Meta-Analysis of Diagnostic Test Data: Modern Statistical Approaches

Inaugural Dissertation

zur

Erlangung des Doktorgrades der

Mathematisch-Naturwissenschaftlichen Fakultät

der Heinrich-Heine Universität Düsseldorf

vorgelegt von

Pablo Emilio Verde

aus Buenos Aires, Argentinien

Juni 2008

Aus der Koordinierungszentrum für Klinische Studien
Heinrich-Heine Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. Christian Ohmann
Koreferent: Prof. Dr. Martin Lercher
Tag der mündlichen Prüfung: 08.07.2008

To
Isabel, Lucía, Heide,
my Mother
and
to the memory of my Father.

Acknowledgments

I am very thankful to Christian Ohmann for supporting my work during these years. He has allowed me to work with the intellectual freedom, that I needed to develop and to accomplish this dissertation. I am very grateful to Martin Lercher for welcoming my dissertation, without his support and trust this work would not be possible. This work was partially financed by the Swiss National Science Foundation (51A240-104890). I thank Johannes Siegrist for supporting this work.

In the summer of 2007, I had the honor to be invited to teach in the Department of Statistics at Stanford University. During my stay I presented an early version of this work. I am immensely grateful to Brad Efron, Trevor Hastie and Ingram Olkin for discussions and several suggestions to improve my work. I am also grateful to Persi Diaconis for numerous enlightening conversations on Bayesian statistics.

My sincere acknowledgement to my professor in Argentina Martin Grondona. He has constantly encouraged me to finalize my dissertation.

I am very thankful to those colleges and friends that during these months of seclusion reinforced my enthusiasm, in particular my thanks go to: Dimitris Venizeleas, Ajrun Shakiri, Uschi Willems, Sumaya Abid and Juan Carlos Carillo.

Finally my immense gratitude to my wife Heide for her love, for her outstanding family management, and for her great support! She has accomplished the almost impossible mission of keeping a happy family, while I was working on this dissertation. All of my gratitude and love go to my daughters Isabel and Lucia for their every day sunshine and to my mother for being a living example.

Abstract

In the last decades, the amount of published results on clinical diagnostic tests has expanded very rapidly. The counterpart to this accelerated technological development has been the formal evaluation and synthesis of diagnostic results. However, published results can be regarded as so far removed from the classical domain of meta-analysis, that they can provide a rather severe test of classical methods. This work concerns the applications of computer intensive statistical methods in meta-analysis of diagnostic test data. These methods are considered from both the classical and Bayesian perspective.

From the classical point of view, bootstrap methods are used to build confidence intervals for complex statistics in meta-analysis. These methods are evaluated extensively by a simulation experiment.

Under the Bayesian perspective, a novel statistical model is presented. This model is general enough to include the presence of studies with different designs, unusual accurate results, large amounts of sparsity data, missing reporting data and heterogeneity between studies' population. These multiple sources of variability are modeled with a Bayesian graphical approach. In this approach a complex model is broken up into manageable sub-models. The full model is built up as a network by exploring the local dependency structure of each model component. A schematic description of this process is presented by Directed Acyclic Graph (DAG), which gives a non-algebraic structure and links computations to Markov chain Monte Carlo (MCMC) techniques.

Statistical computations are implemented in open source and public domain statistical software (BUGS and R) and illustrated with a complex systematic review which evaluates the diagnostic performance of computer tomography scans in diagnostic of appendicitis.

Zusammenfassung

In den letzten Jahrzehnten ist die Menge der publizierten Ergebnisse klinisch - diagnostischer Tests stark angestiegen. Das Pendant zu dieser beschleunigten technologischen Entwicklung war die formale Evaluierung und Synthese diagnostischer Ergebnisse. Publizierte Ergebnisse jedoch können manchmal so stark von der klassischen Domäne der Metaanalyse abweichen, dass die klassischen Methoden ernsthaft auf die Probe gestellt werden.

Das Thema dieser Arbeit ist die Anwendung computer-intensiver statistischer Methoden in der Metaanalyse diagnostischer Test-Daten. Diese Methoden werden sowohl aus der klassischen wie auch aus der Bayesschen Perspektive her betrachtet. In der klassischen Perspektive werden Bootstrap Methoden verwendet, um Konfidenzintervalle für komplexe Statistiken in der Metaanalyse zu konstruieren. Diese Methoden werden in einem Simulations-Experiment intensiv evaluiert.

In der Bayesschen Perspektive wird ein neuartiges statistisches Modell vorgestellt. Dieses Modell ist allgemein genug, um die Präsenz von Studien mit unterschiedlichen Designs, ungewöhnlich genauen Ergebnissen, großen Mengen dünnbesetzter Daten, fehlenden Daten sowie heterogenen Studienpopulationen zu umfassen. Diese Quellen von Variabilität werden anhand eines Bayesschen graphischen Ansatzes modelliert. In diesem Ansatz wird ein komplexes Modell in handhabbarere Unter-Modelle aufgebrochen. Das vollständige Modell wird dann als Netzwerk zusammengesetzt, indem die lokale Abhängigkeits-Struktur einer jeden Modell-Komponente exploriert wird. Eine schematische Beschreibung dieses Prozesses wird anhand Gerichteter Azyklischer Graphen dargestellt, wodurch eine nicht-algebraische Struktur erzeugt wird und die Berechnung mit Markov-Ketten Monte-Carlo Methoden (MCMC) verknüpft wird.

Die statistischen Berechnungen sind in Open Source und Public Domain Statistik Software implementiert (BUGS und R) und werden anhand eines komplexen systematischen Reviews illustriert, der die diagnostische Leistungsfhigkeit von Computer Tomographie bei der Diagnose der Appendizitis evaluiert.

Contents

List of Tables

List of Figures

Introduction

"One of the most important and difficult problems in science is the synthesis of evidence..." -

-David M. Eddy, Vic Hasselblad and Ross Shachter,
Meta-Analysis by the Confidence Profile Method, 1991, pag.1.

The first crucial information in the presence of illness is a medical diagnosis. How good or bad a diagnosis is performed may directly influence the quality of the health care. Accurate evaluation of diagnostic tests contributes to the prevention of unjustified treatment, as well as unnecessary health costs.

In the last decades, the amount of published results on clinical diagnostic tests has expanded very rapidly (Knottnerus et al.(2002)[70]). Naturally, the counterpart to this accelerated development has been the formal evaluation and synthesis of diagnostic results. In this regard, methods for searching and assessing the quality of studies have been established (Whiting et al. [132] 2003), and statistical methods for meta-analysis have been proposed (Gatsonis and Paliwal (2006) [43]).

However, published results can be regarded as so far removed from the classical methodological domain of meta-analysis, that they can provide a rather severe test of classical methods. During the last few years, the statistical study of meta-analysis of diagnostic test has provided a series of interesting problems for applied statistics. Those challenges are the topic of this work.

Special features of meta-analysis of diagnostic data

Meta-analysis of diagnostic test data differs from other type meta-analysis, in at least in two aspects. First is the high complexity of the published data. The context where diagnostic studies have been performed can be very different in terms of study design, population characteristics, study quality or diagnostic setup. These sources of heterogeneity have been investigated by Lijmer et al. (1999)[81], Lijmer et al. (2002) [80] and Westwood et al. (2005) [131]. Moreover, the way that this information is published may differ from paper to paper and in some cases relevant information may be incomplete. Therefore, meta-analysis of diagnostic tests implied the synthesis of imperfect and probably incomplete evidence.

Second, meta-analysis of diagnostic tests involves a small sample of heterogeneous non-normal multivariate data, which is complex to analyze.

Aims of this work

We aim to develop a flexible class of statistical models and techniques to deal with meta-analysis of diagnostic test. Although we use basic mathematics to describe the statistics methods, our aim is not to develop theorems and proofs of theorems, but rather to provide the statistician with statistical tools to make inference from published data.

The starting point is the use of bootstrap methods. We analyze the application of these computational intensive techniques in building confidence intervals for complex summary statistics in meta-analysis. These methods are evaluated extensively by a simulation experiment.

Second, we present a Bayesian statistical framework to model the apparent disparate diagnostic test data. We construct a hierarchical Bayesian model, that realistically reflects the underlying complexities of these types of data. We follow a *Bayesian graphical modeling* approach where a complex model is broken up into small manageable sub-models. The full model is build up as a network by exploring the local dependency structure of each model component. A schematic description of this process is presented by Directed Acyclic Graph (DAG), which gives a non-algebraic structure and links computations to Markov chain Monte Carlo (MCMC) techniques.

The access of *high quality*, *open source* and *free* statistical software has fundamentally changed the way that we develop and communicate our statistical ideas. Another

important aim of this work is to develop the software tools to apply the methods presented in this work. With this aim, two statistical computer languages are used: R [93], the open source implementation of the statistical S language and BUGS (Bayesian Analysis Using Gibbs Sampling), a free available statistical software to implement complex Bayesian modeling (Spiegelhalter et al. (2004)[110]).

Style and typographical conventions

More emphasis is placed on methods and applications than theoretical development, with statistical model building and data analysis playing a central role. Mathematical presentation is kept as elementary as possible and many of the arguments are quite informal. However, some statistical modeling in Chapter 3, Chapter 5 and Chapter 6 may be difficult to understand without substantial experience in applied statistics.

R and BUGS scripts have been developed by ourselves and details are given in Chapter 7. We follow some typographical conventions: R's and BUGS command lines are printed in a monospace typewriter font `like this`. We use the symbol `>` for the prompt of the R's console. In order to save space some of the R output has been edited, when some output lines are omitted we indicate that by `...`

Most of the R output was generated with the option setting

```
options(width = 65, digits = 3)
```

Not all R's functions follow this setting, so sometimes we had to reduce manually the printed numerical precision. Calculations are performed on a PC with a CPU of 3.06GHz and 457.136KB RAM running under the Windows 2000 operative system and they have been tested on Windows XP, VISTA and Linux operating systems respectively.

Overview of the chapters

Chapter 1 presents our running example and a review of the statistical methods. More details of the data are described in Chapter 2, which include the process used to select the studies, the medical databases used, searching scripts, study characteristics, missing information, etc.

Chapter 3 deals with bootstrap methods and their applications in meta-analysis of test data. This chapter starts with a general introduction of bootstrap methods and

how these techniques can be used in meta-analysis. We illustrate the construction of different types of confidence intervals for the area under the summary ROC curve and we evaluate their performance.

A general introduction of Bayesian statistics is presented in Chapter 4. This chapter gives a less technical and more philosophical discussion about Bayesian methods.

The starting point of the model building is presented in Chapter 5. This approach was originally introduced by Verde (2005) [120] and further extended by Verde (2006) [121]. It constitutes a full Bayesian framework to perform meta-analysis of diagnostic test. This chapter includes details in MCMC computations, Bayesian model selection and predictive model checking. We also compare this model with other similar statistical models based on approximate numerical methods (e.g. quadrature for integrals and numerical optimization) implemented in commercial statistical software (SAS and Stata).

Important extensions of the basic Bayesian model are presented in Chapter 6. Using a structural dispersion model we deal with cross-synthesis and quantify the *relative credibility* of studies with different designs. A new concept, called *study relevance*, is presented to assess studies with unusual diagnostic results. Finally, we present a bivariate meta-regression approach to account for systematic variation. These models were presented by Verde (2007a, 2007b, 2008) [122, 123, 124].

Chapter 7 deals with details of the R and BUGS computations. Finally, a general overview of this work is given and further research in this area is discussed.

Chapter 1

A Review of Meta-analysis of Diagnostic Test

''Essentially, all models are wrong, but some are useful -

-George E.P. Box and Normal R. Draper (1987) *Empirical Model-Building and Response Surfaces, p. 424,.*

1.1 Introduction

In this Chapter we introduce some basic concepts of diagnostic test data and how these data are usually combined in a meta-analysis. We start by describing our running example. We briefly explain how a diagnostic test is designed for a single study and how these concepts influence statistical methods to perform meta-analysis of diagnostic tests. Current statistical methods are reviewed and their limitations are highlighted. These limitations motivate the statistical methods that will be described in the next chapters.

1.2 Running Example

Table 1.1 summarizes results of 52 published papers investigating the use of Computer Tomography (CT) scans in the diagnosis of appendicitis (Ohmann et al., 2006) [90]. This disease is one of the most common acute surgical events (Addiss et al., 1990) [1], where traditional clinical examination delivers low diagnostic performance (Kraemer et al., 2000)[71]. Therefore, a new diagnostic technology could reduce the risk of postoperative complications and save health care resources (Flum et al., 2000)[40].

Study-design-id	*tp*	*fp*	*fn*	*tn*	*Study-design-id*	*tp*	*fp*	*fn*	*tn*
Applegate2001 *(R)1*	87	4	2	3	Morris2002*(R)27*	38	8	1	82
Balthazar1998 *(R)2*	111	1	4	30	Mullins2001*(R)28*	64	1	2	128
Bendeck2002 *(R)3*	184	7	8	9	Peck2000*(R)29*	103	1	8	252
Brandt2003 *(R)4*	168	3	1	7	Pickut2001*(P)30*	88	3	5	24
Cakirer2002 *(P)5*	89	3	5	33	Raman2003*(R)31*	137	8	5	402
Cho1999 *(R)6*	21	1	0	14	Raman2002*(R)32*	137	8	5	402
Choi1998 *(R)7*	125	3	0	12	Rao1999*(P)33*	32	2	0	66
Cole2001 *(P)8*	40	4	5	43	Rao1999*(P)34*	114	3	1	211
DIppolito1998 *(P)9*	40	0	4	8	Rao1997*(P)35*	52	1	1	46
Ege2002 *(R)10*	104	3	4	185	Rao1996*(P)36*	17	0	0	18
Fefferman2001 *(R)11*	34	4	1	54	Rao1997*(P)37*	56	2	0	41
Funaki1998 *(P)12*	29	4	1	66	Schuler1998*(R)38*	49	4	1	43
Garcia Pena1999 *(P)13*	28	5	1	74	Sivit2000*(R)39*	58	6	3	87
Hershko2002 *(P)14*	67	7	5	118	Stacher1999*(P)40*	21	0	1	34
Hong2003 *(P)15*	30	3	3	42	Stroman1999*(R)41*	33	11	3	60
Horton2000 *(P)16*	36	1	1	11	Styrud2000*(R)42*	44	3	6	61
Kaiser2002 *(P)17*	131	12	4	170	Torbati2003*(P)43*	43	5	4	166
Kamel2000 *(R)18*	23	0	1	76	Tsai2001*(P)44*	4	0	0	22
Kan2001 *(P)19*	4	2	0	25	Ujiki2002*(R)45*	28	8	3	64
Karakas2000 *(R) 20*	31	2	6	76	Walker2000*(P)46*	30	0	2	25
Lane1999 *(P)21*	110	4	5	181	Weltman2000*(P)47*	47	1	1	51
Lane1997 *(P)22*	37	2	4	66	Weyant2000*(P)48*	183	26	9	24
Lowe2001 *(P)23*	18	1	1	55	Wijetunga2001*(P)49*	28	2	2	68
Lowe2001 *(R)24*	35	0	1	36	Wilson2001*(P)50*	33	3	2	33
Maluccio2001 *(P)25*	28	6	7	63	Wong2002*(P)51*	35	1	2	12
McDonough2002 *(R)26*	9	2	2	16	Yetkin2002*(R)52*	42	3	3	17

Table 1.1: *Cross classified tables for 52 studies reporting diagnostic results of Computer Tomography scans used to diagnose appendicitis.*

In Table 1.1 diagnostic results for the *i*th Study ($i = 1, \ldots, 52$) are denoted by (tp_i, fp_i, fn_i, tn_i), those quantities are the usual information that each study reports and they are summarized in a 2×2 table as follows:

		Patient status	
		With disease	Without disease
Test	+	tp_i	fp_i
outcome	-	fn_i	tn_i
Sum:		$n_{i,1}$	$n_{i,2}$

where tp_i and fn_i are the number of patients with positive and negative diagnostic results in the group with disease and fp_i and tn_i are the number of patients with positive

and negative test results in the group without disease respectively. The total number of patients with disease is $n_{i,1} = tp_i + fn_i$ and the total number of patients without disease is $n_{i,2} = fp_i + tn_i$. In Table 1.1 (R) and (P) indicate that the study has a retrospective or prospective design respectively.

Common summary statistics describing test accuracy can be estimated for each study, the most commonly used are the empirical *true positive rate* or *sensitivity* and the empirical *true negative rate* or *specificity*,

$$\widehat{\text{TPR}}_i = \frac{tp_i}{n_{i,1}}, \quad \widehat{\text{TNR}}_i = \frac{tn_i}{n_{i,2}}, \tag{1.1}$$

and their complementary empirical rates, *the false positive rate* ($\widehat{\text{FPR}}$) and *the false negative rate* ($\widehat{\text{FNR}}$),

$$\widehat{\text{FPR}}_i = \frac{fp_i}{n_{i,2}}, \quad \widehat{\text{FNR}}_i = \frac{fn_i}{n_{i,1}}. \tag{1.2}$$

Another common measure of diagnostic performance is the *diagnostic odds ratio*, which is usually estimated as

$$\widehat{\text{DOR}}_i = \frac{(tp_i + 0.5) \times (tn_i + 0.5)}{(fn_i + 0.5) \times (fp_i + 0.5)}. \tag{1.3}$$

This odds ratio is a measure of the discriminatory power of the test, i.e. how the test correctly classified the presence of disease between disease and non-disease populations[1]. In practice, for high technology discriminatory procedures we may expect values of DOR in the order of hundreds! As usual in statistics, the notation with hat indicates that a theoretical quantity has been estimated from the available data.

One of the main research interests of this systematic review was to give overall measurements of diagnostic accuracy of CT technology. Another one was to explore study characteristics or published information that may influence diagnostic results. In principle any of the above quantities can be pooled to give an overall diagnostic accuracy. However, *sensitivity* and *specificity* are usually interdependent and a marginal combination by averaging or pooling these quantities might be misleading [65]. In the next section we clarify this issue.

1.3 Diagnostic data generation process

In this section we briefly explain how a diagnostic test is usually designed in practice. The intention is to understand how TPR and FPR are tied to the particular threshold

[1] We add the 0.5 value to each frequency to avoid numerical indetermination.

value that defines test outcomes. This underlying relationship between TPR and FPR drives most of the meta-analytic methods for diagnostic test data.

Let y_j be a test measurement (or test score) of patient j and λ the positivity threshold value. For example in our running example, y_j may be the length of the appendix in millimeters calculated in a CT image and λ the length value such as disease is declared, for example 6 mm. In this way the test outcome variable, z_i, is such that

$$z_i = \begin{cases} 1, (\text{test positive}), & y_i \geq \lambda \\ 0, (\text{test negative}), & \text{otherwise.} \end{cases}$$

Now, suppose that we carried out this hypothetical diagnostic test in two groups, one group with n_1 patients with disease and another one with n_2 patients without disease. For a given positivity threshold value λ, we organize diagnostic test results, as usual, in a 2×2 table giving the number of positive and negative test results for patients with and without disease:

		Patient status	
		With disease	Without disease
Test outcome	+	$tp(\lambda)$	$fp(\lambda)$
	-	$fn(\lambda)$	$tn(\lambda)$
Sum:		n_1	n_2

where

$$tp(\lambda) = \#\{z_i = 1 | \text{presence of disease}, \lambda\},$$

denotes the number of patients with positive test results in the group of disease for a threshold value λ. In the same way,

$$tn(\lambda) = \#\{z_i = 0 | \text{absence of disease}, \lambda\},$$

is the number of patients without disease correctly classified for a threshold value λ. Clearly, in this setup we can estimate TPR and FPR for a given value of λ as,

$$\widehat{\text{TPR}}(\lambda) = \frac{tp(\lambda)}{n_1}, \quad \widehat{\text{FPR}}(\lambda) = \frac{fp(\lambda)}{n_2}.$$

A diagnostic test is usually calibrated by plotting the pair $(\widehat{\text{TPR}}(\lambda), \widehat{\text{FPR}}(\lambda))$ that results by changing the possible values of λ. The resulting graphical display is called the *Receiver Operating Characteristic* (ROC) curve. The ROC curve is a well established method in signal detection theory [54]. Modern applications included calibration of statistical classification procedures in machine-learning applications [61, pag.

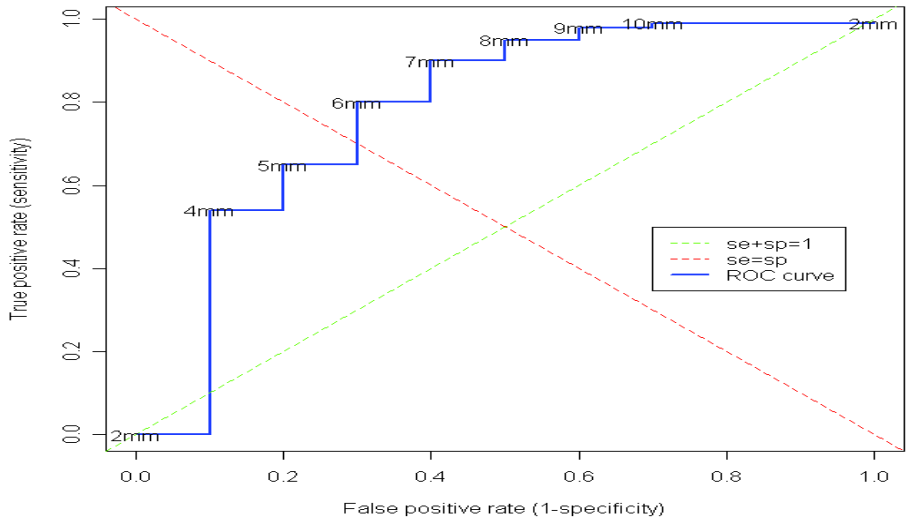

Figure 1.1: *Data for a single study: true positive rates vs false positive rates displayed for different threshold values of a diagnostic procedure.*

277-278]. The medical decision-making community has an extensive literature on the use of the ROC curve (see [133, 134] for reviews).

Figure 1.1 shows an artificial example where these pairs of operating characteristics $(\widehat{\text{TPR}}(\lambda), \widehat{\text{FPR}}(\lambda))$ are plotted at 9 different values of λ. In practice the test threshold λ is chosen as a tread-off between TPR and FPR. Usually, λ will be a value such that the ROC curve intercepts the diagonal line which goes from (0,1) to (1,0), that is the line where $sensitivity = 1 - specificity$.

The ROC curve is a detailed description of test accuracy. It is usually summarized by low dimensional functional statistics, one of the most common is the area under the curve (AUC). The AUC takes values between 0 and 1, and it is interpreted as the probability that in randomly chosen pairs of a diseased and non-diseased case, the disease case is correctly classified as more likely to have the disease. The empirical AUC estimate is in fact the Mann-Whitney version of the Wilcoxon two-sample rank sum statistics [58]. A value of AUC close to 1 indicates high test accuracy and values of AUC less than 0.5 indicates a test accuracy that is less accurate than flipping a coin.

In this way, the construction of diagnostic tests introduces a dependency between TPR and FPR due to λ. A meta-analytic procedure that aims to combine independent studies reporting diagnostic test accuracy should methodologically consider the relationship between TPR and FPR.

1.4 The SROC curve

Probably the most popular meta-analytic method to combine diagnostic results is the Summary Receiving Operation Characteristic (SROC) curve proposed by Moses et al. (1993) [87]. This method is a simple and elegant meta-regression approach inspired by old ideas of analyzing binary data, that can be found in Cox and Snell (1989)[15]. The SROC curve is calculated as follows, the logistic differences

$$\begin{aligned} \widehat{D}_i &= \text{logit}(\widehat{\text{TPR}}_i) - \text{logit}(\widehat{\text{FPR}}_i) && (1.4) \\ &= \log(\widehat{\text{DOR}}_i) \equiv \log\left(\frac{sensitivity \times specificity}{(1-sensitivity) \times (1-specificity)}\right) && (1.5) \end{aligned}$$

are modeled as a function of the logistic sum

$$\widehat{S}_i = \text{logit}(\widehat{\text{TPR}}_i) + \text{logit}(\widehat{\text{FPR}}_i) \equiv \log\left(\frac{sensitivity \times (1-specificity)}{(1-sensitivity) \times specificity}\right), \quad (1.6)$$

by fitting the regression line

$$\widehat{D}_i = A + B \times \widehat{S}_i + \epsilon_i, \quad (1.7)$$

where $\text{logit}(p) = \log(p/(1-p))$. The quantity $\widehat{D}_i$ is the estimated diagnostic odds ratio $\widehat{\text{DOR}}_i$ in the log scale, which summarizes the discriminatory power of the test. The variable $\widehat{S}_i$ is constructed in a way, that describes the overall level of test positive response in each study, it is 0 when sensitivity equals specificity, it is positive for studies with high sensitivity and low specificity and negative for studies with low sensitivity and high specificity. We may think that $\widehat{S}_i$ is a proxy variable for "a threshold" test value for positiveness between studies or as a summary variable for other study characteristics that may influence test positive result, e.g., a diagnostic setup that makes the test very sensitive.

In this model the estimated coefficient $\widehat{A}$ represents a pooled $\widehat{D}_i$ adjusted by some contextual influence captured by $\widehat{S}_i$. Meta-analysis is summarized by transforming back results from (S, D) to (FPR, TPR) by

$$\text{TPR} = \text{logit}^{-1}\left[\frac{\widehat{A}}{(1-\widehat{B})} + \frac{\widehat{B}+1}{(1-\widehat{B})} \times \log\left(\frac{\text{FPR}}{1-\text{FPR}}\right)\right], \quad (1.8)$$

the SROC curve is obtained by calculating TPR in a grid of values of FPR.

Figure 1.2 presents different SROC curves for a fixed value of A and for values of B equal to 0, -0.7, 0.7 and -2. Symmetric SROC curves are obtained when $B = 0$

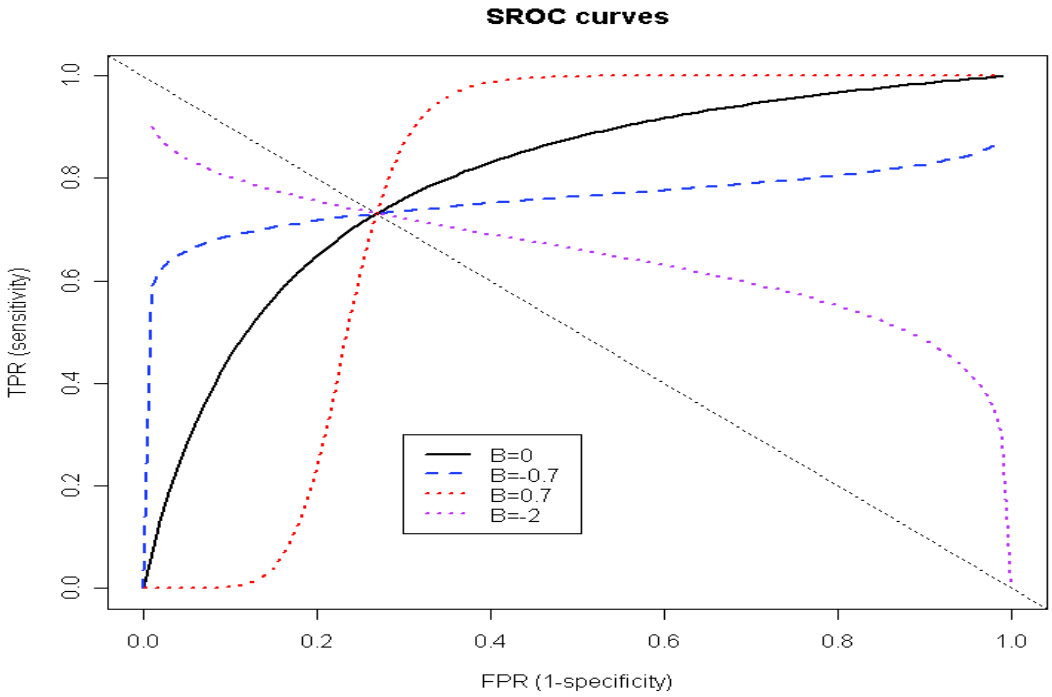

Figure 1.2: *Different resulting SROC curves for a fixed value of A and different values of B.*

indicating studies homogeneity with respect to $\widehat{S}_i$. Asymmetric curves are obtained for different values of B. One surprising feature of the SROC curve is that it may not behave as a ROC curve, we see in Figure 1.2 that for $B = -2$ the curve is monotonic decreasing. This suggests an implausible relationship between TPR and FPR. In practice we expect data yielding values of B within the range (-1,1). For more discussion about mathematical properties of the SROC curve see Walter [126, 127].

The left panel of Figure 1.3 shows the relationship between $\widehat{D}_i$ and $\widehat{S}_i$ for our data. The regression line corresponds to the simple model (1.7), with parameters estimated by ordinal least squares: $\widehat{A} = 5.735\,(0.179)$ and $\widehat{B} = -0.298\,(0.121)$. The right panel of Figure 1.3 presents the SROC curve for our data. The location of the upper left corner of the SROC and a $\widehat{\text{DOR}} = \exp(5.735) = 309.513$ indicates an excellent diagnostic result for the CT techniques. Most of this type of meta-analysis would end at this point, however, as we will see in the following chapters, we can extract more interesting information from this systematic review and summarize results in a more comprehensive way.

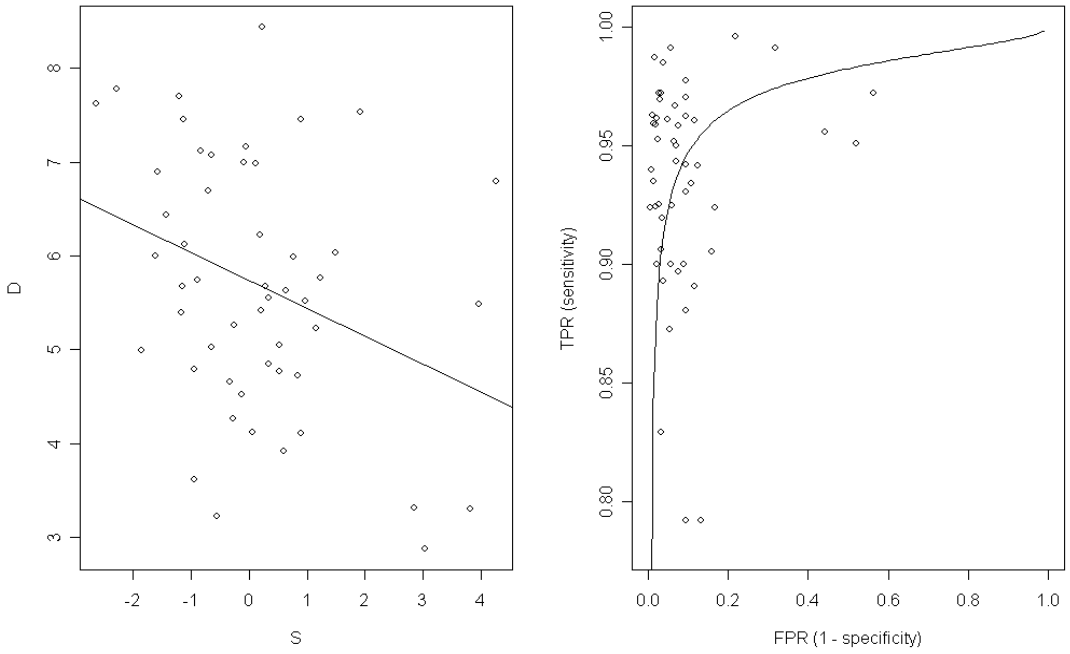

Figure 1.3: *Left panel: Scatter plot of $\widehat{D}_i$ against $\widehat{S}_i$ with a regression line estimated by ordinal least squares. Right panel: Scatter plot for $\widehat{FPR}_i$ and $\widehat{TPR}_i$ with the SROC curve.*

1.4.1 Meta-regression with the SROC curve

Moses et al. (1993)[87] also extended the SROC to a meta-regression approach to include variables that could systematically influence diagnostic test accuracy. We write this meta-regression model as:

$$\widehat{D}_i = A + B \times \widehat{S}_i + \alpha_1 x_{i,1} + \ldots + \alpha_k x_{i,k} + \epsilon_i, \tag{1.9}$$

where now $\widehat{D}_i$ depends also on a vector of covariates $(x_{i,1}, \ldots, x_{i,k})$ and a vector of regression coefficients $(\alpha_1, \ldots, \alpha_k)$ that has to be estimated from the data, usually by ordinal least squares. The regression coefficients are interpreted in the same way as in classical logistic regression analysis [15]. For example if $x_{i,1}$ is a factor variable which indicates study design:

$$x_{i,1} = \begin{cases} 1, & \text{Restrospecive design} \\ 0 & \text{Prospective desing,} \end{cases}$$

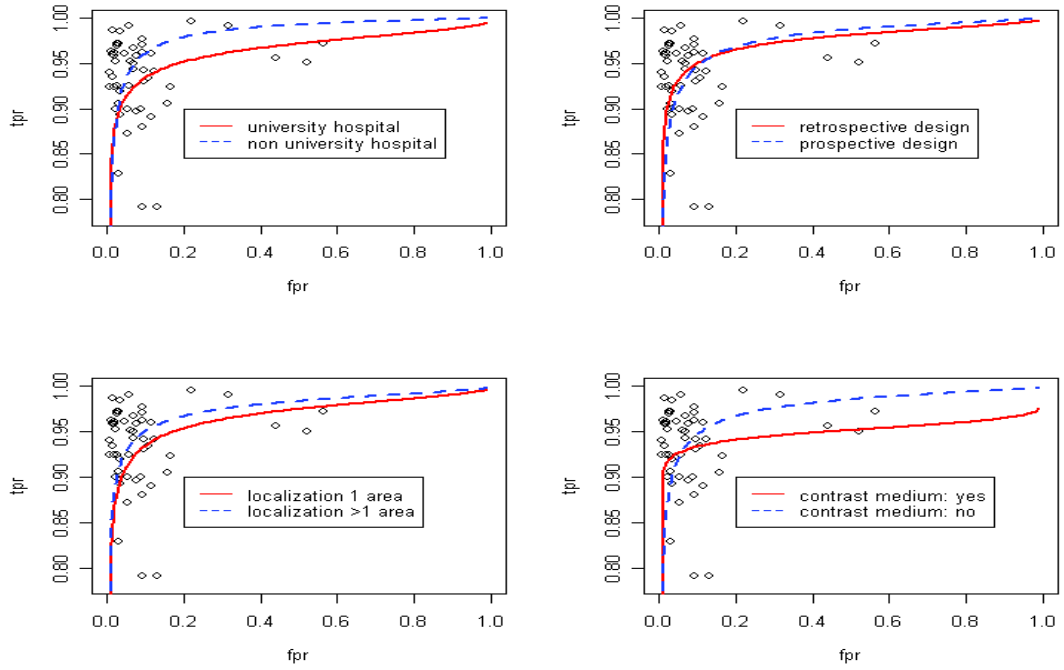

Figure 1.4: *SROC curves for different subgroups of studies. Upper panels: On the left SROC curves for studies in university hospitals vs. studies in non-university hospitals. On the right studies with retrospective design vs. studies with prospective design. Lower panels: On the left studies which use one localization area vs. more than one area. On the right studies with contrast medium vs. studies without contrast medium.*

then $\exp\widehat{\alpha_1}$ is an estimated odds ratio, which indicates increase or decrease of diagnostic accuracy of retrospective design with respect to prospective design. This simple regression approach has been extensively used to explore sources of heterogeneity in systematic reviews of diagnostic tests [65, 81, 80].

In Chapter 2 we give details of covariates that may influence diagnostic results. Now, as an example of the meta-regression model (1.9) we analyze the following 4 covariates: type of hospital, study design, contrast medium and localization. Figure 1.4 shows the effect on the SROC curve after including this covariates and Table 1.2 summarizes numerical results. We see an effect of type of hospital (p-value = 0.0872). Other effects that we see in Figure 1.4 can not be explained by model (1.9). In Chapter 6 we present a novel bivariate meta-regression approach that allows detection of, for instance the influence of contrast medium that is depicted in the lower left panel of Figure 1.4.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.476	1.232	3.634	0.001
S	−0.310	0.127	−2.435	0.019
hosp	0.638	0.365	1.747	0.087
design	−0.082	0.364	−0.227	0.822
contrast	−0.019	0.429	−0.045	0.964
local	0.263	0.413	0.636	0.528

Table 1.2: *Results of a meta-regression with covariates: S, type of hospital, study design, contrast medium and localization*

1.4.2 Summary statistics based on the SROC curve

There are some efforts to summarize results based on the SROC curve in a single number, the most commonly discussed in the bibliography are the Q^* index proposed by Moses et al. (1993) [87] and the Area Under the SROC curve (AUC)proposed by Walter (2002) [126]. The index Q^* summarizes the SROC curve by the point where *sensitivity = specificity*, this point has coordinates

$$\text{TPR}_{se=sp} = \frac{\exp(\widehat{A}/2)}{1+\exp(\widehat{A}/2)} \quad \text{and} \quad \text{FPR}_{se=sp} = \frac{1}{1+\exp(\widehat{A}/2)}.$$

Moses et al. (1993)[87] define

$$Q^* = \frac{\exp(\widehat{A}/2)}{1+\exp(\widehat{A}/2)}, \tag{1.10}$$

the value of TPR where *sensitivity = specificity*. The problem with this quantity is that several SROC curves may have the same Q^* index and be very different. See our example of Figure 1.2, all SROC curves share the same Q^* but they differ in their symmetry. To tackle this issue Walter (2002) [126] proposed to integrate the SROC curve in the range of TPR and summarize results with a single number, the area under the SROC curve (AUC). Later he proposed to integrate the SROC in the range where we have only observed data and he called this summary the partial area under the curve (PAUC) (Walter, 2005 [127]).

Now, suppose that we have performed a meta-analysis and we get estimate values $(\widehat{A}, \widehat{B})$ if we wish to summarize our meta-analysis with the AUC, we need to calculate

$$\widehat{\text{AUC}} = \int_0^1 \text{logit}^{-1}\left[\frac{\widehat{A}}{(1-\widehat{B})} + \frac{\widehat{B}+1}{(1-\widehat{B})} \times \log\left(\frac{x}{1-x}\right)\right] dx, \tag{1.11}$$

which has to be numerically integrated. We can use the function `integrate()` in R, which uses adaptive quadrature in the range of 0 to 1. For our meta-analysis we have

$$\widehat{\text{AUC}} = 0.981.$$

However, here the problem is how to make statistical inference about $\widehat{\text{AUC}}$, for example, how to calculate its standard error, build confidence intervals, make statistical tests, and so on. Walter (2002)[126] gives approximative standard errors based on the delta method (Davison, 2003 pag. 33)[18] for the $\widehat{\text{AUC}}$, which results in the following expression:

$$\begin{aligned}
\text{var}(\widehat{\text{AUC}}) &= \left[\frac{\exp\left(\frac{\widehat{A}}{1-\widehat{B}}\right)}{1-\widehat{B}}\int_0^1 \frac{\left(\frac{x}{1-x}\right)^p}{\left[1+\left(\frac{x}{1-x}\right)^p \exp\left(\frac{\widehat{A}}{1-\widehat{B}}\right)\right]^2}dx\right]^2 \text{var}(\widehat{A}) \\
&+ \left[\frac{\exp\left(\frac{\widehat{A}}{1-\widehat{B}}\right)}{(1-\widehat{B})^2}\int_0^1 \frac{\left(\widehat{A}+2\log\left(\frac{x}{1-x}\right)\right)\left(\frac{x}{1-x}\right)^p}{\left[1+\left(\frac{x}{1-x}\right)^p \exp\left(\frac{\widehat{A}}{1-\widehat{B}}\right)\right]^2}dx\right]^2 \text{var}(\widehat{B}) \\
&+ \left[\frac{\exp\left(\frac{\widehat{A}}{1-\widehat{B}}\right)}{1-\widehat{B}}\int_0^1 \frac{\left(\frac{x}{1-x}\right)^p}{\left[1+\left(\frac{x}{1-x}\right)^p \exp\left(\frac{\widehat{A}}{1-\widehat{B}}\right)\right]^2}dx\right] \times \\
&\times \left[\frac{\exp\left(\frac{\widehat{A}}{1-\widehat{B}}\right)}{(1-\widehat{B})^2}\int_0^1 \frac{\left(\widehat{A}+2\log\left(\frac{x}{1-x}\right)\right)\left(\frac{x}{1-x}\right)^p}{\left[1+\left(\frac{x}{1-x}\right)^p \exp\left(\frac{\widehat{A}}{1-\widehat{B}}\right)\right]^2}dx\right] \times \\
&\times \text{cov}(\widehat{A},\widehat{B}). \qquad (1.12)
\end{aligned}$$

Here, $\text{var}(\widehat{A})$, $\text{var}(\widehat{B})$ and $\text{cov}(\widehat{A},\widehat{B})$ are the components of the estimated variance covariance matrix of the ordinal least square estimate $(X^TX)^{-1}X^TD = (\widehat{A},\widehat{B})^T$:

$$(X^TX)^{-1}\widehat{\text{se}}^2 = \begin{pmatrix} \text{var}(\widehat{A}) & \text{cov}(\widehat{A},\widehat{B}) \\ \text{cov}(\widehat{A},\widehat{B}) & \text{var}(\widehat{B}) \end{pmatrix},$$

where X is the matrix with two columns, the first one with a vector of 1's repeated n times and the the second column is $(\widehat{S}_1,\ldots,\widehat{S}_n)^T$, and

$$\widehat{\text{se}}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(\widehat{D}_i - (\widehat{A}+\widehat{B}\,\widehat{S}_i))^2.$$

The integrals in (1.12) have to be integrated numerically. After evaluating these integrals we have

$$SE(\widehat{\text{AUC}})_{delta} = \sqrt{\text{var}(\widehat{\text{AUC}})} = 0.00358.$$

Walter (2002)[126], also, proposed to simplify (1.12) in the case where $\widehat{B} = 0$, then the integral (1.11) is analytically tractable and the standard error of $\widehat{\text{AUC}}$ is:

$$SE(\widehat{\text{AUC}})_{homog} = \frac{\widehat{\text{OR}}}{(\widehat{\text{OR}} - 1)^3}[(\widehat{\text{OR}} + 1)\log(\widehat{\text{OR}}) - 2(\widehat{\text{OR}} - 1)]SE(\widehat{A}) \qquad (1.13)$$

Here $\exp(\widehat{A}) = \widehat{\text{OR}}$ and $SE(\widehat{A})$ is the standard error of the estimate $\widehat{A}$. For our data we have

$$SE(\widehat{\text{AUC}})_{homog} = 0.0022,$$

which clearly underestimate (1.12) in our data. The reason is clear, our SROC is not symmetric.

Verde (2005)[119] presented a bootstrap analysis for the SROC curve that allows the inference of complex functional statistics like $\widehat{\text{AUC}}$ in realistic situations where $\widehat{B} \neq 0$ and (1.11) is not analytically tractable. Moreover, this bootstrap analysis allows the building of accurate confidence intervals of AUC for practical situations where we have small number of studies (say $n = 10$) in a meta-analysis. This work is presented in Chapter 3.

1.4.3 Classical critics on the SROC methods

The SROC curve has become the "off-the-shelf" procedure for meta-analysis of diagnostic test, however, we should highlight its clear methodological limitations:

- It is a fixed-effect meta-regression, where the assumptions of the linear regression model are usually unrealistic. Other estimation methods based on robust regression techniques and weighted regression have been proposed, but they are subject of debate [87, 126, 99]. Moreover, diagnostic test data present substantial variability between studies that can not be explained by a systematic change in "threshold values" [81].

- To avoid numerical problems, the common practice is to calculate diagnostic rates after adding 1/2 to the cells (tp_i, tp_i, fn_i, tn_i). This procedure induces a downward bias to the estimated rates, e.g. study number 7 has a specificity of 100% that after correction is 96%. This correction could be acceptable with a low amount of sparse data, but as we see in Table 1.1 about 30% of the studies reported 100 % sensitivities or specificities. Therefore, for a high technological diagnostic method it is more convenient to model the frequencies (tp_i, tp_i, fn_i, tn_i) directly without introducing *ad hoc* corrections.

- The variable S_i is assumed to be fixed but it is a sum of two random variables. Ignoring the randomness of S_i can produce a bias toward zero in the estimation of the parameter B [42]. Moreover, given that this variable conveys a different type of information as D_i, it is interesting to know which study information influenced its value. Therefore it is natural to model S_i as an outcome quantity.

- In the author's experience, it is not easy to explain the SROC model to medical researchers, whom usually expect more direct pooled summaries based on *sensitivities* and *specificities*.

1.5 Other statistical methods

Some of the issues mentioned above have been recently tackled with approximate bivariate meta-analysis models for sensitivities and specificities. While the SROC curve is an automatic procedure to make meta-analysis, the models presented in this section require for their correct application a substantial expertise from the statistician.

1.5.1 Bivariate models

Reitsma et al. (2005)[94] proposed to model the pairs $(\text{logit}(\widehat{\text{TPR}}_i), \text{logit}(\widehat{\text{TNR}}_i))^T$ with a bivariate Gaussian model with bivariate Gaussian random effects. This model can be written as

$$(\text{logit}(\widehat{\text{TPR}}_i), \text{logit}(\widehat{\text{TNR}}_i))^T \sim \texttt{Normal}_2(\mu_i, \Sigma_i) \quad i = 1 \ldots, N \tag{1.14}$$

where μ_i is assumed to come from a bivariate normal distribution

$$\mu_i \sim \texttt{Normal}_2(\mu, \Psi). \tag{1.15}$$

In this model the mean value parameter μ and the covariance matrix Ψ are both estimated with the data, the covariance matrix Σ_i is assumed know, diagonal and with diagonal elements given by the asymptotic variance of $\text{logit}(\widehat{\text{TPR}}_i)$ and $\text{logit}(\widehat{\text{TNR}}_i)$ calculated with the delta method, i.e.

$$\Sigma_i = \begin{pmatrix} S_{i,1}^2 & 0 \\ 0 & S_{i,2}^2 \end{pmatrix} \tag{1.16}$$

where

$$S_{i,1}^2 = \frac{1}{n_{i,1}\widehat{\text{TPR}}_i(1 - \widehat{\text{TPR}}_i)} \quad \text{and} \quad S_{i,2}^2 = \frac{1}{n_{i,2}\widehat{\text{TNR}}_i(1 - \widehat{\text{TNR}}_i)}. \tag{1.17}$$

This model is based on previous work on multivariate meta-analysis [117] and they rely upon *ad hoc* continuity corrections for sparse data, that may introduce a severe bias into the analysis [113]. Moreover, normal approximation (1.14) require large numbers of observations per study and stable estimates of μ and Ψ in (1.15) require large numbers of studies in the meta-analysis. In practice both requirements may be difficult to achieve.

To mitigate these issues, Arends (2006) [3] and Chu and Cole (2006) [12] independently proposed a bivariate generalized linear mixed effects model on the pairs (tp_i, tn_i). They presented a Binomial mixed effect model with a logit link function and Gaussian random effects. This model can be represented as following:

$$tp_i \sim \texttt{Bin}(\text{TPR}_i, n_{i,1}), \quad tn_i \sim \texttt{Bin}(\text{TNR}_i, n_{i,2}),$$

where

$$\mu_i = \begin{pmatrix} \text{logit}(\text{TPR}_i) \\ \text{logit}(\text{TNR}_i) \end{pmatrix}$$

and μ_i is assumed to follow a bivariate normal distribution

$$\mu_i \sim \texttt{Normal}_2(\mu, \Psi),$$

with $\mu = (\mu_A, \mu_B)^T$ the mean vector and

$$\Psi = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B \end{pmatrix},$$

the variance-covariance matrix for between studies effects. Note that in this model TPR and FPR are assumed unknown for each study, while in the Rietsma's model these quantities are assumed known and replaced by $\widehat{\text{TPR}}_i$ and $\widehat{\text{TNR}}_i$.

To fit this bivariate GLMM, these authors used the SAS procedure NLMIXED [100] which uses adaptive Gaussian quadrature to approximate the integrated likelihood. This model demands a formidable computational work and it may be sensitive to the numerical setup (e.g. starting values, number of points used in the adaptive quadrature procedure, etc.) [78]. To get stable estimates of Ψ we usually need a large number of studies included in the meta-analysis, this could be problematic in practice. Efficient ways to fit GLMMs are current research topics, for a practical overview of different estimation methods see Venables and Ripley (2002, pag. 292-300) [118].

The GLMM model is restricted to bivariate Gaussian random effects and no model-checking is presented by the authors. Given that in practice we never know which statistical model is correct, model checking is fundamental. In Chapter 5 and Chapter 6

we show, for example, that an appropriate analysis of the link function is fundamental and random effects that follows a bivariate-*t* distribution is more appropriate for our data. Model diagnostic and model-checking for statistical models with multiple sources of variability is a new area of research, some practical ideas have been recently presented by Lee, Nelder and Pawitan (2006, pag. 49-63)[77] and Gelman and Hill (2007, chap. 24) [47].

1.5.2 The HSROC model

Rutter and Gatsonis (2001)[99] introduce the HSROC model, a full Bayesian hierarchical regression approach based on a model for ordinal regression, that has been used to estimate ROC curves in single studies [116]. This model has a less intuitive parametrization, that has limited its popularity in practice. The HSROC model can be written in the following way:

$$tp_i \sim \texttt{Bin}(\text{TPR}_i, n_{i,1}), \quad fp_i \sim \texttt{Bin}(\text{FNR}_i, n_{i,2}),$$

where

$$\text{logit}(\text{TPR}_i) = (\theta_i + \alpha_i X_{i,j}) \exp(-\beta X_{i,j}), \tag{1.18}$$

$$\text{logit}(\text{FPR}_i) = (\theta_i - \alpha_i X_{i,j}) \exp(\beta X_{i,j}). \tag{1.19}$$

In this model the parameters (θ_i, α_i) represent the effects for study i, where θ_i is interpreted as a *cut-point* parameter or positivity criteria and α_i is an accuracy parameter. The last one can be worked out as the difference between TPR_i and FPR_i. The variable $X_{i,j}$ is usually taken as 1/2 and the parameter β is a scale parameter, if $\beta \neq 0$ then the diagnostic odds ratio change with θ_i even if α_i is held fixed. The study effects α_i and θ_i are assumed independent and are modeled as

$$\theta_i \sim \texttt{Normal}_1(\Theta, \sigma_\theta^2), \quad \alpha_i \sim \texttt{Normal}_1(\Lambda, \sigma_\alpha^2).$$

This model is based on a full Bayesian approach and priors for $(\Theta, \Lambda, \sigma_\alpha^2, \sigma_\theta^2)$ are modeled as independent with

$$\Theta \sim \texttt{Normal}_1(\mu_1, \sigma_1^2), \quad \Lambda \sim \texttt{Normal}_1(\mu_2, \sigma_2^2),$$

and

$$\sigma_\alpha^{-2} \sim \texttt{Gamma}(a_1, b_1), \quad \sigma_\theta^{-2} \sim \texttt{Gamma}(a_2, b_2)$$

with hyper parameters $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ and (a_1, b_1, a_2, b_2) considered known and carefully selected to be informative. Posteriors are calculated by Gibbs sampling using BUGS language.

Although, computations with Gibbs sampling can be very intensive, one important advantage of this method is that it does not relay on numerical approximations, estimation of posteriors is exact (up to Monte Carlo error). Therefore, in meta-analysis where we expect small numbers of studies to be included for analysis, Gibbs sampling is a much more reliable method than methods which combine numerical integration with optimization.

Rutter and Gatsonis (2001)[99] proposed to summarize meta-analytic results by recovering a SROC curve from their model. This curve is calculated by allowing the threshold parameter θ_i to vary while holding the accuracy parameter α_i fixed at its mean Λ, then the expected sensitivity for a given specificity is

$$\text{logit(sensitivity)} = E(\Lambda)\exp(E(\beta)/2) - \exp(E(\beta))\,\text{logit(specificity)}, \qquad (1.20)$$

where expectations are calculated over the marginal posterior distribution of Λ and β. The authors suggest that the curve is restricted to the observed rage of estimated specificities and they do not encourage extrapolation beyond this range. When the scale parameter β is 0, then the curve is symmetric around the point where sensitivity is equal to specificity.

Macaskill (2004)[83] proposed to use Empirical Bayes methods to fit the HSROC model and avoid Gibbs sampling, for the data investigated in the paper, results between Gibbs sampling and EB closely agree. The HSROC approach has been extended to handle different numbers of threshold values per study [27].

1.5.3 Relationships between bivariate and HSROC models

Harbord et al.(2007) [59] investigate the relationship between the parametrization used in the HSROC model and the bivariate binomial-gaussian meta-analysis approach. The authors ignore the Bayesian structure of the original HSROC model, and they ignore that the likelihood of the bivariate model. They concentrate on the parametrization of the structural distribution of the studies' effects (θ_i, α_i). In this way, we write the relationship of the studies' effects between the bivariate and the HSROC model as following:

$$\text{logit}(\text{TPR}_i) = b^{-1}\left(\theta_i + 1/2\,\alpha_i\right) \qquad (1.21)$$

$$\text{logit}(\text{TPR}_i) = -b^{-1}\left(\theta_i - 1/2\,\alpha_i\right), \qquad (1.22)$$

where $b^{-1} = \exp(\beta/2)$. Clearly we can express these relationships with matrix notation:

$$\begin{pmatrix} \text{logit}(\text{TPR}_i) \\ \text{logit}(\text{TPR}_i) \end{pmatrix} = S^{-1} \begin{pmatrix} \theta_i \\ \alpha_i \end{pmatrix}, \text{ where } S^{-1} = \begin{pmatrix} b^{-1} & \frac{1}{2}b^{-1} \\ -b & \frac{1}{2}b \end{pmatrix}.$$

Or the invert transformation

$$\begin{pmatrix} \theta_i \\ \alpha_i \end{pmatrix} = S \begin{pmatrix} \text{logit}(\text{TPR}_i) \\ \text{logit}(\text{TPR}_i) \end{pmatrix}, \text{ where } S = \begin{pmatrix} \frac{1}{2}b & -\frac{1}{2}b^{-1} \\ b & b^{-1} \end{pmatrix}. \tag{1.23}$$

S is a transformation matrix, which changes the coordinates from the bivariate model (logit transformed sensitivities and specificities) to the coordinates of the HSROC model (cutpoint and accuracy parameters). We can note that the matrix S is not orthogonal ($S^{-1} \neq S^T$), consequently when plotted in the bivariate model space (logit-ROC space), the axes corresponding to the HROC model are not perpendicular to each other. Taking expectation and variances of both sides of (1.23) we can express the relationship between the means

$$\begin{pmatrix} \Theta \\ \Lambda \end{pmatrix} = S\mu,$$

and variances

$$\begin{pmatrix} \sigma_\theta^2 & 0 \\ 0 & \sigma_\alpha^2 \end{pmatrix} = S\Psi S^T.$$

Thus, S is a non-orthogonal transformation matrix that diagonalizes the variance-covariance matrix of the bivariate model, these off-diagonal elements are zero if and only if $b = \sqrt{\sigma_A/\sigma_B}$ or in terms of β

$$\beta = \log(\sigma_B/\sigma_A).$$

Interestingly, this shape parameter does not depend on the correlation between sensitivity and specificity in the logit scale. In the same way, we can work out the relationship between the other model parameters:

$$\Theta = \frac{1}{2}\left\{(\sigma_B/\sigma_A)^{1/2}\mu_A - (\sigma_A/\sigma_B)^{1/2}\mu_B\right\}, \tag{1.24}$$

$$\Lambda = (\sigma_B/\sigma_A)^{1/2}\mu_A + (\sigma_A/\sigma_B)^{1/2}\mu_B, \tag{1.25}$$

$$\sigma_\theta^2 = \frac{1}{2}(\sigma_A\sigma_B - \sigma_{AB}), \tag{1.26}$$

$$\sigma_\alpha^2 = \frac{1}{2}(\sigma_A\sigma_B + \sigma_{AB}). \tag{1.27}$$

These equations can be inverted to give the five parameters of the bivariate model in terms of those of the HSROC model:

$$\mu_A = b^{-1}(\Phi + \frac{1}{2}\Lambda), \tag{1.28}$$

$$\mu_B = -b(\Phi - \frac{1}{2}\Lambda), \tag{1.29}$$

$$\sigma_A^2 = b^{-2}(\sigma_\theta^2 + \frac{1}{4}\sigma_\alpha^2), \tag{1.30}$$

$$\sigma_B^2 = b^2(\sigma_\theta^2 + \frac{1}{4}\sigma_\alpha^2), \tag{1.31}$$

$$\sigma_{AB} = -(\sigma_\theta^2 - \frac{1}{4}\sigma_\alpha^2). \tag{1.32}$$

The paper of Harbord et al.(2007) [59] is an interesting attempt to investigate the relationship between bivariate models. However, the marginal distribution of (θ_i, α_i) in a binomial-gaussian generalized linear mixed model is not bivariate normal and the random effects are not parametrization invariant. So, the above relationships may hold only asymptotically. Unfortunately, the authors do not mention these issues. Therefore, further work in this direction should be done.

1.5.4 Beta-binomial model for sensitivity and specificities

There is another Bayesian approach that has been recently proposed by Cong et al. (2007)[13]. They propose to model the number of true positive results and true negative results independently with a beta-binomial model. This model can be expressed as following:

$$tp_i \sim \texttt{Bin}(\alpha_i, n_{i,1}), \quad fp_i \sim \texttt{Bin}(\beta_i, n_{i,2}),$$

where

$$\alpha_i \sim \texttt{Beta}(\zeta_1, \zeta_2) \quad \beta_i \sim \texttt{Beta}(\kappa_1, \kappa_2),$$

with independent exponential priors on the parameters $\zeta_1, \zeta_2, \kappa_1$ and κ_2. They compare this model with a bivariate binomial-normal generalized linear model with logistic link function. They conclude that the beta-binomial model delivers results similar to the classical SROC model, while the binomial-normal model shows point estimates higher than those of the SROC approach, although the credible intervals overlapped. A sensitivity analysis showed that the Bayesian models are somewhat sensitive to the

variance of the prior distribution, but their point estimates are more robust than those of the SROC approach. In general they recommend to use a Bayesian approach.

1.6 Further methodological comments

We have seen in this review that new statistical methods has been proposed to overcome the limitations of the SROC curve. However, this simple and elegant model may be quite adequate in several applications. The use of the AUC as summary statistics are still been reported for authors. Until now, only the standard error of the AUC is attached to its estimation. More informative is to present a confidence interval for the AUC. In Chapter 3 we combine the simplicity of these techniques with the power of today computation to build confidence intervals for the AUC.

A common denominator of the models proposed in the bibliography is the universal use on the logistic link function. Although this link function can be suitable in some cases it can also badly fit the data at hand, especially when we have extreme results with very high TPRs as is usually found in high technology diagnostic procedures. In Chapter 6 we present a model with a general link function, the appropriate choice of this function is done by data analysis and not dogmatically imposed.

Droitcour et al. (1993) [25] highlight the restrictions and limitations of analyzing only one type of study design and introduce the term of *cross-design synthesis*, an approach to combine results from studies with different experimental designs (e.g. prospective, retrospective, case-control, etc). However, it is not easy to quantify the contribution of studies with different designs in the meta-analysis. And in our knowledge this has not been solved for multivariate problems. The use of bivariate mixture distributions for random effects presented in Chapter 6 is a practical approach to analyze these type of data.

The default assumption that random effects follow a normal distribution has been technically convenient. The SAS procedure NLMIXED [100] *only* allow us to use normal distributions for random effects and the same is for other statistical software (e.g. `lmer()` function in R). This assumption may not only be unrealistic but also dangerous in practice. Specially if a small number of studies present extreme results in comparison to the rest of the studies. This issue motivated the model extensions to consider multivariate heavy tails distribution in Chapter 6. This extension not only allow us to identify outliers but also automatically download the effects of these studies in the meta-analysis.

All models that we have reviewed in this section can be extended to include covariates to model systematic changes between primary studies. Although this is a simple conceptual extension, the previous authors did *not* highlight the technical complexity behind this point. In our knowledge this topic has not been carefully studied. For example the complex models like the HSROC have only been fitted with one single categorical covariate [99]. Moreover, model fitting, variables selection and inclusion of covariates with missing values in GLMMs are all topics of research at the time we are writing this work. In Chapter 6, we propose how to handle some these issues and we point out other avenues of research.

Modern meta-analysis has been growing during the recent years, a comprehensive introduction is given by Spiegelhalter et al. (2004)[108], an excellent review of complex multiparameter meta-analysis is presented by Ades and Sutton (2006)[2]. The introduction of the *Confidence Profile Method* (CPM) by Eddy, et al. (1992)[33] is a pioneer work in analyzing complex evidence synthesis scenarios.

Chapter 2

Data example: CT scans for diagnosis of appendicitis

"Why does he insist that we must have a diagnosis? Some things are not meant to be known by man."

-Susanna Gregory, *An Unholy Alliance*.

2.1 Introduction

In Chapter 1 we introduce our running example. In this chapter we document further details. These include: how primary studies have been selected, which medical databases have been used, how study quality and other studies' characteristics have been assessed. This information is fundamental to understanding that data collected in a systematic review is fundamentally different when compared to experimental data or observational studies.

It is important to highlight that data at patient level is usually not available in meta-analysis. The variables described in this chapter will be used in Chapter 6 to perform a multivariate meta-regression.

2.2 Information search and data extraction

The data of our running example resulted from an exhaustive search over 13 databases containing online medical publications. Papers were selected to be published in the period running from 1996 to 2003. They had to present diagnostic results for acute

abdominal pain with 10 patients or more. Diagnoses had to be performed with a Computer Tomography (CT) image without restriction on the CT technology used.

2.2.1 Electronic Database Searches

A systematic search was performed on the following databases: Catline (CA66); Cancerlit (CL63); HealthStar (HE75); Medline (ME66; ME60); Somed (SM78); Elsevier Biobase (EB94); Russmed Articles (SU88); CV72; Embase (EM74); Int Health Technology Assessment (HT83); Biosis Preview (BAA93; BA70); Index to Scientific and Technical Proceedings/Index to Social Sciences and Humanities Proceedings (II98; II78); SciSearch (IS74). The search was performed by DIMDI (Deutsches Institut fr Medizinische Dokumentation und Information). The following search patterns were used:

```
(Find "APPENDI?" and "COMPUTER?" and "DIAGN?")
    or(CT="APPEND ..." and CT="DIAGNOS ..." and
         CT="COMPUTER ...")
```

with related keywords, e.g.,

```
"APPENDICAL ABSCESS", "DIAGNOSIS ERROR"
"COMPUTER-ASSISTED DIAGNOSIS"
```

The second pattern was:

```
(Find "APPENDI?" and FT=CT and find "DIAGN?")
```

This searching process showed 1211 papers.

2.2.2 Selection of Studies

Two reviewers independently read articles' abstracts and classified them as: *relevant*, *potentially relevant* and *not relevant*. The articles' assessments were later compared and in case of disagreement a consensus meeting was held with these two reviewers and a third independent researcher. A total of 107 abstracts were classified as *relevant* and 94 as *potentially relevant*. Hard copies of these 201 articles were later classified by a forth independent researcher as *relevant* and *not relevant* . Finally, a total of 52 papers describing results of CT diagnostic for appendicitis were included in this systematic review.

Notation	Variable Name	Value description	NA (%)
x_1	Country	EU and others / USA	0
x_2	Type of hospital	University / Others	0
x_3	Inclusion criteria	Suspected / Appendectomy	0
x_4	Other CT findings included	No / Yes	0
x_5	Study design	Retrospective / Prospective	0
x_6	Contrast medium	No /Yes	0
x_7	Localization	One area / More than one area	0
x_8	Children included	No / Yes	0
z_1	Follow up	No / Yes	13 %
z_2	Valid reference standard	No / Yes	11 %
z_3	Sample	Selected / Random or systematic	21 %
z_4	Gender (majority of women)	No / Yes	26%

Table 2.1: *List of covariates describing study characteristics, patients characteristics, study quality and diagnostic setup.*

2.2.3 Data extraction and quality assessment

Standardized data extraction forms were used to collect the papers' results and to assess quality information (The Cochrane Methods Group on Systematic Review of Screening and Diagnostic Tests, Recommended methods: Screening and diagnostic tests, 2005). [114].

Table 2.1 gives some variables describing study characteristics (*Country*, *Type of hospital*), patients characteristics (*Inclusion criteria*, *Children included* , *Gender*), study quality (*Valid reference standard*, *Sample*, *Follow up*, *Design*) and diagnostic setup (*Contrast medium*, *Localization*). We analyze these data in Chapter 6 with a meta-regression model in order to understand how published information may influence diagnostic results.

Chapter 3

Bootstrap analysis of the SROC curve

"My general feeling about bootstrapping is that I don't like it very much. It's easy for me to say that, because nowadays I don't have to do practical problems for a living."

-Henry Daniels, *Statistical Science, August 1993*.

3.1 Introduction

The SROC curve is a standard method to perform meta-analysis of diagnostic test data (Moses et.al, 1993)[87], one important question is how to summarize meta-analytic results based on this graphical device. In Chapter 1, we presented the area under the SROC curve (AUC) as a comprehensive summary statistics for this type of meta-analysis (Walter, 2002)[126]. The AUC under the SROC curve is interpreted as the probability that in a pair of disease and non-disease subjects, the disease subject will be classified as more likely to have the disease.

We have seen that the current statistical inference for the AUC is based on classical statistical approaches. The application of the delta method presented by Walter (2002) [126] to calculate the standard error of the AUC is a typical example.

The aim of this chapter is to present *Modern Statistical Approaches* to extend the scope of statistical inference based on the SROC curve. For example, we not only expect to calculate standard errors but also confidence intervals and other measures of statistical accuracy. To be more specific, we investigate the following two questions:

- Can we reliably improve upon standard meta-analytic methods to make inference for the AUC?

- Can we make these methods fully automatic, in the sense that they can be summarized in an algorithm applicable in a wide spectrum of realistic applications?

To answer these questions, we present a novel bootstrap analysis for the SROC curve that has been originally introduced by Verde (2005)[119].

This chapter is organized as follows: In Section 3.2 we introduced bootstrap methods in general, in Section 3.3 we show how to apply these techniques to sampling SROC curves and to calculate the standard error and the bias of the AUC. Section 3.4 is a more technical section dedicated to bootstrap confidence intervals and its applications to the AUC. An extensive simulation experiment to evaluate different bootstrap confidence intervals is presented in Section 3.5. Finally, Section 3.6 gives some summary remarks. The implementation in R of all these bootstrap techniques is presented in Chapter 7.

3.2 Bootstrap methods

Bootstrap methods were introduced by Efron (1979)[36][1] as a general statistical approach to improve upon the Jackknife method to assess variability and to calculate standard errors of complex statistics. The main contribution of this seminal paper was that under the bootstrap earlier ideas of sampling techniques were synthesized in a new framework to perform simulation-based statistical analysis.

The main idea of the bootstrap[2] is to replace complicated and often inaccurate approximations of bias, variances, and other measures of uncertainty by computer power. These statistical methods are usually called *Computer Intensive Methods* a term popularized by Diaconis and Efron (1983) [26]. The aim of bootstrap is to answer routine questions that are far too complicated for traditional statistical analysis.

To start, suppose that our observed data $\{y_1, \ldots, y_n\}$ is a realization of a random sample drawn independently and identically distributed from an unknown distribution function F, i.e.

$$Y_1, Y_2, \ldots, Y_n \sim_{i.i.d.} F. \tag{3.1}$$

[1] Holmes et. al. 2003 presented the anecdote that this paper sent to the *Annals of Statistics* and it was turned down. The associate editor, said it that didn't have any theorems in it.

[2] Bootstrapping alludes to a German legend Baron Muenchhausen. According to the stories, the Baron's astounding feats included traveling to the Moon, and escaping from a swamp by pulling himself up by his own hair. In later versions he was using his own bootstraps to pull himself out of the sea, which gave rise to the term bootstrapping. Other more colorful names have been originally proposed. Efron's favorite was the *Shotgun*: *a method that can blow the head off any problem if the statistician can stand the resulting mess* (Efron 1979).

The sample space $\mathcal{Y}$ can be anything at all. For example, in our meta-analytic application $\mathcal{Y}$ is the space of 2×2 contingency tables summarizing diagnostic results in a meta-analysis, see Section 3.3. We are interested in making inference on a scalar parameter

$$\theta(F), \tag{3.2}$$

that will be, estimated from the data by

$$\hat{\theta} = t(\widehat{F}). \tag{3.3}$$

Here $\widehat{F}$ indicates the empirical probability distribution,

$$\widehat{F} : \text{probability mass}\,, 1/n\,, \text{on}\,, y_1, \ldots, y_n. \tag{3.4}$$

This choose of $\widehat{F}$ corresponds to the *non-parametric bootstrap*, which is the most general applicable bootstrap procedure. We further assume that $\hat{\theta}$ is a *symmetric* function of the data, i.e., does not depend on the sample order.

The bootstrap idea is very simple:

1. Think about your data $\{y_1, \ldots, y_n\}$ as a *hypothetical population*.
2. Assume that the data has been originated by the random mechanism induced by $\widehat{F}$.
3. Then, sample *with replacement* from your data. This procedure will generate *a bootstrap sample* say $\{y_1^*, \ldots, y_n^*\}$. The asterisk is used to denote a realization of a bootstrap sample. Note that matches are possible.
4. Now, from the bootstrap sample calculate your statistics of interest $\hat{\theta}$, say, $\hat{\theta}^*$. This is bootstrap replication of $\hat{\theta}$.
5. Repeat steps 3 to 4 a large number of times, say R. That will generate bootstrap values $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_R^*$.

The bootstrap values $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_R^*$ are used to make statistical inference for $\hat{\theta}$. For example, the standard error of $\hat{\theta}$, say $\widehat{\sigma}$, is estimated by the variability of $\hat{\theta}^*$ as:

$$\widehat{\sigma} = \sqrt{\frac{1}{(R-1)} \sum_{r=1}^{R} (\hat{\theta}_r^* - \hat{\theta}_{(\cdot)}^*)^2}, \tag{3.5}$$

where $\hat{\theta}_{(\cdot)}^* = 1/R \sum_{r=1}^{R} \hat{\theta}_r^*$.

In the same way, assessment of bias of $\hat{\theta}$ can be approximated by

$$\text{bias} = \hat{\theta}^*_{(\cdot)} - \hat{\theta}. \tag{3.6}$$

As we can explain with details in Section 3.4, the bootstrap replicates $\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_R$ can be used to construct different types of confidence intervals of θ.

In bootstrap methods *sample from the sample* is used to model *sample from the population*, this idea is implemented in practice by computer simulation. In this regard, bootstrap methods rely on computer power to make the inferential part of the data analysis. We can summarize the bootstrap methods in the following classical schematic form:

$$\begin{array}{ccccc} \textit{Real World} & & & & \textit{Bootstrap World} \\ F & \rightarrow & \mathbf{y} & \Longrightarrow & \widehat{F} \rightarrow \mathbf{y}^* \\ & & \downarrow & & \downarrow \\ & & \widehat{\theta} & & \widehat{\theta}^* \end{array}$$

We wish to estimate the accuracy of statistics $\widehat{\theta}$ for estimating a parameter of interest θ. The point estimates $\widehat{F}$ for F delivers the bootstrap data $\{y^*_1, \ldots, y^*_n\}$. Statistical inference is based on the variability of bootstrap replications $\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_R$. The success of the bootstrap analysis in general depends on the statistical model $\widehat{F}$ which mimics F. In our meta-analytic application we will choose $\widehat{F}$ as the empirical distribution function that puts probability $1/n$ on each study included in the meta-analysis. That corresponds to the *non-parametric bootstrap*, which in this context can also be called a *cluster bootstrap* where studies are interpreted as clusters, which are selected by simple random sampling (see Field and Welsh (2007)[41], Section 3.3).

The two main comprehensive references of bootstrap methods are the introduction book of Efron and Tibshirani (1993) [39] and the most advanced volume of Davison and Hinkley (1997) [19] with more than 500 pages covering statistical applications in several areas. A classical theoretical text is given by Hall (1992) [57], which gives a detailed asymptotic treatment based on Edgeworth Expansions, while Mammen (1992, 1993) [84, 85] describes simulations intended to help when the bootstrap works, and gives theoretical results for various situations. Recent theoretical work from the statistical testing point of view is given by Janssen and Pauls (2003)[67]. In 2003 the journal *Statistical Science*[135] published a special issue on the 25 years of bootstrap methods, where we can find a compilation of new applications. The amount of bibliography on bootstrap methods produced during the last two decades is massive and we

do not intend to review this material. The most relevant publications will be reviewed during the presentation in following sections .

3.3 Bootstrapping the SROC curve model and the AUC

In order to apply the bootstrap methods to meta-analysis of diagnostic test, we proceed as following:

1. We define our data $\{y_1, \ldots, y_{52}\}$ as the set of the 2×2 contingency tables summarizing diagnostic test results for each study included in the meta-analysis, i.e.

$$y_i = (tp_i, tp_i, fn_i, tn_i) \quad \text{for} \quad i = 1, \ldots, 52.$$

2. We give probability 1/52 to each table in $\{y_1, \ldots, y_{52}\}$, i.e. we use a nonparametric bootstrap method.

3. We sample with reposition from $\{y_1, \ldots, y_{52}\}$ and for each bootstrap sample $\{y_1^*, \ldots, y_{52}^*\}$ we calculate a SROC* curve.

4. From each resulting SROC* curve we calculate the area under the curve AUC* integrating numerically (1.11).

5. We repeat from 3 to 4 a large number of times R.

Then the bootstrap values $\widehat{\text{AUC}}_1^*, \ldots, \widehat{\text{AUC}}_R^*$ are used to assess variability about $\widehat{\text{AUC}}$.

Figure 3.1 shows the bootstrap in action. On the left panel we have the first 100 bootstrap replications of the SROC curve, we take a total of R=1000. This amount of bootstrap is about 10 times what we need to calculate standard errors and large enough to calculate confidence intervals, see Efron (1987, Sec. 9), Efron and Tibshirani (1993,pag 50-53, pag 272-275), Davison an Hinkley (1997, pag 34-37, pag 155-156 and pag 248) and Hall (1992, pag. 306-311) for bootstrap sample size.

The right panel of Figure 3.1 gives the bootstrap distribution of $\widehat{\text{AUC}}$, we can see that this bootstrap distribution is not normal, it is asymmetric to the left, also it is bounded between 0 and 1, as we expect for values of the AUC. The variability generated between SROC curves and their AUC results from the different data scenarios produced by each bootstrap sample.

We have seen in Chapter 1 that the estimated value of the AUC was

$$\widehat{\text{AUC}} = 0.981,$$

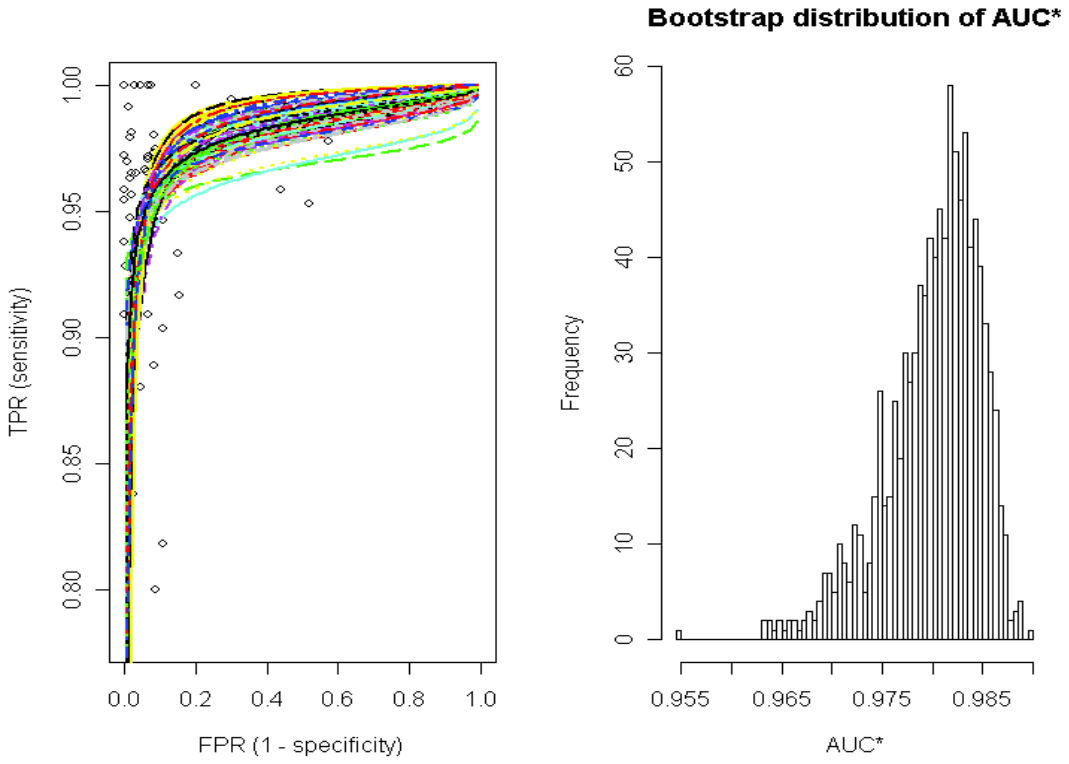

Figure 3.1: *Bootstrapping the SROC curve. Left panel: 100 bootstrap replications for the original SROC curve. Right panel: Bootstrap distribution of the AUC based on R=1000.*

from the bootstrap replicates its standard error calculated from (3.5) is

$$\widehat{\sigma} = 0.0045$$

and the bias of $\widehat{\text{AUC}}$ calculated from (3.6) is

$$\text{bias} = -0.0011.$$

In Chapter 1 we show that the approximated standard error for the $\widehat{\text{AUC}}$ using the delta method [126] yields

$$SE(\widehat{\text{AUC}})_{delta} = 0.0035$$

compared with our bootstrap calculation (1.12) slightly underestimates the variability of $\widehat{\text{AUC}}$.

One important advantage of the bootstrap estimate of standard error is that it is produced fully automatic, without making any special assumption about the stochastic

mechanism which generate the data, and without the formidable analytical calculations with produced formula (1.12). Unfortunately, all of these important advantages are mitigated when we want to build a confidence interval for AUC. As we will see in Section 3.4 much more inside is required from the data analysis point of view to build reliable confidence intervals.

3.4 Bootstrap confidence intervals for the AUC

In the previous section we introduced a simple but effective way to calculate standard errors of $\widehat{\text{AUC}}$ based on bootstrap samples. This section concerns the construction of confidence interval for the AUC. That is, in order to assess the uncertainty about a scalar parameter value θ, we want to construct a random interval, say $I_{1-2\alpha}$ with nominal coverage $1-2\alpha$ such that, if θ is a true parameter value, then

$$\text{Prob}(\theta \in I_{1-2\alpha}) = 1 - 2\alpha. \tag{3.7}$$

There is a small number of cases in applied statistics, were exact confidence intervals can be calculated, for example, the use of the *t*-distribution to calculate the confidence interval of the mean of normal distribution with unknown variance, the use of the *F*-distribution for the ratio of two variance of normal data, etc. Those are all familiar results of applied statistics, that can be found in statistical textbooks (see Mood, Graybill and Boes (1974)[63] Chapter 8).

However, in most applied problems these exact results are not possible and confidence intervals are calculated approximately. The use of bootstrap methods is particular convenient for this task. They try to automatically encapsulate sophisticated statistical thoughts that sometimes provide good solutions to complicated statistical problems.

In this section we review common bootstrap approaches to construct confidence intervals and we apply these techniques to make inference of AUC.

We present these techniques in a general way enabling it to be applied to other quantities of interest in meta-analysis. We also present a specially designed confidence interval (SDCI) for the AUC which results from a careful bootstrap analysis. These techniques are illustrated with the data of Chapter 2 and evaluated with a relatively extensive simulation experiment.

3.4.1 Normal confidence intervals

We can build confidence intervals for θ based on the estimated value $\widehat{\theta}$ and the bootstrap replicates $\widehat{\theta}_r^*$ (for $r = 1, \ldots, R$). The most simple approach is based on the *normal confidence intervals*. Suppose that the distribution of $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_R^*$ is perfectly normal with,

$$\widehat{\theta}_r^* \sim \mathrm{N}(\widehat{\theta}, \widehat{\sigma}^2), \tag{3.8}$$

then a confidence bound with level α is given by

$$\hat{\theta}_{NORM}[\alpha] = \hat{\theta} + z^{(\alpha)}\widehat{\sigma}. \tag{3.9}$$

Here, $z^{(\alpha)}$ is the 100αth percentile of a normal deviate, e.g. $z^{(0.95)} = 1.645$ and $\widehat{\sigma}$ is (3.5) . For example a 95% standard confidence interval has upper and lower limits given by

$$(\hat{\theta}_{NORM}[0.025],\ \hat{\theta}_{NORM}[0.975]).$$

The notation of the confidence intervals bounds emphasizes that we wish to have a coverage at *both ends* of the interval. This is crucial at the moment of comparing different types of approximative confidence intervals.

Figure 3.2 shows the histogram of 1000 bootstrap replicates of $\widehat{\mathrm{AUC}}^*$. The solid lines show the location of the 95% normal confidence interval, which is

$$(\widehat{\mathrm{AUC}}_{NORM}[0.025],\ \widehat{\mathrm{AUC}}_{NORM}[0.975]) = (0.973, 0.991).$$

The asymmetry of the histogram of $\widehat{\mathrm{AUC}}^*$ is evident, which suggests that the model (3.8) is not appropriate for this application.

3.4.2 Percentile confidence interval

A more natural way to construct a confidence interval for θ is to use the quantiles of the empirical distribution function of $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_R^*$, that is

$$\widehat{G}(c) = \frac{\left\{\#\widehat{\theta}_r^* \leq c\right\}}{R}. \tag{3.10}$$

This method is called the *percentile confidence interval*. The α confidence bound is defined as

$$\hat{\theta}_{PERC}[\alpha] = \widehat{G}(\alpha)^{-1}, \tag{3.11}$$

which corresponds to the $R \cdot \alpha$th value in the ordered list of R replications of $\widehat{\theta}^*$. For example if $\alpha = 0.025$ and $R = 1000$, $\hat{\theta}_{PERC}[0.025]$ corresponds to the 25th ordered

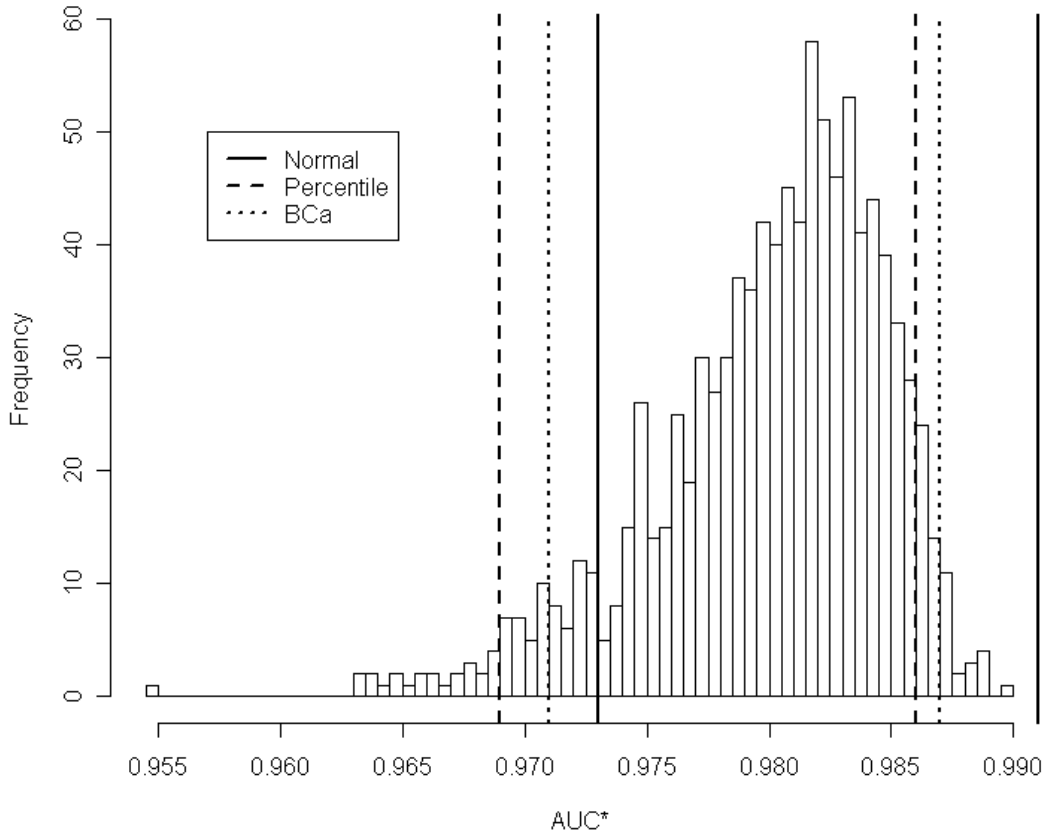

Figure 3.2: *Bootstrap distribution of the AUC. Vertical lines corresponds to different bootstrap confidence intervals based on this histogram: solid line is the Normal CI, dotted lines percentile method and dashed line BC_a method.*

value of bootstrap replications. If $R \cdot \alpha$ is not an integer, we take the *k*th largest value such that $k \leq (R+1)\alpha$ [39, pag 160].

Again with $R = 1000$, the AUC the percentile confidence interval is

$$(\widehat{\text{AUC}}_{PERC}[0.025],\ \widehat{\text{AUC}}_{PERC}[0.975]) = (0.969, 0.986).$$

We can see in Figure 3.2 that the percentile interval corrects the normal interval by shifting the interval to the left. Clearly, when the bootstrap distribution is asymmetric the normal confidence interval (3.9) is not appropriate and it will differ from the percentile confidence interval.

The percentile method generalizes the normal confidence interval to allow asym-

metry in the distribution of $\widehat{\theta}$. It is based on the following theoretical model: Suppose that there is a monotonic increasing function $\phi = m(\theta)$ that perfectly normalizes the distribution of $\widehat{\theta}$:

$$\widehat{\phi} - \phi \sim \mathrm{N}(0, \sigma_{\phi}^2). \tag{3.12}$$

Then under this scale a confidence bound of level α is

$$\hat{\phi}[\alpha] = \hat{\phi} + z^{(\alpha)}\sigma_{\phi}, \tag{3.13}$$

which back transforming to the original scale of θ with $m^{-1}(\cdot)$ gives

$$\hat{\theta}_{PERC}[\alpha] = m^{-1}(\hat{\phi}[\alpha]) = m^{-1}(\hat{\phi} + z^{(\alpha)}\sigma_{\phi}).$$

Finding a scale where the distribution of θ is normalized is a well known method in applied statistics, which improves the practical performance of the normal confidence interval, some commonly used transformations include the logistic transformation of the odds ratio, and the Fisher's transformation of the correlation coefficient. The breakthrough of the percentile method was that in practice we do *not* need to know $m(\cdot)$. This transformation is implicitly constructed by *computational brute force* from the bootstrap values θ^*.

This confidence interval has two very important properties, that the normal method does not share: First, it is *transformation-invariant*, that is the confidence interval for a parameter ψ resulted from a monotonic transformation $g(\theta) = \psi$ is the percentile confidence interval for θ mapped by $g(\theta)$:

$$(\hat{\psi}_{PERC}[\alpha],\, \hat{\psi}_{PERC}[1-\alpha]) = (g(\hat{\theta}_{PERC}[\alpha]),\, g(\hat{\theta}_{PERC}[1-\alpha])).$$

Second, the percentile interval is *range-preserving*, that is the confidence bounds fall within the range of values where θ is defined. For example, the AUC is a probability and we expect that a confidence interval falls within the range [0,1].

3.4.3 BC_a confidence intervals

BC_a stands for *bias corrected and accelerated*, this bootstrap confidence interval has been proposed by Efron (1987) [36] to improve the performance of the percentile confidence interval. It was an answer for the qualms of bootstrap confidence intervals pointed by Schenker (1985) [104]. The BC_a interval corrects the percentile method when the estimate $\widehat{\theta}$ is biased and when its standard error $\widehat{\sigma}$ depends on the value of $\widehat{\theta}$.

Biased estimates with non-constant standard errors are commonly encountered in applied problems (e.g odds ratios, correlation coefficients, etc.), making the BC_a method particularly attractive for practical purposes.

The BC_a method is *transformation-invariant* and *range-preserving* like the percentile method, but it was also a new achievement in bootstrap inference: their limits are *second-order accurate* and they are also *second-order correct*.

Let $\widehat{\theta}_{BC_a}[\alpha]$ be a BC_a interval limit with intended coverage α, by *second-order accurate* we mean that α is actually covered with probability $\alpha + O(1/n)$ i.e.

$$\text{Prob}\{\theta < \widehat{\theta}_{BC_a}[\alpha]\} = \alpha + O(1/n), \tag{3.14}$$

where n is the sample size in a i.i.d. (independent and identically distributed) ideal situation. This result can be compared with the normal and the percentile confidence interval, that they have both a slower *first-order accuracy* of $\alpha + O(1/\sqrt{n})$, i.e.

$$\text{Prob}\{\theta < \widehat{\theta}_{NORM}[\alpha]\} = \alpha + O(1/\sqrt{n}),$$

and

$$\text{Prob}\{\theta < \widehat{\theta}_{PERC}[\alpha]\} = \alpha + O(1/\sqrt{n}).$$

Now, let $\widehat{\theta}_{Exact}[\alpha]$ be a theoretical exact confidence limits with probability α, i.e.

$$\text{Prob}\{\theta < \widehat{\theta}_{Exact}[\alpha]\} = \alpha.$$

Then, a confidence interval limit is said to be second order correct, if it differs from $\widehat{\theta}_{Exact}[\alpha]$ by $O_p(1/n^{3/2})$. The second order accuracy and correctness of the BC_a method was originally proof by Hall (1988) [56].

The BC_a interval has a peculiar model construction which is far to be intuitive, but it is well motivated by the transformation theory that we describe in this section. As the percentile method, the BC_a postulates the existence of a monotonic increasing function $\phi = m(\theta)$ that perfectly normalizes the sampling distribution of $\widehat{\theta}$, with $\widehat{\phi} = m(\widehat{\theta})$ having distribution:

$$\widehat{\phi} - \phi \sim \text{N}(-z_0\,\sigma(\phi), \sigma(\phi)^2), \quad \sigma(\phi) = 1 + a\phi. \tag{3.15}$$

Here the constant z_0 plays the roll of *bias correction factor*. The coefficient a is a skewness correction factor called *acceleration constant*. In order to get $\sigma(\phi) > 0$ we assume that $\phi > -1/a$ if $a > 0$ and $\phi < -1/a$ if $a < 0$. The constant a is typically $|a| < 0.2$ and same is for z_0 [36, Sec. 3]. The use of z_0 and a is similar to the use of

Bartlett correction factors in likelihood inference for parametric models [19, pag 204] and see also [36, Sec. 3].

Under the previous conditions the confidence α level confidence limit for the BC_a method is given by:

$$\hat{\theta}_{BC_a}[\alpha] = \widehat{G}^{-1}\Phi\left(z_0 + \frac{z_0 + z^{(\alpha)}}{1 - a(z_0 - z^{(\alpha)})}\right). \tag{3.16}$$

Formula (3.16) looks strange, but we can see that for the case $z_0 = a = 0$ the confidence limit defined by (3.16) is

$$\hat{\theta}_{BC_a}[\alpha] = \widehat{G}^{-1}(\alpha)$$

the 100αth percentile of the bootstrap distribution. If in addition $\widehat{G}$ is perfectly normal, then

$$\hat{\theta}_{BC_a}[\alpha] = \widehat{\theta} + z^{(\alpha)}\widehat{\sigma},$$

the normal interval.

We can work out the percentile $\hat{\theta}_{BC_a}[\alpha]$ under the scale of ϕ and transform results back to the scale of θ as we did in the percentile method. However, it is more informative in terms of the statistical theory behind $\hat{\theta}_{BC_a}[\alpha]$ to do it in the following way: From (3.15) we can deduce that

$$\widehat{\phi} = \phi + (1 + a\phi)(Z - z_0), \quad Z \sim \mathrm{N}(0, 1),$$

then expanding terms on the right we have

$$\widehat{\phi} = \phi + Z + Za\phi - z_0 - z_0 a\phi,$$

multiplying both sides by a, adding 1 and collecting terms we have

$$1 + a\widehat{\phi} = \{1 + a\phi\}\{1 + a(Z - z_0)\}. \tag{3.17}$$

Taking logarithms results

$$\log(1 + a\widehat{\phi}) = \log(1 + a\phi) + \log(1 + a(Z - z_0))$$

which is a monotonic increasing in ϕ. Then calling $\widehat{\xi} = \log(1 + a\widehat{\phi})$, $\xi = \log(1 + a\phi)$ and $W = \log(1 + a(Z - z_0))$, we put the problem into the standard translation form,

$$\widehat{\xi} = \xi + W, \tag{3.18}$$

where the α confidence limit is given by

$$\widehat{\xi}[\alpha] = \widehat{\xi} - w^{(1-\alpha)}. \tag{3.19}$$

Here $w^{(\alpha)}$ is the $100 \cdot \alpha$ percentile point for W, i.e. Prob$\{W < w^{(\alpha)}\} = \alpha$.

We can go all the way back by mapping $\widehat{\xi}[\alpha]$ to the ϕ scale by noting that

$$\phi = \frac{\exp(\xi) - 1}{a}, \tag{3.20}$$

$$\widehat{\phi} = \frac{\exp(\widehat{\xi}) - 1}{a}, \tag{3.21}$$

$$(Z - z_0) = \frac{\exp(W) - 1}{a}, \tag{3.22}$$

which, after some little algebra results in

$$\widehat{\phi}[\alpha] = \widehat{\phi} + (1 + a\phi)\frac{(z_0 - z^{(\alpha)})}{1 - a(z_0 + z^{(\alpha)})}. \tag{3.23}$$

Now, according to the model (3.16) the α confidence limit of θ is $m(\widehat{\phi}[\alpha])^{-1}$, but of course $m(\cdot)$ is unknown. Using monotonicity of $m(\cdot)$ we have that

$$\widehat{G}(\widehat{\theta}_{BC_a}[\alpha]) = \text{Prob}(\theta^* < \widehat{\theta}_{BC_a}[\alpha]) \tag{3.24}$$

$$= \text{Prob}(\phi^* < \widehat{\phi}[\alpha]) \tag{3.25}$$

$$= \Phi\left(\frac{\widehat{\phi}[\alpha] - \phi}{\sigma(\phi)} + z_0\right) \tag{3.26}$$

$$= \Phi\left(z_0 + \frac{(z_0 - z^{(\alpha)})}{1 - a(z_0 + z^{(\alpha)})}\right), \tag{3.27}$$

and by applying $\widehat{G}(\cdot)^{-1}$ on both sides we have that the BC_a confidence limit is

$$\widehat{\theta}_{BC_a}[\alpha] = \widehat{G}^{-1}\left\{\Phi\left(\widehat{\phi} + \frac{(z_0 - z^{(\alpha)})}{1 - a(z_0 + z^{(\alpha)})}\right)\right\}, \tag{3.28}$$

which expressed in terms of the ordered list of simulated values $\widehat{\theta}^*$ gives

$$\widehat{\theta}_{BC_a}[\alpha] = \widehat{\theta}^*_{(R\tilde{\alpha})} \tag{3.29}$$

with

$$\tilde{\alpha} = \Phi\left(z_0 + \frac{(z_0 - z^{(\alpha)})}{1 - a(z_0 + z^{(\alpha)})}\right). \tag{3.30}$$

When $R \cdot \tilde{\alpha}$ is not an integer we take the largest integer k such that $k \leq (R+1)\tilde{\alpha}$.

We have seen that the theoretical construction of the BC_a limits involve two transformations:

1. First a monotone transformation

$$\theta \rightarrow \phi,$$

which normalizes the the sampling distribution of $\widehat{\theta}$.

2. Second, another monotone transformation

$$\phi \rightarrow \xi,$$

which reduce (3.15) to a translation problem. We can think about this transformation as a variance stabilization transformation.

The BC_a limits are calculated by transforming back from $\xi \rightarrow \phi$ and from $\phi \rightarrow \theta$. The most remarkable aspect, at least in theory, is that these transformations did not need to be known, they are replaced by computational power!

Of course in practice the constants z_0 and a have to be estimated. For z_0 we use the following result (Efron 1987, Section 4)[36]:

$$\text{Prob}(\theta^* < \widehat{\theta}) = \text{Prob}(\phi^* < \widehat{\phi}) = \Phi\left(z_0\right) \tag{3.31}$$

then we have

$$\widehat{z_0} = \Phi^{-1}\left\{\widehat{G}\left(\widehat{\theta}\right)\right\}, \tag{3.32}$$

which in terms of simulated values gives

$$\widehat{z_0} = \Phi^{-1}\left\{\frac{\#\widehat{\theta}^* < \widehat{\theta}}{R}\right\}. \tag{3.33}$$

The most common way to estimate the acceleration constant a is by

$$\widehat{a} = \frac{\sum_{i=1}^{n}(\widehat{\theta}(\cdot) - \widehat{\theta}(i))^3}{6\left\{\sum_{i=1}^{n}(\widehat{\theta}(\cdot) - \widehat{\theta}(i))^2\right\}}, \tag{3.34}$$

where $\widehat{\theta}(i)$ is the ith jackknife value, i.e. the estimation of θ by omitting the ith observation and $\widehat{\theta}(\cdot) = \sum_{i=1}^{n}\widehat{\theta}(i)/n$.

It is not obvious why the formula (3.34) provides an estimate of a. For a single parameter distribution problem it is an empirical estimation of one-sixth the skewness of the score function of model (3.15) evaluated at $\theta = \widehat{\theta} = m^{-1}(\widehat{\phi})$. In a multiple parameter situation, it has the same interpretation but the likelihood is profiled in the direction of the least favorable family (Stein, 1956) [111]. For the full non-parametric situation

a multinomial distribution is induced to reduce $\widehat{F}$ to a multiparameter model and the same theoretical arguments applied. For discussions and details about (3.34) see [36], [19, pag 205-209] and [29, Section 3]. DiCiccio and Romano (1990) [30] formallized the construction of non-parametric confidence intervals by reducing a non-parametric problem to a parametric problem with no nuisance parameters via the construction of a least favorable family. They show that the BC_a interval is a particular case of a general procedure to obtain second order accurate intervals.

We calculate the 95% confidence interval of the AUC with R=1000 and we get the following results:

$$(\widehat{\text{AUC}}_{BCa}[0.025],\ \widehat{\text{AUC}}_{BCa}[0.975]) = (0.971, 0.987).$$

The constants z_0 and a are estimated as

$$(\widehat{z}_0, \widehat{a}) = (0.037, 0.046).$$

The dashed lines of Figure 3.2 shows the effect introduced the corrections of this confidence interval. We can see that the BC_a shrinks the percentile method quite strongly and translate the interval to the right. The advantages of this elaborated theory can be appreciated in Section 3.4.6, where we reproduce similar calculations for smaller sample sizes (e.g. $n = 20$ or $n = 10$) and in the simulation experiment of Section 3.5.

3.4.4 Bootstrap-t intervals

The bootstrap-*t* confidence interval is a conceptually simple algorithm to construct a confidence interval. Its name comes from its analogy with the Student's *t*-statistic. This interval needs an estimate of the standard error $\widehat{\sigma}^*$ of the statistic $\widehat{\theta}^*$ for *each* bootstrap sample. It is based on the *studentized* statistic

$$T^* = \frac{\widehat{\theta}^* - \widehat{\theta}}{\widehat{\sigma}^*}. \tag{3.35}$$

The bootstrap distribution of T^* is used to estimate the distribution of

$$T = \frac{\widehat{\theta} - \theta}{\widehat{\sigma}}, \tag{3.36}$$

which of course is unknown in most situations[3]. By analogy of the Student-*t* confidence interval, the end points of a $1 - 2\alpha$ bootstrap-*t* confidence interval is defined

[3]Usually these pivotal quantities are presented with $n^{1/2}$ in front of the righthand equation, here this constant is absorbed by $\widehat{\sigma}^*$ and $\widehat{\sigma}$ respectively.

as

$$\widehat{\theta}_T[\alpha] = \widehat{\theta} - \widehat{T}^{(1-\alpha)}\,\widehat{\sigma}, \quad \widehat{\theta}_T[1-\alpha] = \widehat{\theta} - \widehat{T}^{(\alpha)}\,\widehat{\sigma}. \tag{3.37}$$

Here $\widehat{T}^{(\alpha)}$ is obtained by αth ordered value of the simulated T_r^* for $r = 1, \ldots, R$. For example if $R = 1000$ and $\alpha = 0.025$ then $\widehat{T}^{(0.025)}$ is the 25th ordered T_r^*.

This method was originally proposed by Efron (1979)[35], but poor numerical behavior reduced its interest. Babu and Singh (1983)[4] gave the first proof of second-order accuracy for the bootstrap-*t*. Hall (1988)[56] showed that the bootstrap-*t* limits are second-order correct and revived its interest. Davison and Hinkley (1997)[19] present extensive use of this technique in several applied problems. Venables and Ripley (2002, pag.137)[118] recommend its use in general applications.

The bootstrap-*t* is computationally very intensive[4]. It requires that we estimate $\widehat{\sigma}^*$ for each bootstrap sample. If we use a second level bootstrap to calculate $\widehat{\sigma}^*$ with R_2 bootstrap replications, then the number of evaluations of $\widehat{\theta}^*$ will be $R_2 \times R$. This computational burden is one of the drawbacks of this method. One remedy is to use the jackkinfe estimate of $\widehat{\sigma}^*$ in each bootstrap sample (see below).

Another drawback is that, unlike the percentile method and the BC_a, this method is not transformation invariant.

More dangerous in practice, the bootstrap-*t* algorithm may be very unstable. Its numerical problem is produced by the fact that $\widehat{\sigma}^*$ could be very small compared to $\widehat{\theta}^* - \widehat{\theta}$, this artefact produces an artificially heavy tailed distribution of T^* resulting in a very long confidence interval. This is particular dangerous in situations where the confidence limits must be bounded to the range where θ is defined.

The application of bootstrap-*t* intervals to the AUC is straightforward. We define the T^* statistics as

$$T^* = \frac{\widehat{\text{AUC}}^* - \widehat{\text{AUC}}}{\widehat{\sigma}^*}, \tag{3.38}$$

where $\widehat{\sigma}^*$ is calculated for each bootstrap sample with the Jackknife estimate

$$\widehat{\sigma}^{*2} = \left\{ \frac{1}{n-1} \sum_{i=1}^{n} \left(\widehat{\text{AUC}}^*(\cdot) - \widehat{\text{AUC}}^*(i) \right)^2 \right\} \tag{3.39}$$

where $\widehat{\text{AUC}}(i)$ is the ith jackknife value, i.e. the estimation of $\widehat{\text{AUC}}$ by omitting the ith observation and $\widehat{\text{AUC}}(\cdot) = \sum_{i=1}^{n} \widehat{\text{AUC}}(i)/n$. The α level confidence limit for AUC is

[4]Measured in computational time, the bootstrap-*t* interval is the most intensive statistical method presented in this work, more intensive than the Bayesian applications of MCMC sampling of the following chapters.

calculated as

$$\widehat{\text{AUC}}_T[\alpha] = \widehat{\text{AUC}} - \widehat{T}^{(1-\alpha)}\,\widehat{\sigma}. \tag{3.40}$$

We calculate the 95% confidence interval of the AUC based on R=1000 and we get the following results:

$$(\widehat{\text{AUC}}_T[0.025],\ \widehat{\text{AUC}}_T[0.975]) = (0.969, 0.992).$$

This calculation takes approximately 12 minutes while the percentile and BC_a takes approximately 12 seconds. Interesting, the lower limit is numerically similar to the BC_a method and the upper limit almost reach the maximum value of AUC.

3.4.5 Specially designed confidence interval for AUC

With its strong theoretical background and worrisome practical drawbacks, the bootstrap-*t* can be used as a starting point to build better bootstrap confidence intervals. There is considerable practical evidence that this method is likely to deliver good results if θ is a location parameter, such as a median or a mean. Tibshirani (1988)[115] proposed an algorithm for transforming a scalar parameter θ to a more location-like parameter $\phi = m(\theta)$, before applying the bootstrap-*t* method. The resulting interval is transformed back to the θ scale via $m^{-1}(\phi) = \theta$. The transformation $m(\cdot)$ is calculated with nonparametric smoothing techniques from the scatter plot of $(\widehat{\theta}^*, \widehat{\sigma}^*)$ and $m^{-1}(\cdot)$ is calculated by numerical integration. For more discussion about this approach see DiCiccio and Romano (1995)[31], Efron and Tibshirani (1993, Section 12.6)[39].

In this section we pursue a less general approach we concentrate *only* in the problem of finding a parametric transformation $h(\cdot)$ which stabilizes the variability of $\widehat{\text{AUC}}$. This is a pure heuristic approach based on recommendations given by recently DiCiccio, Monti and Young (2006)[32] and further applications given by Davison and Hinkley (1997)[19], Sections 3.9 and 5.2.

The plan that we follow is quite simple, first we prepare a *variance parameter plot*, which is the scatter plot of the points

$$(\widehat{\text{AUC}}^*, \text{Var}(\widehat{\text{AUC}})^*). \tag{3.41}$$

Then, we analyze a suitable transformation $h(\cdot)$ such that the T^* statistics

$$T^* = \frac{\text{h}(\widehat{\text{AUC}}^*) - \text{h}(\widehat{\text{AUC}})}{\sqrt{\text{Var}(\text{h}(\widehat{\text{AUC}}))^*}}, \tag{3.42}$$

is approximately a pivotal quantity, i.e. the distribution of T^* does not depend on AUC. We measure the success of $h(\cdot)$ by the correlation coefficient of the *variance parameter plot*. A transformation with correlation close to 0 gives points out a good variance stabilization function (see, DiCiccio, Monti and Young (2006)[32] Section 2.3)

In practice, it is quite difficult to identify a single transformation for this problem and we have had relatively bad empirical results with classical transformations, that includes the logistic, the log, the square root and the Box-Cox transformation. Therefore, we proposed a composed transformation based on the Box-Cox transformation. This transformation is build first by defining the odds of the $\widehat{\text{AUC}}^*$:

$$g(\widehat{\text{AUC}}^*) = \frac{\widehat{\text{AUC}}^*}{1 - \widehat{\text{AUC}}^*}, \tag{3.43}$$

and then by applying a Box-Cox transformation to $g(\widehat{\text{AUC}}^*)$:

$$\text{h}(\widehat{\text{AUC}}^*) = \begin{cases} \dfrac{\left[g\left(\widehat{\text{AUC}}^*\right)\right]^{\lambda} - 1}{\lambda} & \text{if } \lambda > 0 \\ \log\left(g\left(\widehat{\text{AUC}}^*\right)\right) & \text{if } \lambda = 0. \end{cases} \tag{3.44}$$

The parameter λ is estimated by maximum likelihood from the variance plot. Then the bootstrap confidence interval based on T^* is calculated as usual

$$\widehat{\text{AUC}}_{SDCI}[\alpha] = \text{h}^{-1}\left[\text{h}(\widehat{\text{AUC}}) - \widehat{T}^{(1-\alpha)}\sqrt{\text{Var}(\text{h}(\widehat{\text{AUC}}))}\right] \tag{3.45}$$

where $\widehat{T}^{(1-\alpha)}$ is the $(1-\alpha)$th ordered value of the simulated T as in the previous section, but this time calculated in the scale of $h(\cdot)$. Finally, these results are backward transformed by $h(\cdot)^{-1}$ to the scale of AUC.

The left panel of Figure 3.3 shows the scatter plot of $(\widehat{\text{AUC}}^*, \text{Var}(\text{AUC})^*)$[5]. The correlation of these points is -0.522. After applying the transformation $h(\cdot)$ this correlation is reduced to -0.017, the right panel of Figure 3.3 shows the effect of applying $h(\cdot)$ with $\widehat{\lambda} = 0.303$. The resulting 95% confidence interval for AUC is

$$(\widehat{\text{AUC}}_{SDCI}[0.025], \widehat{\text{AUC}}_{SDCI}[0.975]) = (0.967, 0.990).$$

[5] We have scaled the vertical axis to make the variability between panels comparable.

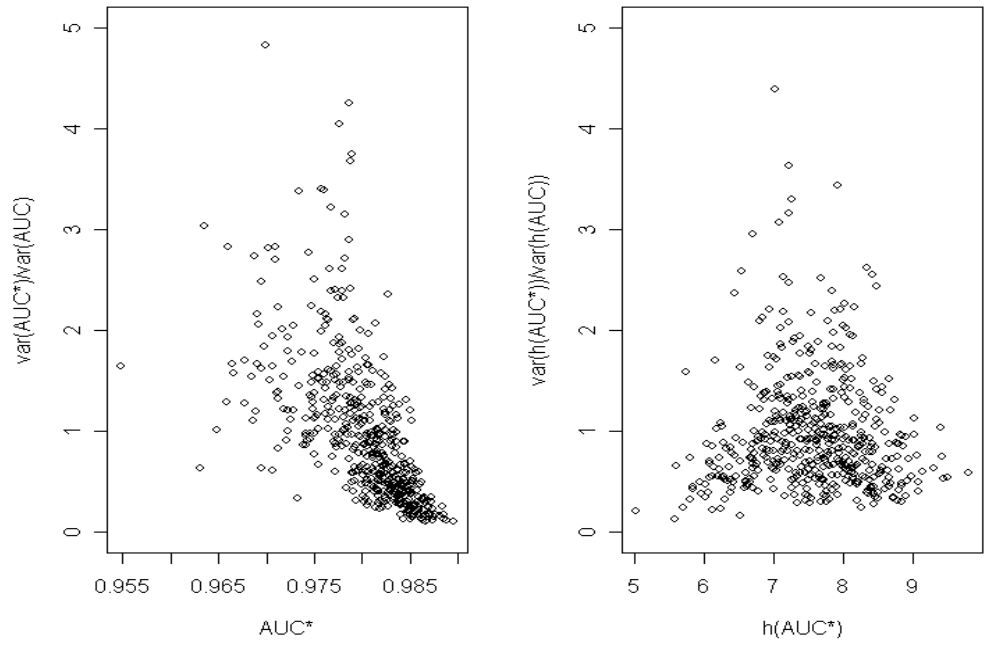

Figure 3.3: *Variance plot for bootstrap-t confidence interval. Left panel: each point represents* $(\widehat{AUC}^*, var(\widehat{AUC}^*))$. *Right panel: variance plot after a variance stabilization transformation* $h(\cdot)$.

As noted by Davison and Hinkley (1997) [19] Section 5.2 we can use a variance stabilization transformation to improve a normal interval as well, which can be calculated as

$$\widehat{\text{AUC}}_{NORMAL-SDCI}[\alpha] = \text{h}^{-1}\left[\text{h}(\widehat{\text{AUC}}) - z^{(1-\alpha)}\sqrt{\text{Var}(\text{h}(\widehat{\text{AUC}})),}\right], \qquad (3.46)$$

where $z^{(\alpha)}$ is the 100αth percentile of a normal deviate. Calculating the normal interval in the scale of $h(\cdot)$ and reporting results in the scale of AUC gives

$$(\widehat{\text{AUC}}_{NORMAL-SDCI}[0.025], \widehat{\text{AUC}}_{NORMAL-SDCI}[0.975]) = (0.971, 0.987).$$

One important advantage of this interval is that we do not need to calculate the bootstrap distribution of T. That substantially reduce the computation burden of this procedure.

Method	Sample size	Original scale		Transformed scale	
		Lower	Upper	Lower	Upper
Normal	n=52	0.973	0.991	0.971	0.988
Bootstrap-t		0.968	0.991	0.966	0.990
Percentile		0.969	0.987		
BC_a		0.969	0.987		
Normal	n=20	0.955	0.997	0.944	0.986
Bootstrap-t		0.928	1.039	0.898	0.998
Percentile		0.946	0.986		
BC_a		0.952	0.988		
Normal	n=10	0.721	1.320	0.959	0.995
Bootstrap-t		0.972	1.006	0.970	0.998
Percentile		0.141	0.990		
BC_a		0.969	0.992		

Table 3.1: *Results of bootstrap confidence intervals for the AUC calculated with different sample sizes (R=1000).*

3.4.6 Comparison of bootstrap confidence intervals with different sample sizes

In the previous subsections, we have seen that we obtain little practical differences between different methods for our data. One possible explanation is that we have a meta-analysis with large number of studies included, $n = 52$.

In this section we explore potentially different results between confidence intervals by reducing the number of studies in the analysis. We take two samples at random from the 52 studies, one with $n = 20$ and the other with $n = 10$. Then, we recalculate all the bootstrap confidence intervals. Table 3.1 summarizes these results.

We can summarize the effect of changing the sample size as following:

- For $n = 10$ and $n = 20$ both the normal and bootstrap-t give confidence bounds out of the range of the AUC, which is a pathological result.
- The percentile method is un-stable for $n = 10$.
- The BC_a presents stable numerical results when we change the sample size.
- The $SDCI$ deliver stable results for different sample sizes.

In the next section we present a simulation experiment to evaluate these confidence intervals for a more general application.

3.5 Simulation experiment

In the previous section we have presented 6 different confidence intervals for the AUC. How do these different confidence intervals perform in finite small samples? In this section we investigate two questions:

1. Which is most reliable, in the sense that the coverage is closest to nominal?
2. Which one does deliver a better quality inference in terms of interval length and range respecting (i.e. $0 \leq \text{AUC} \leq 1$)?

We perform a simulation study to estimate coverage, interval length and rates for exceeding the limits where AUC is defined. We expect that a good statistical method delivers both features good reliability and inferential quality.

3.5.1 Design of the experiment

Simulation experiments with bootstrap methods are computationally extremely demanding and in general it is wise to define carefully an experiment that covers realistic situations.

In order to define simulation scenarios that cover most of the common situations, we roughly define three different SROC curves with different values of A and B:

1. $A = 5$ and $B = 0.1$, this curve represents a meta-analysis where the diagnostic procedure has a very high diagnostic performance and studies are relatively homogeneous with respect to B. The AUC is 0.972.
2. $A = 3$ and $B = 0.20$, this curve represents a meta-analysis where the diagnostic performance is not so extreme like the previous one, but studies may show some substantial heterogeneity with respect to B. The AUC in this case is 0.883.
3. $A = 1.5$ and $B = 0.30$, this curve evidently represents a meta-analysis where studies in general report a lower diagnostic performance, AUC = 0.726, and there is also an important heterogeneity with respect to B causing a substantial asymmetry in the SROC curve.

Figure 3.4 shows these three SROC curves. We clearly see a gradient going from the most extreme homogeneous and highly diagnostic performance to the lower and asymmetry situation. We have chosen these parameter values based on published information ([126] and Walter 2008 personal communication) and our own data.

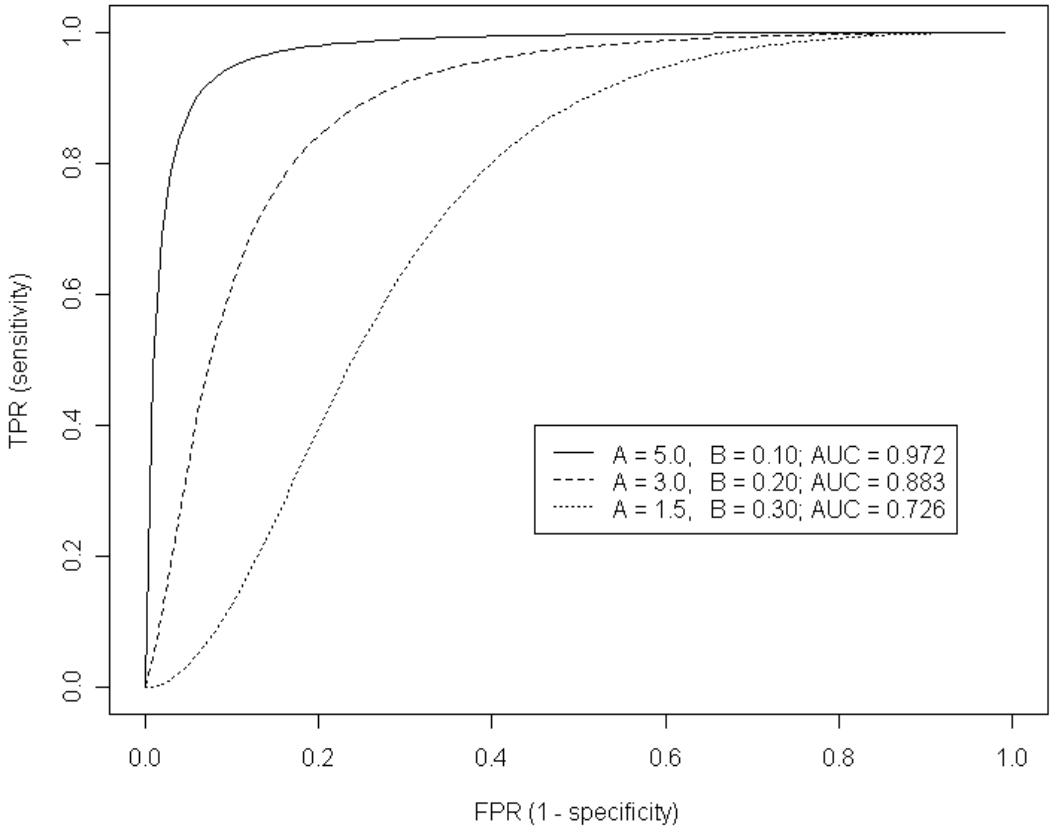

Figure 3.4: *Summary of the SROC behavior of 5 different scenarios for different values of A and B. These 5 cases are studied with 2 different sample sizes n = 10 and n = 20.*

The fact that the AUC is symmetric with respect to B has been helpful to reduce the number of scenarios to the half, e.i. we investigate only positive values of B. Another criteria that we follow to save number of scenarios was not to pretend to estimate the effect of A and B and their interaction. That would demand covering simultaneously different levels of A and different levels of B, this results in combinations of levels with very similar AUC values. For example, taking $A = 5$ and three levels of $B = 0.1, 0.2, 0.3$, results AUC = 0.972, 0.970, 0.967, which are of course an uninteresting difference to analyze.

Another important ingredient in the simulation experiment is the probability structure of the variables D and S which are used to estimate the coefficients A and B in the SROC curve. For this design component we use only information coming from our

own data. Values of (D, S) are simulated with a bivariate normal distribution with the following model:

$$(D_i^{sim}, S_i^{sim}) \sim \texttt{Normal}_2(\mu, \Sigma), \quad s = 1, 2, \ldots, N. \tag{3.47}$$

where

$$\mu = \begin{pmatrix} \mu_D \\ \mu_S \end{pmatrix} = \begin{pmatrix} A \\ \bar{S} \end{pmatrix} \tag{3.48}$$

and $\bar{S} = 0.138$ for our data. The covariance matrix is given by

$$\Sigma = \begin{pmatrix} \sigma_D^2 & \sigma_{D,S} \\ \sigma_{D,S} & \sigma_S^2 \end{pmatrix} \tag{3.49}$$

where using our data we have $\sigma_D^2 = 1.828$ and $\sigma_S^2 = 2.227$. For the covariance parameter we use standard results of regression analysis which gives $\sigma_{D,S} = B\sigma_S^2$. Clearly, the parameter values A and B are used to generate the possible three scenarios described above. Note, that in our experiment both D_i and S_i are random variables, like we expect in practice.

The simulated values (D_i^{sim}, S_i^{sim}) are mapped to the ROC space with

$$\text{TPR}_i^{sim} = (1 + \exp(-0.5\,(S_i^{sim} + D_i^{sim}))^{-1}$$

and

$$\text{FPR}_i^{sim} = (1 + \exp(-0.5\,(S_i^{sim} - D_i^{sim}))^{-1}$$

where TPR_i^{sim} and FPR_i^{sim} are the summary results of a simulated study. Figure 3.5 shows two examples, on the left panel we have 20 simulated studies coming from a SROC curve with $(A, B) = (5, 0.1)$ and the right panel 20 simulated studies with $(A, B) = (1.5, 0.3)$. In both panels the dotted line is the estimated SROC curve and the solid line the true SROC curve. We can see that the simulated data result in plausible real situations.

In order to evaluate the performance of the 6 confidence intervals, we simulate 1000 meta-analysis for each combination of values of (A, B) and with samples $n = 10$ and $n = 20$. These two sample sizes were chosen in order to cover most of the realistic situations, where we expect a small number of studies to be included in the meta-analysis. With this specification, we have a total of 6000 possible scenarios to analyze.

The size of the bootstrap samples is fixed to $R = 1000$ for all confidence intervals. The BCa method may require a larger number of bootstrap samples (Efron 1987, Section 9)[36] to reduce the simulation error of the bias parameter z_0. We also calculated for this method the confidence intervals with $R = 5000$, we did not find important differences and we only reported results with $R = 1000$.

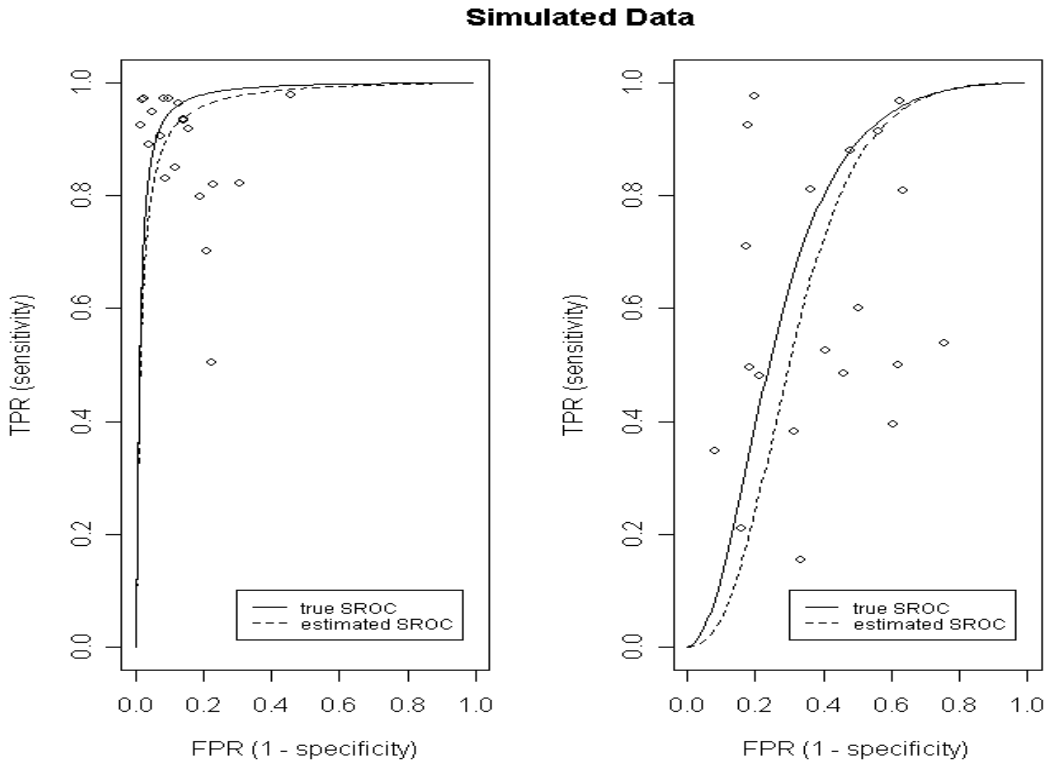

Figure 3.5: *Left panel: simulated meta-analysis with n=20, and parameters of the SROC curve* $A = 5$ *and* $B = 0.1$. *Right panel: simulated meta-analysis with n=20, and parameters of the SROC curve* $A = 1.5$ *and* $B = 0.3$.

3.5.2 Results

Table 3.2 and Table 3.3 summarize the results of our simulation experiment. We start by describing the two sided coverage, the target is that in 1000 different simulated data 95% of the time the true AUC is covered by the interval.

Given that the normal interval is the most commonly applied in statistics we use this interval as a reference method. The coverage of this interval is relatively good across the tables, including the case of $n = 10$ where we expect that the nominal rate could be underestimated. Only for the case $(A, B) = (1.5, 0.3)$ and $n = 10$ the coverage is 91.1%, which underestimate the nominal coverage of 95%.

Both percentile bootstrap methods clearly tend to underestimate the nominal coverage. The BCa does not improve upon the simple percentile method and it did not better than the normal interval. In this case a more sophisticated statistical method has not been justified. This tendency of bootstrap percentile methods to undercover rela-

tively large nominal levels in small samples has been reported in simulation studies for other scalar parameters (see Hall 1992, pag.137; Davison and Hinkley 1997, pag.231 and Canty, Davison and Hinkley 1996).

Both versions of bootstrap-*t* intervals outperform their competitors across the tables, that includes the small sample $n = 10$, where we obtained real coverage very close to 95%. That is a remarkable result!

The normal transformed interval gives also good results. It has the tendency to undercover the proposed nominal level, but outperformed the percentile bootstrap methods in all cases. That is also a very interesting result, this method is computationally very efficient. It computation requires 10% of any percentile method.

Of course coverage in confidence intervals is not the end of the story. We also required that a confidence interval for the AUC gives limits between 0 and 1. Also, we required a notion of *precision* in terms of the interval length. With these respects, the 6 confidence intervals gave very different results in our experiment.

The normal confidence interval has serious problems in the upper bound of the interval in all cases. In particular, for $n = 10$ and $(A, B) = (5, 0.1)$ we have that 85.9% of the time the interval exceeds 1! The bootstrap-*t* interval did not better than the normal one. In fact, for $n = 10$ and $(A, B) = (1.5, 0.3)$ it did worse than the normal interval. By construction, the rest of the intervals respect the range (0,1) of the AUC, and in this regard delivers optimal quality results.

With respect to the length of the interval, the most pronounce result is the large length and the unstable results of the bootstrap-*t* interval. For $n = 10$ it is the worst case and for $n = 20$ it still is the worst with very unstable length. This is explained by the fact that the standard error of the AUC tends to be small and produce numerical unstable results.

The second worst case is the normal interval followed by the percentile bootstrap across the tables. The transformed normal interval did better than the normal and percentile method.

The most sophisticated methods BC_a and bootstrap-t transformed (SDCI), deliver the best results in terms of length and stability.

Clearly, the SDCI is the clear winner of this simulation experiment. It retains the excellent coverage of the bootstrap-*t* interval and adds numerical stability and precision. The construction of a transformation that respects the range of definition of the AUC also makes, that the interval are always bounded to the (0,1) range.

Confidence interval	*(A, B)* AUC	*Lower limit* $AUC < 0$ (%)	*Upper limit* $AUC > 1$ (%)	*Two sided coverage* (%)	*Interval length* $mean \pm sd$
	(5, 0.1) 0.972				
Normal		0.5	85.9	97.5	0.417 ± 0.387
Normal, SDCI		0	0	91.6	0.102 ± 0.117
Bootstrap-t		4.68	54.6	94.3	2.001 ± 9.430
Bootstrap-t, SDCI		0	0	94.0	0.128 ± 0.183
Percentile		0	0	86.9	0.312 ± 0.373
BC_a		0	0	88.1	0.175 ± 0.291
	(3, 0.2) 0.883				
Normal		1.45	63.2	95.4	0.406 ± 0.300
Normal, SDCI		0	0	93.1	0.205 ± 0.123
Bootstrap-t		6.92	47.3	93.9	1.760 ± 6.447
Bootstrap-t, SDCI		0	0	94.2	0.239 ± 0.168
Percentile		0	0	90.0	0.354 ± 0.279
BC_a		0	0	89.0	0.261 ± 0.240
	(1.5, 0.3) 0.726				
Normal		1.25	18.2	91.1	0.336 ± 0.176
Normal, SDCI		0	0	89.9	0.264 ± 0.101
Bootstrap-t		7.8	33.5	95.5	1.083 ± 2.598
Bootstrap-t, SDCI		0	0	94.2	0.349 ± 0.179
Percentile		0	0	90.0	0.316 ± 0.160
BC_a		0	0	89.3	0.305 ± 0.158

Table 3.2: *Results of the simulation experiment. Nominal coverage 95%, number of simulations S=1000, sample size of the meta-analysis n=10, bootstrap sample size R=1000.*

Confidence interval	*(A, B)* AUC	*Lower limit* $AUC < 0$ (%)	*Upper limit* $AUC > 1$ (%)	*Two sided coverage* (%)	*Interval length* $mean \pm sd$
	(5, 0.1) 0.972				
Normal		0	37.9	96.7	0.115 ± 0.165
Normal, SDCI		0	0	93.3	0.039 ± 0.040
Bootstrap-t		0.2	6.6	94.5	0.149 ± 2.360
Bootstrap-t, SDCI		0	0	94.3	0.045 ± 0.052
Percentile		0	0	91.2	0.068 ± 0.157
BC_a		0	0	92.8	0.050 ± 0.119
	(3, 0.2) 0.883				
Normal		0.1	17.9	93.7	0.171 ± 0.164
Normal, SDCI		0	0	92.4	0.113 ± 0.065
Bootstrap-t		0.4	4.8	94.7	0.195 ± 1.132
Bootstrap-t, SDCI		0	0	94.3	0.115 ± 0.055
Percentile		0	0	92.0	0.144 ± 0.165
BC_a		0	0	91.8	0.121 ± 0.142
	(1.5, 0.3) 0.726				
Normal		0.1	2.7	95.1	0.194 ± 0.097
Normal, SDCI		0	0	94.1	0.171 ± 0.056
Bootstrap-t		0.5	5.0	96.4	0.238 ± 0.491
Bootstrap-t, SDCI		0	0	97.1	0.188 ± 0.070
Percentile		0	0	93.5	0.184 ± 0.100
BC_a		0	0	92.5	0.180 ± 0.096

Table 3.3: *Results of the simulation experiment. Nominal coverage 95%, number of simulations S=1000, sample size of the meta-analysis n=20, bootstrap sample size R=1000.*

3.6 Concluding remarks

The area under the summary ROC curve (AUC) has been proposed as a comprehensive summary statistics for meta-analysis of diagnostic test data. It has the appealing interpretation to be the probability that in a pair of disease and non-disease subjects, the disease subject will be classified as more likely to have the disease. The AUC is calculated by numerical integration over the SROC curve in the range of the false positive rate. The available statistical methods for calculating standard errors and confidence intervals are analytically cumbersome and may deliver inaccurate results.

In this chapter we proposed to use bootstrap methods and variance stabilization techniques to improve statistical inference of the AUC. We have reviewed and illustrated several bootstrap techniques and we have proposed a specially designed bootstrap confidence interval for the AUC based on a transformed version of the bootstrap-*t* interval.

We empirically analyzed these methods with an extensive simulation experiment. In this experiment we evaluate nominal coverage, length of the interval and boundary respecting of its limits. The simulation experiment shows that for meta-analysis with small number of studies, standard statistical methods perform poorly, second order accurate bootstrap methods (BC_a and bootstrap-*t*) deliver unstable results and SDCI (transformed bootstrap-*t* and transformed normal) methods are extremely effective.

We conclude that for complex meta-analytic inference a strategy that combines both computer power and careful analytic methods may be an adequate approach to use the bootstrap in practice.

Chapter 4

An Introduction to Bayesian Inferences

"I shall not assume the truth of Bayes' axiom (...) theorems which are useless for scientific purposes."

-Ronald A. Fisher (1935) *The Design of Experiments*, page 6.

4.1 Introduction

There are at least two reasons why Bayesian statistics is particularly important nowadays. One is the ability to combine multiple sources of information in a common synthesis. Bayesian models are particularly well suited for this task, which has a great impact in modern meta-analysis, multi-level models and hierarchical modeling in general. The other reason is the computational revolution produced by the rediscovery of Markov chain Monte Carlo (MCMC) techniques in statistics. As a result, we can construct arbitrary complex statistical models that reflect the complexity for phenomena of interest.

The aim of this chapter is to give a brief overview to Bayesian statistics. We limit the technical aspects of the presentation as much as possible. More technical details are presented in Chapter 5 and Chapter 6.

4.2 Bayes' Theorem and statistical inference

Let us suppose that $y^T = (y_1, y_2, \ldots, y_n)$ is a vector of n observations whose probability distribution $p(y|\theta)$ depends on the values of k *unknown quantities* $\theta^T =$

$(\theta_1, \ldots, \theta_k)$. In classical statistics θ is an unknown fixed quantity, in Bayesian statistics it is an uncertain quantity. This uncertainty about θ is modeled with a probability distribution probability distribution $p(\theta)$. Then,

$$p(y|\theta)p(\theta) = p(y, \theta) = p(\theta|y)p(y). \tag{4.1}$$

Given the observed data y, *Bayes' theorem* says that the conditional distribution of θ is

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}, \tag{4.2}$$

which clearly follows from (4.1). Now $p(y) = c^{-1}$ is just a normalizing constant to ensure that the distribution $p(\theta|y)$ integrates or sums to one. Then, the *Bayes' theorem* is sometimes stated as

$$p(\theta|y) = cp(y|\theta)p(\theta), \tag{4.3}$$

or shortly as

$$p(\theta|y) \propto p(y|\theta)p(\theta), \tag{4.4}$$

where $\propto$ denotes *proportional to*. Expression (4.4) is usually called the *un-normalized posterior distribution*.

The probability distribution $p(\theta)$, which tells us what is known about θ independently of the data y, is called the *prior* distribution of θ, or the distribution of θ *a priory*. The probability distribution $p(\theta|y)$, which tells us what is known about θ given the knowledge of the data, is called the posterior distribution of θ given y, or the distribution of θ *a posteriori*. In this work we sometimes refer to the prior distribution and the posterior distribution simply as the "prior" and the "posterior", respectively.

Given the data y, the probability distribution $p(y|\theta)$ may be regarded *not* as a function of y but of θ, in this way we call it the *likelihood function* of θ. We can clearly see that the Bayes' theorem tells us that the posterior of θ is proportional to the product of the prior and the likelihood, that is,

$$\text{posterior distribution} \propto \text{likelihood} \times \text{prior distribution}.$$

In Chapter 5 and Chapter 6 we will meet models with hierarchical structure where the likelihood function is not well defined, for this reason we prefer to call $p(y|\theta)$ the *data model* distribution function.

Once the Bayesian paradigm is accepted for inference, the posterior distribution $p(\theta|y)$ is used for inference about θ and *no* other concepts are required. For example,

play no role in Bayesian statistics classical concepts like repeated sampling, bias, consistency, sufficiency, etc. Most of the difficulty centers in an adequate specification of the prior distribution $p(\theta)$ and the calculation of $p(\theta|y)$, that could be extremely difficult in practice. The revolutionary use of MCMC techniques and their friendly implementation in several statistical softwares (e.g. WinBUGS, R, SAS, MLWin, etc.) has popularized the use of Bayesian techniques, but the specification of the prior distribution, particularly in multivariate situations, remains a difficult part in this approach.

4.3 Sequential Nature of Bayes' Theorem

Another important result of Bayesian statistics is the sequential nature of the Bayes' Theorem. Suppose that we have an initial sample of observations y_1, then the Bayes' formula gives

$$p(\theta|y_1) \propto p(y_1|\theta)p(\theta). \tag{4.5}$$

Now, suppose that we have a second sample of observations y_2 distributed independently of the first sample, then,

$$\begin{aligned} p(\theta|y_2, y_1) &\propto p(y_2|\theta)p(y_1|\theta)p(\theta) && (4.6)\\ &\propto p(y_2|\theta)p(\theta|y_1). && (4.7) \end{aligned}$$

The last expression shows that the posterior distribution $p(\theta|y_1)$ plays the role of a prior distribution of the second sample y_2. Obviously, this process can be repeated any number of times. Therefore, the Bayes's theorem describes the process of learning from experience, and shows how knowledge about θ is continually modified as new data becomes available.

4.4 Unknown quantities, predictions and model checking

In Bayesian inference the unknown quantity θ is generic, it can denote a vector of parameters, missing data, predictive values, mismeasured data, etc. This approach gives great flexibility at the time of building statistical models that better reflect the complexities encountered in practice. For example, if we are interested in making

predictions, we can write $\theta = (y^{Pred}, \eta)$, after observing y we have

$$p(y^{Pred}|y) = \int p(y^{Pred}, \eta|y)d\eta = \int p(y^{Pred}|\eta, y)p(\eta|y)d\eta \quad (4.8)$$

$$= \int p(y^{Pred}|\eta)p(\eta|y)d\eta. \quad (4.9)$$

The second line clearly shows that the posterior predictive distribution $p(y^{Pred}|y)$ is an average of conditional predictions, $p(y^{Pred}|\eta)$ weighted by the posterior distribution of η. Of course, the integral is calculated over the parameter space of η and it is replaced by sums if the parameter space is discrete.

Posterior predictions play a very important role in model checking assumptions as we will see in Chapter 5 and Chapter 6. The basic point here is, that as usual in practice, we can not guarantee that the model is correct. So in Bayesian modeling we analyze the fitted model by posterior predictive model checking, which consists in simulating data from the fitted model and comparing these quantities with the observed ones. This procedure permits us to understand deficits deficits of the fitted model and correct it by using an updated model. For example, in Chapter 5 we introduce a bivariate normal distribution for random effects in meta-analysis, but in Chapter 6 we will see in that a more elaborated distribution consisting in a mixture of bivariate-*t* distribution will improve the fitness of our model.

The use of simulated data from a model for model checking has a long tradition in data analysis, see for example Bush and Mosteller (1955)[11], and particularly in Bayesian Data Analysis. Posterior predictive assessment was introduced by Guttman (1967)[55], further developments in this are given by Box (1980)[8], applications are given by Rubin (1981)[97], a formalization is given by Rubin (1984)[98]; West (1986)[130] and Gelfand, Dey, and Chang (1992)[45] also present posterior predictive approaches to model evaluation.

For an excellent introduction to Bayesian model checking see Gelman, Carlin , Stern and Rubin (2004)[46] Chapter 6 and Gelman and Hill (2007) Chapter 24 [47].

4.5 Some philosophical aspects

Mathematically, Bayes' formula is a valid statement of conditional probability. What has been in debate for a long time is its applicability for general statistical inference. That is the use of *probabilities* for *inductive inference*. The difficulties are centered in the meaning of probability and in Bayesian statistics the choice of and necessity for

the prior distributions. In this section we concentrate on the first issue, we discuss the issue of prior distributions in the next section.

The two main interpretations of probability are

1. As *frequency distributions*, where there is a physical random mechanism generating the value of θ.

2. As *a subjective measure* of what a particular individual believes about θ.

The first interpretation is uncontroversial and gives an objective application of Bayes theorem. This is the case in some technological applications, for example monitoring the position of an aircraft with Kalman Filtering techniques. In this interpretation of probability we say: *probability of and event*, which indicates that this is *a property of the system* under investigation. However, in most applications, exactly known of objective prior distributions are unusual.

The second interpretation of probability has extensive literature: Ramsay (1931)[92], De Finetti (1937) [21], Savage (1954, 1961a, 1961b) [101, 102, 103] and see Kyburg and Smokler (1964) [72] for a compiled work on subjective probability. In the subjective interpretation of probability, the prior distribution $p(\theta)$ expresses *our* uncertainty about θ without considering the data y and the posterior distribution $p(\theta|y)$ represent an update of *our* knowledge once y is included in the analysis.

Under the subjective interpretation of probability, Bayesian inference implied a *mental construct* where probabilities are used to express our uncertainty concerning the experiment and not a property of the system under investigation. This is why subjectivists use a pedantic: *"probability for an event"*, which indicates that probability is not a property of the system under investigation, but a measure of uncertainty in the relationship between the analyst and the problem at hand. Strictly speaking, we should denote prior probabilities as $p(\theta|H)$, where H indicates the context where the analysis takes place. This context represents the information on which an individual bases his or her own subjective assessment of the *degree of belief*, i.e. probability, of an event occurring (see Spigelhalter, Abrams, and Myles (2004), pag. 11) [108].

In this work we recognize that real data analysis tends to be more clumsy and difficult than a clear philosophical division. In this way we follow a *hybrid position* where depending on the analytical context one or another interpretation of probability is more appropriate. For example, when we build a statistical model to handle meta-analysis of diagnostic data we adopt a clear subjectivistic Bayesian position, but when we perform model checking we adopt a frequentist point of view.

We follow this *pragmatic* and *eclectic* practice with an important dose of *transparency* in the analysis, which include: formal description of the statistical model used, elicitation of priors, description of the algorithms, detailed numerical aspects, sensitivity analysis, etc.

4.6 Exchangeability

As before suppose that $y^T = (y_1, y_2, \ldots, y_n)$ is a vector of n observations whose probability distribution is $p(y|\theta)$. *Exchangeability* is a formal expression of the idea that *we* find no systematic reason to distinguish among the individual variables $Y_1, \ldots, Y_n$ that have produced the outcome y. We say that these variables are similar but not identical. Exchangeability represents a component of *our knowledge* of the data. For example, if we know that a data set results from a mixture of studies with different study design, it would not be reasonable consider these data as exchangeable. Although, numerical results may be similar among studies, the quality of the data is different if a study was performed with a retrospective or a prospective design.

Technically exchangeability implied invariance of $p(y|\theta)$ under permutation of the indexes. Note that we do *not* mean that $Y_1, \ldots, Y_n$ are independent and identically distributed (i.i.d) with a distribution function $p(y|\theta)$, this will mean that $p(y|\theta) = p(y_1|\theta) \ldots p(y_n|\theta)$.

4.7 Priors

Who can deny that there are substantial quantifiable prior beliefs in medicine and other scientific areas? Anyone who has been involved in planing a clinical trial or other experiment knows that there is a clear body of information before an experiment is performed.

To be more precise, in almost all data analysis there are substantial amounts of information, that are external to the data that we wish to analyze. The Bayesian approach formalizes a procedure to combine different sources of information. However, the main issue is how to translate this body of knowledge in a formal distribution function.

Given that Bayesian data analysis is driven by priors and it is worth pointing out some misunderstanding regarding prior distributions, we follow here some points mentioned by Spigelhalter, Abrams, and Myles (2004), pag. 73 [108] and others that are the product of our own experience.

- The name *prior* suggests a temporal relationship, however, this is misleading. The prior distribution models the uncertainty given by the *external evidence*.

> *I was surprised to read that priors must be chosen before the data have been seen. Nothing in the formalism demands this. Prior does not refer to time, but to a situation, hypothetical when we have data, where we assess what our evidence would have been if we had had no data. This assessment may rationally be affected by having seen the data, although there are considerable dangers in this, rather similar to those in frequentist theory.*
> *Cox (1999)[16]*

- The prior is not necessarily unique! In a recent article Lambert et. al. (2005)[73] analyze the use of 13 different priors for the between study variance parameter in random-effects meta-analysis. There is no such thing as *the 'correct' prior*. Bayesian analysis is regarded as transforming prior into posterior opinion, rather than producing *'the' posterior distribution*. For different formal constructions of priors see Kass and Wasserman [69].

- The prior may not be completely specified. In hierarchical models, as we will see in Chapter 5 and Chapter 6, the priors have unknown parameters that have to be estimated.

- Priors can be overparametrized. Sometimes we intentionally overparametrized the priors in order to accelerate convergence of simulation methods, see Gelman, Carlin, Stern and Rubin (2004)[46] Chapter 6 and Gelman and Hill (2007) Chapter 24 [47]. We follow this approach in Chapter 6 to build a prior for the degrees of freedom in a multivariate-t distribution.

- Inference may rely *only* on priors distributions. There are situations where no further data are available to combine with our priors or there is no intention to update the priors. This is the typical case of *risk analysis*, *sample size determination in experiments*, *simulation of complex process* , etc. In these analytical scenarios priors are usually used to simulate hypothetical data and we refer to that *prior predictive analysis*, prior because it is not conditional on observations of the process, predictive because it is the distribution for a quantity that is observable.

- Finally, priors are not necessarily important. In many scientific applications, as the amount of data increases, the prior is overwhelmed by the likelihood and the influence of the prior disappears, see Box and Tiao (1973) (pag. 20-25)[7].

4.8 Modern Bayesian Data Analysis

The statistical model building presented in Chapter 5 and Chapter 6 is based on Modern Bayesian Data Analysis. What we mean by that is the coherent combination of the following statistical and probabilistic techniques:

- The use of hierarchical models to reflect multiple sources of uncertainty in the data analysis.
- The extensive use of MCMC method as a simulation based approach to calculate marginal posterior distributions of quantities of interest. For a general introduction to this topic see Brooks (1998)[5].
- The systematic use of graphical models to give a schematic description of model quantities and their interrelation. This technique is usually combined with automatic algebraic algorithm to factorize complex joint distributions into its conditional marginal distributions. This automatic factorization simplifies the use of MCMC techniques; in particular the application of Gibbs sampling (see Gilks et.al. (1993)[52] and Gilks et.al. (1996)[53]).
- The use of systematic model checking and model diagnostic techniques.
- The ability of using classical and Bayesian techniques when appropriate.

Many of these ideas have been discussed in a paper by Spiegelhalter (1998)[106]. In general bringing these ideas together has substantially changed the way that we make Bayesian statistics. We give technical details about the previous points in our concrete work in Chapter 5 and Chapter 6.

Chapter 5

A Bayesian model for combining diagnostic test data

"Beware: MCMC sampling can be dangerous!"

-David Spiegelhalter, Andrew Thomas, Nicky Best and Dave Lunn
WinBUGS User Manual, January 2003

5.1 Introduction

In Chapter 1 we highlighted some limitations of current statistical methodologies to analyze and combine diagnostic test results. In this chapter we start to develop a statistical model that can tackle most of the current issues. Although this model does not reflect all data complexities encountered in practice, it sets up the mechanic of model fitting and model validation that makes it possible to understand how the model will be extended in practice. These model extensions are presented in Chapter 6.

In this chapter, we present a hierarchical Bayesian model for combining diagnostic test data. Model building starts by specifying a *data model*, $p(y|\theta)$, where $y = (y_1, \ldots, y_N)$ denotes the available data and $\theta = (\theta_1, \ldots, \theta_N)$ all study-specific unknown quantities (e.g. random effects). The set of study-specific parameters θ_i are modeled by the structural distribution $p(\theta_i|\phi)$. Uncertainty about ϕ is modeled by a prior distribution $p(\phi)$. Inferential statements about θ and ϕ are based on the posterior probabilities

$$p(\theta, \phi|y) \propto p(\phi)p(\theta|\phi)p(y|\theta). \tag{5.1}$$

Model checking is performed by simulating unknown quantities from (5.1), these simulations are compared with observed values by visual inspection and more formally

calculating model deviation quantities.

5.2 A Bayesian model

5.2.1 Data model

Following the notation of Chapter 1, let tp_i and fp_i be the true positive and false positive results for study i ($i = 1, \ldots, N$). Conditioning on $n_{i,1}$, which is the number of disease patients, and on $n_{i,2}$ the number of non-disease patients, our data model is a binomial distribution with

$$tp_i \sim \texttt{Bin}(\text{TPR}_i, n_{i,1}), \quad fp_i \sim \texttt{Bin}(\text{FPR}_i, n_{i,2}), \tag{5.2}$$

where TPR_i and FPR_i are the probabilities to observe a positive test result in the disease and non-disease population respectively.

The N pairs of probabilities TPR_i and FPR_i are transformed by a link function g to a scale where they are defined in the range $(-\infty, \infty)$. The canonical link function for binomial data is the logistic link function, but other alternative links, e.g. the complementary log-log (cloglog) link function can be used. Choosing a suitable link function for the data at hand is a delicate modeling problem that will be analyzed in this Chapter.

5.2.2 Structural distribution

We model the variability between studies with a bivariate normal distribution on the differences

$$D_i = g(\text{TPR}_i) - g(\text{FPR}_i)$$

and the sums

$$S_i = g(\text{TPR}_i) + g(\text{FPR}_i)$$

with

$$(D_i, S_i) \sim \texttt{Normal}_2(\mu, \Lambda), \quad i = 1, 2, \ldots, N. \tag{5.3}$$

Where μ is the mean value of (D_i, S_i) and Λ their precision matrix, i.e., $\Lambda = \Sigma^{-1}$ with Σ the variance covariance matrix. This parametrization is convenient in a Bayesian setup.

Clearly, when $g(\cdot)$ is the logistic link function the random effects D_i and S_i can be related with the classical SROC curve. Therefore D_i is the study effect related with diagnostic discriminatory power and S_i is the effect associated with diagnostic threshold

value. Modeling (D_i, S_i) is similar to direct modeling $(g(\texttt{TPR}_i), g(\texttt{FPR}_i))$, however, the linear transformation should leave (D_i, S_i) roughly independent making our inference less sensitive to the prior distribution of Λ. A similar approach is presented by Gelman et al.(2004, pag. 488-491) [46] in another bivariate meta-analysis.

The exchangeability assumption between studies is clearly unsuitable for our data. Study labels contain substantial information that should be included in our model. We address this problem in detail in Chapter 6.

5.2.3 Prior distributions

Our modeling approach is full Bayesian and we need to specify prior distributions on model parameters. We assume independent Normal prior distributions for the components of $\mu = (\mu_D, \mu_S)^T$

$$\mu_D \sim N(\texttt{m}_D, \texttt{v}_D), \quad \mu_S \sim N(\texttt{m}_S, \texttt{v}_S).$$

For the precision matrix Λ we give a prior Wishart distribution with scale matrix R and k degrees of freedom:

$$\Lambda \sim \texttt{Wishart}(R, k).$$

The constants $\texttt{m}_D$, $\texttt{v}_S$, $\texttt{m}_S$, $\texttt{v}_S$, R and k are known. In our data analysis, we take $\texttt{m}_D = \texttt{m}_S = 0$, $\texttt{v}_D = \texttt{v}_S = 0.25$, $R = \texttt{Diag}(1, 1)$ and $k = 3$. We choose these constants in a way that the data model dominates the inference as much as possible. They can also be used for prior elicitation and sensitivity analysis. Table 5.1 summarizes the notation involved in our starting model.

5.2.4 Posterior distribution

Given this model specification all inference is based on the posterior distribution

$$p(\theta, \phi|y) \propto \prod_{i=1}^{N} \left[\binom{n_{i,1}}{tp_i} \text{TPR}_i^{tp_i} (1 - \text{TPR}_i)^{(n_{i,1} - tp_i)} \binom{n_{i,2}}{fp_i} \text{FPR}_i^{fp_i} (1 - \text{FPR}_i)^{(n_{i,2} - fp_i)} \right] \tag{5.4}$$

$$\times \prod_{i=1}^{N} \left\{ \exp\left[-1/2 \, (D_i - \mu_D, S_i - \mu_S)^T \Lambda (D_i - \mu_D, S_i - \mu_S) \right] \right\} \times |\Lambda|^{\frac{N}{2}}$$

$$\times \exp\left[-1/2 \, (v_D \, (\mu_D - m_D)^2 + v_S \, (\mu_S - m_S)^2) \right]$$

$$\times \frac{|\Lambda|^{(N-k-2)/2} \exp\left[-1/2 \, \text{trace}(\Lambda R^{-1}) \right]}{2^{k(N-1)/2} \pi^{k(k-1)/4} |\Lambda|^{(N-1)/2} \prod_{i=1}^{N} \Gamma(1/2(N - i))},$$

Notation	Parameter
tp_i	Frequency of true positive patients
fp_i	Frequency of false positive patients
$n_{i,1}$	Total number of disease patients in the study
$n_{i,2}$	Total number of non disease patients in the study
TPR_i	True positive rate of study i
FPR_i	False positive rate of study i
D_i	link function differences of TPR and FPR
S_i	link function sum of TPR and FPR
μ_D	Mean of D_i
μ_S	Mean of S_i
Λ	The precision matrix of (D_i, S_i)
σ_D^2	Variance of D_i
σ_S^2	Variance of S_i
$\sigma_{D,S}$	Covariance (D_i, S_i)

Table 5.1: *Notation and parameter names for the basic bivariate hierarchical model.*

where (θ, ϕ) is dimension $L = N \times 2 + 2 + 3$ and contains all random effects and all components of μ and Λ.

5.2.5 Summary quantities of interest

In general, we are interested in making inference of particular components of (θ, ϕ), or in functional parameters of components of (θ, ϕ), say $h(\theta, \phi)$. For example, to summarize results *at the group level* we are interested in the posterior distribution of

$$\text{TPR} = g^{-1}[(\mu_D + \mu_S)/2], \quad \text{FPR} = g^{-1}[(\mu_D - \mu_S)/2], \tag{5.5}$$

we called (5.5) *pooled summaries*.

At the study level, we are interested in the marginal posterior distributions

$$p(\text{TPR}_i|y) \quad \text{and} \quad p(\text{TNR}_i|y), \tag{5.6}$$

we call (5.6) the *study summaries*.

Another important summary is the predicted pair of rates (FPR, TPR), for a study that has not been included in the review. In such a case we base inference on the *jointly predictive posterior distribution*

$$p(\text{FPR}^{pred}, \text{TPR}^{pred}|y). \tag{5.7}$$

Based on (5.7), we can graphically report the predictive surface for the pairs (FPR, TPR) at a given credibility level (e.g. 95%), we called this summary the Bayesian predictive surface (BPS). Clearly, in this model framework we can calculate the marginal predictive posteriors $p(\text{FPR}^{pred})$ and $p(\text{TPR}^{pred}|y)$ as well.

5.2.6 Summary ROC curve and the AUC

We can easily recover a Bayesian version of the SROC curve by applying standard results of the bivariate normal distribution. We have that the conditional distribution of $(D_i|S_i = s_i)$ has mean

$$E(D_i|S_i = s_i) = E(\text{A}) + E(\text{B})\,(s_i - E(\mu_S)), \tag{5.8}$$

where the functional parameters A and B are

$$\text{A} = \mu_D + \mu_S\text{B}, \quad \text{B} = \sigma_{D,S}/\sigma_S^2. \tag{5.9}$$

We define the Bayesian SROC curve (BSROC) by transforming back results from (S, D) to (FPR, TPR) with

$$\text{TPR} = \text{BSROC}(\text{FPR}) = \text{g}^{-1}\left[\frac{A}{(1-B)} + \frac{B+1}{(1-B)}\,\text{g}(\text{FPR})\right]. \tag{5.10}$$

As in the classical approach the SROC curve is obtained by calculating TPR in a grid of values of FPR. It is important to highlight, that there is no estimation of variability associated to the SROC curve and to the HSROC curve. Our definition of the BSROC implies a posterior for each value of FPR, therefore it is straightforward to give credibility intervals for the BSROC for each value of FPR.

Finally, we can define a Bayesian area under the SROC curve (BAUC) by integrating numerically the BSROC for all possible values of the false positive rate:

$$\text{BAUC} = \int_0^1 \text{BSROC}(x)\,dx. \tag{5.11}$$

As in the classical approach the BAUC has the appealing interpretation to be the probability that in a pair of disease and non-disease subjects, the disease subject will be classified as more likely to have the disease.

5.2.7 MCMC computations

All of these marginal posteriors and predictive distributions presented in Section 5.2.5 and Section 5.2.6 are not analytically tractable, we based our inference on MCMC

techniques implemented in the BUGS language (Lunn, Thomas, Best and Spiegelhalter; 2000)

BUGS stands for *Bayesian* analysis *Using Gibbs Sampling*, a name that reflects the computational methods originally implemented. It is interesting, that the BUGS project is prior to the seminal paper of Gibbs sampling in Bayesian inference by Gelfand and Smith (1990) [44] and has its roots in modeling complex decision processes with artificial intelligence systems during the 80's (Spiegelhalter, 2006) [107]. This origin has influenced the architecture of the system, which clearly separates *knowledge-base* (e.g. use of declarative programming, graphical modeling, etc.) and *inference-engine* (i.e. draw conclusions in a specific situation). In other words, the philosophy behind `BUGS` is: first build a probabilistic model for the problem at hand, then use the observed data to update the model. If no data is observed, the model can still be used for *prior predictive inference*.

In BUGS we assume that the joint posterior distribution can be represented by a directed acyclic graph (DAG). Figure 5.1 displays the DAG for our basic model. A DAG describes all model components as *nodes* and their relationships by *directed links* between them. Oval nodes represent stochastic quantities and square ones known parameters. *Single arrows* define probability distributions, e.g. $tp_i \sim \texttt{Bin}(\text{TPR}_i, n_{1,i})$, while *double arrows* represent logical or functional relationships, e.g. $\text{TPR}_i = g^{-1}((D_i + S_i)/2)$. The big square containing several nodes and arrows is called a *plate*, which represents replicated sub-model structures over an index, e.g. $i = 1, \ldots, N$.

A DAG defines a set of *parent-child* relationships between stochastic nodes, we say that a node v_k is a parent of a node v_l if an arrow points from v_k to v_l. For example, μ and Λ are parents of (D_i, S_i). This parent-child relationship enables us to factorize the joint probability distribution of the model as a product of the conditional distributions of each node given their parents (Lauritzen, Dawid, Larsen and Leimer, 1990)[75]:

$$p(v_1, ..., v_L) = p(\theta, \phi|y) \propto p(\theta, \phi, y) = \prod_{k=1}^{L} p(v_k|\text{parents}[v_k]). \tag{5.12}$$

This factorization has two practical advantages: first it allows us to build arbitrary complex models by defining their stochastic local structure (*parent-child relationships*) and second it makes the identification of all conditional distributions automatic. Let $v_{(-k)} = (v_1, \ldots, v_{k-1}, v_{k+1}, \ldots, v_L)$ denote all nodes in the DAG except v_k then the full conditional $p(v_k|v_{-k})$ is proportional to the product of the terms in $p(v_1, ..., v_L)$

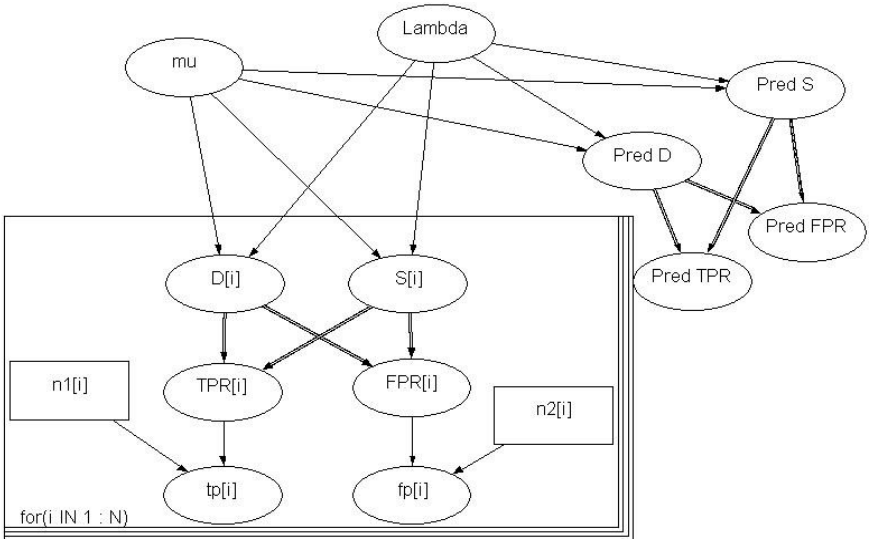

Figure 5.1: *Directed Acyclic Graph (DAG) of the bivariate structural model for meta-analysis of diagnostic test data.*

that contains v_k:

$$p(v_k|v_{-k}) \propto p(v_k|\text{parents}[v_k]) \times \prod_{j \in \text{children}[v_k]} p(v_j|\text{parents}[v_j]). \tag{5.13}$$

The first term $p(v_k|\text{parents}[v_k])$ is called *the prior component* and the second one is called *the data component* or *likelihood component*. These conditional distributions define the kernel of the Gibbs sampler of the model.

Once the model has been specified, BUGS scans each node of the DAG and builds each $p(v_k|v_{-k})$, then the system decides how to sample from each distribution by a hierarchical decision process: if the system detects a conjugate distribution it samples directly by standard algorithms (Ripley, 1987) [95], for non-conjugate problems uses adaptive rejection sampling with log-concave densities (Gilks and Wild, 1992) [51]. If log-concavity is not identified the system uses slice sampling for restricted range supports (Neal, 2003) [88] and Metropolis-Hastings algorithm with a normal proposal distribution otherwise (Metropolis, et al. 1953 [86], Hastings, (1970) [64]). In general, we will denote a sampling value of a node v_k by v_k^*.

This Bayesian graphical model approach is not only a friendly way to perform complex Bayesian computations but a semantic to represent further model structures. Figure 5.2 represent the calculations involved in the posterior distribution of the BSROC (5.10) and the BAUC (5.11). We define two logical nodes pointing from μ and Λ to

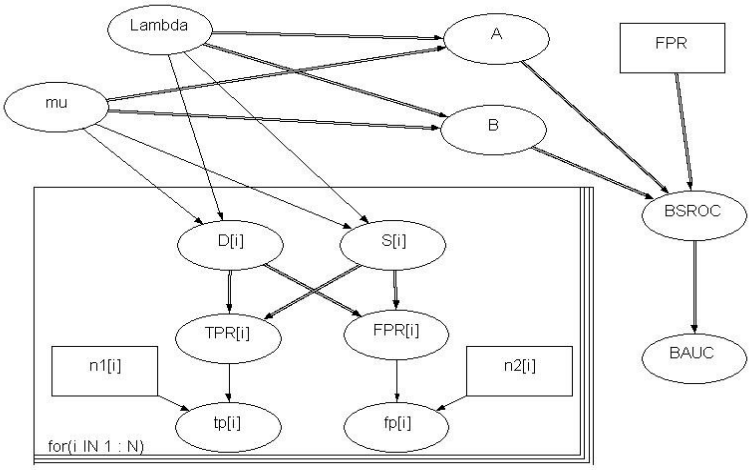

Figure 5.2: *Directed Acyclic Graph (DAG) of the BSROC and BAUC under a bivariate structural model for meta-analysis of diagnostic test data.*

A and B, this represents the mathematical relationship between the mean and the precision matrix and the regression coefficients of equation (5.8). Then a logical node called BSROC is defined as a function of A, B and a constant node FPR. Finally a functional relationship is represented by logical node BAUC.

This idea of developing complex models by DAGs representation will be further extended in Chapter 6.

5.2.8 Assessing convergence of MCMC simulations

The output of a simulated MCMC should be treated with care. The sophisticated Gibbs sampler implemented in BUGS and in general MCMC methods satisfied only sufficient conditions for convergence. In general we need to assess convergence of a simulated Markov chain by empirical methods, which may include a subjective decision of the data analyst.

In this work we base the assessment of convergence of MCMC simulations with three different approaches:

1. Visual analysis via trace plots and autocorrelation functions (ACF). Trace plots are the simulated samples versus the simulation index, these time series apport substantial information about convergence. A trace can tell us if the chain has

converged to its stationary distribution or if it needs a longer burn-in period. Together the trace and the ACF gives us information about stationarity and tells us if a chain mixed well, i.e. its excursions run rapidly across the posterior space. In our experience a quick look at these graphical tools is more valuable than statistical tests designed to assess convergence. At the end of this chapter we will see clear examples where a model with bad parametrization delivers convergence problems.

2. We report the Brooks-Gelman-Rubin (B-G-R) diagnostic (Gelman and Rubin, 1992 [48] and Brooks and Gelman 1997 [9]). This is a test based on analyzing multiple simulated MCMC chains by comparing the variance within each chain and the variance between chains. Large deviation between these two variances indicates convergence problems. In our analysis we run 3 independent chains, starting with random initial values.

3. A measure of mixing is the effective sample size (*ESS*) (Kass et al. 1998 [68]). This number is defined as:

$$ESS = \frac{M}{1 + 2\sum_{k=1}^{L} \hat{\rho}(k)}, \tag{5.14}$$

where M is the length of the simulated chain and $\hat{\rho}(k)$ is the empirical autocorrelation function of the chain. The value of L is taken as a lag where $\hat{\rho}(k) < 0.05$. A low value of *ESS* indicates bad mixing of the Markov chain.

We use the `coda` (Plummer et al., 2007) [91] package in R to analyze convergence .

In our experience, a serious data analysis work needs a combination of methods to assess convergence and mixing. At the end of the chapter we report a case, where the trace plots of a MCMC output clearly indicate non-stationarity and the B-G-R test shows convergence, but the ESS points out bad mixing. In our work we declare convergence if all three methods point out convergence for all parameters in the model.

Classical references about convergence issues in MCMC methods are Cowles and Carlin (1996) [17] and Brooks and Roberts (1998) [10].

5.2.9 Using DIC for model selection

The Deviance Information Criterion (DIC) has been introduced by Spiegelhater et al. (2002) [109] as a Bayesian model selection tool. It is based on trading off between goodness of fit and model complexity:

DIC = measure of fit + complexity penalty

where the measure of goodness of fit is based on the deviance statistic:

$$D(\theta) = -2 \log L(y|\theta) \tag{5.15}$$

and complexity of the model via:

$$\begin{aligned} p_D &= E_{\theta|y}[D] - D\left(E_{\theta|y}[\theta]\right) && (5.16) \\ &= \bar{D} - D(\bar{\theta}) && (5.17) \end{aligned}$$

i.e. posterior mean deviance minus deviance evaluated at the posterior mean of the parameters. p_D should be approximately the true number of parameters. The DIC is defined similarly to AIC (Akaike's information criterion) as

$$DIC = D(\bar{\theta}) + 2 \times p_D = \bar{D} + p_D. \tag{5.18}$$

Models with smaller DIC are better supported by the data, in the sense of short-term predictions like the AIC or the predictive error by cross-validation. As a practical rule, a difference in DIC less than 5 is not considered as an important model improvement. Therefore, just reporting a model with the lowest DIC can be misleading (Spiegelhalter, et al. 2004) [110].

The DIC is trivially calculated from the output of the MCMC. It does not require maximization over the parameter space, like the AIC and BIC (Bayesian Information Criterion). Another advantage is that it can be used to compare non-nested models, models with different hierarchical structures and models which involve dependent data structure. Moreover, the DIC does not require the existence of a "true model" like AIC and BIC (see Ripley, 2004, pag. 159-161) [96].

DIC is based on posterior densities, i.e. it takes prior information into consideration. In this way if a model is based on informative priors it may have a less DIC value than a model based on non-informative priors. In general, in this work we use DIC as a model assessment tool together with careful model diagnostic approaches.

5.3 Data analysis

In this section we analyze the data presented in Chapter 1 and Chapter 2 with the techniques presented in the previous sections. We applied the bivariate random effect model with vague prior distributions and two different link functions: logistic and

Link function	Parameter	Mean	2.5%	50%	97.5%
cloglog	σ_D^2	1.066	0.581	1.016	1.820
	σ_S^2	1.118	0.606	1.067	1.899
	$\sigma_{D,S}$	-1.044	-1.788	-0.996	-0.564
	Sensitivity (pooled)	0.955	0.945	0.955	0.965
	Specificity (pooled)	0.951	0.934	0.952	0.966
	Sensitivity (predicted)	0.951	0.897	0.955	0.986
	Specificity (predicted)	0.924	0.686	0.952	0.994
logistic	σ_D^2	1.147	0.551	1.092	2.070
	σ_S^2	1.388	0.700	1.321	2.476
	$\sigma_{D,S}$	-1.071	-1.890	-1.027	-0.488
	Sensitivity (pooled)	0.955	0.946	0.955	0.963
	Specificity (pooled)	0.950	0.931	0.951	0.966
	Sensitivity (predicted)	0.953	0.915	0.955	0.977
	Specificity (predicted)	0.925	0.693	0.952	0.994

Table 5.2: *Summary results of two fitted models. One model with complementary loglog link and the other with logistic link. Posterior distributions are based on a single chain of length 20,000 with the first 10,000 iterations discarded.*

cloglog. Calculations are based on 3 chains with random starting values and with 20,000 replications. The last 10,000 iterations are used for analysis. We used these 3 chains to assess convergence. Graphical and numerical summaries are based on a single chain of length 10,000. `BUGS/R` scripts and more details of these calculations are in Chapter 7.

Table 5.2 summarizes the numerical results for these models. The variance covariance matrix, in both models, indicates that there is substantial heterogeneity between studies in both components D_i and S_i, with negative correlation as is expected.

Results are very similar in both models, but the model with logistic link function gives larger values for σ_D^2 and σ_S^2 and narrow confidence limits for sensitivity. Figure 5.3 shows this result more clearly. The tails of the posterior distribution for sensitivity summaries (pooled and predicted) are much lighter for the logistic link than for the cloglog link function. The question is: which model should we use for inference?

To answer this question, we picked 200 simulated values $(\text{TPR}^*, \text{TNR}^*)$ from $p(\text{TPR}, \text{TNR}|y)$ and we compare these pairs with $(\widehat{\text{TPR}}_i, \widehat{\text{TPR}}_i)$. Figure 5.4 shows the resulting scatter plots, the left panel corresponds to the model with cloglog link and the right panel to the model with logistic link function.

The scatter of the predicted values are quite similar for both link functions. The

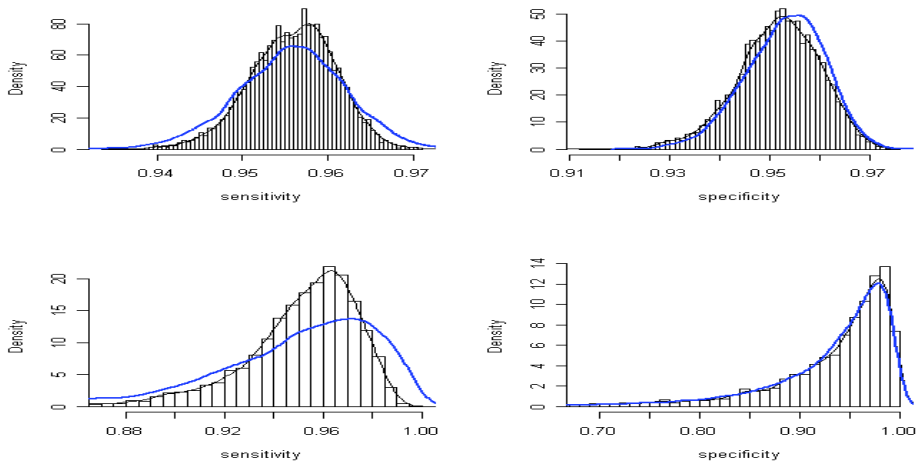

Figure 5.3: *Summary results for sensitivity and specificity. Upper panels correspond to posterior distributions for pooled sensitivity and specificity. Lower panels show predictive posteriors for a study not included in the review. Smoothed histograms correspond to the model with logistic link function and bold line to the posterior densities based on cloglog link function.*

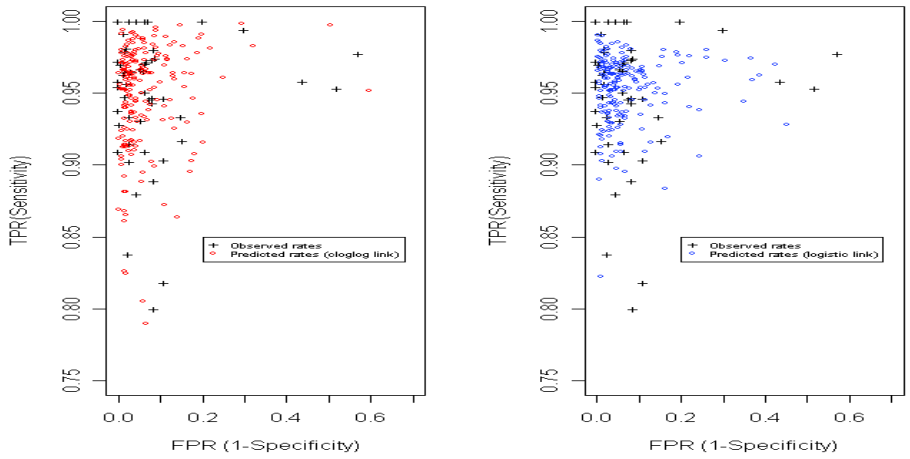

Figure 5.4: *Scatter plots of pairs (FPR, TPR). Crosses denote* $(\widehat{TPR}_i, \widehat{TPR}_i)$*, circles simulated values form* $p(TPR, FPR|y)$*. Left panel: Results for the model with cloglog link function. Right panel: Results for the model with logistic link function.*

DIC for the model with logistic link is 426.5, while the DIC for the model with cloglog link is 425.1. We found that the last model slightly better fits the data, but not substantially. Given that the logistic link is easily interpretable (e.g. as diagnostic odds rations, etc.) we choose this model for further analysis.

We pay particular attention to this posterior model checking, because the logistic link function is applied commonly to binary data, but it may fit the data poorly. In some circumstances, it may be more appropriate to model our uncertainty about the link function, for example by using a mixture of link functions [74, 82].

Another model assumption that we should check is the normality of the random effects. We stress the importance of this type of model checking in meta-analysis, where the normal distribution for random effects is commonly applied without further analysis. Figure 5.5 shows the qq-normal plots for standardized study effects D_i^* and S_i^* and the qq-plot of the distances

$$d_i^* = r_i^{*T} \Lambda^* r_i^*,$$

where $r_i^* = (D_i^* - \mu_D^*, S_i^* - \mu_S^*)$, which are compared to a χ^2 distribution with 2 degrees of freedom. We see that there are clear deviations to the normal distribution and the random effects of some studies are inconsistent with the rest of the studies. In Chapter 6 we model these deviations by incorporating study design information and in general by using a bivariate *t*-distribution for random effects.

Finally, Figure 5.6 presents the trace plots for μ_1, μ_2, and the components of Σ, these traces corresponds to the last 1000 simulated values of three chains with randomly starting values. We can see that convergence is very stable for all parameters in the model. In Chapter 7 we give more details on convergence checking.

BSROC, BAUC and BPS

The left panel of Figure 5.7 presents the BSROC with the 2.5%, 50% and 97.5% percentiles of its posterior distribution at each value of FPR, with wider posteriors for larger values of FPR.

The posterior mean and percentiles of 2.5% and 97.5% of the posterior distribution of the coefficient A is 6.112 (5.724, 6.513) and for B is -0.632 (-0.986 , -0.238), while for the classical SROC curve we have 5.735 (5.384,6.086) and -0.298 (-0.535, -0.06) respectively. The coefficient A is interpreted as a pooled diagnostic odds ratio in the logistic scale, the classical SROC underestimate the diagnostic accuracy of the CT technique in comparison to the BSROC. Both methods give different results with re-

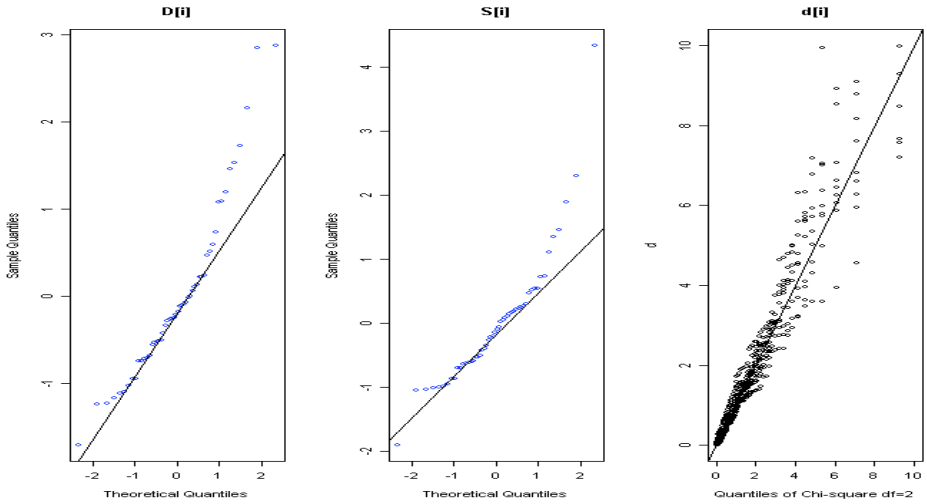

Figure 5.5: *Diagnostic plots for multivariate normality of random effects.*

spect to B, in this case the classical SROC underestimates the effect of the heterogeneity presented in these data. These two results are expected, the bias of the SROC is a well known problem and most of the classical critics of the SROC curve reviewed in Chapter 1 are resolved by the BSROC.

The right panel of Figure 5.7 shows the posterior distribution of the BAUC. The posterior distribution of the BAUC has percentiles of 2.5%, 50% and 97.5% equal to 0.955, 0.975 and 0.987 respectively and mean 0.974. These results are slightly different from the estimated AUC calculated over the SROC which gives $\widehat{\text{AUC}} = 0.981$ and a bootstrap BC_a confidence interval (0.969, 0.987). The reason why the AUC calculated over the SROC has higher value than the posterior mean of the BAUC is explained in Figure 5.7 where we can see that the SROC artificially dominates the BSROC for larger values of FPR, this effect is caused by the underestimation of B. It is very interesting, that the upper limit of the confidence interval based on the BCa gives exactly the same numerical value as the 97.5% percentile of the posterior distribution of BAUC, the theory developed by Efron (1993)[38] showed that the BC_a confidence intervals can be transformed to posterior distribution for a parameter of interest. However, the lower confidence limit of the BC_a underestimate the dispersion of these data.

The ability to predict results of a hypothetical study, that is not included in the systematic review, has fundamental importance in meta-analysis. In the previous section in Table 5.2 and Figure 5.3 we presented the marginal posteriors of sensitivity and

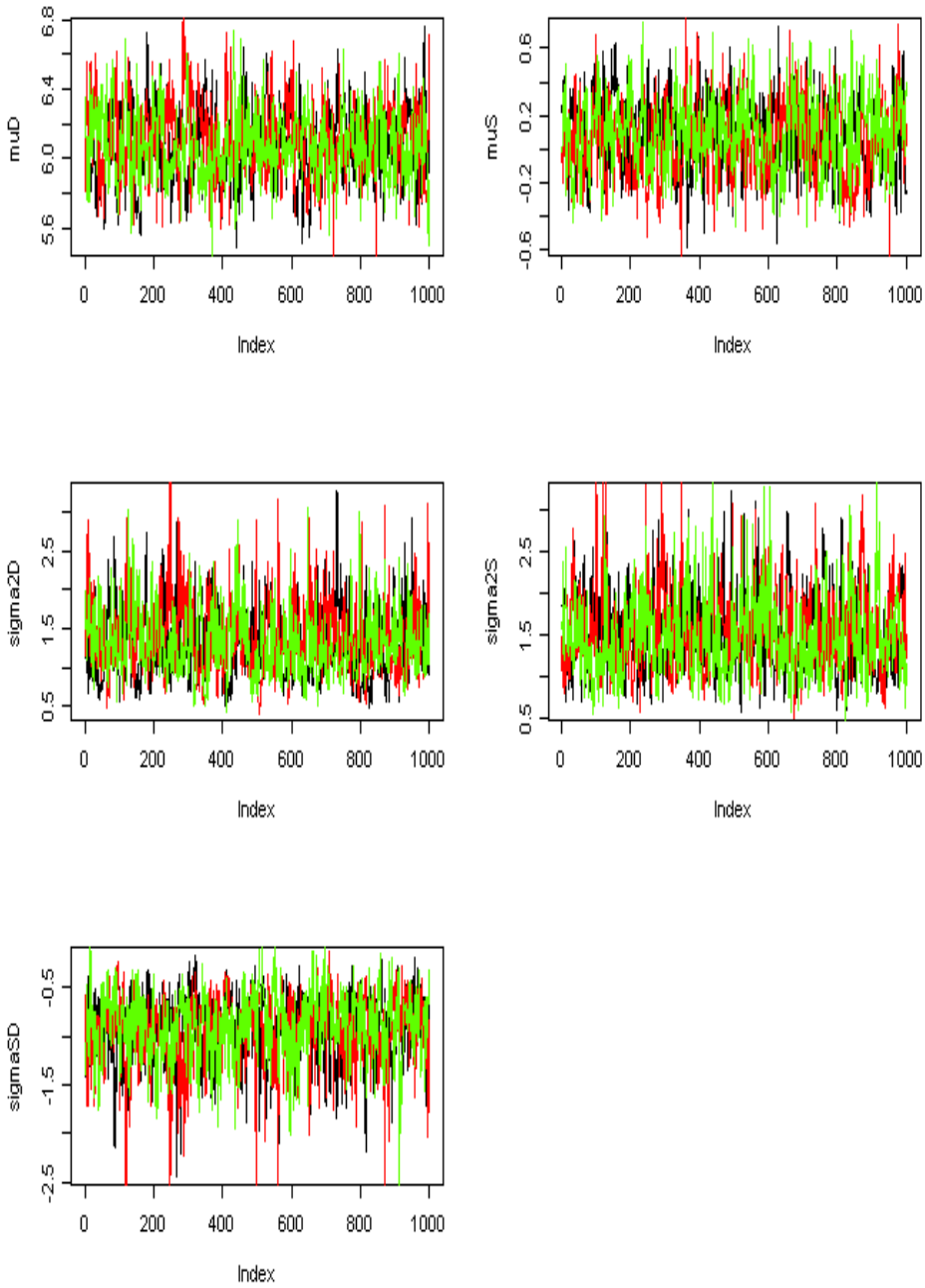

Figure 5.6: *Trace plots for model parameters in the multivariate normality of random effects.*

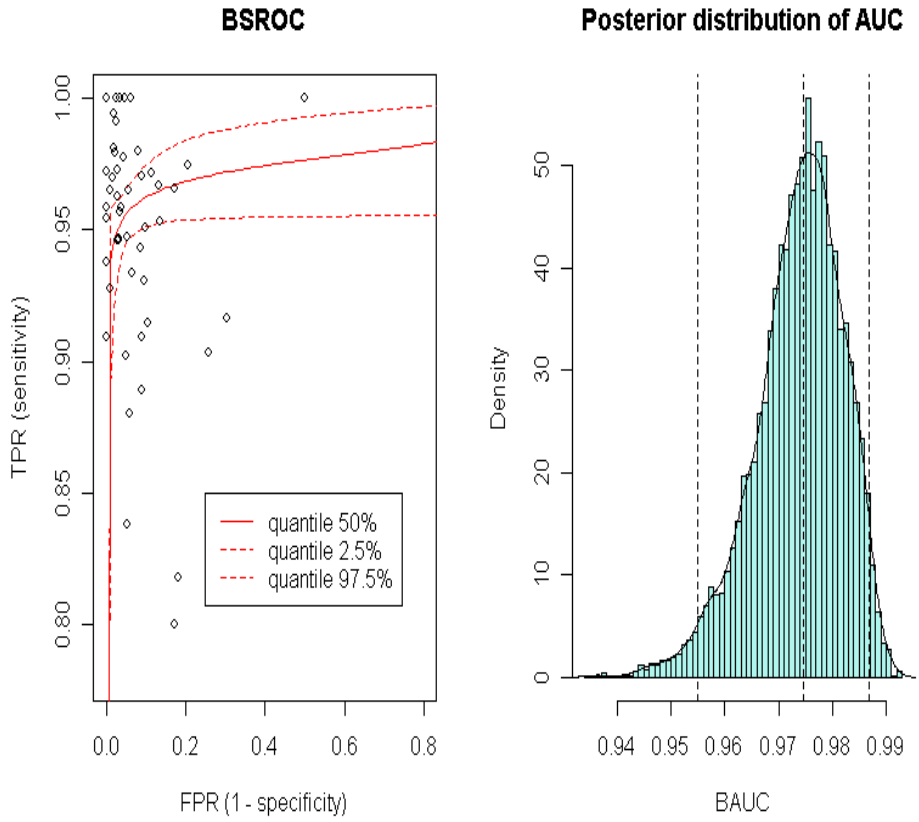

Figure 5.7: *Left panel: crude estimates of the pairs (FPR, TPR) and the BSROC. Right panel: posterior distribution of the Bayesian Area Under the Summary ROC Curve.*

specificity. Figure 5.8 shows the BPS which is based on the jointly posterior distribution of (FPR, TPR). The BPS covers a greater region for FPR than TPR and excludes studies which reported 100% sensitivity or 100 % specificity, the same applies for the credibility intervals of the BSROC. One possible use of these predictive regions is to look along the line where sensitivity = specificity and pick the two interception points. These two points give the better and the worst case studies with a credibility of 95%. For our data, we have that the better predicted case has a combination of sensitivity and specificity of (0.985, 0.985) and the worst case has (0.878, 0.878).

All computations performed in this section took approximately three minutes, that includes the simulation of three independent Markov chains on length 20,000 to assess convergence and the model checking and graphical outputs.

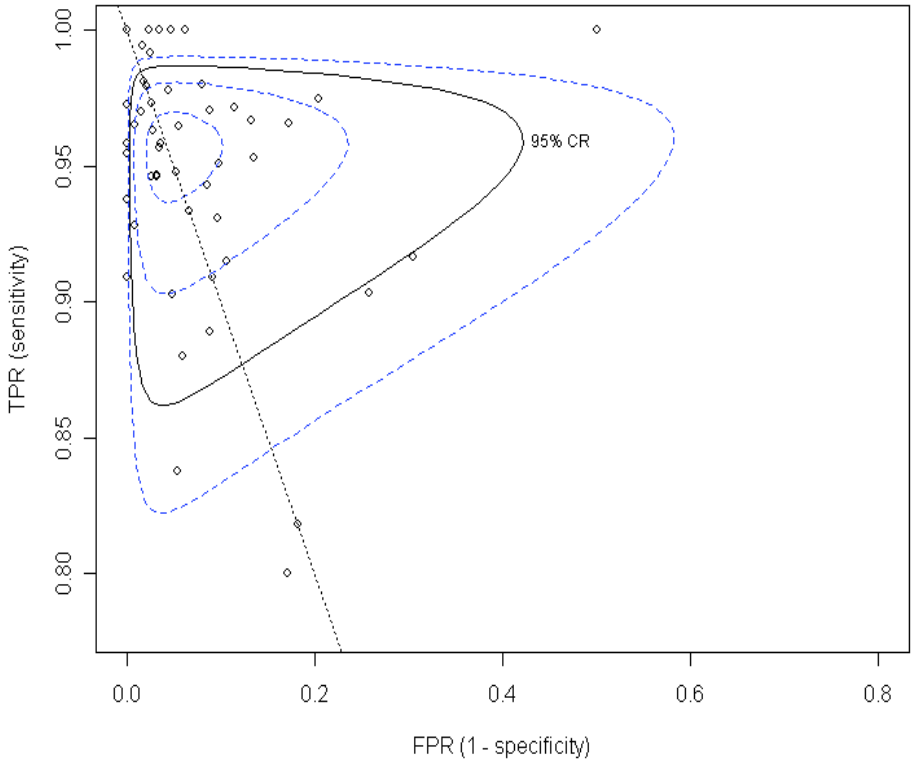

Figure 5.8: *Predictive surface for the pairs (FPR, TPR).*

5.4 Numerical comparison with other approaches

The model introduced in this chapter is similar but not equivalent to other methods reviewed in Chapter 1. Users of these meta-analytic techniques may be interested in the difference between these techniques in practice. In this section, we present numerical and graphical results that can clearly point out advantages and disadvantages of these statistical techniques.

5.4.1 Comparisons with the classical bivariate approach

In Chapter 1 we have seen that a series of bivariate models with bivariate Gaussian random effects has been proposed in the literature (Reitsma et al., 2005 [94]; Arends, 2006 [3]; Chu and Cole, 2006 [12]; Macaskill [83]; and Harbord et.al., 2007 [59]) The current implementation of these techniques are in SAS with using the NLMIXED procedure and in Stata software. Both implementations are based on Adaptive Gaussian Quadrature, the programs have been made available by courtesy of Dr. Roger Harbord.

The classical bivariate model directly models the pairs of (tp_i, tn_i) and uses logistic link function with a bivariate normal distribution for the random effects. We reparametrized our Bayesian model in order to make results comparable in terms of the classical model.

We use two data sets, first one with our data with 52 studies and second one with a small set with 10 randomly selected studies. These 10 studies were selected with the R commands:

```
> set.seed(123)
> sample(1:52, size = 10, replace = FALSE)
 [1] 15 41 21 44 46  3 25 51 48 20
>
```

The SAS and Stata programs are run with the default initial values and without any change in the programs' setup. The Bayesian model is run with an a single chain with length 100,000 with the first 50,000 discarded, this numerical setup gave exactly three decimal values, i.e. values calculated without Monte Carlo error.

Table 5.3 presents the results of this numerical analysis. We can see that for the case on $n = 52$ studies all softwares gave very similar numerical results. Both SAS and Stata implementations deliver fast and reliable results. In general, we can see that the model proposed in this chapter is comparable with previous models and software implementations presented in the literature.

However, when we repeat the calculations for $n = 10$ we found that SAS prints out the warning

	n =52			n=10		
	Coef.	Std.	[95% CI]	Coef.	Std.	[95% CI]
SAS				(*)		
E(logitSe)	3.092	0.120	[2.850, 3.333]	2.694	0.202	[2.228, 3.161]
E(logitSp)	3.023	0.183	[2.655, 3.390]	3.021	0.416	[2.060, 3.981]
Var(logitSe)	0.255	0.135	[-0.016, 0.526]	0.023	0.051	[-0.094, 0.141]
Var(logitSp)	1.170	0.332	[0.504, 1.836]	1.204	0.060	[1.064, 1.343]
Corr(logits)	-0.084	0.243	[-0.572, 0.403]	-1.000	.	[. , .]
Stata				(*)		
E(logitSe)	3.092	0.120	[2.856, 3.327]	2.694	0.202	[2.298, 3.091]
E(logitSp)	3.030	0.186	[2.665, 3.395]	3.030	0.445	[2.157, 3.902]
Var(logitSe)	0.255	0.136	[0.089, 0.726]	0.023	0.0507	[0.001, 1.700]
Var(logitSp)	1.198	0.332	[0.696, 2.063]	1.239	0.822	[0.337, 4.548]
Corr(logits)	-0.075	0.243	[-0.503, 0.383]	-1.000	.	[. , .]
R/WinBUGS						
E(logitSe)	3.097	0.123	[2.866, 3.346]	2.564	0.273	[2.029,3.107]
E(logitSp)	3.010	0.183	[2.656, 3.377]	2.451	0.534	[1.445,3.564]
Var(logitSe)	0.303	0.132	[0.118, 0.624]	0.428	0.299	[0.120,1.188]
Var(logitSp)	1.217	0.358	[0.671, 2.064]	2.421	1.704	[0.733,6.763]
Corr(logits)	-0.066	0.223	[-0.487, 0.373]	-0.286	0.352	[-0.830,0.486]

Table 5.3: *Summary results for a bivariate random effect model. Three softwares SAS, Stata and R/WinBUGS and two sample sizes n=52 and n=10. (*) SAS and Stata reports convergence with warnings.*

```
NOTE: GCONV convergence criterion satisfied.
NOTE: At least one element of the (projected) gradient
      is greater than 1e-3.
WARNING: The final Hessian matrix is full rank but has
      at least one negative eigenvalue. Second-order
      optimality condition violated.
```

indicating convergence problems. In Table 5.3 we can see that the correlation parameter is reported equal to -1 without standard error and without confidence intervals. Stata gave similar results, but without any warning message. Our MCMC calculations gave convergence after about 5,000 simulations and posterior distributions reflect the dispersion caused by reducing the number of studies in the analysis. Moreover, the difficult correlation parameter has a posterior distribution with correct percentiles.

5.4.2 Comparison with the HSROC curve

To compare the bivariate Bayesian model with the HSROC model proposed by Rutter and Gatsonis (2001)[99] we used the BUGS code published by the authors. We applied exactly the same prior specifications presented in the BUGS script, this include a carefully chosen informative prior distribution for the dispersion parameters in the model (see Sections 3.2.1 and 3.2.2 in the paper).

We fit the HSROC curve model with the original data with $n = 52$ by running three randomly initiated chains of length 100,000 and we discarded the first 50,000. The panels of Figure 5.9 gave the trace plots of the last 1,000 simulated values of these chains, we see that the parameters Θ, Λ and $b = \exp(\beta)$ do not converge. Interesting, the B-G-R diagnostic reports convergence, this diagnostic is used for the authors, which for our data gives misleading results. We try to extend the number of simulations, but convergence was not possible. We also try standard non-informative priors distributions for the variance parameters and WinBUGS stops report numerical errors. Given all of these convergence problems we did not perform the analysis with smaller sample size. However, we report the resulting HSROC curve in Figure 5.10, but these results should also be carefully interpreted.

5.5 Concluding remarks

In this chapter we have presented a novel approach for meta-analysis of diagnostic test. This approach is inspired in the classical SROC curve of Moses et al. (1993) [87], however, by using modern Bayesian modeling techniques we avoided the classical critics of the SROC model. Moreover, we presented a uniform modeling approach that gave a Baysian version of the SROC curve, of the area under the curve, and the predictive surface in the ROC space.

We summarize in the following points, which compare our approach with the current ones:

1. Our approach includes all sources of variability like other bivariate meta-analysis models. That is an important aspect in comparison with the SROC curve.

2. We make emphasis in model checking and model diagnostic. This aspect is completely ignored for the current bivariate meta-analytic methods.

3. We get very similar numerical results in comparison with bivariate models implemented in SAS and Stata. But superior results in terms of numerical conver-

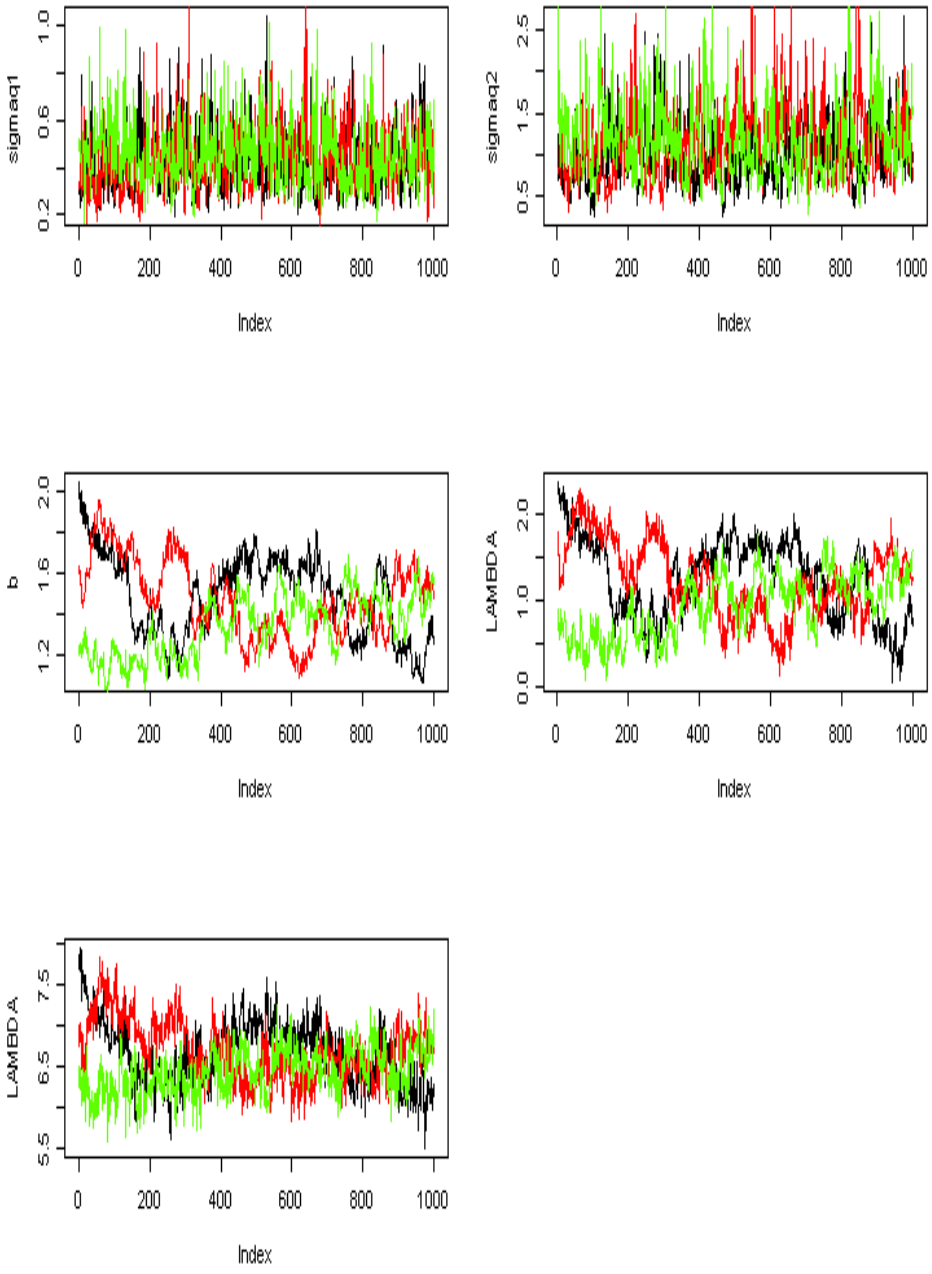

Figure 5.9: *Trace plots for model parameters of the HSROC.*

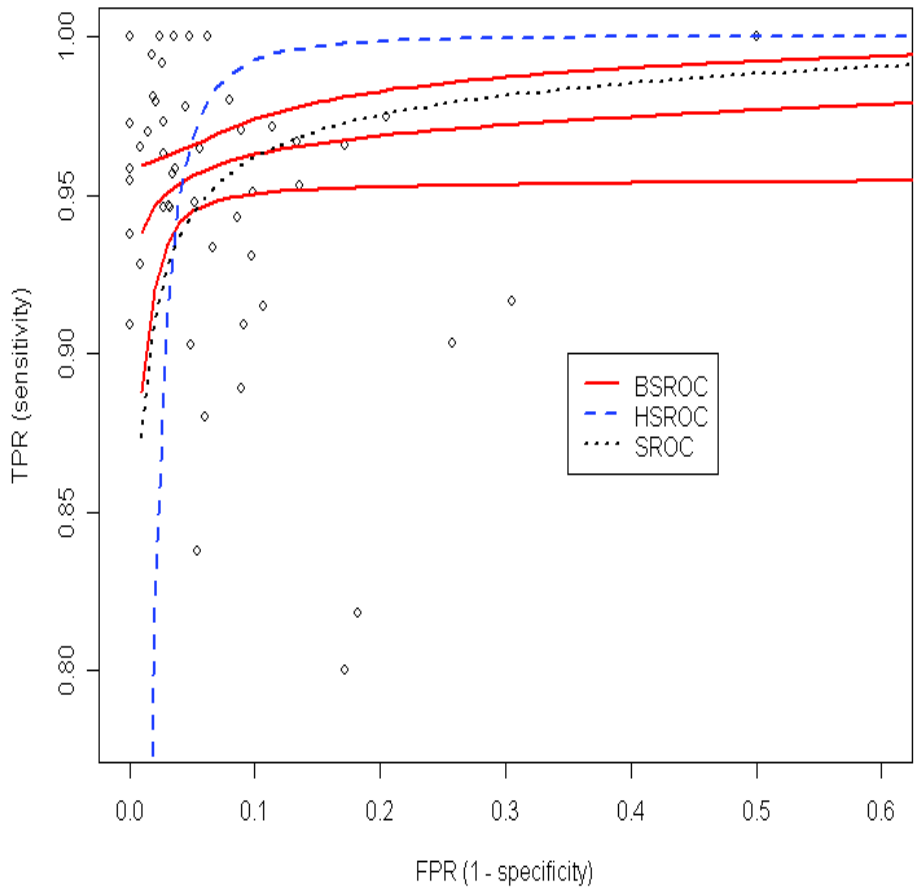

Figure 5.10: *Comparison between three different methods: SROC, HSROC and BSROC.*

gence for small sample size. We make a case study with $n = 10$, which is a very reasonable number of studies to be include in a meta-analysis.

4. With our experience on the SAS and Stata implementation of the bivariate meta-analysis model, we recommend its use for relatively large sample sizes (at least 20 studies included in the meta-analysis).

5. The HSROC curve model is the only complete bivariate Bayesian approach previously presented in the literature. This model has been receiving critics for bing difficult to interpret. We found that it is also numerically unstable and it is based on a poor parametrization. In general, we do not recommend its use in meta-analysis of diagnostic test.

We conclude that MCMC techniques are the method of choice to perform this type of meta-analytical problems. Moreover, the numerical reliability of our model permits us to make further complex extensions that will be presented in the following chapter. These extensions allow us to use non-normal distributions for random effects, fit bivariate meta-regression models with missing values on the covariates and make automatic variable selection. Neither current popular statistical software (e.g. SAS or Stata) nor classical methodological advances (e.g. adaptive Gaussian quadrature) make these extensions possible.

Chapter 6

Modeling extensions

"(...) Box also coined the aphorism "all models are false but some models are useful", raising immediately the questions as to what makes a false model useful and as to why should we bother to criticize models that we know are false anyway."

-D.R. Cox (2004) *Methods and models in statistics: in honor of Professor John Nelder, FRS*, page 13.

6.1 Introduction

In this Chapter we extend the model presented in the previous Chapter. These model extensions are used to summarize the meta-analysis in a novel way. That includes: assessment of credibility of studies with retrospective designs relative to studies with prospective designs, quantification of study relevance and meta-regression to explore the influence of study quality, study population and study characteristics in diagnostic results.

6.2 Accounting for studies with different designs and relative credibility

Our meta-analysis includes studies with retrospective and prospective designs, which is known as *cross-synthesis* meta-analysis. Including different study designs in a meta-analysis may extend the inferential scope, e.g. the spectrum of the population under study at the cost of increasing the complexity of the model.

Given the uncontrolled context where the data of retrospective studies are obtained, we may expect that retrospective studies present substantially more variability than prospective ones. One way to quantify this feature is by modeling the variance matrices of the random effects for each study design separately. Let Σ_R and Σ_P be the between studies covariance matrix for retrospective and prospective studies respectively, then we model random effects as

$$(D_{i,d}, S_{i,d}) \sim \texttt{Normal}_2(\mu, \Lambda_d), \quad i = 1, 2, \ldots, N, \quad d \in \{R, P\}, \tag{6.1}$$

where $\Lambda_d = \Sigma_d^{-1}$. In this model we assume that each study informs us on the same mean parameter μ but at the cost of increasing variability. One attractive feature of this approach is that retrospective studies will be down-weighted if they present more variability than prospective ones. We may use priors on Λ_d to reinforce a weighted schema, but we prefer to model a-priory Λ_d with a common non-informative distribution and leave the data to dominate inference.

We can summarize study design variability by the trace of their covariance matrices Σ_d and define *the relative credibility* (RC) of retrospective studies relative to prospective ones by the ratio

$$RC = \frac{\text{trace}(\Sigma_P)}{\text{trace}(\Sigma_R)}. \tag{6.2}$$

The posterior distribution of RC is used to describe this data feature, mass of probability concentrated away from 1 indicates lower level of evidence of studies with retrospective design compared to studies with prospective design.

Data analysis

Figure 6.1 shows the posterior distribution of RC for our running example. This histogram is based on a single chain with 10,000 iterations after discarding the first 10,000 ones. The posterior mean is 0.552 with 95% credibility interval (0.190, 1.264), which points out that studies with retrospective design have contributed with less information to the analysis. The DIC for this model is 423.8 which indicates an interesting improvement compared to the model with common variability between studies with different design.

Accounting for study design variability has influenced summary diagnostic results as follows: sensitivity predictive summaries are almost the same with 0.951 (0.892, 0.984) for prospective studies and 0.949 (0.871, 0.986) for retrospective ones. Specificity summaries are markedly different, with 0.935 (0.773, 0.992) for prospective studies and 0.913 (0.579, 0.997) for retrospective ones.

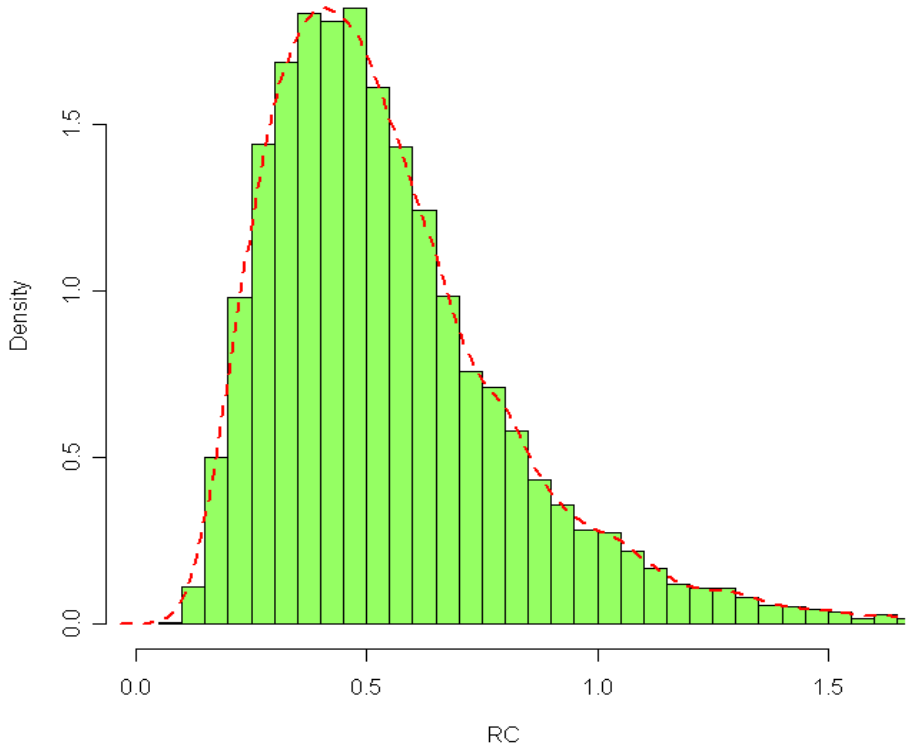

Figure 6.1: *Posterior distribution of* RC. $RC = 1$ *indicates the same credibility between studies with retrospective and prospective designs. Most of the probability mass is over* $RC < 1$*, indicating less credibility for studies with retrospective designs.*

Figure 6.2 presents the 95% predictive posterior surfaces for studies with retrospective and prospective designs, we can clearly see the effect of study design in the meta-analysis. In synthesis, retrospective studies are less specific and more uncertain.

6.3 Non-Gaussian random effects and study relevance

In Section 6.2 we have quantified the relative variability for one predefined group of studies with respect to the rest of the studies, which is relatively simple because we know a-priory how the data have been collected according to their study design. In this section we are interested to identify some particular studies that could influence results and are not simple to find a-priory.

For this sort of outliers detection, we model the random effects (D_i, S_i) with a

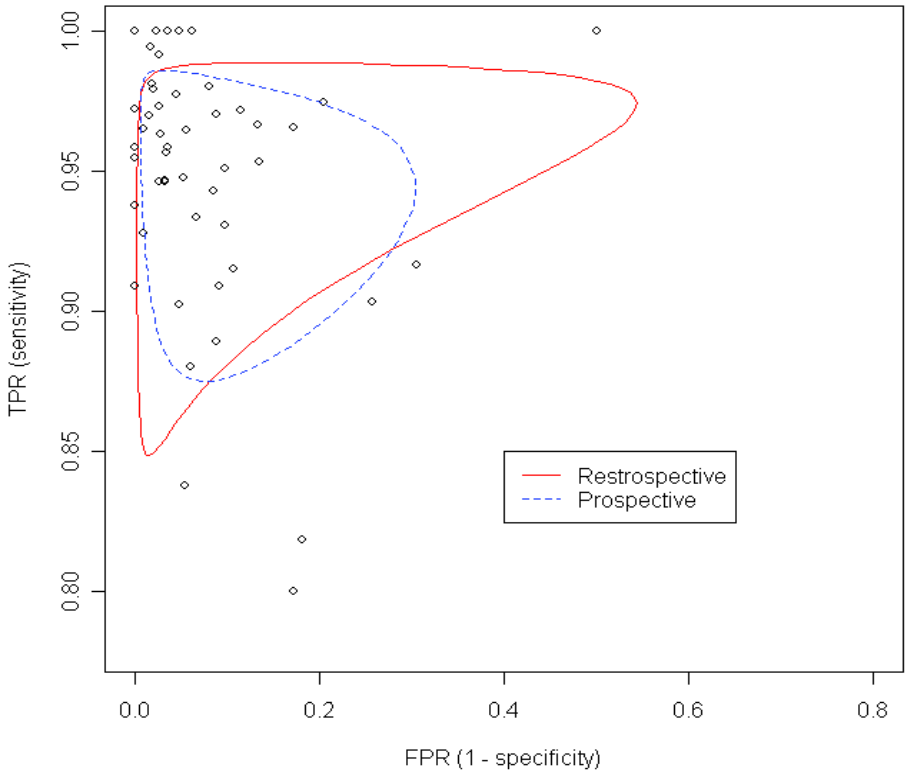

Figure 6.2: *Differences in 95% predictive posteriors surfaces for studies with retrospective and prospective design.*

mixture of two bivariate *t*-distribution with common mean μ, dispersion matrices Λ_R and Λ_P, and common degrees of freedom $\nu > 2$. This multivariate *t*-distribution can be constructed as a mixture of bivariate normal distributions with the following hierarchical structure,

$$(D_{i,d}, S_{i,d}) \sim \texttt{Normal}_2(\mu, \Psi_{i,d}), \quad i = 1, 2, \ldots, N, \quad d \in \{R, P\}, \tag{6.3}$$

$$\Psi_{i,d} = w_i \times \Lambda_d, \tag{6.4}$$

$$w_i \sim \Gamma(\nu/2, \nu/2). \tag{6.5}$$

The weight w_i is an outlier indicator for study i, which in the context of meta-analysis we call *study relevance*, a term borrowed from Efron [34], but used here in a different way. A study is relevant to the systematic review, if its results do not substantially deviate from the rest of the studies. All studies are relevant a-priory with mean relevance

of 1. Less relevant studies will concentrate posterior distribution of w_i with values less than 1.

We may give a fixed small value of ν, e.g. $\nu = 4$ or admit our uncertainty about ν by modeling this parameter with an exponential prior distribution with support on values $\nu > 2$ and parameter τ (Geweke, 1993) [50]. Values of τ between [0.5, 0.01] correspond to degrees of freedom in the range of 2 to 100.

Outliers modeling by a scale mixture approach has a long history in Bayesian data analysis, in linear models it was proposed by De Finetti (1961) [22], in generalized linear models by West (1984, 1986)[128, 129] and for recent modern applications see Congdon (2006) [14].

Data analysis

The data of our example resulted from a carefully performed systematic review and we could expect relatively good results for a relevance analysis. We applied the mixture multivariate *t*-distribution (6.3) by running a chain of length 20,000 and by discarding the first 10,000 iterations as previous sections. The estimated degrees of freedom ν is 8.47 (2.62, 30.96), indicating the presence of studies with lower relevance. Studies number (R)1, (R)3, (R)4, (R)7, (P)25, (P)34 had posterior mean relevance weights of 0.65, 0.71, 0.69, 0.73, 0.73 and 0.77 respectively. Study number (P)48 has relevance 0.25 indicating that it is a clear outlier. Figure 6.3 displays the location of these studies in the ROC space with circles and their study identification number. The DIC = 409.83 shows an important improvement on model building. We see that model design and study relevance are both important features of this meta-analysis.

Going back to study information, we found that these unusual diagnostic results have been produced by a remarkable imbalance between disease and non-disease groups, which is more accentuated in retrospective studies. This analysis points out serious design deficits in published results.

6.4 Meta regression

In this section, we present a meta-regression approach that may be useful to analyze the impact of some published information, like study characteristics or population differences, in diagnostic accuracy. This type of analysis is evidently not possible for single studies and gives an extra pay-off to the meta-analysis. However, it is worth mentioning some limitations of meta-regression methods: results are susceptible to

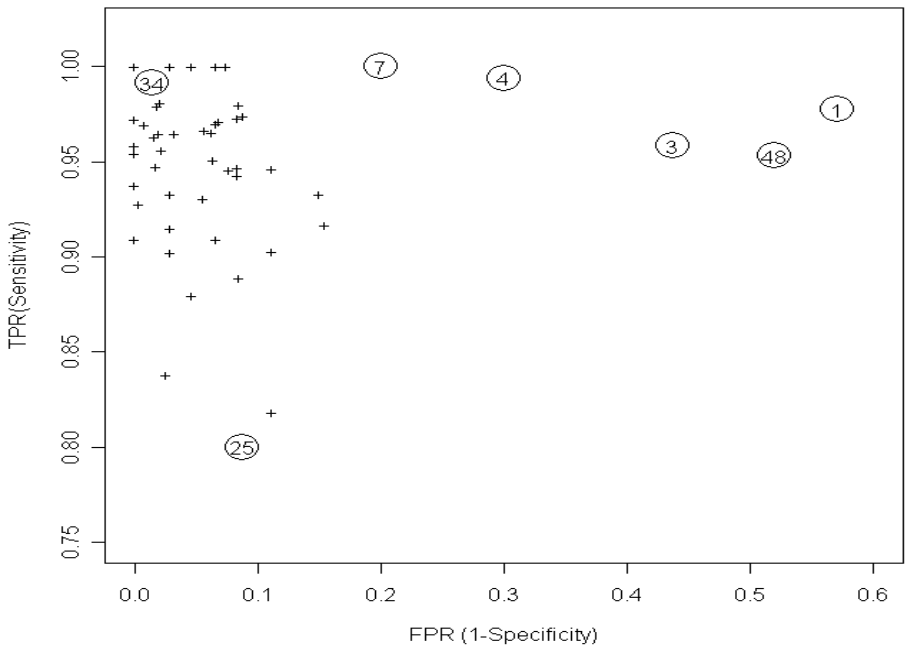

Figure 6.3: *Relevance analysis, circles indicate estimated studies with lower relevance, in particular study number 48 has a relevance of 0.25.*

aggregation or ecological bias, which occurs when study results and published populations' summaries do not directly reflect the relationship between patient characteristics and patients' diagnostic outcomes. In addition, the published available data for analysis may be limited, we mitigate this issue by modeling missing covariate data. In synthesis, meta-regression analysis should be interpreted as a *knowledge discovery* approach where results may be useful to suggest further investigation.

It is easy to include a regression structure to analyze systematic influence of variables in diagnostic results. We write $(\mu_{i,D}, \mu_{i,S})$, the fixed effects of the model, as a system of two regression equations,

$$\mu_{i,D} = \alpha_0 + \alpha_1 x_{i,1} + \ldots + \alpha_p x_{i,p}, \tag{6.6}$$

$$\mu_{i,S} = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p}, \tag{6.7}$$

where each equation depends on a known vector of covariates $(x_{i,1}, \ldots, x_{i,p})$ and unknown regression coefficients vectors $(\alpha_1, \ldots, \alpha_p)$ and $(\beta_1, \ldots, \beta_p)$.

A positive value of α_i corresponds to an improvement of the diagnostic performance and a negative value to a reduction of this feature. They can be directly inter-

preted as an odds ration in the logarithmic scale. In the same line, a positive value of β_i represents an increase in test positive outcome and negative values a reduction. Posterior distributions centered at 0 represent variables that do not have any influence on test results.

6.4.1 Variable selection strategies

Prior distributions for the regression coefficients are essential ingredients in any Bayesian analysis, they encapsulate a variable selection strategy. The nature of our model building allows us to incorporate different variable selection strategies in a complex hierarchical model. We present two well known Bayesian variable selection procedures that will be compared in our data analysis.

Ridge regression

In ridge regression we model α_j and β_j as exchangeable with two independent Gaussian distributions,

$$\alpha_j \sim \mathtt{N}(0, \phi_\alpha), \quad \beta_j \sim \mathtt{N}(0, \phi_\beta). \tag{6.8}$$

Now, for fixed values of ϕ_α and ϕ_β this approach is equivalent to the classical ridge regression, where these parameters are usually estimated by a cross-validation technique, for an excellent introduction to ridge regression see Hastie, Tibshirani and Friedman (2001, pag 59-64)[61]. In Bayesian modeling we prefer to admit uncertainty on these scale parameters and model them with two independent Gamma distributions: $\phi_\alpha \sim \Gamma(r_\alpha, m_\alpha)$ and $\phi_\beta \sim \Gamma(r_\beta, m_\beta)$. We use the constants r_α, r_β, m_α and m_β to make a sensitivity analysis for the regression results.

Stochastic Search Variable Selection

Stochastic Search Variable Selection (SSVS) was introduced by George and McCulloch (1993)[49], it consists of modeling the regression coefficients with mixture a normal distributions with common mean 0 and different precision parameters. The SSVS procedure is also known as *Spike and Slab Variables Selection*, which has been developed from the classical and Bayesian perspectives, for a recent review of these techniques see Ishwaran and Rao (2005)[66].

The method presented in this section is close to Dellaportas et al. (2000, 2002)[23, 24] and based on the implementation in BUGS given by Ntzougras (2002)[89].

In the SSVS approach we model α_j and β_j as exchangeable with two independent mixture Gaussian distributions,

$$\alpha_j \sim \gamma_{\alpha,i}\,\mathtt{N}(0,\phi_\alpha) + (1-\gamma_{\alpha,i})\,\mathtt{N}(0,c_\alpha), \tag{6.9}$$

$$\beta_j \sim \gamma_{\alpha,i}\,\mathtt{N}(0,\phi_\beta) + (1-\gamma_{\beta,j})\,\mathtt{N}(0,c_\beta). \tag{6.10}$$

The random variables $\gamma_{\alpha,i}$ and $\gamma_{\alpha,i}$ are indicator variables, which are modeled a priori as independent with

$$\gamma_{\alpha,i} \sim \mathtt{Ber}(p_{\alpha,i}) \quad \text{and} \quad \gamma_{\beta,j} \sim \mathtt{Ber}(p_{\beta,j}). \tag{6.11}$$

The probabilities $p_{\alpha,i}$ and $p_{\beta,j}$ are hyper-parameters that represent the probability that a covariate x_i will be included in the model. We can set these probabilities equal to 0.5 to indicate prior ignorance, or model our uncertainty with vague priors for $p_{\alpha,i}, p_{\beta,j} \sim \mathtt{Beta}(0.5, 0.5)$ (Box and Tiao, 1973) [7].

The precision parameters c_α and c_β are also considered know a priory. They are set to a large value to indicate a *spike* at 0, which corresponds to a model where the regression coefficients are all equal to 0. In our model we set up $c_\alpha = c_\beta = 100$.

Uncertainty about the scale parameters ϕ_α and ϕ_β is modeled with two independent Gamma distributions: $\phi_\alpha \sim \Gamma(r_\alpha, m_\alpha)$ and $\phi_\beta \sim \Gamma(r_\beta, m_\beta)$ as in the previous section. Low values of ϕ_α and ϕ_β correspond to the *stab* of the model, which allows that the regression coefficients can take any value far from 0.

6.4.2 Accounting for covariates with missing data

One additional problem in meta-regression is the presence of covariates with missing values. Chapter 2 reports 4 covariates with proportions of missing values ranging from 10% to 21%. If we select only studies with completed reported data we reduce our sample form 52 to 23 studies, with the potential problem of bias results. Including a sub-model for missing covariates can alleviate this problem.

In a Bayesian approach, missing covariates data is reduced to a posterior prediction problem. In our running example, we have only missing data in dummy covariates, so for each covariate z_k $(k = 1, \ldots, 4)$ we assign a Bernoulli data model,

$$z_k \sim \mathtt{Ber}(p_{z_k}),$$

with vague priors for $p_{z_k} \sim \mathtt{Beta}(0.5, 0.5)$ (Box and Tiao, 1973) [7]. Posterior distributions of p_{z_k} depend on the observed part of each covariate and the contribution of

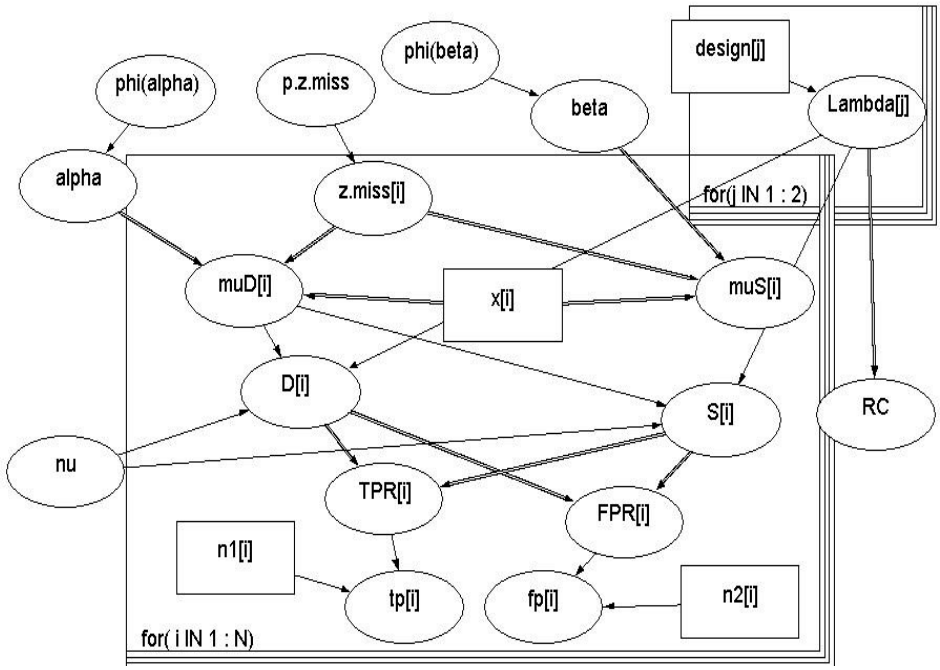

Figure 6.4: *DAG for the meta-regression model including mixture t-distribution.*

the study effect (D_i, S_i). Missing values are imputed by sampling from the predictive posterior distribution of z_k in each MCMC iteration.

Figure 6.4 presents the DAG for the meta-regression model including imputation for missing covariate data and variable selection by the ridge approach.

6.5 Data analysis

We fitted two models, one with ridge regression and another with SSVS approach, covariates with missing values where, also, included in the model.

Figure 6.5 summarizes the results for both models. Each segment corresponds to the 5%, 50% and 95% percentiles of the posterior distribution for each group of coefficients. The left panel corresponds to the posteriors for α_is and the right panel for β_is, solid lines correspond to the model with SSVS and dotted lines correspond to the model with ridge regression. We can see that both variable selection methods point out the same conclusions that we summarize in the following points:

- Studies performed outside *University Hospitals* tend to present better diagnos-

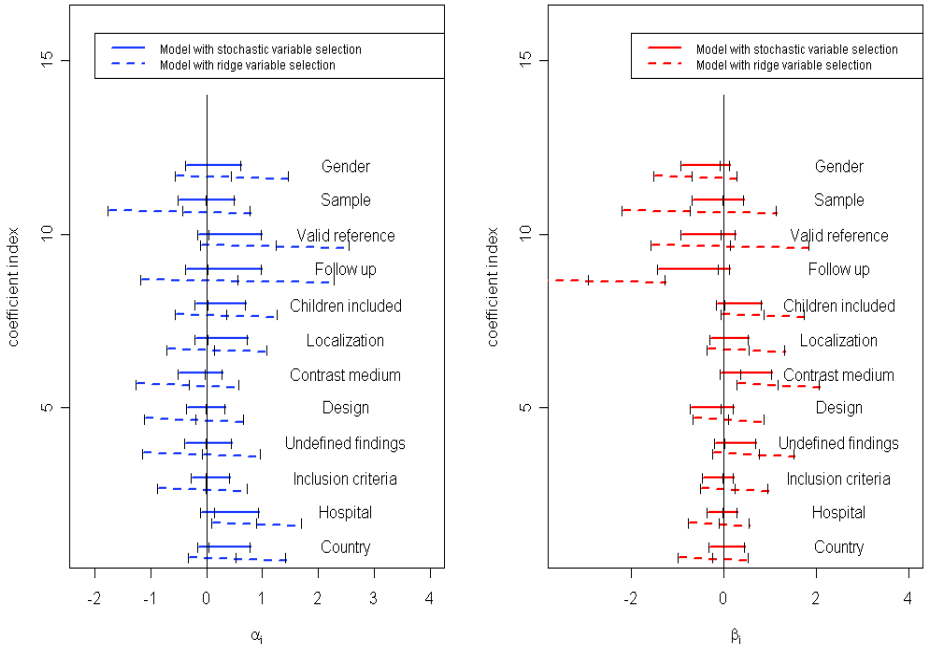

Figure 6.5: *Summary plot for regression analysis. Left panel: regression coefficient α_i explaining influence of test discriminatory power. Right panel: regression coefficient β_i explaining influence of positive test results.*

tic results indicating that some of these clinics may have specialists in this CT technology.

- Studies with *Contrast Medium* do not improve diagnostic results, and increase false positive rates. This result indicates that contrast medium should be avoided.

- For studies with more than 50% of women in the study population the CT diagnostic tends to deliver less positive values. This result show that women can present a variety of alternative pathological findings.

- Studies with *Children included* present better diagnostic results, probably by the fact that this is a common disease in the children population.

- Studies with *Valid reference standard* tend to deliver better results with less increase of false positives rates. This result, together with the trend of studies with *follow-up* to reduce the number of positive test results can be interpreted as a bias correction for lack of quality in the studies.

This analysis shows the importance of the modeling effort to include missing covariates data, some interesting results including a bias correction for low quality performance (no valid reference and no follow-up) are coming from this part of the data. Other modeling characteristics, for example the heavy tails of the random effects remain important, in both models the number of degrees of freedom ν was estimated as 9.476 (2.214, 48.860) for the SVSS and the relative credibility for studies with retrospective design remains low RC 0.380 (0.084, 1.139).

The DIC of this model is 601, which indicates less model fitness than the previous models. However the presence of model components with missing values and the structural distribution for variables selection make this DIC not comparable with the models without covariates. In this case it should be more appropriate to decompose the DIC by model components (see Spiegelhalter, et al. 2004 [110]).

For the scale parameters we have: ϕ_1 2.373 (0.648, 3.799) and ϕ_2 2.288 (1.222, 3.799). We analyze the sensitivity of these results with two priors for the scale parameters ϕ_1 and ϕ_2, one using a less informative $\Gamma(1,1)$ and another with a strong informative $\Gamma(10,5)$. Results remained robust to this prior sensitivity analysis. Furthermore, we run this analysis from 3 chains randomly initiated at different starting points and convergence presented no problem.

6.6 Concluding remarks

In this chapter we have presented a series of innovative modeling extensions. Two new concepts in multivariate meta-analysis are presented, the relative credibility between studies with different designs and the study relevance as a measure of study quality.

We have introduced a bivariate meta-regression model with two different variable selection procedures and with the possibility to include covariates with missing data.

We conclude this chapter with the following summary remarks:

- The use of a mixture of bivariate distributions to model study effects, makes it possible to understand the effect of performing cross-synthesis , i.e. the inclusion of studies with different designs in the meta-analysis.

- We have measured the *relative credibility* of studies with retrospective design in comparison to studies with prospective design. We have seen that results form studies with retrospective design are systematically more disperse.

- The measurement of study relevance through a bivariate-t distribution has shown that some studies are not in concordance with the rest of the meta-analysis and these studies presented possible serious design and quality problems.

- We have presented a bivariate meta-regression model to assess systematic variability through test, study and population characteristics. These models have included covariates with missing data as well. At the present, no other bivariate meta regression approach has been presented in the literature with this level of flexibility.

- Two different Bayesian variable selection procedures have been applied. In general, we have experienced the tremendous problem of finding covariates that point out about clear trends. That may indicate that the published data is not sufficient to resolve these issue.

We conclude that the Bayesian approach to meta-analysis of diagnostic test data allows modeling extensions which reflect the complexity of the published data. These modeling extensions can be implemented with reliable numerical techniques based on MCMC implemented in BUGS and R. The next chapter centers on the important topic of making these methods freely accessible to practitioners.

Chapter 7

Statistical Computations with R and BUGS

"Research statistician proudly have a great idea, write it up, getting glowing referees reports from a good journal, maybe even read a paper to the Society, and then sit back and wait for the idea to conquer the world."

And wait and wait and wait ...

-Brian D. Ripley (2002) *Statistical Methods Need Software: a View of Statistical Computing.* RSS.

7.1 Introduction

In this chapter we describe how to perform the bootstrap calculations of Chapter 3 and to fit the Bayesian models introduced in Chapter 5 and Chapter 6. We use R [93] and WinBUGS [79] statistical systems, both available for free. We assume that the user has experience in both statistical systems. Our statistical analyzes have been done on Windows systems and tested on Windows XP, VISTA and Linux operating systems respectively.

We have prepared two R scripts as complementary material:

- One script to make the bootstrap calculations of Chapter 3. This file is called `meta-boot-Rscript.r`. This script can be run from start to the end without any special setup.

- Another script is called `meta-bayes-Rscript.r` to perform the statistical analysis of Chapter 5 and Chapter 6. This script requires some special setup that is described in the following two sections.

7.2 Getting started

You should setup your computational environment as follows:

1. Install R (2.7.0 or above) and WinBUGS (1.4.3).
2. Within the R's console install the package `R2WinBUGS` and load it:

```
> library(R2WinBUGS)
```

3. Setup your working directory and the WinBUGS directory, e.g.

```
> bugsdir <- "C:/Programs/WinBUGS14"
> workdir <- "C:/meta-analysis"
```

4. The file `models.txt` contains our models in BUGS language. It has to be edited by splitting it up into 6 files, one for each model. For instance, the file `model1.txt` will correspond to our starting model of Chapter 5. Put these model files in your *working directory*.
5. Copy and paste the data section of the script into your R workspace:

```
# true positive
tp <- c(87, 111, 184, 168, ...
# number of patients with disease
n1 <- c(89, 115, 192, 169, ...
...
```

7.3 Bootstrap methods in R

In this section we explain how to make the bootstrap calculations of Chapter 3. There are two packages in R to make bootstrap analysis, one is the `boot` package which is linked to the volume of Davison and Hinkley (1997) [19] and the other package is `bootstrap` which is associated to the introductory book of Efron and Tibshirani (1993) [39]. In our calculations we use `boot`, which in general has a more flexible implementation and is part of the standard distribution of R.

The first step to make bootstrap calculations with the package `boot` is to write the statistics of interest in the *sampling form*. This is done by defining a function with arguments corresponding to the data and to a vector with observations' index as following:

```
theta.boot <- function(dat, Ind){
    S <- dat[Ind, 1]
    D <- dat[Ind, 2]
    f1 <- lm(D~S)
    A <- coef(f1)[1]
    B <- coef(f1)[2]
    f <- function(FPR, A, B){
     x <- A/(1-B) + (B+1)/(1-B)*log(FPR/(1-FPR))
     exp(x)/(1+exp(x))
        }
    theta <- integrate(f, A, B, lower=0, upper=1)$value
    var.theta <-  var.linear(
              empinf(data = dat[Ind, ],
              statistic = theta.abc,
              type = "jack", stype="w")
              )
    return(c(theta, var.theta))
}
```

Here the function `theta.boot` gives the sampling form of the Area Under the Summary ROC curve. It has two arguments, one is `dat`, the dataframe with the original data, and the other is the argument `Ind`, which is used by the function `boot()` to produce the bootstrap values based on the statistic `theta.boot`.

This function is implemented to make all the bootstrap confidence intervals presented in Chapter 3. Its value is the estimated AUC and its variance calculated with the Jackkinfe method for a particular bootstrap sample. To calculate this variance, we use two functions from the package `boot`. One is `empinf()`, which calculates the empirical influence values for the statistic applied to a data set. Its value is a vector of the empirical influence values. The second function is `var.linear()`, which estimates the variance of the statistic from its empirical influence values.

In order to apply `empinf()` to estimate the variance of the AUC in each bootstrap sample, we need to write the statistics AUC in its *weighted form*. We write the function `theta.abc` for this purpose:

```
theta.abc <- function(dat, P) {
   S <- dat[ , 1] ;  D <- dat[ , 2]
   P <- abs(P/sum(P))
  f1 <- lm(D~S, weights = P)
   A <- coef(f1)[1]
   B <- coef(f1)[2]
   f <- function(FPR, A, B)
        {
        x <- A/(1-B) + (B+1)/(1-B)*log(FPR/(1-FPR))
        exp(x)/(1+exp(x))
        }
  integrate(f,A, B, lower=0, upper=1)$value
  }
```

The function `theta.abc` is implemented in a very flexible way. Note that its second argument is `P`, each component of `P`, say `P[i]` is the probability that the observation `dat[i, ]` is in a bootstrap sample. This function can be used to perform Jackknife variance calculations or more sophisticated bootstrap calculations like the non parametric ABC confidence interval (see Efron and Tibshirani 1994, page.188)[39].

The following lines gives the bootstrap analysis with $R = 1000$ bootstrap replications of the AUC:

```
library(boot)

aucboot <- boot(dat, theta.boot,  R=1000)

ci <- boot.ci(aucboot, conf=c(0.90, 0.95,0.99),
     type=c("norm","basic", "stud", "perc",
     "bca"))

ci

Intervals :
Level  Normal            Basic          Studentized
90%   (0.975, 0.990 ) ( 0.976, 0.991 ) (0.972,  0.990)
95%   (0.973, 0.991 ) ( 0.976, 0.993 ) (0.969,  0.992)
99%   (0.970, 0.994 ) ( 0.974, 0.998 ) (0.960,  0.997)

Level     Percentile            BCa
90%   (0.971, 0.986)  (0.973,  0.986)
95%   (0.969, 0.986)  (0.971,  0.987)
99%   (0.965, 0.989)  (0.966,  0.989)
```

In Chapter 3, we have introduced a transformed version of the bootstrap-*t* interval. To implement this bootstrap confidence interval we need to define three functions: the original function $m(\cdot)$, the inverse function $m^{-1}(\cdot)$ to scale back the interval, and the first derivative $m^{'}(\cdot)$ to calculate the variance in the transformed scale:

```
m <- function(x){
      m1 <- x/(1-x)
     (m1^0.303 -1)/0.303
      }

m. <- function(x){
      (x/(1-x))^(0.303-1)*1/(1-x)^2
       }

minv <- function(y){
         a <- 0.303*y+1
         a <- a^{1/0.303}
         a/(1+a)
         }
```

The bootstrap calculations defined on the scale of the function $m(\cdot)$ is performed as following:

```
boot.ci(tmp, conf=c(0.90, 0.95,0.99),h=m, hdot=m., hinv=minv,
        type=c("norm", "stud"))

.......
Intervals :
Level      Normal            Studentized
90%   ( 0.973,  0.987 )   ( 0.971,  0.989 )
95%   ( 0.971,  0.987 )   ( 0.967,  0.990 )
99%   ( 0.967,  0.989 )   ( 0.953,  0.993 )
```

7.4 Bayesian analysis

To perform the statistical analyzes described in Chapter 5 and in Chapter 6, we run BUGS within R with the function `bugs()` from the package `R2WinBUGS` [112]. This approach combines the powerful MCMC calculations implemented in BUGS and gives flexibility for building plots and further summaries within R. It is the recommended form to make this type of Bayesian statistical analysis.

7.4.1 Fitting a model with R and BUGS

We call Model 1 the simple bivariate Generalized Linear Mixed effects model, with logistic link function and bivariate normal random effects distribution. This is the simplest model presented in Chapter 5. In order to fit this model, we assume that the BUGS code for Model 1 is in the file `model1.tex`:

```
model
{
    for( i in 1 : n ) {
        tp[i] ~ dbin(tpr[i], n1[i]); fp[i] ~ dbin(fpr[i], n2[i])
        logit(tpr[i]) <- m[i,2]/2 + m[i,1]/2   # (Di + Si)/2
        logit(fpr[i]) <- m[i,2]/2 - m[i,1]/2   # (Di - Si)/2
        m[i,1:2]~dmnorm(mu[], sigma.inv[1:2 ,1:2 ] )
     }
# Priors ...
    mu[1] ~ dnorm(0, 0.25)
    mu[2] ~ dnorm(0, 0.25)
    sigma.inv[1:2,1:2] ~ dwish(R[1:2,1:2], 3)
# Summary pooled statistics ...
    x <- (mu[1]+mu[2])/2
    y <- (mu[2]-mu[1])/2
    se <- exp(x) / ( 1 + exp(x) )     # with logit link
    sp <- 1 - exp(y) / ( 1 + exp(y) ) # with logit link
# Predictive summaries ...
    m.star[1:2] ~ dmnorm(mu[], sigma.inv[1:2 ,1:2] )
    x.star <- (m.star[1]+m.star[2])/2
    y.star <- (m.star[2]-m.star[1])/2
    se.star <- exp(x.star)/(1 + exp(x.star))
    sp.star <- 1 - exp(y.star) /(1 + exp(y.star))
# Variance covariance matrix for random effects...
    sigma[1:2, 1:2] <- inverse(sigma.inv[ , ])
}
```

To fit this model, we need to specify two R objects, one with the names of *the data* and another with the names of the parameters of interest, so in R we have:

```
> data <- list ("R", "tp", "n1", "fp", "n2","n")
> parameters <- c("sigma", "se", "sp", "se.star", "sp.star")
```

The function `bugs()` has a series of arguments that are needed to run BUGS:

```
ct.m1 <- bugs(data, inits=NULL, parameters, "model1.txt", n.chains = 1,
            n.iter = 20000,  n.thin=1, bugs.directory = bugsdir,
            working.directory = workdir, clearWD=TRUE, debug=FALSE)
```

The first argument refers to the data nodes, the second how initial values are generated (here `NULL` means that BUGS will generate these values randomly), `parameters` is the vector of parameters to monitor and `model1.txt` the BUGS model.

In this example, the argument `n.chains=1` indicates that we generate one chain, `n.inter = 20000` the length of the chain, by default the first `n.inter/2` iterations will be omitted for analysis. For more details see the help files of `bugs()`.

The resulting object `ct.m1` is an R object from the class `mcmc.list`, which can be analyzed using the package `coda` or manually as we do in this section. For example the `print()` function gives a summary of the object:

```
> print(ct.m1, digits.summary = 3)
Inference for Bugs model at "model1.txt", fit using WinBUGS,
 1 chains, each with 20000 iterations (first 10000 discarded)
 n.sims = 10000 iterations saved
               mean     sd    2.5%     25%     50%     75%   97.5%
sigma[1,1]    1.359  0.441   0.687   1.042   1.298   1.605   2.379
sigma[1,2]   -0.952  0.366  -1.827  -1.150  -0.913  -0.698  -0.371
sigma[2,1]   -0.952  0.366  -1.827  -1.150  -0.913  -0.698  -0.371
sigma[2,2]    1.522  0.471   0.805   1.191   1.455   1.774   2.644
se            0.955  0.005   0.946   0.952   0.956   0.959   0.965
sp            0.951  0.008   0.932   0.946   0.951   0.957   0.966
se.star       0.952  0.025   0.892   0.941   0.956   0.968   0.984
sp.star       0.923  0.086   0.680   0.904   0.950   0.976   0.995
mu[1]         6.046  0.212   5.636   5.906   6.046   6.186   6.469
mu[2]         0.099  0.216  -0.319  -0.046   0.098   0.245   0.533
m.star[1]     6.051  1.175   3.770   5.289   6.056   6.804   8.399
m.star[2]     0.135  1.263  -2.320  -0.706   0.132   0.975   2.626
deviance    372.057 14.097 344.800 362.300 371.800 381.025 401.002

DIC info (using the rule, pD = Dbar-Dhat)
pD = 54.4 and DIC = 426.5
```

The following lines show how to access *sensitivity* and *specificity* posterior distributions and plot them:

```
> sensitivity <- ct.m1$sims.array[,1,"se"]
> specificity <- ct.m1$sims.array[,1,"sp"]
> par(mfrow = c(1,2))
> hist(sensitivity, breaks=80, prob=T, main="",
xlab="sensitivity")
```

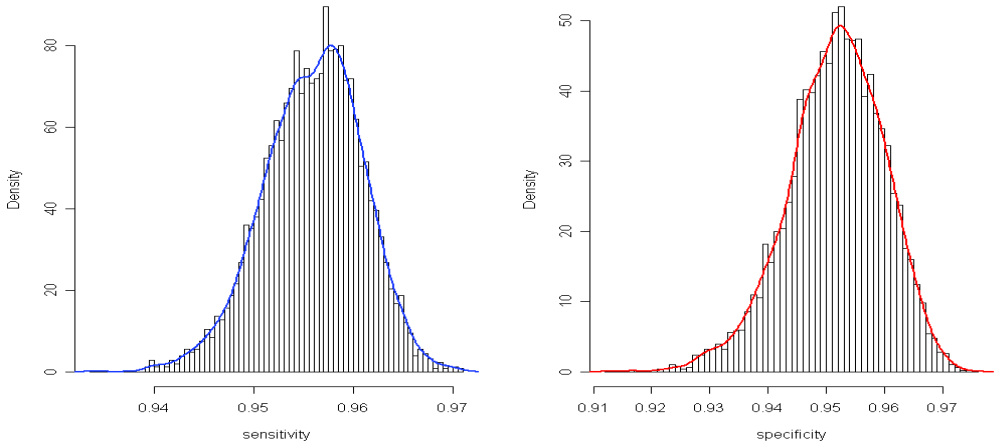

Figure 7.1: *Summary plot. Left panel: Posterior distribution for sensitivity. Right panel: Posterior distribution for specificity.*

```
> lines(density(sensitivity), lwd = 2, col ="blue")
> hist(specificity, breaks=80, prob=T, main="",
xlab="specificity")
> lines(density(specificity), lwd = 2, col ="red")
> par(mfrow = c(1,1))
```

Figure 7.1 displays the resulting histograms. For further analysis see the supplementary R script `meta-bayes-Rscript.r`.

7.4.2 Comments on BUG codes for diverse model extensions

In the Appendix B we give the BUGS code for each model fitted in Chapter 5 and Chapter 6. Here we give the description of these models with some comments on specific BUGS code.

Model 2

This model is similar to Model 1, but we use a complementary log log link function, which is implemented by changing two lines in the model:

```
    cloglog(tpr[i]) <- m[i,2]/2 + m[i,1]/2   # (Di + Si)/2
    cloglog(fpr[i]) <- m[i,2]/2 - m[i,1]/2   # (Di - Si)/2
```

Model 3

This model is Model 1, but we model the variance covariance matrix for the random effects as a function of the study design. We calculate the *Relative Credibility*. This structural variance model is implemented by defining the following random effects distribution:

```
 m[i,1:2] ~ dmnorm(mu[], sigma.inv[ design[i ],  1:2, 1:2])
```

with priors

```
# Priors ...
  sigma.inv[ 1,  1:2,1:2] ~ dwish(R[1:2,1:2], 3)
  sigma.inv[ 2, 1:2,1:2] ~ dwish(R[1:2,1:2], 3)
```

and new summary statistics

```
# Summary statistics
 sigma[1, 1:2, 1:2] <- inverse(sigma.inv[1, 1:2, 1:2]) # Retrospective design
 sigma[2, 1:2, 1:2] <- inverse(sigma.inv[2, 1:2, 1:2]) # Prospective design
 tr.Rt <- sigma[1, 1:1, 1:1] + sigma[1, 2:2, 2:2]
 tr.Pr <- sigma[2, 1:1, 1:1] + sigma[2, 2:2, 2:2]
 RC <- tr.Pr / tr.Rt                                   # Relative Credibility
```

The vector `design[i ]` has value 1 or 2 for retrospective and prospective design respectively. In order to run this model, we need the vector `design[i ]` to be visible in the R workspace. We need also to specify the data and nodes to monitor in R by:

```
data <- list ("R", "tp", "n1", "fp", "n2", "n", "design")
parameters <- c("RC", sigma)
```

Model 4

We extend Model 3 to have bivariate random effects with a *t*-distribution. This model is implemented with scale mixture approach, which is achieved by defining the following random effects distribution:

```
m[i,1:2] ~ dmnorm(mu.0[1:2 ], sigma.inv[ design[i ],  1:2, 1:2])
w[i] ~ dgamma(nu.2, nu.2) I(0.005, 3)
y[i, 1] <- mu[ 1] + m[i, 1] / sqrt(w[i])
y[i, 2] <- mu[ 2] + m[i, 2] / sqrt(w[i])
logit(tpr[i]) <- (y[i, 2] + y[i, 1])/2
logit(fpr[i]) <- (y[i, 2] - y[i, 1])/2
```

The prior distribution for the number of degrees of freedom is given by:

```
nu.2 <- nu/2
nu ~ dexp(eta) I(2, 100 )  # prior for df exponential eta
eta ~ dunif(0.02, 0.5)     # prior for eta 0.02 to 0.5, which
                           # implies df between 2 to 50
```

Model 5

This is a bivariate meta-regression model based on Model 4. This model implements the ridge regression for variable selection and allows to use covariates with missing data. It has three new blocks. One with the bivariate regression equation, one with the priors with the regression coefficients and another one with the sub-model for missing data:

```
# Regression equations ...
mu[i,1] <-  alpha0  + alpha[1, country[i]] + alpha[2,hosp[i]] +
alpha[3,inclus[i]] + alpha[4,indfind[i]] +  alpha[5,design[i] ] +
alpha[6,contr[i]] + alpha[7,localis[i]] + alpha[8, child[i]]

mu[i, 2] <- beta0 + beta[1, country[i]] + beta[2,hosp[i]] +
   beta[3,inclus[i]] + beta[4,indfind[i]] + beta[5,design[i] ] +
     beta[6,contr[i]] + beta[7,localis[i]] + beta[8, child[i]]
#...

# Prior distributions for regression coefficients ...
 for( i in 1:p) {alpha[i, 1] <- 0}
 for( i in 1:p) {
        alpha[i, 2] ~ dnorm(0.0, phi1)
        }

 for( i in 1:p) {beta[i, 1] <- 0}
 for( i in 1:p) {
        beta[i, 2] ~ dnorm(0.0, phi2)
```

```
        }
# Scale parameters ...
phi1 ~ dgamma(10, 5)
phi2 ~ dgamma(10, 5)
#...

# Sub-model for missing covariates ...
for(i in 1:n){
    fup.na[i] ~ dbern(p.fup); fup[i] <- fup.na[i] +1
    }
p.fup ~ dbeta(0.5, 0.5)
#...
```

Model 6

This is the same model as Model 5, but with SSVS method for variable selection. This part of the model is implemented with a mixture normal distribution for each regression coefficient:

```
# Priors for regression coefficients...
 for( i in 1:p) {
        g1[i]~ dbern(0.5)
        alpha[i, 2] ~ dnorm(0.0, tprior1[i] )
        tprior1[i] <-pow(100, 1-g1[i])*phi1
        }
 for( i in 1:p) {
        g2[i]~ dbern(0.5)
        beta[i, 2] ~ dnorm(0.0, tprior2[i])
        tprior2[i] <-pow(100, 1-g2[i])*phi2
        }
 phi1 ~ dgamma(10, 5)
 phi2 ~ dgamma(10, 5)
```

Summary remarks

"There must be, he thought, some key, some crack in this mystery he could use to achieve an answer."

-P.C. Doherty, *Crown in Darkness*

In this final chapter is summarized the results of this work and draws attention to points about further research work in this area.

As a general remark, either from the classical or Bayesian perspective, modern statistics provides important improvement in meta-analysis of diagnostic test. In the author's opinion this line of research will play a fundamental role in the upcoming problems of complex multiparameter meta-analysis (Ades and Sutton, 2006) [2].

In Chapter 1 we have reviewed some clear limitations of the the SROC curve. However, this method is usually applied in practice and complex statistics, like the area under the SROC curve, is currently reported as a summary of the meta-analysis (see Gatsonis and Paliwal (2006) [43]). In Chapter 3 we used bootstrap methods and variance stabilization techniques to improve statistical inference of the AUC. We have reviewed and illustrated several bootstrap techniques and proposed a specially designed bootstrap confidence interval for the AUC, based on a transformed version of the bootstrap-*t* interval.

These confidence intervals were evaluated with an extensive simulation experiment. This experiment shows that for meta-analysis with small number of studies, standard statistical methods perform poorly, second order accurate bootstrap methods (BC_a and bootstrap-*t*) deliver unstable results, but the transformed bootstrap-*t* approach is extremely effective.

Bootstrap methods in meta-analysis could be a clear area of further research. One possibility is to extend the methodology presented in Chapter 3 to make comparisons of AUC. For example, build confidence intervals for the difference of AUC of two

or more different diagnostic tests. What we do not recommend is to follow a bootstrap modeling approach (e.g. using GLMM) in this area, we have found that a direct Bayesian modeling is straightforward and numerically stable.

In Chapter 5 and Chapter 6 we have dealt with the problem of modeling multiple source of uncertainty in meta-analysis. A Bayesian model building approached was applied. The Bayesian paradigm has been applied with pragmatism, which includes the use of posterior model checking, building diagnostic plots and trying to understand deficits of the fitted model. The following are the main aspects of our approach:

- The bivariate model presented in Chapter 5 is simple, comprehensive and it is based on the classical SROC curve. However it improves upon the SROC model in all of its drawbacks.

- New Bayesian summaries statistics have been presented those include: the Bayesian SROC, the posterior distribution of the AUC and the Bayesian predictive surface. All are comparable with other classical approaches.

- In Chapter 5 the bivariate model was compared with the HSROC, an alternative full Bayesian approach. The HSROC produced very unstable results for our data. The bivariate model was, also, compared with other similar models implemented in SAS and Stata. We found no difference in results for large samples. For small samples, the Bayesian approach has a clear numerical advantage.

- Models in meta-analysis of diagnostic test are fitted *without any* sort of model checking or model criticism (e.g. link function, structural distribution, etc.). This is a dangerous practice, which could end up in fitting models that may poorly fit the data at hand. In our Bayesian approach we make particular emphasis in model checking and validation.

- Previous approaches of multivariate meta-analysis have used the normal distribution for random-effects exclusively. The SAS NLMIXED procedure does not allow the fitting of any other random-effect model. This work has shown that the normal distribution is not suitable for these data. We proposed to model random effects as a mixture of normals and mixtures of t-distributions. Important data characteristics are explained with these extensions, like changes of variability between study designs (*relative credibility*) and *relevance* of particular studies.

- Performing a meta-regression is conceptually simple, but operationally very complex. For example, a variable selection procedure should be included as

a model strategy and there are many ways to do this. Here we proposed two approaches one based on ridge regression and another based on SVSS. Although these techniques have been well studied in multiple linear regression (Ishwaran and Rao, 2005)[66]), their application in meta-regression or regression with missing data covariates remains a topic of research.

- MCMC computations have been done with `BUGS`, which assists us to perform complex Bayesian modeling. There may be room to improve sampling techniques in some specific situation, but we can report very stable and efficient results in our applications.

One problem that is still open in meta-analysis of diagnostic test is the assessment of publication bias. Deeks et al. (2005) [20] examined the limitation of current classical techniques (e.g. funnel plots) to detect publication bias in meta-analysis of diagnostic test. Interesting, we have found that studies with *lower relevance* are associated to studies with severe imbalance between the sample size of disease and non-disease groups in the study. We can conjecture that studies with lower relevance in the meta-analysis are associated with publication bias. It could be interesting to have access to more meta-analysis data and to investigate this result.

Combining studies, which reports a single threshold value is the most common case in meta-analysis. However, it could be interesting to extend our approach to the case of studies reporting multiple threshold values. Currently, the only approach presented in the literature is Dukic and Gatsonis (2003) [27], which is based on the HSROC curve model (Rutter and Gatsonis, 2001). These authors pointed out some limitations of this methodology, it could be interesting to develop an alternative approach based on our bivariate Bayesain model.

Finally in Chapter 7 we have explain details on the sophisticated software implementation of our methods. We plan to port these programs and other Bayesian method for meta-analysis to a free available package in R.

Appendix A: Posterior distributions for different model extensions

We have performed our Bayesian data analysis with Gibbs sampling implemented in BUGS, however, some people may be interested to develop their own sampling techniques. In this regard, we give the posteriors distributions that has been used in Chapter 5 and Chapter 6.

The posterior distribution for a model with studies with different designs is

$$p(\theta|y) \propto \prod_{i=1}^{N} \left[\binom{n_{i,1}}{tp_i} \text{TPR}_i^{tp_i} (1-\text{TPR}_i)^{(n_{i,1}-tp_i)} \binom{n_{i,2}}{fp_i} \text{FPR}_i^{fp_i} (1-\text{FPR}_i)^{(n_{i,2}-fp_i)} \right] \tag{7.1}$$

$$\times \prod_{j=1}^{2} \prod_{i=1}^{N} \left[\exp\left(-1/2 \left(D_{i,j} - \mu_D, S_{i,j} - \mu_S \right)^T \Lambda_j (D_{i,j} - \mu_D, S_{i,j} - \mu_S) \right) \right] \times |\Lambda_1|^{\frac{N}{2}} |\Lambda_2|^{\frac{N}{2}}$$

$$\times \exp\left[-1/2 \left(v_D \left(\mu_D - m_D \right)^2 + v_S \left(\mu_S - m_S \right)^2 \right) \right]$$

$$\times \frac{|\Lambda_1|^{(N-k-2)/2} \exp\left[-1/2 \, \text{trace}(\Lambda_1 R^{-1}) \right]}{2^{k(N-1)/2} \prod^{k(k-1)/4} |\Lambda_1|^{(N-1)/2} \prod_{i=1}^{N} \Gamma(1/2(N-i))}$$

$$\times \frac{|\Lambda_2|^{(N-k-2)/2} \exp\left[-1/2 \, \text{trace}(\Lambda_2 R^{-1}) \right]}{2^{k(N-1)/2} \prod^{k(k-1)/4} |\Lambda_2|^{(N-1)/2} \prod_{i=1}^{N} \Gamma(1/2(N-i))}.$$

For a model with Mixture-Multivariate-t distribution the posterior distribution is

$$p(\theta|y) \propto \prod_{i=1}^{N} \left[\binom{n_{i,1}}{tp_i} \text{TPR}_i^{tp_i} (1-\text{TPR}_i)^{(n_{i,1}-tp_i)} \binom{n_{i,2}}{fp_i} \text{FPR}_i^{fp_i} (1-\text{FPR}_i)^{(n_{i,2}-fp_i)} \right] \tag{7.2}$$

$$\times \prod_{j=1}^{2} \prod_{i=1}^{N} \left[\exp\left(-1/2 \left(D_{i,j} - \mu_D, S_{i,j} - \mu_S \right)^T w_{i,j} \Lambda_j (D_{i,j} - \mu_D, S_{i,j} - \mu_S) \right) \right] \times |\Lambda_1|^{\frac{N}{2}} |\Lambda_2|^{\frac{N}{2}}$$

$$\times \exp\left[-1/2 \left(v_D \left(\mu_D - m_D \right)^2 + v_S \left(\mu_S - m_S \right)^2 \right) \right]$$

$$\times \frac{|\Lambda_1|^{(N-k-2)/2} \exp\left[-1/2 \, \text{trace}(\Lambda_1 R^{-1}) \right]}{2^{k(N-1)/2} \prod^{k(k-1)/4} |\Lambda_1|^{(N-1)/2} \prod_{i=1}^{N} \Gamma(1/2(N-i))}$$

$$\times \frac{|\Lambda_2|^{(N-k-2)/2} \exp\left[-1/2 \, \text{trace}(\Lambda_2 R^{-1}) \right]}{2^{k(N-1)/2} \prod^{k(k-1)/4} |\Lambda_2|^{(N-1)/2} \prod_{i=1}^{N} \Gamma(1/2(N-i))}$$

$$\times \prod_{j=1}^{2} \prod_{i=1}^{N} w_{i,j}^{\nu/2-1} \exp\left[-\tau \, \nu \right] \times I(\tau)[0.01, 0.5].$$

Finally for a meta-regression model with ridge approach the posterior distribution results as

$$p(\theta|y) \propto \prod_{i=1}^{N} \left[\binom{n_{i,1}}{tp_i} \text{TPR}_i^{tp_i} (1-\text{TPR}_i)^{(n_{i,1}-tp_i)} \binom{n_{i,2}}{fp_i} \text{FPR}_i^{fp_i} (1-\text{FPR}_i)^{(n_{i,2}-fp_i)} \right] \tag{7.3}$$

$$\times \prod_{j=1}^{2} \prod_{i=1}^{N} \left[\exp\left(-1/2 \left(D_{i,j} - \mu_D, S_{i,j} - \mu_S \right)^T w_{i,j} \Lambda_j (D_{i,j} - \mu_D, S_{i,j} - \mu_S) \right) \right] \times |\Lambda_1|^{\frac{N}{2}} |\Lambda_2|^{\frac{N}{2}}$$

$$\times \exp\left[-1/2 \left(v_D \left(\mu_D - m_D \right)^2 + v_S \left(\mu_S - m_S \right)^2 \right) \right]$$

$$\times \frac{|\Lambda_1|^{(N-k-2)/2} \exp\left[-1/2 \, \text{trace}(\Lambda_1 R^{-1}) \right]}{2^{k(N-1)/2} \prod^{k(k-1)/4} |\Lambda_1|^{(N-1)/2} \prod_{i=1}^{N} \Gamma(1/2(N-i))}$$

$$\times \frac{|\Lambda_2|^{(N-k-2)/2} \exp\left[-1/2 \, \text{trace}(\Lambda_2 R^{-1}) \right]}{2^{k(N-1)/2} \prod^{k(k-1)/4} |\Lambda_2|^{(N-1)/2} \prod_{i=1}^{N} \Gamma(1/2(N-i))}$$

$$\times \prod_{j=1}^{2} \prod_{i=1}^{N} w_{i,j}^{\nu/2-1} \exp\left[-\tau \, \nu \right] \times I(\tau)[0.01, 0.5]$$

$$\times \exp\left[-1/2 \, \alpha_k^2 / \phi_1^2 \right] \, \exp\left[-1/2 \, \beta_k^2 / \phi_2^2 \right]$$

$$\times \phi_1^{m_\alpha - 1} \exp\left[-m_\alpha \phi_2 \right] \times \phi_2^{m_\beta - 1} \exp\left[-m_\beta \phi_2 \right].$$

Appendix B: Statistical models in BUGS language

```
#Model 1 ...
model
{
for( i in 1 : n ) {
tp[i] ~ dbin(tpr[i], n1[i])
fp[i] ~ dbin(fpr[i], n2[i])

   logit(tpr[i]) <- m[i,2]/2 + m[i,1]/2            # (Di + Si)/2
   logit(fpr[i]) <- m[i,2]/2 - m[i,1]/2            # (Di - Si)/2
  m[i,1:2] ~ dmnorm(mu[], sigma.inv[1:2 ,1:2 ] )
     }
#Priors simplest noninformative ...
 mu[1] ~ dnorm(0, 0.25)          # D
 mu[2] ~ dnorm(0, 0.25)          # S
 sigma.inv[1:2,1:2] ~ dwish(R[1:2,1:2], 3)
# Variance covariance matrix for random effects...
sigma.inv[1:2, 1:2] <- inverse(sigma[1:2 ,1:2 ])

# Summary statistics ...
# Pooled summaries...
x <- (mu[1]+mu[2])/2
y <- (mu[2]-mu[1])/2
se <- exp(x) / ( 1 + exp(x) )      # with logit link
sp <- 1 - exp(y) / ( 1 + exp(y) )  # with logit link

# Predictive summaries ...
m.star[1:2] ~ dmnorm(mu[], sigma.inv[1:2 ,1:2] )
x.star <- (m.star[1]+m.star[2])/2
y.star <- (m.star[2]-m.star[1])/2
se.star <- exp(x.star)/(1 + exp(x.star))      # with logit link
sp.star <- 1 - exp(y.star) /(1 + exp(y.star)) # with logit link

# Variance covariance matrix for random effects...
sigma[1:2, 1:2] <- inverse(sigma.inv[ , ])
}
#.............................................................
# Model 2 ...
model
{
for( i in 1 : n ) {
tp[i] ~ dbin(tpr[i], n1[i])
fp[i] ~ dbin(fpr[i], n2[i])
    cloglog(tpr[i]) <- m[i,1]/2 + m[i,2]/2      # (Di + Si)/2
    cloglog(fpr[i]) <- m[i,2]/2 - m[i,1]/2      # (Di - Si)/2
   m[i,1:2]~dmnorm(mu[], sigma.inv[1:2 ,1:2 ] )
     }
# Priors ...
```

```
 mu[1] ~ dnorm(0, 0.25)
 mu[2] ~ dnorm(0, 0.25)
 sigma.inv[1:2,1:2] ~ dwish(R[1:2,1:2], 3)

# Summary statistics ...
x <- (mu[1]+mu[2])/2
y <- (mu[2]-mu[1])/2
se <- 1 - exp(-1*exp(x))                # with cloglog link
sp <- exp(-1*exp(y))                    # with cloglog link

m.star[1:2] ~ dmnorm(mu[], sigma.inv[1:2 ,1:2 ] )
x.star <- (m.star[1] + m.star[2])/2
y.star <- (m.star[2] - m.star[1])/2
se.star <- 1 - exp(-1*exp(x.star ))      # with logit link
sp.star <- exp(-1*exp(y.star ))          # with logit link

sigma[1:2, 1:2] <- inverse(sigma.inv[ , ])
}
#.................................................................
# Bivariate model ...
# This model is compared to the classical bivariate ...

model
{
for( i in 1 : n ) {
tp[i] ~ dbin(tpr[i], n1[i])
tn[i] ~ dbin(tnr[i], n2[i])

   logit(tpr[i]) <- m[i,1]
   logit(tnr[i]) <- m[i,2]
  m[i,1:2] ~ dmnorm(mu[], sigma.inv[1:2 ,1:2 ] )
     }
# Priors ...
 mu[1] ~ dnorm(0, 0.25)
 mu[2] ~ dnorm(0, 0.25)
 sigma.inv[1:2,1:2] ~ dwish(R[1:2,1:2], 3)

# Variance covariance matrix for random effects
sigma[1:2, 1:2] <- inverse(sigma.inv[ , ])

corr <- sigma[1,2]/(pow(sigma[1,1],0.5)*pow(sigma[2,2],0.5))
}
#.................................................................
# HSROC ...
# Statistics in Medicine
# 2001, 20; 2865-2884
# A hierarchical regression approach to meta-analysis of diagnostic
# test accuracy evalations
# Rutter C.M. and Gatsonis C.A.

model
{
# priors ...
 THETA~dunif(-10,10)
 LAMBDA~dunif(-2,20)
 beta~dunif(-5,5)
 prec1 ~ dgamma(2.1, 2)
 prec2 ~ dgamma(2.1, 2)
 sigmaq1 <- 1.0/prec1
 sigmaq2 <- 1.0/prec2
 b <- exp(beta/2)

# data model ...
for(i in 1:n){
  theta[i] ~ dnorm(THETA, prec1)
```

```
  alpha[i] ~ dnorm(LAMBDA, prec2)
  logit(tpr[i]) <- (theta[i] + 0.5*alpha[i])/b
  logit(fpr[i]) <- (theta[i] - 0.5*alpha[i])*b
  tp[i] ~ dbin(tpr[i], n1[i])
  fp[i] ~ dbin(fpr[i], n2[i])
}
}
#.............................................................
# Model 3 ...
model
{
for( i in 1 : n ) {
tp[i] ~ dbin(tpr[i], n1[i])
    fp[i] ~ dbin(fpr[i], n2[i])
    logit(tpr[i]) <- m[i,2]/2 + m[i,1]/2        # (Di + Si)/2
    logit(fpr[i]) <- m[i,2]/2 - m[i,1]/2        # (Di - Si)/2
    m[i,1:2] ~ dmnorm(mu[], sigma.inv[ design[i ],  1:2, 1:2])
     }
# Priors ...
  mu[1] ~ dnorm(0, 0.25)
  mu[2] ~ dnorm(0, 0.25)
  sigma.inv[ 1,  1:2,1:2] ~ dwish(R[1:2,1:2], 3)
  sigma.inv[ 2, 1:2,1:2] ~ dwish(R[1:2,1:2], 3)

# Summary statistics
 sigma[1, 1:2, 1:2] <- inverse(sigma.inv[1, 1:2, 1:2]) # Retrospective design
 sigma[2, 1:2, 1:2] <- inverse(sigma.inv[2, 1:2, 1:2]) # Prospective design

 tr.Rt <- sigma[1, 1:1, 1:1] + sigma[1, 2:2, 2:2]
 tr.Pr <- sigma[2, 1:1, 1:1] + sigma[2, 2:2, 2:2]
 RC <- tr.Pr / tr.Rt                                  # Relative Credibility
 }
#.............................................................
# Model 4 ...
model
{
for( i in 1 : n ) {
       tp[i] ~ dbin(tpr[i], n1[i]);  fp[i] ~ dbin(fpr[i], n2[i])
        m[i,1:2] ~ dmnorm(mu.0[1:2 ], sigma.inv[ design[i ],  1:2, 1:2])
     w[i] ~ dgamma(nu.2, nu.2) I(0.005, 3)
     y[i, 1] <- mu[ 1] + m[i, 1] / sqrt(w[i])
  y[i, 2] <- mu[ 2] + m[i, 2] / sqrt(w[i])
  logit(tpr[i]) <- (y[i, 1] + y[i, 2])/2
          logit(fpr[i]) <- (y[i, 2] - y[i, 1])/2
}

# Priors ...
  mu[1] ~ dnorm(0, 0.25)
  mu[2] ~ dnorm(0, 0.25)
  mu.0[1] <- 0
  mu.0[2] <- 0
  nu.2 <- nu/2
  nu ~ dexp(eta) I(2, 100 )  # prior for df exponential eta
  eta ~ dunif(0.02, 0.5)     # prior for eta 0.02 to 0.5 implies df between 2 to 50

 sigma.inv[ 1,  1:2,1:2] ~ dwish(R[1:2,1:2], 3)
 sigma.inv[ 2, 1:2,1:2] ~ dwish(R[1:2,1:2], 3)

 sigma[1, 1:2, 1:2] <- inverse(sigma.inv[1, 1:2, 1:2])  # Retrospective design
 sigma[2, 1:2, 1:2] <- inverse(sigma.inv[2, 1:2, 1:2])  # Prospective design

#  t-distribution with nu df

for(i in 1:2){
for(j in 1:2){
```

```
  sigma.t[2, i, j] <-  nu / (nu - 2)  * sigma[2, i, j]
  sigma.t[1, i, j] <-  nu / (nu - 2)  * sigma[1, i, j]
}
}

# Summary statistics
trace.R <- sigma.t[1, 1:1, 1:1] + sigma.t[1, 2:2, 2:2]
trace.P <- sigma.t[2, 1:1, 1:1] + sigma.t[2, 2:2, 2:2]

RC <-  trace.P/trace.R
}
#.............................................................
# Model 5 ...
model
{
    for( i in 1 : n ) {
    tp[i] ~ dbin(tpr[i], n1[i]); fp[i] ~ dbin(fpr[i], n2[i])
    m[i,1:2] ~ dmt(mu[i, 1:2], sigma.inv[ design[i ],  1:2, 1:2], nu)
x.s[i] <- (m[i, 1] + m[i, 2])/2
    y.s[i] <- (m[i, 2] - m[i, 1])/2
    logit(tpr[i]) <- x.s[i]
    logit(fpr[i]) <- y.s[i]
mu[i,1] <-  alpha0  + alpha[1, country[i]] + alpha[2,hosp[i]] +
alpha[3,inclus[i]] + alpha[4,indfind[i]] +  alpha[5,design[i] ] +
alpha[6,contr[i]] + alpha[7,localis[i]] + alpha[8, child[i]]

mu[i, 2] <- beta0 + beta[1, country[i]] + beta[2,hosp[i]] +
   beta[3,inclus[i]] + beta[4,indfind[i]] + beta[5,design[i] ] +
     beta[6,contr[i]] + beta[7,localis[i]] + beta[8, child[i]]
 }

# Priors ...
# Regression model ...
 alpha0 ~ dnorm(0, 0.025)
 beta0 ~ dnorm(0, 0.025)
 for( i in 1:p) {alpha[i, 1] <- 0}
 for( i in 1:p) {alpha[i, 2] ~ dnorm(0.0, phi1)}
 for( i in 1:p) {beta[i, 1] <- 0}
 for( i in 1:p) {beta[i, 2] ~ dnorm(0.0, phi2)}
 phi1 ~ dgamma(10, 5)
 phi2 ~ dgamma(10, 5)
# degrees of freedom t-distribution
  nu.2 <- nu/2
  nu ~ dexp(eta) I(2, 100 )# prior for df exponential eta
  eta ~ dunif(0.02, 0.5)   # prior for eta 0.02 to 0.5 implies df between 2 to 50

# variance covariance matrix
 sigma.inv[ 1,  1:2,1:2] ~ dwish(R[1:2,1:2], 3)
 sigma.inv[ 2, 1:2,1:2] ~ dwish(R[1:2,1:2], 3)
}
#.............................................................
# Model 6 ...
model
{
 for( i in 1 : n ) {
 tp[i] ~ dbin(tpr[i], n1[i]); fp[i] ~ dbin(fpr[i], n2[i])
 m[i,1:2] ~ dmt(mu[i, 1:2], sigma.inv[ design[i ],  1:2, 1:2], nu)
 x.s[i] <- (m[i, 1] + m[i, 2])/2
 y.s[i] <- (m[i, 2] - m[i, 1])/2
 logit(tpr[i]) <- x.s[i]
 logit(fpr[i]) <- y.s[i]
 mu[i,1] <- alpha0 + alpha[1, country[i]] + alpha[2,hosp[i]] +
           alpha[3,inclus[i]]  + alpha[4,indfind[i]] +
           alpha[5,design[i]] + alpha[6,contr[i]] +  alpha[7,localis[i]] +
           alpha[8, child[i]] + alpha[9,fup[i]] + alpha[10,refer[i]] +
```

```
          alpha[11,sample[i]] + alpha[12,gender[i]]
 mu[i, 2] <- beta0 + beta[1, country[i]] + beta[2,hosp[i]] + beta[3,inclus[i]]
          + beta[4,indfind[i]] + beta[5,design[i] ] +
       beta[6,contr[i]] + beta[7,localis[i]] + beta[8, child[i]] +
           beta[9,fup[i]] + beta[10,refer[i]] +
           beta[11,sample[i]] + beta[12,gender[i]]
 }
# Priors ...
# Regression model ...
 alpha0 ~ dnorm(0, 0.01)
 beta0 ~ dnorm(0, 0.01)
 for( i in 1:p) {alpha[i, 1] <- 0}
 for( i in 1:p) {
       g1[i]~ dbern(0.5)
       alpha[i, 2] ~ dnorm(0.0, tprior1[i] )
       tprior1[i] <-pow(100, 1-g1[i])*phi1
       }

 for( i in 1:p) {beta[i, 1] <- 0}
 for( i in 1:p) {
       g2[i]~ dbern(0.5)
       beta[i, 2] ~ dnorm(0.0, tprior2[i])
       tprior2[i] <-pow(100, 1-g2[i])*phi2
       }
 phi1 ~ dgamma(10, 5)
 phi2 ~ dgamma(10, 5)

# Missing data models ...
for(i in 1:n){fup.na[i] ~ dbern(p.fup); fup[i] <- fup.na[i] +1 }
p.fup ~ dbeta(0.5, 0.5)
#
for(i in 1:n){refer.na[i] ~ dbern(p.refer); refer[i] <- refer.na[i] +1 }
p.refer ~ dbeta(0.5, 0.5)
#
for(i in 1:n){sample.na[i] ~ dbern(p.sample); sample[i] <- sample.na[i] +1 }
p.sample ~ dbeta(0.5, 0.5)
#
for(i in 1:n){gender.na[i] ~ dbern(p.gender); gender[i] <- gender.na[i] +1 }
p.gender ~ dbeta(0.5, 0.5)

# degrees of freedom t-distribution
nu.2 <- nu/2
nu ~ dexp(eta) I(2, 100) # prior for df exponential eta
eta ~ dunif(0.02, 0.5)   # prior for eta 0.02 to 0.5 implies df between 2 to 50

# variance covariance matrix

 sigma.inv[ 1,  1:2,1:2] ~ dwish(R[1:2,1:2], 3)
 sigma.inv[ 2, 1:2,1:2] ~ dwish(R[1:2,1:2], 3)
}
```

Bibliography

[1] ADDISS, D. G., SHAFFER, N., FOWLER, B. S. AND TAUXE, R. V. (1990). The epidemiology of appendicitis and appendectomy in the United States. *Am J Epidemiol.* **5**, 910–25.

[2] ADES, A. E. AND SUTTON, A. J. (2006) Multiparameter evidence synthesis in epidemilology and medical decision-making: current approaches. *JRSS, A.* **169**, 5–35.

[3] ARENDS, L. R. (2006) *Multivariate meta-analysis: Modeling the heterogeneity. Mixing apples and oranges: Dangerous or delicious? Rotterdam: Erasmus University Rotterdam*, Unpublished doctoral dissertation.

[4] BABU, G.J. AND SINGH, K. (1983). Inference on means using the bootstrap. *Ann. Statisti.* **11** 999-1003.

[5] BROOKS, S. (1998). Markov chain Monte Carlo method and its application. *The Statist.* **47** 69-100.

[6] BOSSUYT, P. M., REITSMA, J. B., BRUNS, D. E., GATSONIS, C. A., GLASZIOU, P. P., IRWIG, L. M., LIJMER, J. G., MOHER, D., RENNIE, D. AND DE VET, H. C. W. (2003) Towards complete and accurate reporting of studies of diangostic accuracy: the STARD initiative *BMJ.* **326**, 41–44.

[7] BOX, G. E. P. AND TIAO, G. C. (1973) *Bayesian Inference in Statistical Analysis.* John Wiley and Sons, Inc. pag. 34–36

[8] BOX G.E.P. (1980). Sampling and Bayes inferece in scientif modelling and robustness. *Journal of the Royal Statistical Society* A **143**, 383-430.

[9] BROOKS, S.P. AND GELMAN, A. (1997) General methods for monitoring convergence of iterative simulations. *J. Comp. Graph. Statisti.*, 7, 434-455.

[10] BROOKS, S.P. AND ROBERTS, G.O. (1998). General methods for monitoring convergence of iterative simulations. *J. Comp. Graph. Statisti.*, 7, 434-455.

[11] BUSH, R. R. AND MOSTELLER, F. (1955) *Stochastic Models for Learning.* New York: Wiley.

[12] CHU, H. AND COLE, S. R. (2006) Bivariate meta-analysis of sensitivity and specificity with sparce data: a generalized linear mixed model approach. *Journal of Clincial Epidemiology.* **59**, 1331–1332.

[13] CONG X., COX D.D. AND CANTOR S.B.(2007) Bayesian meta-analysis of Papanicolaou smear accuracy. *Gynecol Oncol.* **107**, 133-137.

[14] CONGDON, P. (2006) *Bayesian Statistical Modelling.* Wiley Series in Probability and Statistics, pag. 373–374.

[15] COX, D. R. AND SNELL, E. J. (1989) *Analysis of Binary Data, Second Edition.* Chapman and Hall/CRC pag. 76–80.

[16] COX, D. R. (1999) Discussion of "Some statistical heresies (Lindsey)". *The Statistician*, **48**, 30. pag. 76–80.

[17] COWLES, M.K. AND CARLIN, B.P. (1996) Markov chain Monte Carlo convergence diagnostics: a comparative review. *JASA*, 91, 83-904.

[18] DAVISON, A.C. (2003) *Statistical Models* Cambridge Series in Statistical and Probilistic Mathematics. Cambridge University Press.

[19] DAVISON, A.C. AND HINKLEY, A.V. (1997) *Bootstrap Methods and thier Application* Cambridge Series in Statistical and Probilistic Mathematics. Cambridge University Press.

[20] DEEKS, J.J., MACASKILL P. AND IRWIG L. (2005) The performance of test of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol.* Sep, 58(9): 865-6.

[21] DE FINETTI, B. (1937) "La Prevision: Ses Lois Logiques, ses Sources Subjectives", *Ann. Inst. H. Poincare* 7, 1. English translation in *Studies in Subjective Probability*, H. E. Kyburg, Jr. and H. G. Smokler (editors), 1964, New York : Wiley. *Proc. 4th Berkley Symp. Math. Stat. Probab.* **1**, 199–210.

[22] DE FINETTI, B. (1961) The Bayesian approach to the rejection of outliers. *Proc. 4th Berkley Symp. Math. Stat. Probab.* **1**, 199–210.

[23] DELLAPORTAS, P., FORSTER, J.J AND NTZOUFRAS, I. (2000) *Bayesian variable selection using the Gibbs sampler, Generalized Linear Models: A Bayesian Perspective* (D.K. Dey, S. Ghosh, and B. Mallick, eds.) New York: Marcel Dekker, 271-286.

[24] DELLAPORTAS, P., FORSTER, J.J AND NTZOUFRAS, I. (2002) On Bayesian Model and Variable Selection Using MCMC, *Statistics and Computing*, 12, 27-36.

[25] DROITCOUR, J., SILBERMAN, G. AND CHELIMSKY, E. (1993) Cross-design synthesis: a new form of meta-analysis for combining results from randomised clinical trials and medicalpractice databases. *International Journal of Technology Assessment in Health Care.* **9**, 440–9.

[26] DIACONIS, P. AND EFRON, B. (1983)Computer Intensive Methods in Statistics. *Scientific American.***248**, 116-130.

[27] DUKIC, V. AND GATSONIS, C. (2003) Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. *Biometrics.***59**,93646.

[28] DICICCIO T.J. AND EFRON B. (1992) More accurate confidence intervals in exponential famlies. *Biometrika.* **79**, 2, 231-45.

[29] DICICCIO T.J. AND EFRON B. (1996) Bootstrap Confidence Intervals. *Statistical Science.* **11**, 3, 189-228.

[30] DICICCIO T.J. AND ROMANO J.P. (1990) Nonparametric confidence limits by resampling methods and least favorable families. *International Statistical Review*, **58**, 1, pp. 59-76.

[31] DICICCIO T.J. AND ROMANO J.P. (1995) On bootstrap procedures for second-order accurate confidance limits in parametric models. *Statist Sinica*, **5**, 1, pp. 141-160.

[32] DiCiccio T.J., Monti A.C. and Young G.A (2006) Variance stabilization for a scalar parameter. *J.R.Statist.Soc*B, **68**, Part 2, pp. 281-303.

[33] Eddy, D. M., Hasselblad, V. and Shachter, R. (1992) *Meta-analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence.* Academic Press, San Diego, CA.

[34] Efron, B. (1996)Empirical Bayes Methods for Combining Likelihoods (with discussion). *JASA.***91**, 434,538–565.

[35] Efron, B.(1979)Bootstrap methods: another look at the jackknife. *Annals of Statistics.* **7**, 1-16.

[36] Efron, B.(1987)Better boostrap confidence intervals (with Discussion). *Journal of the American Statistical Association.* **82**, 171-200.

[37] Efron, B.(1992)Jackknife-after-bootstrap standard errors and influence functions (with Discussion). *Journal of the Royal Statistical Society series B* **54**, 83-127.

[38] Efron, B.(1993)Bayes and likelihood calculations from confidence intervals. *Biometrika* **80**, 3-26.

[39] Efron, B. and Tibshirani, R. R.(1993)*An Introduction to the Bootstrap.* New York: Chapman and Hall.

[40] Flum, D. R. and Koepsell T. (2002)The clinical and economic correlates of misdiagnosed appendicitis: nationwide analysis.*Arch Surg.* **137(7)**,799–804.

[41] Field C. A. and Welsh A.H. (2007) Bootstrapping clustered data. *J.R. Statist. Soc. B*, **69**, Part 3, pp. 369-390.

[42] Fuller, W. A. (1987) *Mesurement Error Models.* New York: Wiley pag. 3–4.

[43] Gatsonis C. and Paliwal P. (2006). Meta-Analysis of Diagnostic and Screening Test Accuracy Evaluations: Methodologic Primer. *AJR*; 187: 271-281.

[44] Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *JASA.* **85**, 398–409.

[45] Gelfand, A. E., Dey, D. K. and Chang, H. (1992). Model determination using predictive distributions, with implementation via sampling-based methods. In *Bayesian Statiscs* 4,147-167, Oxford University Press.

[46] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004) *Bayesian Data Analysis, Second edition.* Chapman and Hall/CRC. pag. 488-491.

[47] Gelman, A. and Hill, J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press pag. 513–527.

[48] Gelman, A. and Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* 7, 457-511.

[49] George, E.I. and McCullach, R.R. (1993) Variable selection via Gibbs sampling, *JASA*, 88, 711-732.

[50] Geweke, J. (1993) Bayesian treatment of the independent Student-t linear model. *J. Appl. Econometrics.* **8S**, 19–40

[51] GILKS, W. R. AND WILD, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Applied Statistics.* **41**, 337–348.

[52] GILKS W, CLAYTON, D. G., SPIEGELHALTER D., BEST N.G., MCNEIL, A.J. SHARPLES, L.D. AND KIRBY, A.J. (1993) Modelling complexity: applications of Gibbs sampling in medicine (with discussion). *J.R. Statist. Soc.*, B, 55, 39-102.

[53] GILKS W, RICHARDSON S. AND SPIEGELHALTER D. (1996) *Markov chain Monte Carlo in Practice*, Chapman and Hall, London.

[54] GREEN, D.M. AND SWETS, J.M. (1966). Signal detection theory and psychophysics. New York: John Wiley and Sons Inc.

[55] GUTTMAN. I. (1967). The use of the concept of a future observation in goodness-of-fit problems. J.Roy. Statist. Soc. Ser. B 29, 83-100.

[56] HALL, P. (1988). Theoretical comparison of bootstrap confidence intervals (with discussion). *Ann. Statist.* **16**, 927-985.

[57] HALL, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer Series in Statistics.

[58] HANLEY, J. (1998). Receiver operating characteristic curves. In *Encyclopedia of Biostatistics*, P. Armitage and T. Colton (ed), 3738-3745. New York: Wiley.

[59] HARBORD, R. M., DEEKS, J. J., EGGER, M., WHITING, P. AND STERNE, J. A. (2007) A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics.* **8**, 239–251

[60] HOLMES S., MORRIS C. AND TIBSHIRANI R. (2003) Bradley Efron: A Conversation with Good Friends. *Statistical Science*, Vol **18**, No, 2, 268-281.

[61] HASTIE, T., TIBSHIRANI, R. and FRIDMAN J. (2001) *The Elements of Statistical Learning. Data Mining, Inference and Prediction.* Springer Series in Statistics.

[62] Macaskill P. (2004) Empirical Bayes estiamtes generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *J Clin Epidemiol.* **57(9)**, 925-32.

[63] MOOD A. M., GRAYBILL F. A. AND BOES D. C. (1973) *Introduction to the Theory of Statistics*. Third Edition. McGraw-Hill, Inc.

[64] HASTINGS, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika.* **57**, 97–109

[65] IRWIG, L., MACASKILL, P., GLASZIOU, P. AND FAHEY, M. (1995) Meta-analytic methos for diagnostic test accuracy, *J Clin Epidemiol.* **48**, 119–130.

[66] ISHWARAN, H. AND RAO, J. S. (2005) Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics.* **33**, 730–773.

[67] JANSSEN A. AND PAULS T.(2003) How do bootstrap and permutaion tests works?. *The Annals of Statistics.* **31**,No.3, 786–806.

[68] KASS, R.E., CARLIN, B.P., GELMAN, A., AND NEAL, R. (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *The American Statistician*, 52, 93-100.

[69] KASS, R. E. AND WASSERMAN L. (1994) Formal Rules for Selecting Prior Distributions: A Review and Annotated Bibilography. Tecnical Report, Department of Statistics, Cargnegie Mellon University, Pirrsburgh, Pennsylvania 15213-2717.

[70] KNOTTNERUS J.A., ED. (2002) *The evidence base of clicial diagnostis.* London, UK: BMJ Brooks.

[71] KRAEMER, M., OHMANN, C., LEPPERT, R. AND YANG, Q. (2000) Macroscopic assessment of the appendix at diagnostic laparoscopy is reliable. *Surg Endosc.* **7**, 625–33.

[72] KYBURG, H. E. JR, AND SMOKLER H. G. (1964) *Studies in Subjective Probability*, New York: Wiley.

[73] LAMBERT P.C., SUTTON A. J., BURTON P.R. ABRAMS K.R. AND JONES D.R. (2005) How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statist. Med.*, **24**, 2401-2428.

[74] LANG, J. B. (1999) Bayesian ordinal and binary regression models with a parametric family of mixture links. *Computational Statistics and Data Analyis.* **31**, 59–87.

[75] LAURITZEN, S. L., DAWID, A. P., LARSEN, B. N. AND LEIMER, H. G. (1990) Independence properties of directed Markov fields. *Networks.* **20**, 491–505.

[76] LEE, Y. AND NELDER, J. A. (2002) Analysis of ulcer data using hierarchical generalized linear models. *Statist. Med.* **21**, 191–202.

[77] LEE, Y., NELDER, J. A. AND PAWITAN Y. (2006) *Generalized Linear Models with Random Effects Unified Analysis via H-likelihood.* Chapman & Hall/CRC. Taylor & Francis Group. Boca Raton, FL, USA.

[78] LESAFFRE, E. AND SPIESSENS, B. (2001) On the effect of the number of quadrature points in a logistic random effects model: an example. *JRSS, C (Applied Statistics).* **50(3)**, 325-335.

[79] LUNN, D. J., THOMAS, A., BEST, N. AND SPIEGELHALTER, D. J. (2000) WinBUGS - A Bayesain modelling framework: Concepts, structure and extensibility. *Statistics and Computing.* **10**, 325–337.

[80] LIJMER, J. G., MOL, B. W., HEISTERKAMP, S., BONSEL, G. J., PRINS, M. H., VAN DER MEULEN, J. H. P. AND BOSSUYT, P. M. M. (1999) Empirical evidence of design-related bias in studies of diagnostic test.*JAMA.***282**,1061-1066.

[81] LIJMER, J. G., BOSSUYT, P. M. M. AND HEISTERKAMP, S. H. (2002) Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Statistics in Medicine.* **21**,1525-37.

[82] MALLICK, B. AND GELFAND, A. E. (1994) Generalized linear models with unknown link function. *Biometrika.* **81**, 237–245.

[83] MACASKILL P. (2004) Empirical Bayes estiamtes generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *J Clin Epidemiol.*, **57(9)**: 925-32.

[84] MAMMEN, E..(1992) When Does Bootstrap Work? *Asymptotic Results and Simulations.* Vol 77 of *Lecture Notes in Statistics*. New York: Springer.

[85] MAMMEN, E..(1993) Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics.* **21**255-285.

[86] METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. AND TELLER, E. (1953) Equation of state calculations by fast computing machine. *Journal of Chemical Physics.* **21**, 1087–91.

[87] MOSES L. E., SHAPIRO D. AND LITTENBERG B. (1993) Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Statistics in Medicine.* **12**, 1293-316.

[88] NEAL, R.M. (2003), Slice sampling, *The Annals of Statistcs.* **31**, 705–741.

[89] NTZOUGRAS I. (2002) Gibbs Variable Selection Using BUGS. *Journal of Stat. Software*, 7, 7.

[90] OHMANN, C., VERDE, P. E., GILBERS, T., FRANKE, C., FUERST, G., SAUERLAND, S. AND BOEHNER, H. (2006) Systematic Review of CT Investigation in Suspected Acute Appendicitis.Final report *Coordination Centre for Clinical Trials, Heinrich-Heine University. Moorenstr. 5, D-40225 Duesseldorf Germany*

[91] PLUMMER, M., BEST, N., COWLES, K. AND VINES, K. (2007). `coda`: Output analysis and diagnostics for MCMC. R package version 0.12-1.

[92] RAMSEY, F. P.(1931) *The Fundation of Mathematics and Other Logical Essays*, London: Routledge and Kegan Paul.

[93] R DEVELOPMENT CORE TEAM (2007) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing ISBN 3-900051-07-0. `url = http://www.R-project.org`

[94] REITSMA, J. B., GLAS, A. S., RUTJES, A. W. S., SCHOLTEN, R. J, BOSSUYT, P. M. AND ZWINDERMAN, A. H. (2005) Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology.* **58**, 982-90

[95] RIPLEY, B. D. (1987) *Stochastic Simulation.* Wiley, New York.

[96] RIPLEY, B.D. (2004). Selecting amongst large classes of models. *Methods and Models in Statistics: In Honour of Professor John Nelder, FRS.* Imperial College Press.

[97] RUBIN, D. B. (1981). Estimation in parallel randomized experimetns. *J. Educ. Statisti.* **6**, 377-400.

[98] RUBIN, D. B. (1984). Byesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12**, 1151-1172.

[99] RUTTER, C. M. AND GATSONIS, C. A. (2001) A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine.* **20**, 2865-84.

[100] SAS INSTITUTE INC. (2003) The SAS System for Windows. Version 9.1 *SAS Institute Inc, Cary, NC.*

[101] SAVAGE, L. J. (1954), The Foundation of Statistics, New York: Wiley.

[102] SAVAGE, L. J. (1961a), "The Subject Basis of Statistical Practice," unpublished manuscript, the University of Michigan.

[103] SAVAGE, L. J. (1961b), The Subject Basis of Statistics Reconsidered," Proc. 4th Berkeley Symp. 1, 575.

[104] SCHENKER, N. (1985), Qualms about bootstrap confidence intervals. *Journal of the American Statistical Association*, **80**, 360–361.

[105] SMIDT, N., RUTJES, A. W. S., VAN DER WINDT, D. A. W., OSTELO, R. W. J., REITSMA, J. B., BOSSUYT, P. M., BOSSUYT, P. M., BOUTER, L. M. AND DE VET, H. C. W (2005) Quality of reporting of diagnostic accuracy studies. *Radiology.* **235**, 347–353.

[106] SPIEGELHALTER, D. J.(1998) Bayesian graphical modelling: a case-study in monitoring health outcomes. *Appl. Statist* **47**, Part 1, pp. 115-133.

[107] SPIEGELHALTER, D. J.(2006) Two brief topics on modelling with WinBUGS. Presented at IceBUGS conference. Available at: www.math.helsinki.fi/openbugs/IceBUGS/Presentations.

[108] SPIGELHALTER, D. J., ABRAMS, K. R. AND MYLES, J. P. (2004) *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* John Wiley & Sons, Ltd. Chapter 8.

[109] SPIEGELHALTER D.J., BEST N.G., CARLIN B.P. AND VAN DER LINDE A.(2002) Bayesian measures of model complexity and fit (with discussion). *J. Roy. Statist. Soc.* B. 64, 583-640.

[110] SPIEGELHALTER, D. J., THOMAS, A. AND BEST, N.(2004) *WinBUGS, Version 1.4, Upgraded to 1.4.1, User Manual*. MRC Biostatistics Unit: Cambridge.

[111] STEIN, C. (1956). "Efficient Nonparametric Testing and Estimation", *In Proceedings of the Third Berkeley Symposium*, Berkeley: University of California Press, pp. 187-196.

[112] STURTZ, S., LIGGES, U., AND GELMAN, A. (2005). `R2WinBUGS`: A Package for Running WinBUGS from R. Journal of Statistical Software, **12(3)**, 1-16.

[113] SWEETING, M. J., SUTTON, A. J. AND LAMBERT, P. C. (2004) What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine.* **23**, 1351-1375.

[114] THE COCHRANE METHODS GROUP ON SYSTEMATIC REVIEW OF SCREENING AND DIAGNOSTIC TESTS (2005) Recommended methods: Screening and diagnostic tests. Available at: www.cochrane.org/cochrane/sadtdoc1.htm.

[115] TIBSHIRANI, R. (1988) Variance statibilization and the bootstrap. *Biometika* **75**, 433-444.

[116] TOSTESON, A. N. AND BEGG, C. B.(1988) A general regression methodology for ROC curve estimation. *Medical Decistion Making.* **8**, 204–215.

[117] VAN HOUWELINGEN, H. C., ARENDS, L. R. AND STIJNEN, T. (2002) Tutorial in biostatistics. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine.* **21**, 589–624.

[118] VENABLES W.N. AND RIPLEY B.D. (2002) *Modern Applied Statistics with S. Forth edition.* Springer-Verlag New York, Inc.

[119] VERDE, P. E. (2005). Meta-analysis for diagnostic test data. International Biometrical Society, Argentine Region. Invited presentation available at: http://www.unlu.edu.ar/ gab/

[120] VERDE, P. E. (2005). Meta-Analysis for Diagnostic Test Data: a Bayesian Approach. International Biometrical Society, German Region. Bayesian Methods Working Group. Presentation available at: http://ibealt.web.med.uni-muenchen.de/bayes-ag//

[121] VERDE, P. E. (2006). Generalized Evidence Synthesis for Diagnostic Test Data. IceBUGS 2006. Presentation available at:
http://www.math.helsinki.fi/openbugs/IceBUGS/IceBUGSAbstracts.html

[122] VERDE, P. E. (2007).Modern meta-analysis: a case study in combining results of diagnostic test data. Invited presentation at The Department of Statistics, Stanford University. Statistics Seminars 2006-2007, available at: http://www-stat.stanford.edu/seminars/stat/"

[123] VERDE, P. E. (2008). Modern meta-analysis: a case study in combining results of diagnostic test data. *Unpublished manuscript*.

[124] VERDE, P. E. (2008). Meta-analysis of diagnsotic tests and the bootstrap. *Unpublished manuscript*. In preparation.

[125] VERDE, P. E. and Ohmann C. (2007). Meta-regression for diagnostic test studies in presence of missing reporting data: a Bayesian approach. Kongress Medizin und Gesellschaft. Augsburg 17. -21 September 2007. Abstractband, pag. 368

[126] WALTER, S. D. (2002) Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Statistics in Medicine.* **21**, 1237-56.

[127] WALTER, S. D. (2005) The partial area under the summary ROC curve *Statistical in Medicine.* **24**, 2025–2040.

[128] WEST, M. (1984) Outlier models and prior distributions in Bayesain linear regression. *JRSS, B.* **46**, 431–439.

[129] WEST, M. (1985) Generalized linear models: Scale parameters, outlier accomodation and prior distributions. *Bayesian Statistics 2, Proc. 2nd Int. Meet.* 531–558.

[130] WEST, M. (1986) Bayesian model monitoring. *J. Roy. Statist. Soc.* Ser. B, **48**, 70-78.

[131] WESTWOOD, M. E., WHITING, P. F. AND KLEIJNEN, J.(2005) How does study quality affect the results of a diagnostic meta-analysis? *BMC Medical Research Methodology* **5**, 1471-2288.

[132] WHITING P, RUTJES A, REITSMA J, BOSSUYT P, KLEIJNEN J (2003) The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol*.3:25.

[133] M.H. Zweig and G. Campbell (1993). "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine". *Clinical chemistry* **39 (8)**: 561577.

[134] ZOU, K. H. (2002) Receiver operating characteristic (ROC) literature research. Online bibliography available from http://splweb.bwh.harvard.edu:8000/pages/ ppl/zou/roc.html.

[135] *Statistical Science* (2003), **Vol 18**, Number 2, pag 133-268.

Hiermit versichere ich, die Arbeit selbständig erstellt und keine anderen als die angegebenen Hilfsmittel benutzt zu haben.

Pablo Emilio Verde