Defining and Testing Diagnostic Equivalence

Using a Bayesian Hierarchical Model

by

Konstantinos Kalogeropoulos

B.A., Athens University of Economics

and Business, 2001

Thesis

Submitted in partial fulfillment of the requirements for the

Degree of Master of Sciences in the Division of Biology and Medicine

at Brown University

May 2004

SIGNATURE PAGE

This thesis by Konstantinos Kalogeropoulos

is accepted in its present form by the Division of

Biology and Medicine as satisfying the

thesis requirements for the degree of Master of Sciences

Date _____ _____

Constantine Gatsonis, Advisor

Approved by the Graduate Council

Date _____ _____

Karen Newman
Dean of the Graduate School

Table of Contents

Defining and Testing Diagnostic Equivalence Using a Bayesian Hierarchical Model

1. Introduction

1.1 Overview: the question of equivalence in biomedical research

In biomedical research, the assessment of equivalence addresses the question of whether two (or more) quantities are close enough to be considered essentially the same for clinical or health policy decision making. The problem of equivalence appears in several fields. In clinical trials the question of interest is whether a new therapy is as effective as a standard therapy (Blackwelder 1982). In pharmacology the question is one of bioequivalence between two drug formulations (FDA 2001, Selwyn et al 1981, Anderson and Hauck 1990, Schall and Luus 1993). Usually these are compared in terms of their bioavailabilities, which are measured through the drug concentration in the blood by one or more pharmacokinetic variable, such as area under concentration vs time curve and the maximum concentration. In diagnostic medicine, technology advancements often raise the question of whether a new diagnostic test has equivalent diagnostic performance to that of a standard diagnostic modality (Obuchowski 2001). An important aspect of the assessment of equivalence in the diagnostic setting is the variability in the diagnostic performance among test interpreters. A similar issue may also be important in the therapeutic setting, if variability in the effectiveness of the therapy is present among health care providers (e.g. hospitals or surgeons).

1.2 Statistical formulations of equivalence

The statistical formulation of the problem begins by making precise the notion of "essentially the same" As we will see below; this task is not always straightforward. Statistical equivalence between two population parameters A_1 and A_2, is generally assessed by selecting a contrast statement between A_1 and A_2, denoted by *d* (e.g. Euclidean distance), and fixing a maximum acceptable difference δ, within which the parameters will be considered as equivalent. In the frequentist framework, a test for equivalence may take the form:

$$H_0 : d(A_1, A_2) > \delta \text{ vs } H_1 : d(A_1, A_2) \leq \delta \quad (1)$$

Note that since only the probability of type I error is controlled, and therefore the null hypothesis can only be disproved, the alternative hypothesis should contain the statement that the experimenter would hope to prove, which is in our case equivalence of A_1 and A_2 (Blackwelder 1981).

In the Bayesian framework, the joint posterior distribution of (A_1, A_2) can be used to calculate the posterior probability of the equivalence range (Selwyn et al 1981). If this probability is sufficiently high, it would be sensible to infer equivalence. In other words we reject nonequivalence if and only if:

$$\Pr(d(A_1, A_2) \leq \delta \mid \mathrm{Y}) \geq 1 - \alpha \quad (2)$$

where Y represents the data. Note that under both frameworks the choice of δ plays a crucial role.

In practice definitions (1) and (2) cover only some of the types of equivalence that are of interest. Consider for example the comparison of the accuracy of diagnostic tests interpreted by a specially trained radiologist (reader). In such cases, a sensible comparison between two tests should take into account the inherent variability in the diagnostic performance of the readers. As a direct application of the above formulations (1 or 2) one may compare the mean performance, over the population of readers, for example using a simple t-test. However, such an approach provides only a partial answer to the question. Even if the two means were identical, the variability across readers might be significantly different between the two tests, leading to substantially different outcomes from the use of the two tests.

1.3 Assessment of Bioequivalence

A similar problem to diagnostic equivalence has been addressed extensively in pharmacology studies of bioequivalence. A common setting in such studies involves situations in which two drug formulations are compared in terms of their bioavailabilities in a set of subjects Anderson and Hauck (1990) and Schall and Luus (1993) introduced the notions of *population* and *individual* bioequivalence. Population bioequivalence refers to the population of subjects and essentially compares the distributions of the bioavailabilities of drug formulations across subjects. Individual bioequivalence operates at a different level of aggregation and compares the formulations within subjects. Clearly, population bioequivalence does not necessarily imply individual bioequivalence.

The criteria proposed for assessing these types of bioequivalence can be classified into (i) moment-based and (ii) probability-based. To discuss the moment-based criteria, consider the following mixed effects model:

$$Y_{ij} = \mu_i + u_{ij} + e_{ij}$$

Here *i* denotes the formulation (T: formulation being tested and R: reference formulation) *j* denotes the subject (j=1,...*J*, the total number of subjects), and *Y* denotes the measurement on the *j*-th subject under the *i*-th formulation. The vectors (u_{Tj}, u_{Rj}) correspond to the subject's random effect and are assumed to be mutually independent and normally distributed according to a bivariate normal distribution with mean 0, variances σ_{Bi}^2 and correlation ρ. The variables e_{ij}'s correspond to the random error within subjects and are assumed to be jointly independent and normally distributed with mean 0 and variance σ_{Wi}^2. The model also assumes that u_{ij} and e_{ij} are independent given a subject *j*. In the discussion below we denote $\sigma_i^2 :=\text{var}(Y_{ij}) = \sigma_{Bi}^2 + \sigma_{Wi}^2$ and $\sigma_D^2 :=\text{var}(u_{Tj} - u_{Rj})$.

The moment-based criteria require a study design where one of the formulations -the reference formulation- is tested on two different occasions in each subject, resulting in two independent observations Y_{Rj} and Y'_{Rj} for each *j*. This is done in order to establish a benchmark for the discrepancy that may occur even when the two formulations are the same.

Schall and Luus (1993) proposed to assess population bioequivalence using a criterion based on the between subject differences:

$$\mathrm{E}(Y_{Tj} - Y_{Rk})^2 - \mathrm{E}(Y_{Rj} - Y'_{Rk})^2 \le \delta_1^2 \text{ for all } j \neq k \quad \text{or} \quad (\mu_T - \mu_R)^2 + \sigma_T^2 - \sigma_R^2 \le \delta_1^2 \qquad (3)$$

They also proposed to assess individual bioequivalence on the basis of the within subject differences:

$$\mathrm{E}(Y_{Tj} - Y_{Rj})^2 - \mathrm{E}(Y_{Rj} - Y'_{Rj})^2 \le \delta_2^2 \quad \text{or} \quad (\mu_T - \mu_R)^2 + \sigma_D^2 + \sigma_{WT}^2 - \sigma_{WR}^2 \le \delta_2^2 \qquad (4)$$

The quantities in (3), (4) were estimated by replacing the corresponding parameters with their estimates obtained from the mixed model using the method of *Restricted Maximum Likelihood* (REML). To complete the assessment of equivalence one should calculate confidence intervals for the estimates of the relevant quantities. Initially the use of bootstrap was proposed to calculate these estimates but more recently some other approximation methods were proposed (Hyslop et al 2000, FDA 2001). An alternative approach by McNally et al (2003) was based on the use of generalized p-values instead of confidence intervals.

Despite the fact that it has been widely used and adopted in a modified version by FDA guidance (FDA 2001), concerns remain about the Schall and Luus approach. In particular, the quantities in (3), (4) are not easy to interpret quantitatively, because they involve sums of squared mean differences and variance terms. Hence the task of picking a sensible δ is not straightforward.

Wellek (2000b) proposed a different criterion for population bioequivalence, using an intersection-union test. This criterion requires the following two inequalities to be satisfied simultaneously.

$$(\mu_T - \mu_R)^2 / (\sigma_{WT}^2 + \sigma_{WR}^2) \le \delta_1^2 \qquad (5a)$$

and

$$\sigma_T^2 / \sigma_R^2 \leq 1 + \delta_2 \tag{5b}$$

Note that (5a) involves a standardized version of the difference in the two means and provides a basis for the choice of δ_1. However, it is more difficult to develop a rationale for specifying an appropriate value δ_2.

As noted above, probability-based criteria are used in another family of approaches to the assessment of bioequivalence. These criteria are based on the probability of equivalence and are considered to be easier to interpret. In particular, Schall (1995) proposed the criteria:

$$\begin{gathered} \Pr\left(|Y_{Tj} - Y_{Rk}| \leq \delta_1\right) - \Pr\left(|Y_{Rj} - Y'_{Rk}| \leq \delta_1\right) \geq \Delta_1 \quad (j \neq k) \\ \Pr\left(|Y_{Tj} - Y_{Rj}| \leq \delta_2\right) - \Pr\left(|Y_{Rj} - Y'_{Rj}| \leq \delta_2\right) \geq \Delta_2 \end{gathered} \tag{6}$$

for population and individual bioequivalence respectively. The notation in (6) is the same as above and the parameters $\Delta_1, \Delta_2 < 0$ determine the range for population and individual bioequivalence respectively. One may choose $\delta_1 = \gamma\sqrt{2}\sigma_R$ and $\delta_2 = \gamma\sqrt{2}\sigma_{WR}$ where σ_R is the standard deviation of $Y_{Rj} - Y'_{Rk}$ ($j \neq$ k), σ_{WR} is the standard deviation of $Y_{Rj} - Y'_{Rj}$, and γ is a positive constant. Then, under the assumption of normality for the Y_{ij}'s, the second probability term in the above differences is a constant given γ and the first can be written as:

$$\Phi\left(\frac{\gamma\sqrt{2}\sigma_R + \mu_T - \mu_R}{\sqrt{\sigma_R^2 + \sigma_T^2}}\right) - \Phi\left(\frac{-\gamma\sqrt{2}\sigma_R + \mu_T - \mu_R}{\sqrt{\sigma_R^2 + \sigma_T^2}}\right)$$

for population bioequivalence and as :

$$\Phi\left(\frac{\gamma\sqrt{2}\sigma_{WR}+\mu_T-\mu_R}{\sqrt{\sigma_{WR}^2+\sigma_{WT}^2+\sigma_D^2}}\right)-\Phi\left(\frac{-\gamma\sqrt{2}\sigma_{WR}+\mu_T-\mu_R}{\sqrt{\sigma_{WR}^2+\sigma_{WT}^2+\sigma_D^2}}\right)$$

for individual bioequivalence.

The quantities in (6) can be estimated as in (3)-(4) by replacing the corresponding parameters with their estimates obtained from the mixed effect model as before.

A non-parametric probability-based criterion for individual bioequivalence was introduced by Anderson and Hauck (1990) and did not presuppose the replicate design. The criterion was based on Pr ($1-\delta \le m_{Tj}/m_{Rj} \le 1+\delta$), the probability that a randomly selected subject will have equivalent mean bioavailabilities with each formulation (here m_{ij} denotes the mean bioavailability of the formulation *i* on the subject *j*). The use of ratios was adopted because the observed bioavailabilities (Y_{ij}'s) were modeled in the log scale. The probability was estimated by replacing m_{ij} with Y_{ij}, or, in other words, by taking the sample proportion of subjects with equivalent responses. Because of its straightforward interpretation, this criterion has become quite popular in practice. However since it is based on a non-parametric model it has limited power.

Wellek (2000a) used a parametric Bayesian model to estimate a probability similar to the one proposed by Anderson and Hauck. Let $Z_j := (\log Y_{1j} - \log Y_{2j})$ and assume that $Z_j \sim$ N(ζ, σ^2) (*j*=1,…,*J*). Wellek approximated the probability Pr [$(1+\delta)^{-1} \le \exp(Z_j) \le 1+\delta$], using the posterior distribution of (ζ, σ) and numerical integration. If we assign the usual reference prior on (ζ, σ), i.e. $\pi(\zeta, \sigma) \propto \sigma^{-1} I(\sigma > 0)$ (Box and Tiao 1972) , the posterior is:

$$\pi(\zeta,\sigma \mid \underline{Z}) = \left(\sqrt{J/\sigma}\right)\Phi(\sqrt{J}\,(\zeta-\bar{Z})/\sigma)\,\sqrt{J-1}\left(s/\sigma^2\right)g_{J-1}(\sqrt{J-1}\,s/\sigma)$$

where $\bar{Z}$ is the sample mean and $s=[(J-1)^{-1}\sum_j (Z_j-\bar{Z})^2]^{1/2}$, Φ is the density of a standard normal distribution and g_{J-1} is the density of a chi-square distribution with *J*-1 degrees of freedom.

In Chapter 3 of this thesis we introduce a Bayesian hierarchical model that allows separate means for each Z_j. The model serves as the basis for deriving several metrics with meaningful interpretation. Moreover it allows us to consider criteria based directly on the subject –specific means m_{ij} rather than Y_{ij}.

1.4 Equivalence for Diagnostic tests

In the diagnostic setting, studies of equivalence address the question of whether two diagnostic tests are equally effective in terms of their diagnostic performance. A common situation involves a new test that has some practical advantages over a standard test (eg. is less expensive, easier to be implemented etc), thus raising the question whether it would be safe to replace the standard with the new test.

In this thesis we assume that complete test results and reference information (gold standard) are available, and hence measures of diagnostic performance can be estimated for each test. Among the several available measures of diagnostic performance we focus on the area under the Receiver Operating Characteristic (ROC) curve, denoted by A. We note however that formulation of the problem and the methods for assessing equivalence discussed below apply in general to univariate summaries of the ROC curve (e.g. partial areas) or other univariate measures of diagnostic performance. We also comment on how

these approaches can be adapted to handle the case of multivariate measures of diagnostic performance (e.g. pairs of sensitivity/specificity).

The problem of equivalence of diagnostic tests has only recently received attention in the literature, notably in a paper by Obuchowski (2001) (see also Zhou et al (2002) and Alonzo et al (2002). Using ideas from the bioequivalence literature, Obuchowski defined population and individual diagnostic equivalence. Population equivalence was defined as in the Schall and Luus approach (3). Let $\hat{A}_{ijk}$ denote the estimated area under the curve for reader *j* with test *i* on reading occasion *k*. Then the diagnostic population equivalence criterion is:

$$\gamma_P = \mathrm{E}(\hat{A}_{2jk} - \hat{A}_{1j'k})^2 - \mathrm{E}(\hat{A}_{1jk} - \hat{A}_{1j'k'})^2 \leq \Delta_P^2 \qquad (7)$$

Here $\hat{A}_{2jk} - \hat{A}_{1j'k}$ and $\hat{A}_{1jk} - \hat{A}_{1j'k'}$ correspond to the between subject differences involved in (3) and the expectation is taken over the relevant reader and patient populations. Note that the estimates of the areas are correlated even when they refer to different readers (since they are obtained from the same set of patients) and therefore the expression in (7) does not simplify to the corresponding quantity in (3). An unbiased estimate of γ_P, for *k*=2, is given by:

$$\hat{\gamma}_P = (1/J^2)\left\{\sum_j^J \sum_{j'}^J (\hat{A}_{2j1} - \hat{A}_{1j'1})^2 - \sum_j^J \sum_{j'}^J (\hat{A}_{1j1} - \hat{A}_{1j'2})^2\right\} -$$
$$(2/J^2)\left\{\sum_j^J \sum_{j'}^J \hat{\mathrm{cov}}(\hat{A}_{1j1}, \hat{A}_{1j'2})^2 - \sum_j^J \sum_{j'}^J \hat{\mathrm{cov}}(\hat{A}_{2j1} - \hat{A}_{1j'1})^2\right\}$$

Obuchowski used a nonparametric model to obtain estimates for the areas and their covariances (see also Obuchowski 1997), but other models can also be adopted.

Individual reader equivalence was defined at the patient level, representing a type of agreement between the two tests. The area under ROC curve cannot be defined at the patient level so another measure $\hat{\Pi}_{ijkc}$ was used, which refers to the interpretation done by reader *j* with test *i* on reading occasion *k* on a particular case *c*. The criterion for individual diagnostic equivalence was:

$$\gamma_I = \Pr\left[(\hat{\Pi}_{1jkc} - \hat{\Pi}_{1j'k'c})^2 \le \lambda_I^2\right] - \Pr\left[(\hat{\Pi}_{2jkc} - \hat{\Pi}_{1j'kc})^2 \le \lambda_I^2\right] \le \Delta_I^2$$

where λ_I is the upper bound on the acceptable difference between the tests' results for a case. The quantity γ_I can be estimated by averaging the appropriate indicator functions. For confidence intervals of γ_P and γ_I the use of bootstrap was proposed as in the bioequivalence case.

As in the case of the Schall and Luus criteria, criterion (7) combines means and variances into one expression. As a result, the subject matter interpretation of Δ_p^2 is not straightforward. Obuchowski's notion of individual reader equivalence is defined within patient and does not have a clear interpretation in terms of diagnostic performance. However, such a notion may be useful in deciding whether the use of the two tests may lead to similar decisions in specific patients. In this thesis, we consider the reader as the unit of analysis and define *individual* equivalence as equivalence for individual readers.

1.5 Summary

In the next chapter we describe a framework for defining population and individual diagnostic equivalence. In our formulation, *population* and *individual* refer to readers. We begin with a description of the problem and highlight some of its crucial

characteristics, distinguishing diagnostic population equivalence from individual (reader) equivalence. We introduce metrics for assessing these types of equivalence and discuss the rationale of these metrics. In Chapter 3 we present a 3-level Bayesian hierarchical model that can be used to estimate these metrics and discuss model fitting and estimation procedures. We also examine the sensitivity of the findings to the choice of prior distributions. This chapter ends with a description of the possible experimental designs for equivalence studies. In chapter 4 we apply our proposed method to the analysis of data from a study of a Computer Aided Diagnostic (CAD) algorithm for mammography. The chapter also presents results of a simulation study conducted to explore frequentist properties of the model proposed in chapter 3. The final chapter is devoted to a discussion of possible extensions of the Bayesian approach and other issues related to the assessment of test equivalence.

2. Framework for defining equivalence for diagnostic test

2.1 Technology assessment considerations

We will consider two kinds of equivalence of diagnostic tests, defined at the population and individual reader level. The rationale for that is that the two tests may be considered as equivalent as used by a population of readers but may not be equivalent for every particular reader or even for most of them. Therefore we make the assumption that for each test there exists a population mean of a test performance measure, but readers have their own means. Although we use the area under the ROC curve as the main example of a test performance measure, the framework proposed in this thesis applies more generally

to univariate measures of performance, such as the partial area under ROC curve, sensitivity, specificity, etc. We discuss multivariate measures in Chapter 5.

An ordinary comparison of the population means may not yield an adequate criterion for population equivalence. For instance, consider the following hypothetical scenario: Suppose that the distributions of the ROC areas for the population of readers of two tests are both normal with common mean (.75) and different variances (say with standard deviations .02 and .06) as shown in Figure 1. Are these tests equivalent?

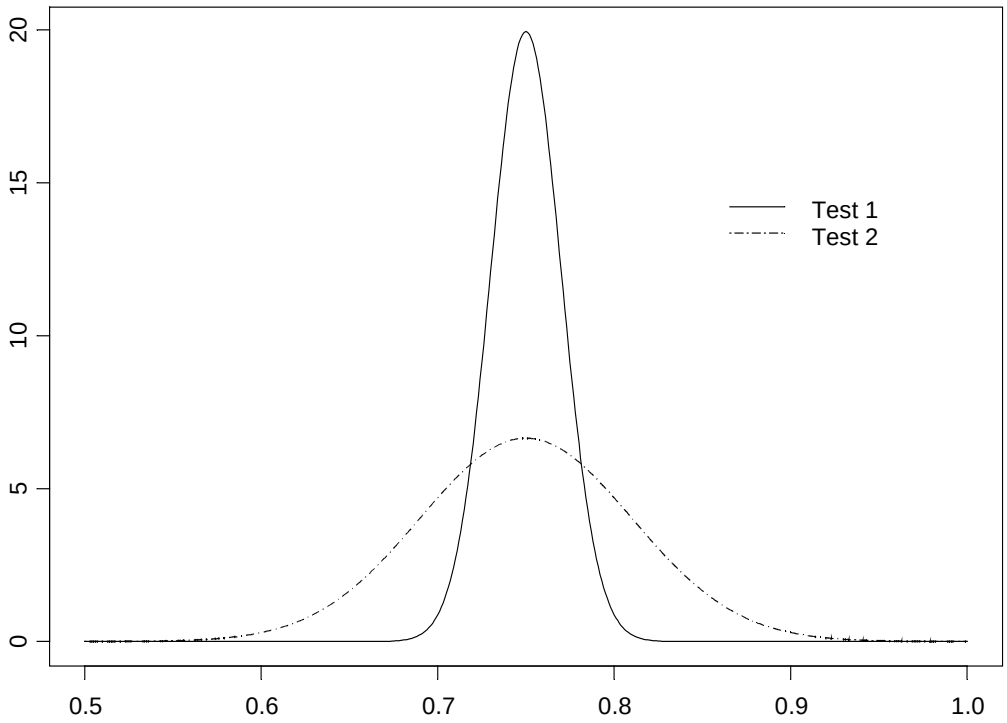

Figure 1: Scenario of normal distributions across readers with identical means (0.75) and different standard errors (test 1: .02, test 2: .06)

If the focus is only on the means, the two tests would be equivalent. However, an argument could be made that test 1 is better, because the proportion of readers with low accuracy (say ROC area below .7) is smaller than the corresponding proportion for test 2. Hence test 1 appears to be safer to adopt. An argument could also be made that test 2 is better because the proportion of readers with high accuracies (say above 0.8) is larger than the corresponding proportion for test 1. Neither argument is wrong. However, the fact that both arguments have merit serves to underline the point that the definition of equivalence must be made on the basis of a specific clinical or health policy objective.

2.2 Criteria for equivalence

Using the notation from Chapter 1, we will reserve the index *i* for tests (*i*= 1,2) and the index *j* for readers. Let μ_i denote the mean for the test *i* over the population of readers of the test, and A_{ij} the true area under the ROC curve mean for reader *j* with test *i*. Assessment of population equivalence refers to the comparison of the distributions of the areas of two tests across the population of readers. In order to compare these distributions, we need to define the characteristics of the distribution to be compared and metrics for measuring their dissimilarity. The characteristics of the distribution may be a set of simple parameters of the distribution (e.g. the mean and the variance) or a function of these parameters (e.g. the interquartile range). However each of these characteristics should be chosen to represent a quantity that is interpretable in the given clinical or health policy context. As we discussed earlier, it is also important to use measures of dissimilarity that have a subject matter interpretation. For instance there exist several measures of distance between two distributions, such as the Hellinger or the Kullback-

Leibler distance. Although such metrics have a mathematical interpretation, their clinical or health policy interpretation is not clear. For example the Kullback-Leibler distance between two normal distributions N(μ_1, σ_1^2), N(μ_2, σ_2^2) is equal to:

$$-1/2 + \log\left(\frac{\sigma_1}{\sigma_2}\right) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2}$$

This quantity does not have a clear subject matter interpretation and, as a consequence, it is not straightforward to determine what a maximum acceptable difference for equivalence δ should be.

The criteria proposed in this thesis can be formulated in the following way: Let e_1, e_2 be the characteristics of the distributions for tests 1,2 and assume that they take values in the sets E_1, E_2 respectively. Also let $d\{e_1, e_2\}$ be the measure of dissimilarity and δ the maximum acceptable dissimilarity between e_1 and e_2. The basis of our criteria for equivalence assessment will be the probability of similarity, defined on $E_1 \times E_2$:

$$\Pr\left(d\{e_1, e_2\} \leq \delta\right).$$

This quantity is easy to interpret in a particular subject matter setting, as long as e_1, e_2 and *d* have quantitative interpretations. Equivalence could then be assessed using a Bayesian framework, as in (2) above.

The choice of the characteristics of the distributions to be compared should be based on the particular health policy or clinical question that led to the considerations of equivalence. Choosing only the mean or the median of the distribution will be appropriate for situations in which the variability across readers is of secondary importance. We denote the probability of similar means by:

$$P_1 := \Pr\left(d(\mu_1, \mu_2) \leq \delta\right)$$

As mentioned earlier however, the variability across readers is often substantial. A comparison that takes such variability into account may be based on some low or high percentiles of these distributions. The corresponding probability would then be defined as:

$$P_2^q := \Pr\left(d(q_1, q_2) \leq \delta \right)$$

Here q_i represents an appropriately chosen percentile of the distribution for test *i*. For instance, if the focus is on the potential for low performance, we may compare a low percentile of these distributions - say the 0.2 percentile –. Population equivalence may then be defined to hold if these percentiles are similar with sufficiently high probability. If the focus is on potential high performance, we may choose to compare the distributions on the basis of a high percentile, say the 0.8 percentile.

A more global assessment of population equivalence can be defined using more than one percentile, some low and some high. For example, if $q_i^{0.2}, q_i^{0.8}$ denote the 0.2 and 0.8 percentile of the distribution for test *i*, equivalence can be defined via the probability:

$$\Pr\left[\{ d(q_1^{0.2}, q_2^{0.2}) \leq \delta_1 \} \cap \{ d(q_1^{0.8}, q_2^{0.8}) \leq \delta_2 \} \right]$$

We will denote the above probability by $P_2^{q1,q2}$.

Ideally, measures of dissimilarity *d* should have a subject matter interpretation and should also be simple enough to allow for an easy specification of values of δ. Two appealing choices, that have been already used, are the Euclidean distance and measures of the form $d\{e_1, e_2\} := |1 - e_1 / e_2|$, which will result in a metric of the type Pr(1 - $\delta \leq e_1 / e_2 \leq 1 +$ δ). Once a choice of d is made, the assessment of equivalence will be based on whether the corresponding probability of similarity is sufficiently high.

Individual reader equivalence compares the tests for each specific reader. This is a stricter type of equivalence because it goes beyond the comparison of the marginal distributions and requires similarities in the conditional (reader-specific) distributions as well. Formally, individual reader equivalence can be defined in matter similar to Anderson and Hauck approach, by computing:

$$I_1 := \Pr\left(d\{A_{1j}, A_{2j}\} \leq \delta\right)$$

Under the Bayesian framework, the probability above is determined by the joint posterior distribution of the A_{ij}'s. According to this criterion the tests will be considered individually equivalent if they are 'similarly' accurate in a sufficiently high proportion of readers.

An alternative way to define individual reader equivalence is by fixing a low percentile of the probabilities of 'similar' areas across readers. Let $p_j = \Pr\left(d\{A_{1j}, A_{2j}\} \leq \delta\right)$ for a given *j*. Given the Bayesian context, consider the distribution of p_j in the population of readers. Then define:

$$I_2^k := \text{k-th percentile of the distribution of } p_j\text{'s}$$

For k=0.2, I_2^k measures the minimum probability of similar areas that will be achieved by 80% of the readers. If this probability is sufficiently high, the tests will be considered individually equivalent.

In order to assess population and individual reader diagnostic equivalence, estimates of the quantities defined above should be obtained. This can be done easily under an appropriate multiple reader study design by fitting a suitable Bayesian model and using

the posterior distributions of the parameters involved. In the next chapter we present such a model and a study design.

3. Assessing equivalence with a Bayesian hierarchical model

In this section we discuss the assessment of equivalence using data that follow a hierarchical model. The specific metric is the area under the ROC curve for a population of readers. As noted above, the approach is applicable more generally, with appropriate modifications of the components of the hierarchical model.

3.1 Model specification

Assume for the purposes of this discussion that a fully crossed design was used to collect degree of suspicion data and estimate ROC curves for J readers in I=2 test modalities. Let Y_{ij} be the estimate of the area under the ROC curve for the reader j ($j = 1, ..., J$) in the test i ($i = 1, 2$). Also let the A_{ij}'s be the corresponding true areas. Since each reader interprets the same set of patients the data are correlated. Therefore in Level I of the model, we assume that, conditional on the hyperparameters, each pair of estimates (Y_{1j}, Y_{2j}) is distributed according to a bivariate normal distribution with its own mean (A_{1j}, A_{2j}) and covariance matrix Σ_j (See Gatsonis and Wang, in preparation). The covariance matrix Σ_j represents variability due to sampling error and can be approximated as a function of the mean areas and some additional quantities including the number of diseased and non-diseased cases (Obuchowski 1997, DeLong et al 1988, Hanley and

McNeil, 1983). For reasons of model parsimony, in this analysis we consider Σ_j to be known and equal to its estimated value. We discuss some possible implications of this choice in Chapter 5. Details of how to obtain these estimates are presented in the Appendix.

In the *Level II* of the model we use a logit transformation of the A_{ij}'s and, conditionally on hyperparameters, we assume again a bivariate normal distribution for the pairs (logit(A_{1j}), logit(A_{2j})) with mean (μ_1, μ_2), the population mean, and covariance matrix B. Here B represents the variability between readers.

To complete model specification, priors on the parameters μ_1, μ_2 and B should be assigned (*Level III*). Since all the realistic values for areas lie between 0.5 and 1, we assigned uniform priors in that region for each $\text{logit}^{-1}(\mu_i)$. To construct a prior for B, we used a hierarchical prior on the correlations, as discussed in Barnard and McCullogh (2001) and Daniels and Kass (1999). This prior has advantages in small samples over the Wishart and enables us to model the variances and the correlations separately, which is very convenient since we have different types of information for these parameters. The prior is set as follows: Consider the decomposition B=diag(S) × R × diag(S), where S is a vector with the standard deviations and R is the correlation matrix. For the elements of S we used log-logistic priors with means c_j, equal to the harmonic means of the variances of the transformed areas across readers for each test. These covariance matrices can be obtained using Delta method. (Du Mouchel 1994). For the prior on the correlation coefficient ρ, we put a U(0,1) on the correlation coefficient ruling out the possibility of negative correlation.

The model can be summarized as follows:

Level I

$$\begin{pmatrix} Y_{1j} \\ Y_{2j} \end{pmatrix} \Big| \, A_{ij} \sim N\left(\begin{pmatrix} A_{1j} \\ A_{2j} \end{pmatrix}, \ \hat{\Sigma}_j \right), \hat{\Sigma}_j \text{ is fixed}$$

Level II

$$\begin{pmatrix} \text{logit}(A_{1j}) \\ \text{logit}(A_{2j}) \end{pmatrix} \Big| \, \mu_i, B \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \ \mathrm{B} \right)$$

Level III

$\text{logit}^{-1}(\mu_i) \sim \text{U}(0.5,1)$

$$B = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}$$

$\sigma_i \sim$ *log-logistic*$(c_i, 1)$,

c_i : harmonic mean of the diagonal entries in the covariance matrices of the transformed areas for each reader

$\rho \sim \text{U}(0,1)$

A Markov Chain Monte Carlo algorithm for this model was written in BUGS. The convergence of the chain needs to be monitored carefully, because the correlations in B are probably going to be high. Alternatively a Metropolis step within Gibbs may be used for B (Daniels and Kass 1999).

3.2 Assessing equivalence using the model

Estimates of the various equivalence criteria can be obtained using the samples from the posterior of the A_{ij}'s. Denote by a_{ijk} the *k*-th draw of the parameter A_{ij} (*k*=1,…,*n*). Then P_1 can be approximated by the formula:

$$\frac{\sum_k I\{d(\bar{a}_{1\bullet k}, \bar{a}_{2\bullet k}) \le \delta\}}{n}, \qquad \bar{a}_{i\bullet k} = \left(\sum_j a_{ijk}\right)/n,$$

where $I(\cdot)$ is an indicator function. In order to approximate P_2^q and $P_2^{ql,qh}$, we should calculate samples from the relevant percentiles. Let ql_{ik} and qh_{ik} be the relevant low and high percentiles of the distribution across readers for test *i*, obtained from the draw a_{ijk}.

Then

$$P_2^{ql} = \frac{\sum_k I\{d(ql_{1k}, ql_{2k}) \le \delta\}}{n}, \qquad P_2^{qh} = \frac{\sum_k I\{d(qh_{1k}, qh_{2k}) \le \delta\}}{n}$$

$$P_2^{ql,qh} = \frac{\sum_k I\{d(ql_{1k}, ql_{2k}) \le \delta\} \times I\{d(qh_{1k}, qh_{2k}) \le \delta\}}{n}$$

To approximate I_2^m, we may calculate the probabilities of similar accuracies p_j for each reader, by averaging over all draws for each *j*:

$$p_j = \frac{\sum_k I\{d(a_{1jk}, a_{2jk}) \le \delta\}}{n}$$

Then I_2^m can be obtained by taking the *m*-th percentile of the p_j's.

For estimating I_1 we should average over the draws and the readers *j*:

$$I_1 = \frac{\sum_j \sum_k I\{d(a_{1jk}, a_{2jk}) \le \delta\}}{n \times J}$$

Equivalently we calculate the average of the probabilities p_j over j.

3.3 Sensitivity to distributional assumptions

The normality assumption in *Level I* of the model is based on the asymptotic distribution of the estimates Y_{ij}'s and holds quite generally. Concerns about this assumption may arise when the sample size of normal and abnormal cases are small and when the values of the true areas are close to 1. The normality assumption at *Level II* is similar to the assumption made in the mixed models discussed in section 1. This assumption needs closer scrutiny because it is an important determinant of the posterior estimates of individual reader performance measures. The normality assumption may be difficult to validate on the basis of empirical information, unless the equivalence study involves a large number of readers. Of course, when the sample size of cases is large, the variance at *Level I* will be small and the shrinkage of *Level I* parameters will be relatively small for most reasonable choices of the *Level II* distribution. However, this is often not the case in practice. A heavier tailed multivariate t distribution with few degrees of freedom would be a more "robust" choice for the *Level II* prior. In section 4.1 we perform a sensitivity analysis to this particular distributional assumption.

3.4 Study design considerations

Because of the conditional independence assumption made in *Level I*, the hierarchical model described above is suitable for the analysis of data from designs in which the readers interpret different sets of patients. This situation is common in data from two of

the most likely sources of information for equivalence studies, multi-center trials and meta-analyses of published studies

The hierarchical model described in this chapter can also be used to analyze data from a typical multi-reader study where each reader interprets both tests, performed on the same set of cases (patients) and the set of cases is the same for each reader.

There are three separate types of correlation that may be induced by this study design (Obuchowski 1995, Toledano and Gatsonis 1996): (i) because scans of the common set of cases are read by the same reader for both tests, (ii) because scans on the set of cases are read by different readers for the same test, and (iii) because scans on the common set of cases are read by different readers for different tests. These correlations can be estimated directly from the primary scan interpretation data. The first appears explicitly in $\hat{\Sigma}_j$. The other two are included only implicitly in our model by the marginal correlations induced by the hierarchical model.

We note that the formulations of equivalence considered in this thesis apply to studies in which each scan is interpreted by each reader on only one occasion. Designs of this type are commonly used in diagnostic and screening test evaluation. .

4. Applications

4.1 Example: Transfer of Intelligence Technologies to Breast Imaging study

We illustrate the use of the proposed model in the Transfer of Intelligence technologies to Breast Imaging study. In this study 10 radiologists interpreted a set of mammograms on 900 women in two settings, (i) mammogram alone and (ii) mammogram with a computer aided diagnostic tool (CAD). Reference information about the disease status of the

participants was based on pathology data for those with a biopsy and on follow-up for those who did not have a biopsy. 5 of the radiologists were specialized in mammography and the other 5 were general radiologists. The data (Table 4.1), consist of estimates of the ROC areas for each reader and the estimates of the covariance matrices ($\hat{\Sigma}_j$) obtained from the Binormal model (Metz 1978).

The t-test statistics and confidence intervals for the difference of the average areas were computed using the mixed model approach of Obuchowski (1995), which accounts for the various correlations described in section 3.4. This test suggests that the difference in the two means is not significant. However, as mentioned before, this may not be an adequate criterion to infer equivalence without further consideration of the variability across readers in a more elaborate analysis.

	Reader	**Plain**	**S.E**	**CAD**	**S.E.**	**Correlation**
Radiologists specialized in Mammography	**#1**	0.8406	0.0232	0.8176	0.0257	0.4998
	#3	0.8808	0.0154	0.8953	0.0151	0.6111
	#4	0.9040	0.0137	0.8955	0.0147	0.6870
	#7	0.8706	0.0161	0.8805	0.0157	0.6052
	#8	0.8402	0.0177	0.8553	0.0166	0.6538
General Radiologists	**#2**	0.7826	0.0209	0.7654	0.0220	0.5802
	#5	0.7663	0.0252	0.8304	0.0193	0.4667
	#6	0.8683	0.0160	0.8831	0.0146	0.6197
	#9	0.8354	0.0184	0.8513	0.0178	0.5570
	#10	0.8527	0.0186	0.8346	0.0209	0.5321
Overall Average		**Plain**	**CAD**	**t-value**	**P-value**	**95 % C.I.**
		0.8442	0.8509	-0.5286	0.6099	[-0.022 0.036]

Table 4.1: Areas under fitted ROC curves, t-test, p-value and 95% confidence interval for the difference in average areas

We used 3 variants of the hierarchical model resulting from different distributional assumptions in *Level II* – a bivariate normal and two bivariate t distributions with 5 and 1

degrees of freedom. Since a bivariate t with 1 degree of freedom has no moments (is similar to a bivariate Cauchy distribution), the μ_i's and *B* were just location and scale parameters in that case. After a large burn-in sample of 10,000 iterations, and the appropriate check for convergence, 10,000 samples were drawn from the posterior distribution of the parameters in each model. We used a thinned sample by 5 and thus estimated the quantities of interest using 2,000 nearly independent draws.

The estimates of the probabilities that correspond to population equivalence assessment are shown in table 4.2. We chose a low percentile of 0.2 and a high of 0.8 and for the probability $P_2^{0.2,0.8}$ and used the same value of δ for each percentile.

Tables 4.2a and 4.2b summarize the results for population equivalence. To assess population equivalence we should first define the contrast metric *d*, the minimum practical difference δ and the probability 1-a. As noted above, these choices need to be made on the basis of subject matter considerations. Once the choices are made, equivalence assessment is relatively straightforward.

For example, suppose that we choose the Euclidean distance contrast and we set δ=0.05 and 1-a=0.95. Then the estimates of the probabilities presented in table 4.2a will suggest population equivalence, since the means, the low and the high percentiles of the distributions of the areas across readers are "similar" with sufficiently high probability. However for δ=0.03 the probability that the low percentiles are similar is smaller and $P_2^{0.2}$ or $P_2^{0.2,0.8}$ less suggestive of population equivalence. The probability estimates do not change appreciably when heavier tailed distributions are used in *Level II* and, in any case they remain above 90%.

	Bivariate Normal		Bivariate t with 5 d.f.		Bivariate t with 1 d.f.	
	δ =0.05	δ =0.03	δ =0.05	δ =0.03	δ =0.05	δ =0.03
P_1	1	1	1	1	1	1
$P_2^{0.2}$	0.9995	0.9465	0.9990	0.9400	0.9975	0.9055
$P_2^{0.8}$	1	1	1	0.9985	1	0.9990
$P_2^{0.2-0.8}$	0.9995	0.9465	0.9990	0.9385	0.9975	0.9045

Table 4.2a: Metrics related to population equivalence with $d(e_1, e_2) = |e_1 - e_2|$

	Bivariate Normal		Bivariate t with 5 d.f.		Bivariate t with 1 d.f.	
	δ =0.1	δ =0.05	δ =0.1	δ =0.05	δ =0.1	δ =0.05
P_1	1	1	1	1	1	1
$P_2^{0.2}$	1	0.9955	1	0.9940	1	0.9875
$P_2^{0.8}$	1	1	1	1	1	1
$P_2^{0.2,0.8}$	1	0.9955	1	0.9940	1	0.9875

Table 4.2b: Metrics related to population equivalence with $d(e_1, e_2) = |1 - e_1 / e_2|$

Estimates of the metrics that correspond to individual reader equivalence are shown in Table 4.3. The values of the estimated probabilities are quite high, indicating that individual equivalence would be inferred for many reasonable choices of δ and a.

	Bivariate Normal		Bivariate t with 5 d.f.		Bivariate t with 1 d.f.	
	δ =0.05	δ =0.03	δ =0.05	δ =0.03	δ =0.05	δ =0.03
I_1	0.9750	0.8921	0.9753	0.8991	0.9688	0.8905
$I_2^{0.2}$	0.9892	0.8938	0.9912	0.9038	0.9855	0.8828

Table 4.3a: Metrics related to individual reader equivalence with $d(e_1, e_2) = |e_1 - e_2|$

	Bivariate Normal		Bivariate t with 5 d.f.		Bivariate t with 1 d.f.	
	δ =0.1	δ =0.05	δ =0.1	δ =0.05	δ =0.1	δ =0.05
I_1	0.9990	0.9554	0.9979	0.9572	0.9964	0.9473
$I_2^{0.2}$	1	0.9605	0.9998	0.9658	0.9995	0.9513

Table 4.3b: Metrics related to individual reader equivalence with $d(e_1, e_2) = |1 - e_1 / e_2|$

Here 1-a has a slightly different interpretation and it may be sensible to set it at a lower value. The inference seems to be robust to the choice of the *Level II* prior in every case. One explanation for this similarity in the results between population and individual equivalence may be the high correlations in the estimates of the ROC areas, which tend to make the pairs move "together" and therefore the differences within pairs, are not affected by the choice of the *Level II* distribution.

4.2 Simulations

We used simulations to investigate the behavior of the model and the metrics proposed in section 3. Datasets analogous to the Transfer of Intelligence to Breast Imaging study were generated under two different scenarios and analyzed each one of them. These scenarios contained different choices for the distributions of the accuracies for each test across readers, which are shown in Figure 2. The first scenario assumed that the areas of both tests (in the logit scale) had normal distributions across readers with identical means and different variances. The second scenario assumed again normal distributions for the areas (in the logit scale), but one test had larger mean and larger variance than the other. In both scenarios the areas exhibited high correlation in order to correspond to the study designs described in section 3.4. We included a small number of readers (10) in these study designs, as this is usually the case. For every scenario we generated 100 datasets and at each one of them we used a Gibbs sampler to draw from the posterior and analyze the data. In order to ensure convergence in all datasets we applied a large burn-in period of 10,000 iterations. After that we proceeded as in the analysis of example 4.1, drawing 10,000 samples and thinning the sample by 5. To assess population equivalence we used

the Euclidean based metrics with δ=0.05, the ratio-based with δ=0.08 and we examined two possibilities for a (.05 and 0.1). The percentiles that we chose were again to be 0.2 and 0.8 and for the probability: $P_2^{0.2,0.8}$ the same δ was used for each percentile. For individual reader equivalence we used the same *d*'s and δ's, a's of 0.2 and 0.1 and the 0.2 percentile for $I_q^{0.2}$.

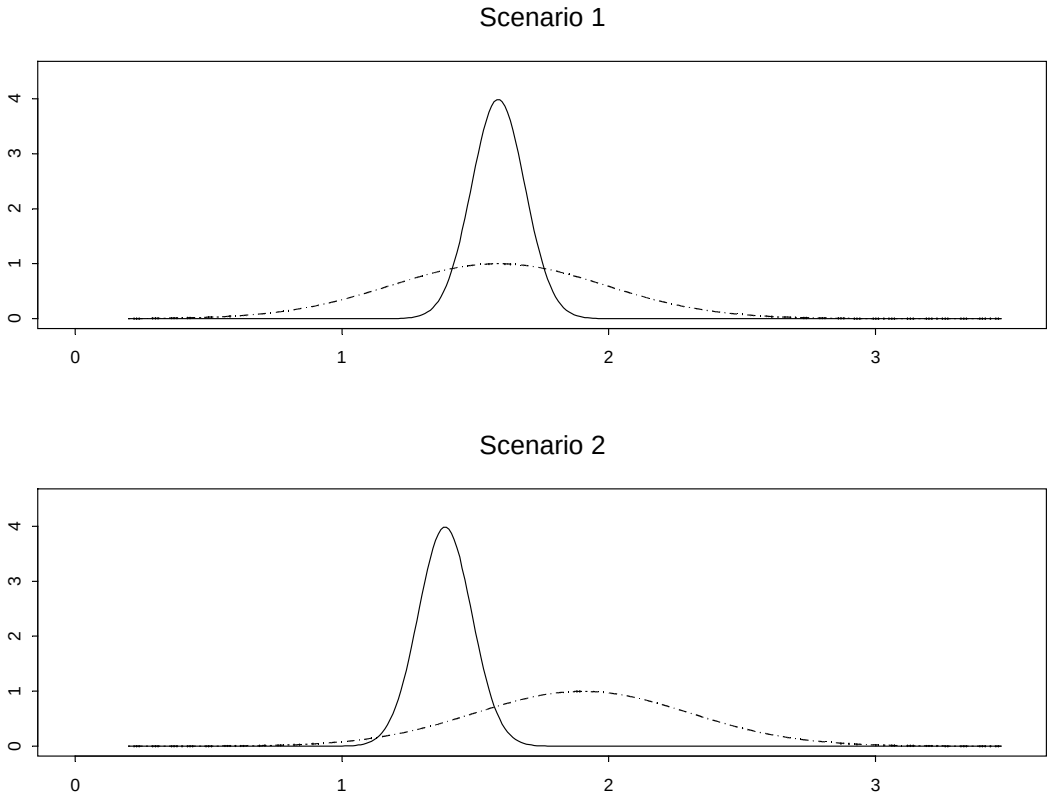

Figure 2: Hypothetical marginal distributions of the areas across reader for test 1 (solid line) and test 2 (dashed line) under the two simulation scenarios.

Scenario 1: In this setting one of the tests is more likely to have extreme (high or low) reader performance than the other The logits of the true areas (A_{ij} 's) for each reader were drawn from a bivariate normal distribution with mean logit(0.83) and covariance matrix

with standard deviations .1 and .4, and correlation .6. The estimates of the areas for each reader (Y_{ij}'s) were generated by adding noise to the A_{ij}'s according to a bivariate normal distribution with mean 0 and covariance matrices similar to the Σ_j's in the example. Tables 4.4a and 4.4b summarize the results of this set of simulations:

	Criteria with $d(e_1, e_2) = \lvert e_1 - e_2 \rvert$, δ =.05		Criteria with $d(e_1, e_2) = \lvert 1 - e_1 / e_2 \rvert$, δ =.08	
	=0.1	=0.05	=0.1	α =0.05
Pr ({ P_1 >1-α })	1	1	1	1
Pr ({ $P_2^{0.2}$ >1-α })	0.67	0.61	0.79	0.71
Pr ({ $P_2^{0.8}$ >1-α })	0.65	0.55	0.89	0.83
Pr ({ $P_2^{0.2,0.8}$ >1-α })	0.50	0.44	0.69	0.55

Table 4.4a: Summary of simulations under scenario 1 for population equivalence. The entries show the proportion of times each criterion was larger than 1-α

	Criteria with $d(e_1, e_2) = \lvert e_1 - e_2 \rvert$, δ =.05		Criteria with $d(e_1, e_2) = \lvert 1 - e_1 / e_2 \rvert$, δ =.08	
	α =0.1	α =0.05	α =0.1	α =0.05
Pr ({ I_1 >1-α })	0.16	0.08	0.38	0.23
Pr ({ $I_2^{0.2}$ >1-α })	0.11	0.07	0.42	0.35

Table 4.4b: Summary of simulations under scenario 1 for individual reader equivalence. The entries show the proportion of times each criterion was larger than 1-α

As expected the metric P_1 suggests equivalence almost always, whereas the other metrics may suggest otherwise. In this scenario, one may have expected lower values for the percentile-based probabilities than what was reported in the table. An explanation for this discrepancy may be that the number of readers (10) is small, making it difficult to estimate the percentiles with high precision. It is reassuring however that the estimates of I_1 and $I_2^{0.2}$ are fairly low, making it very unlikely to conclude individual equivalence in

this case. This is a situation where the tests may be equivalent in the population of readers but not in the individual reader level.

Scenario 2: In this setting, one of the tests is generally better than the other. However reader performance with the "better" test is more variable. The logits of the true areas were drawn from a bivariate normal distribution with a different mean for each test (logit(0.80) and logit(0.87)) and a covariance matrix with standard deviations 0.1 and 0.4, and correlation 0.6. The estimates of the areas were generated in the same way as in the previous scenario. The results of the analysis are presented in Tables 4.5a and 4.5b. Most entries in Table 4.5 do not suggest equivalence, with the exception of $P_2^{0.2}$. The latter is not surprising because the 0.2 percentiles of the distributions of the two tests are in fact quite close, despite the fact that most other characteristics of the distributions are not.

The assessment of equivalence may rest primarily on the 0.2 percentile criterion, for example, if the new test is considerably more expensive or invasive that the standard test. In such a situation a conservative strategy may suggest to stay with the current test, as long as the new does not improve the low end of the test performance.

	Criteria with $d(e_1, e_2) = \lvert e_1 - e_2 \rvert$, $\delta = .05$		Criteria with $d(e_1, e_2) = \lvert 1 - e_1 / e_2 \rvert$, $\delta = .08$	
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.05$
Pr ($\{P_1 > 1-\alpha\}$)	0.11	0.07	0.52	0.46
Pr ($\{P_2^{0.2} > 1-\alpha\}$)	0.45	0.37	0.74	0.67
Pr ($\{P_2^{0.8} > 1-\alpha\}$)	0	0.01	0.02	0
Pr ($\{P_2^{0.2,0.8} > 1-\alpha\}$)	0	0	0.02	0

Table 4.5a: Summary of simulations under scenario 2 for population equivalence. The entries show the proportion of times each criterion was larger than $1-\alpha$

	Criteria with $d(e_1, e_2) = \lvert e_1 - e_2 \rvert$, δ=.05		Criteria with $d(e_1, e_2) = \lvert 1 - e_1 / e_2 \rvert$, δ =.08	
	α =0.1	α =0.05	α =0.1	α =0.05
Pr ({ I_1 >1-α })	0	0	0	0
Pr ({ $I_2^{0.2}$ >1-α })	0	0	0	0

Table 4.5b: Summary of simulations under scenario 1 for individual reader equivalence. The entries show the proportion of times each metric was larger than 1-α

5. Discussion

In this thesis we presented population and individual formulations of equivalence of two diagnostic tests, as interpreted by a population of readers. We also discussed how the relevant quantities could be estimated from a Bayesian hierarchical model.

We assumed paired designs for the tests under comparison because such designs are commonly used and because they provide better control of confounding. We also focused the development of criteria and models on studies in which each scan is interpreted once by a given reader. However, the equivalence criteria may be applied to other study designs and other statistical models as well. For instance, the population equivalence metrics can be applied to uncorrelated data resulting from studies in which each reader uses only each test on a separate group of cases. The approach can also be extended to studies in which scans are interpreted repeatedly by the same reader.

Elaborations or simplifications of the basic hierarchical model can be considered to fit data arising from particular study designs or to incorporate additional sources of information. For example, the covariance matrices Σ_j could be treated as unknown parameters of the model instead of assuming them as known and equal to their estimates.

Such an elaboration of the model would make use of known approximations of the Σ_j as functions of the true areas and other parameters (e.g. *a* and *b* in the Binormal model – see Appendix). The choice of considering the *Level I* matrices Σ_j as known is common in hierarchical model analysis, primarily because it results in considerable reduction of the number of the parameters in the model. Such parsimony is particularly important in modeling data from diagnostic test studies, because such data are often based on a limited number of readers and allow a relatively small number of degrees of freedom.

The choice of the *Level II* distribution needs careful consideration. As noted earlier this distribution affects, to a large extent, the posterior estimates of individual reader performance measures. However, the evaluation of this assumption may be difficult because the sample size of readers will typically be limited. Normality of the reader effects is a common assumption in practice, both in the hierarchical model used in our analysis and in the mixed models used in other work on equivalence, as described in Chapter 1. One safeguard available to the analyst is to perform a sensitivity analysis using several *Level II* distributional assumptions. The alternatives to be considered may include unimodal symmetric distributions with heavier tails or non-symmetric distributions as well as multimodal distributions. However, an extended sensitivity analysis should be guided by the available information regarding the population of readers. It may also be reasonable to consider an analysis that is stratified by reader characteristics. For example in the example of section 4.1, 5 of the readers were general radiologists and 5 were specialized in mammography.

The hierarchical model described in Chapter 3 can be modified easily to accommodate other measures of diagnostic performance, as long as the normality assumption in *Level I*

can be supported. We note that the normality assumption will be met asymptotically by most estimates of measures of diagnostic performance as well as by suitable transformations, such as logits. Hence, the hierarchical model can be used with the appropriate specification of the covariance matrix $\hat{\Sigma}_j$. The form of this covariance matrix becomes simpler when the metric is sensitivity or specificity.

The proposed approach can be extended to handle situations in which multivariate measures of diagnostic performance are used. An important special case arises when both the sensitivity and the specificity are considered simultaneously. Equivalence may then be assessed through the appropriate probability Pr ($d\{e_1, e_2\} \leq \delta$) as defined in section 2.2. In this case, e_1 and e_2 are vectors containing the elements of the distributions (across readers) of the sensitivities and specificities for each test.

The general formulation of the equivalence criteria and they hierarchical model analysis proposed in this thesis can be applied beyond the diagnostic equivalence setting. Indeed, the approach can be used to define and assess equivalence in biomedical settings where patients are grouped and performance is measured for each grouping. For example, it may be of interest to assess equivalence in the effectiveness of two therapeutic procedures (e.g. two types of surgery), as performed in various hospitals. In that case population equivalence will compare the distributions of the response rates across hospitals and individual equivalence will be defined in the hospital level.

It should also be noted that the approach discussed in this thesis may also be used to handle the assessment of *non-inferiority*. In studies of non-inferiority the question of interest is whether one test (or therapy) is at least as good as the other, instead of whether the two tests are equivalent. The assessment of non-inferiority can proceed essentially as

the assessment of equivalence, after modifying the metrics by removing the absolute values from the *d*'s. For example, in order to test whether test 2 is non-inferior of test 1 in terms of their population means, we may estimate the value of the corresponding metric P_1, which in this case will be equal to Pr($d(\mu_1, \mu_2) \leq \delta$).

APPENDIX

We will illustrate how to obtain estimates of Σ_j 's in the case where the Y_{ij} 's are obtained from the binormal model with test results on an ordinal scale. Under that model assume that X_D and X_N refer to the underlying continuous test results of the diseased and non-diseased patients respectively and that they are normally distributed with means μ_D, μ_N and variances σ_D^2, σ_N^2. Let $\alpha = ((\mu_D - \mu_N)/\sigma_D$ and $b = \sigma_N / \sigma_D$. Then the area under the ROC curve equals $A = \int (\alpha + b\nu)\phi(\nu)d\nu$, and we can estimate it by $Y = \Phi[\hat{a}/(\sqrt{1+\hat{b}^2})]$, where $\hat{a}, \hat{b}$ are the MLE's of α and *b* respectively.

In our case there is a pair of correlated estimates since the observations are taken from the same reader and patients. Hence we have $\hat{a}_1, \hat{a}_2, \hat{b}_1, \hat{b}_2$ which as MLE's are asymptotically normal with mean their rue value and variance the Fisher Information matrix. Using Delta method, we can approximate their variance Σ_j by the following formulae:

$$\text{var}(Y_i) = f_i^2 \, \text{var}(\hat{a}_i) + g_i^2 \, \text{var}(\hat{b}_i) + 2 f_i \, g_i \, \text{cov}(\hat{a}_i, \hat{b}_i) \qquad (8)$$

$$\operatorname{cov}(Y_1, Y_2) = f_1 f_2 \operatorname{cov}(\hat{a}_1, \hat{a}_2) + g_1 g_2 \operatorname{cov}(\hat{b}_1, \hat{b}_2) + f_1 g_2 \operatorname{cov}(\hat{a}_1, \hat{b}_2) + f_2 g_1 \operatorname{cov}(\hat{a}_2, \hat{b}_1) \quad (9)$$

where:

$$f_i = \frac{\exp\left(-\frac{a_i^2}{2(1+b_i^2)}\right)}{\sqrt{2\pi(1+b_i^2)^3}}, \quad g_i = \frac{-a_i b_i \exp\left(-\frac{a_i^2}{2(1+b_i^2)}\right)}{\sqrt{2\pi(1+b_i^2)^3}}$$

In order to compute an approximation we can use the observed information matrix evaluated at the MLE estimate, together with some estimates for the variances and covariances involved. Obuchowski and McClish (1997), approximated those estimates with Taylor expansion series:

$$\widehat{\operatorname{var}}(\hat{a}_i) = \frac{1 + \hat{b}_i^2 / R + \hat{a}_i^2 / 2}{N_D}, \quad \widehat{\operatorname{var}}(\hat{b}_i) = \frac{1 + \frac{\hat{b}_i^2 (1+R)}{2R}}{N_D}, \quad \widehat{\operatorname{cov}}(\hat{a}_i, \hat{b}_i) = \frac{\hat{a}_i \hat{b}_i}{2N_D}$$

$$\widehat{\operatorname{cov}}(\hat{a}_1, \hat{a}_2) = \frac{\hat{r}_D + \frac{\hat{r}_N \hat{b}_1 \hat{b}_2}{R} + \frac{\hat{r}_D^2 \hat{a}_1 \hat{a}_2}{2}}{N_D}, \quad \widehat{\operatorname{cov}}(\hat{b}_1, \hat{b}_2) = \frac{\hat{b}_1 \hat{b}_2 (\hat{r}_N + R\, \hat{r}_D^2)}{2RN_D}$$

$$\widehat{\operatorname{cov}}(\hat{a}_1, \hat{b}_2) = \frac{\hat{r}_N^2 \hat{a}_1 \hat{b}_2}{2N_D}, \qquad \widehat{\operatorname{cov}}(\hat{b}_1, \hat{a}_2) = \frac{\hat{r}_N^2 \hat{b}_1 \hat{a}_2}{2N_D}$$

where N_D, N_N is the number of diseased and non-diseased patients respectively, *R* the ratio N_N / N_D and $\hat{r}_D, \hat{r}_N$ are the sample correlation coefficients of the test results between the two tests in the diseased and non-diseased patients respectively.

An estimate of Σ_j can be obtained by substituting the unknown terms with their estimates in (8) and (9). It is not difficult to show that if the observed information matrix is positive definite then $\hat{\Sigma}_j$ will also be positive definite.

Finally, an estimate of Σ_j can be obtained without the assumption of the binormal model. For example if the Y_{ij}'s represent non-parametric estimates (obtained from the Mann-Whitney statistic) we can again calculate $\hat{\Sigma}_j$ in a similar fashion (Delong et al)**.**

References

1. Alonzo T., Pepe M. and Moskowitz C. (2002). Sample size calculations for comparative studies of medical tests for detecting presence of disease. *Statistics in Medicine* 21, 835-852

2. Anderson S. and Hauck W. (1990) Consideration of individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics* 18, 259-273

3. Barnard, J., McCulloch R. and Meng, X. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage. *Statistica Sinica* 10, 1281-1311

4. Blackwelder, W. (1982) Proving the null hypothesis in clinical trials. *Controlled Clinical Trials* 3, 345-353

5. Box G.P.E Tiao G. C (1973) Bayesian Inference in Statistical Analysis. Reading, Mass.: Addison-Wesley

6. Daniels M.J. and Kass R.J. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of American Statistical Association*, 94, 1254-1263

7. DeLong E.R., DeLong D.M. and Clarke-Pearson D.L. (1988). Comparing the areas under two or more correlated ROC curves: A Nonparametric approach. *Biometrics*, 44, 837-845

8. Du Mouchel W. Hierarchical Bayes linear models for meta-analysis. Technical no. 27, *National Institute of Statistical Sciences Research Triangle Park, NYC*

9. Hanley J.A. and McNeil B.J. (1983). A method for comparing areas under ROC curves derived from the same set of cases. *Radiology*, 148, 839-843

10. Hyslop T., Hsuan F. and Holder D.J. (2000). A small sample confidence interval approach to assess individual bioequivalence. *Statistics in Medicine*, 19, 2885-2897

11. McNally R.J., Iyer H. and Mathew T. (2003). Tests for individual and population bioequivalence based on generalized p-values. *Statistics in Medicine*, 22, 31-53

12. Metz C.E. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8, 283-298

13. Obuchowski N.A. (2001). Can electronic medical images replace hard copy film? Defining and testing the equivalence of diagnostic tests. *Statistics in Medicine*, 20, 2845-2863

14. Obuchowski N.A. (1995). Multireader, multimodality ROC curve studies: Hypothesis testing and sample size estimation using an analysis of variance approach with dependent observations. *Academic Radiology*, 2, 709-716

15. Obuchowski N.A. and McClish D.K. (1997). Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. *Statistics in Medicine*, 16, 1529-1542

16. Schall R. and Luus H. (1993). On population and individual bioequivalence. *Statistics in Medicine* 12, 1109-1124

17. Schall R. (1995). Assessment of individual and population bioequivalence using the probability that bioavaillabilities are similar. *Biometrics* 51, 615-626

18. Selwyn M, Dempster A. and Hall N (1981) A Bayesian approach to bioequivalence for the 2×2 changeover design. *Biometrics* 37, 11-21

19. D J Spiegelhalter and A Thomas and N G Best (1999). WinBUGS Version 1.2 User Manual. MRC Biostatistics Unit

20. Toledano A.Y. and Gatsonis C. (1996). Ordinal regression methodology for ROC curves derived from correlated data. *Statistics in Medicine*, 15, 1807-1826

21. US Food and Drug Administration. Statistical approaches to establishing equivalence. Rockville, MD

22. Wang F. and Gatsonis C. Hierarchical models in multi-reader multi-modality ROC studies. *In preparation.*

23. Wellek S. (2000a) Bayesian construction of an improved parametric test for probability-based individual bioequivalence. *Biom. J*. 42, 1039-1052

24. Wellek S. (2000b) On a reasonable disaggregate criterion of population bioequivalence admitting of resampling-free testing procedures. *Statistics in Medicine*, 19, 2755-1767

25. Zhou X., McClish D. and Obuchowski N. (2002). Statistical methods in Diagnostic medicine. Wiley, New York