flowerdew (2000). "Management of patients with suspected deep vein thrombosis in the emergency department: combining use of a clinical diagnosis model with d-dime testing". The journal of emergency medicine **19** (3):225-230.

T7. Anderson, D., M. Kovacs, G. Kovacs, I. Stiell, M. Mitchell, V. Khoury, J. Dryer, J. Ward, and P. Wells (2002). "Combined use of clinical assessment and D-dimer to improve the management of patients presenting to the emergency department with suspected deep vein thrombosis (the EDITED study)." Journal of Thrombosis and Haemostasis **1**:645-651.

T8. Bates, S., C. Kearon, M. Crowther, L. Linkins, M. O'Donnell, J. Douketis, A. Lee, J. Weitz, M. Johnston, and J. Ginsberg (2003). "A diagnostic strategy

involving a quantitative latex D-dimer assay excludes deep vein thrombosis." Annals of internal medicine **138**:787-794.

T9. Rio Solá, M., J. Gonzalez Fajardo, M. Martin Pedrosa, V. Gutierrez, S. Carrera, and C. Vaqueto Puerta (1999). "Evaluacion clinica del dimero-D en el diagnostico de infermedad tromboembolica venosa." Angiologia **6**:251-258.

T10. Williams, D., A. Lee, H. Clark, J. Webster, and H. Watson (2005). "A comparison of computerised strain gauge plethismography with ddimer testing in screening for deep vein thrombosis." British journal of haematology **131**:253-257.

T11. Yamaki, T., M. Nozaki, H. Sakurai, Y. Kikuchi, K. Soejima, T. Kono, A. Hamahata, and K. Kim (2009). "Combined use of pre-test clinical probability score and latex agglutinatino d-dimer testing for excluding acute deep vein thrombosis." journal of vascualr surgery **50**:1099-1105.

Type B

T12. Borg (1997). "Rapid quantitative d-dimer assay and clinical evaluation fot the diagnosis of clinically suspected Deep Vein Thrombosis." Thrombosis and Haemostasis **77**(3):600-609.

T13. Dewar, C., C. Selby, K. Jamieson, and S. Rogers (2008). "Emergency department nurse-based outpatients diagnosis of DVT using an evidence-based protocol." Journal of emergency medicine **25**:441-416.

T14. Elf, J., K. Strandberg, C. Nilsson, and P. Svensson (2009). "Clinical probability assessment and ddimer determination in patients with suspected deep vein thrombosis, a prospective multicenter management study." Thrombosis research **123**:612-616.

Type C

T15. Aguilar Franco, C., A. Martinez Benedicto, A. Martinez Santabarbara, C. del Rio Mayor, V. del Villar Sordo, M. Vazquez Salvado, and F. Rodriguez Recio (2002)a. "Valor diagnostico del dimero-D enpacientes con baja probabilidad clinical de thrombosis venosa profunda en miembros inferiores." Med Clin (Barcelona) **118**(14):539-542.

T16. Walsh, K., N. Kelaher, K. Long, and P. Cervi (2009). "An algorithm for the investigation and management of patients with suspected deep venous thrombosis at a district general hospital." Postgrad Med J **78**:742-745.

T17. Aguilar Franco, C., A. Martinez Benedicto, A. Martinez Santabarbara, C. del Rio Mayor, M. Vazquez Salvado, and F. Rodriguez Recio (2002)b. "Diagnostic value of Ddimer i npatients with a moderate pretest probability of deep venous thrombosis." British journal of haemostasis **118**:275-277.

T18. Bucek, R., N. Koca, M. Reiter, M. Haumer, T. Zontsich, and E. Minar (2002). "Algorithms for the diagnosis of deep vein thrombosis inpatients with low clinical pre-test probability." Thrombosis research **105**:43-47.

Type D (Wells Score used individually)

T19. Anderson, D.R., et al. (1999). "Thrombosis in the emergency department: Use of a clinical diagnosis model to safely avoid the need for urgent radiological investigation." Archives of Internal Medicine. **159**(5): 477-482.

T20. Arrivé, L., et al. (2002). "Combination of rapid D-Dimer testing and simple clinical model for the diagnosis of deep vein thrombosis." J Radiol, **83**(3): 337-340.

T21. Bozic, M., A. Blinc, and M. Stegnar. (2002). "D-dimer, other markers of haemostasis activation and soluble adhesion molecules in patients with different clinical probabilities of deep vein thrombosis." Thrombosis Research, **108**(2-3): 107-114.

T22. Bucek, R.A., et al. (2001). "Results of a new rapid D-dimer assay (Cardiac D-Dimer) in the diagnosis of deep Vein thrombosis." Thrombosis Research, **103**(1): 17-23.

T23. Constans, J., et al. (2003). "Comparison of four clinical prediction scores for the diagnosis of lower limb deep venous thrombosis in outpatients." American Journal of Medicine, **115**(6): 436-440.

T24. Constans, J., et al. (2001). "Clinical prediction of lower limb deep vein thrombosis in symptomatic hospitalized patients." Thrombosis and Haemostasis, **86**(4): 985-990.

T25. D'Angelo, A., et al. (1996). "Evaluation of a new rapid quantitative D-dimer assay in patients with clinically suspected deep vein thrombosis." Thrombosis and Haemostasis, **75**(3): 412-416.

T26. Dryjski, M., et al. (2001). "Evaluation of a screening protocol to exclude the diagnosis of deep venous thrombosis among emergency department patients." Journal of Vascular Surgery, **34**(6): 1010-1015.

T27. Fisher, B.W., S.R. Majumdar, and F.A. McAlister (2002). "Clinical prediction of deep venous thrombosis using two risk assessment methods in combination with rapid quantitative D-dimer testing." American Journal of Medicine, **112**(3): 198-203.

T28. Funfsinn, N., et al. (2001). "Rapid D-dimer testing and pre-test clinical probability in the exclusion of deep venous thrombosis in symptomatic outpatients." Blood Coagulation and Fibrinolysis, **12**(3): 165-170.

T29. Kilroy, D.A., et al (2003). "Emergency department investigation of deep vein thrombosis." Emergency Medicine Journal, **20**(1): 29-32.

T30. Kraaijenhagen, R.A., et al (2002). "Simplification of the diagnostic management of suspected deep vein thrombosis." Archives of Internal Medicine, **162**(8): 907-911.

T31. Miron, M.J., A. Perrier, and H. Bounameaux (2000). "Clinical assessment of suspected deep vein thrombosis: Comparison between a score and empirical assessment." Journal of Internal Medicine, **247**(2): 249-254.

T32. Oudega, R., K.G.M. Moons, and A.W. Hoes (2005). "Limited value of patient history and physical examination in diagnosing deep vein thrombosis in primary care." Family Practice, **22**(1): 86-91.

T33. Ruiz-Gimenez, N., et al (2002). "Deep venous thrombosis of lower extremities in an emergency department. Utility of a clinical diagnosis model." Medicina Clinica, **118**(14): 529-533.

T34. Wells, P., et al (1997). "Value of assessment of pretest probability of deep-vein thrombosis in clinical management." Lancet, **350**(9094):1795-1798.

T35. Wells, P.S., et al (1999). "Application of a diagnostic clinical model for the management of hospitalized patients with suspected deep-vein thrombosis." Thrombosis and Haemostasis, **81**(4): 493-497.

T36. Wells, P.S., et al (1995). "Accuracy of clinical assessment of deep-vein thrombosis." Lancet, **345**(8961): 1326-1330.

Type E (Ddimer used individually)

T37. Kilroy, D.A., et al (2003). "Emergency department investigation of deep vein thrombosis." Emergency Medicine Journal, **20**(1): 29-32.

T38. Gosselin, R.C., et al (2003). "Comparison of six D-dimer methods in patients suspected of deep vein thrombosis." Blood Coagulation and Fibrinolysis, **14**(6): 545-550.

T39. Cini, M., et al (2003). "A new rapid bedside assay for D-dimer measurement (Simplify D-dimer) in the diagnostic work-up for deep vein thrombosis." Journal of thrombosis and haemostasis: JTH, **1**(12): 2681-2683.

T40. Bucek, R.A., et al (2003). "Thrombus precursor protein, endogenous thrombin potential, von-Willebrand factor and activated factor VII in suspected deep vein thrombosis: Is there a place for new parameters?" British Journal of Haematology, **120**(1): 123-128.

T41. Walsh, K., et al (2002) "An algorithm for the investigation and management of patients with suspected deep venous thrombosis at a district general hospital." Postgraduate Medical Journal, **78**(926): 742-745.

T42. Shitrit, D., et al (2002) "Appropriate indications for venous duplex scanning based on D-dimer assay." Annals of Vascular Surgery, **16**(3): 304-308.

T43. Shields, G.P., et al (2002). "Validation of the Canadian clinical probability model for acute venous thrombosis." Academic Emergency Medicine, **9**(6): 561-566.

T44. Schutgens, R.E.G., et al (2002). "Usefulness of a semiquantitative D-dimer test for the exclusion of deep venous thrombosis in outpatients." American Journal of Medicine, **112**(8): 617-621.

T45. Miranda, C. and M. Harden (2002). "Comparison of D-dimer findings with venous duplex examinations: A work in progress." Journal of Vascular Technology, **26**(2): 103-105.

T46. Larsen, T.B., et al (2002). "Validity of D-dimer tests in the diagnosis of deep vein thrombosis: A prospective comparative study of three quantitative assays." Journal of Internal Medicine, **252**(1): 36-40.

T47. Johanning, J.M., et al (2002). "D-dimer and calf circumference in the evaluation of outpatient deep venous thrombosis." Journal of Vascular Surgery, **36**(5): 877-880.

T48. Gosselin, R.C., et al (2002). "Evaluation of a new automated quantitative d-dimer, advanced D-dimer, in patients suspected of venous thromboembolism." Blood Coagulation and Fibrinolysis, **13**(4): 323-330.

T49. Fisher, B.W., S.R. Majumdar, and F.A. McAlister (2002). "Clinical prediction of deep venous thrombosis using two risk assessment methods in combination with rapid quantitative D-dimer testing." American Journal of Medicine, **112**(3): 198-203.

T50. Bucek, R.A., et al (2002). "C-reactive protein in the diagnosis of deep vein thrombosis." British Journal of Haematology, **119**(2): 385-389.

T51. Bucek, R.A., et al (2002). "Algorithms for the diagnosis of deep-vein thrombosis in patients with low clinical pretest probability." Thrombosis Research, **105**(1): 43-47.

T52. Boziç, M., A. Blinc, and M. Stegnar (2002). "D-dimer, other markers of haemostasis activation and soluble adhesion molecules in patients with different clinical probabilities of deep vein thrombosis." Thrombosis Research, **108**(2-3): 107-114.

T53. Arrivé, L., et al (2002). "Combination of rapid D-Dimer testing and simple clinical model for the diagnosis of deep vein thrombosis." Journal de Radiologie, **83**(3): 337-340.

T54. Arancibia, F., et al (2002). "The clinical usefulness of D-dimer testing in cancer patients with suspected deep venous thrombosis." Archives of Internal Medicine, **162**(16): 1880-1884.

T55. Siragusa, S., et al (2001). "A rapid D-dimer assay in patients presenting at an emergency room with suspected acute venous thrombosis: Accuracy and relation to clinical variables." Haematologica, **86**(8): 856-861.

T56. Shitrit, D., et al (2001). "Diagnostic value of the D-dimer test in deep vein thrombosis: Improved results by a new assay method and by using discriminate levels." Thrombosis Research, **102**(2): 125-131.

T57. Harper, P.L., et al (2001). "The rapid whole blood agglutination d-dimer assay has poor sensitivity for use as an exclusion test in suspected deep vein thrombosis." New Zealand Medical Journal, **114**(1126): 61-64.

T58. Funfsinn, N., et al (2001). "Rapid D-dimer testing and pre-test clinical probability in the exclusion of deep venous thrombosis in symptomatic outpatients." Blood Coagulation and Fibrinolysis, **12**(3): 165-170.

T59. Dryjski, M., et al (2001). "Evaluation of a screening protocol to exclude the diagnosis of deep venous thrombosis among emergency department patients." Journal of Vascular Surgery, **34**(6): 1010-1015.

T60. Bucek, R.A., et al (2001). "Results of a new rapid D-dimer assay (Cardiac D-Dimer) in the diagnosis of deep Vein thrombosis." Thrombosis Research, **103**(1): 17-23.

T61. Bates, S.M., et al (2001). "A latex D-dimer reliably excludes venous thromboembolism." Archives of Internal Medicine, **161**(3): 447-453.

T62. Villa, P., et al (2000). "Quantification of D-dimer using a new fully automated assay: Its application for the diagnosis of deep vein thrombosis." Haematologica, **85**(5): 520-524.

T63. Van Der Graaf, F., et al (2000). "Exclusion of deep venous thrombosis with D-Dimer testing. Comparison of 13 D-Dimer methods in 99 outpatients suspected of deep venous thrombosis using venography as reference standard." Thrombosis and Haemostasis, **83**(2): 191-198.

T64. Trujillo-Santos, A.J., et al (2000). "Clinical analytic diagnostic assessment of deep vein thrombosis of lower limbs." Med Clin, **114**(2): 46-49.

T65. Sadouk, M., et al (2000). "Comparison of diagnostic performance of three new fast D-dimer assays in the exclusion of deep vein thrombosis." Clinical Chemistry, **46**(2): 286-287.

T66. Permpikul, C., et al (2000). "Whole Blood Agglutination D - Dimer Test for the Diagnosis of Deep Vein Thrombosis." Journal of the Medical Association of Thailand, **83**(7): 732-736.

T67. LaCapra, S., et al (2000). "The use of thrombus precursor protein, D-dimer, prothrombin fragment 1.2, and thrombin antithrombin in the exclusion of proximal deep vein thrombosis and pulmonary embolism." Blood Coagulation and Fibrinolysis, **11**(4): 371-377.

T68. Hein-Rasmussen, R., C.D. Tuxen, and N. Wiinberg (2000). "Diagnostic value of the Nycocard, Nycomed D-dimer assay for the diagnosis of deep venous thrombosis and pulmonary embolism: A retrospective study." Thrombosis Research, **100**(4): 287-292.

T69. Gosselin, R.C., et al (2000). "A new method for measuring D-dimer using immunoturbidometry: A study of 255 patients with suspected pulmonary embolism and deep vein thrombosis." Blood Coagulation and Fibrinolysis, **11**(8): 715-721.

T70. Farrell, S., T. Hayes, and M. Shaw (2000). "A negative simpliRED D-dimer assay result does not exclude the diagnosis of deep vein thrombosis or pulmonary embolus in emergency department patients." Annals of Emergency Medicine, **35**(2): 121-125.

T71. Bradley, M., J. Bladon, and H. Barker (2000). "D-dimer assay for deep vein thrombosis: Its role with colour doppler sonography." Clinical Radiology, **55**(7): 525-527.

T72. Anderson, D.R., et al (2000). "Management of patients with suspected deep vein thrombosis in the Emergency Department: Combining use of a clinical diagnosis model with D- dimer testing." Journal of Emergency Medicine, **19**(3): 225-230.

T73. Wells, P.S., et al (1999). "Application of a diagnostic clinical model for the management of hospitalized patients with suspected deep-vein thrombosis." Thrombosis and Haemostasis, **81**(4): 493-497.

T74. Wahlander, K., et al (1999). "Comparison of various D-dimer tests for the diagnosis of deep venous thrombosis." Blood Coagulation and Fibrinolysis, **10**(3): 121-126.

T75. Scarano, L., et al (1999). "Failure of soluble fibrin polymers in the diagnosis of clinically suspected deep venous thrombosis." Blood Coagulation and Fibrinolysis, **10**(5): 245-250.

T76. Lindahl, T.L., T.H. Lundahl, and S.G. Fransson (1999). "Evaluation of an automated micro-latex D-dimer assay (Tina-quant on Hitachi 911 analyser) in symptomatic outpatients with suspected DVT." Thrombosis and Haemostasis, **82**(6): 1772-1773.

T77. Lennox, A.F., et al (1999). "Combination of a clinical risk assessment score and rapid whole blood D- dimer testing in the diagnosis of deep vein thrombosis in symptomatic patients." Journal of Vascular Surgery, **30**(5): 794-804.

T78. Legnani, C., et al (1999). "Performance of a new, fast D-dimer test (IL Test, D-Dimer) for the management of outpatients with suspected deep vein thrombosis in emergency situations." Fibrinolysis, **13**(3): 139-141.

T79. Legnani, C., et al (). "Contribution of a new, rapid, quantitative and automated method for D- dimer measurement to exclude deep veto thrombosis in symptomatic outpatients." Blood Coagulation and Fibrinolysis, **10**(2): 69-74.

T80. Le Blanche, A.F., et al (1999). "Ruling out acute deep vein thrombosis by ELISA plasma D-dimer assay versus ultrasound in inpatients more than 70 years old." Angiology, **50**(11): 873-882.

T81. Del Rio Sola, M.L., et al (1999). "Clinical evaluation of D-Dimer in the diagnosis of thromboembolic disease." Angiology, **51**(6): 251-258.

T82. Carter, C.J., et al (1999). "Rapid fibrin D-dimer tests for deep venous thrombosis: Factors affecting diagnostic utility." Journal of Emergency Medicine, **17**(4): 605-610.

T83. Aschwanden, M., et al (1999). "The value of rapid D-dimer testing combined with structured clinical evaluation for the diagnosis of deep vein thrombosis." Journal of Vascular Surgery, **30**(5): 929-935.

T84. Wildberger, J.E., et al (1998). "Bedside testing (simpliRED) in the diagnosis of deep vein thrombosis: Evaluation of 250 patients." Investigative Radiology, **33**(4): 232-235.

T85. Wijns, W., et al (1998). "Evaluation of two D-Dimer assays in the diagnosis of venous thromboembolism." Acta Clinica Belgica, **53**(4): 270-274.

T86. Wells, P.S., et al (1998). "SimpliRED D-dimer can reduce the diagnostic tests in suspected deep vein thrombosis." Lancet, **351**(9113): 1405-1406.

T87. Mauron, T., et al (1998). "SimpliRED D-dimer assay: Comparability of capillary and citrated venous whole blood, between-assay variability, and performance of the test for exclusion of deep vein thrombosis in symptomatic outpatients." Thrombosis and Haemostasis, **79**(6): 1217-1219.

T88. Lindahl, T.L., et al (1998). "Clinical evaluation of a diagnostic strategy for deep venous thrombosis with exclusion by low plasma levels of fibrin degradation product D-dimer." Scandinavian Journal of Clinical and Laboratory Investigation, **58**(4): 307-316.

T89. Lee, A.Y.Y., et al (1998). "Diagnostic accuracy of SimpliRED (R) D-dimer testing in cancer patients with clinically suspected deep vein thrombosis." Blood, **92**: 177.

T90. Khaira, H.S. and J. Mann, Plasma (1998). "D-Dimer measurement in patients with suspected DVT - a means of avoiding unnecessary venography." European Journal of Vascular and Endovascular Surgery, **15**(3): 235-238.

T91. Escoffre-Barbe, M., et al (1998). "Evaluation of a new rapid D-dimer assay for clinically suspected deep venous thrombosis (Liatest D-dimer)." American Journal of Clinical Pathology, **109**(6): 748-753.

T92. Scarano, L., et al (1997). "Accuracy of two newly described D-dimer tests in patients with suspected deep venous thrombosis." Thrombosis Research, **86**(2): 93-99.

T93. Mayer, W., R. Hirschwehr, and H. Partsch (1997). "The D-dimer bedside test 'SimpliRED(R)' as diagnostic aid in suspected thrombosis." Vasomed, **9**(4): 256-257.

T94. Leroyer, C., et al (1997). "Diagnostic value of a new sensitive membrane based technique for instantaneous D-Dimer evaluation in patients with clinically suspected deep venous thrombosis." Thrombosis and Haemostasis, **77**(4): 637-640.

T95. Legnani, C., et al (1997). "Comparison of new rapid methods for D-dimer measurement to exclude deep vein thrombosis in symptomatic outpatients." Blood Coagulation and Fibrinolysis, **8**(5): 296-302.

T96. Kozman, H., M.C. Flemmer, and M. Rahnama (1997). "Deep venous thrombosis: Prediction by D-dimer?" Southern Medical Journal, **90**(9): 907-910.

T97. Knecht, M.F. and F. Heinrich (1997). "Clinical evaluation of an immunoturbidimetric D-dimer assay in the diagnostic procedure of deep vein thrombosis and pulmonary embolism." Thrombosis Research, **88**(5): 413-417.

T98. Killick, S.B., et al (1997). "Comparison of immunofiltration assay of plasma D-dimer with diagnostic imaging in deep vein thrombosis." British Journal of Haematology, **96**(4): 846-849.

T99. Janssen, M.C.H., et al (1997). "Factor VIIA determination compared to D-dimer in diagnosis of deep venous thrombosis." Thrombosis Research, **86**(5): 423-426.

T100. Janssen, M.C.H., et al (1997). "Reliability of five rabid D-Dimer assays compared to ELISA in the exclusion of deep venous thrombosis." Thrombosis and Haemostasis, **77**(2): 262-266.

T101. Jacq, F., et al (1997). "Evaluation of a rapid blood test for the exclusion of venous thromboembolism in symptomatic outpatients." Presse Med, **26**(24): 1132-1134.

T102. Guazzaloca, G., et al (1997). "Deep vein thrombosis. Validation of a non-invasive diagnostic procedure based on compression ultrasound sonography and measurement of D-dimer plasma level." Minerva Cardioang, **45**(6): 259-266.

T103. Fiessinger, J.N., et al (1997). "Rapid blood test for the exclusion of venous thromboembolism in symptomatic outpatients." Thrombosis and haemostasis, **77**(5): 1042-1043.

T104. Crippa, L., et al (1997). "Clinical pre test probability and D-dimer in the diagnosis of deep vein thrombosis." Thromb Haemost, June**5**:PD647.

T105. Borg, J.Y., et al (1997). "Rapid quantitative D-dimer assay and clinical evaluation for the diagnosis of clinically suspected deep vein thrombosis." Thrombosis and Haemostasis, **77**(3): 602-603.

T106. Leroyer, C., et al (1996). "Diagnostic value of plasma D-Dimer measurement, using ELISA test, in front of a suspected deep venous thrombosis." European Journal of Internal Medicine, **7**(2): 99-103.

T107. Gavaud, C., et al (1996). "Dosage of the D-dimers in the diagnosis of deep vein thrombosis and/or pulmonary embolism. Review based on 80 consecutive patients seen at an emergency unit." J Mal Vasc, **21**(1): 22-30.

T108. Elias, A., et al (1996). "D-Dimer test and diagnosis of deep vein thrombosis: A comparative study of 7 assays." Thrombosis and Haemostasis, **76**(4): 518-522.

T109. D'Angelo, A., et al (1996). "Evaluation of a new rapid quantitative D-dimer assay in patients with clinically suspected deep vein thrombosis." Thrombosis and Haemostasis, **75**(3): 412-416.

T110. Wells, P.S., et al (1995). "A novel and rapid whole-blood assay for D-dimer in patients with clinically suspected deep vein thrombosis." Circulation, **91**(8): 2184-2187.

T111. Legnani, C., et al (1995). "Validation of a new quantitative automated latex photometric immunoassay of plasma D-dimer (LPIA-100)." Thrombosis and Haemostasis, **73**: 1101.

T112. Brenner, B., et al (1995). "Application of a bedside whole blood D-dimer assay in tne diagnosis of deep vein thrombosis." Blood Coagulation and Fibrinolysis, **6**(3): 219-222.

T113. Bouman, C.S.C., S.T. Ypma, and J.P.H.B. Sybesma (1995). "Comparison of the efficacy of D-dimer, fibrin degradation products and prothrombin fragment 1+2 in clinically suspected deep venous thrombosis." Thrombosis Research, **77**(3): 225-234.

T114. Tengborn, L., et al (1994). "D-dimer and thrombin/antithrombin III complex-diagnostic tools in deep venous thrombosis?" Haemostasis, **24**(6): 344-350.

T115. Hansson, P.O., et al (1994). "Can laboratory testing improve screening strategies for deep vein thrombosis at an emergency unit?" Journal of Internal Medicine, **235**(2): 143-151.

T116. Dale, S., et al (1994). "Comparison of three D-dimer assays for the diagnosis of DVT: ELISA, latex and an immunofiltration assay (NycoCard D-Dimer)." Thrombosis and Haemostasis, **71**(3): 270-274.

T117. Pini, M., et al (1993). "Combined use of strain-gauge plethysmography and latex D-dimer test in clinically suspected deep venous thrombosis." Fibrinolysis, **7**(6): 391-396.

T118. Carter, C.J., et al (1993). "Investigations into the clinical utility of latex D-Dimer in the diagnosis of deep venous thrombosis." Thrombosis and Haemostasis, **69**(1): 8-11.

T119. Brenner, B., et al (1993). "Application of SimpliRED D-dimer in the diagnosis of deep vein thrombosis." Thrombosis and Haemostasis, **69**: 832.

T120. Ibrahim, K.M.A., A.I. O'Neill, and D.J. Parkin (1992). "D-dimer (latex) assay in the diagnosis of deep venous thrombosis." Medical Journal of Australia, **157**(8): 574-575.

T121. Heijboer, H., et al (1992). "The use of the D-dimer test in combination with non-invasive testing versus serial non-invasive testing alone for the diagnosis of deep-vein thrombosis." Thrombosis and Haemostasis, **67**(5): 510-513.

T122. Kroneman, H., et al (1991). "Diagnostic value of D-dimer for deep venous thrombosis in outpatients." Haemostasis, **21**(5): 286-292.

T123. Grau, E., et al (1991). "Utility of D dimer in the diagnosis of deep venous thrombosis in outpatients." Thrombosis and Haemostasis, **66**(4): 510.

T124. De Boer, W.A., et al (1991). "D-Dimer latex assay as screening method in suspected deep venous thrombosis of the leg. A clinical study and review of the literature." Netherlands Journal of Medicine, **38**(2): 65-69.

T125. Chang-Liem, G.S., F.A.T. Lustermans, and J.W.J. Van Wersch (1991). "Comparison of the appropriateness of the latex and Elisa plasma D-dimer determination for the diagnosis of deep venous thrombosis." Haemostasis, **21**(2): 106-110.

T126. Boneu, B., et al (1991). "D-dimers, thrombin antithrombin III complexes and prothrombin fragments 1 + 2: Diagnostic value in clinically suspected deep vein thrombosis." Thrombosis and Haemostasis, **65**(1): 28-32.

T127. Mossaz, A., et al (1990). "Value of D-dimer assays in the emergency diagnosis of deep venous thrombosis." Presse Med, **19**(22): 1055.

T128. Elias, A., et al (1990). "Assessment of D-dimer measurement by ELISA or latex methods in deep vein thrombosis diagnosed by ultrasonic duplex scanning." Fibrinolysis, **4**(4): 237-240.

T129. Chapman, C.S., et al (1990). "The use of D-Dimer assay by enzyme immunoassay and latex agglutination techniques in the diagnosis of deep vein thrombosis." Clinical and Laboratory Haematology, **12**(1): 37-42.

T130. Bounameaux, H., et al (1989). "Measurement of plasma D-dimer for diagnosis of deep venous thrombosis." American Journal of Clinical Pathology, **91**(1): 82-85.

T131. Ott, P., et al (1988). "Assessment of D-dimer in plasma: Diagnostic value in suspected deep venous thromobosis of the leg." Acta Medica Scandinavica, **224**(3): 263-267.

T132. Rowbotham, B.J., et al (1987). "Measurement of crosslinked fibrin derivatives - Use in the diagnosis of venous thrombosis." Thrombosis and Haemostasis, **57**(1): 59-61.

T133. Heaton, D.C., J.D. Billings, and C.M. Hickton (1987). "Assessment of D dimer assays for the diagnosis of deep vein thrombosis." Journal of Laboratory and Clinical Medicine, **110**(5): 588-591.

Appendix C- Table of the data and references to the studies excluded from the meta-analysis in Chapter 8

Type F and G

T134. Funfsinn, N., C. Caliezi, F. Demarmels Biasutti, W. Korte, A. Z'Brun, I. Baumgartner, M. Ulrich, C. Cottier, B. Lammle, and W. Wuillemin (2001). "Rapid D-dimer testing and pre-test clinical probability in the exclusion of deep venous thrombosis in symptomatic outpatients." Blood coagulation and fibrinolysis, **12**: 165-170.

Type H

T135. Bozic, M., A. Blinc, and M. Stegnar (2003). "D-dimer, other markers of haemostasis activation and soluble adhesion molecules in patients with different clinical probabilities of Deep Vein Thrombosis." Thrombosis research **108**:107-114.

Type I

T136. Martí-Mestre, F., M. Cairols-Castellote, A. Romera, and C. Herranz (2005). "Diagnostico an urgencias de la thrombosis venosa de miembros inferiores: valor de los criterios clinicos unidos al simero-D." Angiologia **57**:219-224.

Study, author and year	Ddimer assay	Ddimer Accuracy data		Wells score	
		Sensitivity (%)	Specificity (%)	score level	(Disased / Total)
Type F study					
T134. Funfssin 2001	LATEX	NA	NA	Low	0/14
		97.5^	78.8	Moderate	10/36
		NA	NA	High	30/42
	Rap. ELISA	NA	NA	Low	0/14
		100	76.9	Moderate	10/36
		NA	NA	High	30/42
	ELISA	NA	NA	Low	0/14
		97.5	78.8	Moderate	10/36
		NA	NA	High	30/42
Type G study	**(a*d hoc* Thresholds)**				
T134. Funfssin 2001	LATEX 100mg/l	NA	NA	Low	0/14
		100	76.9	Moderate	10/36
		NA	NA	High	30/42
Type H study	**(a*d hoc* Thresholds)**				
T135. Bozic 2003	ELISA 137mg/l	100 (100;100)	81 (70;92)	Low	3/45
		100 (100;100)	87 (77;97)	Moderate	18/48
		100 (100;100)	41 (26;56)	High	31/42
	ELISA 876mg/l	100 (100;100)	79 (67;91)	Low	3/45
		100 (100;100)	80 (69;91)	Moderate	18/48
		100 (100;100)	30 (16;44)	High	31/42
	WBA 114mg/l	100 (100;100)	24 (12;36)	Low	3/45
		100 (100;100)	20 (9;31)	Moderate	18/48
		100 (100;100)	20 (8;32)	High	31/42
	WBA 344mg/l	100 (100;100)	64 (50;78)	Low	3/45
		100 (100;100)	67 (54;80)	Moderate	18/48
		100 (100;100)	30 (16;44)	High	31/42
	SimpliRED 200mg/l	100 (100;100)	71 (58;84)	Low	3/45
		94 (91;97)	57 (43;71)	Moderate	18/48
		97 (94;100)	27 (14;40)	High	31/42
	LATEX 250mg/l	100 (100;100)	55 (40;70)	Low	3/45
		94 (91;97)	60 (46;74)	Moderate	18/48
		100 (100;100)	10 (1;19)	High	31/42

Table E1 Data extracted by the systematic review of WS and DD used in combination; type F (partial proportions for DD), type G (partial

proportions for DD, accuracy at the best threshold), type H (complete proportions for DD, accuracy at the best threshold).

Study, Author and year of publication	Ddimer Accuracy data					Wells score	
	Ddimer Assay	TP	FP	FN	TN	WS level	(Disased/ Total)
Type I studies							
T19. Martí-Mestre 2005	LATEX	NA	0	NA	223	low	NA/223
		8	5	4	10	moderate	12/27
		41	1	0	3	high	41/45

Table E2 Data extracted by the systematic review of WS and DD used in combination; type I data excluded for poor reporting.

Appendix D - Publications, presentations, and posters produced during the PhD project

This section contains al list of journals and scientific events where the findings of this PhD project were used for dissemination.

Publicationsrelated to PhD work:

1. Novielli, N., N. Cooper, et al. (2010). "How Is Evidence on Test Performance Synthesized for Economic Decision Models of Diagnostic Tests? A Systematic Appraisal of Health Technology Assessments in the UK since 1997" Value in Health:13(8).
2. Novielli, N., N. J. Cooper, et al. (2010). "Bayesian model selection for meta-analysis of diagnostic test accuracy data: Application to Ddimer for deep vein thrombosis." Res. Synth. Method: Article first published online: 21 NOV 2010.

A third paper has been drafted on the systematic review of the conditionl accuracy of DD given WS. Provisory title: The effect of Wells score test for clinical assessment on the diagnostic accuracy of Ddimer test for Deep Vein Thrombosis: a systematic review and meta-analysis. Journal: British Medical Journal

A fourth paper is being written that presents the modelling framework developed in Chapter 8. Provisory title: A meta-analytic framework for the estimation of the

accuracy of combinations of tests that accounts for conditional dependence between tests, and for the inclusion in a cost-effectiveness analysis. Journal: Statistics in Medicine.

Scientific Talks:

1. **Nicola Novielli**, Nicola J. Cooper , Alex J. Sutton, Keith R. Abrams.(3rd-7th October 2008) Which meta-analysis model best fits my diagnostic test data? Use of model fit statistics. Cochrane colloquium 2008, Freiburg, Germany
2. **Nicola Novielli**, Nicola J. Cooper , Alex J. Sutton, Keith R. Abrams. (1st-3rd July 2010) Meta-analysis of the accuracy of sequences of diagnostic tests by allowing for between tests correlation. Second international diagnostic tests and biomarkers symposium, Birmingham, UK.

Poster presentations:

1. **Nicola Novielli**, Nicola J. Cooper , Alex J. Sutton, Keith R. Abrams. (From 30th of January to 1st of February 2008) Bayesian model selection criteria in Meta-analysis of Diagnostic test accuracy. Bayesian Biostatistics conference, MD Anderson cancer center, Houston Texas, USA.
2. **Nicola Novielli**, Nicola J. Cooper , Alex J. Sutton, Keith R. Abrams. (June 2008) Meta-analysis of diagnostic test accuracy data and Bayesian model

choice criteria: Deep Venous Thrombosis example. First international diagnostic tests and biomarkers symposium, Birmingham, UK.

3. **Nicola Novielli**, Nicola J. Cooper , Alex J. Sutton, Keith R. Abrams. (24th June 2010) How likely is misdiagnosis? Festival of postgraduate research, Leicester.
4. **Nicola Novielli**, Nicola J. Cooper , Alex J. Sutton, Keith R. Abrams. (from 26th to 28st of January 2011) A Meta-analytic framework for the accuracy of combinations of diagnostic tests: the case of Ddimer and Wells score for Deep Vein Thrombosis. Bayesian Biostatistics conference, MD Anderson cancer center, Houston Texas, USA.

Other publications not directly associated to the PhD research topic:

1. Oddone, F., G. Virgili, M. Parravano, M. Brazzelli, **N. Novielli**, M. Michelessi (2010). "Optic nerve head and fibre layer imaging for diagnosing glaucoma." Cochrane Database of Systematic Reviews:11.
2. Virgili, G., **N. Novielli**, et al. (2010). "Pharmacological Treatments for Neovascular Age-Related Macular Degeneration: Can Mixed Treatment Comparison Meta-Analyses be Useful?" Current drug target: 12

Appendix E – Published paper 1: based on Chapter 4 and Chapter 5

Novielli, N., Cooper, N.J., Sutton, A.J. and Abrams, K.R., 'Bayesian model selection for meta-analysis of diagnostic test accuracy data: Application to Ddimer for deep vein thrombosis' in Research Synthesis Methods, 2010, 1 (3-4), pp. 226-238. DOI: 10.1002/jrsm.15.

Appendix F – Published paper 2: based on Chapter 6

Novielli, N., Cooper, N.J., Abrams, K.R. and Sutton, A.J., 'How Is Evidence on Test Performance Synthesized for Economic Decision Models of Diagnostic Tests? A Systematic Appraisal of Health Technology Assessments in the UK Since 1997' in Value in Health, 2010, 13 (8), pp. 952-957. DOI: 10.1111/j.1524-4733.2010.00762.x.

Bibliography

Ades, A. E. and S. Cliffe (2002). "Markov Chain Monte Carlo Estimation of a Multiparameter Decision Model: Consistency of Evidence and the Accurate Assessment of Uncertainty." Medical Decision Making **22**(4): 359-371.

Ades, A. E., G. Lu, et al. (2004). "Expected Value of Sample Information Calculations in Medical Decision Modeling." Medical Decision Making **24**(2): 207-227.

Ades, A. E. and A. J. Sutton (2006). "Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches." Journal of the Royal Statistical Society: Series A (Statistics in Society) **169**(1): 5-35.

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle." proc. 2nd Int. Symp. Information theory **1**(1973): 267-281.

Alexandersson, A. (2004). "Graphing confidence ellipses: An update of ellip for Stata 8." Stata Journal **4**(3): 242-256.

Altman, D. G. and J. M. Bland (1994). "Statistics Notes: Diagnostic tests 1: sensitivity and specificity." BMJ **308**(6943): 1552.

Altman, D. G. and J. M. Bland (1994). "Statistics Notes: Diagnostic tests 2: predictive values." BMJ **309**(6947): 102.

Anderson, D. R., P. S. Wells, et al. (2000). "Management of patients with suspected deep vein thrombosis in the Emergency Department: Combining use of a clinical diagnosis model with D- dimer testing." Journal of Emergency Medicine **19**(3): 225-230.

Arends, L. R., T. H. Hamza, et al. (2008). "Bivariate random effects meta-analysis of ROC curves." Medical Decision Making **28**(5): 621 - 638.

Bamber, D. (1975). "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph." Journal of Mathematical Psycology **12**: 387-415.

Bate, C. M., S. A. Riley, et al. (1999). "Evaluation of omeprazole as a cost-effective diagnostic test for gastro-oesophageal reflux disease." Alimentary Pharmacology and Therapeutics **13**(1): 59-66.

Bernardo, J. and A. Smith (1994). Bayesian theory. Chichester, Wiley and son ltd.

Bipat, S., A. Zwinderman, et al. (2007). "Multivariate Random-Effects Approach: For Meta-Analysis of Cancer Staging Studies." Acad Radiol **14**: 974-984.

Bland, J. M. and D. G. Altman (2000). "The odds ratio." BMJ **320**(7247): 1468.

Blyth, C. (1972). "Subjective vs. Objective Methods in Statistics." The Americal statistician **26**(3): 20-22.

Bossuyt, P. M., L. Irwig, et al. (2006). "Comparative accuracy: Assessing new tests against existing diagnostic pathways." British Medical Journal **332**(7549): 1089-1092.

Brenner, H. and O. Gefeller (1997). "Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence." Statistics in Medicine **16**(9): 981-991.

Briggs, A., K. Claxton, et al. (2006). Decision modelling for health economic evaluation. Oxford, Oxford University Press.

Brooks, S. and A. Gelman (1997). "General methods for monitoring convergence of iterative simulations." Journal of Computational and Graphical Statistics **7**: 434-455.

Celeux, G., F. Forbes, et al. (2006). "Deviance Information Criteria for Missing Data Models." Bayesian Analysis **na**(na): na.

Chappell, F., G. Raab, et al. (2009). "When are summary ROC curves appropriate for diagnostic meta-analyses." Statistics in Medicine **28**: 2653-2668.

Chu, H., L. Nie, et al. (2009). "Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: alternative parameterization and model selection." Statistics in Medicine **28**: 2384-2399.

Claxton, K. (1999). "The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies." Journal of Health Economics **18**(3): 341-364.

Cooper, N., A. Sutton, et al. (2009). "Including covariates in a mixed treatment comparison framework: Application to stroke prevention treatments in individuals with non-rheumatic Atrial Fiblillation." Statistics in Medicine **28**: 1861-1881.

Cooper, N. J., A. J. Sutton, et al. (2004). "Comprehensive decision analytical modelling in economic evaluation: a Bayesian approach." Health Economics **13**(3): 203-226.

Cosh, E., A. Girling, et al. (2007). "Investing in New Medical Technologies: A decision framework." Journal of Commercial Biotechnology **13**(4): 263-271.

Deeks, J. J. (2001). "Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests." BMJ **323**(7305): 157-162.

Deeks, J. J. and D. G. Altman (2004). "Statistics notes - Diagnostic tests 4: Likelihood ratios." British Medical Journal **329**(7458): 168-169.

Dempster, A. (1997). "The direct use of likelihood for significance testing." Statistics and Computing **7**: 247-252.

Dendukuri, N., A. Hadgu, et al. (2009). "Modeling conditional dependence between diagnostic tests: A multiple latent variable model." Statistics in Medicine **28**(3): 441-461.

Dendukuri, N. and J. Lawrence (2001). "Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests." Biometrics **57**: 158-167.

Drummond, M., M. Sculpher, et al. (2005). Methods for the economic evaluations of health care programmes. Oxford, Oxford University Press.

Dukic, V. and C. Gatsonis (2003). "Meta-analyis of Diagnostic test accuracy assessment studies with varying number of thresholds." Biometrics **59**: 936 - 946.

Egger, M., G. D. Smith, et al. (2001). Systematic reviews in health care. Meta analysis in context., BMJ Books.

Enoe, C., M. P. Georgiadis, et al. (2000). "Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown." Preventive Veterinary Medicine **45**(1-2): 61-81.

Fryback, D. G. (1978). "Bayes' theorem and conditional nonindependence of data in medical diagnosis." Computers and Biomedical Research **11**(5): 423-434.

Gardner, I. A., H. Stryhn, et al. (2000). "Conditional dependence between tests affects the diagnosis and surveillance of animal diseases." Preventive Veterinary Medicine **45**(1-2): 107-122.

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." Bayesian Analysis **1**(3): 515-533.

Gelman, A. (2009). "Bayes, Jeffreys, Prior Distributions and the Philosophy of Statistics." Statistical Science **24**(2): 176-178.

Gelman, A., J. Carlin, et al. (2003). Bayesian data analysis. Dallas, Texas, USA, Chapman and Hall (CRC).

Gelman, A. and D. Rubin (1992). "Inference from iterative simulation using multiple sequences." Statistical Science **7**: 457-511.

George, F. B., L. Oscar, et al. (2009). "The evidence provided by a single trial is less reliable than its statistical analysis suggests." Journal of Clinical Epidemiology **62**(7): 711-715.e711.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. Bayesian Statistics 4 (ed JM Bernado, JO Berger, AP Dawid and AFM Smith). C. Press. Oxford, UK.

Gilks, W., S. Richardson, et al. (1996). Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics. London.

Glas, A. S., J. G. Lijmer, et al. (2003). "The diagnostic odds ratio: a single indicator of test performance." Journal of Clinical Epidemiology **56**(11): 1129-1135.

Goodacre, S., F. C. Sampson, et al. (2005). "Variation in the diagnostic performance of D-dimer for suspected deep vein thrombosis." Quarterly Journal of Medicine **98**(7): 513-527.

Goodacre, S., A. J. Sutton, et al. (2005). "Meta-analysis: The value of clinical assessment in the diagnosis of deep venous thrombosis." Annals of Internal Medicine **143**(2): 129-139+I-140.

Goodacre, S. S., F ; Stevenson,M;, A. S. Wailoo, A; Thomas,S; Locker,T;, et al. (2006). "Measurement of the clinical and cost-effectiveness of non-invasive diagnostic testing strategies for deep vein thrombosis " Health Technology Assessment **10**(15): 168.

Halfon, P., Y. Eggli, et al. (2002). "Measuring potentially avoidable hospital readmissions." Journal of Clinical Epidemiology **55**(6): 573-587.

Hamza, T., L. Arends, et al. (2009). "Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds." BMC Medical Research Methodology **9**(1): 73.

Hamza, T., L. Arends, et al. (2009). "Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds." BMC Medical Research Methodology **9**(1): 1-15.

Hamza, T. H., J. B. Reitsma, et al. (2008). "Meta-Analysis of Diagnostic Studies: A Comparison of Random Intercept, Normal-Normal, and Binomial-Normal Bivariate Summary ROC Approaches." Medical Decision Making **28**(5): 639-649.

Hamza, T. H., H. C. van Houwelingen, et al. (2009). "Associating explanatory variables with summary receiver operating characteristic curves in diagnostic meta-analysis." Journal of Clinical Epidemiology **62**(12): 1284-1291.

Hamza, T. H., H. C. van Houwelingen, et al. (2008). "Random effects meta-analysis of proportions: The binomial distribution should be used to model the within-study variability." Journal of Clinical Epidemiology **61**(1): 41 - 51.

Harbord, R. and P. Whitting (2009). "metandi: Meta-analysis of diagnostic accuracy using hierarchical logistic regression." Stata Journal **9**: 211-229.

Harbord, R. M., J. J. Deeks, et al. (2007). "A unification of models for meta-analysis of diagnostic accuracy studies." Biostatistics **8**(2): 1 - 21.

Harbord, R. M., J. J. Deeks, et al. (2007). "A unification of models for meta-analysis of diagnostic accuracy studies." Biostat **8**(2): 239-251.

Harbord, R. M., P. Whiting, et al. (2008). "An empirical comparison of methods for emta-analysis of diagnostic accuracy showed hierarchical models are necessary." Journal of Clinical Epidemiology **61**: 1095-1103.

Heidelberger, P. and P. Welch (1983). "Simulation run length control in the presence of an initial transient." Opns Res. **31**: 1109-1144.

Hellmich, M., K. Abrams, et al. (1999). "Classical and Bayesian approaches to meta-analysis of ROC curves: a comparative review. ." Medical Decision Making **19**: 252-264.

Henschke, C. and J. Whalen (1994). "Evaluation of competing diagnostic tests: sequences for the diagnosis of pulmonary embolism, Part II." Clinical imaging **18**(4): 248-254.

Higgins, J. P. T., S. G. Thompson, et al. (2009). "A re-evaluation of random-effects meta-analysis." Journal of the Royal Statistical Society: Series A (Statistics in Society) **172**(1): 137-159.

Hinkley, D., N. Reid, et al. (1991). Statistical theory and modelling. Bury St Edmund, England.

Huang, X., G. Qin, et al. (2010). "Optimal Combinations of Diagnostic Tests Based on AUC." Biometrics: no-no.

Hui, S. L. and S. D. Walter (1980). "Estimating the Error Rates of Diagnostic Tests." Biometrics **36**(1): 167-171.

Hull, R. D., G. E. Raskob, et al. (1986). "Prophylaxis of Venous Thromboembolism." Chest **89**(5 Supplement): 374S-383S.

Jackson, D., I. R. White, et al. (2010). "Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses." Statistics in Medicine **29**(12): 1282-1297.

Jin, H. and Y. Lu (2008). "A Procedure for Determining Whether a Simple Combination of Diagnostic Tests May Be Noninferior to the Theoretical Optimum Combination." Medical Decision Making **28**(6): 909-916.

Jones, C. M. and T. Athanasiou (2009). "Diagnostic accuracy meta-analysis: review of an important tool in radiological research and decision making." The British Journal of Radiology **82**: 441-446.

Joseph, L., T. W. Gyorkos, et al. (1995). "Bayesian Estimation of Disease Prevalence and the Parameters of Diagnostic Tests in the Absence of a Gold Standard." American Journal of Epidemiology **141**(3): 263-272.

Kass, E. and L. Wasserman (1996). "The Selection of Prior Distributions by Formal Rules." Journal of the American Statistical Association **91**(435): 1343-1370.

Khalil, E. L. (2010). "The Bayesian fallacy: Distinguishing internal motivations and religious beliefs from other beliefs." Journal of Economic Behavior & Organization **75**(2): 268-280.

Knorr-Held, R. and N. Best (1999). A shared component model for detecting joint and selective clustering of two diseases. International Conference on the Analysis and Interpretation of Disease Clusters and Ecological Studies. B. P. LTD. **164:** 73-85.

Lambert, P. C., A. J. Sutton, et al. (2005). "How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS." Statistics in Medicine **24**(15): 2401-2428.

Lau, J., J. P. A. Ioannidis, et al. (1998). "Summing up evidence: one answer is not always enough." Lancet **351**(9096): 123-127.

Lau, J., C. H. Schmid, et al. (1995). "Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care." Journal of Clinical Epidemiology **48**(1): 45-57.

Leeflang, M. M. G., J. J. Deeks, et al. (2008). "Systematic Reviews of Diagnostic Test Accuracy." Annals of Internal Medicine **149**(12): 889-897.

Lijmer, J. G., B. W. Mol, et al. (1999). "Empirical Evidence of Design-Related Bias in Studies of Diagnostic Tests." JAMA: The Journal of the American Medical Association **282**(11): 1061-1066.

Lilford, R. and D. Braunholtz (1996). "The statistical basis of public policy: a paradigm shift is overdue." BMJ **313**(7057): 603-607.

Littenberg, B. and L. E. Moses (1993). "Estimating diagnostic-accuracy from multiple conicting reports - a new meta-analytic method." Medical Decision Making **13**: 313 - 321.

Liu, A., E. F. Schisterman, et al. (2005). "On linear combinations of biomarkers to improve diagnostic accuracy." Statistics in Medicine **24**(1): 37-47.

Lunn, D. J., A. Thomas, et al. (2000). "WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility." Statistics and Computing **10**: 325-337.

Macaskill, P., S. D. Walter, et al. (2002). "Assessing the gain in diagnostic performance when combining two diagnostic tests." Statistics in Medicine **21**(17): 2527-2546.

Mallett, S., J. J. Deeks, et al. (2006). "Systematic reviews of diagnostic tests in cancer: review of methods and reporting." British Medical Journal **333**: 413-416.

McAteer, H., E. Cosh, et al. (2007). "Cost-effectiveness analysis at the development phase of a potential health technology: examples based on tissue engineering of bladder and urethra." Journal of Tissue Engineering and Regenerative Medicine **1**(5): 343-349.

McIntosh, M. W. and M. S. Pepe (2002). "Combining Several Screening Tests: Optimality of the Risk Score." Biometrics **58**(3): 657-664.

Moses, L. E., D. Shapiro, et al. (1993). "Combining independent studies of a disgnostic-test into a summary ROC curve - data-analytic approaches and some additional considerations." Statistics in Medicine **12**(14): 1293-1316.

Mosteller, F. and G. A. Colditz (1996). "Understanding Research Synthesis (Meta-Analysis)." Annual Review of Public Health **17**(1): 1-23.

Newcombe, R. G. (1998). "Two -sided confidence intervals for the single proportion: comparison of seven methods." Statistics in Medicine **17**: 857-872.

NICE (2008). "Guide to the methods of technology appraisal." National Institute for Health and Clinical Excellence.

NICE (2010). "Measuring effectiveness and cost effectiveness: the QALY." National Institute for Health and Clnical Excellence.

Novielli, N., N. J. Cooper, et al. (2010). "How Is Evidence on Test Performance Synthesized for Economic Decision Models of Diagnostic Tests? A Systematic Appraisal of Health Technology Assessments in the UK Since 1997." Value in Health **13**(8): 952-957.

Novielli, N., N. J. Cooper, et al. (2010). "Bayesian model selection for meta-analysis of diagnostic test accuracy data: Application to Ddimer for deep vein thrombosis." Research Synthesis Methods **1**(3-4): 226-238.

Ntzoufras, I. (2010). Bayesian modeling using WinBUGS.

Numans, M. E., J. Lau, et al. (2004). "Short-Term Treatment with Proton-Pump Inhibitors as a Test for Gastroesophageal Reflux Disease: A Meta-Analysis of Diagnostic Test Characteristics." Annals of Internal Medicine **140**(7): 518-527+I551.

O'Hagan, A., C. E. Buck, et al. (2006). Uncertain judgements: eliciting experts' probabilities. Chicester.

Paul, M., A. Riebler, et al. (2009). "Bayesian bivariate meta-analysis of diagnostic test studies using integrated nested Laplace approximations." Statistics in Medicine **29**(12): 1325-1339.

Pepe, M. (2003). The Statistical Evaluation of Medical Tests for Classification and Prediction Oxford University Press.

Pepe, M. S. and M. L. Thompson (2000). "Combining diagnostic test results to increase accuracy." Biostatistics **1**(2): 123-140.

Phillips, C. and G. Thompson (1998). "What is a QALY?" Evidence based medicine **1**(6).

Press, S. (2002). Subjective and Objective Bayesian Statistics: Principles, Models, and Applications. New York, Wiley.

Principato, F., A. Vullo, et al. (2010). "On implementation of the Gibbs sampler for estimating the accuracy of multiple diagnostic tests." Journal of Applied Statistics **37**(8): 1335-1354.

Putter, H., M. Fiocco, et al. (2009). "Meta-Analysis of Diagnostic Test Accuracy Studies with Multiple Thresholds using Survival Methods." Biometrical Journal **52**(1): 95-110.

Qin, J. and B. Zhang (2010). "Best combination of multiple test for screening purposes." Statistics in Medicine **29**: 2905-2919.

R Foundation for Statistical Computing (2005). R: A Language and Environment for Statistical Computing. Vienna, Austria, R Development Core Team.

Racine-Poon, A. and J. Wakefield (1996). Bayesian Analysis of population pharmacokinetic and instantaneous pharmacodynamic relationships. Bayesian Biostatistics (Berry, DA, Stangl, DK)

Bayesian Biostatistics

Taylor and Francis (CRC). **151**.

Raftery, A. E. and S. M. Lewis (1992). "One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo." Statistical Science **7**: 493-497.

Raftery, A. E. and S. M. Lewis (1995). The number of iterations, convergence diagnostics and generic Metropolis algorithms. Practical Markov Chain Monte Carlo (W.R. Gilks, D.J. Spiegelhalter and S. Richardson, eds.). C. a. Hall. London, UK.

Reitsma, J. B., A. S. Glas, et al. (2005). "Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews." Journal of Clinical Epidemiology **58**: 982 - 990.

Rhea, J., S. DeLuca, et al. (1982). "Evaluation of a sequence of diagnostic tests using the workup of ureteral stone as a model." Medical care **20**(8): 843-848.

Riley, R. D., S. R. Dodd, et al. (2008). "Meta-analysis of diagnostic test studies using individual patient data and aggregate data." Statistics in Medicine **27**(29): 6111-6136.

Rucker, G. and M. Schumacher (2010). "Summary ROC curve based on a weighted Youden index for selecting an optimal cutpoint in meta-analysis of diagnostic accuracy." Statistics in Medicine **29**: 3069-3078.

Rutter, C. M. and C. A. Gatsonis (1995). "Regression methods for meta-analysis of diagnostic test data." Academic Radiology **2 Suppl 1**: S48-56; discussion S65-47.

Rutter, C. M. and C. A. Gatsonis (2001). "A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations." Statistics in Medicine **20**(19): 2865-2884.

Severens, J. L., P. F. de Vries Robbé, et al. (1999). "Optimizing Diagnostic Test Sequences: The Probability Modifying Plot." Methods of Information in Medicine **38**: 50-55.

Shen, Y., D. Wu, et al. (2001). "Testing the Independence of Two Diagnostic Tests." Biometrics **57**(4): 1009-1017.

Siadaty, M. S., J. T. Philbrick, et al. (2004). "Repeated-measures modeling improved comparison of diagnostic tests in meta-analysis of dependent studies." Journal of Clinical Epidemiology **57**(7): 698-711.

Snedecor, G. and G. Cochran (1967). Statistical methods. Ames, Iowa, USA, The Iowa State University Press.

Spiegelhalter, D., K. Abrams, et al. (2004). Bayesia approaches to Clinical trials and Health-Care evaluation Chichester, UK, Wiley.

Spiegelhalter, D., N. Best, et al. (2002). "Bayesian measure of model complexity and fit." Journal of the Royal Statistical Society **64**: 583-639.

Spiegelhalter, D., A. Thomas, et al. (2003). "WinBUGS user manual: Version 1.4." Cambridge: MRC Biostatistics Unit.

Spiegelhalter DJ, K. Abrams, et al. (2004). Bayesian Approaches to Clinical Trials and Health-Care Evaluation, Wiley.

Sterne, J. A. (18 November 2009). Meta-analysis of test accuracy studies: how can we make sense of extreme heterogeneity? Meta-Analysis in Diagnostic Testing, Reading university.

Sterne, J. A. C., D. Gavaghan, et al. (2000). "Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature." Journal of Clinical Epidemiology **53**(11): 1119-1129.

Su, J. Q. and J. S. Liu (1993). "Linear Combinations of Multiple Diagnostic Markers." Journal of the American Statistical Association **88**(424): 1350-1355.

Sutton, A. J., K. R. Abrams, et al. (2000). Exploring between study heterogeneity. In: Methods for meta-analysis in medical research. Chichester, Wiley**:** xvii, 317.

Sutton, A. J., N. J. Cooper, et al. (2008). "Integration of Meta-analysis and Economic Decision Modeling for Evaluating Diagnostic Tests." Medical Decision Making **28**: 650-667.

Sweeting, M. J., A. J. Sutton, et al. (2004). "What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data." Statistics in Medicine **23**(9): 1351-1375.

The BUGS project http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/dicpage.shtml#q9.

Thompson, M. L. (2003). "Assessing the diagnostic accuracy of a sequence of tests." Biostat **4**(3): 341-351.

Vacek, P. M. (1985). "The Effect of Conditional Dependence on the Evaluation of Diagnostic Tests." Biometrics **41**(4): 959-968.

Van Houwelingen, H. C., K. H. Zwinderman, et al. (1993). "A bivariate approach to meta-analysis." Statistics in Medicine **12**(24): 2273-2284.

Verde, P. (2010). "Meta-analysis of diagnostic test data: a bivariate Bayesian modeling approach." Statistics in Medicine **29**(30): 3088-3102.

Wallace, B., C. Schmid, et al. (2009). "Meta-Analyst: software for meta-analysis of binary, continuous and diagnostic data." BMC Medical Research Methodology **9**(1): 80.

Wells, P., D. R. Anderson, et al. (1997). "Value of assessment of pretest probability of deep-vein thrombosis in clinical management." Lancet **350**(9094): 1795-1798.

Wells, P. S., J. Hirsh, et al. (1995). "Accuracy of clinical assessment of deep-vein thrombosis." Lancet **345**(8961): 1326-1330.

Wells, P. S., C. Owen, et al. (2006). "Does This Patient Have Deep Vein Thrombosis?" JAMA: The Journal of the American Medical Association **295**(2): 199-207.

Welton, N., N. Cooper, et al. (2008). "Mixed treatment comparison with multiple outcomes reported inconsistently across trials: evaluation of antivirals for treatment of influenza A and B." Statistics in Medicine **27**: 5620-5639.

Whiting, P., A. W. S. Rutjes, et al. (2004). "Sources of Variation and Bias in Studies of Diagnostic Accuracy: A Systematic Review." Ann Intern Med **140**(3): 189-202.

Whiting, P. F. S., Jonathan AC; Westwood, Marie E; Bachmann,Lucas M; Harbord, Roger; Egger, Matthias; Deeks ,Jonathan J (2008). "Graphical presentation of diagnostic information." Med Res Methodol(8): 20.

Zhou, X.-H., D. K. McClish, et al. (2002). Statistical Methods in Diagnostic Medicine, Wiley.

Zou, K. H., J. G. Bhagwat, et al. (2006). "Statistical Combination Schemes of Repeated Diagnostic Test Data." Academic radiology **13**(5): 566-572.

Zweig, M. and G. Campbell (1993). "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine." Clinical Chemistry **39**(4): 561-577.

Zwinderman, A. H. and P. M. Bossuyt (2008). "We should not pool diagnostic likelihood ratios in systematic reviews." Statistics in Medicine **27**(5): 687-697.