Maastricht University | universiteit hasselt

2015 | Faculty of Sciences

DOCTORAL DISSERTATION

On the estimation and validation of biomarker-index' accuracy

Doctoral dissertation submitted to obtain the degree of
Doctor of Science: Statistics, to be defended by

Leandro García Barrado

Promoter: Prof. Dr Tomasz Burzykowski

D/2015/2451/54

universiteit hasselt | I-BioStat | KU LEUVEN

Interuniversity Institute for Biostatistics
and statistical Bioinformatics

Acknowledgements

Vooraleer een belangrijke levensgebeurtenis te beeïndigen is het vaak goed om even achterom te kijken en een woord van dank uit te spreken aan iedereen die heeft bijgedragen aan het resultaat van deze 4-jarige tocht.

First and foremost, I would like to thank my promoter, Tomasz. I consider it a huge privilege to have been given the opportunity to think together with you for the last four years. Without your help, suggestions, exceptional clear insight, and unmeasurable enthusiasm, this would have been four long years. In één adem zou ik hierbij ook Els willen bedanken. Ik zou niet kunnen indenken hoe deze doctoraatsopleiding verlopen zou zijn zonder de dagen aan vergaderingen en brain-storm sessies om schijnbaar onoplosbare problemen op te lossen. The memory of our numerous trips to our friends of the north as the three-headed BRAVO-team will always bring a smile to my face, despite the occasional frustration we have shared. Many thanks, Tomasz; ontzettend bedankt, Els.

Also many thanks, to all the members of the jury for taking the time to read the dissertation and for providing valuable comments and suggestions on a draft version of the following text.

I also want to express my gratitude to the University of Hasselt and Censtat, for giving me the opportunity to meet many interesting people at numerous conferences and seminars both national and international. In particular, I want to thank all my Censtat officemates and colleagues. The helpfulness, good fellowship, and friendly environment over the last four years, have been highly appreciated.

A special thanks also to IDDI, for partly financing my PhD. But even more so for the extremely warm, hospitable, and enthusiastic atmosphere at any of my bi-weekly visits. Thank you very much, for all the educational lunch breaks, the short coffee-break talks and especially 'les cours de français'.

Deze vier jaren waren ook nooit tot een goed einde gekomen zonder de welgekomen afleiding van vrienden en familie. De feestjes, uitstappen, muzikale uitspattingen of gewoon de simpele rustige gesprekjes. Merci allemaal voor deze generische medicijnen voor een gezonde geest!

Zonder mijn ouders en grote kleine broer had ik dit avontuur, of alle anderen, niet eens kunnen aanvatten. Mama en papa, bedankt voor het nooit eindigende vertrouwen, het altijd klaarstaan en het aanwakkeren van een kritische blik op de wereld maar steeds met respect en eerlijkheid. Merci, Fernando voor de oneindige mini-avonturen, humor, opgewektheid en als belangrijkste uitdager van de meegekregen kritische blik. Er is in de statistiek veel onzekerheid, maar dat ik uit het beste gezin ter wereld kom, dat staat vast.

Als laatste gaat mijn dank uit naar Ann, mijn levensmaatje. Bedankt voor de onmetelijke steun en interesse in alles wat ik doe. Het 'jou' zijn en mij steeds terug te brengen naar de dingen die belangrijk zijn. Mijn eeuwige dank!

And to everybody I've failed to acknowledge, thank you very much!

Leandro García Barrado
Diepenbeek, 18 september 2015

List of Publications

The materials presented in the dissertation are based on the following publications:

García Barrado L., Coart E., Burzykowski, T. (2013) Estimation of optimally combined-biomarker accuracy in the absence of a gold standard reference test. In: Lanzarone, E.; Ieva, F. (Ed.). The contribution of young researchers to Bayesian statistics: Proceedings of BAYESM2013, p7-10. *Springer proceedings in Mathematics & Statistics, 63.*

Coart E., **García Barrado L.**, Duits, F.H., Scheltens, P., van der Flier, W.M, Teunissen, C.E., van der Vies, S..M., Burzykowski, T. (2015) Correcting for the absence of a gold standard improves diagnostic accuracy of biomarkers in Alzheimer's disease. *Journal of Alzheimer's Disease* **46** (4), 889-899.

García Barrado L., Coart E., Burzykowski, T. (2015) Development of a diagnostic test based on multiple continuous-biomarkers with an imperfect reference-test. *Accepted for publication in Statistics in Medicine.*

García Barrado L., Coart E., Burzykowski, T. (2015) Estimation of diagnostic accuracy of a combination of continuous biomarkers allowing for conditional dependence between the biomarkers and the imperfect reference test. *Revision submitted to: Biometrics.*

García Barrado L., Coart E., Burzykowski, T. (2015) A Bayesian Framework Allowing Incorporation of Retrospective Information in Prospective Diagnostic Biomarker-validation Designs. *Submitted to: Statistics in Biopharmaceutical Research.*

García Barrado L., Coart E., Vanderstichele H.M.J., Burzykowski, T. (2015) Transferring cut-off values between assays for cerebrospinal fluid Alzheimer's disease biomarkers. *Accepted for publication in Journal of Alzheimer's Disease.*

Contents

List of abbreviations

Here, we give a list of the most often used abbreviations in the dissertation.

GS	:	Gold Standard
Se	:	Sensitivity
Sp	:	Specificity
ROC	:	Receiver Operating Characteristic
AD	:	Alzheimer's Disease
CSF	:	Cerebrospinal Fluid
p-tau	:	Phosphorylated tau
ADNI	:	Alzheimer's Disease Neuroimaging Initiative
VUmc	:	Vrije Universiteit Amsterdam Medisch Centrum
NIA	:	National Institute on Aging
NIBIB	:	National Institute of Biomedical Imaging and Bioengineering
FDA	:	Food and Drug Administration
SMC	:	Subjective Memory Complaints
EEG	:	Electroencephalography
MRI	:	Magnetic Resonance Imaging
MMSE	:	Mini Mental State Examination
NINCDS-ADRDA	:	National Institute of Neurological and Communicative Disorders and Stroke - Alzheimer's Disease and Related Disorders Association
NIA-AA	:	National Institute on Aging - Alzheimer's Association

Chapter 1

Overview of the dissertation

The current and future importance and impact of Alzheimer's Disease (AD) on society is hard to overstate. The number of dementia patients is about 9.2 million in Europe, about 177.000 in the Belgian population, more or less 100.000 of which live in Flanders, with Alzheimer's accounting for the majority of the cases [31, 120].

Advances in AD research are hindered by issues related to the diagnosis of the disease. Still no treatment for AD is available and the efficacy of currently available symptom-directed treatments depend highly on how early in the disease process they are administered. Current AD diagnosis heavily depends on clinical manifestation of the disease. Hence, patients are only diagnosed in later stages of the disease, resulting in small symptom-directed treatment benefits. Moreover, the diagnosis is prone to misclassification, making it imperfect. Therefore, it is not very well suited to be used in the development of new drugs. Alternatively, the gold standard (GS) diagnostic test of AD can only be established post-mortem by brain tissue dissection. This AD diagnosis is costly to obtain while from a diagnostic point of view, it is useless.

Consequently, AD research is steering into the direction of the development of diagnostic biomarkers. Preferably, these biomarkers should be measurable relatively easy, e.g., in cerebrospinal fluid (CSF), blood or making use of imaging techniques, and be related to the pathological process preceding the manifestation of clinical symptoms. Formal development of such biomarkers has been found disappointing. Many biomarkers fail to statistically show adequate and satisfactory diagnostic accuracy, while clinicians feel they perform very well in practice. The discrepancy between study results and practical performance may not be surprising since in these studies the current clinical diagnosis of AD is wrongfully considered to be a GS reference-test,

leading to biased results of the biomarkers' accuracy.

Because of the lack of suitable validation methods and enormous costs due to huge sample sizes, established biomarkers are usually applied in practice without any formal validation. In addition, currently measured AD CSF-biomarkers lead to different values on different commercially available platforms. This implies that for every platform, different diagnostic cut-offs should be developed. In practice, to avoid the costly collection of GS post-mortem data, a cut-off for one platform is usually translated to another using linear-regression-based methods. The characteristics of these methods have never been investigated in detail and the low precision of the resulting estimates is usually ignored.

To address these issues, some novel methods are presented in the following dissertation. An introduction to diagnostic test development and Alzheimer's Disease research is presented in Chapter 2. Chapter 3 introduces a Bayesian latent-class mixture model to allow for the development of a diagnostic biomarker-index in the absence of a GS reference-test. In Chapter 4, the method is extended to allow for conditional dependence between the continuous biomarkers of interest and the dichotomous imperfect reference-test. A novel Bayesian validation approach is developed in Chapter 5. We show that by allowing for a reasonable dependence between the development and validation data, a large gain in the efficiency of biomarker accuracy validation is obtained, dramatically reducing the validation-study sample size. Concerning the transfer of AD CSF-biomarker cut-offs from one platform to another, we investigate the characteristics of the currently applied transfer method in Chapter 6. In addition, we develop a new Bayesian transfer method and show that it is unbiased in all considered settings while being more efficient than the current linear-regression-based transfer method.

In conclusion, Chapter 7 contains a general discussion on the proposed methods and introduces topics for further research.

Chapter 2

Introduction

The following sections contain a general introduction to the field of diagnostic test development and Alzheimer's Disease (AD) research. The fundamental concepts underlying diagnostic tests are introduced along with measures of overall diagnostic accuracy in Section 2.1. In Section 2.2, we zoom in on the construction of a diagnostic index and discuss the advantages of considering a combination of diagnostic tests over single diagnostic tests. Special care has to be taken to estimate the performance of a new diagnostic test when no gold-standard reference-test is available. Models resolving this issue are discussed in Section 2.3. In Section 2.4, the preference for the Bayesian method is clarified and currently applied diagnostic accuracy methods are described. Biomarkers are excellent candidates for the construction of diagnostic tests; the importance of biomarkers in Alzheimer's disease research is scrutinized in Section 2.5. Finally, two data sets are introduced which are closely investigated throughout the dissertation. Discussion of these data sets can be found in Section 2.5.2.

2.1 Diagnostic test performance

A diagnostic test is a test constructed to discriminate between patients with and without a certain condition. Within the field of diagnostic medicine a good diagnostic test can serve several purposes [133]. Among others, it provides health care providers with essential information about a patient's condition, guides the health care provider in setting up an appropriate treatment plan, and allows understanding of disease mechanisms by investigating changes in diagnostic outcomes over time. In order

for health care providers to be able to select the appropriate diagnostic test, its performance should be investigated. Diagnostic test performance can be evaluated by several diverse measures, such as costs to society, diagnostic accuracy, effect on patient outcome, etc. Fryback and Thornbury [34] have proposed to hierarchically order these measures such that if a diagnostic test is considered non-efficacious on a lower level it is deemed non-efficacious at all higher levels. The main focus of this dissertation will be on the accuracy of diagnostic tests (level 2 according to Fryback and Thornbury's [34] hierachy as shown in Table 2.1).

Table 2.1: A hierarchical model of efficacy applied to medical diagnostic imaging according to Fryback and Thornbury. Taken from [34].

Level 1. Technical efficacy
Resolution of line pairs
Modulation transfer function change
Gray-scale range
Amount of mottle
Sharpness
Level 2. Diagnostic accuracy efficacy
Yield of abnormal or normal diagnoses in a case series
Diagnostic accuracy (percentage correct diagnosis in case series)
Predictive value of positive or negative examination (in a case series)
Sensitivity and specificity in a defined clinical problem setting
Measures of ROC curve height (d') or area under the curve A_z
Level 3. Diagnostic thinking in efficacy
Number (percentage) of cases in a series in which image judged 'helpful' to making the diagnosis
Entropy change in differential diagnosis probability distribution
Difference in clinicians' subjectively estimated diagnosis probability pre- to posttest information
Empirical subjective log-likelihood ratio for test positive and negative in a case series
Level 4. Therapeutic efficacy
Number (percentage) of times image judged helpful in planning management of the patient in a case series
Percentage of times medical procedure avoided due to image information
Number or percentage of times clinicians' prospectively stated therapeutic choices changed after test information
Level 5. Patient outcome efficacy
Percentage of patients improved with test compared with without test
Morbidity (or procedures) avoided after having image information
Change in quality-adjusted life expectancy
Expected value of test information in quality-adjusted life years (QALYs)
Cost per QALY saved with image information
Level 6. Society efficacy
Benefit-cost analysis from societal viewpoint
Cost-effectiveness analysis from societal viewpoint

Zweig and Campbell [136] define diagnostic-test accuracy as the ability of the test to correctly classify subjects into clinically meaningful subgroups. For example, if diseased and healthy patients have indistinguishable test results, the test is said to have negligible accuracy; if test results show no overlap, the test has perfect accuracy. Of course, most tests have some discriminative ability while displaying overlapping results so their accuracy will be somewhere in between these two extremes. In order to quantify the accuracy of a diagnostic test, diagnostic-test accuracy studies are performed. Generally, these diagnostic-test accuracy studies are defined by the following three key attributes. First, subjects who will undergo the diagnostic test of interest have to be sampled. Second, some sort of interpretation or decision rule based on the test results has to be constructed. Finally, the results of the obtained diagnostic test results have to be compared to the results of a reference test. Preferably, this reference test has perfect accuracy and is then termed as a *gold standard* (GS) reference-test. Results based on these three attributes can be used to construct the concepts defining diagnostic test accuracy.

2.1.1 Estimating accuracy

The measures of accuracy considered in the following section describe the *intrinsic accuracy* of the diagnostic test. Intrinsic accuracy is concerned with the inherent ability of the test to correctly identify a condition when it is present and at the same time correctly identify the absence of a condition when it is not present [133]. In particular, these measures compare the test results to the true condition state of the patients. It is important to note that the intrinsic accuracy of a diagnostic test is independent of the condition's prevalence in the test sample. For the remainder of the dissertation we will assume that there are only two mutually exclusive clinically relevant subgroups, a group for which the *condition is present* and another for which the *condition is absent*. Furthermore, the *condition is present* and *condition is absent* groups will generally be referred to as the *case* and *control* groups, respectively, without implying any specifics to case-control studies.

If the diagnostic test under evaluation is dichotomous in nature, resulting in a subgroup membership value for each subject in the test sample, its accuracy can be represented by two probabilistic statements. Denote the true condition state of the N patients in the study sample by GS reference-test result variable D, so that $D = 0$ for controls and $D = 1$ for cases. Next, represent the results of the dichotomous diagnostic test by variable T, so that $T = 0$ represents diagnosis to the control group and $T = 1$ indicates diagnosis to the case group. Subsequently we can define the probability of correctly identifying cases as $P(T = 1|D = 1)$, referred to as the

Sensitivity of test T (Se_T). The probability of correctly identifying controls is then defined by $P(T = 0|D = 0)$, known as the *Specificity* of the dichotomous diagnostic test T (Sp_T).

An example of a diagnostic accuracy study on N subjects can be found in Table 2.2. From the N subjects, n_1 suffer from the condition and for n_0 the condition is absent as indicated by the GS reference-test results. The results from the diagnostic test of interest categorize m_1 subjects to the case group while m_0 subjects receive the control label. The number of subjects for which the results from the GS reference-test and diagnostic test agree are indicated by s_1 and r_0, for the case and control diagnosis, respectively. Finally, s_0 subjects for which the condition is present receive a control diagnosis, while r_1 condition-free subjects receive the case diagnosis.

The Se_T of the diagnostic test under evaluation can then be estimated by the proportion of subjects for which the condition is present who receive the case diagnosis: $\widehat{Se}_T = s_1/n_1$. The proportion of condition-free subjects who receive the control diagnosis is then an estimate of the Sp_T of the diagnostic test in the considered example: $\widehat{Sp}_T = r_0/n_0$.

Table 2.2: Example of results of a dichotomous diagnostic test study.

	Test Result:		
True Condition State	**Negative ($T = 0$)**	**Positive ($T = 1$)**	**total**
Absent ($D = 0$)	r_0	r_1	n_0
Present ($D = 1$)	s_0	s_1	n_1
total	m_0	m_1	N

When the diagnostic test under evaluation T is a continuous test, it is not possible to express diagnostic accuracy by a single Se_T and Sp_T pair. Every value within the range of possible diagnostic test result values can serve as a cut-off c to dichotomize T. Specifically, a subject can be classified to the *case* group when $T > c$ and to the control group when $T \leq c$, resulting in a Se and Sp pair for every threshold c. The accuracy of such a diagnostic test can be described by the receiver operating characteristic (ROC) curve [135, 136, 114]. The ROC curve shows the probability of correctly identifying a case, i.e. Se, against the probability of falsely identifying a control as a case, i.e., $1 - Sp$ for ordered values of c.

Methods to estimate ROC curves for continuous diagnostic-tests largely fall into two groups, non-parametric and parametric. Empirical or non-parametric methods involve plotting pairs of Se and $1-Sp$, calculated using the empirical survival curves of the cases and controls, for each possible cut-off value c. No particular parametrisation or distributional assumptions regarding the continuous test need to be specified [136].

The disadvantage of these methods is that the obtained ROC curves are usually not smooth; see for example the black empirical ROC curve in Figure 2.1. For this reason, non-parametric ROC curve methods involving kernel density functions, which allow obtaining smooth ROC curves, were proposed [134, 58, 132, 100]. Rufibach [100] proposes to use an estimator based on log-concave density estimates which, among others, is implemented in the R-package pROC [97]. The red curve in Figure 2.1 is the log-concave smoothed counterpart of the black empirical ROC curve.

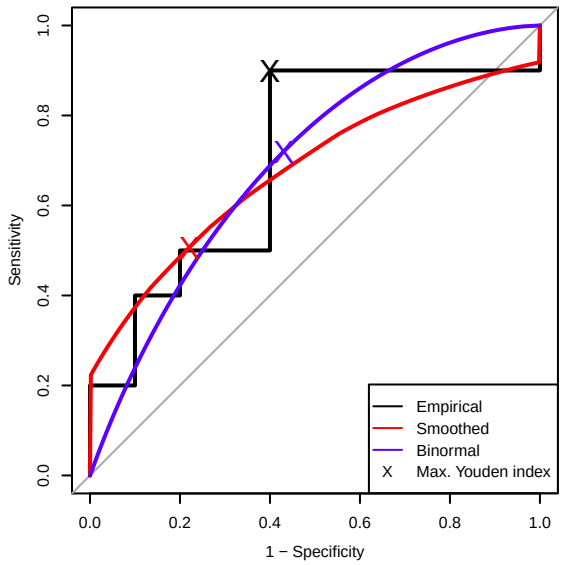

Figure 2.1: Example of ROC curve estimates. Empirical ROC curve is depicted in black, the smoothed and binormal ROC curve in red and blue, respectively. The grey line indicates the unity diagonal. Maximum Youden index for each ROC curve estimate is indicated by the respectively coloured X.

Parametric methods assume a particular distributional form for the diagnostic-test values for cases and controls. Often, normal distributions are assumed. Assume that a continuous diagnostic-test T is normally distributed conditionally on the value of D. Specifically, $T|D = d \sim N(\mu_d, \sigma_d^2)$ for $d = 0$ for true controls and $d = 1$ for true cases. Under this so-called *binormal* assumption, the ROC curve of continuous diagnostic-test T is defined by

$$Se_{(1-Sp)} = \Phi\left(\alpha + \beta\Phi^{-1}\left(1 - Sp\right)\right), \tag{2.1}$$

where $Se_{(1-Sp)}$ denotes the sensitivity for each value of $1 - Sp$, $\Phi(\cdot)$ indicates the cumulative standard-normal distribution function, $\alpha = (\mu_1 - \mu_0)/\sigma_1$ and $\beta = \sigma_0/\sigma_1$. Maximum-likelihood estimates of means and variances of the normal distributions can be used directly to estimate the ROC curve [70, 71, 133] and its statistics [61]. An example of a binormal ROC curve is depicted by the blue solid line in Figure 2.1.

Alternatively, estimates can be obtained by using a Bayesian estimation approach. A Bayesian regression method has been proposed by O'Malley et al. [78], while Gu and Ghosal [41] have developed a method using a rank likelihood approach.

2.1.2 Summary measures of accuracy

Although the ROC curve is an elegant way to represent the properties of a diagnostic test, in practice, summary measures based on the curve are often used. One of such measures is the maximum Youden index (J) [130]. The Youden index is defined as $Se_c + Sp_c - 1$ for a particular cut-off c. The point on the ROC curve furthest away from no differentiation — the identity diagonal in ROC-space (grey line in Figure 2.1) — constitutes the maximal Youden index. Because of its link to a specific cut-off, the maximal Youden index is also proposed as an optimal diagnostic test cut-off criterion [86]. The maximum Youden indices are indicated in Figure 2.1 by the Xs in the corresponding ROC curve estimates.

Another summarizing measure is the area under the ROC curve (AUC). It can be interpreted as the probability that, when provided with a random case and a random control, the case will have a higher diagnostic test value than the control [7]. Faraggi and Reiser [32] discuss the estimation of AUC for several ROC estimation methods. Under the binormal ROC curve assumption, the AUC can be defined based on Equation 2.1 as follows:

$$AUC = \Phi\left(\frac{\alpha}{\sqrt{1+\beta^2}}\right). \tag{2.2}$$

2.2 Diagnostic index

One can expect that a combination of diagnostic tests can offer a better diagnostic accuracy than a single test. Therefore, much interest has been focused on developing and evaluating performance of diagnostic tests based on a combination of several tests [85, 81, 82]. Specifically, improvement of diagnostic accuracy has been shown for CSF AD-biomarkers as well [63]. Often, combinations of tests are constructed by using logistic-regression-type models, in which different tests are related to the

disease status [133]. Another approach is to combine tests in such a way that the obtained combination is optimal with respect to some measure of diagnostic accuracy [135]. In particular, the AUC can be used as the criterion, which can be optimized by using a model-free approach [85, 87, 44], a discriminant-function approach [77], or a fully-parametric approach [112]. All of these approaches combine several diagnostic tests into a single composite diagnostic test, the values of which will be denoted by the term *diagnostic index*.

2.2.1 Fully-parametric approach to combine tests

In this dissertation we consider the fully-parametric approach to combine diagnostic tests proposed by Su and Liu [112] and revisited by Liu, Schisterman and Zhu [57]. In this approach, which is defined under the binormal assumption, the linear combination maximizing AUC is defined by coefficients $\boldsymbol{a}$ proportional to a function of the assumed normal distribution parameters. Consider measurements on K diagnostic tests $\boldsymbol{y} = (Y_1, ..., Y_K)^T$ which are assumed to have a joint normal distribution conditional on true disease status D. For the true controls ($D = 0$) the K-variate normal distribution is defined by mean vector $\boldsymbol{\mu_0} = (\mu_{0,Y_1}, ..., \mu_{0,Y_K})^T$ and variance-covariance matrix $\boldsymbol{\Sigma_0}$ defined as:

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} \sigma^2_{0,Y_1} & \cdots & \rho_{0,Y_1,Y_K}\sigma_{0,Y_1}\sigma_{0,Y_K} \\ \vdots & \ddots & \vdots \\ \rho_{0,Y_K,Y_1}\sigma_{0,Y_K}\sigma_{0,Y_1} & \cdots & \sigma^2_{0,Y_K} \end{pmatrix}. \tag{2.3}$$

$\boldsymbol{\Sigma}_0$ contains the test-specific variances $\sigma^2_{0,k}$ and between-test correlations ρ_{0,Y_k,Y_j} for $k, j \in (1, \dots K)$ and $k \neq j$. The diagnostic test values of the true cases are assumed to follow a K-variate normal distribution with mean vector equal to $\boldsymbol{\mu_1} = (\mu_{1,Y_1}, ..., \mu_{1,Y_K})^T$ and variance-covariance matrix $\boldsymbol{\Sigma}_1$, which is defined as in Equation 2.3 but with test-specific variances $\sigma^2_{1,k}$ and between-test correlations ρ_{1,Y_k,Y_j} for $k, j \in (1, \dots K)$ and $k \neq j$.

The linear combination maximizing AUC is then defined as $\boldsymbol{y}^T\boldsymbol{a}$ where $\boldsymbol{y}$ is the column-vector containing the observations for all K diagnostic tests and $\boldsymbol{a}$ is defined as [57]

$$\boldsymbol{a} \propto (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1} (\boldsymbol{\mu_1} - \boldsymbol{\mu_0}). \tag{2.4}$$

From Equation 2.4 we can see that the coefficients leading to the linear combination which maximizes AUC are proportional to the difference in means between the case

and control normal distributions scaled by the inverse of the sum of the respective variance-covariance matrices. The AUC of the so-obtained diagnostic index is then given by a complex function of parameters of the case and control K-variate normal distributions:

$$AUC_a = \Phi\left\{\sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}\right\}. \tag{2.5}$$

2.3 The absence of a GS reference-test

An important issue in the development of diagnostic tests, and therefore also of diagnostic indices, is the availability of the correct case and control labels. Often, it is assumed that a GS reference-test is available. As stated before, a GS reference-test provides perfect discrimination between cases and controls. In practice, such a reference test may not be available. For example, in the context of dementia and Alzheimer's disease (AD), only post-mortem pathological confirmation on brain tissue can be regarded as a GS reference-test [102]; however, the confirmation is useless from a diagnostic perspective.

Hence, in practice, the case and control labels are often based on the result of an imperfect reference-test. Such a test may misclassify cases and controls. If the misclassification is ignored in the development of a diagnostic test or combination of tests, the parameters describing the accuracy of the developed test may be severely biased.

2.3.1 Falsely assuming GS information leads to bias

The direction of the bias is related to the correlation between the misclassification errors of the reference test and of the diagnostic test of interest [117]. When, conditionally on the true disease status, the misclassification error of the reference test is independent of that of the diagnostic test of interest, the accuracy of the diagnostic test will be underestimated. This instance is referred to as the *conditional independence assumption*. In case the misclassification errors are correlated, the size of the correlation will determine the direction and magnitude of bias [59].

The following examples, based on Zhou et al. [133], illustrate the effect of falsely assuming a reference test to be a GS reference-test and the impact conditional dependence may have on the direction of this effect. The first example shows how true underlying Se_T and Sp_T of a new dichotomous diagnostic test T may be underestimated. Consider that the imperfect reference-test has sensitivity $Se_R = 0.9$ and

specificity $Sp_R = 0.7$. Moreover, assume that the new test and imperfect reference-test are independent given the patients' true disease status (the conditional independence assumption). The underlying true Se_T and Sp_T of the new test, the parameters we actually want to estimate, are 0.8 and 0.6, respectively. Assume that the study sample consists of 100 patients with the condition and 100 patients without the condition. Given the Se_R and Sp_R of the imperfect reference-test, of all patients with the condition, 90 are expected to respond positively and 10 negatively to the imperfect reference-test. In the group of patients without the condition, the imperfect reference-test will find 30 positive and 70 negative responders. This means that in total, the imperfect reference-test will have identified 120 cases (90 correctly identified true cases + 30 misclassified true controls) and 80 controls (10 misclassified true cases + 70 correctly identified true controls). Given that the true underlying Se_T of the new test is 0.8, 72 of the 90 true cases correctly identified by the imperfect reference-test will respond positively to the new test; the remaining 18 will respond negatively. Of the 30 true controls who have responded positively to the imperfect reference, 12 respond positively to the new test and 18 have a negative test results; because the true Sp_T of the new test is 0.6. This yields a total of 84 (= 72 + 12) positive and 36 (= 18 + 18) negative responders to the new test in the group of 120 patients having a positive imperfect reference-test outcome ($R = 1$ row in Table 2.3).

Similarly, 28 of the 70 correctly identified true controls by the imperfect reference-test will test positive on the new test, while the remaining 42 will respond negatively. From the 10 misclassified true cases, 8 will respond positively and 2 negatively to the new test. In total, from the 80 negative imperfect reference-test responders 36 (= 28 + 8) will have a positive and 44 (= 42 + 2) a negative new test result ($R = 0$ row in Table 2.3). Using the imperfect reference-test as a GS reference-test would imply to use the results contained in Table 2.3 to estimate Se_T and Sp_T of the new test. This would lead to an estimate of Se_T equal to 0.7 (84/120) and Sp equal to 0.55 (44/80). As compared to the true values of 0.8 and 0.6, respectively, these estimates are too small.

Table 2.3: Example of biased estimation of diagnostic-test accuracy when falsely assuming GS information under the conditional independence assumption.

	Test Result:		
Reference test	**Negative ($T = 0$)**	**Positive ($T = 1$)**	**total**
$\boldsymbol{R = 0}$	44	36	80
$\boldsymbol{R = 1}$	36	84	120
total	80	120	200

The second example is an illustration of how the Se_T and Sp_T of a new test can be overestimated by falsely considering GS information from an imperfect reference-test. The operating characteristics of both the imperfect reference-test and new dichotomous diagnostic-test are the same as before: $Se_R = 0.9$, $Sp_R = 0.7$, $Se_T = 0.8$, and $Sp_T = 0.6$. To achieve overestimation, the new test and imperfect reference-test are now assumed to be conditionally dependent. In other words, the new test and imperfect reference-test have a tendency to misclassify the same patients. The dependence is defined as shown in Table 2.4; for the true cases ($D = 1$), 10% will negatively respond to the imperfect reference-test as well as the new test $[P(R = 0, T = 0|D = 1) = 0.1]$, none will have a negative result on the imperfect reference-test and a positive result on the new test $[P(R = 0, T = 1|D = 1) = 0]$, and 10% will have a positive result of the imperfect reference but a negative outcome for the new test $[P(R = 1, T = 0|D = 1) = 0.1]$. Similarly for patients without the condition ($D = 0$), 10% will have a negative outcome of the imperfect reference-test and a positive result for the new test $[P(R = 0, T = 1|D = 0) = 0.1]$, none will have a positive imperfect reference-test result and a negative result of the new test $[P(R = 1, T = 0|D = 0) = 0]$ and 30% will have a positive result of both the imperfect reference and new test $[P(R = 1, T = 1|D = 0) = 0.3]$. The marginal Se and Sp of the imperfect reference and new test can be derived from the conditional joint probabilities summarized in Table 2.4; for instance, the true Se_R of the imperfect reference-test is then defined as $P(R = 1|D = 1) = P(R = 1, T = 0|D = 1) + P(R = 1, T = 1|D = 1) = 0.1 + 0.8 = 0.9$.

Table 2.4: Conditional joint probabilities of classification for an imperfect reference-test with marginal $Se_R = 0.9$ and $Sp_R = 0.7$ and new dichotomous diagnostic-test with $Se_T = 0.8$ and $Sp_T = 0.6$.

Reference test	$\boldsymbol{D=0}$		$\boldsymbol{D=1}$	
	$\boldsymbol{T=0}$	$\boldsymbol{T=1}$	$\boldsymbol{T=0}$	$\boldsymbol{T=1}$
$\boldsymbol{R=0}$	0.6	0.1	0.1	0
$\boldsymbol{R=1}$	0	0.3	0.1	0.8

In a study population consisting of 100 true cases and 100 true controls, the conditional joint probabilities from Table 2.4, will lead to the observed frequencies in Table 2.5 using a reasoning similar to the one used for the conditional independence example. Again, one can now estimate the Se_T and Sp_T of the new test considering the imperfect reference-test as a GS reference-test. This leads to overestimated values for $Se_T = 0.8$ and $Sp_T = 0.6$: $\widehat{Se}_T = 0.92\,(110/120)$ and $\widehat{Sp}_T = 0.88\,(70/80)$.

Table 2.5: Observed joint frequencies in the case of conditional dependence between the imperfect reference-test R and new dichotomous diagnostic test T as summarized in Table 2.4, for a sample of 100 true cases and controls.

	Test Result:		
Reference test	**Negative ($T = 0$)**	**Positive ($T = 1$)**	**total**
$\boldsymbol{R = 0}$	70	10	80
$\boldsymbol{R = 1}$	10	110	120
total	80	120	200

2.3.2 Latent-class models with two classes

To overcome the absence of a GS reference-test, latent-class models with two latent classes have been proposed to assess the accuracy of diagnostic tests. The models employ the EM-algorithm to obtain maximum likelihood estimates of diagnostic test accuracy and require certain strict identifiability restrictions. A traditional latent-class analysis was proposed by Rindskopf et al. [95] for dichotomous tests assuming conditional independence. Qu et al. [89] and Yang et al. [128] extended these ideas to allow for conditional dependence between dichotomous diagnostic tests by introducing continuous random effects modelling the dependence.

2.4 Preference for the Bayesian method

The traditional latent-class models mentioned in Section 2.3 usually ignore any reference test information and include only new-test data. Instead of completely ignoring imperfect reference-test information, preferably, one would like to weigh the information according to the prior knowledge about the accuracy of the test. For example, in AD-biomarker research, clinical diagnosis of AD can be regarded as an imperfect reference-test. Reports about the accuracy of the diagnosis are available in the literature [126, 11]. Using this information might be instrumental in obtaining more reliable estimates of biomarker accuracy. This is possible within the Bayesian framework. However, care has to be taken with respect to the way the prior information is conveyed through the prior distribution. For instance, for parameters resulting from non-linear functions, flat priors can lead to both silly as well as overly informative prior distributions [19, 106]. Nevertheless, Bayesian analysis is becoming more common in diagnostic science [15]. Bayesian models proposed for diagnostic accuracy studies are described in the following paragraph, grouped according to whether a GS reference-test is assumed available or not.

2.4.1 Proposed models

In case a GS reference-test is available, fully-parametric Bayesian inference was introduced by O'Malley et al. [78] for univariate diagnostic tests, and extended to multiple correlated tests by O'Malley et al. [77]. For the case of an imperfect reference-test, a non-parametric Bayesian method to estimate the accuracy of continuous diagnostic-tests was proposed by Ladouceur et al. [54]. Branscum et al. [14] proposed a Bayesian semi-parametric model allowing inclusion of additional information in the form of covariates or imperfect diagnostic-tests. A fully-parametric method for bivariate continuous biomarker-based diagnostic-tests was proposed by Choi et al. [18]. Bayesian latent-class mixture models for categorical diagnostic tests were developed by Joseph et al. [50] and extended by Scott et al. [105] to the case of several dichotomous and one univariate continuous diagnostic-test. A Bayesian latent-class mixture model for a single continuous test, which allows inclusion of a dichotomous imperfect reference-test, was proposed by Wang et al. [123]. Yu et al. [131] developed a Bayesian latent-class mixture model to estimate the optimal linear-combination of multiple continuous tests.

2.5 Alzheimer's disease research

Biomarkers aimed at diagnosing AD will be the guiding example throughout this dissertation. In the following paragraphs, AD CSF-biomarkers and two data sets of these biomarkers are briefly introduced.

2.5.1 Biomarkers

Often, diagnostic tests are developed based on biomarkers. A biomarker is "a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to therapeutic interventions" [22]. Biomarkers can be applied in diagnostic tests, in assessing severity or prognosis of disease, or in monitoring response to a therapeutic intervention [91].

Recently, the importance of biomarkers in diagnosing AD was acknowledged by including AD-categories based on CSF biomarker results [66]. CSF biomarkers for AD fall in two classes defined by the biological change they relate to. A first class is linked to the process of brain amyloid-beta ($A\beta$) protein deposition. The deposition of this protein is observable as a low concentration of $A\beta_{42}$ in the CSF. A second class of CSF biomarkers is involved with downstream neuronal degeneration or injury, leading to elevated CSF concentrations of total tau and phosphorylated tau (p-tau)

[20]. Changes in these CSF biomarkers reflect the ongoing pathophysiological AD mechanism underlying the clinical AD dementia. In particular, the pathophysiological AD mechanism precedes the clinical AD symptoms, making the CSF biomarkers ideal candidates to speed up the timing of AD diagnosis. This also makes them important targets for therapeutic intervention [121].

2.5.2 Data sets

To investigate the applicability of the methods developed in this dissertation, they will be applied to two sets of data from Alzheimer's disease patients. One of the AD data sets is the publically available data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The other is the Vrije Universiteit Amsterdam Medisch Centrum (VUmc) data set, which consists of patients from the memory-clinic-based Amsterdam Dementia Cohort [29]. These data sets contain observations for the three biomarkers discussed in Section 2.5.1. In addition, imperfect reference-test information is available in the form of the clinical diagnosis of AD for each patient. The clinical diagnosis of AD is imperfect because it suffers from classification errors (misdiagnosis) [11] and because the onset of the pathogenic process, as reflected in biomarker changes (see Section 2.5.1), can precede the manifestation of clinical symptoms by at least a decade [109]. Hence, a clinical non-AD diagnosis does not exclude underlying AD-pathology and the clinical diagnosis of AD does not predict underlying pathology, as was recently shown in the phase-III study with Bapineuzimab [101].

2.5.2.1 ADNI

The ADNI data set includes patients from ADNI-I. ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and non-profit organizations. ADNI-I subjects who (i) agreed to undergo a lumbar puncture, (ii) had results for all three CSF biomarkers at baseline, and (iii) belonged to either the control or AD group at baseline, were selected for the current study. This selection resulted in a data set including 96 AD and 109 control subjects. The CSF biomarker data were obtained by using the xMAP platform (Luminex Corp, Austin, Texas) and INNO-BIA AlzBio3 research use only reagents [76]. Baseline characteristics of the data are provided in Table 2.6 while histograms of the ADNI-data are shown in Figure 2.2.

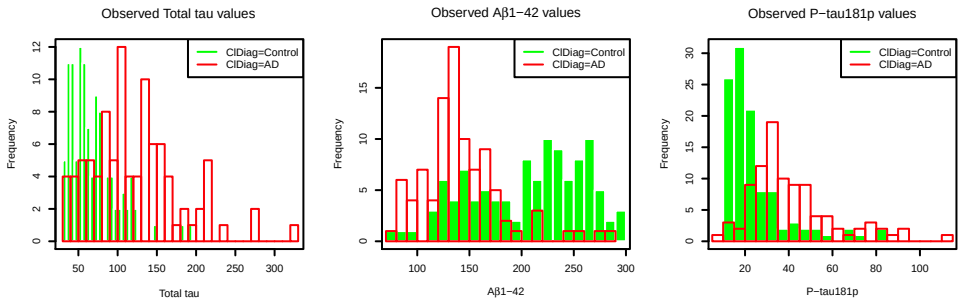

Figure 2.2: Observed total tau, Aβ_{1-42}, and p-tau$_{181p}$ values from the ADNI data set. Clinically diagnosed controls are indicated by the green histograms, clinical cases by the red histograms.

Table 2.6: Baseline characteristics of the study populations contained in the ADNI and VUmc data sets. MMSE = Mini Mental State Examination score. (mean ± SD).

Dataset	Group	n	Age(y)	Female (%)	MMSE	Aβ_{42}* (pg/mL)	total tau* (pg/mL)	p-tau$_{181}$* (pg/mL)
ADNI	Control	109	76 ± 5.3	55 (50)	29 ± 1.0	206 ± 54.4	69 ± 30.2	25 ± 14.8
	AD	96	75 ± 8.0	40 (42)	24 ± 1.9	142 ± 4.0	122 ± 57.0	42 ± 19.8
VUmc	SMC	251	64 ± 6.6	104 (41)	28 ± 1.5	874± 251.0	302 ± 197.7	52± 24.0
	AD	96	75 ± 8.0	40 (42)	24 ± 1.9	142 ± 4.0	122 ± 57.0	42 ± 19.8

*CSF-levels of Aβ_{1-42}, total tau, and p-tau$_{181p}$ were determined using commercially available single-parameter ELISA kits (INNOTEST® AMYLOID(1-42), INNOTEST® hTAU Ag, INNOTEST® PHOSPHOTAU(181P) and using the xMAP platform (Luminex Copr, Austin, Texas) and INNOBIA AlzBio3 reagens at VUmc and ADNI, respectively.

2.5.2.2 VUmc

The VUmc data set contains patients who received a diagnosis of either subjective memory complaints (SMC) or probable AD. Baseline CSF was collected between October 1999 and November 2011. All patients underwent standard dementia screening at baseline, including physical and neurological examination, electroencephalography (EEG), magnetic resonance imaging (MRI), and laboratory tests. Cognitive screening included a Mini Mental State Examination (MMSE) and a comprehensive neuropsychological test battery. Diagnoses were made by consensus in a multidisciplinary team without knowledge of the CSF results. The label of SMC was given when results of all clinical examinations were normal, and there was no psychiatric diagnosis. Patients with subjective complaints were considered as controls, but were only included when the diagnosis was confirmed at follow-up visits. This resulted in 251 SMC subjects. Probable AD (n=631) was diagnosed according to the criteria of the National Insti-

tute of Neurological and Communicative Disorders and Stroke-Alzheimer's Disease and Related Disorders association (NINCDS-ADRDA), and all patients met the core clinical National Institute on Aging - Alzheimer's Association (NIA-AA) criteria [27]. More details about this cohort have been provided elsewhere [29]. All subjects gave written informed consent for the use of their clinical data for research purposes. The study was approved by the local ethical review board. CSF levels of A$\boldsymbol{\beta}_{1-42}$, total tau, and p-tau$_{181p}$ were determined using commercially available single-parameter ELISA kits (respectively, INNOTEST® AMYLOID(1-42),INNOTEST® hTAU Ag, INNOTEST® PHOSPHOTAU(181P)) and were not used for diagnosis. The VUmc-data are shown in 2.3 with baseline characteristics of the data summarized in Table 2.6.

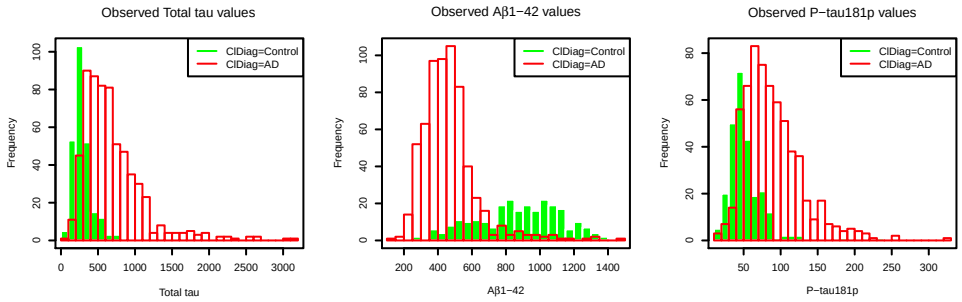

Figure 2.3: Observed total tau, Aβ_{1-42}, and p-tau$_{181p}$ values from the VUmc data set. Clinically diagnosed controls are indicated by the green histograms, clinical cases by the red histograms.

Chapter 3

Estimating continuous diagnostic-index accuracy in the presence of an imperfect reference-test

In this chapter, a Bayesian latent-class mixture model is proposed to estimate the accuracy of a diagnostic index composed of continuous biomarkers when only imperfect reference-test information is available. The problem setting is discussed in Section 3.1. The model is developed in Section 3.2. Section 3.3 discusses important issues related to the choice of the prior distributions for the proposed model, enabling control over the amount of prior accuracy information while addressing important issues concerning model non-identifiability. Model performance is investigated by considering a simulation study and by applying the model to real AD-data. The simulation study and real-data application are developed in Section 3.4 and Section 3.5, respectively. The results of both the simulation study and data application are summarized and discussed in Section 3.6. Finally, conclusions are presented in Section 3.7.

3.1 Problem setting

To overcome the problem of estimating diagnostic-biomarker accuracy when only imperfect reference-test information is available (see Section 2.3), we propose a Bayesian

latent-class mixture model. As discussed in Section 2.2, combining several continuous biomarkers into a diagnostic index may increase diagnostic accuracy as compared to the use of individual biomarkers. Therefore, we propose a Bayesian latent-class mixture model to develop a diagnostic test based on an optimal linear-combination of multiple continuous biomarkers while incorporating imperfect reference-test information in the estimation.

On the one hand, the model is an extension of the approaches developed by O'Malley et al. [77] who proposed a Bayesian mixture model for multiple continuous biomarkers when GS reference-test information is available. Moreover, the model can also be seen as an extension of the Bayesian latent-class mixture model of Yu et al. [131] to estimate diagnostic accuracy of a combination of multiple continuous biomarkers when no reference test information is available. On the other hand, we extend the model of Wang et al. [123] who proposed the inclusion of imperfect reference-test information when estimating the diagnostic accuracy of a single continuous biomarker.

When developing our model, we consider a suitable parametrisation. As mentioned in Section 2.1, the parameter of interest, AUC, is a complex function of other parameters. As a consequence, care is needed when choosing the prior distributions. We propose to consider specific prior distributions for particular functions of parameters that allow for a more controlled way of introducing prior information into the model. Moreover, this type of latent-class mixture models generally suffers from model non-identifiability. In this respect, informative priors may be required to mitigate these issues without enforcing strict parameter constraints.

Finally, we show that the proposed model could prove an important tool in AD-biomarker research, where admitting the imperfect nature of the clinical diagnosis could be essential in obtaining reliable estimates of the accuracy of diagnostic biomarkers.

3.2 Methodology

Let Y denote a multivariate continuous-biomarker variable representing the results for K biomarkers. We assume that Y follows a K-variate normal distribution with mean and variance-covariance depending on true disease status D:

$$Y|D = d \sim N_K\left(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d\right) \quad \text{for} \quad d \in \{0, 1\}, \tag{3.1}$$

where $d = 0$ indicates a true control and $d = 1$ denotes a true case.

Sensitivity (Se_T) and Specificity (Sp_T) of the reference test T can then be defined as the probability of observing $T = 1$ for a true case $[P(T = 1|D = 1)]$ and observing $T = 0$ for a true control $[P(T = 0|D = 0)]$, respectively. Specifically, we assume that variable T follows a Bernouilli distribution conditional on true disease status D, defined as follows:

$$\begin{aligned} T|D = d \sim Bern(\pi_d) \quad \text{for} \quad d \in \{0, 1\}, \\ \pi_0 = P(T = 1|D = 0) \equiv 1 - Sp_T, \\ \pi_1 = P(T = 1|D = 1) \equiv Se_T. \end{aligned} \tag{3.2}$$

As shown in Equation (3.2), the conditional distribution of T can be defined in terms of its sensitivity and specificity, Se_T and Sp_T, respectively.

As mentioned in Section 2.2, we consider the AUC as the measure of diagnostic accuracy and seek a linear combination of the K biomarkers that maximizes AUC_a. The method we will consider is the fully-parametric approach proposed by Su et al. [112], leading to linear combination coefficients as in Equation (2.4). Applying these coefficients to the biomarker data leads to a diagnostic index with diagnostic accuracy denoted by AUC_a in Equation (2.5).

If reference test T is a GS reference-test, both Se_T and Sp_T are equal to 1. In such a case, estimation of AUC_a is straightforward, since the maximum likelihood estimates of $\boldsymbol{\mu}_d$ and $\boldsymbol{\Sigma}_d$ could be simply plugged into Equation (2.5). If reference test T can not be assumed a GS reference-test, estimation of $\boldsymbol{\mu}_d$ and $\boldsymbol{\Sigma}_d$ becomes difficult, because D is unobserved. We consider it a latent variable, from now on denoted by $\tilde{D}$. We propose a latent-class mixture model that uses all observed information contained in Y and T to come up with estimates of $\boldsymbol{\mu}_d$ and $\boldsymbol{\Sigma}_d$.

Finally, we assume $\tilde{D}$ to be Bernouilli-distributed with probability of succes equal to $\theta = P(\tilde{D} = 1)$, which can be interpreted as the prevalence of disease.

3.2.1 Full-data likelihood

Assume that we have observed biomarker values and results of the reference test for a sample of N individuals (indexed by i). For the remainder of this dissertation assume $\boldsymbol{Y}$ to be the $N \times K$ matrix containing the values of K continuous biomarkers from the N subjects. Denote by $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iK})^T$ the K continuous biomarker measurements for subject i. The results of the reference test T for the N individuals are contained in column vector $\boldsymbol{t} = (t_1, \ldots, t_N)^T$. The latent true disease-status indicators are collected in a column vector $\tilde{\boldsymbol{d}} = \left(\tilde{d}_i, \ldots, \tilde{d}_N\right)^T$.

The full-data likelihood can be factorized as

$$P(\boldsymbol{Y}, \boldsymbol{t}, \tilde{\boldsymbol{d}}) = P(\boldsymbol{t}|\boldsymbol{Y}, \tilde{\boldsymbol{d}}) \times P(\boldsymbol{Y}|\tilde{\boldsymbol{d}}) \times P(\tilde{\boldsymbol{d}}). \tag{3.3}$$

Moreover, under the assumption of *conditional independence* of the biomarkers and imperfect reference-test values, conditional on true latent disease status, equation (3.3) simplifies to

$$P(\boldsymbol{Y}, \boldsymbol{t}, \tilde{\boldsymbol{d}}) = P(\boldsymbol{t}|\tilde{\boldsymbol{d}}) \times P(\boldsymbol{Y}|\tilde{\boldsymbol{d}}) \times P(\tilde{\boldsymbol{d}}). \tag{3.4}$$

Under the assumed biomarker data distribution expressed in Equations (3.1) and (3.2), and the assumption of the Bernouilli distributed latent disease status variable $\tilde{D}$, the full-data likelihood takes the following form:

$$\begin{aligned} &L\left(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1, \theta, Se_T, Sp_T | \boldsymbol{Y}, \boldsymbol{t}, \tilde{\boldsymbol{d}}\right) = \\ &\prod_{i=1}^{N} \left[Se_T^{t_i} \left(1 - Se_T\right)^{1-t_i} \frac{1}{\sqrt{(2\pi)^K \left|\boldsymbol{\Sigma_1}\right|}} exp\left\{ -\frac{1}{2} \left(\boldsymbol{y}_i - \boldsymbol{\mu}_1\right)^T \boldsymbol{\Sigma}_1^{-1} \left(\boldsymbol{y}_i - \boldsymbol{\mu}_1\right) \right\} \theta \right]^{\tilde{d}_i} \\ &\times \left[Sp_T^{1-t_i} \left(1 - Sp_T\right)^{t_i} \frac{1}{\sqrt{(2\pi)^K \left|\boldsymbol{\Sigma_0}\right|}} exp\left\{ -\frac{1}{2} \left(\boldsymbol{y}_i - \boldsymbol{\mu}_0\right)^T \boldsymbol{\Sigma}_0^{-1} \left(\boldsymbol{y}_i - \boldsymbol{\mu}_0\right) \right\} (1 - \theta) \right]^{1-\tilde{d}_i}. \end{aligned} \tag{3.5}$$

In Equation (3.5), $\tilde{d}_i$ and t_i denote, respectively, the true and reference test disease status of individual i, with $\tilde{d}_i = t_i = 1$ for cases and 0 for controls. Moreover, θ denotes the prevalence of disease, and Se_T and Sp_T denote, respectively, the sensitivity and specificity of imperfect reference-test T. The parameters of interest are $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_0$, and $\boldsymbol{\Sigma}_1$, as they define the AUC_a, in accordance with Equation (2.5).

The direct use of the full-data likelihood for estimation of the parameters of interest is not feasible as the indicators of the true disease status $\tilde{d}_i$ are not observed. Moreover, without any information about Se_T, Sp_T, or θ, these three parameters are not identifiable. Hence, to estimate the model, some information about Se_T, Sp_T, and/or θ has to be provided.

One could derive the observed-data likelihood by simply marginalizing (3.5) over $\tilde{D}$. By defining identifying restrictions for Se_T, Sp_T, and/or θ it would then be possible to estimate the parameters of the model with the help of, e.g., the EM-algorithm [24]. An alternative is to apply a Bayesian approach using the full-data likelihood directly. This is the approach that we will consider.

Non-identifiability is an important issue for mixture models [67], as well as for

the estimation of accuracy of imperfect diagnostic-tests [26]. In Bayesian statistics, in order to fit a model, model identifiability is not strictly required as long as proper priors are used. To ensure sensible posterior inference, however, non-identifiability can be mitigated by including non-diffuse prior information [48, 39]. Where and how much information should be included, will depend on the particular problem and data at hand [14, 52]. In our case, available scientific prior knowledge on Se_T, Sp_T, and/or θ could be included to overcome non-identifiability, as proposed by Joseph et al. [50] for a binary diagnostic-test. For model research purposes, we propose to use as diffuse priors as possible, but keeping a clinical research setting in mind. Moreover, disease prevalence is not assumed to be extreme and the imperfect reference-test is assumed to be the best diagnostic-tool available. In particular, Se_T and Sp_T are expected to be larger than 0.5.

3.3 Prior distributions

Table 3.1 summarizes all proposed prior distributions for the parameters included in the model. A discussion of the choice of these distributions follows below.

Prevalence

For θ, a Bayes-Laplace $Beta\,(1,1)$ prior distribution, truncated between $1/N$ and $(1-(1/N))$, may be assumed [10]. The truncation is introduced to avoid problems with the Bayesian-fitting of the model defined in (3.5) due to values of θ at the boundary of the parameter space. This may result in convergence issues when MCMC algorithms get stuck in a one-component solution instead of a mixture, a possible indication of non-identifiability [96]. Moreover, the truncation can be interpreted as ensuring that at least one true control and one true case is included in the data. Other truncation limits could also be chosen. In a case-control setting, for example, a $0.1 \leq \theta \leq 0.9$ truncation could be considered. Although more restrictive, this truncation could make sense in a setting where the proportion of cases is known up to a misclassification error and is not extreme.

Table 3.1: Structure of the considered prior distributions (assuming $K = 3$ biomarkers).

Parameter	**Prior distribution**
Prevalence	
θ	$U\left(\frac{1}{N}, \left(1 - \frac{1}{N}\right)\right)$
Parameters of the dichotomous reference test	
Se_T	$Beta\,(a, b)\,trunc(0.5, 1)$
Sp_T	$Beta\,(c, d)\,trunc(0.5, 1)$
Mean of biomarker values control group	
$\boldsymbol{\mu}_0$	$N_3\left(\mathbf{0}, \mathbf{I}_3 10^6\right)$
Naïve AUC_a prior	
Mean of biomarker values case group	
$\boldsymbol{\mu}_1$	$N_3\left(\mathbf{0}, \mathbf{I}_3 10^6\right)$
Biomarker variance-covariance matrices	
$\boldsymbol{\Sigma}_k$	$Wishart(K, I_K)$
'Controlled' AUC_a prior	
Scaled difference biomarker distribution means	
$\boldsymbol{\delta}$	$N_3\,(\boldsymbol{\kappa}, \boldsymbol{\Psi})$
Biomarker-distribution standard deviations	
$\sigma_{d,k}$	$U\,(0, 1000)$
Cholesky-factor values of correlation matrix $\boldsymbol{R}_d$	
$l_{d,21}$	$U\,(-1, 1)$
$l_{d,31}$	$U\,(-1, 1)$
$l_{d,32}$	$U\left(-\sqrt{1 - l_{d,31}^2}, \sqrt{1 - l_{d,31}^2}\right)$

Se_T and Sp_T

For Se_T and Sp_T, a truncated $Beta(a, b)$ distribution can be used. A possible truncation could be to restrict $Se_T + Sp_T > 1$, expressing a larger true- than false-positive rate [48]. The implementation of this restriction is not trivial. It requires the choice of a joint distribution of Se_T and Sp_T with consequences for the marginal prior distributions, because of the potential dependence between Se_T and Sp_T induced by the joint distribution.

One might attempt to implement the restriction while retaining the flat standard-uniform marginal distributions for Se_T and Sp_T. This is, however, not possible. A formal argumentation could be built considering the copula representation of the joint distribution function of two standard-uniform marginal distribution functions. From the representation it follows that the joint distribution function has to be bounded by

the Fréchet lower-bound (C_L) and upper-bound (C_U) copulas. These are expressed as

$$C_L(u_1, u_2) = max\{0, u_1 + u_2\}, \quad (u_1, u_2) \in [0,1]^2$$
$$\text{and}$$
$$C_U(u_1, u_2) = min\{u_1 + u_2\}, \quad (u_1, u_2) \in [0,1]^2.$$

It can be shown that for C_L we get $P(u_1 + u_2 = 1) = 1$, while for C_U we have $P(u_1 + u_2 > 1) = 0.5$. In other words, no joint distribution with standard-uniform marginals will lead to $P(u_1 + u_2 > 1) = 1$ for the desired restriction.

For the independent Se_T and Sp_T case, one could think of two uniform marginal distributions of Se_T and Sp_T restricted to be strictly higher than 0.5 to ensure that their sum exceeds 1. Although restrictive, this truncation could be thought of as a reasonable choice for the case of an imperfect reference-test for which both Se_T and Sp_T can be expected to be large.

To allow Se_T and Sp_T assuming values smaller than 0.5, a possible solution is to use a standard-uniform marginal distribution for Se_T or Sp_T and define a suitable conditional distribution of the other parameter. For instance,

$$\begin{aligned} Se_T &\sim U(0,1), \\ Sp_{T|Se_T} &\sim U(1.001 - Se_T, 1). \end{aligned} \tag{3.6}$$

Note that this solution is 'asymmetric' in that one of the parameters is selected to be uniformly distributed on the $(0,1)$ interval. Obviously, other implementations of the $Se_T + Sp_T > 1$ restriction are also possible. They will differ in terms of informativeness of the marginal Se_T and Sp_T distributions and the amount of dependence between Se_T and Sp_T. The choice of the implementation may require a careful consideration of the characteristics of the particular problem at hand. Both the 'independent but restrictive' and 'the liberal but asymmetric' truncations solve the problem of label switching, often encountered in Bayesian finite-mixture modelling due to model non-identifiability [67, 36]. Both restrictions are considered and compared in the remainder of the chapter.

$\boldsymbol{\Sigma_0}$ and $\boldsymbol{\Sigma_1}$

For the variance-covariance matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$, one could specify the prior distributions using the scaled Wishart-distribution for the precision matrices $\boldsymbol{\Sigma}_0^{-1}$ and $\boldsymbol{\Sigma}_1^{-1}$. The scaled Wishart-distribution is a popular prior for precision matrices

[77, 131, 36, 60]. In particular, the scaled Wishart-distribution with scaling matrices equal to a K-dimensional identity matrix and degrees of freedom equal to K could be applied. This choice results in a prior distribution of the variance-covariance matrices that is often claimed to be 'uninformative' [77, 131]. However, whether a scaled Wishart-distribution with the number of degrees of freedom equal to the rank of the scaling matrix can be regarded as uninformative is debatable [125, 37]. Figure 3.1 shows the results of a simple simulation exercise for the case $K = 3$. In the exercise, 100,000 draws from a scaled Wishart-distribution with three degrees of freedom and the 3×3 identity scaling-matrix were obtained. Panel *a* of Figure 3.1 presents the histogram of the three simulated variances (note that only the lower 90% of the simulated variances is shown; extreme values would make the histograms unreadable if included). Panel *b* shows the histograms of the three simulated correlation coefficients. From panel *a* it can be seen that most of the probability mass for the distribution of variances is located below 10, which can hardly be considered an uninformative. The simulated correlation coefficients show U-shaped histograms which favour extreme correlations of -1 and 1.

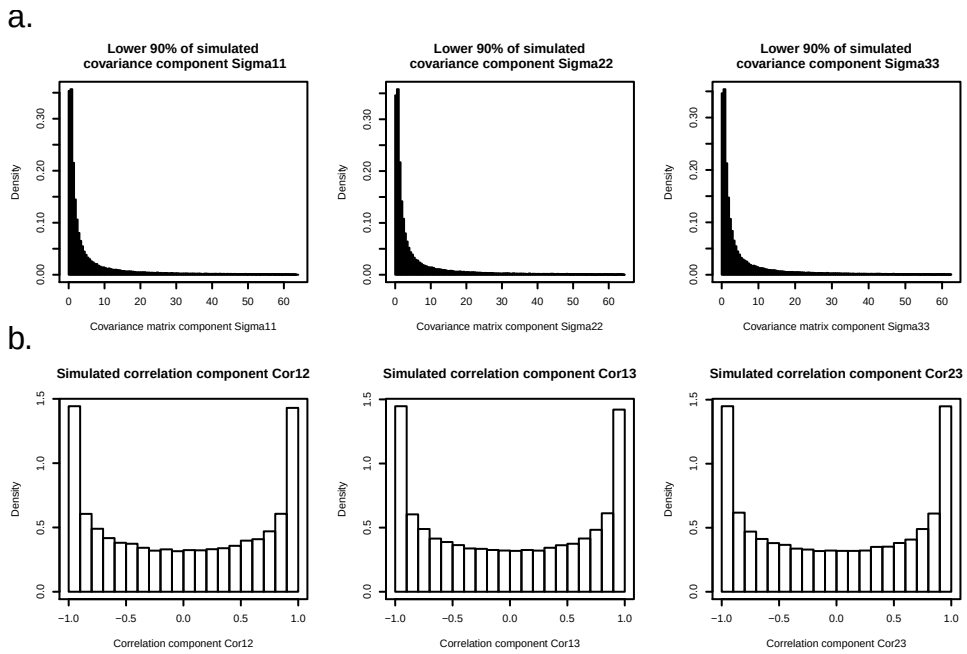

Figure 3.1: Results from 100,000 draws from a scaled $Wishart\,(3, S = \mathbf{I}_3)$ distribution. **a**. Histograms of lower 90% of simulated variances. **b**. Histograms of simulated correlation coefficients.

For this reason we consider an alternative specification of the prior distributions for the variance-covariance matrices, proposed by Wei et al. [125]. The specification is based on the following decomposition of the variance-covariance matrix, also known as the 'separation strategy', proposed by Barnard et al. [8]:

$$\boldsymbol{\Sigma} = \boldsymbol{SRS},$$

where $\boldsymbol{S}$ is a diagonal matrix of standard deviations and $\boldsymbol{R}$ is the correlation matrix. For example, for a 3×3 variance-covariance matrix,

$$\boldsymbol{S} = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{pmatrix}, \quad \boldsymbol{R} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{pmatrix}.$$

In the next step, the correlation matrix $\boldsymbol{R}$ is represented by a Cholesky decomposition:

$$\boldsymbol{R} = \boldsymbol{L}\boldsymbol{L}^T,$$

where $\boldsymbol{L}$ is a lower-triangular matrix. For example, for a 3×3 correlation matrix:

$$\boldsymbol{L} = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix}.$$

The correlation coefficients contained in $\boldsymbol{R}$ can be written in terms of the elements of $\boldsymbol{L}$ as follows:

$$\begin{aligned} \rho_{12} &= l_{21}, \\ \rho_{13} &= l_{31}, \\ \rho_{23} &= \rho_{12}\rho_{13} + l_{22}l_{33}. \end{aligned}$$

Using the separation strategy, it is possible to construct less 'informative' prior distributions for variances and correlation coefficients than those obtained for the 'uninformative' scaled Wishart-distribution prior (see Figure 3.1). In particular, one can use flat-prior distributions directly on the diagonal elements (standard deviations) of matrix $\boldsymbol{S}$ and, additionally, on $(K-1)$ of the $\left(K^2 - K\right)/2$ non-zero off-diagonal elements of the Cholesky-decomposition matrix $\boldsymbol{L}$. For example of a 3×3 variance-

covariance matrix, the prior distributions may be specified as follows:

$$\begin{aligned}
&\sigma_k \sim U\left(0, 1000\right), \\
&l_{11} = 1, \\
&l_{21} \sim U\left(-1, 1\right), \\
&l_{31} \sim U\left(-1, 1\right), \\
&l_{32} \sim U\left(-\sqrt{1 - l_{31}^2}, \sqrt{1 - l_{31}^2}\right), \\
&l_{22} = \sqrt{1 - l_{21}^2}, \\
&l_{33} = \sqrt{1 - l_{31}^2 - l_{32}^2},
\end{aligned}$$

where $U\left(a, b\right)$ denotes the uniform distribution over the interval (a, b). Figure 3.2 shows results obtained from 100,000 draws from the distributions specified above. The histograms of the three variances, presented in panel *a* (note that only the lower 90% of simulated variances is included), show that the bulk of the probability mass ranges now from 0 to 8^5. The histograms for the correlation coefficients, presented in panel *b*, show that for the first two correlations the probability mass is equally spread between -1 and 1, in contrast to what can be observed in panel *b* of Figure 3.1.

AUC_a

Another aspect of the prior distribution specification is how to control the amount of prior information assumed for the parameter of interest AUC_a. As indicated in (2.5), the AUC_a is a function of the means $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ and variance-covariance matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$. It follows that prior distributions for those parameters together imply a prior distribution for AUC_a. Given the complexity of the function it is not straightforward to deduce the prior distribution for AUC_a.

To this aim a simulation exercise may be used. First, we consider the set of prior distributions proposed by O'Malley et al. and Yu et al. [77, 131]. Figure 3.3 presents the histogram of 100,000 values of AUC_a simulated by using the flat normal priors for $\boldsymbol{\mu}_0$ as well as $\boldsymbol{\mu}_1$ and the 'uninformative' scaled Wishart-prior distributions for $\boldsymbol{\Sigma}_0^{-1}$ and $\boldsymbol{\Sigma}_1^{-1}$ (see Table 3.1). The resulting prior distribution for AUC_a is a point-mass distribution centered at 1. Upon reflection, this is not surprising. If the flat priors for $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ are assumed then, *a priori*, no overlap between the normal distributions for the biomarkers for cases and controls can be expected, irrespectively of the biomarker

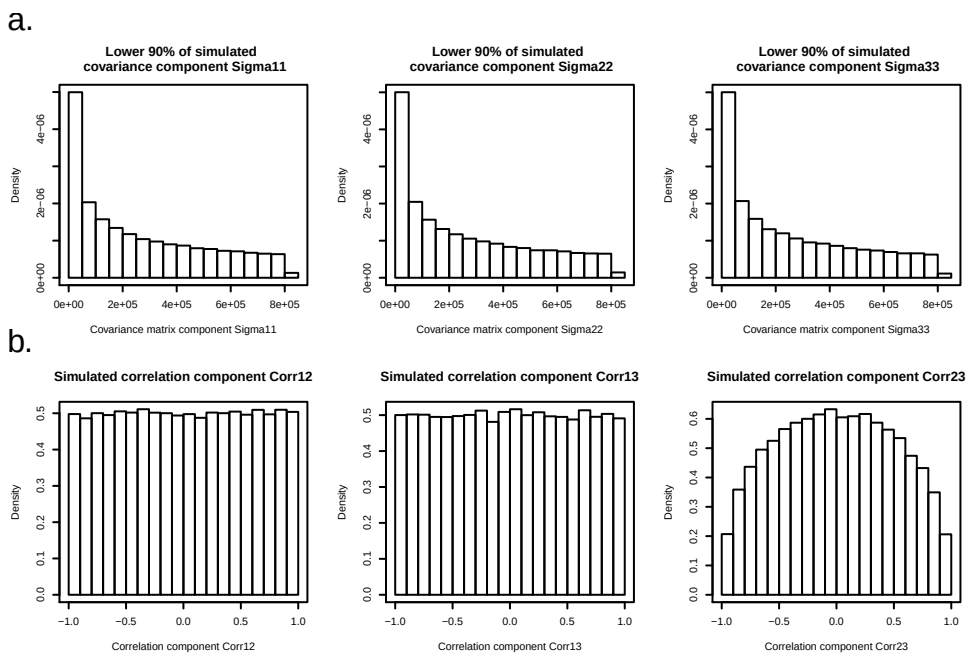

Figure 3.2: Results of 100,000 draws from the 'controlled' Wei et al. [125] variance-covariance prior distribution. **a**. Histograms of the lower 90% of simulated variance. **b**. Histograms of simulated correlation coefficients.

variances σ_k^2. However, this implies that, with a high probability, $AUC_a = 1$.

Given that AUC_a is the main parameter of interest, the priors for $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ have to be specified in such a way that the resulting prior distribution for AUC_a is controlled. To this aim, a different parametrisation of the model is proposed. By considering the Cholesky decomposition $\boldsymbol{Q}^T\boldsymbol{Q}$ of $(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1}$, AUC_a can be expressed as

$$AUC_a = \Phi\left\{\sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{Q}^T\boldsymbol{Q}\,(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}\right\}, \tag{3.7}$$

with $\boldsymbol{Q}$ defined as an upper-triangular matrix. It follows that, upon defining the scaled difference $\boldsymbol{\delta} = \boldsymbol{Q}\,(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$,

$$AUC_a = \Phi\left(\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}\right). \tag{3.8}$$

The new parametrisation consists of $\boldsymbol{\delta}$, $\boldsymbol{\mu}_0$, and $\boldsymbol{Q}$. Note that, since $\boldsymbol{\mu}_1 = \boldsymbol{Q}^{-1}\boldsymbol{\delta} + \boldsymbol{\mu}_0$, a complex prior is implied for $\boldsymbol{\mu}_1$, which should be verified. This prior distribution

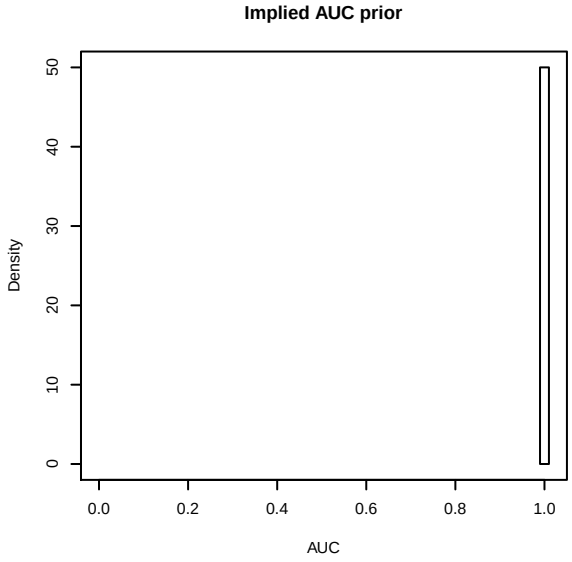

Figure 3.3: Simulation-based prior distribution for AUC_a implied by assuming flat normal prior distributions for $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ and the 'uninformative' scaled-Wishart priors for $\boldsymbol{\Sigma}_0^{-1}$ and $\boldsymbol{\Sigma}_1^{-1}$.

results from the flat prior distribution for $\boldsymbol{\mu}_0$, the distribution for $\boldsymbol{Q}$ implied by the Wei et al. [125] priors for $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$, and the prior distribution for $\boldsymbol{\delta}$.

Now, assume a K-variate normal prior distribution for $\boldsymbol{\delta}$ with mean $\boldsymbol{\kappa}$ and variance-covariance matrix $\boldsymbol{\Psi}$. Under this assumption, an approximate distribution for AUC_a can be derived based on the distribution of quadratic forms [99, 62]. In particular, the distribution can be represented by an expansion in non-central chi-square distributions:

$$f(x) = \sum_{k=0}^{\infty} c_k \chi^2_{(p+2k;\zeta)} \Phi^{-1}(x)^2 \times \left| \frac{2\Phi^{-1}(x)}{\phi\left(\Phi^{-1}(x)\right)} \right| \quad \text{for } x \in [0.5, 1], \tag{3.9}$$

with

$$\begin{aligned}
\zeta &= \sqrt{\sum_{j=0}^{p} b_j^2}, \\
c_0 &= \prod_{j=1}^{p} \sqrt{\frac{1}{\lambda_j}}, \\
c_k &= \frac{1}{2k} \sum_{r=0}^{k-1} d_{k-r} c_r, \\
d_1 &= \sum_{j=1}^{p} \left(1 - b_j^2\right) \left(1 - \frac{1}{\lambda_j}\right), \\
d_k &= \sum_{j=1}^{p} \left(1 - \frac{1}{\lambda_j}\right)^k + k \sum_{j=1}^{p} \left(\frac{b_j^2}{\lambda_j}\right) \left(1 - \frac{1}{\lambda_j}\right)^{k-1},
\end{aligned}$$

where $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $\boldsymbol{\Psi}$ and the vector $\boldsymbol{b} = (b_1, \ldots, b_p)^T = \left(\boldsymbol{P}^T \boldsymbol{\Psi}^{-\frac{1}{2}} K\right)^T$, which is a by-product of diagonalizing $\boldsymbol{\Psi}$ by $\boldsymbol{P}$, the $p \times p$ orthogonal matrix of eigenvectors of $\boldsymbol{\Psi}$. Moreover, $\phi(\cdot)$ denotes the standard-normal density function and $\Phi^{-1}(\cdot)$ the probit function. By using a finite number of terms in the series expansion, it is possible to approximate the distribution of AUC_a for different choices of $\boldsymbol{\kappa}$ and $\boldsymbol{\Psi}$ with arbitrary precision.

For instance, assume that $\boldsymbol{\kappa} = (0, 0, 0)^T$ and that standard deviations and correlation coefficients resulting from the variance-covariance matrix $\boldsymbol{\Psi}$ vary between 0.1 and 1, and 0 and 0.9, respectively. Figure 3.4 presents the histograms of 100,000 simulated values of AUC_a for the 90 resulting combinations of standard deviations and correlation coefficients. Additionally, the figure presents approximations computed from the approximate expansion of 200 non-central chi-square distributions as defined in (3.9). From the figure it can be seen that the approximations correspond closely to the histograms when standard deviations are larger than 0.4 and correlation coefficients are smaller than 0.7. For more extreme values, the approximation tends to break down at the upper tail of the distribution. Despite the issues for some of the more extreme cases, the series-expansion can be used to explore the specified prior distribution for AUC_a.

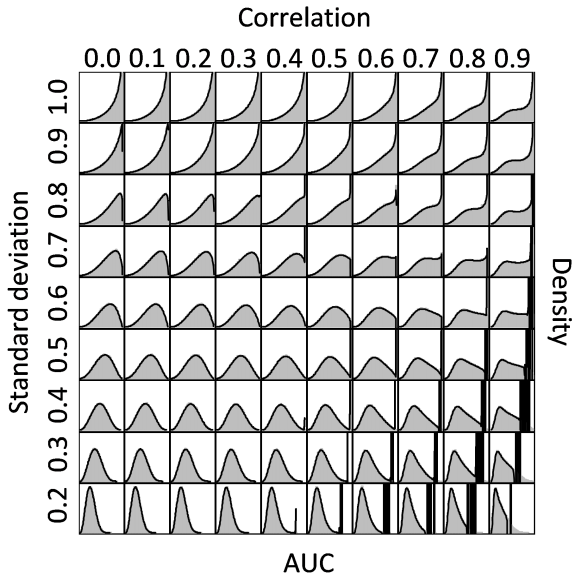

Figure 3.4: Histograms (grey) and approximated densities (solid black line) of the simulated AUC values.

As an example, Figure 3.5 presents the histogram of the simulated AUC_a values and the corresponding approximation of the density of the distribution of AUC_a obtained by setting $\boldsymbol{\kappa} = (0, 0, 0)^T$ and by assuming that standard deviations and correlation coefficients resulting from the variance-covariance matrix $\boldsymbol{\Psi}$ are equal to 0.7 and 0.6, respectively. Clearly, the distribution is much less informative than the one implied by assuming flat normal prior distributions for $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ and the 'uninformative' scaled-Wishart priors for the $\boldsymbol{\Sigma}_0^{-1}$ and $\boldsymbol{\Sigma}_1^{-1}$ (see Figure 3.3).

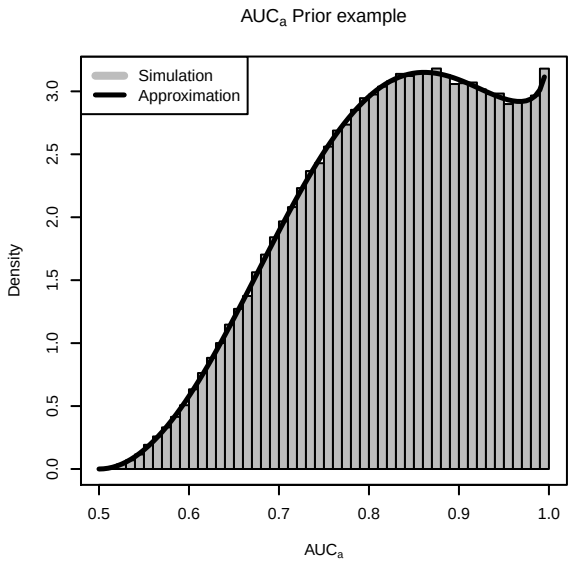

Figure 3.5: Simulated (histogram) and approximated (solid line) 'controlled' prior distribution example for AUC_a.

$\boldsymbol{\mu_0}$ and $\boldsymbol{\mu_1}$

For the prior distribution for the components of $\boldsymbol{\mu}_0$, we will follow the developments of O'Malley et al. [77] and Yu et al. [131]. In these references, normal distributions with mean 0 and variance 10^6 are proposed. These distributions are essentially flat priors.

To evaluate the implied prior distribution for $\boldsymbol{\mu}_1$, 100,000 simulations were drawn for $\boldsymbol{\mu}_1 = \boldsymbol{Q}^{-1}\boldsymbol{\delta} + \boldsymbol{\mu}_0$ by assuming the proposed priors for $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$, $\boldsymbol{\Sigma}_1$, and $\boldsymbol{\delta}$. The histograms for the three components of $\boldsymbol{\mu}_1$ are shown in Figure 3.6, together with the assumed normal prior distribution for $\boldsymbol{\mu}_0$. The implied prior distribution for $\boldsymbol{\mu}_1$ is at least as flat (note that the x-axis ranges from -15,000 to 15,000) as the one for $\boldsymbol{\mu}_0$.

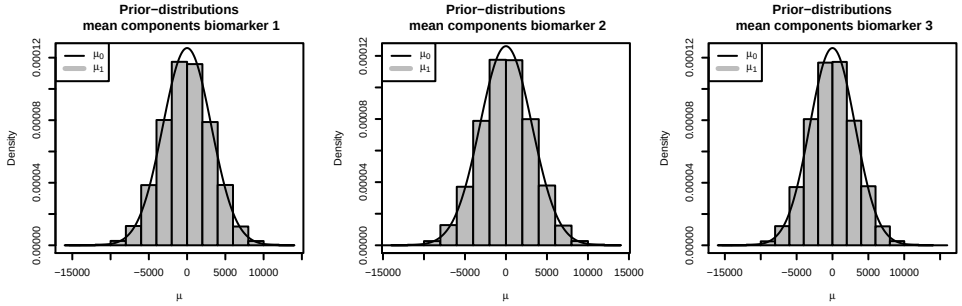

Figure 3.6: Simulated prior distribution for $\boldsymbol{\mu}_1$ (histogram) based on proposed prior distributions for $\boldsymbol{\delta}$, $\boldsymbol{\Sigma}_0$, $\boldsymbol{\Sigma}_1$, and $\boldsymbol{\mu}_0$ (solid line).

3.4 Simulation study

To evaluate the proposed method, the model was fitted to simulated data. Different data-generating models, sample sizes, as well as prior distributions were implemented to investigate model robustness and applicability.

3.4.1 Data

Data for three biomarkers were simulated. Underlying true parameter values are presented in Table 3.2. The true control-group mean-vector $\boldsymbol{\mu}_0$ was set equal to $(0, 0, 0)^T$. The mean vector of the cases $\boldsymbol{\mu}_1$ was derived based on (2.5), applied separately for each biomarker, by fixing the true value of AUC for each individual biomarker at 0.75.

Four different variance-covariance structures for the biomarkers were considered. In the homoscedastic simulation-setting, the biomarker variance-covariance matrices for the controls and cases were assumed equal. The variances of all biomarkers were assumed to be equal to 1. Additionally, the biomarkers were assumed independent or dependent (correlated). In the heteroscedastic setting, different variance-covariance matrices for the controls and cases were assumed, and again independent or dependent biomarkers, were considered (see Table 3.2).

For the imperfect reference-test, the values of Se_T and Sp_T were fixed at 0.85. Finally, the prevalence of disease, θ, was assumed to be equal to 0.5.

Table 3.2: Parameter values underlying the sampled simulation data sets.

Parameter	**Value**
Multivariate parameters	
$\boldsymbol{\mu}_0$	$(0,0,0)^T$
$\boldsymbol{\mu}_1$	$(1.1683, 1.3490, 1.5082)^T$
Homoscedastic case ($\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$)	
Independent biomarkers $\left(\boldsymbol{\rho} = (0,0,0)^T\right)$	
$\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
Dependent biomarkers $\left(\boldsymbol{\rho} = (0.5, 0.9, 0.5)^T\right)$	
$\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$	$\begin{pmatrix} 1 & 0.5 & 0.9 \\ 0.5 & 1 & 0.5 \\ 0.9 & 0.5 & 1 \end{pmatrix}$
Heteroscedastic case ($\boldsymbol{\Sigma}_0 \neq \boldsymbol{\Sigma}_1$)	
Independent biomarkers $\left(\boldsymbol{\rho} = (0,0,0)^T\right)$	
$\boldsymbol{\Sigma}_0$	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix}$
$\boldsymbol{\Sigma}_1$	$\begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix}$
Dependent biomarkers $\left(\boldsymbol{\rho} = (0.5, 0.9, 0.5)^T\right)$	
$\boldsymbol{\Sigma}_0$	$\begin{pmatrix} 1 & 0.87 & 1.27 \\ 0.87 & 3 & 1.22 \\ 1.27 & 1.22 & 2 \end{pmatrix}$
$\boldsymbol{\Sigma}_1$	$\begin{pmatrix} 2 & 0.71 & 2.2 \\ 0.71 & 1 & 0.87 \\ 2.2 & 0.87 & 3 \end{pmatrix}$
Parameters of the dichotomous reference test	
Sp_T	0.85
Se_T	0.85
Prevalence	
θ	0.5
Functions of multivariate parameters	
AUC_1	0.75
AUC_2	0.75
AUC_3	0.75
$\boldsymbol{a}$	$(0.1594, 0.2237, 0.0972)^T$
Optimal combination parameters	
$\mu_{a,0}$	0
$\mu_{a,1}$	0.6347
$\sigma^2_{a,0}$	0.3490
$\sigma^2_{a,1}$	0.2857
AUC_a	0.7872

For each of the four simulation scenarios, corresponding to the assumed variance-covariance structures (see Table 3.2), three different sample sizes were considered: 100, 400, and 600. This led up to 12 different simulation scenarios. For each scenario 100 data sets were generated.

In addition, to investigate robustness of the model to the violation of the underlying normality assumption, 100 skew-normal [4] data sets, consisting of 400 observations each, were simulated. The underlying characteristics of these data were matched to the normal data, with biomarker-specific AUC of around 0.77, leading to a combined AUC_a of 0.9. For the imperfect reference-test, underlying sensitivity and specificity of 0.85 were maintained. The underlying true marginal distributions are shown in Figure 3.7.

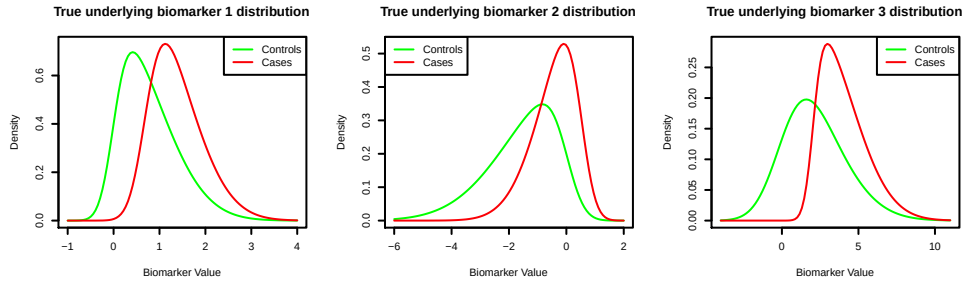

Figure 3.7: Underlying skew-normal biomarker distributions. Green solid line denotes the control distribution, red solid line denotes the case distribution.

3.4.2 Prior distributions

The prior distributions for the components of the mean-vector of the true control group $\boldsymbol{\mu}_0$ were assumed as defined in Table 3.1. For the prevalence of disease, θ, the more restrictive $U(0.1, 0.9)$ prior distribution was used to allow for more stable results, especially in the small data sets.

In order to investigate the sensitivity of the model results to prior information, several prior distributions were considered. For the sensitivity and specificity of the imperfect reference-test, two priors were assumed (see Figure 3.8). First, a flat truncated $Beta(1, 1)$ distribution was assumed as shown in panel a of Figure 3.8. Second, an informative truncated $Beta(10, 1.765)$ distribution was assumed. This distribution, shown in panel b of Figure 3.8, is centred around 0.85 with an equal-tail 95% interval of $(0.608, 0.983)$. In both cases, truncation to the $[0.51, 1)$ interval, as discussed in Section 3.3, was applied (see Figure 3.8).

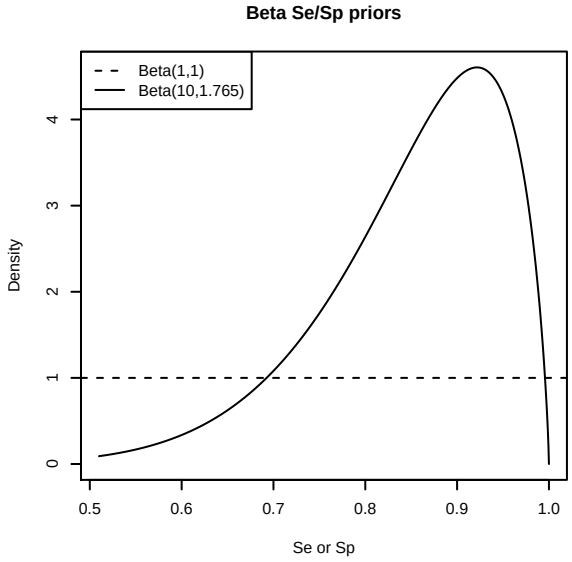

Figure 3.8: Se_T/Sp_T-prior distributions lower-truncated to $[0.51, 1)$. Dotted line shows the flat Se_T/Sp_T-prior distribution based on the $Beta(1, 1)$ distribution. Solid line denotes the informative Se_T/Sp_T-prior distribution based on the $Beta(10, 1.765)$ distribution.

For the assumed normal-distribution parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_0$, and $\boldsymbol{\Sigma}_1$, the 'naïve' as well as the proposed 'controlled' AUC-prior distributions were assumed. In the 'naïve AUC_a'-prior setting, a flat $N\left(0, 10^6\right)$ prior was assumed for $\boldsymbol{\mu}_1$ while the 'uninformative' scaled Wishart-distribution, with degrees of freedom equal to the number of biomarkers (K) and scaling matrix equal to the $K \times K$ identity matrix, was used for $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$. These prior distributions implied the point-mass prior distribution for AUC_a as shown in Figure 3.3. In case the 'controlled' AUC_a-prior distribution was assumed, two $\boldsymbol{\delta}$ prior distributions were investigated. An *optimistic* AUC_a-prior distribution was defined by considering a $\boldsymbol{\delta}$-prior distribution with $\boldsymbol{\kappa} = (0, 0, 0)^T$ and variance-covariance matrix $\boldsymbol{\Psi}$ resulting in standard deviations and correlation coefficients equal to $(0.7, 0.7, 0.7)$ and $(0.6, 0.6, 0.6)$, respectively. With a mean of 0.827 and 95% equal-tail interval of [0.608.0.992], this distribution disfavours small values of AUC_a, as indicated in panel a of Figure 3.9. Alternatively, a *conservative* AUC_a-prior distribution is assumed by considering the same $\boldsymbol{\kappa}$ as above and $\boldsymbol{\Psi}$ defined as resulting in standard deviations and correlation coefficients equal to $(0.5, 0.5, 0.5)$ and $(0.3, 0.3, 0.3)$, respectively. As shown in panel b of Figure 3.9, this distribution favours moderate values of AUC_a with a mean value of 0.772 and 95% equal-tail interval of [0.588;0.943].

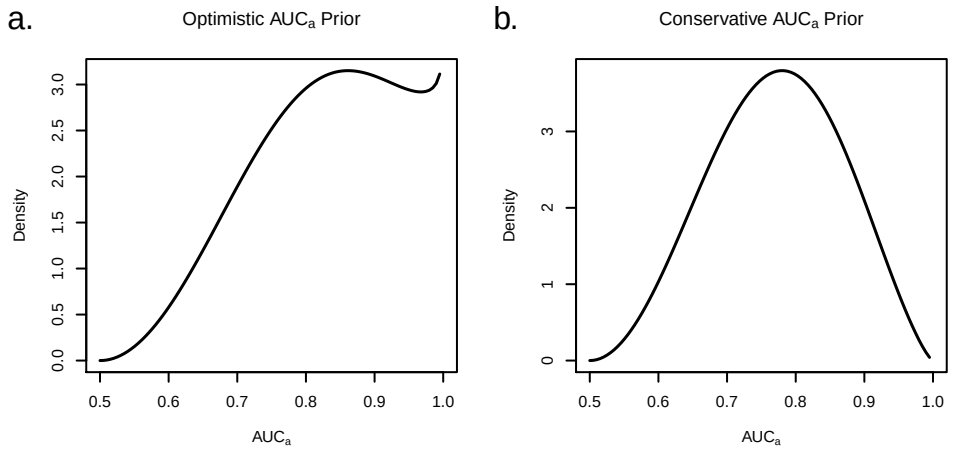

Figure 3.9: Considered AUC_a-prior distributions. **a**. 'Optimistic' AUC_a-prior distribution. **b**. 'Conservative' AUC_a-prior distribution.

3.4.3 Analysis setting

Overall, the model defined in (3.5) was fitted six times to each of the simulated data sets. In particular, it was fitted two times for each of the three AUC_a-prior settings ('naïve', 'controlled optimistic', and 'controlled conservative'): once assuming the flat truncated $Beta(1, 1)$ prior for Se_T and Sp_T, and once with the informative truncated $Beta(10, 1.765)$ distribution.

The estimates of the coefficients of the model were obtained by using 10,000 samples from the posterior distribution after a burn-in period of 10,000 samples from five independent MCMC chains. Starting values for the MCMC chains were fixed at plausible data-based values for all parameters with exception of Se_T and Sp_T which were started at the midpoint of their parameter space, i.e., 0.75. Starting values for $\boldsymbol{\mu}_0$ were based on the observed mean values for the controls. In the same line, the starting values for the standard deviations and the Cholesky-decomposition components of the correlation matrices and the case-control distribution differences scaled by the inverse of the sum of the variance-covariance matrices were computed from their observed counterparts. The starting values for the latent-disease indicator variable $\tilde{D}$ were taken to be the observed imperfect reference-test results.

After fitting, the results were first checked by general diagnostic-tools in order to assess convergence of the MCMC chains. Convergence over chains was investigated by the Gelman-Rubin convergence index, for which a cut-off value of 1.1 was applied [38]. Chain-by-chain convergence was checked by using the Geweke convergence-criterion

[40]. Fits for which the Gelman-Rubin index suggested non-convergence were excluded from the results, while the Geweke index was monitored to ensure that, on average, no more than two out of five chains were considered as non-converged for each parameter over all simulated data sets.

The models were fitted by using OpenBUGS 3.2.1 [60]. Annotated BUGS model codes can be found in Section A.1 of Appendix A. Results were analyzed and summarized by using R 3.0.1 (x64) [90]. The R-package R2OpenBUGS [111] was used as an interface between R 3.0.1 and OpenBUGS. Fitting times depended on sample size and were equal to, approximately, 2, 8, and 12 hours for sample sizes of 100, 400, and 600, respectively, on a 64-bit, 2.8GHz, 8GB RAM machine.

3.5 Real data

For the real data application, the publicly available data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) data base and the data set from the Vrije Universiteit Amsterdam Medisch Centrum (VUmc), which consists of patients from the memory-clinic-based Amsterdam Dementia Cohort, were considered (see Section 2.5.2).

Inspection of the histograms of the observed ADNI and VUmc data in panel *a* of Figures 3.10 and 3.11, leads to the observation that the histograms for total tau and p-tau$_{181p}$ show right-skewness for the cases as well as the controls. Generally this type of skewness is observed for biomarkers measured on a strictly positive scale having values close to zero. For this reason we considered to log-transform total tau and p-tau$_{181p}$ for both the ADNI and VUmc data. The log-transformation resolved the right-skewness in the histograms, as shown in panel *b* of Figures 3.10 and 3.11.

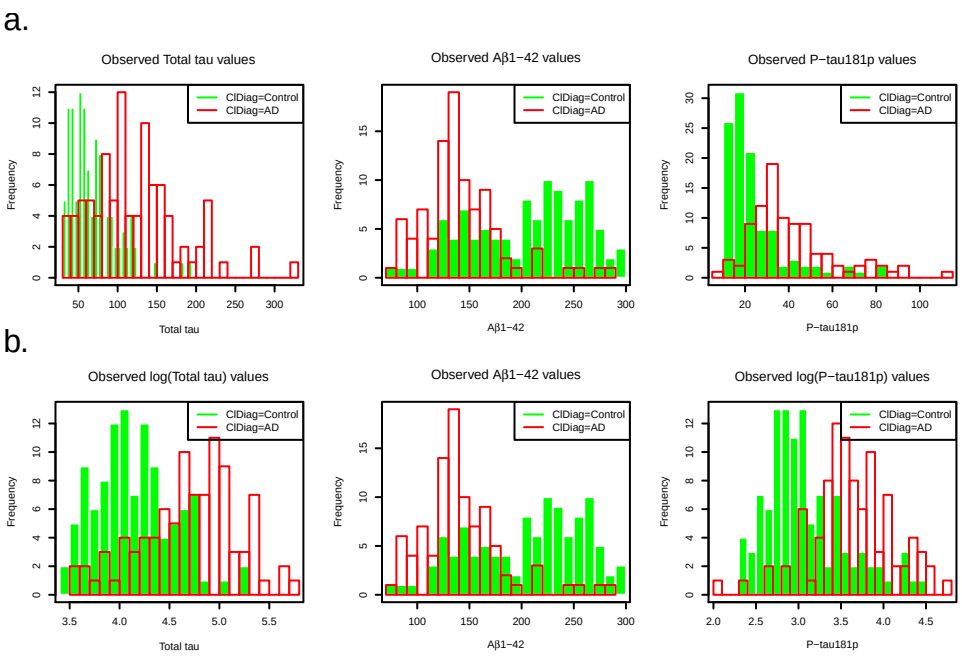

Figure 3.10: Observed distributions of the ADNI CSF-biomarker data (histograms) by clinical diagnosis (green=clinical control; red=clinical case). **a**. Raw data. **b**. Log-transformed total tau and p-tau$_{181p}$.

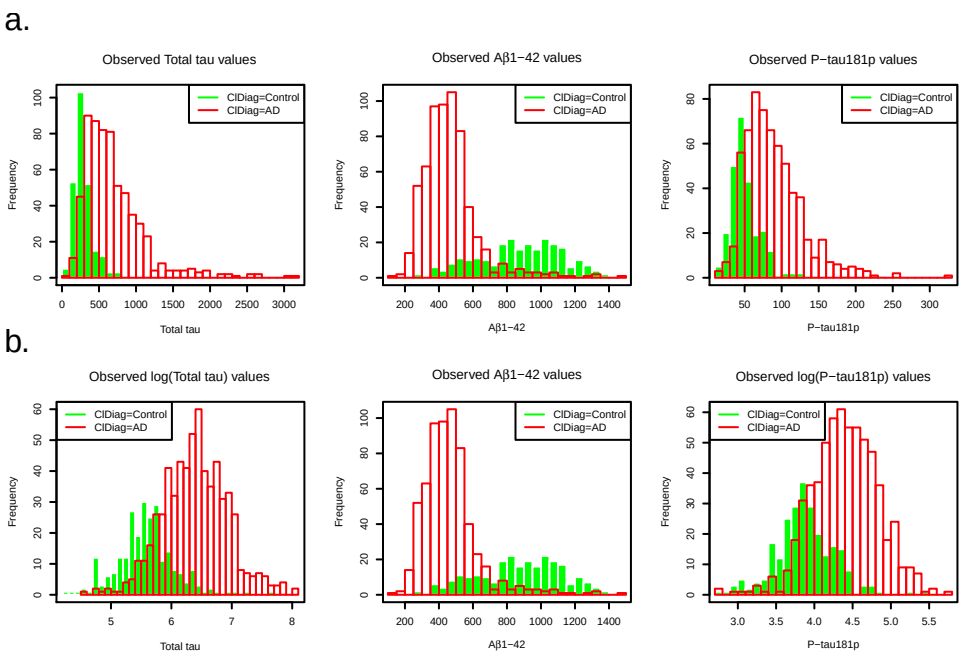

Figure 3.11: Observed distributions of the VUmc CSF-biomarker data (histograms) by clinical diagnosis (green=clinical control; red=clinical case). **a**. Raw data. **b**. Log-transformed total tau and p-tau$_{181p}$.

The real data were analysed by applying the same proposed model that was used for the analysis of the simulated data, as defined in (3.5). The prior distribution for the prevalence of disease θ was equal to the one specified in Table 3.1. For Se_T and Sp_T, we considered two types of truncation: $Se_T > 0.5$ and $Sp_T > 0.5$, and $Se_T + Sp_T > 1$. In particular, to restrict $Se_T > 0.5$ and $Sp_T > 0.5$, the truncated flat $Beta\,(1, 1)$ distribution for both Se_T and Sp_T (Figure 3.8) was assumed. On the other hand, the $Se_T + Sp_T > 1$ restriction was implemented by assuming a flat $Beta\,(1, 1)$ prior for Se_T and the restricted conditional distribution of Sp_T given Se_T, as defined in (3.6). This prior is shown in Figure 3.12. Moreover, the AUC_a-prior distribution was also varied as in the simulation study, considering the optimistic, as well as the conservative, AUC_a-prior distributions.

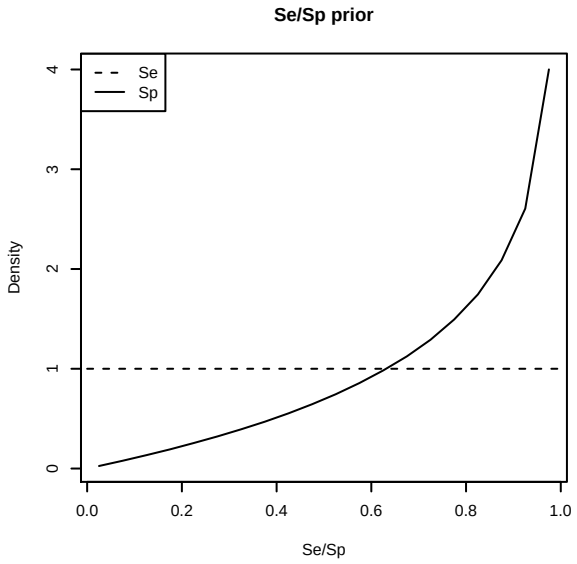

Figure 3.12: Se_T- and Sp_T-prior restricted to $Se_T + Sp_T > 1$. Dotted line shows the flat Se_T prior distribution based on the $Beta(1, 1)$ distribution. Solid line denotes the Sp_T prior distribution conditional on Se_T truncated to $[1.001 - Se_T, 1)$.

After fitting, convergence-diagnostics measures similar to those used in the analysis of the simulated data were applied.

To show the impact of ignoring the imperfectness of the reference test, the logistic regression model relating the clinical diagnosis to the three CSF-biomarkers was additionally considered [104]. The AUC was computed based on log-condense smoothing of the empirical ROC curve as described by Rufibach [100] and implemented in the R-package pROC [97]. The resulting AUC distribution was obtained by bootstrapping. By definition, this model considers the clinical diagnosis as a GS reference-test.

3.6 Results

3.6.1 Simulation study

Tables 3.3 to 3.5 present the averages of the posterior medians for AUC_a from all simulation scenarios for $N = 100$, $N = 400$, and $N = 600$, respectively. Note that, for each scenario, six sets of results are presented: three obtained for the analysis using the flat prior for sensitivity and specificity of the reference test (FLAT) for every AUC_a prior-distribution, and three obtained for the analysis using the informative

prior (INF) (see Figure 3.8).

Table 3.3: Mean of posterior AUC_a medians with corresponding (standard deviation of posterior AUC_a medians) based on [number of converged data sets] for the simulated data sets of size $N = 100$. FLAT — the analysis using the flat prior for sensitivity and specificity of the reference test; INF — the analysis using the informative prior for sensitivity and specificity of the reference test (see Figure 3.8). Results obtained with the naïve, 'conservative' (Cons.) and 'optimistic' (Opt.) AUC_a prior.

Data Gen. Model		**Se/Sp**	**True**	$\boldsymbol{AUC_a - Prior}$		
VarCov	**Corr**	**prior**	**AUC**	**Naïve**	**Cons.**	**Opt.**
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.879	0.915 (0.030) [95]	0.859 (0.038) [68]	0.877 (0.036) [79]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.879	0.906 (0.037) [97]	0.845 (0.042) [85]	0.864 (0.045) [92]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.784	0.896 (0.039) [83]	0.762 (0.044) [16]	0.823 (0.065) [16]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.784	0.874 (0.048) [94]	0.754 (0.050) [35]	0.799 (0.050) [54]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.879	0.912 (0.037) [95]	0.859 (0.042) [93]	0.876 (0.041) [95]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.879	0.902 (0.042) [99]	0.858 (0.044) [98]	0.871 (0.044) [97]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.787	0.855 (0.050) [96]	0.754 (0.040) [66]	0.789 (0.048) [84]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.787	0.836 (0.052) [99]	0.742 (0.043) [84]	0.777 (0.052) [96]

The results shown in Tables 3.3, 3.4, and 3.5 indicate non-convergence problems. The number of converged data sets, indicated in square brackets in the tables, ranged from 16 to 100. As already mentioned, non-convergence was defined by observing a Gelman-Rubin convergence index > 1.1 for any of the considered parameters. The problems were occurring for the $N = 100$ case and for the case of correlated biomarkers with the same variance-covariance matrix for cases and controls irrespective of sample size for the two 'controlled' AUC_a priors. In general, the use of the informative Se_T and Sp_T prior distributions decreased the rate of non-convergence. Moreover, for the 'naïve' AUC_a-prior, the non-convergence rate was significantly reduced while there was no clear indication for a difference in the convergence rate when selecting the 'conservative' and 'optimistic' AUC_a-priors.

It is clear from the tables that considering the 'naïve' AUC_a-prior leads to biased estimates of AUC_a. In all simulation settings, for both Se_T/Sp_T priors and all sample

Table 3.4: Mean of posterior AUC_a medians with corresponding (standard deviation of posterior AUC_a medians) based on [number of converged data sets] for the simulated data sets of size $N = 400$. FLAT — the analysis using the flat prior for sensitivity and specificity of the reference test; INF — the analysis using the informative prior for sensitivity and specificity of the reference test (see Figure 3.8). Results obtained with the 'naïve', 'conservative' (Cons.) and 'optimistic' (Opt.) AUC_a prior.

Data Gen. Model		**Se/Sp**	**True**	$\boldsymbol{AUC_a - Prior}$		
VarCov	**Corr**	**prior**	**AUC**	**Naïve**	**Cons.**	**Opt.**
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.879	0.843 (0.032) [100]	0.868 (0.026) [100]	0.877 (0.025) [100]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.879	0.885 (0.026) [100]	0.865 (0.027) [100]	0.872 (0.026) [100]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.784	0.843 (0.032) [93]	0.766 (0.031) [61]	0.798 (0.032) [87]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.784	0.825 (0.033) [98]	0.754 (0.031) [88]	0.782 (0.032) [97]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.879	0.885 (0.022) [100]	0.874 (0.022) [100]	0.879 (0.022) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.879	0.882 (0.022) [100]	0.871 (0.022) [100]	0.876 (0.022) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.787	0.806 (0.032) [100]	0.770 (0.029) [100]	0.787 (0.030) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.787	0.802 (0.031) [100]	0.769 (0.028) [100]	0.783 (0.030) [100]

sizes, the average posterior medians are overestimating the true value. For both 'controlled' AUC_a-priors, the average posterior medians of AUC_a are very close to the true values. Note that this conclusion is based only on the data sets for which convergence was observed. Thus, if the model converges, it provides a reliable estimate of AUC_a if a sensible prior distribution for AUC_a is assumed. The priors for sensitivity and specificity of the reference test and the 'controlled' AUC_a-priors seem to have negligible effect on precision of the estimates.

For all other parameters, average posterior-median estimates are also close to the true underlying values (results are shown in Appendix B). The proportion of cases when the true parameter value is contained in the 95% credible interval varies between 0.89 and 1 for all parameters over all simulation settings.

For the skew-normal data, the average posterior-median for AUC_a was equal to 0.907 (true $AUC_a = 0.9$), with the standard deviation of the posterior medians equal to 0.021. Fits for two of the 100 simulated data sets did not converge and for about

Table 3.5: Mean of posterior AUC_a medians with corresponding (standard deviation of posterior AUC_a medians) based on [number of converged data sets] for the simulated data sets of size $N = 600$. FLAT — the analysis using the flat prior for sensitivity and specificity of the reference test; INF — the analysis using the informative prior for sensitivity and specificity of the reference test (see Figure 3.8). Results obtained with the 'naïve', 'conservative' (Cons.) and 'optimistic' (Opt.) AUC_a prior.

Data Gen. Model		**Se/Sp**	**True**	$\boldsymbol{AUC_a - Prior}$		
VarCov	**Corr**	**prior**	**AUC**	**Naïve**	**Cons.**	**Opt.**
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.879	0.883 (0.022) [100]	0.866 (0.024) [99]	0.872 (0.024) [100]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.879	0.879 (0.023) [100]	0.863 (0.25) [100]	0.869 (0.024) [100]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.784	0.829 (0.029) [88]	0.770 (0.022) [66]	0.793 (0.030) [92]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.784	0.813 (0.029) [89]	0.755 (0.024) [74]	0.780 (0.028) [99]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.879	0.887 (0.019) [100]	0.880 (0.019) [100]	0.883 (0.019) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.879	0.886 (0.019) [100]	0.877 (0.019) [100]	0.881 (0.019) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.787	0.796 (0.026) [100]	0.775 (0.024) [100]	0.786 (0.025) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.787	0.794 (0.025) [100]	0.773 (0.024) [100]	0.784 (0.025) [100]

94% of the remaining fits the true underlying AUC_a was contained in the 95% credible interval.

3.6.2 ADNI data

Assuming *a priori* that both Se_T and $Sp_T > 0.5$ or that $Se_T + Sp_T > 1$ leads to essentially the same posterior estimates (Table C.1 in Appendix C). Considering a 'conservative' or 'optimistic' AUC_a-prior leads to the same posterior estimates as well (C.1 in Appendix C). Hence, in what follows, only the estimates for the Se_T/Sp_T prior restricting $Se_T + Sp_T > 1$ (see Figure 3.12) in combination with the 'optimistic' AUC_a-prior (see Figure 3.9) will be discussed. The posterior density of the AUC of the optimal linear-combination of the biomarker of interest, AUC_a, for the logistic regression model, as well as for the proposed latent-class mixture model, are shown in Figure 3.13. For the logistic regression model, the median AUC_a was estimated

to be equal to 0.883 with the 95% bootstrap-interval equal to [0.831; 0.928]. The proposed Bayesian latent-class mixture model, accounting for the imperfect nature of the clinical diagnosis (and using the prior for Se_T and Sp_T as in Figure 3.12), resulted in the median estimate of 0.984 with the 95% credible interval equal to [0.959; 0.994].

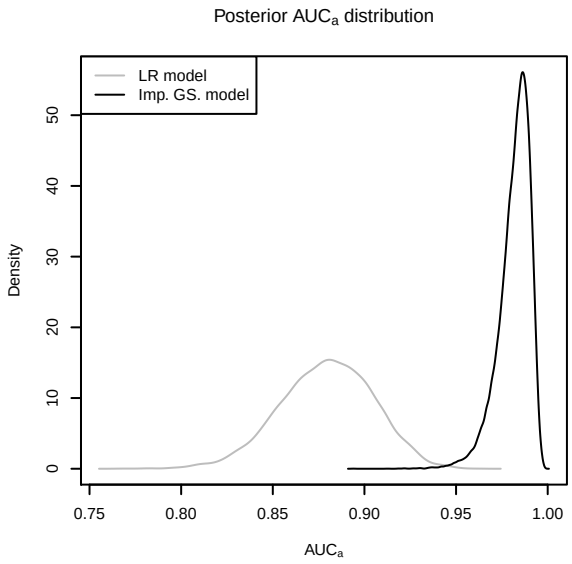

Figure 3.13: Posterior AUC_a distribution for the ADNI-data fitted with a logistic regression model (grey line) and the proposed-imperfect reference-test model (black line) with the flat Se_T and Sp_T prior distributions with $Se_T + Sp_T > 1$ restriction and the 'optimistic' AUC_a-prior distribution.

Panel a of Figure 3.14 presents the posterior distributions for the sensitivity and specificity of imperfect reference-test T. The posterior medians for Se_T and Sp_T were estimated to be equal to 0.826 and 0.888, respectively. For Se_T, the 95% credible interval was equal to [0.730; 0.905], while for Sp_T it was equal to [0.803; 0.951], confirming that the clinical diagnosis is indeed not a GS reference-test. The posterior distribution for prevalence of disease θ, corresponding to considering the uniform prior truncated between $1/N = 0.005$ and $1 - (1/N) = 0.995$, is shown in panel b of Figure 3.14. The posterior median estimate for θ was equal to 0.500 with the 95% credible interval [0.415; 0.587].

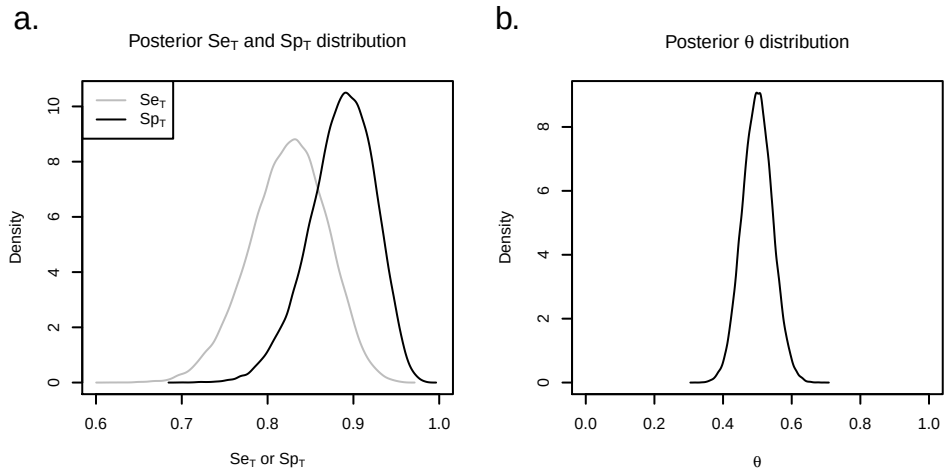

Figure 3.14: Posterior distribution for the ADNI data fitted with the proposed imperfect reference-test model, with the flat Se_T and Sp_T prior distributions with $Se_T + Sp_T > 1$ restriction and the 'optimistic' AUC_a-prior distribution. **a**. Posterior Se_T (grey line) and Sp_T (black line). **b**. Posterior prevalence of disease (θ).

Figure 3.15 shows the estimated probability of AD for the resulting score based on the optimal combination of the three biomarkers by clinical diagnosis. This probability is based on the posterior median estimates of the component parameters and optimal combination coefficients in line with Scott et al. [105]. Hereby, the probability of AD is constrained to increase monotonically with the diagnostic score, consistent with the results from a logistic regression model. The plot in panel *a* of Figure 3.15 illustrates the imperfect nature of the clinical diagnosis, i.e., potential misclassification of several individuals. In particular, nine subjects diagnosed as having AD with a diagnostic score smaller than 13.7 have less than 50% probability of being truly AD patients. On the other hand, 17 subjects diagnosed clinically as not having AD, but with a diagnostic score larger than 15.5, have more than 50% probability of being truly AD patients.

Panel *b* of Figure 3.15 shows the posterior probability of AD as a function of diagnostic score by clinical diagnosis. These probabilities are defined as the posterior means of the true disease status for each subject based on the combined information of biomarkers and clinical diagnosis. This way, the posterior probability of AD is not constrained to monotonically increase with diagnostic score and has a clear Bayesian interpretation. The plot in panel *b* of Figure 3.15 also clearly illustrates the imperfect nature of the clinical diagnosis. Results show that for ten subjects diagnosed as AD patients, the posterior probability of being truly AD is less than 50%. Of the subjects

not clinically diagnosed as AD patients, 16 have a posterior probability of AD of more than 50%.

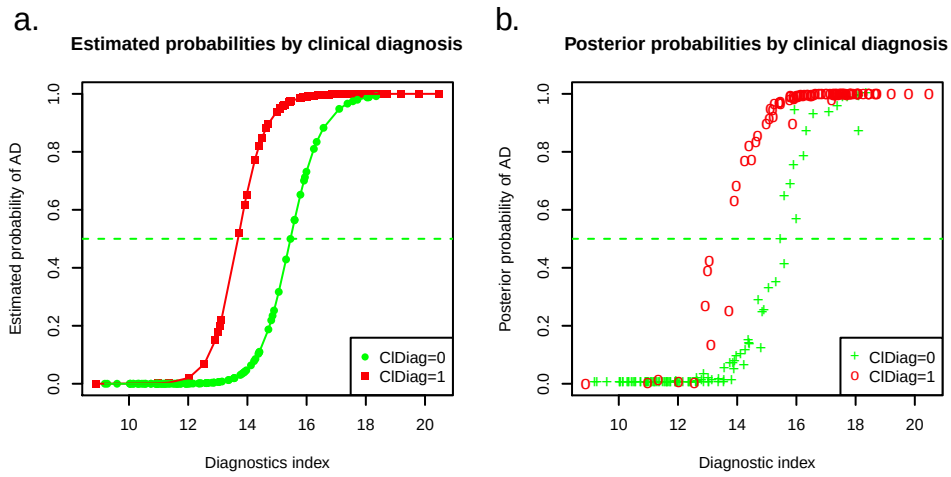

Figure 3.15: Probability of AD for estimated optimal combination score by clinical diagnosis (ClDiag) for the ADNI data set. Clinical controls indicated in green, clinical cases in red. **a**. Estimated probability of AD. **b**. Posterior probability of AD.

3.6.3 VUmc data

As it was the case for the ADNI data set results, both the Se_T and $Sp_T > 0.5$ and $Se_T + Sp_T > 1$ restrictions and considered AUC_a priors lead to essentially the same posterior estimates (Table C.2 and Figure C.2 in Appendix C). Hence, in what follows, only the estimates for the prior restricting $Se_T + Sp_T > 1$ (see Figure 3.12) will be discussed. The posterior density for AUC_a of the optimal combination of the three CSF-biomarkers is shown in Figure 3.16. For the logistic regression model, the median AUC_a was estimated to be equal to 0.88 with the 95% bootstrap interval equal to [0.828; 0.926]. The proposed Bayesian latent-class mixture model, accounting for the imperfect nature of the clinical diagnosis (and using the prior for Se_T and Sp_T as in Figure 3.12), resulted in the median estimate of 0.995 with the 95% credible interval equal to [0.991; 0.998].

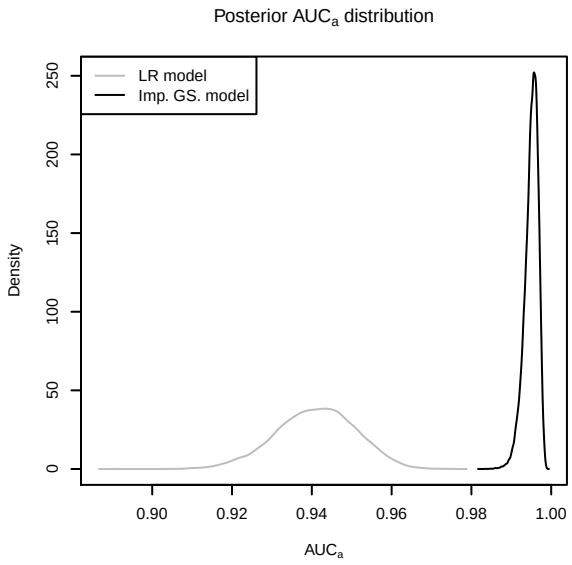

Figure 3.16: Posterior AUC_a distribution for the VUmc-data fitted with a logistic regression model (grey line) and the proposed imperfect reference-test model (black line) with the flat Se_T and Sp_T prior distributions with $Se_T + Sp_T > 1$ restriction and the 'optimistic' AUC_a-prior distribution.

The posterior distributions for sensitivity and specificity of the imperfect reference-test for the VUmc data set are presented in panel a of Figure 3.17. The posterior medians for Se_T and Sp_T were estimated to be equal to 0.957 and 0.853, respectively. For Se_T, the 95% credible interval was equal to [0.936; 0.974], while for Sp_T it was equal to [0.803; 0.896]. The posterior distribution for prevalence of disease θ, corresponding to the uniform prior truncated between $1/N = 0.001$ and $1-(1/N) = 0.999$, is shown in panel b of Figure 3.17. The posterior median estimate for θ was equal to 0.701 with the 95% credible interval [0.667; 0.733].

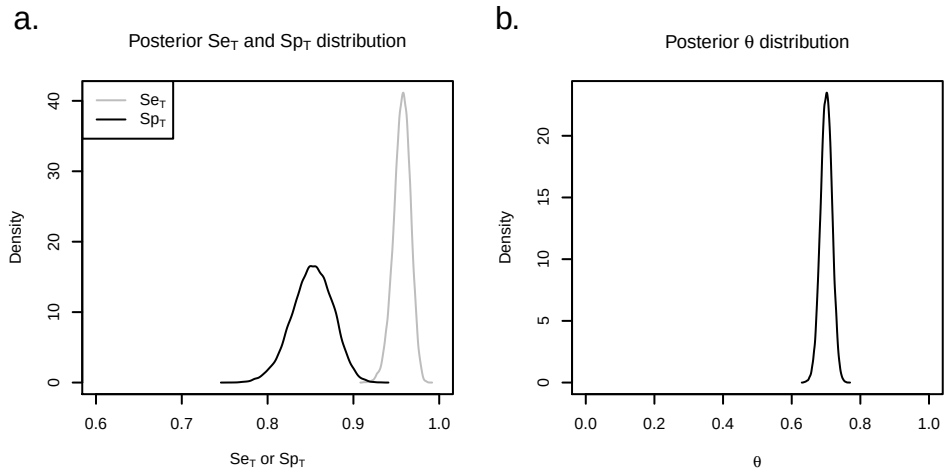

Figure 3.17: Posterior distribution for the VUmc data fitted with the proposed imperfect-reference-test model, with the flat Se_T and Sp_T prior distributions with $Se_T + Sp_T > 1$ restriction and the 'optimistic' AUC_a-prior distribution. **a**. Posterior Se_T (grey line) and Sp_T (black line). **b**. Posterior prevalence of disease (θ).

Figure 3.18 shows the estimated probability of AD for the resulting score based on the optimal combination of the three biomarkers by clinical diagnosis. As it was the case for the ADNI-data set, the plot in panel a of 3.18 illustrates the imperfect nature of the clinical diagnosis, i.e., potential misclassification of several individuals. In particular, 48 subjects diagnosed as having AD with a diagnostic score smaller than 22.62 have less than 50% probability of being truly AD patients. On the other hand, 36 subjects diagnosed clinically as not having AD, but with a diagnostic score larger than 24.7, have more than 50% probability of being truly AD patients.

Panel b of Figure 3.18 shows the posterior probability of AD as a function of diagnostic score by clinical diagnosis. These posterior probabilities express the probability of having AD considering the information from both the biomarkers and the clinical diagnosis results. Results show that for 35 subject diagnosed as AD patients, the posterior probability of being truly AD is less than 50%. Of the subjects not clinically diagnosed as AD patients, 26 have a posterior probability of AD of more than 50%.

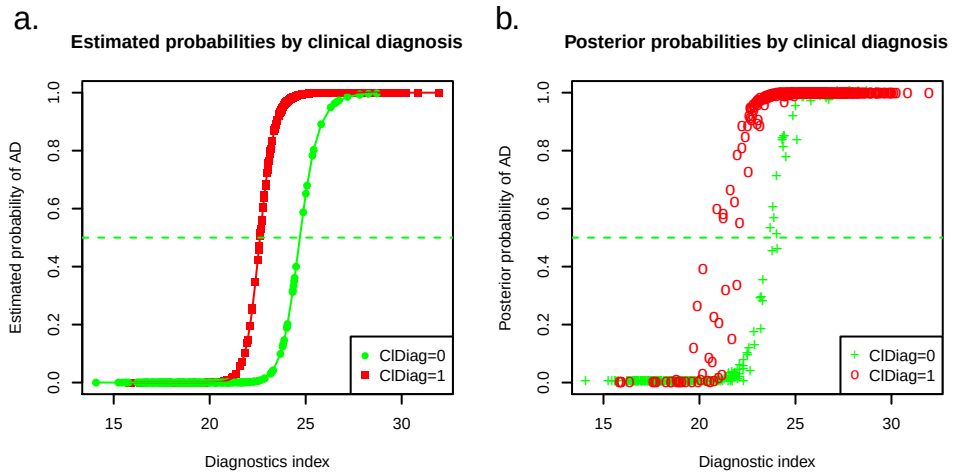

Figure 3.18: Probability of AD for estimated optimal combination score by clinical diagnosis (ClDiag) for the VUmc data set. Clinical controls indicated in green, clinical cases in red. **a**. Estimated probability of AD. **b**. Posterior probability of AD.

3.7 Conclusions

In this chapter, we have proposed a Bayesian latent-class mixture model which estimates the accuracy of an optimal linear-combination of continuous biomarkers while accounting for the use of an imperfect reference-test. Moreover, we have proposed a parametrisation that allows a more controlled way of introducing prior information to the model.

Application of the model may encounter non-convergence problems, especially in small data sets. In the simulations and particular examples considered in this chapter, the model provided unbiased estimates of AUC of the optimal linear-combination of biomarkers when convergence was obtained. The performance of the model was also satisfactory for simulated skew-normal data. Hence, the model is shown to be robust against some violation of the binormal assumption concerning the biomarker distributions conditional on true disease status.

In both the ADNI and VUmc data application, inspection of the posterior results for Se_T, Sp_T, and θ shows that the prior information is substantially updated by the data. This observation provides evidence that both of the suggested truncations (Se_T and $Sp_T > 0.5$ or $Se_T + Sp_T > 1$) were successful in allowing estimation of these parameters in the particular data application [35]. Although this does not guarantee identifiability in every application, it demonstrates that in some applications relatively

little prior information may be sufficient to obtain sensible results.

The results obtained for the two forms of truncated prior distributions for Se_T and Sp_T were essentially equal. It is important to note that the different forms lead to different marginal prior distributions for Se_T and Sp_T. These differences can play a role when data contain less information to update the prior information.

While accounting for the imperfectness of clinical diagnosis in the analysis of the ADNI and VUmc data sets, substantially higher estimates of the accuracy of the combination of the CSF-biomarkers were obtained as compared to the analysis which assumed that the diagnosis was perfect. Given the conditional independence assumption, this is an expected result [59]. In the next chapter, we will investigate the effect of the assumption on the results.

Chapter 4

Allowing for conditional dependence between biomarkers and the imperfect reference-test

In the current chapter, an extension of the Bayesian latent-class mixture model, developed in Chapter 3, is proposed. The extension allows taking dependence between biomarkers and the imperfect reference-test into account. The problem setting is discussed in Section 4.1. The Bayesian latent-class model allowing for conditional dependence is developed in Section 4.2. Particular prior distributions allowing for the introduction of sensible prior information, while mitigating model non-identifiability, are proposed in Section 4.3. The performance and applicability of the model are investigated by performing a simulation study in Section 4.4 and by applying it to two AD data sets in Section 4.5. The results from these applications are discussed in Section 4.7. Concluding remarks are formulated in Section 4.8.

4.1 Problem setting

All developments and results from Chapter 3 are based on the conditional independence assumption. This assumption implies independence between the continuous biomarker observations and the results from the dichotomous imperfect reference-

test, conditionally on the true disease status of the subjects. Because conditional dependence describes how misclassification by the imperfect reference-test and the biomarkers are related, it has an impact on the bias in accuracy estimates when ignoring the imperfectness of the reference test. The examples presented in Section 2.3 of Chapter 2 show that the introduction of conditional dependence results in overestimation of the accuracy of a new test, while underlying conditional independence resulted in underestimation of the accuracy of the same test.

To account for possible conditional dependence in estimating the diagnostic accuracy of several dichotomous tests in latent-class analysis, Yang and Becker [128] have extended the ideas of Rindskopf and Rindskopf [95] by introducing continuous random effects. Xu and Craig [127] proposed a probit latent-class model to account for the conditional dependence between dichotomous diagnostic-tests. These models employ the EM-algorithm to obtain maximum-likelihood estimates of the diagnostic-test accuracy. They also allow inclusion of the imperfect reference-test information in the form of covariate information, but, as mentioned before, the models require certain strict identifiability restrictions to do so.

Fully-parametric Bayesian latent-class models allowing for conditional dependence between dichotomous tests were also proposed. The models suggested by Menten et al. [69] and Dendukuri et al. [25] extend the model proposed by Joseph et al. [50] to account for conditional dependence.

In this chapter, we propose an extension of the Bayesian latent-class mixture model developed in Chapter 3 to allow for conditional dependence between the continuous biomarkers and the results of a dichotomous imperfect reference-test.

4.2 Methodology

To extend the model proposed in Chapter 3, we start by the decomposition of the full-data likelihood $P\left(\boldsymbol{Y}, \boldsymbol{t}, \tilde{\boldsymbol{d}}\right)$, as defined in (3.3). Instead of simplifying (3.3) by making the conditional independence assumption, we propose a form of $P\left(\boldsymbol{t}|\boldsymbol{Y}, \tilde{\boldsymbol{d}}\right)$, i.e. the distribution of the imperfect reference-test results $\boldsymbol{t}$ conditional on the biomarker values $\boldsymbol{Y}$ and the latent true disease status $\tilde{\boldsymbol{d}}$.

To allow for dependence between the imperfect reference-test T and biomarkers $\boldsymbol{y}$, we propose to model the dependence through a latent continuous tolerance variable $\tilde{T}$, underlying T. In particular, we assume that T is the result of dichotomizing $\tilde{T}$, with

$$\tilde{T}|\tilde{D} = \tilde{d} \sim N\left(\mu_{\tilde{T}_{\tilde{d}}}, 1\right),$$

where $\tilde{d} = 0$ for true controls and $\tilde{d} = 1$ for true cases.

Consequently, the probability of observing a positive imperfect reference-test result conditional on latent true disease-status, π_0 and π_1 (3.2) for $\tilde{d} = 0$ and $\tilde{d} = 1$, respectively, is expressed as follows:

$$\pi_0 = 1 - \Phi\left(-\mu_{\tilde{T}_0}\right),$$
$$\pi_1 = 1 - \Phi\left(-\mu_{\tilde{T}_1}\right),$$

where $\mu_{\tilde{T}_{\tilde{d}}}$ denotes the mean of the continuous latent tolerance distribution for group $\tilde{d}$. Note that, without loss of generality, the variance of the tolerance distribution can be fixed to 1 (see, e.g., Renard et al. [94]).

By considering the joint distribution of $\tilde{T}$ and $\boldsymbol{y}$ conditional on the latent true disease status $\tilde{D}$, their correlation can be introduced directly. Assume that, conditionally on $\tilde{D}$, $\tilde{T}$ and $\boldsymbol{y}$ are jointly normally distributed:

$$\begin{pmatrix} \tilde{T} \\ \boldsymbol{y} \end{pmatrix} | \tilde{D} = \tilde{d} \sim N_{K+1}\left(\begin{pmatrix} \mu_{\tilde{T}_{\tilde{d}}} \\ \boldsymbol{\mu}_{Y_{\tilde{d}}} \end{pmatrix}, \overline{\boldsymbol{\Sigma}}_{\tilde{d}}\right),$$

with

$$\overline{\boldsymbol{\Sigma}}_{\tilde{d}} = \begin{pmatrix} 1 & \boldsymbol{\tau}_{\tilde{d}}^T \\ \boldsymbol{\tau}_{\tilde{d}} & \boldsymbol{\Sigma}_{\tilde{d}} \end{pmatrix} \tag{4.1}$$

and

$$\boldsymbol{\tau}_{\tilde{d}} = \left(\rho_{\tilde{d},1}\sigma_{\tilde{d},1}, \ldots, \rho_{\tilde{d},K}\sigma_{\tilde{d},K}\right)^T.$$

In (4.1), $\mu_{\tilde{T}_{\tilde{d}}}$ and $\boldsymbol{\mu}_{Y_{\tilde{d}}}$ are the mean value and mean vector of $\tilde{T}$ and $\boldsymbol{y}$ in group $\tilde{d}$, respectively; $\overline{\boldsymbol{\Sigma}}_{\tilde{d}}$ is the overall variance-covariance matrix of $\tilde{T}$ and $\boldsymbol{y}$ in group $\tilde{d}$, containing the variance of $\tilde{T}$ (fixed at 1), the variance-covariance matrix $\boldsymbol{\Sigma}_{\tilde{d}}$ of the continuous-biomarker vector $\boldsymbol{y}$, and the vector of covariances $\boldsymbol{\tau}_{\tilde{d}}$. The covariance of $\tilde{T}$ and the k-th biomarker is expressed as the product of the correlation coefficient $\rho_{\tilde{d},k}$ and biomarkers' standard deviation $\sigma_{\tilde{d},k}$.

By using the joint normal distribution (4.1), the distribution of the imperfect reference-test T, conditional on $\boldsymbol{y}$ and $\tilde{D}$, can be defined by considering the distribution of $\tilde{T}$ conditional on $\boldsymbol{y}$ and $\tilde{D}$. As this conditional distribution has mean

$\mu_{\tilde{T}_{\tilde{d}}} + \boldsymbol{\tau}_{\tilde{d}}^T \boldsymbol{\Sigma}_{\tilde{d}}^{-1} \left(\boldsymbol{y} - \boldsymbol{\mu}_{Y_{\tilde{d}}} \right)$ and variance $1 - \boldsymbol{\tau}_{\tilde{d}}^T \boldsymbol{\Sigma}_{\tilde{d}}^{-1} \boldsymbol{\tau}_{\tilde{d}}$, it follows that

$$T|\boldsymbol{y}, \tilde{D} = \tilde{d} \sim Bern\left(\pi_{\tilde{d}}\left(\boldsymbol{y}\right)\right), \tag{4.2}$$

where

$$\pi_0\left(\boldsymbol{y}\right) = 1 - \phi\left(\frac{-\mu_{\tilde{T}_0} + \boldsymbol{\tau}_0^T \boldsymbol{\Sigma}_0^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}_{Y_0}\right)}{\sqrt{1 - \boldsymbol{\tau}_0^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\tau}_0}}\right) \equiv 1 - Sp_T\left(\boldsymbol{y}\right),$$

$$\pi_1\left(\boldsymbol{y}\right) = 1 - \phi\left(\frac{-\mu_{\tilde{T}_1} + \boldsymbol{\tau}_1^T \boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}_{Y_1}\right)}{\sqrt{1 - \boldsymbol{\tau}_1^T \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\tau}_1}}\right) \equiv Se_T\left(\boldsymbol{y}\right).$$

From (4.2) it follows that the imperfect reference-test T has a different sensitivity $\left[Se_T\left(\boldsymbol{y}\right) = P\left(T = 1|\boldsymbol{y}, \tilde{D} = 1\right)\right]$ and specificity $\left[Sp_T\left(\boldsymbol{y}\right) = P\left(T = 0|\boldsymbol{y}, \tilde{D} = 0\right)\right]$ for each possible value $\boldsymbol{y}$, which introduces the dependence between T and $\boldsymbol{y}$ conditionally on true disease status $\tilde{D}$.

Combining all the developments, we arrive at the following full-data likelihood function for a data set including observations for N individuals (indexed by i) on K biomarkers and an imperfect reference-test T:

$$L\left(\boldsymbol{\mu}_{Y_0}, \boldsymbol{\mu}_{Y_1}, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1, \mu_{\tilde{T}_0}, \mu_{\tilde{T}_1}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \theta | \boldsymbol{Y}, \boldsymbol{t}, \tilde{\boldsymbol{d}}\right) =$$

$$\prod_{i=1}^{N} \left(\left\{1 - Se_T\left(\boldsymbol{y}_i\right)\right\}^{\left(1-t_i\right)} \left\{Se_T\left(\boldsymbol{y}_i\right)\right\}^{t_i} \frac{exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_i - \boldsymbol{\mu}_{Y_1}\right)^T \boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{y}_i - \boldsymbol{\mu}_{Y_1}\right)\right\}}{\sqrt{\left(2\pi\right)^K \left|\boldsymbol{\Sigma}_1\right|}} \theta \right)^{\tilde{d}_i}$$

$$\times \left(\left\{1 - Sp_T\left(\boldsymbol{y}_i\right)\right\}^{t_i} \left\{Sp_T\left(\boldsymbol{y}_i\right)\right\}^{\left(1-t_i\right)} \frac{exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_i - \boldsymbol{\mu}_{Y_0}\right)^T \boldsymbol{\Sigma}_0^{-1}\left(\boldsymbol{y}_i - \boldsymbol{\mu}_{Y_0}\right)\right\}}{\sqrt{\left(2\pi\right)^K \left|\boldsymbol{\Sigma}_0\right|}} \left(1 - \theta\right) \right)^{1-\tilde{d}_i},$$

where $\tilde{\boldsymbol{d}} = \left(\tilde{d}_1, \ldots, \tilde{d}_N\right)^T$ and $\boldsymbol{t} = \left(t_1, \ldots, t_N\right)^T$ are the vectors containing, respectively, the true (unobserved) disease-status indicators and observed reference-test results for the N individuals, while $\boldsymbol{Y} = \left(\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_N^T\right)^T$ is the $N \times K$ matrix containing the observed biomarker values. The parameters of interest are $\boldsymbol{\mu}_{Y_0}$, $\boldsymbol{\mu}_{Y_1}$, $\boldsymbol{\Sigma}_0$, and $\boldsymbol{\Sigma}_1$, because together they define AUC_a, as indicated in (2.5).

4.3 Prior distributions

Prior distributions are specified essentially in the same way as an in Chapter 3. To mitigate model non-identifiability and allow the introduction of sensible prior infor-

mation, prior distributions as shown in Table 4.1 are proposed.

Table 4.1: Structure of the considered prior distributions (assuming $K = 3$ biomarkers).

Parameter	**Prior distribution**
Prevalence	
θ	$U\left(\frac{1}{N}, \left(1-\frac{1}{N}\right)\right)$
Parameters of the dichotomous reference test	
Se_T	$Beta\,(a,b)\,trunc(0.5,1)$
Sp_T	$Beta\,(c,d)\,trunc(0.5,1)$
Mean of biomarker values control group	
$\boldsymbol{\mu}_{Y_0}$	$N_3\left(\mathbf{0}, \mathbf{I}_3 10^6\right)$
Scaled difference biomarker distribution means	
$\boldsymbol{\delta}$	$N_3\left(\boldsymbol{\kappa}, \boldsymbol{\Psi}\right)$
Biomarker-distribution standard deviations	
$\sigma_{\tilde{d},k}$	$U\,(0,1000)$
Cholesky-factor values of correlation matrix $\boldsymbol{R}_{\tilde{d}}$	
$l_{\tilde{d},21}$	$U\,(-1,1)$
$l_{\tilde{d},31}$	$U\,(-1,1)$
$l_{\tilde{d},41}$	$U\,(-1,1)$
$l_{\tilde{d},32}$	$U\left(-\sqrt{1-l^2_{\tilde{d},31}}, \sqrt{1-l^2_{\tilde{d},31}}\right)$
$l_{\tilde{d},42}$	$U\left(-\sqrt{1-l^2_{\tilde{d},41}}, \sqrt{1-l^2_{\tilde{d},41}}\right)$
$l_{\tilde{d},43}$	$U\left(-\sqrt{1-l^2_{\tilde{d},41}-l^2_{\tilde{d},42}}, \sqrt{1-l^2_{\tilde{d},41}-l^2_{\tilde{d},42}}\right)$

The considered prior distributions largely coincide with the *controlled* prior distributions discussed in the previous chapter, though some particularities due to the conditional dependence case are worth mentioning.

The prior distributions for the means of the latent continuous tolerance variable $\tilde{T}$, $\mu_{\tilde{T}_0}$ and $\mu_{\tilde{T}_1}$, are derived from the considered prior distributions for Se_T and Sp_T of the dichotomous imperfect reference-test T. This way, restrictions can be enforced on $\mu_{\tilde{T}_0}$ and $\mu_{\tilde{T}_1}$, leading to a sensible interpretation of Se_T and Sp_T. For case-control data, a sensible choice for Se_T and Sp_T prior-distributions is to use independent Laplace Beta-distributions [10] restricted to the $(0.5, 1]$ interval. Based on the relationship defined in (4.2), this leads to the following prior distributions ($\phi\,(\cdot)$

denotes the standard-normal density function):

$$Se_T \sim Beta(a,b)trunc\left[0.51,1\right],$$

$$f_{\mu_{\tilde{T},1}}\left(\mu_{\tilde{T},1}\right) = \begin{cases} \frac{1}{B(a,b)}\left(\Phi\left(\mu_{\tilde{T},1}\right)\right)^{(a-1)}\left(1-\Phi\left(\mu_{\tilde{T},1}\right)\right)^{(b-1)}\left|\phi\left(\mu_{\tilde{T},0}\right)\right| & \text{if } \mu_{\tilde{T},1} \in \left(\Phi\left(0.51\right),+\infty\right] \\ 0 & \text{otherwise} \end{cases}$$

$$Sp_T \sim Beta(c,d)trunc\left[0.51,1\right],$$

$$f_{\mu_{\tilde{T},0}}\left(\mu_{\tilde{T},0}\right) = \begin{cases} \frac{1}{B(c,d)}\left(\Phi\left(-\mu_{\tilde{T},0}\right)\right)^{(c-1)}\left(1-\Phi\left(-\mu_{\tilde{T},0}\right)\right)^{(d-1)}\left|-\phi\left(-\mu_{\tilde{T},0}\right)\right| & \text{if } \mu_{\tilde{T},0} \in \left(-\infty,-\Phi\left(0.51\right)\right] \\ 0 & \text{otherwise} \end{cases}.$$

As discussed in Section 3.3, other restrictions for the prior distributions of Se_T and Sp_T could be considered as well. Different choices may imply a dependence between Se_T and Sp_T, which could not be trivial to interpret and/or implement. For this reason, we limit ourselves in the current chapter to assuming that Se_T and Sp_T are both strictly larger than 0.5. This restriction resolves the label-switching problem observed for mixture models [67] and mitigates the over-parametrisation with multiple imperfect reference-tests [26], two consequences of model non-identifiability.

For the prior distributions for the biomarker variance-covariance matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$, we propose again to construct flat prior-distributions as shown in [125]. As described in Section 3.3, this entails considering flat priors directly on the standard deviations and correlation coefficients of $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$. In the proposed parametrisation for the conditional dependence case (4.1), $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ are contained in the overall variance-covariance matrices $\overline{\boldsymbol{\Sigma}}_0$ and $\overline{\boldsymbol{\Sigma}}_1$. Therefore, prior distributions have to be considered for $\overline{\boldsymbol{\Sigma}}_0$ and $\overline{\boldsymbol{\Sigma}}_1$ to include prior distributions for the correlation coefficients describing the potential conditional dependence between the biomarkers and the imperfect reference-test. In particular, the overall variance-covariance matrix is decomposed as $\overline{\boldsymbol{\Sigma}}_{\tilde{d}} = \boldsymbol{S}_{\tilde{d}}\boldsymbol{R}_{\tilde{d}}\boldsymbol{S}_{\tilde{d}}$, where $\boldsymbol{S}_{\tilde{d}}$ and $\boldsymbol{R}_{\tilde{d}}$ are, respectively, the diagonal matrix of standard deviations and the correlation matrix for disease group $\tilde{d}$. Additionally, $\boldsymbol{R}_{\tilde{d}}$ is expressed as $\boldsymbol{R}_{\tilde{d}} = \boldsymbol{L}_{\tilde{d}}\boldsymbol{L}_{\tilde{d}}^T$, where $\boldsymbol{L}_{\tilde{d}}$ is a lower-triangular matrix. Subsequently, wide uniform distributions are put directly on the biomarker standard deviations $\sigma_{\tilde{d}}$ included in $\boldsymbol{S}_{\tilde{d}}$, while the remaining tolerance standard deviations in $\boldsymbol{S}_{\tilde{d}}$ are fixed to 1. Finally, flat priors are put on K of the $\left(\left(K+1\right)^2-\left(K+1\right)\right)/2$ non-zero off-diagonal elements of the Cholesky decomposition-factor $\boldsymbol{L}_{\tilde{d}}$ of $\boldsymbol{R}_{\tilde{d}}$ (see Table 4.1).

4.4 Simulation study

To investigate the performance of the model, we carried out a simulation study. The goal of this simulation study was to show adequate model performance in the conditional-dependence case and subsequently, to investigate the impact of violating the conditional-independence assumption.

4.4.1 Data

We simulated 400 data sets of size 600 under the conditional-dependence setting. The underlying parameter settings are summarized in Table 4.2. These parameter values yield biomarker data (for $K = 3$ biomarkers) with an underlying true AUC_a of 0.787 and an imperfect reference-test with $Se_T = Sp_T = 0.85$. The underlying latent tolerance and biomarkers' correlation coefficients were set as follows: $\rho_{1,0} = \rho_{1,1} = 0$, $\rho_{2,0} = \rho_{2,1} = 0.7$, and $\rho_{3,0} = \rho_{3,1} = 0.3$.

Table 4.2: Parameter values underlying the sampled simulation data sets.

Parameter	**Value**
Multivariate parameters	
$\boldsymbol{\mu}_0$	$(0, 0, 0)^T$
$\boldsymbol{\mu}_1$	$(1.1683, 1.3490, 1.5082)^T$
Dependent biomarkers $\left(\boldsymbol{\rho} = (0.5, 0.9, 0.5)^T\right)$	
$\boldsymbol{\Sigma}_0$	$\begin{pmatrix} 1 & 0.5 \times \sqrt{1 \times 3} & 0.9 \times \sqrt{1 \times 2} \\ 0.5 \times \sqrt{1 \times 3} & 3 & 0.5 \times \sqrt{3 \times 2} \\ 0.9 \times \sqrt{1 \times 2} & 0.5 \times \sqrt{3 \times 2} & 2 \end{pmatrix}$
$\boldsymbol{\Sigma}_1$	$\begin{pmatrix} 2 & 0.5 \times \sqrt{2 \times 1} & 0.9 \times \sqrt{2 \times 3} \\ 0.5 \times \sqrt{2 \times 1} & 1 & 0.5 \times \sqrt{1 \times 3} \\ 0.9 \times \sqrt{2 \times 3} & 0.5 \times \sqrt{1 \times 3} & 3 \end{pmatrix}$
Conditional dependent correlations	
$\rho_{0,1} = \rho_{1,1}$	0
$\rho_{0,2} = \rho_{1,2}$	0.7
$\rho_{0,3} = \rho_{1,3}$	0.5
Parameters of the reference test	
Se_T	0.85
Sp_T	0.85
Prevalence	
θ	0.5
Functions of multivariate parameters	
AUC_1	0.75
AUC_2	0.75
AUC_3	0.75
$\boldsymbol{a}$	$(0.1594, 0.2237, 0.0972)^T$
Optimal combination parameters	
$\mu_{a,0}$	0
$\mu_{a,1}$	0.6347
$\sigma^2_{a,0}$	0.3490
$\sigma^2_{a,1}$	0.2857
AUC_a	0.7872

The underlying (unobserved) true joint distribution for biomarker 1 and the latent-tolerance from the simulation study, together with two conditional tolerance distributions given a particular biomarker value, are shown in Figure 4.1. The joint distribution for the biomarker and the latent-tolerance is shown in panel a of Figure 4.1. It is clear that, conditional on true disease status, the biomarker values and latent-tolerance values are independent ($\rho_{1,0} = \rho_{1,1} = 0$). Furthermore, the con-

ditional independence is also shown in panels b and c, where different conditional latent-tolerance distributions ($y_1 = 1$ and $y_1 = 2$) lead to the same conditional sensitivity and specificity. Moreover, these conditional sensitivities and specificities are equal to the marginal $Se_T = 0.85$ and $Sp_T = 0.85$, respectively.

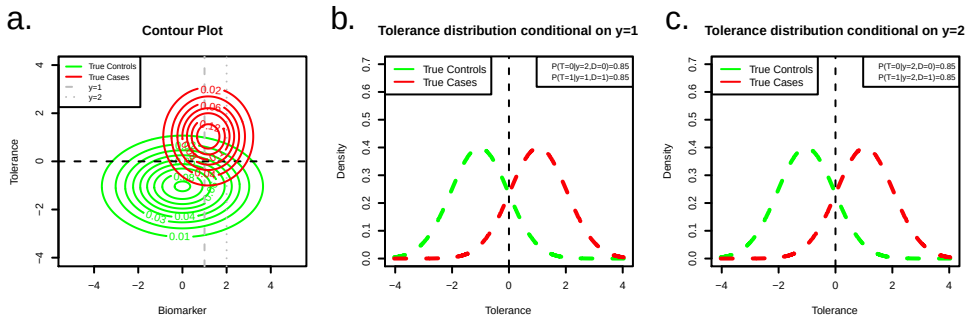

Figure 4.1: True underlying biomarker and latent-tolerance distributions for biomarker 1 in the simulation study, independent of the latent-tolerance variable conditional on true disease status. **a**. Joint distribution for the biomarker and latent-tolerance. **b**. Latent-tolerance distributions conditional on $y_1 = 1$. **b**. Latent-tolerance distributions conditional on $y_1 = 2$. True-control distributions are indicated by the green dashed line, true-case distributions by the red dashed line. Grey dashed and dotted line indicate $y_1 = 1$ and $y_1 = 2$, respectively.

Figure 4.2 shows the true underlying joint distribution and two conditional latent-tolerance distributions for biomarker 2 from the simulation study. This biomarker is correlated with the imperfect dichotomous reference-test, conditional on true disease status ($\rho_{2,0} = \rho_{2,1} = 0.7$). The correlation is easily observed from the elliptical contours for the joint distributions in panel a of Figure 4.2. In panels b and c, the latent-tolerance distributions are shown conditional on $y_2 = 1$ and $y_2 = 2$, respectively. From these panels the conditional dependence can be seen by observing that sensitivity and specificity of the dichotomized tolerance variable depend on the particular value of y_2. Moreover, both of the indicated conditional sensitivity and specificity values are different from the marginal Se_T and Sp_T.

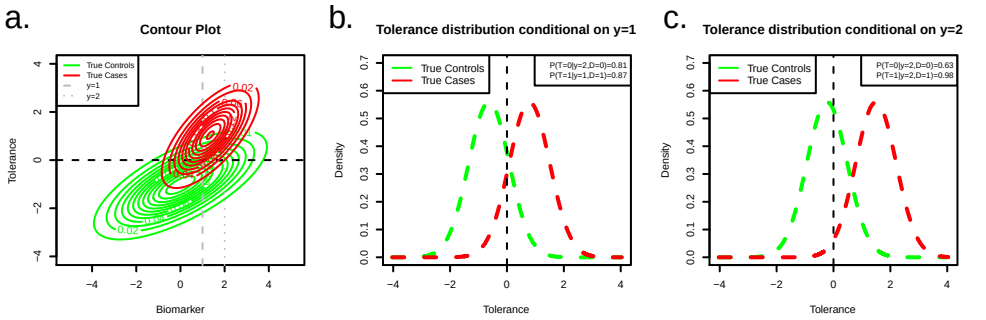

Figure 4.2: True underlying biomarker and latent-tolerance distributions for biomarker 2 in the simulation study, dependent on the latent-tolerance variable conditional on true disease status. **a**. Joint distribution for the biomarker and latent-tolerance. **b**. Latent-tolerance distributions conditional on $y_2 = 1$. **b**. Latent-tolerance distributions conditional on $y_2 = 2$. True-control distributions are indicated by the green dashed line, true-case distributions by the red dashed line. Grey dashed and dotted line indicate $y_2 = 1$ and $y_2 = 2$, respectively.

4.4.2 Prior distributions

Prior distributions for the simulation study were considered as summarized in Table 4.1. The prior distribution for the prevalence of disease θ was a uniform distribution between 0.1 and 0.9. Parameters a, b, c, and d of the Beta prior distributions for Se_T and Sp_T were all set to 1, leading to flat-uniform priors. Moreover, these prior distributions were restricted to ensure that both Se_T and $Sp_T > 0.5$ (see Section 3.3). For AUC_a, the 'optimistic' $\boldsymbol{\delta}$ prior was considered with $\boldsymbol{\kappa} = (0, 0, 0)^T$ and $\boldsymbol{\Psi}$ such that its standard deviations and correlation coefficients were all equal to 0.7 and 0.6, respectively. This prior distribution disfavours small AUC_a values, as seen in panel a of Figure 3.9, but still allows a wide range of plausible AUC_a values.

4.5 Real data

The applicability of the proposed model was investigated by fitting it to two data sets containing data from AD patients: the VUmc (VU University Medical Center) data set, which consists of patients from the memory-clinic-based Amsterdam Dementia Cohort, and the publically available data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). Both of these data sets have been discussed in Section 2.5.2 of Chapter 2.

As in the previous chapter, the measurements of total tau and p-tau$_{181p}$ from both data sets were log-transformed. The resulting observed data distributions are shown in Figures 3.10 and 3.11 for the ADNI and VUmc data, respectively.

4.5.1 Prior distributions

For both data sets, the parameters of the Beta prior distributions for Se_T and Sp_T were set so that they allowed capturing the available information from literature [126, 11, 104, 116]. Three studies [11, 104, 116] reported high sensitivity of the clinical AD diagnosis (ranging from 81.8% to 100%) in a mixed dementia setting; another study [126] reported much worse sensitivities ranging from 39% to 95% and specificities ranging from 33% to 100%. Based on this information, we formulated conservative informative prior distributions for Se_T and Sp_T. In particular, $Beta(c = 4.15, d = 2.54)$ and $Beta(a = 2.69, b = 1.99)$ distributions were used, respectively, truncated to the $[0.51, 1]$ interval for Se_T and Sp_T. These literature-based informative priors are shown in Figure 4.3. The prior distributions have a mean value of 0.620 and 0.575 for Se_T and Sp_T, respectively. The 95% equal-tail interval for Se_T ranges between 0.258 and 0.915, while the corresponding interval for Sp_T is equal to [0.164;0.927].

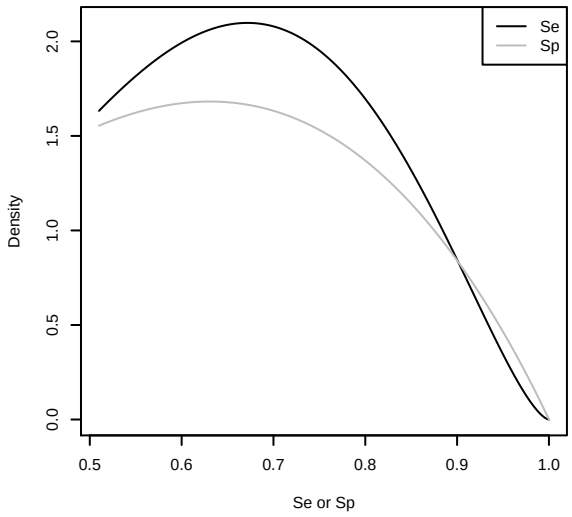

Figure 4.3: Considered literature-based informative priors for Se_T and Sp_T. Se_T prior denoted by the solid black line, Sp_T prior by the solid grey line.

As in Chapter 3, the prior information for AUC_a was varied to investigate sensitivity of the results to the choice of the prior distribution. In addition to the 'optimistic' AUC_a prior (see Section 4.4.2), also a 'conservative' prior, favouring moderate AUC_a values, was considered by setting $\boldsymbol{\kappa} = (0, 0, 0)^T$ and defining $\boldsymbol{\Psi}$ by standard deviations

equal to 0.5 and setting correlation coefficients to 0.3 (see panel *b* of Figure 3.9).

The prior distributions for the remaining parameters were considered as in Table 4.1.

4.6 Analysis settings

The proposed model, allowing for conditional dependence, was fitted to the simulated data and the two case studies. The model assuming conditional independence, as developed in the previous chapter, was fitted to the data as well. The latter model can be obtained from the former by fixing the correlation parameters, $\rho_{\tilde{d},k}$, expressing dependence between the latent continuous tolerance variable $\tilde{T}$ and biomarkers, to zero.

Each fit was obtained by sampling 10,000 iterations from the posterior distributions from five independent MCMC chains after discarding the first 10,000 as a burn-in. Starting values of the MCMC chains were set, as described in Section 3.4 of Chapter 3, by fixing them at plausible data-based values for all parameters with exception of Se_T and Sp_T, which were started at the midpoint of their parameter space, i.e., 0.75.

After fitting the models, the results were first checked by general diagnostic-tools in order to assess convergence of the MCMC chains. Convergence over chains was investigated by the Gelman-Rubin convergence index, for which a cut-off value of 1.1 was applied [38]. Chain-by-chain convergence was checked by using the Geweke convergence criterion [40]. The Geweke criterion was monitored to ensure that overall at least three out of five chains converged.

The models were fitted by using OpenBUGS 3.2.1 [60]. Annotated BUGS model codes can be found in Section A.2 of Appendix A. Results were analyzed and summarized using R 3.0.1 (x64) [90]. The R-package R2OpenBUGS [111] was used as an interface between R 2.14.2 and OpenBUGS. For the proposed conditional-dependence model, fitting times were equal to 20h for each simulated and VUmc data set, and 15h for the ADNI data, on a 64-bit, 2.8 GHz, 8GB RAM machine.

4.7 Results

4.7.1 Simulation study

The results of the simulation study are summarized in Table 4.3. For the proposed conditional-dependence model (the third column of the table), the mean of the 400

posterior medians is very close to the true underlying values for all parameters. The fourth column of Table 4.3 presents the results for the conditional-independence model. It can be concluded that, on average, AUC_a (true value of 0.787), Se_T (true value of 0.85), and Sp_T (true value of 0.85), are significantly overestimated with mean posterior-medians equal to 0.892, 0.982, and 0.995, respectively.

Table 4.3: Mean of posterior medians, (standard deviations), and [empirical 95% confidence intervals] for the simulated data (400 data sets of size N=600). Results are shown for both the conditional-dependence and conditional-independence models considering the 'optimistic' AUC_a-prior distribution.

		Model	
Parameter	**True**	**Conditional Dep.**	**Conditional Ind.**
$\boldsymbol{AUC_a}$	0.787	0.777 (0.023) [0.728;0.819]	0.892 (0.011) [0.869; 0.912]
$\boldsymbol{Se_T}$	0.85	0.834 (0.031) [0.770;0.894]	0.982 (0.013) [0.940; 0.993]
$\boldsymbol{Sp_T}$	0.85	0.837 (0.027) [0.782;0.887]	0.995 (0.001) [0.993; 0.996]
$\boldsymbol{\theta}$	0.5	0.505 (0.023) [0.458;0.554]	0.510 (0.017) [0.479; 0.542]
$\boldsymbol{\rho_{0,1}}$	0	0.022 (0.090) [-0.146;0.213]	0
$\boldsymbol{\rho_{0,2}}$	0.7	0.702 (0.046) [0.608;0.782]	0
$\boldsymbol{\rho_{0,3}}$	0.3	0.318 (0.079) [0.165;0.470]	0
$\boldsymbol{\rho_{1,1}}$	0	0.020 (0.088) [-0.147;0.197]	0
$\boldsymbol{\rho_{1,2}}$	0.7	0.705 (0.052) [0.596;0.795]	0
$\boldsymbol{\rho_{1,3}}$	0.3	0.310 (0.078) [0.158;0.457]	0

4.7.2 VUmc data

Table 4.4 presents the medians of the posterior distributions obtained for the VUmc data, together with their posterior standard deviations and 95%-credible intervals. Results are shown for both models (assuming conditional dependence and independence) and for both the 'conservative' (Cons) and 'optimistic' (Opt) AUC_a-priors. For the conditional-independence model, the medians of the AUC_a distributions corresponding to the 'conservative' and 'optimistic' AUC_a-priors are both equal to 0.995, with respective 95%-credible intervals equal to [0.990;0.997] and [0.991;0.998], respectively. For the conditional-dependence model, the corresponding posterior AUC_a medians are equal to 0.996 and 0.997, respectively, with 95%-credible intervals [0.992;0.998] and [0.993;0.998], respectively. Because both AUC_a-priors lead to practically the same results for both models, in what follows only the results for the 'optimistic' AUC_a-prior are discussed in more detail.

Allowing for conditional dependence leads to posterior median Se_T and Sp_T

Table 4.4: Posterior medians, (standard deviations), and [95%-credible intervals] for the VUmc data. In the respective columns results are shown for the conditional-dependence and conditional-independence models considering both the 'conservative' and 'optimistic' AUC_a-prior distribution. Correlation coefficients: $\rho_{0,1}$: total tau in the control group; $\rho_{1,1}$: total tau in the AD group; $\rho_{0,2}$: $A\beta_{1-42}$ in the control group; $\rho_{1,2}$: $A\beta_{1-42}$ in the AD group; $\rho_{0,3}$: p-tau$_{181p}$ in the control group; $\rho_{1,3}$: p-tau$_{181p}$ in the AD group.

	Model (AUC_a Prior)			
Parameter	**Cond. Dep. (Cons.)**	**Cond. Dep. (Opt.)**	**Cond. Ind. (Cons.)**	**Cond. Ind. (Opt.)**
AUC_a	0.996 (0.002) [0.992;0.998]	0.997 (0.001) [0.993;0.998]	0.995 (0.002) [0.990;0.997]	0.995 (0.002) [0.991;0.998]
Se_T	0.940 (0.012) [0.915;0.960]	0.940 (0.012) [0.915;0.960]	0.955 (0.010) [0.934;0.973]	0.954 (0.010) [0.934;0.971]
Sp_T	0.818 (0.027) [0.762;0.868]	0.823 (0.026) [0.768;0.870]	0.850 (0.024) [0.799;0.894]	0.850 (0.024) [0.799;0.894]
θ	0.705 (0.017) [0.671;0.738]	0.706 (0.017) [0.672;0.739]	0.700 (0.016) [0.668;0.732]	0.701 (0.016) [0.668;0.732]
$\rho_{0,1}$	0.366 (0.091) [0.176;0.520]	0.358 (0.093) [0.156;0.530]	0	0
$\rho_{0,2}$	0.110 (0.103) [-0.100;0.306]	0.104 (0.103) [-0.103;0.301]	0	0
$\rho_{0,3}$	0.034 (0.095) [-0.154;0.206]	0.030 (0.100) [-0.171;0.235]	0	0
$\rho_{1,1}$	0.293 (0.091) [0.102;0.454]	0.280 (0.097) [0.068;0.451]	0	0
$\rho_{1,2}$	0.186 (0.093) [-0.009;0.353]	0.186 (0.092) [-0.007;0.356]	0	0
$\rho_{1,3}$	0.199 (0.091) [0.014;0.366]	0.186 (0.098) [-0.018;0.366]	0	0

estimates of 0.940 and 0.823, respectively, with 95%-credible intervals equal to [0.915;0.960] and [0.768;0.870], respectively. The posterior distribution for θ has a median of 0.706 with a 95%-credible interval of [0.672;0.739]. The posterior distributions, shown in Figure 4.4, indicate that the assumed prior distributions for Se_T, Sp_T, and θ are well updated by the data, even for the 'conservative' AUC_a prior, suggesting a successful mitigation of model non-identifiability [35].

Moreover, the results show significant dependence between the clinical diagnosis

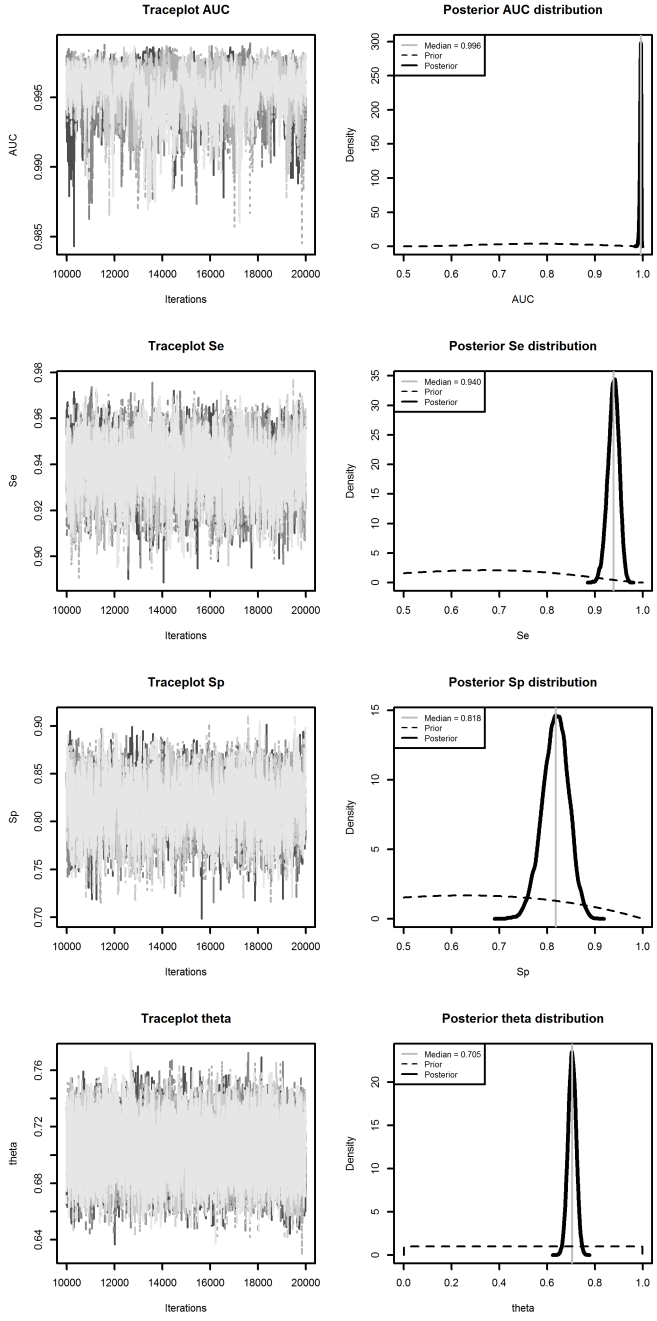

Figure 4.4: Traceplots and posterior distributions for AUC_a, Se_T, Sp_T and θ from the VUmc-data fit with the 'conservative' AUC_a prior-distribution.

of AD and total tau: the 95%-credible intervals for the correlation between the latent tolerance and total tau in the control ($\rho_{0,1}$) and AD ($\rho_{1,1}$) groups are equal to [0.156;0.530] and [0.068;0.451], respectively, and they both exclude the value of zero. For $A\beta_{1-42}$ ($\rho_{0,2}$ and $\rho_{1,2}$) and p-tau$_{181p}$ ($\rho_{0,3}$ and $\rho_{1,3}$) the 95%-credible intervals do not suggest any dependence.

Despite the correlation between the latent-tolerance and total tau, no important difference in the posterior medians of AUC_a is found between the results of the conditional-dependence and conditional-independence models. In particular, the medians are equal to 0.997 and 0.995 for the former and the latter, respectively, with overlapping respective 95%-credible intervals of [0.993;0.998] and [0.991;0.998].

4.7.3 ADNI data

Table 4.5 presents the results for the ADNI data. In this case, the resulting MCMC-samples for the conditional-dependence model defined by using the 'conservative' AUC_a-prior require some attention. Even after 300,000 iterations, the OpenBUGS MCMC-algorithms do not seem to have converged. Though both the Gelman-Rubin, and Geweke convergence criteria are satisfied, inspection of the posterior distributions shown in Figure 4.6 reveals that, for AUC_a, Se_T, Sp_T, and θ, bi-modal posteriors are obtained caused by erratic jumps of the MCMC chains which were not able to converge to one of two stationary distributions as indicated by the traceplots in Figure 4.6. This may be taken as implying that the use of the 'conservative' prior for AUC_a does not provide enough information to overcome potential non-identifiability of the model given the limited size of the data set. For this reason, the ADNI data were also fitted with an 'intermediate' AUC_a-prior (see Figure 4.5). This AUC_a-prior distribution is characterized by $\boldsymbol{\kappa} = (0, 0, 0)^T$ and $\boldsymbol{\Psi}$ defined by standard deviations and correlation coefficients equal to $(0.6, 0.6, 0.6)$ and $(0.5, 0.5, 0.5)$, respectively. The 'intermediate' AUC_a-prior distribution has a mean of 0.802 and 95% equal-tail interval of [0.598;0.978]. Table 4.5 contains the results obtained for the 'optimistic' as well as the 'intermediate' AUC_a-prior distribution.

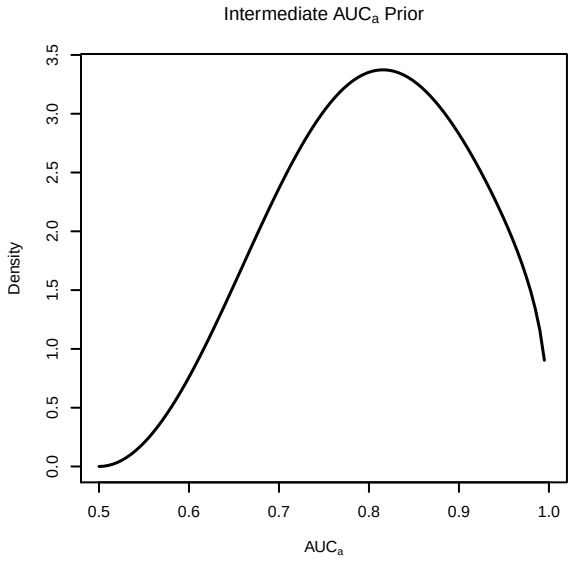

Figure 4.5: Considered 'intermediate' AUC_a-prior distribution.

No significant differences between the results obtained with the 'intermediate' and 'optimistic' AUC_a-prior distributions can be observed. Therefore, only the results from the 'optimistic' AUC_a-prior setting will be discussed. The posterior medians of Se_T and Sp_T are equal to 0.808 and 0.835, respectively. The corresponding 95%-credible intervals are equal to [0.688;0.905] and [0.691;0.930], respectively. The results obtained for the proposed conditional-dependence model do not indicate any significant correlation between the biomarkers and the latent tolerance underlying the AD diagnosis. The posterior AUC_a distributions for the conditional-dependence and conditional-independence models show only a slight difference. In particular, the posterior medians of AUC_a are equal to 0.979 and 0.983, respectively, with the respective 95% credible intervals of [0.939;0.994] and [0.961;0.994].

Table 4.5: Posterior medians, (standard deviations), and [95%-credible intervals] for the ADNI data. In the respective columns results are shown for the conditional-dependence and conditional-independence models considering the 'intermediate' and 'optimistic' AUC_a-prior distribution. Correlation coefficients: $\rho_{1,0}$: total tau in the control group; $\rho_{1,1}$: total tau in the AD group; $\rho_{0,2}$: $A\beta_{1-42}$ in the control group; $\rho_{1,2}$: $A\beta_{1-42}$ in the AD group; $\rho_{0,3}$: p-tau$_{181p}$ in the control group; $\rho_{1,3}$: p-tau$_{181p}$ in the AD group.

	Model (AUC_a Prior)			
Parameter	**Cond. Dep. (Int.)**	**Cond. Dep. (Opt.)**	**Cond. Ind. (Int.)**	**Cond. Ind. (Opt.)**
AUC_a	0.976 (0.022) [0.912;0.994]	0.979 (0.014) [0.939;0.994]	0.982 (0.009) [0.958;0.993]	0.983 (0.009) [0.961;0.994]
Se_T	0.805 (0.061) [0.669;0.906]	0.808 (0.056) [0.688;0.905]	0.818 (0.044) [0.726;0.898]	0.818 (0.044) [0.724;0.896]
Sp_T	0.829 (0.077) [0.627;0.940]	0.835 (0.061) [0.691;0.930]	0.880 (0.037) [0.798;0.942]	0.879 (0.037) [0.798;0.941]
θ	0.463 (0.091) [0.230;0.614]	0.466 (0.071) [0.317;0.597]	0.499 (0.043) [0.413;0.582]	0.498 (0.043) [0.414;0.582]
$\rho_{0,1}$	0.142 (0.225) [-0.297;0.549]	0.121 (0.185) [-0.264;0.459]	0	0
$\rho_{0,2}$	0.295 (0.250) [-0.323;0.637]	0.277 (0.223) [-0.288;0.584]	0	0
$\rho_{0,3}$	0.186 (0.222) [-0.257;0.560]	0.161 (0.183) [-0.226;0.494]	0	0
$\rho_{1,1}$	0.322 (0.182) [-0.073;0.621]	0.314 (0.177) [-0.318;0.394]	0	0
$\rho_{1,2}$	0.053 (0.195) [-0.352;0.414]	0.048 (0.182) [-0.318;0.394]	0	0
$\rho_{1,3}$	-0.079 (0.273) [-0.580;0.468]	-0.083 (0.252) [-0.543;0.412]	0	0

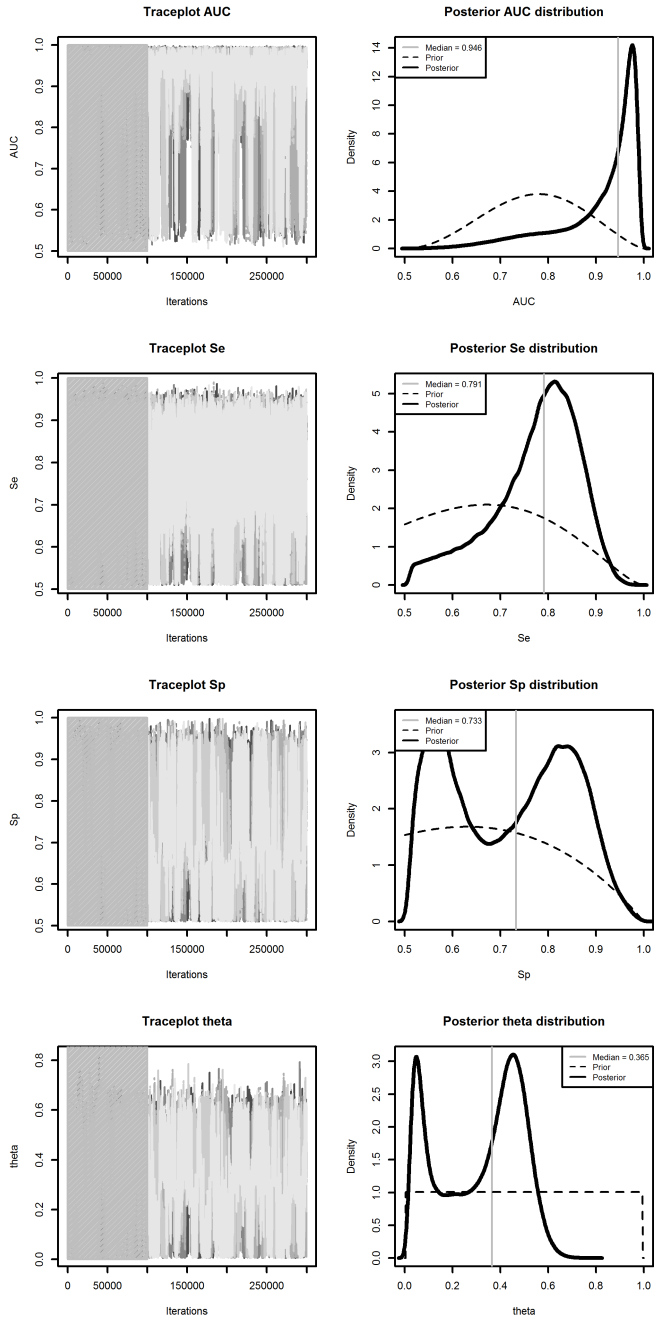

Figure 4.6: Traceplots and posterior distributions for AUC_a, Se_T, Sp_T and θ from the ADNI-data fit with the 'conservative' AUC_a prior-distribution. Shaded area in the traceplots denotes first 100,000 iterations burn-in.

4.8 Conclusions

In this chapter we have proposed a novel latent-class Bayesian mixture-model for construction of a diagnostic biomarker-index in the presence of a dichotomous imperfect reference-test. Importantly, in contrast to the currently available models, the model does not require the conditional-independence assumption, because it explicitly allows for a correlation between the results of the reference test and biomarkers.

The simulation study results showed adequate model performance leading to unbiased estimates of the model parameters. Given that the posterior distributions were substantially updated as compared to the assumed prior distributions, we concluded that non-identifiability was mitigated [35]. On the other hand, simulation study results showed that falsely assuming conditional independence may lead to substantial bias in the estimates of biomarker-index accuracy. The observed overestimation of AUC_a may be related to the statistically significant overestimation of Se_T and Sp_T. This can be explained by the model trying to capture the excess correlation of the conditional dependence by increasing the imperfect reference-test sensitivity estimate. Since the marginal positive rate for the reference test is fixed together with a stable disease prevalence estimate, specificity is overestimated as well. A similar observation was made for dichotomous tests in the paper by Pepe and Janes [84].

For the VUmc data set, the proposed parameter restrictions on Se_T, Sp_T, and θ, as well as the chosen prior distributions for Se_T, Sp_T, and AUC_a, addressed the model non-identifiability issues. From the results it is clear that prior distributions were substantially updated by the data for all parameters. Moreover, the effect of the type of prior distribution assumed for AUC_a did not seem to effect the results. Interestingly, the model suggested dependence between clinical diagnosis and total tau. Despite the dependence, there was not much difference in the posterior-median AUC obtained when assuming conditional independence or allowing for conditional dependence. A possible explanation could be that the bias in AUC_a is too small to be picked-up by the model [6]. Another explanation could be that the importance of total tau in the linear combination comprising the diagnostic index is limited [66]. In fact, the median of the posterior distribution for coefficient $\hat{a}_{totaltau}$ was equal to 9.43 and was smaller as compared to the medians for the two other biomarkers ($\hat{a}_{A\beta_{1-42}} = 15.04$; $\hat{a}_{p-tau_{181p}} = 21.92$). This observation shows that, in absolute terms, the results for total tau are the least important in the construction of the continuous diagnostic biomarker-index.

For the ADNI data, non-identifiability issues were resolved when the proposed parameter restrictions were considered in combination with the 'intermediate' or 'op-

timistic' AUC_a-prior. When the 'conservative' AUC_a-prior distribution was used, the OpenBUGS MCMC-algorithms failed to converge. This indicates that the use of the model may require a substantial sample size or, otherwise, a substantial amount of prior information to provide reliable results.

Chapter 5

Incorporation of retrospective information in prospective diagnostic biomarker-validation designs

The current chapter describes a Bayesian framework to estimate and validate the accuracy of a continuous diagnostic biomarker-index in an efficient way. Section 5.1 contains the problem setting for which the proposed framework provides a solution. The framework itself is described in Section 5.2. To investigate whether the proposed approach does indeed provide a more efficient way of validating biomarker accuracy a simulation study is performed as described in Section 5.3. The results of this simulation study are described in Section 5.4. Finally, a short concluding note is provided in Section 5.5.

5.1 Problem setting

A biomarker is 'a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to therapeutic interventions' [22]. As healthcare and drug-development costs continue to increase, many consider the identification and development of biomarkers as a solution to overcome this financial burden and ensure high-quality healthcare

in the future [5, 30]. To be useful, candidate biomarkers have to be properly validated. Unfortunately, statistical validation of biomarkers is challenging and costly when properly executed [16, 91].

In this chapter, we focus on the prospective validation of a diagnostic-biomarker index. As in previous chapters, interest lies in the accuracy of the index in distinguishing cases in a case-control population. The index can consist of a single biomarker or a combination of biomarkers. Combining continuous biomarkers into indices has been shown to increase the diagnostic accuracy over considering biomarkers alone [85, 81, 82, 63].

There is no clear general strategy for diagnostic-test development and validation [93]. Several suggestions for such a strategy have been made [83, 91]. For instance, according to Zhou, Obuchowski, and McClish [133], the assessment of the accuracy of diagnostic tests should be structured in three phases: an 'Exploratory', a 'Challenge', and a 'Clinical' phase. These phases categorize biomarker research in identification, development, and validation of diagnostic tests. As a starting point, we consider a set of identified diagnostic biomarkers from which we want to develop a biomarker index. This, usually limited, development study results in a biomarker index maximizing some diagnostic property.

As discussed in Section 2.1, the diagnostic properties of a continuous diagnostic-biomarker index can be summarized by a receiver operator characteristic (ROC) curve. Moreover, because it is more convenient to interpret a single measure than a complete curve, the area under the ROC curve (AUC) is often considered to represent the accuracy of the continuous diagnostic-biomarker index. The AUC can be interpreted as the probability that the observed value of a randomly selected case will be larger than that of a randomly selected control [7, 133, 135].

In Chapter 3 we have proposed to use a Bayesian latent-class mixture model to select the appropriate continuous diagnostic-biomarker index and to estimate its ROC curve and the corresponding AUC_a. This model allows estimating the accuracy of the diagnostic test even in the absence of a gold-standard (GS) reference-test. A GS reference-test provides perfect discrimination between cases and controls. In practice, such a reference test may not be available and this can potentially lead to biased accuracy estimates (see Section 3.6).

Moreover, the proposed Bayesian latent-class mixture model enables estimating the accuracy of a continuous diagnostic-index based on a linear combination of several biomarkers. By following the suggestion of Su and Liu [112], the combination is such that it maximizes the accuracy of the index expressed by AUC. After the development study, validation proceeds by setting up a prospective study aimed at re-estimation

of the accuracy of the developed continuous diagnostic-biomarker index. A validation criterion is agreed upon and, after gathering new data, the validation study confirms or disproves the validity of the index. Whether the sampling populations of the development and validation study have to be completely independent or not, depends on the goal of the validation at hand. For example, Zhou et al. [133] proposed overlapping population definitions for each of the diagnostic-test accuracy study phases. In principle, the entire range between completely independent and completely equal sampling populations can be of interest in the validation of diagnostic biomarkers.

In cases when development and validation populations are not required to be completely independent, one could think about introducing information from the development study into the validation study. Frequentist approaches can only manage such an introduction under the assumption of complete equality between the sampling populations, which leads to data pooling. Applying the Bayesian framework allows a more focused introduction of prior information [110]. By summarizing available knowledge in a prior distribution, the knowledge can be updated by new data resulting in a posterior distribution representing the updated knowledge. It is by virtue of this mechanism that the Bayesian framework has already gained popularity in diagnostic science [15].

In a Bayesian approach, population comparability can be defined using the concept of 'exchangeability', crucial in the context of introducing external evidence into data analysis. Spiegelhalter et al. [110] discuss a continuum of the relevance of historical information based on this concept. At one extreme of this continuum, complete independence between the historical studies and the current one may be assumed. At the other extreme, individual measurements can be considered exchangeable, i.e., sampling populations can be assumed equal. In the latter case, historical information could be included by considering the historical posterior information as the current prior, which is equivalent to pooling the current and historical data. Between the two extremes, there is a range of possible solutions. When the assumption of exchangeable observations is considered too strong, discounting methods based on the use of, e.g., power priors, could be considered to down-weigh historical evidence [45, 74]. If exchangeability can only be assumed at the level of the parameters, historical information can be included by considering the posterior predictive distribution as prior information [110]. By making one of the aforementioned assumptions, information from the development study can be effectively introduced as prior information into the validation study, with an appropriate weighting that depends on the degree of comparability between the respective populations.

The focus of the current chapter is two-fold. First, we develop a Bayesian

method which allows incorporating the development-study information about the accuracy of the diagnostic-biomarker index into the design and analysis of a prospective validation-study. Second, we investigate the potential gain regarding the design of the validation study, related to incorporating the development-study information.

5.2 Methodology

5.2.1 Development-study analysis

In the development study, the main goal is to select the appropriate optimal biomarker-combination that could serve as the index. We assume that we are interested in constructing a diagnostic-biomarker index which maximizes the AUC in the absence of a GS reference-test.

Consider data from N subjects, for which values of K biomarkers and the result of an imperfect diagnostic-test T are available. In this setting, we can apply the model proposed in Chapter 3.

The Bayesian model is fitted to the data by MCMC, implemented using the OpenBUGS [60] software. Annotated BUGS code can be found in Section A.1 of Appendix A. MCMC sampling results in empirical samples of the posterior distributions for all defined parameters and all of their functions. As a result, empirical posterior distributions for the optimal-combination coefficients contained in $\mathbf{a}$, are obtained. Additionally, we also obtain the empirical posterior distribution for AUC_a. This distribution already contains information about the performance of the to-be-validated diagnostic-biomarker index.

By using the aforementioned Bayesian approach, two valuable pieces of information become available after the development study. First, we obtain the estimates $\hat{\boldsymbol{a}}$ (e.g., as the posterior medians) of the coefficients of the linear combination of biomarkers maximizing the AUC. Second, an empirical posterior distribution for the value of the AUC of the optimal linear-combination, AUC_a, becomes available.

5.2.2 Validation-study analysis

The estimated optimal linear-combination coefficients are used in the prospective validation-study including n subjects. In particular, for the i-th subject from the validation data set, the value of the diagnostic index $y_{\mathbf{a},i}$ is obtained as follows: $y_{\mathbf{a},i} = \boldsymbol{y}_i^T \times \hat{\boldsymbol{a}}$, where $\boldsymbol{y}_i$ denotes a K-dimensional column-vector of biomarker values for the i-th subject. Denote by μ_0 and σ_0^2, respectively, the mean value and variance of the diagnostic index for controls. Similarly, let μ_1 and σ_1^2 denote, respectively, the mean

value and variance of the diagnostic index for cases. In addition, for every subject in the validation study also the result of the imperfect reference-test T, with Se_T and Sp_T, is available. Then, the full-data likelihood for the validation study is expressed as follows:

$$L\left(\mu_0, \mu_1, \sigma_0^2, \sigma_1^2, \theta, Se_T, Sp_T | \boldsymbol{y_a}, \boldsymbol{t}, \tilde{\boldsymbol{d}}\right) = \prod_{i=1}^{n} \left[Se_T^{t_i} \left(1 - Se_T\right)^{1-t_i} \frac{1}{\sqrt{2\pi\sigma_1^2}} exp\left\{ \frac{1}{2\sigma_1^2} \left(y_{\mathbf{a},i} - \mu_1\right)^2 \right\} \theta \right]^{\tilde{d}_i} \times \left[Sp_T^{1-t_i} \left(1 - Sp_T\right)^{t_i} \frac{1}{\sqrt{2\pi\sigma_0^2}} exp\left\{ \frac{1}{2\sigma_0^2} \left(y_{\mathbf{a},i} - \mu_0\right)^2 \right\} (1 - \theta) \right]^{1-\tilde{d}_i}, \quad (5.1)$$

where $\boldsymbol{y_a} = \left(y_{\mathbf{a},1}, \ldots, y_{\mathbf{a},n}\right)^T$, $\boldsymbol{t} = \left(t_1, \ldots, t_n\right)^T$ is the vector containing the imperfect reference-test results for all subjects, and $\tilde{\boldsymbol{d}} = \left(\tilde{d}_1, \ldots, \tilde{d}_n\right)^T$ is the vector containing the (latent) indicators of the true disease-status. Note that we make the conditional independence assumption, i.e., that the imperfect reference-test results $\boldsymbol{t}$ and the diagnostic biomarker-index observations $\boldsymbol{y_a}$ are independent, conditionally on latent true disease-status $\tilde{\boldsymbol{d}}$.

Furthermore, the likelihood is reparametrised by introducing parameter

$$\gamma = \sqrt{\frac{\left(\mu_1 - \mu_0\right)^2}{\sigma_0^2 + \sigma_1^2}}.$$

The AUC of the diagnostic index of the validation study, AUC_a^*, can then be defined as follows:

$$AUC_a^* = \Phi\left[\sqrt{\frac{\left(\mu_1 - \mu_0\right)^2}{\sigma_0^2 + \sigma_1^2}} \right] = \Phi\left(\gamma\right). \quad (5.2)$$

Using γ as a parameter allows introduction of the information about the AUC of the optimal linear-combination of biomarkers, obtained in the development study. Toward this aim, a prior distribution for γ has to be constructed. This is what we consider next.

5.2.3 Transfer from posterior to prior distribution for AUC_a^*

Translating the information about AUC_a obtained in the development study into a prior distribution for AUC_a^* in the validation study requires some consideration. AUC_a is a complex function of other parameters, as can be seen from (2.5). As

shown in (3.7), it is possible to express AUC_a in terms of a scaled difference between the mean biomarker-values of the cases and controls. A similar approach is applied in the validation study, as shown in (5.2). Upon comparing equations (3.8) and (5.2), one can conclude that the posterior distribution for $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$ from the development study can be used to construct a prior distribution for γ in the validation study. The form of the posterior distribution for $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$ is not obvious, as it depends on $\boldsymbol{\delta} = \boldsymbol{Q}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$. Given that $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$ only takes positive values on the real line, a log-normal distribution could be considered as a plausible approximation.

This suggestion was investigated by means of simulations. Data sets of small ($N = 150$) and moderate ($N = 400$) size with different underlying AUC_a values were simulated. In particular, AUC_a equal to 0.6, 0.75, and 0.9 was considered. These values span the parameter space of AUC_a, excluding the unlikely values of 0.5 and 1. On the AUC-scale, the posterior distributions for $AUC_a = 0.6$ and $AUC_a = 0.9$ are expected to be skewed due to the proximity to the parameter-space boundaries (0.5 and 1). On the other hand, the posterior distribution for the $AUC_a = 0.75$ data is more likely to be symmetrically shaped. The resulting posterior distributions of $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$ were approximated by a normal and log-normal distribution with mean and variance equal to the respective mean and variance of the obtained MCMC samples of $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$ and its logarithm. The left-hand-side column of Figure 5.2 shows the resulting $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$ posterior distributions with the normal and log-normal approximations for $N = 150$. The right-hand-side column of Figure 5.1 shows the results for $N = 400$.

Figure 5.1 shows that, in most cases, the difference between the histograms and the log-normal and normal approximations is very subtle. In the case of $AUC_a = 0.6$, the best approximation seems to depend on sample size. For $N = 400$ it appears that the normal approximation is slightly better in capturing the mode of the posterior distribution. For $N = 150$, the log-normal approximation seems to outperform the normal approximation. The empirical evidence from this simulation exercise indicates that the posterior distribution of $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$ can be approximated well by a normal distribution with mean and variance equal to the respective mean and variance of the MCMC samples of $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$ obtained in the development study if sample size is sufficient. When only limited data is available and AUC_a is small, a log-normal approximation could be considered. Transforming the $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$ posterior distributions and their approximations to the AUC-scale results in conclusions regarding the quality of approximation similar to those expressed above for $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$ (see Figure 5.2).

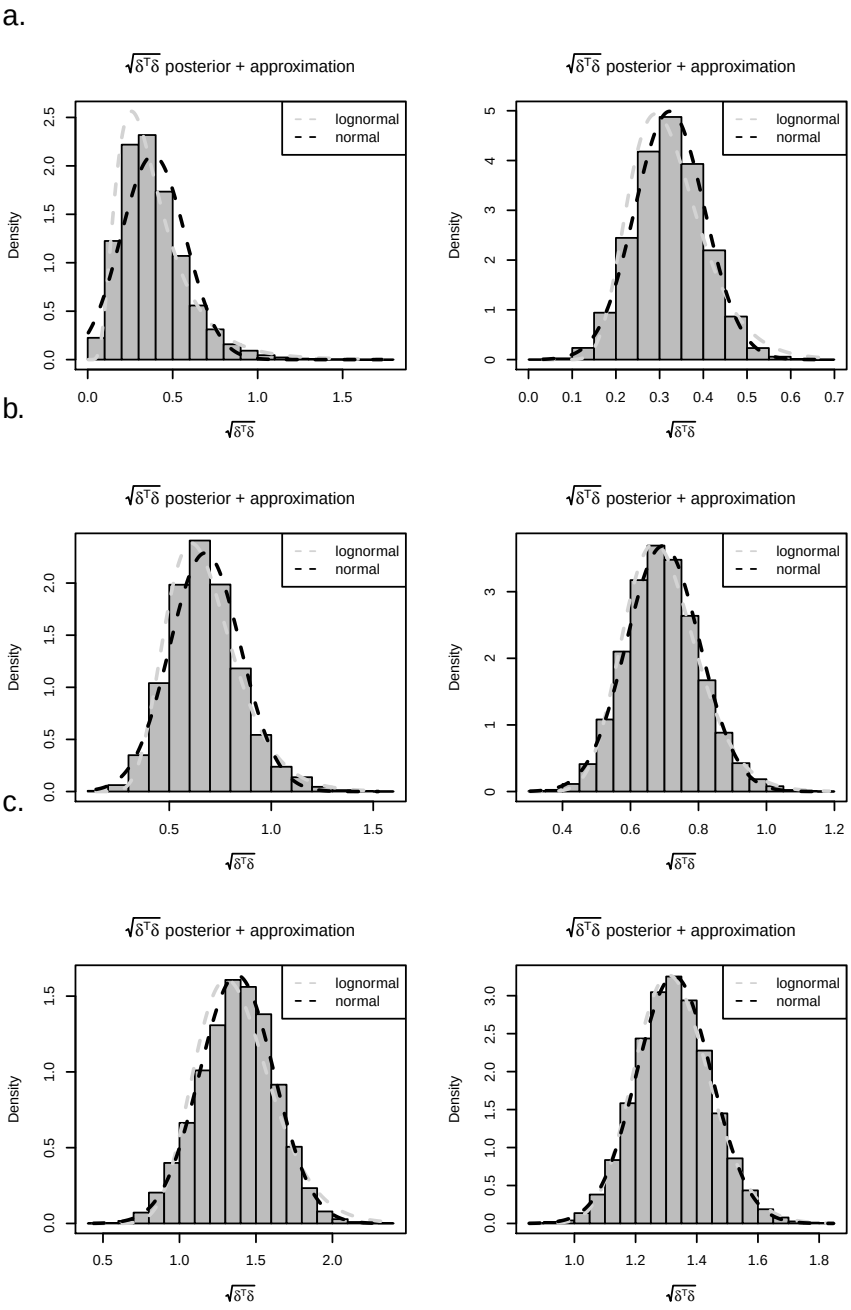

Figure 5.1: Empirical posterior distributions for $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$ resulting from data with different accuracy (rows) and sample sizes (columns). The dashed lines denote the log-normal (grey) and normal (black) approximation. **a**. Posterior of data with $AUC_{\mathbf{a}} = 0.6$. **b**. $AUC_{\mathbf{a}} = 0.75$ and **c**. $AUC_{\mathbf{a}} = 0.9$. The left hand side column shows results for $N = 150$, the right hand side column for $N = 400$.

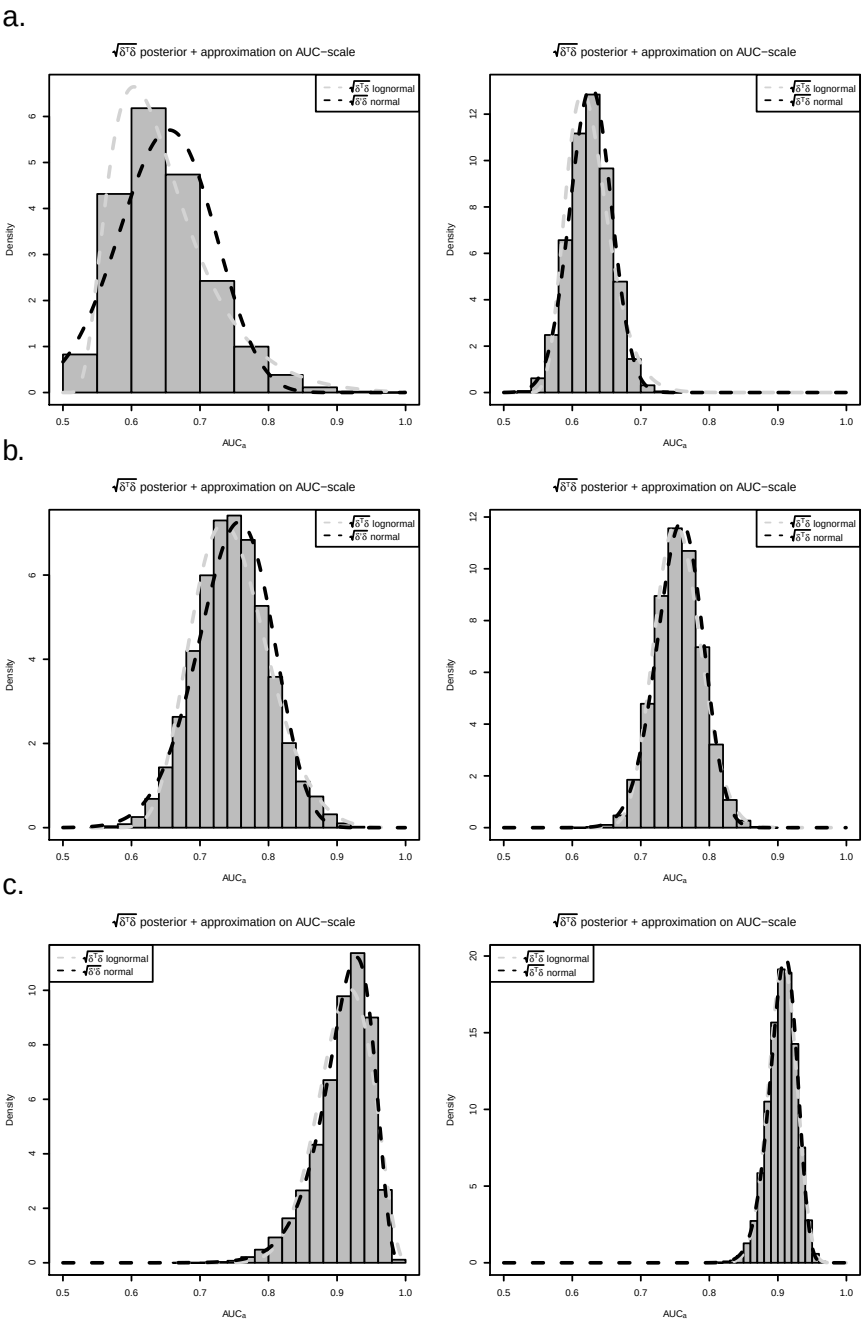

Figure 5.2: Empirical posterior distributions for $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$ resulting from data with different accuracy (rows) and sample sizes (columns) transformed to AUC_a-scale. The dashed lines denote the transformed log-normal (grey) and transformed normal (black) approximation. **a**. Posterior of data with $AUC_a = 0.6$. **b**. $AUC_a = 0.75$ and **c**. $AUC_a = 0.9$. The left hand side column shows results for $N = 150$, the right hand side column for $N = 400$.

If the normal approximation to the $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$ posterior from the development study is used, the following form of the prior distribution for AUC_a^* in the validation study can be proposed:

$$f_{AUC_a^*}(AUC_a) = f_y\left(\Phi^{-1}(AUC_a)\right) \times \left|\frac{1}{\phi\left(\Phi^{-1}(AUC_a)\right)}\right| \quad \text{for } AUC_a \in [0.5;1], \quad (5.3)$$

where $f_y(\cdot)$ is a normal distribution with parameters μ and σ equal to the posterior mean and standard deviation of $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$, $\Phi^{-1}(\cdot)$ denotes the probit function, and $\phi(\cdot)$ is the density of the standard-normal distribution. In case we want to ignore the development study information, it is clear from (5.3) that assuming a standard-normal distribution for γ leads to a constant function for AUC_a^*, i.e., to a flat prior distribution for AUC_a^*.

As summarized in Table 5.1, the prior distributions for the remaining parameters introduced in (5.1) have the same form as for the development study (see Section 3.3 of Chapter 3 for the argumentation about the choice of these priors).

Table 5.1: Prior distributions for all parameters in the model for the validation study. $\hat{x}_{\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}|\boldsymbol{Y}}$ and $s^2_{\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}|\boldsymbol{Y}}$ denote the empirical mean and variance of the posterior distributions of $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$ coming from the development study.

Parameter	**Prior distribution**
Prevalence	
θ	$U(0.1, 0.9)$
Parameters of the dichotomous reference test	
Se_T	$Beta(a,b)\,trunc(0.5,1)$
Sp_T	$Beta(c,d)\,trunc(0.5,1)$
Mean of biomarker values control group	
$\boldsymbol{\mu}_0$	$N_3\left(\mathbf{0}, \mathbf{I}_3 10^6\right)$
Scaled difference biomarker distribution means	
$\boldsymbol{\gamma}$	$N\left(\hat{x}_{\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}|\boldsymbol{Y}}, s^2_{\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}|\boldsymbol{Y}}\right)$
Biomarker-distribution standard deviations	
$\sigma_{\tilde{d}}$	$U(0, 1000)$

In order to obtain posterior results for the considered parameters, the proposed model is fitted by MCMC, as implemented in the OpenBUGS software. The annotated BUGS code is provided in Section A.3 of Appendix A.

5.2.4 Validation criterion

By combining likelihood (5.1) with the proposed prior distributions (Table 5.1), a posterior distribution for AUC_a^* is obtained. The posterior can be used to decide whether the diagnostic-biomarker index can be considered validated or not. Toward this end, a validation criterion is required. We consider a Bayesian hypothesis testing rule as the validation criterion [110]. In particular, let us consider the following null- and alternative-hypotheses:

$$H_0 : AUC \leq \tau,$$
$$H_1 : AUC > \tau,$$

where τ is a fixed value, possibly depending on the biomarkers and application of interest. If, for example, current diagnostic tests are already very accurate, a larger validation criterion would be selected than when the current tests are only of moderate accuracy. Based on the posterior distribution of AUC_a^*, we compute the probability of H_0. If this probability is smaller than α, say, we consider the diagnostic index as validated.

5.3 Simulation study

To compare the effect of including information about the accuracy from the development study, as opposed to ignoring this information, we performed a simulation study. In this simulation study, we evaluated the power to reject the proposed null hypothesis, i.e., to satisfy the validation criterion, for a series of sample sizes. Power was defined as the proportion of simulated data sets for which the null hypothesis was rejected. For each sample size, the power was investigated for a study using the informative prior distribution for AUC_a^* based on the development study, and for a study using the flat prior for AUC_a^* resulting from a standard-normal distribution for γ.

We simulated data for three correlated biomarkers, which followed a different normal distribution for cases and controls (for a summary of all underlying parameters, see Table 5.2). For each of the biomarkers, the individual diagnostic-performance was described by $AUC = 0.75$. The true underlying optimal linear-combination of the biomarkers, defined by the vector of coefficients $\boldsymbol{a} = (0.1594, 0.2237, 0.0972)^T$, yields $AUC_a = 0.7872$. In addition, we assumed that an imperfect reference-test was available with sensitivity and specificity of 0.85. Furthermore, the prevalence of cases in the sample, θ, was considered to be equal to 0.5.

Table 5.2: Parameter values underlying the sampled simulation data sets.

Parameter	Value
Multivariate parameters	
$\boldsymbol{\mu}_0$	$(0, 0, 0)^T$
$\boldsymbol{\mu}_1$	$(1.1683, 1.3490, 1.5082)^T$
Dependent biomarkers $\left(\boldsymbol{\rho} = (0.5, 0.9, 0.5)^T\right)$	
$\boldsymbol{\Sigma}_0$	$\begin{pmatrix} 1 & 0.5 \times \sqrt{1 \times 3} & 0.9 \times \sqrt{1 \times 2} \\ 0.5 \times \sqrt{1 \times 3} & 3 & 0.5 \times \sqrt{3 \times 2} \\ 0.9 \times \sqrt{1 \times 2} & 0.5 \times \sqrt{3 \times 2} & 2 \end{pmatrix}$
$\boldsymbol{\Sigma}_1$	$\begin{pmatrix} 2 & 0.5 \times \sqrt{2 \times 1} & 0.9 \times \sqrt{2 \times 3} \\ 0.5 \times \sqrt{2 \times 1} & 1 & 0.5 \times \sqrt{1 \times 3} \\ 0.9 \times \sqrt{2 \times 3} & 0.5 \times \sqrt{1 \times 3} & 3 \end{pmatrix}$
Parameters of the reference test	
Se_T	0.85
Sp_T	0.85
Prevalence	
θ	0.5
Functions of multivariate parameters	
AUC_1	0.75
AUC_2	0.75
AUC_3	0.75
$\boldsymbol{a}$	$(0.1594, 0.2237, 0.0972)^T$
Optimal combination parameters	
$\mu_{a,0}$	0
$\mu_{a,1}$	0.6347
$\sigma^2_{a,0}$	0.3490
$\sigma^2_{a,1}$	0.2857
AUC_a	0.7872

We assumed that, prior to the validation study, results from a single development study consisting of 400 observations were available. In particular, the results from the model developed in Chapter 3, with the 'optimistic' AUC_a-prior and flat Se_T/Sp_T prior distributions with the Se_T and $Sp_T > 0.5$ restriction, was considered. Due to sampling variability, the realizations of a single development study can lead to different posterior distributions for AUC_a which can influence validation results. To acknowledge this, three development studies were selected based on their sampling probability. In particular, 200 development studies were sampled and ordered by their posterior median estimate of AUC_a. Subsequently, the development studies

corresponding to the 2.5-, 50-, and 97.5-percentile were selected. The 2.5- and 97.5-percentile development studies led to posterior distributions that underestimated and overestimated, respectively, the AUC_a value with the posterior median AUC_a equal to 0.7296 and 0.8469, respectively (see the histograms in Figure 5.3). The 50-percentile development study yielded an unbiased posterior median AUC_a estimate equal to 0.7873. For each of these development studies, the estimate of the optimal linear-combination coefficients vector $\hat{\boldsymbol{a}}$ (see Table 5.3) and the posterior $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$ distribution were retained. Because of the moderate size of the development studies ($N = 400$), the normal approximation to the $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$-posterior distribution was considered.

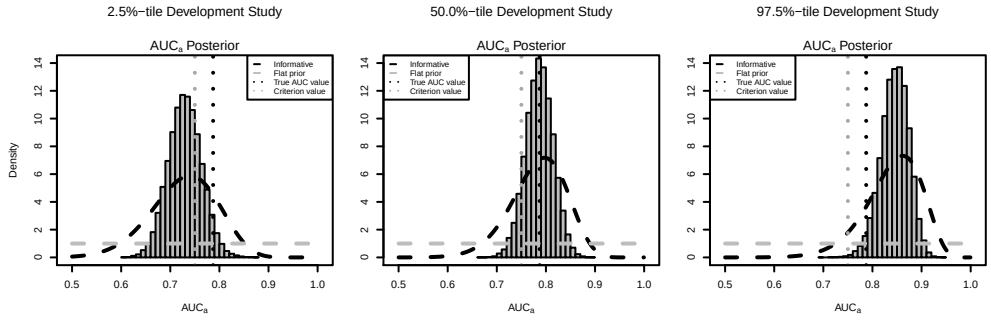

Figure 5.3: Posterior distributions (histograms) for the three selected development studies, providing underestimated (**a**.), unbiased (**b**.), and overestimated (**c**.) value of AUC_a. The dashed black line presents the AUC_a^*-prior corresponding to the (discounted by 200%) normal approximation to the posterior $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$ distribution. The dashed grey line denotes the flat AUC_a^*-prior corresponding to the standard-normal prior for γ. Validation criterion (τ) and true underlying AUC_a^* are indicated by the dotted grey and black vertical line, respectively.

Under the assumption that AUC_a and AUC_a^* are exchangeable, but that the individual observations from the development and validation study are not, the posterior predictive-distribution for $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$ would have been the preferred prior for γ. As this distribution is not obtainable when only one development study is available, we discounted the posterior distribution of $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$ to account for sampling variability. In particular, the standard deviations of the approximating normal distributions for the three simulated development-studies were multiplied by two. The resulting discounted informative prior-distributions for AUC_a are shown, together with the flat prior, in Figure 5.3.

To calculate the power of the validation study, 200 data sets were simulated with sample sizes equal to 100, 400, 600, and 800. In each data set, the simulated biomarker values for each subject were transformed into the diagnostic index by using the estimates of $\boldsymbol{a}$ obtained in each of the three selected development studies (Table 5.3).

In this way, for each development study and sample size combination, 200 validation data sets with 'univariate' biomarker-based index were obtained.

Table 5.3: Estimates (based on posterior medians) of the optimal linear-coefficient vector $\hat{\boldsymbol{a}}$ for the three considered development data studies and their corresponding true underlying values.

	Development data set			
$\hat{\boldsymbol{a}}$	**2.5%-tile**	**50%-tile**	**97.5%-tile**	**True**
$\hat{\boldsymbol{a}}_{\mathbf{1}}$	0.0610	0.0985	0.0200	0.1594
$\hat{\boldsymbol{a}}_{\mathbf{2}}$	0.1673	0.2101	0.2557	0.2237
$\hat{\boldsymbol{a}}_{\mathbf{3}}$	0.0966	0.1271	0.3046	0.0972

For the analysis of the validation data sets, the prior distributions were defined as indicated in Table 5.1. In particular, for μ_0, a flat normal prior $N(0, 10^6)$ was considered. For σ_0 and σ_1, uniform prior distributions defined between 0 and 1000 were used. As in the development study, sensitivity and specificity of the imperfect reference-test were given $Beta(1, 1)$ priors, left truncated to be larger than 0.5 in order to ensure identifiability. Finally, a uniform distribution between 0.1 and 0.9 was assumed as the prior distribution for θ. Avoiding the extremes of the parameter space for θ alleviates convergence problems with the MCMC algorithms deployed to fit the model.

To compare the power estimates for the proposed design with the power of the 'traditional' frequentist design, we investigated whether the well-known fact that Bayesian analysis is asymptotically identical to frequentist analysis if a flat prior is assumed applies to the case of the moderate data set sizes considered in this chapter. To this aim, we considered data with the true disease status information. In particular, the Bayesian power simulation with a flat AUC_a^*-prior was compared to the frequentist power simulation, as well as to the results of a frequentist power calculation [133]. The underlying true settings were as before (Table 5.2) and the diagnostic index was constructed based on $\hat{\boldsymbol{a}}$ of the development study that yielded the unbiased estimate of $\boldsymbol{a}$ (see Table 5.3).

The frequentist power simulation was performed by bootstrapping the 60%-confidence interval (note that the significance level (α) of 20% was assumed) around the point estimate of AUC_a^* under the binormal assumption [81]. In case the lower limit of this interval exceeded 0.75, H_0 was considered to be rejected and the diagnostic-biomarker index was considered validated. This type of hypothesis testing based on the 60%-confidence interval is equivalent to a one-sided hypothesis test with the type-I error probability equal to 0.2. The proportion of the cases when H_0 was

rejected provided an estimate of the frequentist power.

For the frequentist power calculation, the following general equation was considered:

$$1-\beta=\Phi^{-1}\left\{\frac{\sqrt{n_c\left(\theta_0-\theta_1\right)^2}-z_\alpha\sqrt{V_0\left(\hat{\theta}\right)}}{\sqrt{V_1\left(\hat{\theta}\right)}}\right\}.$$

In the equation, (1-β) is the desired power, $\Phi^{-1}(\cdot)$ is the probit function, θ_0 and θ_1 represent the conjectured AUC_a^* under H_0 and H_1, respectively, n_c denotes the considered number of cases, and z_α indicates the upper α-quantile of a standard-normal distribution. The number of controls is accounted for by the definition of the variance functions, depending on the ratio of controls versus cases. $V_0\left(\hat{\theta}\right)$ and $V_1\left(\hat{\theta}\right)$ represent the variance function of AUC_a^* under H_0 and H_1, respectively. Estimates of $V_0\left(\hat{\theta}\right)$ and $V_1\left(\hat{\theta}\right)$ are expressed as indicated in Obuchowski et al. [75], under the assumption that the data are observed on a truly continuous scale and follow an underlying binormal distribution:

$$\hat{V}\left(\hat{\theta}\right)=\left[0.0099\times exp\left(-\frac{a^2}{2}\right)\right]\times\left[\left(5a^2+8\right)+\frac{\left(a^2+8\right)}{R}\right]-\left[0.0398\times a^2 exp\left(-\frac{a^2}{2}\right)\right],$$

where $a=\Phi^{-1}\left(\hat{\theta}\right)\times 1.414$, and R denotes the ratio of patients without the disease to those with the disease. In the considered example with prevalence equal to 0.5, $R=1$.

As for the validation criterion, we considered $\tau=0.75$. This can be interpreted as considering a hypothesis test to investigate whether the biomarker index is significantly more accurate than any of the biomarkers alone. For the proposed model, taking development study accuracy information into account, α was set equal to 0.2. In other words, we regarded the diagnostic-biomarker index to be validated if the posterior probability that $AUC\leq 0.75$ was smaller than 0.2 (see the illustration in Figure 5.4). In order to ensure valid power comparisons, type-I error characteristics for both the proposed and the 'traditional' prior setting were investigated by considering results under the null hypothesis (see Appendix D). Figure D.1 in Appendix D shows that for $\alpha=0.2$, a significant difference in type-I error between the two prior settings is observed. Therefore, α was lowered to 0.06 and 0.15 for the 'traditional' prior setting considering the 2.5- and 50-percentile development studies, respectively. For the 97.5-percentile development study, α was increased to 0.4. Adjusting α results

in type-I errors which are statistically non-significant for all considered sample sizes (Figure D.2 in Appendix D), with exception of the N=100 case for the 97.5-percentile development study setting.

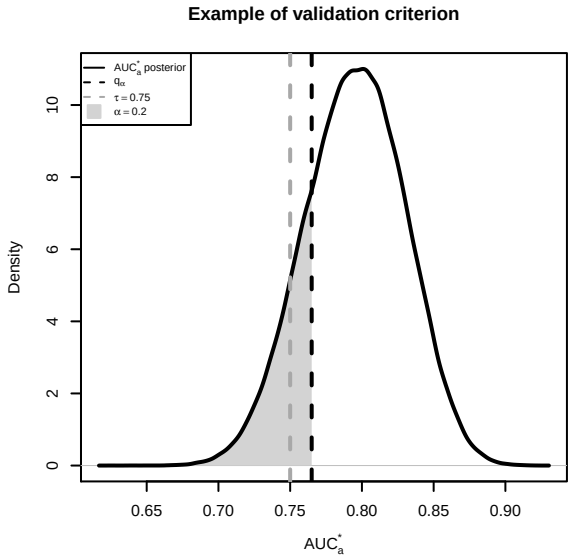

Figure 5.4: Example of a validated diagnostic-biomarker index for $\tau = 0.75$ and $\alpha = 0.2$. The posterior distribution of AUC_a^* is denoted by the solid black line. The dashed grey line indicates the validation criterion τ and the dashed black line indicates the observed α-quantile of the posterior AUC_a^* distribution.

According to the discounted prior-distributions shown in Figure 5.3, the prior probability that $AUC \leq 0.75$ considering the proposed model, taking the development study accuracy information into account, is equal to 0.62, 0.26, and 0.007 for the underestimating, unbiased, and overestimating development-study, respectively. Since the prior for the 'traditional' setting is a flat prior between 0.5 and 1, the prior probability that $AUC \leq 0.75$ in this case, is equal to 0.5.

Thus, in total, each of the 200 validation-study data sets was analysed seven times: ignoring the prior information about AUC_a^*, but including the true disease status (to compare the flat-prior Bayesian and frequentist results); ignoring the prior information about AUC_a^*, while including the imperfect reference-test result (for each of the three selected development studies, to obtain power estimates for the 'traditional' study design); and including the prior information about AUC_a^* and the imperfect reference-test result (for each of the three selected development studies, to obtain

power estimates for the proposed validation-study design).

The models were fitted using OpenBUGS 3.2.1 [60]. Annotated BUGS-model code can be found in Section A.3 of Appendix A. Results were analyzed and summarized using R 3.0.1 (x64) [90]. The R-package R2OpenBUGS [111] was used as an interface between R 2.14.2 and OpenBUGS. For all data fits, five chains were considered, with 10,000 iterations retained after a 10,000 iteration burn-in. For the development study, fitting time was approximately 8 hours. Fitting times for the validation studies depended on the sample size and were equal to, approximately, seven, 30, 35, and 45 minutes for sample sizes of 100, 400, 600, and 800, respectively, on a 64-bit, 2.6 GHz, 8GB RAM machine.

5.4 Results

Table 5.4 presents the results for the frequentist power obtained from calculations ('Freq Calc'), simulations ('Freq Sim'), and simulations for the Bayesian analysis with a flat AUC_a^*-prior ('Freq'). There were no convergence problems when fitting the Bayesian model in any of the scenarios. Table 5.4 does not indicate any statistically significant differences between the obtained power estimates. Thus, it can be concluded that, for the considered sample sizes, the frequentist approach and the Bayesian flat-prior analysis provide equivalent results.

Table 5.4: Results of the comparison between the frequentist and Bayesian flat-prior power simulations/calculations. Proportions (standard errors) of validated diagnostic biomarker-indices are given for all considered validation study sample sizes for the frequentist calculation ('Freq Calc'), simulation ('Freq Sim'), and the Bayesian flat prior simulation ('Freq').

	Validation study sample size			
Model	**n=100**	**n=400**	**n=600**	**n=800**
Freq Calc	0.48	0.79	0.88	0.93
Freq Sim	0.50 (0.04)	0.77 (0.03)	0.89 (0.02)	0.96 (0.01)
Freq	0.38 (0.04)	0.74 (0.03)	0.87 (0.02)	0.94 (0.02)

The results of the simulation of the validation-study power are presented in Figure 5.5. For each of the three considered development data sets (panels a – c), the empirical power is presented as a function of the validation study size for the case when the informative AUC_a^*-prior (shown in Figure 5.3) is used, as well as when the flat AUC_a^*-prior (corresponding to the 'traditional' analysis) is applied. Note that, in all considered cases, not more than 1% of non-convergence was observed (data not shown).

Overall, power tends to increase with an increasing validation-study sample size. However, when the prior information is too optimistic, as it is the case of the development study that overestimates the AUC_a^* value, the power decreases with an increasing validation-study sample size (panel *c* of Figure 5.5). This is understandable, because substantial validation data is needed to 'correct' the information provided by the severely-biased prior distribution.

As compared to the 'traditional' flat-prior approach, including the prior information from the development study increases the power, as long as the prior information does not (severely) underestimate the true diagnostic accuracy (panel *a* of Figure 5.5). Note that the situation depicted in panel *a* of Figure 5.5 is observed for a development study corresponding to the 2.5-percentile of the distribution of the posterior median estimates of AUC_a. Thus, it is an unlikely situation. In our setting, when using the prior obtained from a development study that correctly estimates the AUC_a, approximately 100 subjects would be required in the validation study to obtain a power of 0.53 (panel *b* of Figure 5.5). In a 'traditional' flat-prior study, reaching the same power would require 600 subjects.

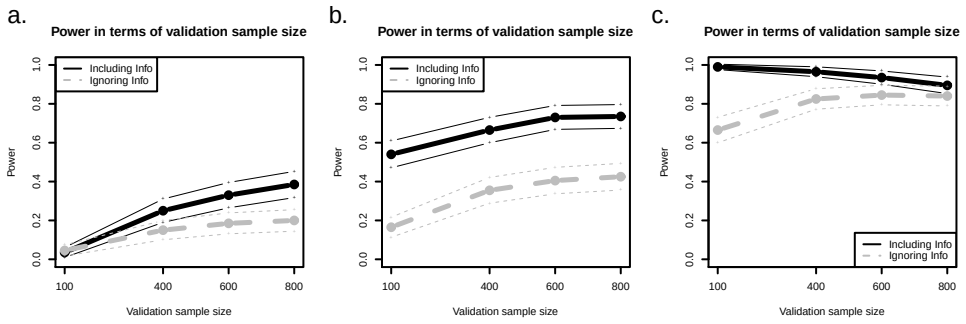

Figure 5.5: Empirical power as a function of the validation study sample size for the underestimated (**a**.), unbiased (**b**.), and overestimated (**c**.) development study information. Results are indicated for the flat (grey circles connected by the dashed line) and informative (black circles connected by the solid line) AUC_a^*-priors. The 95% confidence interval for each estimate is indicated by the respective plus-signs.

5.5 Conclusions

We have proposed a novel Bayesian latent-class mixture model to validate a diagnostic index based on a combination of correlated biomarkers. The model allows incorporation of prior information about the accuracy of the index. Toward this aim, the posterior information about $\sqrt{\boldsymbol{\delta}^T\boldsymbol{\delta}}$ from the development study of a moderate sample

size can be approximated by a normal distribution and used as a prior distribution for γ in the validation study. Based on the exchangeability principle, the normal approximation can be used to specify a range of priors, from a completely flat one to the one matching the posterior distribution from the development study. Depending on the particular implementation and comparability of the development and validation study sampling populations, a prior between these two exchangeability extremes can be selected [110, 17, 56].

In the presented simulation study, it was assumed that a validation study will be performed after results from a single development study were obtained. To investigate the range of expected results, the posterior information from the development studies corresponding to the 2.5-, 50-, and 97.5-percentiles of the posterior median AUC_a range was considered for the use as prior information. To account for sampling variability in a single application, the standard deviation of the approximated posterior distribution was discarded by a factor of two. Results show that using the prior information derived from a development study can lead to substantial gains in efficiency. In particular, when the 50-percentile development study information was included in the validation study, rejecting H_0 at a similar significance level with power of about 0.53 required about 20% of the validation sample size needed for a 'traditional' study, i.e., when ignoring this information. Overall, efficiency was gained in all settings, except when the underestimating development study was considered in combination with validation study sample sizes smaller than 400 observations.

Chapter 6

Transferring cut-off values between assays for Alzheimer's disease CSF-biomarkers

In the following chapter, a novel two-stage Bayesian model to transfer AD CSF-biomarkers cut-off values from a current to a new assay is proposed. The problem setting is discussed in Section 6.1. The currently applied linear-regression-based cut-off transfer method and novel two-stage Bayesian model are discussed and developed, respectively, in Section 6.2. To investigate model performance and its applicability to real data sets, the proposed model is applied to simulated and two real Alzheimer's disease CSF-biomarker data sets. A description of the simulation study can be found in Section 6.3, while the real data set application is discussed in Section 6.4. Results from both the simulation study and the real data applications can be found in Section 6.5. Finally, concluding remarks can be found in Section 6.6.

6.1 Problem setting

Over the past two decades, numerous studies have assessed potential applications of cerebrospinal fluid (CSF) biomarkers in the field of Alzheimer's disease (AD). There is a general agreement across studies that an initial decrease in CSF-Aβ_{1-42}, followed

by an increase in total tau and/or phosphorylated tau, is a reflection of ongoing neuropathology (amyloidopathy, tauopathy) of the AD type in the brain of affected subjects [47, 79]. In addition, CSF biomarker analysis was integrated in (research) criteria for diagnosis of AD [66, 28]. Position Emission Tomography (PET) imaging has been approved by the Food and Drug Administration (FDA) to identify subjects with ongoing amyloidopathy [129]. The European Medicine Agency (EMA) qualified the combination of CSF-Aβ_{1-42} and total tau for use as a tool for patient stratification and patient enrichment in clinical trials [68].

However, these guidance documents do not specify the use of a specific assay or technology, nor do they provide advice for the manufacturer on acceptance criteria for analytical and diagnostic-performance requirements. A laboratory that desires integration of AD biomarker quantification in its portfolio has the choice among several commercially available assays, which differ with respect to their world-wide availability, ease-of-use, technology, design, critical raw materials (antibodies, calibrators), regulatory status, as well as the level of validation [3]. At present, there is no reference standard or reference method that is fully representative for the endogenous peptide [73] and the same assay can generate different values across different laboratories [119, 115, 12, 108]. Efforts to improve the assays' between-lab variability are already on-going [119, 64, 21].

The use of a biomarker assay for patient classification implies the need to establish a cut-off value to assign patients to the desired categories (e.g. No-AD/AD). At present, these cut-offs cannot be derived universally and each laboratory has to establish its own cut-off for the assay of choice [51, 13]. These efforts are time-consuming and require well-characterized samples, preferably from subjects with autopsy-confirmed AD. It has been shown that cut-off values derived by using the clinical diagnosis as the reference test lead to a shift in the cut-off value as compared to using the autopsy confirmed diagnosis [116], with suboptimal sensitivity and specificity as a result. However, well-characterized samples like those obtained from clinical trials or world-wide consortia, are not widely available in quantities sufficient for repeated testing. Therefore, the data sets used to derive a cut-off value are often small with a total of 100 to 200 measurements [107, 118, 72, 80, 43], resulting in cut-off values with high uncertainty. These precision aspects are typically ignored in practice [9] and, to our knowledge, have not been reported for AD CSF-biomarkers. When a laboratory wants to convert a test procedure for a specific analyte to a newer assay, the cut-off value has to be derived for this new assay. The best way to do this is by testing samples of well-diagnosed subjects. In the absence of these samples, however, it is a common practice to test available samples 'side-by-side' with the current and

new assay and to transfer the measurements and/or the cut-off value of the current assay to the new assay by means of a linear-regression formula [42, 65, 46, 124, 49]. To our knowledge, the validity and the effect of this cut-off transfer method on the clinical performance of the biomarker measured with the new assay has not been studied.

In this chapter, we study the properties of the linear-regression-based method of transferring the cut-off value of a current assay to a new assay. Moreover, we compare it to a novel Bayesian method that we have developed. Toward this aim, we undertake a simulation study and apply the methods to two sets of data with Aβ_{1-42} measurements from the BIODEM lab of the University of Antwerp. In the process, we also evaluate the precision of the obtained cut-off estimates as a function of the size of data sets.

6.2 Methods

In the next section the currently applied linear-regression-based cut-off transfer method is discussed and a novel two-stage Bayesian method is proposed. As a starting point we first discuss the assumed data structure, assumptions, and notation used in the remainder of this chapter.

6.2.1 Data structure, assumptions, and notation

Consider two assays that measure the same biomarker (Y), a current one (which generates data indicated by Y_c) and a new one (producing Y_n). Moreover, assume that, conditional on true disease status D, these biomarker values are distributed according to a bivariate normal distribution:

$$\begin{pmatrix} Y_c \\ Y_n \end{pmatrix} | D = d \sim N_2 \left(\begin{pmatrix} \mu_{c,d} \\ \mu_{n,d} \end{pmatrix}, \begin{pmatrix} \sigma^2_{c,d} & \rho_d \sigma_{c,d} \sigma_{n,d} \\ \rho_d \sigma_{n,d} \sigma_{c,d} & \sigma^2_{n,d} \end{pmatrix} \right),$$

where $\mu_{c,d}$ and $\mu_{n,d}$ are the mean of the biomarker distribution in disease group d ($d = 0$ for controls; $d = 1$ for cases) for the current and new assay, respectively. More specifically, we assume that $\mu_{j,1} \geq \mu_{j,0}$ for $j \in \{n, c\}$. The variances of the respective current- and new-assay biomarker distributions for group d are denoted by $\sigma^2_{c,d}$ and $\sigma^2_{n,d}$. Finally, the true disease-status-dependent correlation between both assays is denoted by ρ_d.

Based on the assumption of normally-distributed biomarker data and a particular order of the biomarker-distribution means with respect to true disease status D, the

assay-specific optimal cut-offs maximizing the Youden-index (indexed by $j \in \{c, n\}$) can be defined by [103]

$$c_j = \frac{\left(\mu_{j,1}\sigma^2_{j,0} - \mu_{j,0}\sigma^2_{j,1}\right) - \sigma_{j,0}\sigma_{j,1}\sqrt{(\mu_{j,0} - \mu_{j,1})^2 + (\sigma_{j,0} - \sigma_{j,1})\, ln\left(\sigma^2_{j,0}/\sigma^2_{j,1}\right)}}{\left(\sigma^2_{j,0} - \sigma^2_{j,1}\right)}. \quad (6.1)$$

Assuming equal variances [132], (6.1) simplifies to:

$$c_j = \frac{\mu_{j,0} + \mu_{j,1}}{2}. \quad (6.2)$$

In general, we assume that two sets of data are available. One data set contains measurements only from the current assay on which the cut-off, c_c, for the current assay is estimated. This data set will be referred to as the *current-assay cut-off* data set. For this data set assume that GS reference-test information is available. The second data set, referred to as the *new-assay-transfer* data set contains data from samples measured side-by-side on both assays, but for which GS reference-test information is lacking.

In the next section, the overall strategy is to combine the information about c_c from the current-assay cut-off data set, with the information contained in the new-assay data set to come up with an estimate for c_n. In general, a linear-regression-based transfer method (see Section 6.2.2) is applied. In order to avoid biased estimates, we propose a two-stage Bayesian latent-class model as developed in Section 6.2.3.

6.2.2 Linear-regression-based cut-off transfer

Current-assay cut-off estimation

A variety of ROC curve estimation and cut-off selection methods exist [132] when the information about the disease-status of the subjects is available. It is beyond the scope of this dissertation to compare the performance of all methodologies. To come up with an estimate of $\hat{c}_c$ from the current-assay cut-off data set, we focused on the fully non-parametric (empirical) direct estimation method, in which the cut-off value is obtained by selecting the (observed) biomarker value with the highest Youden-index [86, 130]. This approach is often used to establish cut-off values for AD biomarkers [107, 118, 72, 80]. It is appropriate because sensitivity and specificity are deemed of equal importance in AD diagnosis [13]. Hence, the prevalence of AD and the relative cost of a false-negative classification, as compared to a false-positive classification, are not included in the selection of an 'optimal' cut-off [9, 42, 86, 33]. The standard error

(SE) and 95% confidence interval (CI) of the estimated cut-off value can be estimated by bootstrapping.

New-assay cut-off estimation

If data containing biomarker measurements from the new-assay were available, together with results from a GS reference-test, the new-assay cut-off could be estimated applying the methods described in the previous paragraph. However, usually no GS reference-test information is available in the new-assay-transfer data set, but measurements on both assays are available. In this case, one could think of using the information about the linear relationship between the current and new assay to translate $\hat{c}_c$ to the new assay as an estimate for $\hat{c}_n$. This is exactly what is done in the linear-regression-based cut-off transfer method. An estimate of the relation between the current- and new-assay is obtained by fitting the following linear-regression model:

$$Y_{n,i} = \beta_0 + \beta_1 Y_{c,i} + \epsilon_i \quad \text{with } \epsilon_i \sim N\left(0, \sigma^2\right), \tag{6.3}$$

where $(Y_{c,i}, Y_{n,i})$ is the pair of measurements for subject i obtained with the current (c) and new (n) assay, contained in the new-assay-transfer data set. The cut-off value of the new assay $(\hat{c}_n)$ is then obtained by transforming the cut-off value of the current assay $(\hat{c}_c)$ as follows:

$$\hat{c}_n = \hat{\beta}_0 + \hat{\beta}_1 \hat{c}_c, \tag{6.4}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated coefficients of linear-regression model (6.3).

The SE of the new-assay cut-off value can be estimated with the delta method [98], taking into account the SE of the current-assay cut-off, as well as the uncertainty about the estimated regression parameters.

Note that the regression model (6.3) is fitted to the entire data set, which contains a mixture of control and AD subjects. Thus, the method assumes that the same linear relationship holds in the control and AD populations. If this is not the case, however, model (6.3) is wrong. This can be shown analytically with the following simplified example.

Assume that assay-specific variances are equal, irrespective of true disease status D, for assay $j \in \{c, n\}$:

$$\sigma_{j,0} = \sigma_{j,1} \equiv \sigma_j. \tag{6.5}$$

Then the cut-off of assay $j \in \{c, n\}$, c_j, is defined by (6.2) and we can define the

sensitivity (Se_j) and specificity (Sp_j) of c_j as

$$Se_j = 1 - \Phi\{(c_j - \mu_{j,1})/\sigma_j\},$$

$$Sp_j = \Phi\{(c_j - \mu_{j,0})/\sigma_j\},$$

such that

$$Se_j = Sp_j \equiv \Phi\{(\mu_{j,1} - \mu_{j,0})/2\sigma_j\},$$

where Se_j and Sp_j are sensitivity and specificity of assay $j \in \{c, n\}$, respectively, and $\Phi(\cdot)$ is the cumulative distribution function of the standard-normal distribution.

Under (6.5), we get that Y_n, conditionally on $Y_c = y$ and $D = d$, follows a normal distribution with mean $\mu_{n,d} - \rho_d \frac{\sigma_n}{\sigma_c}\mu_{c,d} + \rho_d \frac{\sigma_n}{\sigma_c} y$ and variance $\sigma_n^2(1 - \rho_d^2)$. Note that $\mu_{n,d} - \rho_d \frac{\sigma_n}{\sigma_c}\mu_{c,d}$ is the intercept and $\rho_d \frac{\sigma_n}{\sigma_c}$ is the slope of the regression of Y_n on Y_c.

Consider a sample, for which $P(D = 1) \equiv \pi$. If we estimate the regression of Y_n on Y_c without the information about the disease status of the subjects, we will be estimating the following regression line:

$$\pi \left(\mu_{n,1} - \rho_1 \frac{\sigma_n}{\sigma_c}\mu_{c,1} + \rho_1 \frac{\sigma_n}{\sigma_c} y \right) + (1 - \pi) \left(\mu_{n,0} - \rho_0 \frac{\sigma_n}{\sigma_c}\mu_{c,0} + \rho_0 \frac{\sigma_n}{\sigma_c} y \right). \tag{6.6}$$

That is, we will be estimating a mixture of the regression lines for cases ($D = 1$) and controls ($D = 0$).

Now, assume that we will compute the cut-off for the novel assay, c_n^*, say, by applying the regression equation (6.6) to the cut-off c_c, given in (6.2). We then obtain:

$$\begin{aligned} c_n^* &= \pi \left(\mu_{n,1} - \rho_1 \frac{\sigma_n}{\sigma_c}\mu_{c,1} + \rho_1 \frac{\sigma_n}{\sigma_c} \frac{\mu_{c,1} + \mu_{c,0}}{2} \right) \\ &\qquad + (1 - \pi) \left(\mu_{n,0} - \rho_0 \frac{\sigma_n}{\sigma_c}\mu_{c,0} + \rho_0 \frac{\sigma_n}{\sigma_c} \frac{\mu_{c,1} + \mu_{c,0}}{2} \right) \\ &= \mu_{n,0} + \pi(\mu_{n,1} - \mu_{n,0}) + \sigma_n \frac{\mu_{c,1} - \mu_{c,0}}{2\sigma_c} \{(1 - \pi)\rho_0 - \pi\rho_1\}. \end{aligned} \tag{6.7}$$

In general, c_n^*, defined in (6.7), is not equal to c_n, given in (6.2). In other words, transfer of the cut-off for the current assay (c_c) to a cut-off for the novel assay by using a regression fitted to the mixed sample without using the disease-labels will, in general, not result in the correct cut-off (c_n), which provides the maximum Youden-index.

Assume that $\rho_0 = \rho_1 \equiv \rho$, i.e., that the slopes of the regression lines for the $D = 0$

and $D = 1$ sub-samples are equal. Then

$$c_n^* = \mu_{n,0} + \pi(\mu_{n,1} - \mu_{n,0}) + (1 - 2\pi)\rho\sigma_n \frac{\mu_{c,1} - \mu_{c,0}}{2\sigma_c}. \tag{6.8}$$

Now, if $\pi = 0.5$, then c_n^*, defined in (6.8), becomes equal to c_n, given in (6.2). Thus, if the sample is an equal mixture of cases and controls and the slopes of the regression lines are equal, the regression-based transfer will yield the correct cut-off. However, if $\pi \neq 0.5$, the result will be biased.

Finally, if we assume that $\rho_0 = \rho_1 \equiv \rho$ and that $\mu_{n,1} - \rho\sigma_n \frac{\mu_{c,1}}{\sigma_c} = \mu_{n,0} - \rho\sigma_n \frac{\mu_{c,0}}{\sigma_c}$, i.e., that the slopes and the intercepts of the regression lines for the $D = 0$ and $D = 1$ sub-samples are equal, then c_n^*, defined in (6.8), becomes equal to c_n, given in (6.2), irrespectively of π. Thus, in case the regression lines for the $D = 0$ and $D = 1$ sub-samples are equal, the regression-based transfer method will yield the correct cut-off.

Consequently, using the estimated coefficients from (6.4) may lead to a biased cut-off value for the new assay. In fact, as shown above, assuming bivariate normality with equal variances, unbiased results are obtained only if (1) the regression lines have the same intercept and slope in the AD and control groups, or if (2) the regression lines have the same slope in both groups and the data set contains an equal number of diseased and control subjects.

An example of a biased setting is shown in Figure 6.1. In panel *a* of this figure, the underlying theoretical binormal distributions are indicated in green for the controls and red for the cases. The different dependence between the current and new platform observations conditional on the true disease status – indicated by the red and green solid-line, respectively – is apparent. The cut-off estimated by the linear-regression-based transfer method is indicated by the blue dashed-line. The true underlying cut-off of the new platform is indicated by the black dashed-line. It is clear from panel *a* of Figure 6.1 that the new-assay cut-off $\hat{c}_n$ obtained by the linear-regression transfer method is biased. Panel *b* shows a random sample of 64 observations from the theoretical setting displayed in panel *a*. Without any indication of the true disease status of the subjects contained in the data, it is very difficult to conclude from panel *b* of Figure 6.1 whether there is a different linear relationship between the current- and new-assay for AD and control patients.

In order to obtain unbiased and more efficient estimation of the new cut-off estimate $\hat{c}_n$, we propose a two-stage Bayesian approach which is discussed in the next section.

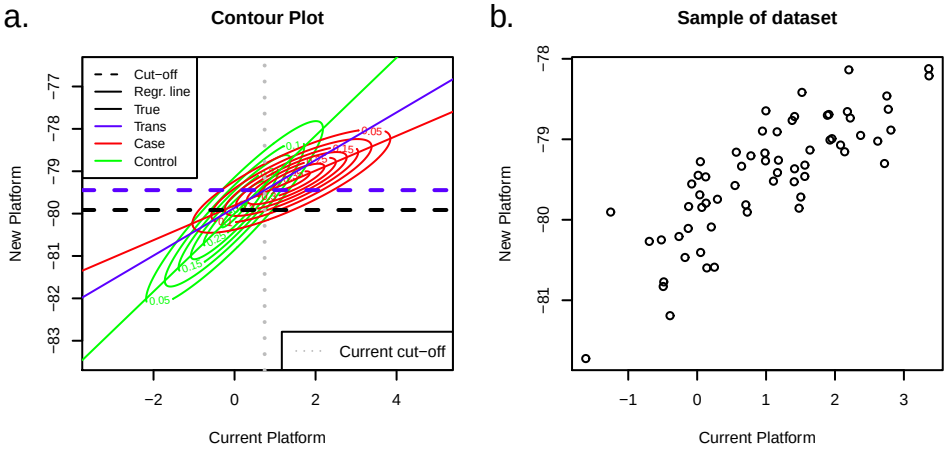

Figure 6.1: Illustration of a possible underlying true setting where the application of the linear-regreassion based cut-off transfer method would lead to a biased estimate of the new-assay cut-off. **a.** True underlying bivariate normal distributions by true disease status. Control and case distributions are indicated in green and red, respectively. Underlying linear relations are indicated by the solid lines colored by the respective disease group – estimated linear-regression relation indicated in blue. The new-assay cut-off is indicated by the dashed line, colored based on its estimation method – blue = linear-regression-based; black = true cut-off. **b.** Example of simulated data ($N = 64$) based on the true underlying setting shown in panel *a*.

6.2.3 Two-stage Bayesian cut-off transfer method

This approach is a novel method that allows transferring the current-assay cut-off value to a new assay when the disease-status of the subjects, for which measurements for the current and new assays were collected, is unavailable or only based on an imperfect clinical-diagnosis.

Stage 1

In the first stage, information about the distribution of the current-assay measurements is obtained by analysing the current-assay cut-off data set, in which GS reference-test information (autopsy-confirmation) of the diagnosis is also available. In particular, normal distributions are fitted to the measured biomarker-values (or transformations thereof) for the AD and control group by using a Bayesian model with as uninformative prior distributions as possible.

The likelihood of this latent-class model with two classes is defined as follows:

$$L\left(\mu_{c,0}, \mu_{c,1}, \sigma_{c,0}, \sigma_{c,1} | \boldsymbol{y}_c, \boldsymbol{d}\right) = \prod_{i=1}^{N_c} \left[\frac{1}{\sqrt{2\pi\sigma_{c,1}^2}} exp\left\{ \frac{1}{2\sigma_{c,1}^2} \left(y_{c,i} - \mu_{c,1}\right)^2 \right\} \right]^{d_i} \times \left[\frac{1}{\sqrt{2\pi\sigma_{c,0}^2}} exp\left\{ \frac{1}{2\sigma_{c,0}^2} \left(y_{c,i} - \mu_{c,0}\right)^2 \right\} \right]^{1-d_i}, \quad (6.9)$$

where $\mu_{c,d}$ and $\sigma_{c,d}^2$ are the mean and variance of the AD ($d = 1$) and control ($d = 0$) current-assay biomarker distributions, respectively. Vector $\boldsymbol{y}_c = \left(y_{c,1}, \ldots, y_{c,N_c}\right)^T$ contains the biomarker data on N_c subjects (indexed by i) measured with the current-assay and $\boldsymbol{d} = \left(d_1, \ldots, d_{N_c}\right)^T$ contains their GS reference-test values.

As in Chapter 5, we propose to reparametrise the problem in terms of $\gamma_c = \Phi^{-1}\left(AUC_c\right)$. We assume a flat prior distribution for the AUC of the current-assay (AUC_c), while the prior for the mean of the AD group is defined as $\mu_{c,1} = \left[\Phi^{-1}\left(AUC_c\right) \times \sqrt{\sigma_{c,0}^2 + \sigma_{c,0}^2}\right] + \mu_{c,0}$. As shown in Section 5.2 of Chapter 5, a standard-normal distribution prior for γ_c leads to a flat prior for AUC_c. The prior distributions for the remaining parameters $\mu_{c,0}$, $\sigma_{c,0}$, and $\sigma_{c,1}$ are summarized in Table 6.1.

Table 6.1: Proposed prior distributions for stage 1 of the two-stage Bayesian cut-off estimation approach.

Parameter	**Prior distribution**
Mean of biomarker values control group	
$\mu_{c,0}$	$N\left(0, 10^6\right)$
Scaled difference biomarker distribution means	
γ_c	$N\left(0, 1\right)$
Biomarker-distribution standard deviations	
$\sigma_{c,d}$	$U\left(0, 1000\right)$

After fitting the model defined in (6.9), results are obtained for parameters $\mu_{c,0}$, γ_c, $\sigma_{c,0}$, and $\sigma_{c,1}$ in the form of posterior distributions, which can be used in the second stage of the analysis.

Stage 2

In the second stage, a Bayesian latent-class model with two classes is fitted to the new-assay-transfer data set. In this data set, measurements of the current and new assays are available for all subjects, but there is no GS information for the subjects' diagnosis. The latent-class model predicts the unknown disease status (AD or control) of the subjects using the biomarker values obtained with both assays.

The full-data likelihood of the Bayesian latent-class model in the second stage has a similar structure as the model presented in Section 3.2 of Chapter 3. The full-data likelihood expressed with the parameters of the problem at hand is defined as follows:

$$L\left(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1, \theta | \boldsymbol{Y}, \tilde{\boldsymbol{d}}\right) = \prod_{i=1}^{N} \left[\frac{1}{\sqrt{(2\pi)^K |\boldsymbol{\Sigma_1}|}} exp\left\{ -\frac{1}{2} \left(\boldsymbol{y}_i - \boldsymbol{\mu}_1\right)^T \boldsymbol{\Sigma}_1^{-1} \left(\boldsymbol{y}_i - \boldsymbol{\mu}_1\right) \right\} \theta \right]^{\tilde{d}_i} \times \left[\frac{1}{\sqrt{(2\pi)^K |\boldsymbol{\Sigma_0}|}} exp\left\{ -\frac{1}{2} \left(\boldsymbol{y}_i - \boldsymbol{\mu}_0\right)^T \boldsymbol{\Sigma}_0^{-1} \left(\boldsymbol{y}_i - \boldsymbol{\mu}_0\right) \right\} \{1 - \theta\} \right]^{1-\tilde{d}_i}, \quad (6.10)$$

where $\boldsymbol{\mu}_{\tilde{d}} = \left(\mu_{c,\tilde{d}}, \mu_{n,\tilde{d}}\right)^T$ are the mean vectors containing the current- and new-assay means in the respective latent-disease status groups ($\tilde{d} = 0$ for controls, $\tilde{d} = 1$ for cases). Moreover, $\boldsymbol{Y} = \left(\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_N^T\right)^T$, with $\boldsymbol{y}_i = \left(y_{c,i}, y_{n,i}\right)^T$ containing the biomarker values for the current- and new-assay for subject i. Vector $\tilde{\boldsymbol{d}} = \left(d_1, \ldots, d_N\right)^T$ contains the latent-disease status indicators for all N subjects, respectively. Finally, the overall variance-covariance matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ are defined as

$$\boldsymbol{\Sigma}_{\tilde{d}} = \begin{pmatrix} \sigma^2_{c,\tilde{d}} & \rho_{\tilde{d}} \sigma_{c,\tilde{d}} \sigma_{n,\tilde{d}} \\ \rho_{\tilde{d}} \sigma_{n,\tilde{d}} \sigma_{c,\tilde{d}} & \sigma^2_{n,\tilde{d}} \end{pmatrix} \quad \text{for } \tilde{d} \in \{0, 1\},$$

where the variances of the respective current- and new-assay biomarker distributions for disease group $\tilde{d}$ are denoted by $\sigma^2_{c,\tilde{d}}$ and $\sigma^2_{n,\tilde{d}}$. Finally, $\rho_{\tilde{d}}$ denotes the correlation between the current- and new-assay biomarker values conditional on latent-true disease status $\tilde{d}$. As in the first stage, the model is reparametrised such that $\gamma_c = \Phi^{-1}\left(AUC_c\right)$ and $\gamma_n = \Phi^{-1}\left(AUC_n\right)$.

The proposed prior distributions for the second-stage model are summarized in Table 6.2. In order to improve the efficiency of the estimation of $\hat{c}_n$, we propose to include information from the current-assay cut-off data set. The prior distributions

for the parameters of the current-assay normal distributions in the second stage are constructed by considering informative normal distributions with mean and variance based on the posterior distributions from stage 1. These parameters are set equal to the empirical mean and variance from the obtained posterior samples from the first stage. To ensure strictly positive values for $\sigma_{c,0}$ and $\sigma_{c,1}$, the considered normal prior distributions are truncated to be larger than 0. For the remaining parameters, flat priors are assumed.

Table 6.2: Proposed prior distributions for stage 2 of the two-stage Bayesian cut-off estimation approach. $\hat{x}_{\mu_{c,0}|\boldsymbol{y}_c}$ and $s^2_{\mu_{c,0}|\boldsymbol{y}_c}$, and $\overline{x}_{\gamma_c|\boldsymbol{y}_c}$ and $s^2_{\gamma_c|\boldsymbol{y}_c}$ denote the empirical mean and variance of the posterior distributions of $\mu_{c,0}$ and γ_c, respectively, coming from the first stage.

Parameter	**Prior distribution**
Prevalence	
θ	$U(0.1, 0.9)$
Mean of biomarker values control group	
$\mu_{c,0}$	$N\left(\overline{x}_{\mu_{c,0}\|\boldsymbol{y}_c}, s^2_{\mu_{c,0}\|\boldsymbol{y}_c}\right)$
$\mu_{n,0}$	$N\left(0, 10^6\right)$
Scaled difference biomarker distribution means	
γ_c	$N\left(\overline{x}_{\gamma_c\|\boldsymbol{y}_c}, s^2_{\gamma_c\|\boldsymbol{y}_c}\right)$
γ_n	$N(0, 1)$
Biomarker-distribution standard deviations	
$\sigma_{c,d}$	$N\left(\overline{x}_{\sigma_{c,d}}, s^2_{\sigma_{c,d}}\right) trunc(0, +\infty)$
$\sigma_{n,d}$	$U(0, 1000)$
Correlation between current- and new-assay	
ρ_d	$U(-1, 1)$

The model then estimates the normal-distribution parameters (means and variances) for the biomarker values of both assays in the AD and control populations. Using the posterior estimates in (6.1) leads to a posterior distribution of $\hat{c}_n$, from which a point-estimate can be selected. The precision of the new-assay cut-off value is estimated by the standard deviation of the posterior distribution of the cut-off value.

6.2.4 Data sets

Two data sets were available for which AD CSF-biomarker cut-off transfer was considered. The first data set was obtained from BIODEM, the Reference Center for

Biological Markers of Dementia (University of Antwerp, Belgium) and contained information about two AD cohorts. The second data set, contained data for one cohort from Euroimmun AG.

INNOTEST-EUROIMMUN data

This set of data consists of two parts. The first part is a data set with CSF Aβ_{1-42} values of 42 age-matched control and 42 autopsy-confirmed-AD subjects measured with the ELISA kit INNOTEST® − βAMYLOID$_{(1-42)}$ tested in the BIODEM lab (referred to as unpublished CSF-data in [118]). This is the current-assay cut-off data set (see Section 6.2), as the information of the autopsy-confirmed-AD status of subjects is available. Figure 6.2 presents the histograms of the INNOTEST measurements for the control and AD groups.

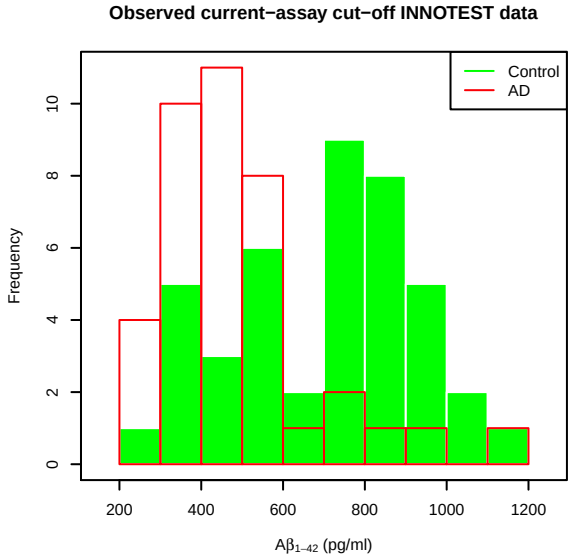

Figure 6.2: Histograms of the INNOTEST measurements for the control (green) and AD (red) groups (current-assay cut-off data set).

The second part was a data set consisting of CSF Aβ_{1-42} values of 64 samples, tested side-by-side with the INNOTEST (current) assay and the EUROIMMUN AG (new) assay (see Figure 6.3). This is the new-assay transfer data set, because there is no autopsy-confirmed diagnosis available for the samples.

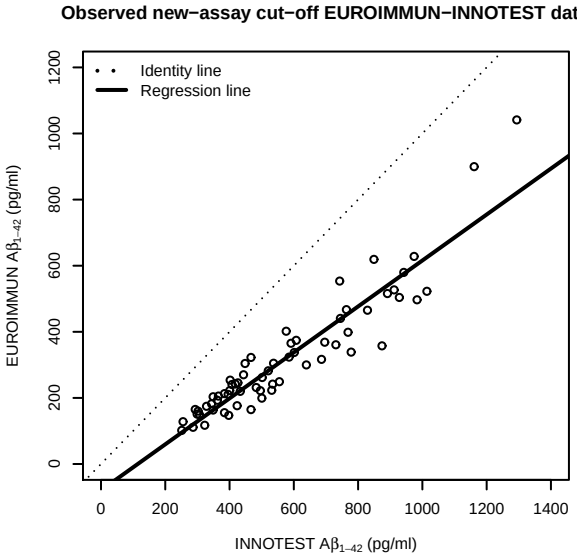

Figure 6.3: Scatter plot of the INNOTEST and EUROIMMUN measurements (new-assay cut-off data set).

INNOTEST-INNOBIA data

This set of data contained CSF Aβ_{1-42} values of 95 control and 51 autopsy-confirmed AD subjects, measured with the commercially available single-parameter ELISA kit INNOTEST® − βAMYLOID$_{(1-42)}$ (current-assay) and the multiplex xMAP format (Luminex Corp, Austin, Texas) with INNO-BIA AlzBio3 (new-assay). In this data set (described in [72]) information about the autopsy-confirmed-AD status of subjects is available for both the current- and the new-assay. Figure 6.4 presents the histograms of the INNOTEST and INNOBIA measurements for the control and AD groups. The scatter-plot of the measurements together with estimated regression lines for both diagnosis groups is shown in Figure 6.5.

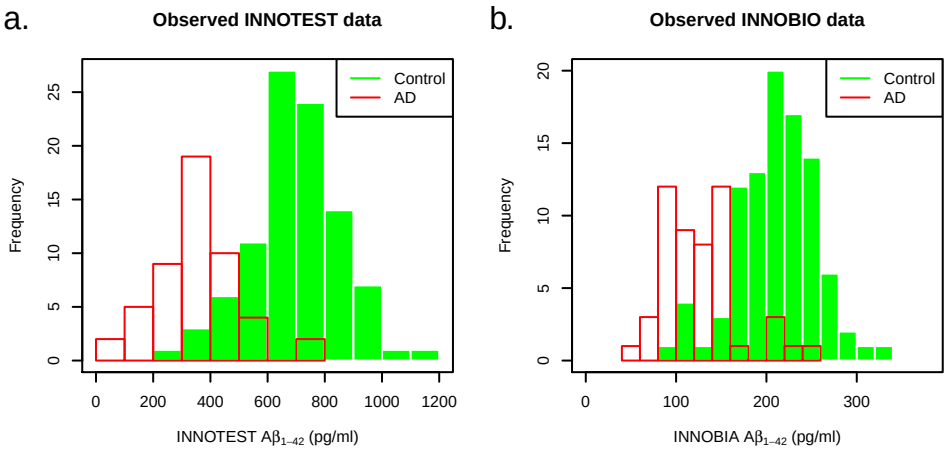

Figure 6.4: Histograms of the observed measurements for the control (green) and AD (red) groups. **a**. INNOTEST data. **b**. INNOBIA data.

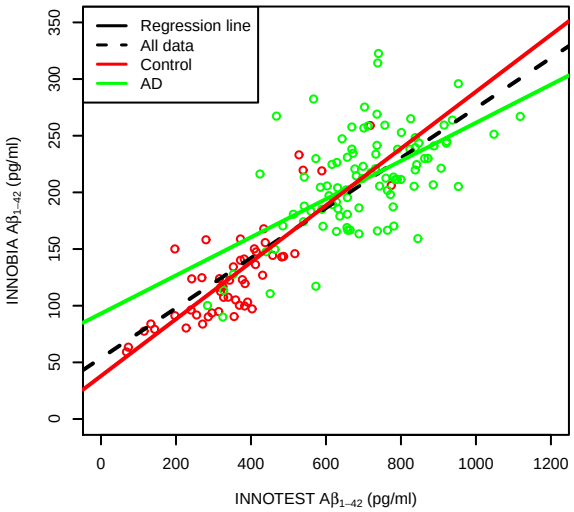

Figure 6.5: Scatter plot of INNOTEST and INNOBIA measurements, with a common regression line (dashed black line) and regression lines by group (repectively colored solid lines).

6.3 Simulation study

To check the performance of the proposed two-stage Bayesian model described in the previous section, a simulation study was performed. Correlated bivariate normal data, $Y_d = (Y_{c,d}, Y_{n,d})^T$, expressing a linear relationship between $Y_{c,d}$ and $Y_{n,d}$ conditionally on group d, were generated. We assumed a perfect linear-relationship between the parameters of the marginal normal distributions for each group d. Specifically, we assumed $\mu_{n,d} = \beta_{0,d} + \beta_{1,d}\mu_{c,d}$ and $\sigma^2_{n,d} = \beta^2_{1,d}\sigma^2_{c,d}$, for the marginal means and variances, respectively. The assumed correlation between $Y_{c,d}$ and $Y_{n,d}$, denoted by ρ_d, effects the relationship between the parameters such that

$$Y_{n,d} = \beta^*_{0,d} + \beta^*_{1,d} Y_{c,d},$$

with

$$\begin{aligned}\beta^*_{0,d} &= \beta_{0,d} + (1-\rho_d)\beta_{1,d}\mu_{c,d},\\ \beta^*_{1,d} &= \rho_d \beta_{1,d}.\end{aligned}$$

$\beta^*_{0,d}$ and $\beta^*_{1,d}$ are referred to as the *observed* intercept and slope of the linear relation between the current- and new-assay in the respective disease status groups.

6.3.1 Simulation scenarios

Two simulation scenarios were considered. In the first setting, data were simulated assuming a similar 'observed' relationship between the current- and new-assay measurements for both the AD and the control populations. The second scenario corresponds to the example already shown in Figure 6.1, where a different relationship between the current- and new-assay measurements was assumed depending on true disease status d. Table 6.3 contains the parameter values of both scenarios.

Scenario 1

In the first scenario (see the second column of Table 6.3), data were simulated assuming a similar 'observed' relationship between the current- and new-assay measurement for both disease populations and a disease prevalence of 0.5. In order to obtain the same 'observed' regression line in both disease groups, i.e. that, $\beta^*_{0,0} = \beta^*_{0,1}$, and $\beta^*_{1,0} = \beta^*_{1,1}$ we need to ensure that

$$\begin{aligned}\rho_0 &= \rho_1 \equiv \rho,\\ \beta_{1,0} &= \beta_{1,1} \equiv \beta_1,\\ \beta_{0,1} &= \beta_{0,0} + (1+\rho)\beta_1\left(\mu_{c,0} - \mu_{c,1}\right).\end{aligned}$$

As can be seen from the second column of Table 6.3, the considered parameters comply with these prerequisites. In terms of the operating characteristics, the first simulation scenario represents data with $AUC = 0.76$ for both the current- and new-assay, while at the optimal cut-off, sensitivity and specificity are equal to 0.83 and 0.6, respectively, for both assays.

Figure 6.6 shows the underlying true distributions (panel *a*) and an example of a simulated data set (panel *b*) based on the underlying true distributions. As expected, only a negligible difference between the true cut-off of the new-assay (dashed black line) and the estimated cut-off using the linear regression translation method (dashed blue line) can be observed.

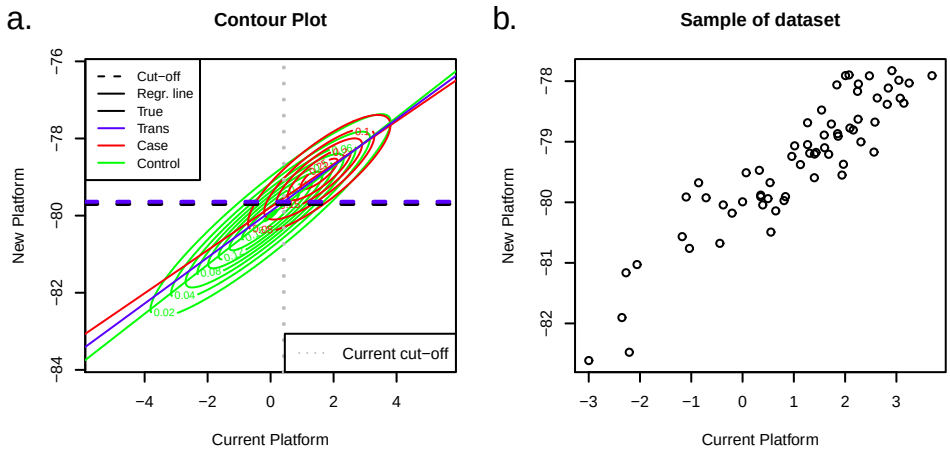

Figure 6.6: Illustration of a possible underlying true setting where the application of the linear-regression-based cut-off transfer method would lead to a satisfactory unbiased estimate of the new-assay cut-off. **a.** True underlying bivariate normal distributions by true disease status. Control and case distributions are indicated in green and red, respectively. Underlying linear relations are indicated by the solid lines colored by the respective disease group – estimated linear-regression relation indicated in blue. The new-assay cut-off is indicated by the dashed line, colored based on its estimation method – blue = linear-regression-based; black = true cut-off. **b.** Example of simulated data ($N = 64$) based on the true underlying setting shown in panel *a*.

Scenario 2

The parameters underlying the second simulation scenario are summarized in the third column of Table 6.3. In this scenario, the prevalence of disease is assumed to be 0.8 and the parameter values were selected as to show a clear bias in the case the linear-regression transfer method would be applied. For this simulation setting the AUC of the current and new assay are equal to 0.83 and 0.72, respectively. At the cut-off maximizing the Youden index, the sensitivity and specificity are equal to 0.87 and 0.53, respectively.

Table 6.3: Underlying true parameter values used to simulate the data for the simulation study.

Parameter	**Values for Setting 1**	**Values for Setting 2**
Distribution parameters		
$\mu_{c,0}$	0.00	0.00
$\mu_{c,1}$	1.54	1.40
$\mu_{n,0}$	-80.00	-80.00
$\mu_{n,0}$	-78.90	-78.23
$\sigma^2_{c,0}$	3.30	1.30
$\sigma^2_{c,1}$	1.48	1.48
$\sigma^2_{n,0}$	1.57	1.30
$\sigma^2_{n,1}$	0.68	0.37
Correlation coefficients		
ρ_0	0.92	0.92
ρ_1	0.82	0.82
Linear coefficients		
$\beta_{0,0}$	-80	-80
$\beta_{0,1}$	-79.96	-79.93
$\beta_{1,0}$	0.69	1
$\beta_{1,1}$	0.68	0.5
$\hat{\beta}_{0,0}$	-80	-80
$\hat{\beta}_{0,1}$	-79.78	-79.81
$\hat{\beta}_{1,0}$	0.64	0.92
$\hat{\beta}_{1,1}$	0.56	0.41

Simulation-study size

For each simulation scenario, current-assay cut-off data and new-assay transfer data were simulated. The current-assay cut-off data sets contained (simulated) biomarker values measured with the current assay and the AD-status of the subjects. To investigate the effect of sample size on the estimation of the new-assay cut-off, we considered sample sizes of 84, 150, and 300 subjects. The smallest sample size (84) was chosen to be equal to the available real data in the INNOTEST-EUROIMMUN current-assay cut-off data set (see Section 6.2.4).

The new-assay-transfer data sets contained (simulated) biomarker values measured with the current- and new-assay, but no true disease status of the subjects. As for the current-assay cut-off data sets, sample size was varied and sample sizes of 64, 150, and 300 subjects were considered. Again, the smallest sample size (64) corresponds to the available new-assay-transfer data contained in the INNOTEST-EUROIMMUN

data (see Section 6.2.4). For each considered data set type and sample size, 400 data sets were simulated.

6.3.2 Model fitting and diagnostics

For each current-assay cut-off and new-assay-transfer data set combination, the cut-off for the new-assay was estimated with the linear-regression transfer model and the proposed two-stage Bayesian model. In addition, also the results for the implied sensitivity and specificity of the cut-off are considered.

The SE and 95% CI for the linear-regression transfer-method estimates were obtained by bootstrapping. Ten-thousand bootstrapped data sets were sampled, cut-off values were obtained, and the sample standard deviation of the resulting 10,000 cut-off values was taken as the estimate of the SE. The 95% confidence interval for the estimated cut-off value was obtained by selecting the 2.5 and 97.5% percentiles of the bootstrapped cut-off values.

For the two-stage Bayesian model, the estimates of the parameters of interest were obtained by using 10,000 samples from the posterior distribution after a burn-in period of 10,000 samples from five independent MCMC chains. Median posterior values were considered as point-estimates and retained for each data set fit. Starting values for the MCMC chains were fixed at plausible data-based values for all parameters.

After fitting, the results were first checked by general diagnostic-tools in order to assess convergence of the MCMC chains. Convergence over chains was investigated by the Gelman-Rubin convergence index, for which a cut-off value of 1.1 was applied [38]. Chain-by-chain convergence was checked by using the Geweke convergence criterion [40]. Fits for which the Gelman-Rubin index suggested non-convergence were excluded from the results, while the Geweke index was monitored to ensure that, on average, no more than two out of five chains were considered as non-converged for each parameter over all simulated data sets.

The models were fitted by using OpenBUGS 3.2.1 [60]. Annotated BUGS code can be found in Section A.4 of Appendix A. Results were analyzed and summarized using R 3.0.1 (x64) [90]. The R-package R2OpenBUGS [111] was used as an interface between R 3.0.1 and OpenBUGS.

6.4 Data application

Model applicability was also investigated by fitting the two data sets described in Section 6.2.4.

For the INNOTEST-EUROIMMUN data set, autopsy-confirmed-AD status is only available for the current-assay data; hence, the cut-off for the new-assay has to be transferred from the current-assay.

Since in the INNOTEST-INNOBIA data set autopsy-confirmed-AD status is available for both the current- and the new-assay, it is possible to directly estimate the cut-off value for the new-assay. By discarding the autopsy-confirmed-AD status in this data set, we can consider it a new-assay-transfer data set and use it to transfer the current-assay cut-off estimate from the INNOTEST-EUROIMMUN current-assay data set. This allows us to compare the transferred new-assay cut-off to the estimate directly from the INNOTEST-INNOBIA data set. Moreover, since autopsy-confirmed-AD status is available, it is possible to directly estimate and compare sensitivity and specificity related to the differently estimated cut-off values.

6.4.1 Model fitting and diagnostics

For each data set, the cut-off for the new-assay was estimated with the linear-regression transfer model and the proposed two-stage Bayesian model. In addition, the results for the implied sensitivity and specificity of the cut-off are also considered.

The SE and 95% CI for the linear-regression transfer method estimates were obtained as in the simulation study (see Section 6.3.2).

As for the simulation study, the estimates of the coefficients of the model were obtained by using 10,000 samples from the posterior distribution after a burn-in period of 10,000 samples from five independent MCMC chains. After fitting, convergence-diagnostics measures similar to those used in the analysis of the simulated data were applied (see Section 6.3.2).

6.5 Results

6.5.1 Simulation study

We investigated the performance of the linear-regression-based method (see Section 6.2.2) and the Bayesian approach (see Section 6.2.3) using the simulated data (see Section 6.3). Figure 6.7 shows the means and 95% empirical intervals of the 400 cut-off estimates for all settings in the two considered simulation study scenarios. In the first simulation scenario, a similar linear relationship between the current- and new-assay measurements in the control and AD populations was assumed. As expected, in this case, the left panel of Figure 6.7 shows that both methods result in unbiased estimates of the cut-off value, i.e., the estimated values are on average equal to the

true value for the linear-regression-based transfer as well as the two-stage Bayesian method.

In terms of efficiency, the two-stage Bayesian approach provides more efficient results in all sample size settings as compared to the linear-regression-based transfer method, as indicated by the narrower 95% CIs. Within each approach, sample size seems to affect efficiency differently. For the linear-regression-based transfer method, it is the sample size of the current-assay cut-off data set which is most important in terms of efficiency. Increasing the current-assay cut-off data set size decreases the length of the 95% CI more than increasing the size of the new-assay-transfer data set. For the two-stage Bayesian approach the opposite is observed. Efficiency is increased more by considering larger new-assay-transfer data sets than increasing the size of the current-assay cut-off data set.

To make the results more interpretable, summary statistics of the sensitivity and specificity, corresponding to the estimated cut-off values, are presented in Table 6.4. In the table, the means and empirical 95% CIs of the cut-off estimates as well as the corresponding sensitivities and specificities are contained for all simulation settings for the first simulation scenario. Results show that for the largest data set combination, $N_c = N_t = 300$, the empirical 95% CI of the new-assay cut-off estimate is $[-80.092; -79.292]$ and $[-79.925; -79.193]$ for the linear-regression-based and two-stage Bayesian approach, respectively. The corresponding sensitivity and specificity of these cut-offs lead to 95%-CIs of $[0.692; 0.946]$ and $[0.474; 0.718]$, and $[0.7262; 0.896]$ and $[0.525; 0.659]$ for the linear-regression-based and two-stage Bayesian approach, respectively. Although the two-stage Bayesian approach results are consistently more precise, the 95% CIs of the sensitivity and specificity estimates are still considerably wide.

Table 6.4: Means and empirical 95% confidence intervals of the estimated cut-off ($\hat{c}_c$) and corresponding sensitivity ($Se_{\hat{c}_c}$) and specificity ($Sp_{\hat{c}_c}$) for the linear-regression-based and two-stage Bayesian methods for simulation scenario 1.

Sample size		**Linear-regression cut-off transfer**			**Two-stage Bayesian cut-off transfer**		
N_c	N_t	$\hat{c}_n$	$Se_{\hat{c}_n}$	$Sp_{\hat{c}_n}$	$\hat{c}_n$	$Se_{\hat{c}_n}$	$Sp_{\hat{c}_n}$
84	**64**	-79.685 $[-80.297; -79.073]$	0.808 $[0.618; 0.998]$	0.597 $[0.413; 0.781]$	-79.692 $[-80.109; -79.275]$	0.819 $[0.680; 0.958]$	0.596 $[0.473; 0.719]$
	150	-79.690 $[-80.281; -79.097]$	0.810 $[0.618; 1.002]$	0.600 $[0.417; 0.773]$	-79.700 $[-80.022; -79.376]$	0.820 $[0.725; 0.921]$	0.590 $[0.494; 0.694]$
	300	-79.697 $[-80.260; -79.134]$	0.814 $[0.634; 0.994]$	0.593 $[0.424; 0.762]$	-79.699 $[-79.983; -79.415]$	0.827 $[0.735; 0.919]$	0.591 $[0.505; 0.677]$
150	**64**	-79.685 $[-80.222; -79.148]$	0.811 $[0.637; 0.985]$	0.597 $[0.436; 0.758]$	-79.693 $[-80.073; -79.313]$	0.820 $[0.697; 0.943]$	0.596 $[0.482; 0.710]$
	150	-79.692 $[-80.223; -79.161]$	0.814 $[0.645; 0.983]$	0.595 $[0.434; 0.756]$	-79.705 $[-80.001; -79.409]$	0.826 $[0.736; 0.916]$	0.592 $[0.500; 0.684]$
	300	-79.706 $[-80.210; -79.002]$	0.819 $[0.664; 0.974]$	0.591 $[0.438; 0.744]$	-79.707 $[-79.942; -79.472]$	0.828 $[0.755; 0.901]$	0.592 $[0.519; 0.665]$
300	**64**	-79.696 $[-80.139; -79.253]$	0.818 $[0.683; 0.953]$	0.595 $[0.460; 0.730]$	-79.700 $[-79.994; -79.406]$	0.825 $[0.735; 0.915]$	0.594 $[0.504; 0.684]$
	150	-79.681 $[-80.104; -79.258]$	0.815 $[0.682; 0.948]$	0.599 $[0.470; 0.728]$	-79.699 $[-79.962; -79.36]$	0.825 $[0.741; 0.909]$	0.594 $[0.514; 0.674]$
	300	-79.692 $[-80.092; -79.292]$	0.819 $[0.692; 0.946]$	0.596 $[0.474; 0.718]$	-79.709 $[-79.925; -79.193]$	0.829 $[0.762; 0.896]$	0.592 $[0.525; 0.659]$
True values		**-79.699**	**0.828**	**0.595**	**-79.699**	**0.828**	**0.595**

In the second simulated scenario, different linear relationships between the current- and new-assay measurements in the control and AD populations were assumed. In this case, the linear-regression-based method leads to biased estimates of the cut-off value, for the reasons explained in Section 6.2. On the other hand, the Bayesian approach provides unbiased estimates of the cut-off value (Figure 6.7).

From Figure 6.7 it is also clear that, in terms of efficiency, increasing current-assay data set size increases efficiency for the linear-regression-based approach while considering larger new-assay-transfer data set size improves efficiency of the two-stage Bayesian approach.

Similar conclusions can be drawn for the sensitivity and specificity corresponding to the estimated cut-off values (see Table 6.5). From the table, the bias for the linear-regression-based approach is apparent for the estimated sensitivity and specificity corresponding to the new-assay cut-off. In terms of precision, for the $N_c = N_t = 300$ setting, the empirical 95%-CIs of $Se_{\hat{c}_n}$ and $Sp_{\hat{c}_n}$ are $[0.574; 0.880;]$ and $[0.537; 0.729]$, and $[0.818; 0.916]$ and $[0.484; 0.578]$ for the linear-regression-based and two-stage Bayesian approach, respectively.

Table 6.5: Means and empirical 95% confidence intervals of the estimated cut-off ($\hat{c}_c$) and corresponding sensitivity ($Se_{\hat{c}_c}$) and specificity ($Sp_{\hat{c}_c}$) for the linear-regression-based and two-stage Bayesian methods for simulation scenario 2.

Sample size		**Linear-regression cut-off transfer**			**Two-stage Bayesian cut-off transfer**		
N_c	N_t	$\hat{c}_n$	$Se_{\hat{c}_n}$	$Sp_{\hat{c}_n}$	$\hat{c}_n$	$Se_{\hat{c}_n}$	$Sp_{\hat{c}_n}$
84	**64**	-79.941 $[-80.129; -79.153]$	0.734 $[0.491; 0.977]$	0.621 $[0.462; 0.780]$	-79.851 $[-80.161; -79.541]$	0.838 $[0.713; 0.963]$	0.551 $[0.445; 0.657]$
	150	-79.590 $[-80.035; -79.141]$	0.710 $[0.470; 0.948]$	0.640 $[0.493; 0.783]$	-79.880 $[-80.092; -79.664]$	0.850 $[0.771; 0.935]$	0.540 $[0.468; 0.616]$
	300	-79.611 $[-80.048; -79.174]$	0.721 $[0.498; 0.944]$	0.631 $[0.488; 0.774]$	-79.903 $[-80.062; -79.744]$	0.863 $[0.806; 0.920]$	0.534 $[0.479; 0.589]$
150	**64**	-79.609 $[-80.009; -79.209]$	0.722 $[0.516; 0.928]$	0.632 $[0.501; 0.763]$	-79.864 $[-80.125; -79.603]$	0.846 $[0.744; 0.948]$	0.547 $[0.457; 0.637]$
	150	-79.600 $[-79.986; -79.214]$	0.718 $[0.516; 0.920]$	0.635 $[0.508; 0.772]$	-79.889 $[-80.083; -79.695]$	0.857 $[0.786; 0.928]$	0.539 $[0.470; 0.608]$
	300	-79.610 $[-79.988; -79.232]$	0.724 $[0.522; 0.926]$	0.632 $[0.509; 0.755]$	-79.905 $[-80.060; -79.750]$	0.864 $[0.809; 0.919]$	0.533 $[0.478; 0.588]$
300	**64**	-79.590 $[-79.925; -79.255]$	0.715 $[0.535; 0.895]$	0.639 $[0.529; 0.749]$	-79.863 $[-80.098; -79.628]$	0.846 $[0.752; 0.940]$	0.548 $[0.468; 0.628]$
	150	-79.592 $[-79.896; -79.288]$	0.717 $[0.552; 0.882]$	0.639 $[0.539; 0.739]$	-79.897 $[-80.079; -79.715]$	0.860 $[0.793; 0.927]$	0.536 $[0.473; 0.599]$
	300	-79.609 $[-79.901; -79.317]$	0.727 $[0.574; 0.880]$	0.633 $[0.537; 0.729]$	-79.911 $[-80.048; -79.774]$	0.867 $[0.818; 0.916]$	0.531 $[0.484; 0.578]$
True values		**-79.914**	**0.869**	**0.530**	**-79.914**	**0.869**	**0.530**

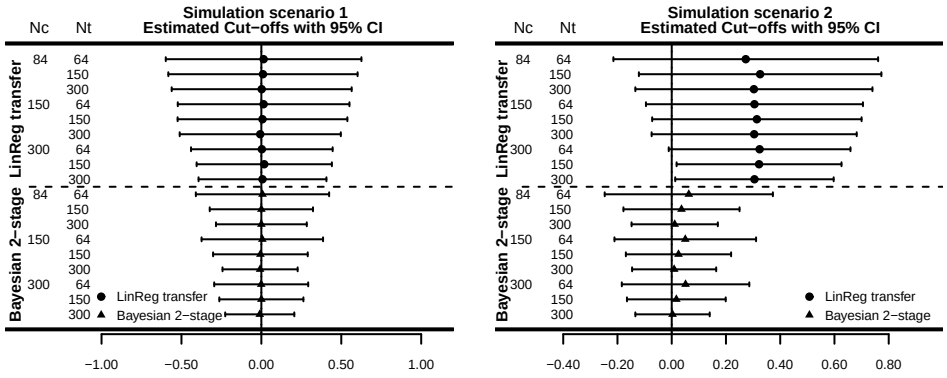

Figure 6.7: Means and 95% confidence intervals (based on empirical estimates of the mean and variance obtained from the 400 cut-off values for the simulated data sets) for estimated cut-off values for the new assay obtained with the linear-regression and Bayesian method, for the first and second simulation scenario (see text). Cut-off values were rescaled to obtain a true cut-off values equal to 0. Nc: Sample size 'current-assay cut-off' data, Nt: Sample size 'new-assay-transfer' data.

6.5.2 INNOTEST-EUROIMMUN data set

For the INNOTEST assay, the cut-off value was estimated directly from the diagnostic labels in the current-assay cut-off data set. In particular, the estimated value was equal to 638.5 (SE=55.39, 95% CI = [508.5, 728.0]), the same value as reported previously [118]. The sensitivity and specificity corresponding to the estimated cut-off value were equal to, respectively, 0.87 (SE=0.083, 95% CI = [0.64, 0.90]) and 0.62 (SE=0.069, 95% CI = [0.52, 0.79]). Worth noting are wide CIs for the estimated values.

For the EUROIMMUN assay, the cut-off value was obtained by transferring the INNOTEST cut-off value by using the linear-regression-based method and the Bayesian approach. In particular, the cut-off value obtained by the linear-regression-based method was equal to 364.4 (SE=39.48, 95% CI = [269.7, 426.8]), while for the Bayesian approach it was equal to 402.8 (posterior-distribution SD=31.68, 95% credible interval = [348.0, 473.9]). The obtained cut-off values are quite different, but given their precision, they cannot be seen as statistically significantly different.

To visualize the importance of the uncertainty about the derived cut-offs for the INNOTEST and EUROIMMUN assays, the cut-off values and the accompanying 95% CIs were plotted on the scatter plot of biomarker values measured with both assays

(Figure 6.8). The 95% CI for the INNOTEST cut-off value ranges from 508.5 pg/ml to 728.0 pg/ml or, expressed on a relative scale, from -20% to +14% of the estimated cut-off value. The 95% CI for the EUROIMMUN-assay cut-off value obtained by the linear-regression-based method ranges from 269.7 to 426.8 pg/ml or from -26% to +17% of the estimated cut-off value. The wide 95%CI for both assays imply that, by assuming different cut-off values within the CIs, the $A\beta_{1-42}$ -based disease-status could be potentially altered for 13 of 64 subjects (20%) for the INNOTEST assay and 17 of 64 subjects (27%) for the EUROIMMUN assay.

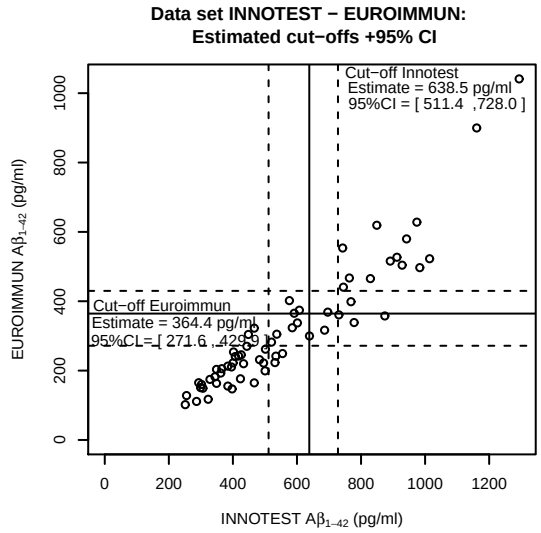

Figure 6.8: Scatter plot of $A\beta_{1-42}$ values measured with INNOTEST (X-axis) and EUROIMMUN (Y-axis) with estimated cut-off values and 95% CIs. INNOTEST: cut-off value based on the non-parametric ROC curve estimate and maximal Youden-index. EUROIMMUN: cut-off value obtained by the linear-regression-based method.

6.5.3 INNOTEST-INNOBIA data set

We first obtained cut-off values and their SEs for both assays using the fully non-parametric ROC-curve (see Section 6.2) estimated based on the available diagnosis information (Table 6.6, first two columns). Worth noting are wide 95% CIs for the estimated cut-off values, which indicate substantial uncertainty due to the limited sample size of the data set. This is similar to the case of the INNOTEST-EUROIMMUN current-assay cut-off data set (see Section 6.4).

Table 6.6: Estimates of cut-off values and the corresponding sensitivity and specificity for the INNOTEST-INNOBIA data set. Standard errors (SE) and 95% confidence (CI) or credible (CrI) intervals are indicated between round and squared brackets, respectively. First 2 columns: Estimated directly with non-parametric ROC analysis; last 2 columns: Estimates for the INNOBIA assay obtained by transferring the INNOTEST cut-off with linear regression and the novel two-stage Bayesian approach.

Parameter	**INNOTEST Non-para ROC Estimate (SE) [95% CI]**	**INNOBIA Non-para ROC Estimate (SE) [95% CI]**	**INNOBIA Lin. Reg. Estimate (SE) [95% CI]**	**INNOBIA 2-Stage Bayes Estimate (SE) [95% CrI]**
Cut-off	539.5 (37.69) [437.0; 553.1]	159.15 (3.13) [147.6; 168.5]	172.8 (8.9) [147.6; 179.6]	167.5 (5.58) [156.1; 178.0]
Sensitivity	0.94 (0.046) [0.80; 0.94]	0.88 (0.022) [0.78; 0.90]	0.90 (0.035) [0.78; 0.90]	0.91 (0.027) [0.85; 0.95]
Specificity	0.88 (0.028) [0.85; 0.95]	0.92 (0.016) [0.84; 0.94]	0.80 (0.051) [0.78; 0.94]	0.90 (0.024) [0.84; 0.94]

The estimated cut-off value for INNOTEST Aβ_{1-42} (539.5 pg/ml) is the same as the published value (see [72]) and somewhat smaller than the value obtained in the INNOTEST-EUROIMMUN current-assay cut-off data set (638.5 pg/ml). However, taking into account the considerable uncertainty associated with the estimates presented in Table 6.6, the difference could be either due to random variation or could be caused by changes over time in lab equipment or assay reagents.

Similarly to the 95% CIs for the cut-off values, the CIs for sensitivity and specificity, implied by the estimated cut-off values, are also wide. They indicate substantial uncertainty about the diagnostic performance of the assays.

In the next step, we estimated the cut-off value for the INNOBIA assay by applying the linear-regression-based and Bayesian methods to the INNOTEST-INNOBIA new-assay-transfer data set (see Section 6.4). The slopes of the linear relationship between the observed values of the assays were significantly different between AD and control cohorts ($p = 0.0374$), regression lines shown in Figure 6.5. The results are presented in Table 6.6 (last two columns). The obtained cut-off values are equal to 172.8 and 167.5 for the linear-regression-based method and the novel Bayesian approach, respectively. They are similar to the value of 159.1 obtained by the direct estimation (see Table 6.6, column 2), especially taking into account the limited precision of the obtained estimates.

6.6 Conclusions

The upcoming commercialization of a new generation of immunoassays for CSF AD-biomarkers will include the full automation of tests, improved between-center and between-lot variability, link to a reference method, and availability of run-validation or proficiency panels [49, 53, 55, 23]. It is hoped that these improved assays will enable the introduction of universal cut-off levels for the AD CSF-biomarkers [13]. However, due to the lack of left-over samples from the most important observational studies which have been used to document the value for the markers, it will be difficult or almost impossible to confirm the clinical utility of the new biomarker assays using samples which have been analysed previously with the first generation of assays.

In this chapter, we have proposed a novel, Bayesian approach to the problem of transferring a cut-off value to a new assay. Results of the simulation study suggest that the method performs better than the often-used linear-regression-based method. In particular, the latter requires that there exists a common linear relationship between the current- and new-assay measurements in the control and AD populations. If this assumption is violated, the method produces incorrect estimates of the cut-off value for the new assay. The validity of the common linear relationship cannot be verified if no reliable clinical-diagnosis information is available; yet, this is exactly the reason why a transfer of an existing cut-off value may be needed. Note that for the INNOTEST-INNOBIA Aβ_{1-42} data set, the assumption could be verified and was shown not to hold (Figure 6.5).

The proposed Bayesian approach does not make an assumption of the linear relationships. In addition, the Bayesian method results in unbiased and less variable estimates of the cut-off as compared to the linear-regression method. The comparison of widths of the 95% CI for different sample sizes (Figure 6.7, Table 6.4) demonstrates that the differences in precision between both methods are substantial, with the precision of the Bayesian cut-off for the smallest sample sizes (84 and 64) almost equal to the precision of the linear regression cut-off for the largest sample sizes (300 and 300).

Given that the Bayesian method provides unbiased cut-off estimates regardless of the linear relationships between assay results and makes better use of the available data, it is preferred over the linear regression method when a cut-off needs to be transferred.

The Bayesian approach does require that the biomarker measurements (or a transformed version thereof) are normally distributed. The Box-Cox transformation as a way of normalizing biomarker values has been shown to perform well in the ROC con-

text [80, 33]. If needed, the method could be adapted to a semi-parametric approach as the mixture of Dirichlet processes, which was proposed by [113] to establish an optimal threshold using a Bayesian approach when the disease status is known.

Chapter 7

Concluding remarks and future work

7.1 Concluding remarks

In this dissertation, we have proposed Bayesian models and approaches to accommodate issues related to the diagnosis of Alzheimer's Disease. The proposed methods are aimed at facilitating the development and validation of CSF-biomarker-based indices when only imperfect reference-test information is available.

Overall, the use of a Bayesian approach offers important flexibility. By construction, it can accommodate any prior information related to, e.g., the AUC of the combination of biomarkers or the diagnostic performance of the imperfect reference-test. Moreover, issues related to model non-identifiability can be mitigated by introducing a small amount of information over many parameters. In contrast, a frequentist approach would require strict restrictions on particular parameters. Throughout the dissertation, these important characteristics of Bayesian statistics have been applied.

Diagnostic accuracy estimation

Estimating diagnostic biomarker-index' accuracy when only imperfect reference-test information is available is not straightforward. Ignoring the imperfectness of the reference test results in biased estimates which may lead to the rejection of important and useful biomarkers. In Chapters 3 and 4 we have proposed a Bayesian latent-class model which provides unbiased estimates of the accuracy of a biomarker index,

even when the biomarkers underlying the index are correlated with the imperfect reference-test (see Chapter 4). The results obtained in these chapters suggest that the reports indicating disappointing results of diagnostic performance for the AD CSF-biomarkers might be due in part to the fact that clinical diagnosis was treated as a GS reference-test.

Validation

After the development of a diagnostic biomarker-index, the index should be validated. Currently, validation is rarely performed because of the need for large sample sizes or expensive data to reach adequate power of the validation study and the lack of an efficient statistical framework. In particular, GS reference-test data is usually scarce and expensive to obtain, while imperfect reference-test information contains less useful information for validation, causing validation sample sizes to increase. In Chapter 5, we have proposed such a framework allowing efficient validation of a diagnostic biomarker-index. Based on the exchangeability assumption of the parameters of the development and validation studies, a large reduction of the required sample size was shown possible. When exchangeability of observations or parameters is less obvious, the proposed method could still be applied. In principle, any informed prior distribution concerning the accuracy of a diagnostic index can be combined with validation study data. For example, diagnostic tools for children could be validated by including information from adults [92, 122]. Such priors should be constructed with care and accompanied by a clear discussion about the included available information. In our opinion, sacrificing independence between development and validation could be warranted when doing so would render a validation study feasible.

Cut-off transfer

As discussed in Chapter 6, developing and validating a diagnostic AD CSF-biomarker cut-off for a particular commercially available assay does not imply the applicability of the cut-off on other assays measuring the same biomarker. If new development and validation studies are to be avoided, the current cut-off can be transferred to a new assay. Underlying linear-relation assumptions of the currently applied method may lead to biased estimates of the new cut-off, resulting in a diagnostic test with different operating characteristics depending on the applied platform. Therefore, a novel two-stage Bayesian method has been proposed in Chapter 6. We have shown that this method leads to unbiased and more precise estimates than the currently applied linear-regression-based method. However, with the current size of develop-

ment and validation studies, only imprecise cut-off estimates, in terms of operating characteristics, are available. This is generally overlooked and not communicated in practice, but extremely important to acknowledge.

7.2 Topics for future work

Several assumptions and methodological aspects underlying the developed approaches are worth further investigation. In particular, the following sections discuss possible extensions and generalisations of the models developed in this dissertation. Possible future developments for the Bayesian latent-class mixture model are discussed in Section 7.2.1. Extensions of the developed validation and cut-off transfer method are proposed in Sections 7.2.2 and 7.2.3, respectively.

7.2.1 Bayesian latent-class mixture model

Weaknesses latent-class mixture models

Several weaknesses regarding latent-class mixture models in the context of binary or ordinal tests have been discussed [1, 84, 88, 2]. A first issue relates to the definition of disease. In the absence of a GS reference-test, disease is considered a latent concept. In this case, it is not easy to define disease explicitly as in this type of models, it is the entity which simply links the tests. Secondly, although accounting for conditionally dependent tests can resolve biased estimates with respect to test accuracy, different proposals for dependence models lead to different results in different contexts. How these limitations translate to the models proposed in this dissertation remains to be investigated. The first issue may be mitigated by introduction of partial gold standard reference-test information [2]. By following the ideas of [88] one could think of investigating the second issues by defining other dependence models and comparing the impact of these models on the results.

Transformation to normality

Currently, the Bayesian latent-class mixture model has been developed for biomarkers which are normally distributed or for which a transformation to normality exists, conditional on true disease status. For many biomarkers such a transformation is required (e.g., see the application in Chapter 3). At this point it is up to the user to select an appropriate transformation which should be applied before the model can be fitted to the data. This is generally not an easy task especially when only imperfect

reference-test information is available. For example, under the assumption of conditional independence between biomarker and imperfect reference-test, the observed distribution of the biomarker, conditional on the imperfect reference-test values, will be skewed, as shown in Figure 7.1. Although the underlying biomarker distributions for true cases and controls is normal (indicated by the dashed lines), the misclassification in the reference test causes the observed distributions to be skewed towards each other (shown by the histograms).

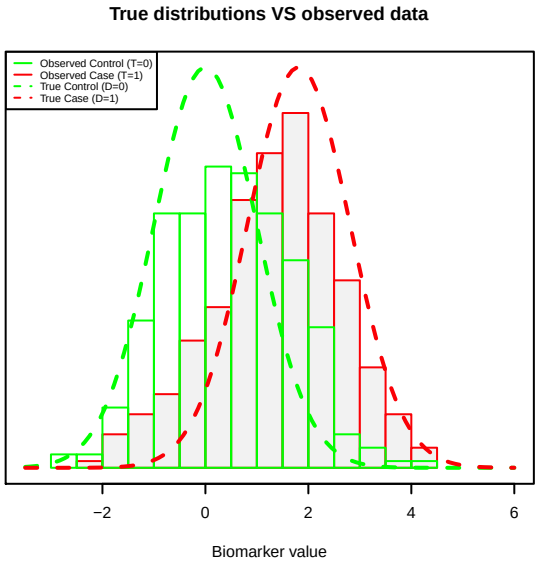

Figure 7.1: Underlying true (dashed lines) and observed (histograms) biomarker distributions by imperfect reference-test. Cases (red) and controls (green) are misclassified under the conditional independence assumption between biomarker and imperfect reference-test.

To overcome the need of pre-specifying the form of transformation to normality, an extension of the proposed model could be developed. In particular, by including a Box-Cox power transformation into the model, it should be possible to estimate the most appropriate transformation to normality while considering this on the scale of the true latent-disease status. This may not be straightforward, however, because of non-identifiability issues related to the introduction of this power parameter.

Selection of optimal biomarker combination

In current applications, it has been assumed that the biomarkers of interest, constituting the diagnostic index, are already identified and that interest only lies in how to optimally combine these biomarkers into a linear combination. A possible future use of the model could be to investigate a pool of biomarkers and select the linear combination of 'optimal' biomarkers. This selection could be made based on the estimated values of the elements in $\hat{\boldsymbol{a}}$. In order to do this, the model should be able to perform with a large number of biomarkers, many of which might not be useful for purposes of classification.

7.2.2 Validation

Model-based weighting of development-study information

As an alternative to discounting development-study posterior information as prior information in the validation study, model-based approaches could be considered. One could extend the proposed model by implementing the power-prior [45, 74] or meta-analytic approach methods developed to include historical information into Bayesian analysis. In order to account for conflicting prior and data information, these methods can even be extended by considering the commensurate and robustified versions for the power- and meta-analytic-prior method, respectively. Specifically, the methods allow estimation of the amount of prior information that will be included in the analysis of new data according to how well the prior and new data agree.

Validation criterion

Alternative criteria to conclude to validation can be proposed as well. Given the Bayesian setting one could think of reformulating the hypothesis test into a criterion based on the posterior AUC_a distribution which could be assessed continuously during the validation process. The validation process could then be concluded when enough information has been obtained to consider validation or not.

Conditional dependence setting

Conditional dependence between the biomarkers and imperfect reference-test will also effect the diagnostic accuracy estimates in the validation model. Future work could entail the extension of the validation model in a similar way as proposed in Chapter 4. More specifically, one could allow for conditional dependence between the continuous diagnostic-index and the imperfect reference-test by considering a latent tolerance variable underlying the imperfect reference-test results.

7.2.3 Cut-off transfer

One-stage Bayesian approach

An obvious extension of the proposed two-stage Bayesian model to transfer a current-assay cut-off to a new-assay cut-off, is to consider a general one-stage approach. This should be possible by considering a joint model for the current-assay and new-assay-transfer data sets. One could consider the combined data set where, for one set of observations, only measurements for the current assay are available together with true disease status information. For the remaining observations, measurements

for both assays are available, but true disease status information is missing. By virtue of the Bayesian method, these missing components are considered as parameters characterised by posterior information, enabling estimation of the parameters defining the assumed underlying mixture of multivariate normal distributions.

Model-based weighting of stage one information

As for the validation model, the two-stage Bayesian transfer model also includes prior information for several parameters coming from previously obtained posterior distributions. Currently, the introduction of this information is only based on heuristic arguments and approximations. Further work is needed to investigate the impact of these approximations and the possibility to invoke the (commensurate) power- and/or (robustified) meta-analytic approach-prior definitions to end up with a model-based weighting of prior information.

Bibliography

[1] P.S. Albert and L.E. Dodd. Cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*, 60:427–435, 2004.

[2] P.S. Albert and L.E. Dodd. On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. *Journal of the American statistical association*, 103:61–73, 2008.

[3] U. Andreasson, E. Vanmechelen, L.M. Shaw, H. Zetterberg, and H. Vanderstichele. Analytical aspects of molecular alzheimer's disease biomarkers. *Biomarkers in medicine*, 32:377–389, 2012.

[4] A. Azzalini and A. Capitano. *The skew-normal and related families.* Cambridge Press, 2014.

[5] M. Baker. In biomarkers we trust? *Nature biotechnology*, 23:297–304, 2005.

[6] S.G. Baker, E. Schuit, E.W. Steyerberg, M.J. Pencina, A. Vickers, K.G.M. Moons, B.W.J. Mol, and K.S. Lindeman. How to interpret a small increase in auc with an additional risk prediction marker: decision analysis comes through. *Statistics in medicine*, 33(22):3946–3959, 2014.

[7] D. Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology*, 12:387–415, 1975.

[8] J. Barnard, R. McCulloch, and X.L. Meng. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica sinica*, 10:1281–1331, 2000.

[9] J.W. Bartlett, C. Frost, N. Mattsson, T. Skillbäck, K. Blennow, H. Zetterberg, and J.M. Schott. Determining cut-points for alzheimer's disease biomarkers: statistical issues, methods and challenges. *Biomarkers in medicine*, 6(4):391–400, 2012.

[10] T.T. Bayes. An essay towards solving a prblem in the doctrine of chances. *Philosophical transactions of the royal society of London*, 53:370–418, 1763.

[11] T.G. Beach, S.E. Monsell, L.E. Philips, and W. Kukull. Accuracy of the clinical diagnosis of alzheimer disease at the national institute on aging alzheimer disease centers, 2005-2010. *Journal of neuropathology & experimental neurology*, 71:566–573, 2012.

[12] M. Bjerke, E. Portelius, M. Lennart, A. Wallin, H. Anckars ater, R. Anckars ater, N. Andreasen, H. Zetterberk, U. Andreassen, and K. Blennow. Confounding factors influencing amyloid beta concentration in cerebrospinal fluid. *International journal of Alzheimer's disease*, 2010.

[13] K. Blennow, B. Dubois, A.M. Fagan, P. Lewczuk, M.J. de Leon, and H. Hampel. Clinical utility of cerebrospinal fluid biomarkers in the diagnosis of early alzheimer's disease. *Alzheimer's & Dementia*, 11(1):58 – 69, 2015.

[14] A.J. Branscum, W.O. Johnson, T.E. Hanson, and I.A. Gardner. Bayesian semiparametric roc curve estimation and disease diagnosis. *Statistics in medicine*, 27:2474–2496, 2008.

[15] L.D. Broemeling. *Advanced Bayesian methods for medical test accuracy.* Chapman & Hall, 2012.

[16] M. Buyse, D.J. Sargent, A. Grothey, A. Matheson, and A. de Gramont. Biomarkers and surrogate end points - the challenge of statistical validation. *Nature review: Clinical oncology*, 7:309–317, 2010.

[17] B.P. Carlin and T.A. Louis. *Bayesian methods for data analysis.* Chapman & Hall, 2009.

[18] Y.K. Choi, W.O. Johnson, M.T. Collins, and I.A. Gardner. Bayesian inference for receiver operating characteristic curves in the absence of a gold standard. *American statistical association and the international biometric society journal of agricultural, biological and environmental statistics*, 11:210–229, 2006.

[19] R. Christensen, W. Johnson, A. Branscum, and T.E. Hanson. *Bayesian ideas and data analysis: An introduction for scientists and statisticians.* Chapman & Hall//CRC, 2011.

[20] R. Craig-Schapiro, A.M. Fagan, and D.M. Holtzman. Biomarkers of alzheimer's disease. *Neurobiology of disease*, 35(2):128–140, 2009.

[21] V.C. Cullen, R.A. Fredenburg, C. Evans, P.R. Conliffe, and M.E. Solomon. Development and advanced validation of an optimized method for the quantitation of aβ_{42} in human cerebrospinal fluid. *The AAPS Journal*, 14(3):510–518, 2012.

[22] Biomarkers definitions working group. Biomarkers and surrogate endpoint: Preferred definitions and conceptual framework. *Clinical pharmacology & therapeutics*, 69:89–95, 2001.

[23] M. del Campo, B. Mollenhauer, A. Bertolotto, S. Engelborghs, H. Hampel, A.H. Simonsen, E. Kapaki, N. Kruse, N. Le Bastard, S. Lehmann, et al. Recommendations to standardize preanalytical confounding factors in alzheimer's and parkinson's disease cerebrospinal fluid biomarkers: an update. *Biomarkers in medicine*, 6(4):419–430, 2012.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society, series B*, 39(1):1–38, 1977.

[25] N. Dendukuri, A. Hadgu, and L. Wang. Modeling conditional dependence between diagnostic tests: A multiple latent variable model. *Statistics in medicine*, 28:441–461, 2009.

[26] N. Dendukuri and L. Joseph. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics*, 57:158–167, 2001.

[27] B. Dubois, H.H. Feldman, C. Jacova, S.T. Dekosky, H. Barberger-Gateau, P. adn Cummings, A. Delacourte, D. Galasko, S. Gauthier, G. Jicha, K. Meguro, J. O'Brien, F. Pqsuier, P. Robert, M. Rossor, S. Salloway, Y. Stern, P. Visser, and P. Scheltens. Research criteria for the diagnosis of alzheimer's disease: revising the nincds-adrda criteria. *Lancet neurology*, 6:734–746, 2007.

[28] B. Dubois, H.H. Feldman, C. Jacova, H. Hampel, J.L. Molinuevo, K. Blennow, S.T. DeKosky, S. Gauthier, D. Selkoe, R. Bateman, S. Cappa, S. Crutch, S. Engelborghs, G.B. Frisoni, N.C. Fox, D. Galasko, M.O. Habert, G.A. Jicha, A. Nordberg, F. Pasquier, G. Rabinovici, P. Robert, C. Rowe, S. Salloway,

M. Sarazin, S. Epelbaum, L.C. de Souza, B. Vellas, P.J. Visser, L. Schneider, Y. Stern, P. Scheltens, and J.L. Cummings. Advancing research diagnostic criteria for alzheimer's disease: the iwg-2 criteria. *The Lancet Neurology*, 13(6):614–629, 2014.

[29] F.H. Duits, C.E. Teunissen, F.H. Bouwman, P.J. Visser, N. Mattsson, H. Zetterber, K. Blennow, O. Hansson, L. Minthon, N. Andreasen, J. Marcusson, A. Wallin, M.O. Rikkert, M. Tsolaki, L. Parnetti, S.K. Herukka, H. Hampel, M.J. De Leon, J. Schröder, D. Aarsland, M.A. Blankenstein, P. Scheltens, and W. van der Flier. The cerebrospinal fluid "alzheimer profile": Easily said, but what does it mean? *Alzheimer's & dementia*, 10(6):713–723, 2014.

[30] A.M. Epstein, C.B. Begg, and B.J. McNeil. The use of ambulatory testing in prepaid and fee-for-service group practices. relation to perceived profitability. *The New England journal of medicine*, 314(17):1089–1094, April 1986.

[31] Alzheimer Europe. The prevalence of dementia in europe. Webpage, 2014. `http://www.alzheimer-europe.org/Policy-in-Practice2/Country-comparisons/The-prevalence-of-dementia-in-Europe`.

[32] D. Faraggi and B. Reiser. Estimation of the area under the roc curve. *Statistics in medicine*, 21:3093–3106, 2002.

[33] R. Fluss, D. Faraggi, and B. Reiser. Estimation of the youden index and its associated cutoff point. *Biometrical Journal*, 47(4):458–472, 2005.

[34] D.G. Fryback and J.R. Thornbury. The efficacy of diagnostic imaging. *Medical decision making*, 11(2):88–94, 1991.

[35] E.S. Garret and S.L. Zeger. Latent class model diagnosis. *Biometrics*, 56:1055–1067, 2000.

[36] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis.* Chapman & Hall., 2004.

[37] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press, 2006.

[38] A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, 7:457–472, 1992.

[39] M.P. Georgiadis, W.O. Johnson, I.A. Gardner, and R. Singh. Correlation adjusted estimation of sensitivity and specificity of two diagnostic tests. *Applied statistics*, 52:63–76, 2003.

[40] J. Geweke. *Bayesian statistics 4*, chapter Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Oxford University Press, 1992.

[41] J. Gu and S. Ghosal. Bayesian roc curve estimation under binormality using a rank likelihood. *Journal of statistical planning and inference*, 139:2076–2083, 2009.

[42] O. Hansson, H. Zetterberg, P. Buchhave, E. Londos, K. Blennow, and L. Minthon. Association between csf biomarkers and incipient alzheimer's disease in patients with mild cognitive impairment: a follow-up study. *The Lancet Neurology*, 5(3):228–234, 2006.

[43] J. Hertze, L. Minthon, H. Zetterberg, E. Vanmechelen, K. Blennow, and O. Hansson. Evaluation of csf biomarkers as predictors of alzheimer's disease: a clinical follow-up study of 4.7 years. *Journal of Alzheimer's Disease*, 21(4):1119–1128, 2010.

[44] X. Huang, G. Qin, and Y. Fang. Optimal combinations of diagnostic tests based on auc. *Biometrics*, 67:568–576, 2011.

[45] J.G. Ibrahim and M.H. Chen. Power prior distributions for regression models. *Statistical Science*, 15(1):46–60.

[46] D.J. Irwin, C.T. McMillan, J.B. Toledo, S.E. Arnold, L.M. Shaw, L.-S. Wang, V. Van Deerlin, V.M.-Y. Lee, J.Q. Trojanowski, and M. Grossman. Comparison of cerebrospinal fluid levels of tau and $a\beta_{1-42}$ in alzheimer disease and frontotemporal degeneration using 2 analytical platforms. *Archives of neurology*, 69(8):1018–1025, 2012.

[47] C.R. Jr. Jack, D.S. Knopman, W.J. Jagust, R.C. Petersen, M.W. Weiner, P.S. Aisen, L.M. Shaw, P. Vemuri, H.J. Wiste, S.D. Weigand, T.G. Lesnick, V.S. Pankratz, M.C. Donohue, and J.Q. Trojanowski. Tracking pathophysiological processes in alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *The Lancet Neurology*, 12(2):207–216, 2013.

[48] G. Jones, W.O. Johnson, T.E. Hanson, and R. Christensen. Identifiability of models for multiple diagnostic testing in the absence of a gold standard. *Biometrics*, 66:855–863, 2010.

[49] W. Jongbloed, M.I. Kester, W.M. van der Flier, R. Veerhuis, P. Scheltens, M.A. Blankenstein, and C.E. Teunissen. Discriminatory and predictive capabilities of enzyme-linked immunosorbent assay and multiplex platforms in a longitudinal alzheimer's disease study. *Alzheimer's & Dementia*, 9(3):276–283, 2013.

[50] L. Joseph, T.W. Gyorkos, and L. Coupal. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American journal of epidemiology*, 141:263–272, 1995.

[51] J-H. Kang, M. Korecka, J.B. Toledo, J.Q. Trojanowski, and L.M. Shaw. Clinical utility and analytical challenges in measurement of cerebrospinal fluid amyloid-β_{1-42} and τ proteins as alzheimer disease biomarkers. *Clinical Chemistry*, 59(6):903–916, 2013.

[52] R.E. Kass, B.P. Carlin, A. Gelman, and R.M. Neal. Markov chain monte carlo in practice: A roundtable discussion. *The American statistician*, 52:93–100, 1998.

[53] M. Korecka, T. Waligorska, M. Figurski, J.B. Toledo, S.E. Arnold, M. Grossman, J.Q. Trojanowski, and L.M. Shaw. Qualification of a surrogate matrix-based absolute quantification method for amyloid-β 42 in human cerebrospinal fluid using 2d uplc-tandem mass spectrometry. *Journal of Alzheimer's Disease*, 41(2):441–451, 2014.

[54] M. Ladouceur, L. Rahme, P. Bélisle, A.N. Scott, K. Schwartzman, and L. Joseph. Modeling continuous diagnostic test data using approximate dirichlet process distributions. *Statistics in medicine*, 30:2648–2662, 2011.

[55] A. Leinenbach, J. Pannee, T. Dülffer, A. Huber, T. Bittner, U. Andreasson, J. Gobom, H. Zetterberg, U. Kobold, E. Portelius, et al. Mass spectrometry–based candidate reference measurement procedure for quantification of amyloid-β in cerebrospinal fluid. *Clinical chemistry*, 60(7):987–994, 2014.

[56] E. Lesaffre and A. Lawson. *Bayesian biostatistics*. John Wiley & Sons, Ltd., 2012.

[57] A. Liu, E.F. Schisterman, and Y. Zhu. On linear combinations of biomarkers to improve diagnostic accuracy. *Statistics in medicine*, 24:37–47, 2005.

[58] C.J. Lloyd. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the american statistical association*, 93(444):1356–1364, 1998.

[59] Y. Lu, N. Dendukuri, I. Schiller, and L. Joseph. A bayesian approach to simultaneously adjusting for verification and reference standard bias in diagnostic test studies. *Statistics in medicine*, 29:2532–2543, 2010.

[60] D. Lunn, D. Spiegelhalter, A. Thomas, and N. Best. The bugs project: Evolution, critique, and future directions. *Statistics in medicine*, 28:3049–3067, 2009.

[61] N.A. MacMillan, C.M. Rotello, and J.O. Miller. The sampling distributions of gaussian roc statistics. *Perception & Psychophisics*, 66(3):406–421, 2004.

[62] A.M. Mathai and S.B. Provost. *Quadratic forms in random variables.* Marcel dekker: inc., 1992.

[63] J. Matilla, J. Koikkalainen, A. Virkki, M. van Gils, G. Waldenmar, H. Soininen, and J. Lötjönen. A disease state fingerprint for evaluation of alzheimer's disease. *Journal of Alzheimer's disease*, 27:163–176, 2011.

[64] N. Mattsson, U. Andreasson, S. Persson, H. Arai, S.D. Batish, S. Bernardini, L. Bocchio-Chiavetto, M.A. Blankenstein, M.C. Carrillo, S. Chalbot, E. Coart, D. Chiasserini, N. Cutler, G. Dahlfors, S. Duller, A.M. Fagan, O. Forlenza, G.B. Frisoni, D. Galasko, D. Galimberti, H. Hampel, A. Handberg, M.T. Heneka, A.Z. Herskovits, S-K. Herukka, D.M. Holtzman, C. Humpel, B.T. Hyman, K. Iqbal, M. Jucker, S.A. Kaeser, E. Kaiser, E. Kapaki, D. Kidd, P. Klivenyi, C.S. Knudsen, M.P. Kummer, J. Lui, A. Lladó, P. Lewczuk, Q-X. Li, R. Martins, C. Masters, J. McAuliffe, M. Mercken, A. Moghekar, J.L. Molinuevo, T.J. Montine, W. Nowatzke, R. O'Brien, M. Otto, G.P. Paraskevas, L. Parnetti, R.C. Petersen, D. Prvulovic, H.P.M. de Reus, R.A. Rissman, E. Scarpini, A. Stefani, H. Soininen, J. Schr oder, L.M. Shaw, A. Skinningsrud, B. Skrogstad, A. Spreer, L. Talib, C. Teunissen, J.Q. Trojanowski, H. Tumani, R.M. Umek, B. Van Broeck, H. Vanderstichele, L. Vecsei, M.M. Verbeek, M. Windisch, H. Zhang, H. Zetterberg, and K. Blennow. The alzheimer's association external quality control program for cerebrospinal fluid biomarkers. *Alzheimer's & Dementia*, 7(4):386 – 395, 2011.

[65] N. Mattsson, H. Zetterberg, O. Hansson, N. Andreasen, L. Parnetti, M. Jonsson, S.-K. Herukka, W.M. van der Flier, M.A. Blankenstein, M. Ewers, et al. Csf biomarkers and incipient alzheimer disease in patients with mild cognitive impairment. *The journal of the American Medical association*, 302(4):385–393, 2009.

[66] G.M. McKhann, D.S. Knopman, H. Chertkow, B.T. Hyman, Jr. Jack, C.R., C.H. Kawas, W.E. Klunk, W.J. Koroshetz, J.J. Manly, R. Mayeux, R.C. Mohs, J.C. Morris, M.N. Rossor, P. Scheltens, M.C. Carillo, B. Thies, S. Weintraub, and C.H. Phelps. The diagnosis of dementia due to alsheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimers & dementia*, 7:263–269, 2011.

[67] G. McLachlen and D. Peel. *Finite mixture models.* Wiley inc., 2004.

[68] European medicines agency. Qualification opinion of alzheimer's disease novel methodologies/biomarkers for bms-708163. Report, 2011. `http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2011/02/WC500102018.pdf`.

[69] J. Menten, M. Boelaert, and E. Lesaffre. Bayesian latent class models with conditionally dependent diagnostic tests: a case study. *Statistics in medicine*, 27:4469–4488, 2008.

[70] C.E. Metz, B.A. Herman, and J.H Shen. Maximum likelihood estimation of receiver operating characteristic (roc) curves from continuously-distributed data. *Statistics in medicine*, 17:1033–1053, 1998.

[71] C.E. Metz and X. Pan. "proper" binormal roc curves: theory and maximum-likelihood estimation. *Journal of mathematical psychology*, 43:1–33, 1999.

[72] Le Bastard N., E. Coart, H. Vanderstichele, E. Vanmechelen, J.J. Martin, and S. Engelborghs. Comparison of two analytical platforms for the clinical qualification of alzheimer's disease biomarkers in pathologically-confirmed dementia. *Journal of Alzheimer's disease*, 33:117–131, 2013.

[73] Mattssonn N., I. Zegers, U. Andreasson, M. Bjerke, M.A. Blankenstein, R. Bowser, M.C. Carrillo, J. Gobom, T. Heath, R. Jenkins, A. Jeromin, J. Kaplow, D. Kidd, O.F. Laterza, A. Lockhart, M.P. Lunn, R.L. Martone, K. Mills, J. Pannee, M. Ratcliffe, L.M. Shaw, A.J. Siman, H. Soares, C.E. Teunissen, M.M. Verbeek, R.M. Umek, H. Vanderstichele, H. Zetterberg, K. Blennow, and E. Portelius. Reference measurement procedures for alzheimer's disease cerebrospinal fluid biomarkers: definitions and approaches with focus on amyloid β42. *Biomarkers in medicine*, 6:409–417, 2012.

[74] B. Neuenschwander, M. Branson, and D.J. Spiegelhalter. A note on the power prior. *Statistics in Medicine*, 28(28):3562–3566, 2009.

[75] N.A. Obuchowski and D.K. McCLISH. Sample size determination for diagnostic accuracy studies involving binormal roc curve indices. *Statistics in medicine*, 16(13):1529–1542, 1997.

[76] A. Olsson, H. Vanderstichele, N. Andreasen, G. De Meyer, A. Wallin, B. Holmberg, L. Rosengren, E. Vanmechelen, and K. Blennow. Simultaneous measurement of beta-amyloi(1-42), total tau, and phosphorylated tau (thr181) in cerebrospinal fluid by the xmap thechnology. *Clinical chemistry*, 51:336–345, 2005.

[77] A.J. O'Malley and H.K. Zou. Bayesian multivariate hierarchical transformation models for roc analysis. *Statistics in medicine*, 25:495–479, 2006.

[78] A.J. O'Malley, K.H. Zou, J.R. Fielding, and C.M.C. Tempany. Bayesian regression methodology for estimating a receiver operating characteristic curve with two radiologic applications: Prostate biopsy and spiral ct of ureteral stones. *Academic radiology*, 8:713–725, 2001.

[79] S. Palmqvist, H. Zetterberg, K. Blennow, S. Vestberg, U. Andreasson, D.J. Brooks, R. Owenius, D. H a gerstr om, P. Wollmer, L. Minthon, and O. Hansson. Accuracy of brain amyloid detection in clinical practice using cerebrospinal fluid β-amyloid 42: A cross-validation study against amyloid positron emission tomography. *Journal of the American medical association: Neurology*, 71(10):1282–1289, 2014.

[80] L. Parnetti, D. Chiasserini, P. Eusebi, D. Giannandrea, G. Bellomo, C. De Carlo, C. Padiglioni, S. Mastrocola, V. Lisetti, and P. Calabresi. Performance of aβ 1-40, aβ 1-42, total tau, and phosphorylated tau as predictors of dementia in a cohort of patients with mild cognitive impairment. *Journal of Alzheimer's Disease*, 29(1):229–238, 2012.

[81] M.S. Pepe. *The statistical evaluation of medical tests for classification and prediction.* Oxford university press, 2004.

[82] M.S. Pepe, T. Cai, and G. Longton. Combining predictors for classification using the erea under the receiver operating characteristic curve. *Biometrics*, 62:221–229, 2006.

[83] M.S. Pepe, R. Etzioni, Z. Feng, J.D. Potter, M.L. Thompson, M. Thornquist, M. Winget, and Y. Yasui. Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute*, 93(14):1054–1061, 2001.

[84] M.S. Pepe and H. Janes. Insights into latent class analysis of diagnostic test performance. *Biostatistics*, 8:474–484, 2007.

[85] M.S. Pepe and M.L. Thompson. Combining diagnostic test results to increase accuracy. *Biostatistics*, 1:123–140, 2000.

[86] N.J. Perkins and E.F. Schisterman. The inconsistency of 'optimal' cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American journal of epidemiology*, 163(7):670–675, 2006.

[87] R.M. Pfeiffer and E. Bura. A model-free approach to combining biomarkers. *Biometrical journal*, 1:123–140, 2008.

[88] Albert P.S. Random effects modeling approaches for estimating roc curves from repeated ordinal tests without a gold standard. *Biometrics*, 63:593–602, 2007.

[89] T. Qu and M. Kutner. Random effects models in latent class analysis for evaluation of accuracy of diagnostic tests. *Biometrics*, 52:797–810, 1996.

[90] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[91] P. Ray, Y. Le Manach, B. Riou, and T.T. Houle. Statistical evaluation of a biomaker. *Anesthesiaology*, 112:1023–1040, 2010.

[92] J.S. Read and the Committee on Pediatric AIDS. Diagnosis of hiv-1 infection in children younger than 18 months in the united states. *Pediatrics*, 120(6):e1547–e1562, 2007.

[93] M. Reid, M.S. Lachs, and A.R. Feinstein. Use of methodological standards in diagnostic test research: Getting better but still not good. *Journal of the american medicial society*, 274(8):645–651, 1995.

[94] D. Renard, H. Geys, G. Molenberghs, T. Burzykowski, and M. Buyse. Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. *Biometrical journal*, 44(8):921–935, 2002.

[95] D. Rindskopf and W. Rindskopf. The value of latent class analysis in medical diagnosis. *Statistics in medicine*, 5:21–27, 1986.

[96] C.P. Robert and C. Soubiran. Estimation of a normal mixture model through gibbs sampling and prior feedback. *Test*, 2:125–146, 1993.

[97] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. Sanchez, and M. Muller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12(77):1–8, 2011.

[98] E. Rosenblueth. Point estimates for probability moments. *Proceedings of the National Academy of Sciences*, 72(10):3812–3814, 1975.

[99] H. Ruben. Probability of content of regions under spherical normal distributions, iv: the distribution of homogeneous and non-homogeneous quadratic functions of normal variables. *The annals of mathematical statistics*, 33:542–570, 1962.

[100] K. Rufibach. A smooth roc curve estimator based on log-concave density estimates. *The internationl journal of biostatistics*, 8(1):1–29, 2012.

[101] S. Salloway, R. Sperling, N.C. Fox, K. Blennow, W. Klunk, M. Raskind, M. Sabbagh, L.S. Honig, A.P. Porsteinsson, S. Ferris, et al. Two phase 3 trials of bapineuzumab in mild-to-moderate alzheimer's disease. *New England Journal of Medicine*, 370(4):322–333, 2014.

[102] P. Scheltens and K. Rockwood. How golden is the gold standard of neuropathology in dementia. *Alzheimer's & dementia*, 7:486–489, 2011.

[103] E.F. Schisterman and N. Perkins. Confidence intervals for the youden index and corresponding optimal cut-point. *Communications in Statistics-Simulation and Computation®*, 36(3):549–563, 2007.

[104] N.S. Schoonenboom, F.E. Reesink, N.A. Verwey, M.I. Kester, C.E. Teunissen, P.M. van de Ven, Y.A.L. Pijnenburg, M.A. Blankenstein, A.J. Rozemuller, P. Scheltens, and W.M. van der Flier. Cerebrospinal fluid markers for differential dementia diagnosis in a large memory clinic cohort. *Neurology*, 78(1):47–54, 2012.

[105] A.N. Scott, L. Joseph, P. Bélisle, M.A. Behr, and K. Schwartzman. Bayesian modeling of tuberculosis clustering from dna fingerprint data. *Statistics in medicine*, 27:140–156, 2007.

[106] J.W. Seaman, J.W. Seaman, and J.D. Stamey. Hidden dangers of specifying noninformative priors. *The American statistician*, 66:77–84, 2012.

[107] L.M. Shaw, H. Vanderstichele, M. Knapik-Czajka, C.M. Clark, P.S. Aisen, R.C. Petersen, K. Blennow, H. Soares, A. Simon, P. Lewczuk, R. Dean, E. Siemers, W. Potter, V.M.-Y. Lee, and J.Q. Trojanowski. Cerebrospinal fluid biomarker

signature in alzheimer's disease neuroimaging initiative subjects. *Annals of Neurology*, 65(4):403–413, 2009.

[108] L.M. Shaw, H. Vanderstichele, M. Knapik-Czajka, M. Figurski, E. Coart, K. Blennow, H. Soares, A.J. Simon, P. Lewczuk, R.A. Dean, E. Siemers, W. Potter, V.M-Y. Lee, and J.Q. Trojanowski. Qualification of the analytical and clinical performance of csf biomarker analyses in adni. *Acta Neuropathologica*, 121(5):597–609, 2011.

[109] R. A. Sperling, P.S. Aisen, L.A. Beckett, D.A. Bennett, S. Craft, A.M. Fagan, T. Iwatsubo, C.R. Jack, J. Kaye, T.J. Montine, et al. Toward defining the preclinical stages of alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & Dementia*, 7(3):280–292, 2011.

[110] D.J. Spiegelhalter, K.R. Abrams, and J.P. Myles. *Evidence Synthesis, in Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* John Wiley & Sons, Ltd., 2003.

[111] S. Sturtz, U. Ligges, and A. Gelman. R2winbugs: A package for running winbugs from r. *Journal of Statistical Software*, 12(3):1–16, 2005.

[112] J.Q. Su and J.S. Liu. Linear combinations of multiple diagnostic markers. *Statistical association*, 88:1350–1355, 1993.

[113] F. Subtil and M. Rabilloud. Estimating the optimal threshold for a diagnostic biomarker in case of complex biomarker distributions. *BMC medical informatics and decision making*, 14(1):53, 2014.

[114] J.A. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240(4857):1285–1293, 1988.

[115] C.E. Teunissen, N.A. Verwey, M.I. Kester, K. van Uffelen, and M.A. Blankenstein. Standardization of assay procedures for analysis of the csf biomarkers amyloid β_{1-42}, tau, and phosphorylated tau in alzheimer's disease: Report of an international workshop. *International journal of Alzheimer's disease*, pages 1–6, 2010.

[116] J.B. Toledo, J. Brettschneider, M. Grossman, S.E. Arnold, W.T. Hu, S.X. Xie, V.M.-Y. Lee, L.M. Shaw, and J.Q. Trojanowski. Csf biomarkers cutoffs: the importance of coincident neuropathological diseases. *Acta Neuropathologica*, 124(1):23–35, 2012.

[117] P.N. Valenstein. Evaluating diagnostic tests with imperfect standards. *American journal of clinical pathology*, 93:252–258, 1990.

[118] S. Van der Mussele, E. Fransen, H. Struyfs, J. Luyckx, P. Mari en, J. Saerens, N. Somers, J. Goeman, P.P. De Deyn, and S. Engelborghs. Depression in mild cognitive impairment is associated with progression to alzheimer's disease: A longitudinal study. *Journal of Alzheimer's disease*, 42:1239–1250, 2014.

[119] H.M. Vanderstichele, L. Shaw, M. Vandijck, A. Jeromin, H. Zetterberg, K. Blennow, C. Teunissen, and S. Engelborghs. Alzheimer disease biomarker testing in cerebrospinal fluid: A method to harmonize assay platforms in the absence of an absolute reference standard. *Clinical Chemistry*, 59(4):710–712, 2013.

[120] J. Vandeurzen. Naar een dementievriendelijk vlaanderen. Webpage, 2010. `http://www.jovandeurzen.be/sites/jvandeurzen/files/dementieplan2010-2014.pdf`.

[121] S.J.B. Vos, C. Xiong, P.J. Visser, M.S. Jasielec, J. Hassentab, E.A. Grant, N.J. Cairns, J.C. Morris, D.M. Holtzman, and A.M. Fagan. Preclinical alzheimer's disease and its outcome: a longitudinal cohort study. *Lancet neurology*, 12:657–965, 2013.

[122] A. Waldman, A. Ghezzi, A. Bar-Or, Y. Mikaeloff, M. Tardieu, and B. Banwell. Multiple sclerosis in children: an update on clinical diagnosis, therapeutic strategies, and research. *Lancet neurology*, 13:936–948, 2014.

[123] C. Wang, B.W. Turnbull, Y.T. Gröhn, and S.S. Nielsen. Estimating receiver operating characteristic curves with covariates when there is no perfect reference test for diagnosis of johne's disease. *Journal of dairy science*, 89:3038–3046, 2006.

[124] L.-S. Wang, Y.Y. Leung, S.-K. Chang, S. Leight, M. Knapik-Czajka, Y. Baek, L.M. Shaw, V.M-Y Lee, J.Q. Trojanowski, and C.M. Clark. Comparison of xmap and elisa assays for detecting cerebrospinal fluid biomarkers of alzheimer's disease. *Journal of Alzheimer's Disease*, 31(2):439–445, 2012.

[125] Y. Wei and P.T. Higgins. Bayesian multivariate meta-analysis with multiple outcomes. *Statistics in medicine*, 32:2911–2934, 2013.

[126] D.E. Wollman and I. Prohovnik. Sensitivity and specificity of neuroimiging for the diagnosis of alzheimer's disease. *Diaglogues in clinical neuroscience*, 5:89–99, 2003.

[127] H. Xu and B.A. Craig. A probit latent class model with general correlation structures for evaluating accuracy of diagnostic tests. *Biometrics*, 65:1145–1155, 2009.

[128] I. Yang and M.P. Becker. Latent variable modeling of diagnostic accuracy. *Biometrics*, 53:948–958, 1997.

[129] L. Yang, D. Rieves, and C. Ganley. Brain amyloid imaging — fda approval of florbetapir f18 injection. *New England Journal of Medicine*, 367(10):885–887, 2012. PMID: 22931256.

[130] W.J. Youden. Index for rating diagnostic tests. *Cancer*, 3:32–35, 1950.

[131] B. Yu, C. Zhou, and S. Bandinelli. Combining multiple continuous tests for the diagnosis of kidney impairment in the absence of a gold standard. *Statistics in medicine*, 30:1712–1721, 2011.

[132] X.H. Zhou and J. Harezlak. Comparison of bandwidth selection methods for kernel smoothing of roc curves. *Statistics in medicine*, 21:2045–2055, 2002.

[133] X.H. Zhou, N. Obuchowski, and D.K. McClish. *Statistical methods in diagnostic medicine*. John Wiley & Sons, Inc., Hoboken, New Jersey: USA, 2011.

[134] K.H. Zou, W.J. Hall, and D.E. Shapiro. Smooth non-parametric receiver operating characteristic (roc) curves for continuous diagnostic tests. *Statistics in medicine*, 16:2143–2156, 1997.

[135] K.H. Zou, A. Liu, A.I. Bandos, and H.E. Ohno-Machado. *Statistical evaluation of diagnostic performance: Topics in ROC analysis*. Chapman & Hall, Boca Raton, New York: USA, 2012.

[136] M.H. Zweig and G. Campbell. Receiver-operating characteristic (roc) plots: A fundamental evaluation tool in clinical medicine. *Clinical chemistry*, 39(4):561–577, 1993.

Appendix A

R Codes

A.1 BUGS model for the Bayesian latent-class model assuming conditional independence

The following code relates to the models developed in Sections 3.2.1 and 5.2.1 of Chapters 3 and 5, respectively.

```
model{
  ## Hyperpriors
  # Prevalence
  theta ~ dunif(Trunc_Par[1],Trunc_Par[2]);

  ### Priors
  ## Latent true disease status
  for(i in 1:N){
    D[i] ~ dbern(theta);
    D_ind[i] <- D[i] + 1;
  }

  ## Precision matrices
  for(Prec.index in 1:2){
    # Cholesky decomposition of correlation matrix
    L[Prec.index,1,1] <- 1;

    L[Prec.index,1,2] ~ dunif(-1,1);
    L[Prec.index,1,3] ~ dunif(-1,1);

    Lim[Prec.index] <- sqrt(1-(pow(L[Prec.index,1,3],2)));
    L23[Prec.index] ~ dunif(-1,1);
    L[Prec.index,2,3] <- Lim[Prec.index]*L23[Prec.index];

    L[Prec.index,2,2] <- sqrt(1-pow(L[Prec.index,1,2],2));
    L[Prec.index,3,3] <- sqrt(1-((pow(L[Prec.index,1,3],2))+(pow(L[Prec.index,2,3],2))));
```

```
  L[Prec.index,2,1] <- 0;
  L[Prec.index,3,1] <- 0;
  L[Prec.index,3,2] <- 0;

  for(k1 in 1:K){
    for(k2 in 1:K){
      R[Prec.index,k1,k2] <- inprod(L[Prec.index,1:K,k1],L[Prec.index,1:K,k2]);
    }
  }

  # Standard deviation matrix
  for(k in 1:K){
    Sd[Prec.index,k] ~ dunif(0,1000);
  }

  # Definition of Variance covariance matrix Sigma
  for(k1 in 1:K){
    for(k2 in 1:K){
      Sigma[Prec.index,k1,k2] <- (equals(k1,k2) * pow(Sd[Prec.index,k1],2))
       + ((1-equals(k1,k2))*Sd[Prec.index,k1]*R[Prec.index,k1,k2]*Sd[Prec.index,k2]);
    }
  }

  # Define precision matrix
  Prec[Prec.index,1:K,1:K] <- inverse(Sigma[Prec.index,1:K,1:K]);
}

## Scaled difference vector
ScD[1:K] ~ dmnorm(Kappa[1:K],Psy_Inv[1:K,1:K]);

## Cholesky decomposition of inverse of pooled variance covariance matrix
# Pooled Variance Covariance matrix
for(k1 in 1:K){
  for(k2 in 1:K){
    Sigma_Pooled[k1,k2] <- Sigma[1,k1,k2] + Sigma[2,k1,k2];
  }
}

# Inverse of pooled variance covariance matrix
Sigma_Pooled_Inv[1:K,1:K] <- inverse(Sigma_Pooled[1:K,1:K]);

# Cholesky decomposition of inverse of pooled variance covariance matrix
L2[1,1] <- sqrt(Sigma_Pooled_Inv[1,1]);
L2[2,1] <- (1/L2[1,1]) * Sigma_Pooled_Inv[2,1];
L2[3,1] <- (1/L2[1,1]) * Sigma_Pooled_Inv[3,1];
L2[1,2] <- 0;
L2[2,2] <- sqrt(Sigma_Pooled_Inv[2,2]-pow(L2[2,1],2));
L2[3,2] <- (1/L2[2,2]) * (Sigma_Pooled_Inv[3,2] - (L2[3,1]*L2[2,1]));
L2[1,3] <- 0;
L2[2,3] <- 0;
L2[3,3] <- sqrt(Sigma_Pooled_Inv[3,3]-(pow(L2[3,1],2)+pow(L2[3,2],2)));

# Inverse of cholesky factor of inverse of pooled variance covariance matrix
L2_Inv[1:K,1:K] <- inverse(L2[1:K,1:K]);
```

```
  ## Normal component means
  for(k in 1:K){
    Mu[k,1] ~ dnorm(0,1.E-6);
    Mu[k,2] <- inprod(L2_Inv[1:K,k],ScD[1:K]) + Mu[k,1];
  }

  ## Clinical diagnosis parameters
  Se ~ dbeta(Se_Prior[1],Se_Prior[2])T(0.51,);
  Sp ~ dbeta(Sp_Prior[1],Sp_Prior[2])T(0.51,);

  ## Likelihood
  for(i in 1:N){
    # Continuous Biomarker part
    Y[i,1:K] ~ dmnorm(mean[i,],Prec[D_ind[i],,]);
    for(k in 1:K){
      mean[i,k] <- (Mu[k,1] * (2-D_ind[i])) + (Mu[k,2] * (D_ind[i]-1));
    }

    # Clinical diagnosis part
    T[i] ~ dbern(ProbT[i]);
    ProbT[i] <- (Se * (D_ind[i]-1)) + ((1-Sp) * (2-D_ind[i]));
  }

  # Biomarker performance measure AUC
  for(k in 1:K){
    Mu_diff[k] <- Mu[k,2] - Mu[k,1];
    a[k] <- inprod(Sigma_Pooled_Inv[k,1:K],Mu_diff[1:K]);
  }

  AUC <- phi(pow(inprod(a[1:K],Mu_diff[1:K]),0.5));
}
```

A.2 BUGS model for the Bayesian latent-class model allowing for conditional dependence

The represented code relates to the model developed in Section 4.2 of Chapter 4.

```
model{
  #### Hyperpriors
  ### Prevalence
  theta ~ dunif(Trunc_Par[1],Trunc_Par[2]);


  #### Priors
  ### Latent true disease status
  for(i in 1:N){
    D[i] ~ dbern(theta);
    D_ind[i] <- D[i] + 1;
  }

  ### Latent tolerance distribution parameters
  ## Means
```

```
# Beta Se/Sp priors
dummy_Mu1 <- 0;
dummy_Mu1 ~ dloglik(logLike_Mu1);
logLike_Mu1 <- loggam(Sp_Prior[1] + Sp_Prior[2]) - loggam(Sp_Prior[1])
 - loggam(Sp_Prior[2]) + log(pow(phi(-MuT[1]),(Sp_Prior[1]-1))
 * pow((1-phi(-MuT[1])),(Sp_Prior[2]-1)) * abs(exp(-0.5*pow(MuT[1],2))/sqrt(2*3.14)));
MuT[1] ~ dflat()T(,-0.02506891);

dummy_Mu2 <- 0;
dummy_Mu2 ~ dloglik(logLike_Mu2);
logLike_Mu2 <- loggam(Se_Prior[1] + Se_Prior[2]) - loggam(Se_Prior[1])
 - loggam(Se_Prior[2]) + log(pow((1-phi(-MuT[2])),(Se_Prior[1]-1))
 * pow((1-(1-phi(-MuT[2]))),(Se_Prior[2]-1)) * abs(exp(-0.5*pow(MuT[2],2))/sqrt(2*3.14)));
MuT[2] ~ dflat()T(0.02506891,);

### Overall distribution parameters
## Standard deviations
Sd[1,1] <- 1;                    # Latent Tolerance
Sd[2,1] <- 1;
Sd[1,2] ~ dunif(0,1000);         # Biomarker 1
Sd[2,2] ~ dunif(0,1000);
Sd[1,3] ~ dunif(0,1000);         # Biomarker 2
Sd[2,3] ~ dunif(0,1000);
Sd[1,4] ~ dunif(0,1000);         # Biomarker 3
Sd[2,4] ~ dunif(0,1000);

## Cholesky decomposition of correlation-matrix
L[1,1,1] <- 1;
L[2,1,1] <- 1;

# Correlations
L[1,1,2] ~ dunif(-1,1);
L[2,1,2] ~ dunif(-1,1);
L[1,1,3] ~ dunif(-1,1);
L[2,1,3] ~ dunif(-1,1);
L[1,1,4] ~ dunif(-1,1);
L[2,1,4] ~ dunif(-1,1);

# L23
Lim1[1] <- sqrt(1-(pow(L[1,1,3],2)));
Lim1[2] <- sqrt(1-(pow(L[2,1,3],2)));

L23[1] ~ dunif(-1,1);
L23[2] ~ dunif(-1,1);

L[1,2,3] <- Lim1[1]*L23[1];
L[2,2,3] <- Lim1[2]*L23[2];

# L24
Lim2[1] <- sqrt(1-(pow(L[1,1,4],2)));
Lim2[2] <- sqrt(1-(pow(L[2,1,4],2)));

L24[1] ~ dunif(-1,1);
L24[2] ~ dunif(-1,1);
```

```
L[1,2,4] <- Lim2[1]*L24[1];
L[2,2,4] <- Lim2[2]*L24[2];

# L34
Lim3[1] <- sqrt(1-(pow(L[1,1,4],2) + pow(L[1,2,4],2)));
Lim3[2] <- sqrt(1-(pow(L[2,1,4],2) + pow(L[2,2,4],2)));

L34[1] ~ dunif(-1,1);
L34[2] ~ dunif(-1,1);

L[1,3,4] <- Lim3[1]*L34[1];
L[2,3,4] <- Lim3[2]*L34[2];


L[1,2,2] <- sqrt(1-pow(L[1,1,2],2));
L[2,2,2] <- sqrt(1-pow(L[2,1,2],2));

L[1,3,3] <- sqrt(1-((pow(L[1,1,3],2))+(pow(L[1,2,3],2))));
L[2,3,3] <- sqrt(1-((pow(L[2,1,3],2))+(pow(L[2,2,3],2))));

L[1,4,4] <- sqrt(1-((pow(L[1,1,4],2))+(pow(L[1,2,4],2))+(pow(L[1,3,4],2))));
L[2,4,4] <- sqrt(1-((pow(L[2,1,4],2))+(pow(L[2,2,4],2))+(pow(L[2,3,4],2))));


L[1,2,1] <- 0;
L[2,2,1] <- 0;
L[1,3,1] <- 0;
L[2,3,1] <- 0;
L[1,3,2] <- 0;
L[2,3,2] <- 0;
L[1,4,1] <- 0;
L[2,4,1] <- 0;
L[1,4,2] <- 0;
L[2,4,2] <- 0;
L[1,4,3] <- 0;
L[2,4,3] <- 0;

# Recreate Overall correlation matrix
for(k1 in 1:K){
  for(k2 in 1:K){
    R[1,k1,k2] <- inprod(L[1,1:K,k1],L[1,1:K,k2]);
    R[2,k1,k2] <- inprod(L[2,1:K,k1],L[2,1:K,k2]);
  }
}

# Overall Variance-Covariance matrix Sigma
for(k1 in 1:K){
  for(k2 in 1:K){
    Sigma[1,k1,k2] <- (equals(k1,k2) * pow(Sd[1,k1],2))
     + ((1-equals(k1,k2))*Sd[1,k1]*R[1,k1,k2]*Sd[1,k2]);
    Sigma[2,k1,k2] <- (equals(k1,k2) * pow(Sd[2,k1],2))
     + ((1-equals(k1,k2))*Sd[2,k1]*R[2,k1,k2]*Sd[2,k2]);
  }
}
```

```
# Define overall precision matrix
Prec[1,1:K,1:K] <- inverse(Sigma[1,1:K,1:K]);
Prec[2,1:K,1:K] <- inverse(Sigma[2,1:K,1:K]);

### Biomarker distribution parameters
## Means
# Scaled difference vector
ScD[1:(K-1)] ~ dmnorm(Kappa[1:(K-1)],Psy_Inv[1:(K-1),1:(K-1)]);

## Cholesky decomposition of inverse of pooled biomarker variance-covariance matrix SigmaX
# Biomarker variance-covariance matrix SigmaX
for(k1 in 2:K){
  for(k2 in 2:K){
    SigmaX[1,k1-1,k2-1] <- (equals(k1,k2) * pow(Sd[1,k1],2))
     + ((1-equals(k1,k2))*Sd[1,k1]*R[1,k1,k2]*Sd[1,k2]);
    SigmaX[2,k1-1,k2-1] <- (equals(k1,k2) * pow(Sd[2,k1],2))
     + ((1-equals(k1,k2))*Sd[2,k1]*R[2,k1,k2]*Sd[2,k2]);
  }
}

# Define Biomarker precision matrix
PrecX[1,1:(K-1),1:(K-1)] <- inverse(SigmaX[1,1:(K-1),1:(K-1)]);
PrecX[2,1:(K-1),1:(K-1)] <- inverse(SigmaX[2,1:(K-1),1:(K-1)]);

# Pooled Biomarker Variance-Covariance matrix
for(k1 in 1:(K-1)){
  for(k2 in 1:(K-1)){
    SigmaX_Pooled[k1,k2] <- SigmaX[1,k1,k2] + SigmaX[2,k1,k2];
  }
}

# Inverse of pooled Biomarker variance-covariance matrix
SigmaX_Pooled_Inv[1:(K-1),1:(K-1)] <- inverse(SigmaX_Pooled[1:(K-1),1:(K-1)]);

# Cholesky decomposition of inverse of pooled Biomarker variance-covariance matrix
L2[1,1] <- sqrt(SigmaX_Pooled_Inv[1,1]);
L2[2,1] <- (1/L2[1,1]) * SigmaX_Pooled_Inv[2,1];
L2[3,1] <- (1/L2[1,1]) * SigmaX_Pooled_Inv[3,1];
L2[1,2] <- 0;
L2[2,2] <- sqrt(SigmaX_Pooled_Inv[2,2]-pow(L2[2,1],2));
L2[3,2] <- (1/L2[2,2]) * (SigmaX_Pooled_Inv[3,2] - (L2[3,1]*L2[2,1]));
L2[1,3] <- 0;
L2[2,3] <- 0;
L2[3,3] <- sqrt(SigmaX_Pooled_Inv[3,3]-(pow(L2[3,1],2)+pow(L2[3,2],2)));

# Inverse of cholesky factor of inverse of pooled Biomarker variance-covariance matrix
L2_Inv[1:(K-1),1:(K-1)] <- inverse(L2[1:(K-1),1:(K-1)]);

# Biomarker Means
for(k in 1:(K-1)){
  MuX[1,k] ~ dnorm(0,1.E-6);
  MuX[2,k] <- inprod(L2_Inv[1:(K-1),k],ScD[1:(K-1)]) + MuX[1,k];
}

### Remaining shared latent tolerance and biomarker Sigma entries
```

```
  ## Latent tolerance - biomarker covariance vector
  Tau[1,1] <- Sigma[1,1,2];
  Tau[2,1] <- Sigma[2,1,2];

  Tau[1,2] <- Sigma[1,1,3];
  Tau[2,2] <- Sigma[2,1,3];

  Tau[1,3] <- Sigma[1,1,4];
  Tau[2,3] <- Sigma[2,1,4];

  # Latent-tolerance covariance - biomarker precision product
  for(k in 1:(K-1)){
    Cov_Prec_Prod[1,k] <- inprod(Tau[1,1:(K-1)],PrecX[1,1:(K-1),k]);
    Cov_Prec_Prod[2,k] <- inprod(Tau[2,1:(K-1)],PrecX[2,1:(K-1),k]);
  }

  Cov_Prec_Cov_Prod[1] <- inprod(Cov_Prec_Prod[1,1:(K-1)],Tau[1,1:(K-1)]);
  Cov_Prec_Cov_Prod[2] <- inprod(Cov_Prec_Prod[2,1:(K-1)],Tau[2,1:(K-1)]);


  #### Likelihood
  for(i in 1:N){
    ### Continuous Biomarker part
    Y[i,1:(K-1)] ~ dmnorm(MuX[D_ind[i],],PrecX[D_ind[i],,]);

    ### Clinical diagnosis part
    T[i] ~ dbern(ProbT[i]);

    for(k in 1:(K-1)){
      Obs_Mean_Diff[i,k] <- Y[i,k] - MuX[D_ind[i],k];
    }

    Cov_Prec_Cen_Mean[i] <- inprod(Cov_Prec_Prod[D_ind[i],1:(K-1)],Obs_Mean_Diff[i,1:(K-1)]);

    ProbT[i] <- 1-phi(-(MuT[D_ind[i]] + Cov_Prec_Cen_Mean[i]) / sqrt(1-Cov_Prec_Cov_Prod[D_ind[i]]));
  }

  ### Parameter transformations of interest
  # Reference test Se and Sp
  Se <- 1-phi(-MuT[2]);
  Sp <- phi(-MuT[1]);

  # Biomarker performance measure AUC and linear coefficients a
  for(k in 1:(K-1)){
    Mu_diff[k] <- MuX[2,k] - MuX[1,k];
    a[k] <- inprod(SigmaX_Pooled_Inv[k,1:(K-1)],Mu_diff[1:(K-1)]);
  }

  AUC <- phi(pow(inprod(a[1:(K-1)],Mu_diff[1:(K-1)]),0.5));
}
```

A.3 BUGS model for the Bayesian latent-class model of a validation study under the conditional independence assumption

The following code relates to the model developed in Section 5.2.2 of Chapter 5.

```
model{
  ### Priors
  ## Precisions
  # Standard deviations
    for(k in 1:2){
      Sd[k] ~ dunif(0,1000);
      Prec[k] <- pow(Sd[k],-2);
      Var[k] <- pow(Sd[k],2);
    }

  sum_var <- Var[1] + Var[2];
  pooled_var <- pow(sum_var,0.5);

  ## Scaled difference vector
  AUC_star ~ dnorm(mean_star,prec_star);

  ## Normal component means
  Mu[1] ~ dnorm(0,1.E-6);
  Mu[2] <- (AUC_star*pooled_var) + Mu[1];

  ## Creating indicator
  for(i in 1:N){
    D_ind[i] <- D[i] + 1;
  }

  ## Likelihood
  for(i in 1:N){
    # Continuous Biomarker part
    Y[i] ~ dnorm(mean[i],Prec[D_ind[i]]);
    mean[i] <- (Mu[1] * (2-D_ind[i])) + (Mu[2] * (D_ind[i]-1));
  }

  # Biomarker performance measure AUC
  AUC <- phi(AUC_star);
}
```

A.4 BUGS model for the Bayesian two-stage approach to estimate the optimal new-assay cut-off

The represented code relates to the model developed in Section 6.2.3 of Chapter 6.

```
## STAGE 1
```

```
model{
  ## Priors
  # Standard deviation matrix
  for(Sd.index in 1:2){
    Sd[Sd.index] ~ dunif(0,1000);
    Sigma[Sd.index] <- pow(Sd[Sd.index],2);
    Prec[Sd.index] <- pow(Sd[Sd.index],-2);
  }

  ## AUCst parameterization priors
  # Priors on individual AUCs of platforms
  AUCst ~ dnorm(0,1);

  ## Normal component means
  Mu[1] ~ dnorm(0,1.E-6);
  Mu[2] <- (AUCst * sqrt(Sigma[1] + Sigma[2])) + Mu[1];

  ## Likelihood
  # Validation data set
  for(i in 1:Nc){
    D_ind[i] <- Dc[i] + 1;
    Y_c[i] ~ dnorm(Mu[D_ind[i]],Prec[D_ind[i]]);
  }

  # Calculation of AUC
  AUC <- phi(AUCst);
}

### STAGE 2
model{
  ### Priors
  ## Prevalence Hyperprior
  theta ~ dunif(Trunc_Par[1],Trunc_Par[2]);

  ## Latent true disease status
  for(i in 1:Nt){
    D[i] ~ dbern(theta);                  # Prevalance parameter for complete D-vector
    D_ind[i] <- D[i] + 1;                 # Create indicator 1-2
  }

  ## Precision matrices
  # Standard deviation matrix [1st platform]
  Sd[1,1] ~ dnorm(Pr_Mu_SdC0,Pr_Tau_SdC0)T(0,);
  Sd[2,1] ~ dnorm(Pr_Mu_SdC1,Pr_Tau_SdC1)T(0,);

  for(Prec.index in 1:2){
    # Cholesky decomposition of correlation matrix
    L[Prec.index,1,1] <- 1;
    L[Prec.index,1,2] ~ dunif(-1,1);      # Prior on correlation
    L[Prec.index,2,2] <- sqrt(1-pow(L[Prec.index,1,2],2));
    L[Prec.index,2,1] <- 0;

    # Construct correlation matrix from cholesky factors
    for(k1 in 1:K){
      for(k2 in 1:K){
```

```
      R[Prec.index,k1,k2] <- inprod(L[Prec.index,1:K,k1],L[Prec.index,1:K,k2]);
    }
  }

  # Standard deviation matrix [2nd platform]
  Sd[Prec.index,2] ~ dunif(0,1000);

  # Definition of Variance covariance matrix Sigma through multiplication by Sd and R
  for(k1 in 1:K){
    for(k2 in 1:K){
      Sigma[Prec.index,k1,k2] <- (equals(k1,k2) * pow(Sd[Prec.index,k1],2))
       + ((1-equals(k1,k2))*Sd[Prec.index,k1]*R[Prec.index,k1,k2]*Sd[Prec.index,k2]);
    }
  }

  # Define precision matrix
  Prec[Prec.index,1:K,1:K] <- inverse(Sigma[Prec.index,1:K,1:K]);
}

## AUCst parameterization priors
# Priors on individual AUCs of platforms
AUCst[1] ~ dnorm(Pr_Mu_AUCst,Pr_Tau_AUCst);
AUCst[2] ~ dnorm(0,1);

## Normal component means
Mu[1,1] ~ dnorm(Pr_Mu_MuC0,Pr_Tau_MuC0);
Mu[1,2] ~ dnorm(0,1.E-6);

for(k in 1:K){
  Mu[2,k] <- (AUCst[k] * sqrt(Sigma[1,k,k] + Sigma[2,k,k])) + Mu[1,k];
}

## Likelihood
# Translation data set
for(i in 1:Nt){
  Y_t[i,1:K] ~ dmnorm(Mu[D_ind[i],],Prec[D_ind[i],,]);
}

# Calculation of AUC
AUC[1] <- phi(AUCst[1]);
AUC[2] <- phi(AUCst[2]);
}
```

Appendix B

Simulation results

In the following tables, the results are shown for the remaining parameters (Se, Sp, and θ) estimated in the simulation study described in Section 3.4 of Chapter 3. In this chapter the results for AUC_a are discussed in Section 3.6.

Sensitivity (Se)

$N = 100$

Table B.1: Mean of posterior Se medians with corresponding (standard deviation of posterior Se medians) based on [number of converged data sets] for the simulated data sets of size $N = 100$. FLAT — the analysis using the flat prior for sensitivity and specificity of the reference test; INF — the analysis using the informative prior for sensitivity and specificity of the reference test (see Figure 3.8). Results obtained with the 'naïve', 'conservative' (Cons.), and 'optimistic' (Opt.) AUC_a prior.

Data Gen. Model		**Se/Sp**	**True**	$\boldsymbol{AUC_a - Prior}$		
VarCov	**Corr**	**prior**	**AUC**	**Naïve**	**Cons.**	**Opt.**
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.85	0.779 (0.078) [95]	0.826 (0.062) [68]	0.810 (0.067) [79]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.85	0.835 (0.057) [97]	0.854 (0.043) [85]	0.845 (0.049) [92]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.85	0.723 (0.082) [83]	0.762 (0.098) [16]	0.786 (0.094) [16]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.85	0.806 (0.069) [94]	0.841 (0.061) [35]	0.832 (0.063) [54]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.85	0.800 (0.062) [95]	0.815 (0.055) [93]	0.815 (0.056) [95]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.85	0.836 (0.050) [99]	0.854 (0.040) [98]	0.853 (0.041) [97]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.85	0.778 (0.077) [96]	0.823 (0.695) [66]	0.804 (0.067) [84]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.85	0.831 (0.056) [99]	0.859 (0.045) [84]	0.851 (0.046) [96]

$N = 400$

Table B.2: Mean of posterior Se medians with corresponding (standard deviation of posterior Se medians) based on [number of converged data sets] for the simulated data sets of size $N = 400$. FLAT — the analysis using the flat prior for sensitivity and specificity of the reference test; INF — the analysis using the informative prior for sensitivity and specificity of the reference test (see Figure 3.8). Results obtained with the 'naïve', 'conservative' (Cons.), and 'optimistic' (Opt.) AUC_a prior.

Data Gen. Model		**Se/Sp**	**True**	$\boldsymbol{AUC_a - Prior}$		
VarCov	**Corr**	**prior**	**AUC**	**Naïve**	**Cons.**	**Opt.**
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.85	0.842 (0.046) [100]	0.858 (0.045) [100]	0.850 (0.045) [100]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.85	0.856 (0.038) [100]	0.873 (0.037) [100]	0.865 (0.037) [100]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.85	0.806 (0.084) [93]	0.874 (0.054) [61]	0.847 (0.068) [87]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.85	0.842 (0.064) [98]	0.880 (0.043) [88]	0.863 (0.051) [97]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.85	0.838 (0.033) [100]	0.845 (0.032) [100]	0.842 (0.032) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.85	0.848 (0.030) [100]	0.854 (0.029) [100]	0.853 (0.029) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.85	0.837 (0.053) [100]	0.852 (0.046) [100]	0.840 (0.052) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.85	0.848 (0.048) [100]	0.856 (0.045) [100]	0.852 (0.046) [100]

$N = 600$

Table B.3: Mean of posterior Se medians with corresponding (standard deviation of posterior Se medians) based on [number of converged data sets] for the simulated data sets of size $N = 600$. FLAT — the analysis using the flat prior for sensitivity and specificity of the reference test; INF — the analysis using the informative prior for sensitivity and specificity of the reference test (see Figure 3.8). Results obtained with the 'naïve', 'conservative' (Cons.), and 'optimistic' (Opt.) AUC_a prior.

Data Gen. Model		**Se/Sp**	**True**	$\boldsymbol{AUC_a - Prior}$		
VarCov	**Corr**	**prior**	**AUC**	**Naïve**	**Cons.**	**Opt.**
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.85	0.833 (0.043) [100]	0.51 (0.044) [99]	0.842 (0.044) [100]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.85	0.846 (0.038) [100]	0.859 (0.038) [100]	0.852 (0.038) [100]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.85	0.803 (0.059) [88]	0.855 (0.053) [66]	0.832 (0.057) [92]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.85	0.818 (0.050) [89]	0.875 (0.041) [74]	0.850 (0.044) [99]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.85	0.837 (0.026) [100]	0.841 (0.026) [100]	0.838 (0.026) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.85	0.842 (0.024) [100]	0.846 (0.024) [100]	0.845 (0.024) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.85	0.833 (0.026) [100]	0.847 (0.037) [100]	0.841 (0.037) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.85	0.845 (0.033) [100]	0.854 (0.033) [100]	0.851 (0.033) [100]

Specificity (Sp)

$N = 100$

Table B.4: Mean of posterior Sp medians with corresponding (standard deviation of posterior Sp medians) based on [number of converged data sets] for the simulated data sets of size $N = 100$. FLAT — the analysis using the flat prior for sensitivity and specificity of the reference test; INF — the analysis using the informative prior for sensitivity and specificity of the reference test (see Figure 3.8). Results obtained with the 'naïve', 'conservative' (Cons.), and 'optimistic' (Opt.) AUC_a prior.

Data Gen. Model		**Se/Sp**	**True**	$\boldsymbol{AUC_a - Prior}$		
VarCov	**Corr**	**prior**	**AUC**	**Naïve**	**Cons.**	**Opt.**
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.85	0.776 (0.086) [95]	0.839 (0.080) [68]	0.826 (0.078) [79]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.85	0.839 (0.059) [97]	0.864 (0.051) [85]	0.856 (0.053) [92]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.85	0.684 (0.096) [83]	0.754 (0.101) [16]	0.738 (0.105) [16]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.85	0.779 (0.081) [94]	0.848 (0.087) [35]	0.842 (0.078) [54]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.85	0.796 (0.075) [95]	0.815 (0.068) [93]	0.810 (0.072) [95]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.85	0.833 (0.058) [99]	0.849 (0.050) [98]	0.848 (0.051) [97]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.85	0.760 (0.083) [96]	0.787 (0.070) [66]	0.784 (0.074) [84]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.85	0.820 (0.063) [99]	0.842 (0.055) [84]	0.839 (0.056) [96]

$N = 400$

Table B.5: Mean of posterior Sp medians with corresponding (standard deviation of posterior Sp medians) based on [number of converged data sets] for the simulated data sets of size $N = 400$. FLAT — the analysis using the flat prior for sensitivity and specificity of the reference test; INF — the analysis using the informative prior for sensitivity and specificity of the reference test (see Figure 3.8). Results obtained with the 'naïve', 'conservative' (Cons.), and 'optimistic' (Opt.) AUC_a prior.

Data Gen. Model		**Se/Sp**	**True**	$\boldsymbol{AUC_a - Prior}$		
VarCov	**Corr**	**prior**	**AUC**	**Naïve**	**Cons.**	**Opt.**
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.85	0.842 (0.055) [100]	0.858 (0.053) [100]	0.855 (0.053) [100]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.85	0.857 (0.046) [100]	0.867 (0.043) [100]	0.866 (0.044) [100]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.850	0.801 (0.072) [93]	0.868 (0.070) [61]	0.845 (0.066) [87]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.85	0.829 (0.058) [98]	0.874 (0.042) [88]	0.860 (0.053) [97]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.85	0.842 (0.038) [100]	0.849 (0.038) [100]	0.0848 (0.038) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.85	0.852 (0.034) [100]	0.859 (0.034) [100]	0.856 (0.034) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.85	0.838 (0.046) [100]	0.849 (0.043) [100]	0.847 (0.044) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.85	0.849 (0.040) [100]	0.862 (0.038) [100]	0.859 (0.038) [100]

$N = 600$

Table B.6: Mean of posterior Sp medians with corresponding (standard deviation of posterior Sp medians) based on [number of converged data sets] for the simulated data sets of size $N = 600$. FLAT — the analysis using the flat prior for sensitivity and specificity of the reference test; INF — the analysis using the informative prior for sensitivity and specificity of the reference test (see Figure 3.8). Results obtained with the 'naïve', 'conservative' (Cons.), and 'optimistic' (Opt.) AUC_a prior.

Data Gen. Model		**Se/Sp**	**True**	$\boldsymbol{AUC_a - Prior}$		
VarCov	**Corr**	**prior**	**AUC**	**Naïve**	**Cons.**	**Opt.**
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.85	0.845 (0.042) [100]	0.861 (0.042) [99]	0.852 (0.042) [100]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.85	0.857 (0.037) [100]	0.868 (0.036) [100]	0.864 (0.036) [100]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.850	0.807 (0.070) [88]	0.884 (0.042) [66]	0.846 (0.060) [92]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.85	0.850 (0.054) [89]	0.889 (0.032) [74]	0.864 (0.042) [99]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.85	0.845 (0.031) [100]	0.850 (0.031) [100]	0.0846 (0.031) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.85	0.852 (0.029) [100]	0.857 (0.029) [100]	0.856 (0.029) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.85	0.844 (0.032) [100]	0.849 (0.031) [100]	0.846 (0.031) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.85	0.851 (0.029) [100]	0.857 (0.029) [100]	0.854 (0.029) [100]

Prevalence (θ)

$N = 100$

Table B.7: Mean of posterior θ medians with corresponding (standard deviation of posterior θ medians) based on [number of converged data sets] for the simulated data sets of size $N = 100$. FLAT — the analysis using the flat prior for sensitivity and specificity of the reference test; INF — the analysis using the informative prior for sensitivity and specificity of the reference test (see Figure 3.8). Results obtained with the 'naïve', 'conservative' (Cons.) and 'optimistic' (Opt.) AUC_a prior.

Data Gen.	**Model**	**Se/Sp**	**True**	$\boldsymbol{AUC_a - Prior}$		
VarCov	**Corr**	**prior**	**AUC**	**Naïve**	**Cons.**	**Opt.**
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.5	0.500 (0.131) [95]	0.514 (0.102) [68]	0.518 (0.104) [79]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.5	0.504 (0.095) [97]	0.508 (0.070) [85]	0.508 (0.077) [92]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.5	0.455 (0.186) [83]	0.487 (0.202) [16]	0.441 (0.198) [16]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.5	0.477 (0.156) [94]	0.500 (0.140) [35]	0.493 (0.133) [54]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.5	0.499 (0.104) [95]	0.504 (0.079) [93]	0.506 (0.089) [95]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.5	0.500 (0.090) [99]	0.502 (0.064) [98]	0.502 (0.067) [97]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.5	0.490 (0.131) [96]	0.472 (0.093) [66]	0.482 (0.108) [84]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.5	0.490 (0.103) [99]	0.483 (0.079) [84]	0.486 (0.082) [96]

$N = 400$

Table B.8: Mean of posterior θ medians with corresponding (standard deviation of posterior θ medians) based on [number of converged data sets] for the simulated data sets of size $N = 400$. FLAT — the analysis using the flat prior for sensitivity and specificity of the reference test; INF — the analysis using the informative prior for sensitivity and specificity of the reference test (see Figure 3.8). Results obtained with the 'naïve', 'conservative' (Cons.), and 'optimistic' (Opt.) AUC_a prior.

Data Gen. Model		**Se/Sp**	**True**	$\boldsymbol{AUC_a - Prior}$		
VarCov	**Corr**	**prior**	**AUC**	**Naïve**	**Cons.**	**Opt.**
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.5	0.501 (0.060) [100]	0.500 (0.054) [100]	0.498 (0.56) [100]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.5	0.497 (0.050) [100]	0.499 (0.045) [100]	0.498 (0.047) [100]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.5	0.511 (0.121) [93]	0.503 (0.081) [61]	0.508 (0.092) [87]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.5	0.495 (0.090) [98]	0.503 (0.050) [88]	0.499 (0.068) [97]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.5	0.501 (0.039) [100]	0.0501 (0.037) [100]	0.499 (0.038) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.5	0.500 (0.036) [100]	0.500 (0.035) [100]	0.501 (0.035) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.5	0.503 (0.062) [100]	0.504 (0.053) [100]	0.507 (0.056) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.5	0.501 (0.055) [100]	0.504 (0.048) [100]	0.504 (0.050) [100]

$N = 600$

Table B.9: Mean of posterior θ medians with corresponding (standard deviation of posterior θ medians) based on [number of converged data sets] for the simulated data sets of size $N = 600$. FLAT — the analysis using the flat prior for sensitivity and specificity of the reference test; INF — the analysis using the informative prior for sensitivity and specificity of the reference test (see Figure 3.8). Results obtained with the 'naïve', 'conservative' (Cons.), and 'optimistic' (Opt.) AUC_a prior.

Data Gen. Model		**Se/Sp**	**True**	$\boldsymbol{AUC_a - Prior}$		
VarCov	**Corr**	**prior**	**AUC**	**Naïve**	**Cons.**	**Opt.**
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.5	0.503 (0.052) [100]	0.500 (0.049) [99]	0.501 (0.050) [100]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.5	0.502 (0.045) [100]	0.504 (0.042) [100]	0.502 (0.043) [100]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.5	0.498 (0.090) [88]	0.512 (0.049) [66]	0.507 (0.071) [92]
$\boldsymbol{\Sigma_0 = \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.5	0.518 (0.068) [89]	0.507 (0.037) [74]	0.509 (0.048) [99]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**FLAT**	0.5	0.506 (0.034) [100]	0.507 (0.033) [100]	0.506 (0.033) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho = 0}$	**INF**	0.5	0.507 (0.032) [100]	0.506 (0.032) [100]	0.506 (0.032) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**FLAT**	0.5	0.510 (0.040) [100]	0.510 (0.037) [100]	0.509 (0.038) [100]
$\boldsymbol{\Sigma_0 \neq \Sigma_1}$	$\boldsymbol{\rho \neq 0}$	**INF**	0.5	0.510 (0.036) [100]	0.507 (0.034) [100]	0.508 (0.035) [100]

Appendix C

Se/Sp prior sensitivity

Results of the ADNI and VUmc data fits using the two different forms of prior distribution restrictions for Se_T and Sp_T and considering the conservative (Cons.) and optimistic (Opt.) AUC_a prior distributions, are summarized in Tables C.1 and C.2. Posterior medians of AUC_a, Se_T, Sp_T, and θ are essentially the same, irrespective of the assumed prior distribution restriction or AUC_a prior distribution, for both data sets. Therefore, we can conclude that the more restrictive Se_T and $Sp_T > 0.5$ does not introduce enough extra prior information to significantly change the posterior results.

Moreover, the difference between the conservative and optimistic AUC_a priors is not informative enough to affect the posterior results. AUC_a posterior distributions for the logistic regression (LR), optimistic (Opt.) and conservative (Cons.) model fits considering the Se_T and $Sp_T > 0.5$ restriction are shown in Figures C.1 and C.2. Comparing the posterior distributions from the results considering the optimistic (black solid line) and conservative AUC_a-prior (black dashed line) confirms that the AUC_a prior does not significantly effect the posterior results.

Table C.1: Posterior median results of the parameters of interest after fitting the ADNI data with both the Se_T and $Sp_T > 0.5$, and the $Se_T + Sp_T > 1$ prior distribution restriction (see Figure 3.12).

Parameter	AUC_a Prior	Se_T/Sp_T Prior	
		$Beta\,(1,1)\,T\,[0.51,1)$	$Se \sim Beta\,(1,1)$ $Sp_{\|Se} \sim Beta\,(1,1)\,T\,[1.001-Se,1)$
AUC_a	Cons.	0.978[0.948,0.991]	0.978[0.948,0.992]
AUC_a	Opt.	0.983[0.958,0.994]	0.983[0.960,0.994]
Se_T	Cons.	0.832[0.736,0.913]	0.831[0.735,0.914]
Se_T	Opt.	0.825[0.730,0.908]	0.826[0.730,0.905]
Sp_T	Cons.	0.890[0.807,0.952]	0.890[0.806,0.953]
Sp_T	Opt.	0.889[0.804,0.952]	0.888[0.803,0.951]
θ	Cons.	0.496[0.411,0.581]	0.497[0.411,0.583]
θ	Opt.	0.500[0.413,0.586]	0.500[0.415,0.587]

Table C.2: Posterior median results of the parameters of interest after fitting the VUmc data with both the Se_T and $Sp_T > 0.5$, and the $Se_T + Sp_T > 1$ prior distribution restriction (see Figure 3.12).

Parameter	AUC_a Prior	Se_T/Sp_T Prior	
		$Beta\,(1,1)\,T\,[0.51,1)$	$Se \sim Beta\,(1,1)$ $Sp_{\|Se} \sim Beta\,(1,1)\,T\,[1.001-Se,1)$
AUC_a	Cons.	0.994[0.990;0.997]	0.995[0.989,0.997]
AUC_a	Opt.	0.995[0.991;0.998]	0.995[0.991,0.998]
Se_T	Cons.	0.958[0.937;0.975]	0.958[0.937,0.975]
Se_T	Opt.	0.957[0.936;0.974]	0.957[0.936,0.974]
Sp_T	Cons.	0.853[0.801;0.897]	0.854[0.803,0.897]
Sp_T	Opt.	0.853[0.802;0.897]	0.853[0.803,0.896]
θ	Cons.	0.700[0.667;0.732]	0.700[0.667,0.732]
θ	Opt.	0.701[0.668;0.733]	0.701[0.667,0.733]

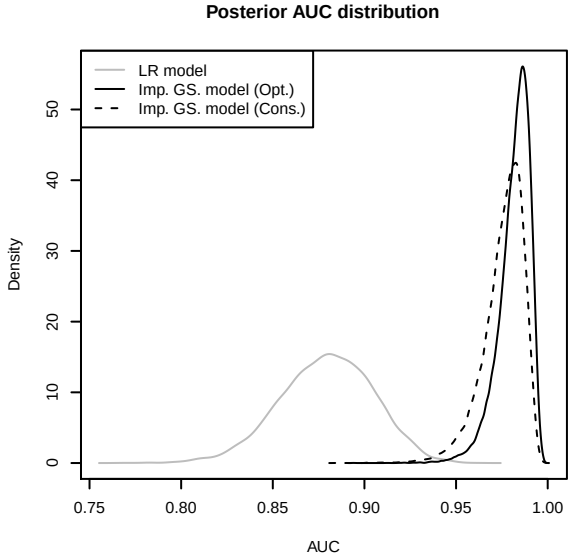

Figure C.1: Posterior AUC_a distribution for the ADNI-data fitted with a logistic regression model (grey line) and the proposed-imperfect reference-test model (black line) with the flat Se_T and Sp_T prior distributions with $Se_T + Sp_T > 1$ restriction, considering the 'optimistic' (solid line) and 'conservative' (dashed line) AUC_a-prior distribution.

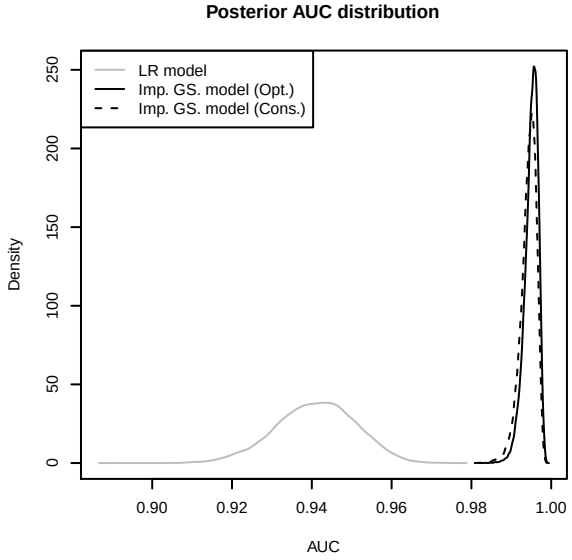

Figure C.2: Posterior AUC_a distribution for the VUmc-data fitted with a logistic regression model (grey line) and the proposed-imperfect reference-test model (black line) with the flat Se_T and Sp_T prior distributions with $Se_T + Sp_T > 1$ restriction, considering the 'optimistic' (solid line) and 'conservative' (dashed line) AUC_a-prior distribution.

Appendix D

Type-I error investigation

In order to ensure that the resulting power estimates from the simulation study considered in section 5.3 of Chapter 5 can be compared, type-I error characteristics have to be investigated. By considering the same simulated data sets as in Chapter 5, characterised by a true AUC_a of 0.787, estimates of type-I error probability can be obtained by setting $\delta = 0.79$. The proportion of simulated data sets for which the null hypothesis: $AUC_a \leq 0.79$, is rejected with a particular value of α can then be assumed an estimate of type-I error probability.

Figure D.1 shows the empirical type-I error probabilities considering the four validation-study sample size settings for the three considered development studies. In Figure D.1, the grey dashed line denotes the empirical probabilities in case development study accuracy information is ignored while the black solid line denotes results for the case when this information is included. The results shown in the figure assume that in both prior settings $\alpha = 0.2$. It is clear from the figure that for the 2.5- and 50-percentile development studies, including development study accuracy information into the validation study leads to a decreased type-I error probability. For the over-estimating 97.5-percentile development study, including development-study accuracy information increases the probability of type-I error.

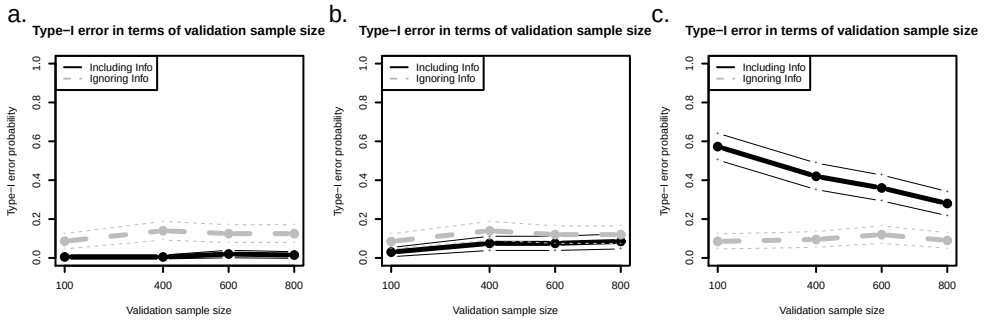

Figure D.1: Empirical type-I error probability as a function of the validation study sample size for the underestimated (**a**.), unbiased (**b**.), and overestimated (**c**.) development study information. Results are indicated for the flat (grey circles connected by the dashed line) and informative (black circles connected by the solid line) AUC_α^*-priors in case $\alpha = 0.2$ for both the flat and informative setting. The 95% confidence interval for each estimate is indicated by the respective plus-signs.

By adjusting the α values for the 'traditional' prior setting in the considered development study cases, it is possible to reach statistically non-significant type-I error probability differences. Figure D.2 contains the results for the empirical type-I error probabilities when α was lowered to 0.06 and 0.15 for the 'traditional' prior setting in case the 2.5- and 50-percentile development studies were considered, respectively. In case validation was performed of the 97.5-percentile development study, α was increased to 0.4. With exception of the case of a validation study sample size of $N = 100$ of the 97.5-percentile development study, all type-I error probabilities are statistically non-significant between the two prior settings. Considering these adjusted α values for the flat-prior setting, allows for valid power comparisons.

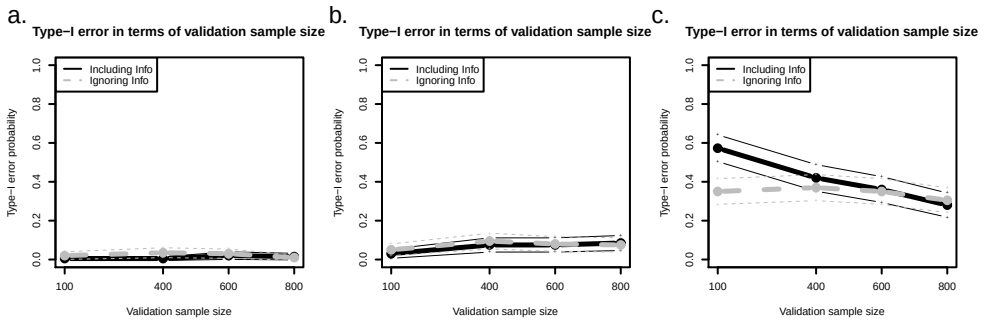

Figure D.2: Empirical type-I error probability as a function of the validation study sample size for the underestimated (**a**.), unbiased (**b**.), and overestimated (**c**.) development study information. Results are indicated for the flat (grey circles connected by the dashed line) and informative (black circles connected by the solid line) AUC_a^*-priors in case $\alpha = 0.2$ for the informative setting and $\alpha = 0.06$, $\alpha = 0.15$, and $\alpha = 0.4$ for the flat setting after the respective development studies. The 95% confidence interval for each estimate is indicated by the respective plus-signs.

Summary

Alzheimer's Disease (AD) is an enormous burden on society and future perspectives foresee this burden only to increase. The need for a treatment for AD is growing but at the same time advances in AD research are hindered by issues related to the diagnosis of the disease. The currently used clinical diagnosis of AD is known to be imperfect while the perfect post-mortem diagnosis is expensive and useless from a diagnostic point of view. Therefore, the need for easily measurable biomarkers is high but many fail to show statistically adequate diagnostic accuracy. One of the reasons may be biased estimation of biomarker accuracy due to the use of the imperfect clinical diagnosis as a reference test without acknowledging this.

The main goal of this dissertation was the development of methods facilitating the development of biomarker-based diagnostic tests for AD. The first research question focuses on how to efficiently estimate the accuracy of a diagnostic biomarker-index. Because of the lack of a gold-standard reference-test, currently available methods making use of the true disease labels would lead to biased accuracy estimates. Therefore, we propose the use of a Bayesian latent-class mixture model in Chapter 3. The model allows to include the information from an imperfect reference-test while accounting for its imperfectness. Care has to be taken with respect to the inclusion of prior information since a combination of uninformative priors may lead to an extremely informative prior for the parameter of interest. Therefore, an alternative parametrisation is proposed to allow the inclusion of prior information directly on the accuracy of the diagnostic-biomarker index. We show that, when appropriate priors are chosen, this model provides unbiased estimates of the diagnostic biomarker-index' accuracy. Moreover, the results suggest that the reports indicating disappointing results of diagnostic performance of the AD CSF-biomarkers might by due in part to the fact that the clinical diagnosis was treated as a GS reference-test.

The assumption that the considered biomarkers are independent of the reference test, conditionally on the true disease status, is untestable and only heuristically enforceable. Therefore, the proposed Bayesian latent-class model is extended in Chapter 4. By considering that the imperfect reference-test is a dichotomized version of an underlying continuous latent-tolerance variable, conditional dependence between the biomarkers and the reference test are modelled directly. Assuming that the continuous tolerance variable and the biomarkers are jointly normally distributed, their correlation can be estimated. Therefore, the estimated accuracy of the diagnostic biomarker-index is corrected for any possible conditional dependence between the biomarkers and reference test without the need for any untestable heuristic argumentation. In terms of the AD application, it is shown that, although statistically significant conditional dependence is observed, it has no significant impact on the accuracy estimate of the diagnostic biomarker-index.

The focus of the second research question is on the validation of a developed diagnostic biomarker-index. Because of the need for large sample sizes or expensive data to reach adequate power of the validation study together with the lack of an efficient statistical framework, validation is rarely performed. In Chapter 5 we propose a Bayesian framework allowing efficient validation of a diagnostic biomarker-index. By making use of the exchangeability assumption of the parameters of the development and validation studies, accuracy information obtained in the development study can be included into the validation study. In particular, an approximation to the posterior distribution of the accuracy parameter from the development study, is carried over to the validation study. Validation is defined as an hypothesis test, testing whether a particular validation criterion value can be rejected. Before comparing the proposed analysis to a 'traditional' analysis in which the development-study information is ignored, significance levels of the hypothesis test are adjusted to obtain comparable type-I error probabilities. We show that, although the information from the development study is discarded by doubling its standard deviation, a large reduction of the required sample size is possible. In particular, the considered settings shows a reduction to about 20% of the required sample size compared to a validation study ignoring the development-study accuracy information to reach a power of approximately 0.53.

The development and validation of a diagnostic AD CSF-biomarker cut-off for a particular commercially available assay does not imply the applicability of the cut-off on other assays, measuring the same biomarker. This would imply setting up time-consuming and expensive studies. Therefore, the third research question investigates the transfer of the cut-off value of an AD CSF-biomarker from a currently used assay to a new one, without having to conduct new development and valida-

tion studies. The validity and the effect of the currently applied linear-regression transfer-method on the clinical performance of the biomarker measured with a new assay, have never been investigated. In Chapter 6 we establish that if the underlying assumptions of the linear-regression-based transfer-method are violated the results are biased. This entails that the diagnostic biomarker has different operating characteristics depending on the assay on which it is measured. Therefore, we propose a novel two-stage Bayesian approach which leads to unbiased and more precise estimates than the linear-regression-based transfer-method. The approach first estimates the distributional characteristics of the diagnostic-biomarker on the current assay based on the results of a GS reference-test. Next, the posterior information is introduced in the second stage as prior information. In the second stage, the cut-off of the new-assay is estimated by considering data measured on both assays side-by-side. Because of the introduction of the information on the current assay in the first stage, no GS information is required to end up with unbiased estimates. The proposed Bayesian approach provides more precise cut-off estimates than the linear-regression-based transfer-method. Though, with the limited sample size of currently considered development and validation studies, only imprecise cut-off estimates are available. This means that the currently used cut-offs have large uncertainty in terms of operating characteristics, which is rarely acknowledged.

Samenvatting

De ziekte van Alzheimer heeft een enorme impact op onze huidige samenleving en voorspellingen menen dat deze impact enkel zal toenemen. De nood aan een doeltreffende behandeling voor Alzheimer neemt toe terwijl vooruitgang in het onderzoek naar de ziekte belemmerd wordt door moeilijkheden met de diagnose van de ziekte. Van de momenteel gehanteerde klinische diagnose van Alzheimer, weet men dat deze niet perfect is. De post-mortem diagnose is dan weer wel perfect, maar kostelijk en vanuit diagnostisch oogpunt onbruikbaar. Daarom is de nood aan eenvoudig te meten biomerkers groot. Vele biomerkers, echter, slagen er niet in om statistisch voldoende diagnostische nauwkeurigheid aan te tonen. Een van de redenen zou kunnen zijn dat men, door de foutieve aanname dat de klinische diagnose een perfecte referentietest zou zijn, tot een vertekende schatting van diagnostische nauwkeurigheid komt.

Het hoofddoel van deze verhandeling was om methoden te ontwikkelen die de ontwikkeling van diagnostische tests voor Alzheimer op basis van biomerkers, zouden kunnen ondersteunen. De eerste onderzoeksvraag richt zich op het efficiënt schatten van de nauwkeurigheid van een diagnostische index gebaseerd op biomerkers. Bij gebrek aan een gouden standaard referentietest, leiden de huidige methoden, die gebruik maken van de ware onderliggende ziekte status, mogelijks tot vertekende schattingen van de nauwkeurigheid. Om die reden stellen wij een *Bayesian latent-class mixture* model voor in Hoofdstuk 3. Dit model laat toe om de informatie vervat in een niet-perfecte referentietest, toch op te nemen tijdens de schatting van diagnostische nauwkeurigheid, terwijl er rekening mee wordt gehouden dat deze niet perfect is. Men dient zorg te besteden aan de wijze waarop men *prior* informatie aan het model toevoegt. Een combinatie van niet-informatieve *priors* kan immers leiden tot een zeer informatieve *prior* voor de parameter waarin men geïnteresseerd is. Om dit te vermijden, stellen wij een alternatieve parameterizatie voor, die het toelaat om

de *prior* informatie rechtstreeks te veronderstellen op het niveau van de nauwkeurigheid van de diagnostische biomerker-index. We tonen aan dat, wanneer geschikte *priors* worden gekozen, het voorgestelde model een niet-vertekende schatting van de nauwkeurigheid van de diagnostische biomerker-index kan maken. De resultaten suggereren ook dat de voorgaande teleurstellende resultaten inzake de nauwkeurigheid van de Alzheimer biomerkers mogelijk te wijten zijn aan het foutief beschouwen van de klinische diagnose als een perfecte referentietest.

De aanname dat de beschouwde biomerkers onafhankelijk zijn van de referentietest, gegeven de ware ziekte status, is niet testbaar en kan enkel via heuristische argumentatie aannemelijk gemaakt worden. Hierom, wordt het *Bayesian latent-class* model uitgebreid in Hoofdstuk 4. Door aan te nemen dat de niet-perfecte referentietest een dichotome versie is van een onderliggende continue latente tolerantievariabele, kan de conditionele afhankelijkheid tussen de biomerkers en de referentietest rechtstreeks gemodelleerd worden. Onder de veronderstelling dat de gezamenlijke distributie van de continue tolerantievariabele en de biomerkers een multi-variate normaal distributie is, kan de beschouwde correlatie geschat worden. Hierdoor is de geschatte nauwkeurigheid van de diagnostische biomerker-index gecorrigeerd voor een mogelijke conditionele afhankelijkheid tussen de biomerkers en de referentietest zonder zich te hoeven beroepen op heuristische argumentatie. Wat betreft een applicatie op data van Alzheimer patiënten, tonen we aan dat hoewel er sprake is van statistisch significante conditionele afhankelijkheid, dit geen effect heeft op de schatting van de nauwkeurigheid van de diagnostische biomerker-index.

De tweede onderzoeksvraag betreft de validatie van een ontwikkelde diagnostische biomerker-index. Om tot een toereikende power van de validatiestudie te komen, zijn er momenteel zulke grote steekproefgroottes of kostelijke data nodig in combinatie met een gebrek aan efficiënte statistische modellen, dat er zelden tot validatie wordt overgegaan. In Hoofdstuk 5 stellen wij een Bayesiaans model voor dat toelaat om de nauwkeurigheid van een diagnostische biomerker-index efficiënt te valideren. Door gebruik te maken van de uitwisselbaarheidsassumptie, wordt het mogelijk om nauwkeurigheidsinformatie, vergaard in de ontwikkelingsstudie, te introduceren in de validatiestudie. In het bijzonder kan een benadering van de *posterior* distributie van de nauwkeurigheidsparameter, geschat in de ontwikkelingsstudie, overgedragen worden als *prior* informatie voor de validatiestudie. Validatie is gedefinieerd in de vorm van een hypothese-toets die nagaat of een bepaald validatie criterium al dan niet kan worden weerlegd. Bij het vergelijken van de voorgestelde analyse met de 'traditionele' analyse, waarbij de nauwkeurigheidsinformatie van de ontwikkelingsstudie buiten beschouwing gelaten wordt, worden eerst de significantie niveaus van de hypothese-toets

aangepast zodanig dat vergelijkbare kansen op een type-I fout worden bekomen. We tonen aan dat, hoewel de informatie van de ontwikkelingsstudie wordt gereduceerd door de standaarddeviatie van deze informatie te verdubbelen, een significante reductie mogelijk is van de benodigde steekproefgrootte. In het beschouwde voorbeeld, kan deze reductie tot ongeveer 20% van de 'traditioneel' benodigde steekproefgrootte gaan om een zelfde power van ongeveer 0.53 te bekomen.

De ontwikkeling en validatie van de drempelwaarde van een diagnostische Alzheimer biomerker voor een bepaald commercieel beschikbaar platform, impliceert niet automatisch de overdraagbaarheid van de drempelwaarde naar een ander platform, dat dezelfde biomerker meet. Dit betekent dat er nieuwe tijdrovende en kostelijke studies moeten worden opgezet. Om te vermijden dat er nieuwe ontwikkelings- en validatiestudies zouden moeten uitgevoerd worden, spitst de derde onderzoeksvraag zich toe op de overdraagbaarheid van biomerker drempelwaarden van een huidig toegepast platform naar een nieuw platform. De geldigheid en het effect van de huidige toegepaste lineaire-regressie overdrachtsmethode op de klinische nauwkeurigheid van de biomerker, gemeten op het nieuwe platform, zijn tot op heden nooit onderzocht. In Hoofdstuk 6 stellen we vast dat wanneer de onderliggende aannames van de huidige overdrachtsmethode geschonden zijn, vertekende resultaten bekomen worden. Deze vaststelling houdt in dat, afhankelijk van het platform waarop de biomerker gemeten wordt, diens klinische nauwkeurigheid varieert. Daarom stellen wij een nieuwe tweefase Bayesiaanse aanpak voor die tot onvertekende en preciezere resultaten leidt dan de huidige overdrachtsmethode. De voorgestelde methode schat eerst de distributionele kenmerken van de diagnostische biomerker, gemeten op het huidige platform, in combinatie met de resultaten van een gouden standaard referentietest. Vervolgens wordt de *posterior* informatie in de tweede fase aangebracht als *prior* informatie. In de tweede fase wordt de drempelwaarde van het nieuwe platform geschat door middel van data gemeten op beide platforms. Omdat de informatie omtrent het huidige platform via de eerste fase wordt binnengebracht, is er geen nood meer aan gouden standaard referentietest informatie om tot onvertekende schattingen te komen. Hoewel de voorgestelde Bayesiaanse methode tot preciezere schattingen leidt dan de huidige overdrachtsmethode, blijven de schattingen nog steeds zeer onzeker omwille van de beperkte steekproefgroottes van de ontwikkelings- en validatiestudies. Dit betekent dat er op dit ogenblik grote onzekerheid bestaat rond de gebruikte drempelwaarden in termen van klinische nauwkeurigheid, onzekerheid die zelden bekrachtigd wordt.