PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE GRADUADOS

Estimation of the disease prevalence when diagnostic tests are subject to classification error: Bayesian Approach

TESIS PARA OPTAR POR EL GRADO DE MAGISTER EN ESTADÍSTICA

Presentado por:

Evelyn Patricia Gutierrez Ayala

Asesor: Victor Giancarlo Sal y Rosas Celi

Miembros del jurado:
Dr. Luis Hilmar Valdivieso Serrano
Dr. Cristian Luis Bayes Rodriguez
Dr. Victor Giancarlo Sal y Rosas Celi

Lima, Diciembre 2016

Dedicatoria

Dedico esta tesis a mis padres por quererme y apoyarme en todo momento.

A mis mejores amigas, Jessica y Sandra, quienes fueron el gran apoyo emocional durante el tiempo en que escribía esta tesis.

Agradecimentos

Agradezco a los profesores de la Maestría de Estadística de la Pontificia Universidad Católica del Perú (PUCP), quienes con sus enseñanzas y consejos me han dado el soporte teórico necesario para poder desarrollar este trabajo.

A mi asesor el Dr. Giancarlo Sal y Rosas Celi por su guia y apoyo durante el tiempo que escribía esta tesis.

A mi buen amigo Elvis quien me apoyo y alentó desde el inicio y en los momentos cuando parecia que me iba a rendir.

A Jessica, Sandra, Maria Andrea, Rosemary, Fabiola y Cesar quienes con sus consejos han sido de gran ayuda y apoyo emocional en los momentos dificiles mientras escribía esta tesis.

Resumen

La estimación de la prevalencia de una enfermedad, la cual es definida como el número de casos con la enfermedad en una poblacion dividida por el numero de elementos en esta, es realizado con gran precisión cuando existen pruebas 100% exactas, tambien llamadas *gold standard*. Sin embargo, en muchos casos, debido a los altos costos de las pruebas de diagnóstico o limitaciones de tecnología, la prueba *gold standard* no existe y debe ser reemplazada por una o más pruebas diagnósticas no tan caras pero con bajos niveles de sensibilidad o especificidad. Este estudio esta enfocado en el estudio de dos enfoques bayesianos para la estimacion de prevalencia cuando no es factible tener resultados de una prueba 100% exacta. El primero es un modelo con dos parametros que toman en cuenta la asociación entre los resultados de las pruebas. El segundo es un enfoque que propone el uso del *Bayesian Model Averaging* para combinar los resultados de cuatro modelos donde cada uno de estos tiene suposiciones diferentes sobre la asociación entre los resultados de las pruebas diagnosticas. Ambos enfoques son estudiados mediante simulaciones para evaluar el desempeño de estos bajo diferentes escenarios. Finalmente estas tecnicas serán usadas para estimar la prevalencia de enfermedad renal crónica en el Perú con datos de un estudio de cohortes de CRONICAS (Francis et al., 2015).

Palabras-clave: Análisis Bayesiano, Prevalencia, Pruebas diagnosticas, Sensibilidad, Especificidad, Bayesian Model Averaging, Modelo de Efectos fijos.

Abstract

The estimation of a disease prevalence, which is defined as the number of cases with the disease in a population divided by the number of the elements in it, is done with high precision when we have a 100% accuracy (also known as gold standard) test. However, in many cases, due to the high cost of the diagnostic tests or limited technology, a gold standard test can not be used and should be replaced by one or two non expensive ones with usually a limited level of accuracy. (i.e. low levels of sensitivity or specificity). This study is focused on two Bayesian approaches to estimate the prevalence of a disease when it is not possible to have the results from a 100% accurate diagnostic test. The first approach is a model with two parameters that account for the association between test results. The second approach is a model that proposes the use of the Bayesian Model Averaging to combine four models where each one of the four models has different assumptions on the association of test results. Both approaches will be studied under different scenarios using simulated datasets to assess their performance; and finally, based on the results of simulation study, we will be able to use them to estimate the prevalence of chronic kidney disease in Peru using data from the CRONICAS cohort study (Francis et al., 2015).

Keywords: Prevalence, Bayesian Approach, Diagnostic Tests, Sensitivity, Specificity, Bayesian Model Averaging, Fixed Effects model.

Contents

List of Abbreviations

BMA	Bayesian Model Averaging .
FE	Fixed Effects.
HPD	High posterior density.
RMSE	Root Mean Square Error.
MCMC	Monte Carlo Markov Chain.
RJMCMC	Reversible Jump MCMC.
CKD	Chronic Kidney Disease.
PCR	Protein-Creatinine Ratio.
GFR	Glomerular Filtration Rate.
eGFR	Estimated Glomerular Filtration Rate.
CRONICAS	Centro de Excelencia en Enfermedades Crónicas.

List of Symbols

π	Prevalence.
S_i	Test i Sensitivity.
C_i	Test i Specificity.
$P_{ij\|k}$	Conjoint probability of the results in test 1 and test 2 given the status of disease is k.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Preliminary Considerations

The estimation of a disease prevalence, which is defined as the number of cases with the disease in a population divided by the number of the elements in it, is done with high precision when we have a 100% accuracy (also known as gold standard) test. However, in many cases, due to the high cost of the diagnostic tests or limited technology, a gold standard test should be replaced by one or two non expensive tests which usually have a limited level of accuracy, i.e. low levels of sensitivity or specificity.

The problem of prevalence estimation has followed different approaches. One is the frequentist approach when only with two diagnostic tests the problem becomes not identifiable. This has been managed by using a minimum of four diagnostic tests. Another way to manage the identifiability problem has been by setting values to some of the parameters so that the problem becomes identifiable. In this case, the problem that follows is to decide the parameters and the values that will be assumed. The answer is not always unique and give us the freedom to chose which of the variables we want to make assumptions on. This could lead to different estimations depending not only on the variable chosen but also on the values assumed for those.

The other way to handle the problem of estimation is to consider a Bayesian approach. Here, the identifiability in the model is handle by adding prior knowledge on some of the parameters. This means the inclusion of probabilistic models for those parameters for which we have some information on. Gustafson (2005) believe that in many cases models with some prior information work better than simpler models or non identifiable models.

In the Bayesian approach, we found different ways of estimating the prevalence. Two of them are studied here. Dendukuri and Lawrence (2001) proposed a model that takes into account the association between the diagnostic tests given the true disease status. On the other hand, Black and Craig (2002) proposed a procedure that combines four models where those four models account for different assumptions on the association of the tests.

One of the challenges in the Bayesian approach is the specification of the model. The association between the diagnostic tests by disease status sometimes is not clear and some assumptions have to be made; however, Gustafson (2005) and Albert and Dodd (2004) show that a misspecification of the model could lead to biased prevalence estimations. The two models presented here do not make any assumptions on the association between test results so that we do not have to make any assumption on the association between the test results.

1.2 Objectives

The objective of this study is to compare two Bayesian models described in the literature. The model proposed by Dendukuri and Lawrence (2001), which we will also name the FE approach, is a general model that takes into account any association between the diagnostic tests. while the BMA model proposed by Black and Craig (2002) takes into account different cases of association to handle the specification uncertainty. Combining those four models, it is expected to reduce the uncertainty of model specification and the estimation bias.

Understanding the assumptions and model properties allow us to understand the performance of those models under different scenarios and find the best approach to estimate the disease prevalence. The models will be studied using simulations of different scenarios of association between the diagnostic tests results.

In particular, these models will be applied to the problem of estimating the prevalence of chronic kidney disease (CKD) in Peru. The research on this disease has been studied at the Cayetano Heredia University in Peru, centro de enfermedades cronicas (CRONICAS), and recently, the researchers estimated the prevalence of this disease to be 16.8% [95% CI: 13.5-20.9%] (Francis et al., 2015). However, the diagnostic tests used have reduced levels of accuracy which could have led to a biased estimation of the prevalence.

1.3 Organization of the thesis

This work is divided as follows: In chapter 2, we present the data structure, statistical models and inference. Chapter 3 is dedicated to present the implemented simulations. Chapter 5 shows the application of the models to the estimation of the prevalence of Chronic Kidney Disease in Peru and chapter 6 presents the conclusions and suggestions for further research. Finally, we end with an appendix that describes technical developments in the Gibbs algorithms that have been used in this work.

Chapter 2

Data Structure, Models and Inference

2.1 Data Structure

Let D be a latent variable which represents the real status of the disease. $D = 1$ means that the subject has the disease, and $D = 0$ means that the subject does not have the disease. The true prevalence is denoted by π ($\pi = P(D = 1)$). Let T_j be the result of test j, $j \in \{0, 1\}$, which could take two values: $\{0\}$ if the test result is negative or $\{1\}$ if the result is positive. The sensitivity and specificity of test i will be denoted by S_i and C_i, respectively and are defined as follows: $S_i = P(T_i = 1|D = 1)$ and $C_i = P(T_i = 0|D = 0)$.

Let N be the total number of observations (patients) we have in the sample, and let $\boldsymbol{n}$ be a vector that contains the numbers of patients at each combination of the two test results: (positive, positive), (positive, negative), (negative, positive) or (negative, negative) respectively. Table 1 shows a 2×2 contingency table that has the terminology of the count of patients classified by their results in the tests.

Table 2.1: Contingency table showing all patients classified by their results on the two diagnostic tests

		Test 2		
		Positive	Negative	Total
Test 1	Positive	n_{11}	n_{10}	$n_{11} + n_{10}$
	Negative	n_{01}	n_{00}	$n_{01} + n_{00}$
	Total	$n_{11} + n_{01}$	$n_{10} + n_{00}$	N

Let Y be a latent variable that represents the number of patients that have the disease. Y could also be distributed in a 2×2 table depending on the two diagnostic test results. For convenience, we will define the vector $\mathbf{y}$, where $\mathbf{y} = (y_{11}, y_{10}, y_{01}, y_{00})$, and it will be a vector of patients with the disease at each cell of the 2×2 table. Table 2 shows the terminology we use for the patients with the disease in a 2×2 table.

Table 2.2: Contingency table showing patients with the disease classified by their results on the two diagnostic tests

		Test 2		
		Positive	Negative	Total
Test 1	Positive	y_{11}	y_{10}	$y_{11} + y_{10}$
	Negative	y_{01}	y_{00}	$y_{01} + y_{00}$
	Total	$y_{11} + y_{01}$	$y_{10} + y_{00}$	Y

Throughout this study, we will be also using the conjoint probabilities of test results

conditional to the status of the disease: $P_{t_i t_j|d} = P(T_i = t_i, T_j = t_j|D = d)$. However, in some cases, to simplify notation, we will be using the vector of probabilities $\boldsymbol{p_1}$ and $\boldsymbol{p_0}$, which are vectors of probabilities for patients with and without the disease respectively. $\boldsymbol{p_1} = (P_{11|1}, P_{10|1}, P_{01|1}, P_{00|1})$ and $\boldsymbol{p_0} = (P_{11|0}, P_{10|0}, P_{01|0}, P_{00|0})$.

As Y is the number of patients with the disease, it is easy to identify its distribution:

$$Y \propto Binomial(N, \pi)$$

and similarly, for $\mathbf{y}$, the vector of patients with the disease divided by the results in the two diagnostic tests:

$$\mathbf{y} \propto Multinomial(\boldsymbol{n}, \boldsymbol{p_1})$$

Those distributions are taken into account to define the augmented likelihood:

$$L(\mathbf{n}; \theta, \mathbf{y}) \propto \pi^Y P_{11|1}^{y_{11}} P_{10|1}^{y_{10}} P_{01|1}^{y_{01}} P_{00|1}^{y_{00}} (1-\pi)^{(N-Y)} P_{11|0}^{(n_{11}-y_{11})} P_{10|0}^{(n_{10}-y_{10})} P_{01|0}^{(n_{01}-y_{01})} P_{00|0}^{(n_{00}-y_{00})} \quad (2.1)$$

The models we present in the following section take as a starting point this augmented likelihood. The models discussed here consider different parametrization and prior distributions of the parameters of interest to estimate the prevalence of a disease. In the following section, we describe both approaches to estimate the prevalence.

2.2 Fixed Effects Approach

In this section, we describe a model proposed by Dendukuri and Lawrence (2001) which takes into account the association between the test results and which we name also named Fixed Effects approach (FE approach). In order to model this, it includes two parameters that model the relationship that could exist between the results of two diagnostic tests. The first parameter accounts for the relationship between test results among patients who have the disease ($covS$) and the second for the relationship between test results among patients without the disease ($covC$). The parameters are defined as follows:

$$covS = P_{11|1} - P_{1.|1} P_{.1|1} = P_{11|1} - S_1 S_2$$

$$covC = P_{00|0} - P_{.0|0} P_{0.|0} = P_{00|0} - C_1 C_2$$

where we could notice that ($covS$) and ($covC$) are defined as the difference between the conjoint probability of test results and the conjoint probability as if the tests would be independent, i.e. a multiplication of S_1 and S_2 or C_1 and C_2. Those parameter show the association between the test results. The larger they are, the more association between the test results exists.

By using the sensibility, specificity of both tests and the two additional parameters, it is possible to rewrite $\boldsymbol{p_0}$ and $\boldsymbol{p_1}$ in terms of (S_i, C_i, $covS$, $covC$):

$$\begin{aligned} P_{11|1} &= S_1 S_2 + covS \\ P_{10|1} &= S_1(1 - S_2) - covS \end{aligned} \quad (2.2)$$

$$\begin{aligned} P_{01|1} &= (1-S_1)S_2 - covS \\ P_{00|1} &= (1-S_1)(1-S_2) + covS \end{aligned}$$

Similarly, for patients without the disease:

$$\begin{aligned} P_{11|0} &= (1-C_1)(1-C_2) + covC \\ P_{10|0} &= (1-C_1)C_2 - covC \\ P_{01|0} &= C_1(1-C_2) - covC \\ P_{00|0} &= C_1C_2 + covC \end{aligned} \tag{2.3}$$

Replacing (2.2) and (2.3) in (2.1), the augmented likelihood can be rewritten as follows:

$$\begin{aligned} L(\mathbf{n};\theta,Y,\mathbf{y}) \propto\ & \pi^Y(S_1S_2+covS)^{y_{11}}(S_1(1-S_2)-covS)^{y_{10}} \\ & (S_1(1-S_2)-covS)^{y_{01}}((1-S_1)(1-S_2)+covS)^{y_{00}} \\ & (1-\pi)^{(N-Y)}((1-C_1)(1-C_2)+covC)^{(n_{11}-y_{11})} \\ & ((1-C_1)C_2-covC)^{(n_{10}-y_{10})}(C_1(1-C_2)-covC)^{(n_{01}-y_{01})} \\ & (C_1C_2+covC)^{(n_{00}-y_{00})} \end{aligned}$$

Given the new augmented likelihood, we assume some convenient prior distribution:

The prior distribution for the prevalence is supposed to follow a beta distribution: $\pi \sim beta(\alpha_\pi, \beta_\pi)$. The sensitivities and specificities of both tests also follow a beta distribution: $S_i \sim beta(\alpha_{S_i}, \beta_{S_i})$ and $C_i \sim beta(\alpha_{C_i}, \beta_{C_i})$, respectively. Finally, the prior distribution of $covS$ and $covC$ are chosen to follow a generalized beta distribution of first kind, which are distributions similar to the beta distribution with the only difference that this is restricted to a certain range. This is due to the fact that the range of $covS$ and $covC$ are not between 0 and 1 but in a reduced space. More specifically, $covS \sim genbeta(\alpha_{covS}, \beta_{covS}, \mu_s)$, where $\mu_s = min((S_1, S_2) - S_1S_2)$ and $covC \sim genbeta(\alpha_{covC}, \beta_{covC}, \mu_c)$, where $\mu_c = min((C_1, C_2) - C_1C_2)$.

Finally, with the priors defined above, the posterior distribution is the following:

$$\begin{aligned} P \propto\ & (S_1(1-S_2)-covS)^{y_{01}}((1-S_1)(1-S_2)+covS)^{y_{00}} \\ & (1-\pi)^{(N-Y)}((1-C_1)(1-C_2)+covC)^{(n_{11}-y_{11})} \\ & ((1-C_1)C_2-covC)^{(n_{10}-y_{10})}(C_1(1-C_2)-covC)^{(n_{01}-y_{01})} \\ & (C_1C_2+covC)^{(n_{00}-y_{00})}\pi^{\alpha_\pi-1}(1-\pi)^{\beta_\pi-1} \\ & S_1^{\alpha_{S_1}-1}(1-S_1)^{\beta_{S_1}-1}S_2^{\alpha_{S_2}-1}(1-S_2)^{\beta_{S_2}-1} \\ & C_1^{\alpha_{C_1}-1}(1-C_1)^{\beta_{C_1}-1}C_2^{\alpha_{C_2}-1}(1-C_2)^{\beta_{C_2}-1} \\ & covS^{\alpha_{covS}-1}(u_s-covS)^{\beta_{covS}-1}covC^{\alpha_{covC}-1}(u_c-covC)^{\beta_{covC}-1} \end{aligned} \tag{2.4}$$

(2.5)

In order to estimate the prevalence, we need to get a sample of values from the posterior distribution which is not so easy given that not all the conditional probabilities are well known. It is easy to see that the posterior distribution of π, conditional on the rest of parameters, follows a beta distribution; however, for the remaining parameters, it is not easy

to recognize their conditional distributions. To draw samples from the parameters that do not have a well known distribution, we will use a Sampling Importance Resampling (SIR) (Gelman et al., 1995) and the Gibbs sampling algorithm will be used as always to sample get a sample from the posterior distribution. The Gibbs and SIR algorithm is detailed in the appendix A.2.

In this approach, we see that the parameter that account for the positive association between test results is a difference between two probabilities, similar to a distance to the case of independence. Given the fact that this parameters($covS$ and $covC$) varies between 0 and 1 and is usually in a shorter interval, it is used a beta distribution as a prior distribution of the parameter, which could allow any information that we could have about the association between test results. For example, if we are not completely sure about the association between the test results, the parameters of the beta distribution could be assumed in order to have high probability around zero for the distributions of $covC$ or $covS$, which is the case of independence, and less probability for values away from zero. In the applications of this work, we have used the parameters $alpha$=1 and $beta$=1, so that the distribution is flat and reflects the fact that we do not have much information about the association between test results. In practice, it will be common to use a non informative prior because those parameters from which we need the information, $covS$ and $covC$, are not easy to understand by medical practitioner.

2.3 Bayesian Model Averaging Approach (BMA model)

This approach was proposed by Black and Craig (2002) and it was motivated by the idea that we usually do not know whether there is or not association between the test results. This approach proposes to ensemble four models where each one of the models considers different scenarios of association between the test results. Using the technique of Bayesian Model Averaging, the four models are combined in order to get better estimates of the prevalence and to reduce the uncertainty about the association of test results.

The first model (Model 1) makes the assumption that the test results are independent from each other. The second one (Model 2) considers the association of the test results only for patients with the disease. The third model (Model 3) considers the association of test results only for patients who do not have the disease; and, fourth model (Model 4) considers the association between test results for all patients (patients who do not have the disease and patients who do have the disease).

Model 1: Independence of the test results.

In this first model we assume independence between test results. It means, the result of test 1 does not affect in any way the result of test 2 and vice versa. This, in terms of probability, means we can express the conjoint probability of test results by multiplying the marginal probabilities of test results, i.e.: $P_{ij|k} = P_{i.|k}P_{.j|k}$. As a consequence, the elements of $\boldsymbol{p_1}$ and $\boldsymbol{p_0}$ could be written in terms of: S_1, S_2, C_1, and C_2. For patients with the disease,

the probabilities are written as follows:

$$\begin{aligned} P_{11|1} &= P_{1.|1}P_{.1|1} = S_1S_2 \\ P_{10|1} &= P_{1.|1}(1-P_{.1|1}) = S_1(1-S_2) \\ P_{01|1} &= (1-P_{1.|1})P_{.1|1} = (1-S_1)S_2 \\ P_{00|1} &= (1-P_{1.|1})(1-P_{.1|1}) = (1-S_1)(1-S_2) \end{aligned} \tag{2.6}$$

similarly, for patients without the disease:

$$\begin{aligned} P_{11|0} &= P_{1.|0}P_{.1|0} = (1-C_1)(1-C_2) \\ P_{10|0} &= P_{1.|0}(1-P_{.1|0}) = (1-C_1)C_2 \\ P_{01|0} &= P_{0.|0}P_{.1|0} = C_1(1-C_2) \\ P_{00|0} &= P_{0.|0}P_{.0|0} = C_1C_2 \end{aligned} \tag{2.7}$$

Replacing $\boldsymbol{p_1}$ and $\boldsymbol{p_0}$ from (2.6) and (2.7) in (2.1), the likelihood is written as follows:

$$\begin{aligned} L(\mathbf{n};\theta_{M_1},T) \propto\ & \pi^Y[S_1{}^{(y_{11}+y_{10})}(1-S_1)^{(y_{01}+y_{00})}][S_2{}^{(y_{11}+y_{01})}(1-S_2)^{(y_{10}+y_{00})}] \\ & (1-\pi)^{(N-Y)}[C_1{}^{((n_{10}-y_{10})+(n_{00}+y_{00}))}(1-C_1)^{(n_{11}-y_{11})+(n_{10}-y_{10})}] \\ & [C_2{}^{(n_{10}-y_{10})+n_{00}-y_{00}}(1-C_2)^{(n_{11}-y_{11})+n_{01}-y_{01}}], \end{aligned}$$

where $\theta_{M_1} = (S_1, S_2, C_1, C_2)$

Once we have the model, the prior distribution are conveniently assumed to follow the distributions below:

$$\begin{aligned} S_1 \sim Beta(\alpha_{S_1},\beta_{S_1}) \quad &, \quad S_2 \sim Beta(\alpha_{S_2},\beta_{S_2}) \\ C_1 \sim Beta(\alpha_{C_1},\beta_{C_1}) \quad &, \quad C_2 \sim Beta(\alpha_{C_2},\beta_{C_2}) \\ \pi \sim Beta(\alpha_\pi,\beta_\pi) & \end{aligned} \tag{2.8}$$

And, finally, it is easy to combine the prior distributions and the model to obtain the posterior distribution:

$$\begin{aligned} P(\theta_{M_1},T|n) \propto\ & [\pi^{(\alpha_\pi+Y-1)}(1-\pi)^{\beta_\pi+N-Y-1}] \\ & [S_1{}^{\alpha_{S_1}+y_{11}+y_{10}-1}(1-S_1)^{\beta_{S_1}+y_{01}+y_{00}-1}] \\ & [S_2{}^{\alpha_{S_2}+y_{11}+y_{01}-1}(1-S_2)^{\beta_{S_1}+y_{10}+y_{00}-1}] \\ & [C_1{}^{\alpha_{C_1}+n_{01}+n_{00}-y_{01}-y_{00}-1}(1-C_1)^{\beta_{C_1}+n_{11}+n_{10}-y_{11}-y_{10}-1}] \\ & [C_2{}^{\alpha_{C_2}+n_{10}+n_{00}-y_{10}-y_{00}-1}(1-C_2)^{\beta_{C_2}+n_{11}+n_{01}-y_{11}-y_{01}-1}] \end{aligned} \tag{2.9}$$

This model was initially studied by Lawrence et al. (1995). The prior distribution are chosen as beta distribution for two reasons. The first one is the flexibility of modeling any prior information with a beta distribution, and the second reason is that by using beta distributions, the posterior probability are easy-to-recognize so that it is possible to use a

Gibbs sampling to get samples from there. The Gibbs sampling algorithm is described in Appendix A.2.

Model 2: Positive association between test results in patients with the disease

In this case, one assumes a positive association between the test results for patients with the disease. That means, it is more likely to have the same results in both tests in patients with the disease. In this case, besides S_1, S_2, C_1 and C_2, the model add an additional parameter ($P_{11|1}$) which model the association between the test results of patients with the disease. Furthermore, to make sure that there is a positive association between the test results, the following restriction will be forced:

$$S_1 S_2 < P_{11|1} < min(S1, S2) \tag{2.10}$$

Now, $\boldsymbol{p_1}$, defined by 2.2 in the first model, can be rewritten adding the new parameter $P_{11|1}$:

$$\begin{aligned} P_{10|1} &= P_{+\cdot|1} - P_{11|1} = S_1 - P_{11|1} \\ P_{01|1} &= P_{\cdot+|1} - P_{11|1} = S_2 - P_{11|1} \\ P_{00|1} &= 1 - P_{10|1} - P_{01|1} - P_{11|1} = 1 - S_1 - S_2 + P_{11|1} \end{aligned} \tag{2.11}$$

On the other hand, for patients without the disease, one assumes independence between the test results (2.7).

Replacing $\boldsymbol{p_1}$ from (2.11), and $\boldsymbol{p_0}$ from (2.7), the likelihood is written now in terms of $\theta_{M_2} = (S_1, S_2, C_1, C_2, P_{11|1})$:

$$\begin{aligned} L(\mathbf{n}; \theta_{M_2}, \mathbf{y}, Y) \propto\ & \left(\pi^Y (1-\pi)^{N-Y}\right) \\ & \left(P_{11|1}^{y_{11}} (S_1 - P_{11|1})^{y_{10}} (S_2 - P_{11|1})^{y_{01}} (1 - S_1 - S_2 + P_{11|1})^{y_{00}}\right) \\ & \left({C_1}^{(n_{01}-y_{01})+(n_{00}-y_{00})} (1-C_1)^{(n_{11}-y_{11})+(n_{10}-y_{10})}\right) \\ & \left({C_2}^{(n_{10}-y_{10})+(n_{00}-y_{00})} (1-C_2)^{(n_{11}-y_{11})+(n_{01}-y_{01})}\right) \end{aligned}$$

For this model we will assume the same priors than in the previous model and additionally a prior distribution for $P_{11|1}$. Assuming that we do not know much about $P_{11|1}$, it will be used a uniform distribution to model this parameter: $P_{11|1} \sim Unif(S_1 S_2, min(S_1, S_2))$. Then, the posterior distribution is written as follows:

$$\begin{aligned} P(\theta_{M_2}, T|n) \propto\ & \left(\pi^{(\alpha_\pi + Y - 1)} (1-\pi)^{\beta_\pi + N - Y - 1}\right) \\ & \left(P_{11|1}^{y_{11}} (S_1 - P_{11|1})^{y_{10}} (S_2 - P_{11|1})^{y_{01}} (1 - S_1 - S_2 + P_{11|1})^{y_{00}}\right) \\ & \left({S_1}^{\alpha_{S_1}-1} (1-S_1)^{\beta_{S_1}-1}\right) \left({S_2}^{\alpha_{S_2}-1} (1-S_2)^{\beta_{S_1}-1}\right) \left(\frac{1}{min(S_1, S_2) - S_1 S_2}\right) \\ & \left({C_1}^{\alpha_{C_1}+(n_{01}-y_{01})+(n_{00}-y_{00})-1} (1-C_1)^{\beta_{C_1}+(n_{11}-y_{11})+(n_{10}-y_{10})-1}\right) \\ & \left({C_2}^{\alpha_{C_2}+(n_{10}-y_{10})+(n_{00}-y_{00})-1} (1-C_2)^{\beta_{C_2}+(n_{11}-y_{11})+(n_{01}-y_{01})-1}\right) \end{aligned} \tag{2.12}$$

It can be easily seen that the conditional distribution of π, C_1, and C_2 are known distribu-

tions (beta distributions) but the conditional distributions of S_1, S_2 and $S_{00|0}$ are unknown. For those unknown distributions, we will use the Metropolis-Hastings algorithm. To draw samples from the posterior distribution, one will use a Gibbs Sampling algorithm with a Metropolis Hasting algorithm to draw samples from $(S_1, S_2, P_{11|1})$ as they do not have an easy-to-identify distribution. (See Appendix A.3)

Model 3: Positive association of test results for patients without the disease

In this model one assumes a positive association between tests results of patients who do not have the disease. In other words, the results in test 1 and test 2 are likely to be the same for patients who do not have the disease.

To account for a positive association between test results of patients without the disease, we will be using the parameter $P_{00|0}$ and will impose a restriction which is the following:

$$C_1C_2 \leq P_{00|0} \leq min(C_1, C_2) \tag{2.13}$$

Then, the probabilities of test results for patients without the disease (elements of $\mathbf{p_0}$) can be written in terms of C_1, C_2, and $P_{00|0}$:

$$\begin{aligned} P_{10|0} &= P_{.-|0} - P_{00|0} = C_2 - P_{00|0} \\ P_{01|0} &= P_{-.|0} - P_{00|0} = C_1 - P_{00|0} \\ P_{11|0} &= 1 - P_{10|1} - P_{01|0} - P_{00|0} = 1 - C_1 - C_2 + P_{00|0} \end{aligned} \tag{2.14}$$

On the other hand, the test results in patients with the disease ($\boldsymbol{p_1}$) are assumed to be independent, just as in model 1.

Then, replacing $\boldsymbol{p_0}$ from (2.14), and $\boldsymbol{p_1}$ from (2.6), the likelihood is written in terms of a set of parameters $\theta_{M_3} = (S_1, S_2, C_1, C_2, P_{00|0})$:

$$\begin{aligned} L(\mathbf{n}; \theta_{M_3}, \mathbf{y}) \propto\ & \left(\pi^Y(1-\pi)^{N-Y}\right) \\ & \left({S_1}^{(y_{11}+y_{10})}(1-S_1)^{(y_{01}+y_{00})}\right)\left({S_2}^{(y_{11}+y_{01})}(1-S_2)^{(y_{10}+y_{00})}\right) \\ & (1-C_1-C_2+P_{00|0})^{(n_{11}-y_{11})}(C_2-P_{00|0})^{(n_{10}-y_{10})} \\ & (C_1-P_{00|0})^{(n_{01}-y_{01})}P_{00|0}^{(n_{00}-y_{00})} \end{aligned}$$

The prior distribution are chosen to be the same as in the independent model(2.8) and for $P_{00|0}$, assuming that we don't know much about $P_{00|0}$, it will follow a uniform distribution:

$$P_{00|0} \sim Unif(C_1C_2, min(C_1, C_2)) \tag{2.15}$$

Then, the posterior distribution is written as:

$$\begin{aligned} P(\theta_{M_3}|n) \propto\ & \left(\pi^{(\alpha_\pi+Y-1)}(1-\pi)^{(\beta_\pi+N-Y-1)}\right) \\ & \left({S_1}^{\alpha_{S_1}+y_{11}+y_{10}-1}(1-S_1)^{\beta_{S_1}+y_{01}+y_{00}-1}\right) \end{aligned}$$

$$\begin{aligned}
&\left({S_2}^{\alpha_{S_2}+y_{11}+y_{01}-1}(1-S_2)^{\beta_{S_1}+y_{10}+y_{00}-1}\right)\\
&\left({C_1}^{\alpha_{C_1}-1}(1-C_1)^{\beta_{C_1}-1}\right)\left({C_2}^{\alpha_{C_2}-1}(1-C_2)^{\beta_{C_2}-1}\right)\\
&(1-C_1-C_2+P_{00|0})^{(n_{11}-y_{11})}(C_2-P_{00|0})^{(n_{10}-y_{10})}\\
&(C_1-P_{00|0})^{(n_{01}-y_{01})}P_{00|0}^{(n_{00}-y_{00})}\left(\frac{1}{min(C_1,C_2)-C_1C_2}\right)
\end{aligned} \tag{2.16}$$

It is easy to see that the conditional distribution of π, S_1, and S_2 are beta distributions. However, the conditional distributions of C_1, C_2 and $P_{00|0}$ are unknown. For those, we will use the Metropolis-Hastings algorithm. To draw samples from this posterior distribution, one will use a Gibbs Sampling algorithm that uses a Metropolis-Hastings algorithm inside similar to the one described in Model 2. This time, the Metropolis-Hastings algorithm will draw samples from the set (C_1, C_2, $P_{00|0}$). More details are given in appendix A.4.

Model 4: Positive association of test results for all patients

In this model, we assume there is a positive association between the test results of all patients, with and without the disease. In order to add this information to the model, we will add the restrictions from Model 2 (2.10) and Model 3 (2.13). Therefore, $\boldsymbol{p_1}$ and $\boldsymbol{p_0}$ are now in terms of $(S_1, S_2, P_{11|1})$ and $(C_1, C_2, P_{00|0})$ respectively as shown in (2.11) and (2.14).

Then, the likelihood function can be written in terms of $\theta_{M_4} = (S_1, S_2, P_{11|1}, C_1, C_2, P_{00|0})$:

$$\begin{aligned}
L(\mathbf{n};\theta_{M_4},\mathbf{y}) \quad \propto \quad & \left(\pi^{Y}(1-\pi)^{N-Y}\right)\\
& P_{11|1}^{y_{11}}(S_1-P_{11|1})^{y_{10}}(S_2-P_{11|1})^{y_{01}}(1-S_1-S_2+P_{11|1})^{y_{00}}\\
& (1-C_1-C_2+P_{00|0})^{(n_{11}-y_{11})}(C_2-P_{00|0})^{(n_{10}-y_{10})}(C_1-P_{00|0})^{(n_{01}-y_{01})}P_{00|0}^{(n_{00}-y_{00})}
\end{aligned}$$

As this model is similar to the rest of the models, we would assume the prior distribution from (2.8) and the prior distribution of $p_{11|1}$ and $p_{00|0}$ as in model 2 and 3, respectively:

$$\begin{aligned}
P_{11|1} \quad &\sim \quad Unif(S_1S_2, min(S_1,S_2))\\
P_{00|0} \quad &\sim \quad Unif(C_1C_2, min(C_1,C_2))
\end{aligned}$$

Then, the posterior distribution can be written as:

$$\begin{aligned}
P(\theta_{M_4}|\mathbf{n}) \quad \propto \quad & \left(\pi^{(\alpha_\pi+Y-1)}(1-\pi)^{\beta_\pi+N-Y-1}\right)\\
& P_{11|1}^{y_{11}}(S_1-P_{11|1})^{y_{10}}(S_2-P_{11|1})^{y_{01}}(1-S_1-S_2+P_{11|1})^{y_{00}}\\
& (1-C_1-C_2+P_{00|0})^{(n_{11}-y_{11})}(C_2-P_{00|0})^{(n_{10}-y_{10})}(C_1-P_{00|0})^{(n_{01}-y_{01})}P_{00|0}^{(n_{00}-y_{00})}\\
& \left({S_1}^{\alpha_{S_1}-1}(1-S_1)^{\beta_{S_1}-1}\right)\left({S_2}^{\alpha_{S_2}-1}(1-S_2)^{\beta_{S_2}-1}\right)\left(\frac{1}{min(S_1,S_2)-S_1S_2}\right)\\
& \left({C_1}^{\alpha_{C_1}-1}(1-C_1)^{\beta_{C_1}-1}\right)\left({C_2}^{\alpha_{C_2}-1}(1-C_2)^{\beta_{C_2}-1}\right)\left(\frac{1}{min(C_1,C_2)-C_1C_2}\right)
\end{aligned}$$

To sample from this posterior distribution we need to use a Gibbs Sampling and the Metropolis-Hastings algorithm. This time, only the prevalence(π) has known distribution. We use the Metropolis-Hastings algorithm twice to sample from the rest of parameters. First to sample $\{S_1, S_2, P_{11|1}\}$ and then to sample $\{C_1, C_2, P_{00|0}\}$. Details are given in appendix A.5.

Bayesian Model Averaging (BMA)

The Bayesian model Averaging (BMA) is a technique that will be useful to combine the four initial models described above and which will be useful to get better estimation of the prevalence. Bayesian Model Averaging (BMA)(Hoeting et al., 1999) starts from the following idea:

$$P(\pi|n) = \sum_{i=1}^{4} P(\pi|M_k, n)P(M_k|n)$$

where we see that the posterior distribution of the parameter of interest (π) given the data(n) is written as a weighted average of the posterior distribution of π given each model.

As it is not easy to get samples only from $P(\pi|n)$, we will get samples from the vector (θ_k, M_k), where θ_k are all the parameters within the model k. Then, the marginal distribution of π will be obtained from that chain.

Since we need to sample from (θ_k, M_k), which has different parameter for each of the models, we will use the Reversible Jump MCMC (RJMCMC) algorithm to generate the chain. The steps are detailed in appendix A.6.

The BMA approach, compared to the FE model, considers the conjoin probability of test results, $p_{11|1}$ and $p_{00|0}$, as parameters that accounts for the association of test results. In this approach they are assumed to follow a uniform distribution because it is based on the assumption that, in practice, it is difficult to obtain information on the prior distribution of the conjoint probabilities.

Furthermore, the BMA approach, in contrast to the FE approach, allows to defines prior information about the probability of the models. For example, if we are not completely sure about whether or not there is an association between test results, we could incorporate that information into the model by defining the prior probability of model 1 to be higher than the rest.

In summary, both models have their differences in terms of the parameters used in the model and their algorithms. FE approach is simply a generalization of all the scenarios in the BMA approach and needs SIR to sample from the complex posterior distributions. The BMA approach on the other hand, is a kind of expansion of the FE approach, where 4 simpler models have emerged in order to account for the uncertainty of the association between test results, and expecting to have better estimates of the prevalence. We will see later, in the simulation, whether BMA approach could have better estimations of the prevalence or not.

Chapter 3

Simulations

Six different scenarios were simulated in this section. In this simulations, for each one of the scenarios, we have used a prevalence of 30%. Sensitivity of test 1 and 2 were assumed to be 0.90 and 0.75 respectively while the specificities of test 1 and test 2 were 0.65 and 0.80 respectively.

Table 3.1: Parameters used in the simulations

		Prior Parameters	
Parameter	Value	α	β
π	0.30	1	1
S_1	0.90	90	10
C_1	0.65	65	35
S_2	0.75	60	20
C_2	0.80	80	20

Except for the parameters that account for the association between test results ($covS$ and $covC$), all parameters are the same between scenarios. The parameters considered for the simulations are shown in the table 3.2.

Table 3.2: Simulated parameters of association ($covS$ and $covC$) for each simulated scenario

Parameter	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
$covS$	0	0.065	0	0.065	0	0.065
$covC$	0	0	0.120	0.120	0.050	0.050

At each simulated scenario, the same prior distributions were considered for the parameters of interest. The parameters used in each scenario alongside to the prior parameters that were used in the simulations are specified in table 3.1. We can see that for the prevalence, a non informative prior distribution was used ($Beta(\alpha = 1, \beta = 1)$).

At each simulation, a chain of 20,500 observations was drawn, and after the removal of the first 500, and sampling each 20 observations, we ended up with a chain of 1,000 not correlated observations.

The performance of the models was assessed by looking at the length and coverage of the 95% high posterior density (HPD) interval and the RMSE for the median estimation (3.3). These results show that the FE model has good coverage in all the simulated cases with more than 95% coverage except in scenario 1 where there is a slightly reduced coverage (94.8%). The BMA approach, on the other hand, has good coverage (more than 95%) in scenario 1

and scenario 2; however, the coverage of the BMA for scenario 3 and 4 are both less than 80% which suggests that this approach does not have good coverage in scenarios where the association between tests are significatively high.

In scenario 5 and 6, the real value of *covC* was set to an smaller value, and in those cases, the BMA approach reached better coverage.

Table 3.3: Performance indicators for both models at different simulated scenarios

Scenario	covS	covC	Model	Coverage HPD(%)	Length HPD(%)	RMSE.Median
1	0	0	BMA	95.80	23.05	1.19
			FE	94.80	26.22	1.35
2	0.065	0	BMA	96.00	21.92	1.16
			FE	97.00	26.12	1.23
3	0	0.120	BMA	78.60	19.23	1.72
			FE	97.80	27.72	1.28
4	0.065	0.120	BMA	78.20	18.89	1.68
			FE	96.20	29.66	1.42
5	0	0.050	BMA	94.20	20.96	1.21
			FE	98.80	26.30	1.16
6	0.065	0.050	BMA	88.20	20.18	1.42
			FE	98.20	26.74	1.20

The results also show that BMA results show smaller HDP intervals than FE model and RMSE. The problem with the BMA approach appears when the association between tests for non disease subject is somewhat higher. As we could see, in scenario 3.2 the *covC* has been reduced on purpose and the result is that BMA model get good coverage in that case.

Furthermore, we analyzed the distribution of posterior weights for the simulation of the BMA approach. At each one of the simulations, the posterior weights of the four models were calculated. In each scenario, a boxplot shows the distribution of the posterior weights by model in the 1000 simulations (Figure 3.1). At each scenario, we expected high weight in the model that represents the true association in the simulated data (i.e., for scenario 1, it is expected to find high weight in model 1 and small weights in the rest of the models as it is represents the real structure of association between tests). However, in scenario 1, even though there is a high weight in model 1, the remaining models also have significant weights. In scenario 2, the weights are well distributed between the four models which is not what we expected based on how we simulated the data. In scenario 3 and scenario 4, the weights are not distributed as we expected (i.e. high weights in model 3 for scenario 3 and high weight in model 4 for scenario 4). In those scenarios, the weights are high in both model 3 and model 4 which is not totally consistent with the data we simulated.

The results on the distribution of the model weights show that the BMA model is not always adequate to identify the case of association between diagnostic tests. Therefore, it could lead to inappropriate results if it is used to identify the underlying structure of association between diagnostic tests.

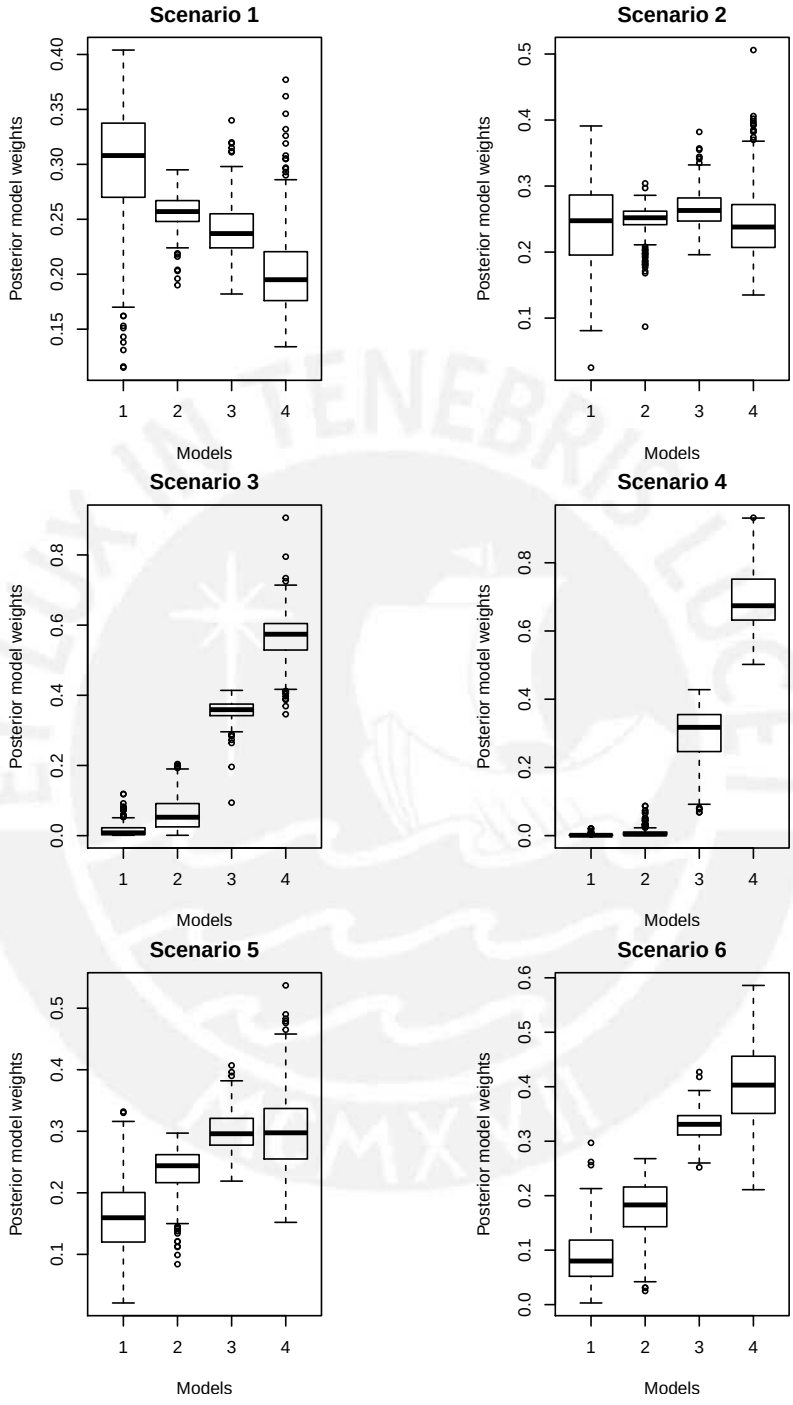

Figure 3.1: Distribution of posterior weights by model for each simulated scenario

Chapter 4

Applications

Strongyloides Infection

In this section, we will study the case of the Strongyloides infection which was first studied by Lawrence et al. (1995). In this problem the prevalence is estimated by using two standard diagnostic tests: Stool examination and Sereologic test. Both tests were applied to 162 Cambodian refugees arriving in Montreal, Canada, between July 1982 and February 1983.

The parameters for the prior distributions are from Lawrence et al. (1995). They are summarized in table 4.1. It was assumed a non informative prior for the prevalence (where the parameters α and β are both equal to one), and the prior parameters for the sensitivity and specificity are based on prior knowledge.

Table 4.1: Prior parameters α and β alongside with their mean and precision.

Parameter	Mean μ	Precision ϕ	Alpha α	Beta β
π	0.5	2	1	1
S_1	0.8	27.45	21.96	5.49
C_1	0.7	5.86	4.1	1.76
S_2	0.25	17.75	4.44	13.31
C_2	0.95	75	71.25	3.75

With the priors specified above, the BMA and FE models were used to estimate the prevalence. The results are described in tables 4.2 and 4.3:

Table 4.2: Estimation of the prevalence of Strongyloides by using the BMA and FE approach

	BMA Approach			Fixed Effects Approach		
	Median	95% C.I.		Median	95% C.I.	
	$P_{50\%}$	$P_{2.5\%}$	$P_{97.5\%}$	$P_{50\%}$	$P_{2.5\%}$	$P_{97.5\%}$
π	0.817	0.559	0.978	0.874	0.561	0.992
S_1	0.880	0.770	0.955	0.820	0.729	0.918
C_1	0.759	0.358	0.980	0.700	0.312	0.942
S_2	0.295	0.227	0.403	0.275	0.200	0.381
C_2	0.963	0.898	0.991	0.933	0.853	0.979
$CovS$	0.018	0.000	0.048	0.031	0.005	0.060
$CovC$	0.000	0.000	0.042	0.019	0.001	0.074

C.I.: Credible Interval

The estimation of the prevalence using the BMA approach is 81.7% while using the FE approach, the estimate of the prevalence was 87.4%, a difference of 5.7% between them. On

Table 4.3: Details of the estimation of Strongyloides infection prevalence using the BMA approach

	Model 1			Model 2			Model 3			Model 4		
	Median	$P_{2.5\%}$	$P_{97.5\%}$	Median	$P_{2.5\%}$	$P_{97.5\%}$	Median	$P_{2.5\%}$	$P_{97.5\%}$	Median	$P_{2.5\%}$	$P_{97.5\%}$
π	0.781	0.545	0.899	0.834	0.561	0.976	0.775	0.512	0.892	0.840	0.564	0.986
S_1	0.907	0.828	0.958	0.872	0.762	0.952	0.908	0.827	0.958	0.865	0.763	0.952
C_1	0.766	0.403	0.977	0.759	0.367	0.983	0.745	0.374	0.982	0.758	0.348	0.979
S_2	0.310	0.245	0.432	0.292	0.216	0.396	0.309	0.242	0.419	0.288	0.226	0.394
C_2	0.970	0.921	0.992	0.963	0.900	0.992	0.966	0.907	0.989	0.958	0.898	0.991
$Covs$	-	-	-	0.023	0.003	0.049	-	-	-	0.023	0.002	0.051
$Covc$	-	-	-	-	-	-	0.011	0.000	0.050	0.011	0.000	0.053
$P(M\|\mathbf{n})$	0.114			0.375			0.139			0.371		

the other hand, the length of the credible interval for the BMA approach is $0.987 - 0.559 = 0.419$ while the length for the FE approach is $0.992 - 0.561 = 0.431$.

Based on the results of our simulations in the previous section, we expected the BMA approach to have more error in its estimates and less coverage for some scenarios. Therefore, in this case, we consider that the FE approach is a better estimation for the prevalence of strongyloides disease, which is estimated to be around 87.4% with a 95% credible interval of 56.1% - 99.2%.

Chronic Kidney Disease

Estimation of chronic kidney disease (CKD) prevalence is performed using data from the CRONICAS cohort study group, which was previously used for an estimation of CKD by Francis et al. (2015). The dataset has test results from 404 adults in Peru where the mean age was 55 years (sd=12.972); fifty percent were males, and they came from two cities in Peru: 50.2% from Lima and 49.8% from Tumbes.
To define CKD, we used the definition given by *KDIGO 2012, Clinical Practice Guideline for the Evaluation and Managment of Chronic Kidney Disease* (2013) that states that a patient has CKD when either of these conditions are met:

1. The Protein Excretion Rate (proteinuria) is more than 150 mg/24h, or
2. The Glomerular Filtration Rate is less than 60 $ml/min/1.73\ m^2$.

Francis et al. used the protein-to creatinine ratio (PCR) and the estimated Glomerular Rate(eGFR), both measured from a sample of early morning urine. The two used test were: (1) the PCR $\geq$ 15 mg/mol and (2) eGFR $<$ 60 ml/min/1.73 m^2. These authors took into account a 'worst case' scenario, where a patient was considered to have the disease when either one of the test resulted positive.

In this study, however, we are taking into account the sensitivity and specificity of the tests to have a better estimation of the CKD prevalence. In addition, we are using a variation of the first test which is considered to be the optimal test to detect when the proteinuria is more than 150 mg/24h.

An study of the PCR and the optimal cut-offs for detecting more than 150mg/24h of proteinuria is described by Guy et al. (2009) in which they considered that a cut-off of 23 mg/mmol (or 230 mg/g) in PCR is an optimal cut-off to identify patients with proteinuria $\geq$ 150 mg/24h. We used this test instead of the one used by Francis et. al. where the cut-off

was 15 mg/mmol (or 150 mg/g). The associated sensitivity and specificity for this optimal test are 78% and 79%, respectively. On the other hand, the second evidence of CKD is a Glomerular Filtration Rate greater than 60 ml/min/1.73 m^2 which is estimated by using a second test, the same that was use by Francis et al., which is the estimated Glomerular Filtration Rate (eGFR) with a cutoff of 60, i.e. eGFR$<$ 60 ml/min/1.73 m^2. The accuracy of this tests was studied by Murata et al. (2011) where they found this test has 50% sensitivity and 98% specificity.

The results of both tests for this study are showed in Table 4.4 (using the cutoff points described by Francis et al., 2015) and Table 4.5 (using a different cutoff point for test 1, described by Guy et al., 2009).

Table 4.4: Test results used in this study to estimate CKD prevalence

	eGFR$<$ 60 ml/min/1.73 m^2		
	Positive	Negative	Total
PCR $\geq$ 23 mg/mmol Positive	2	26	28
Negative	6	370	376
Total	8	396	404

Table 4.5: Test results used in Francis et al. (2015) to estimate CKD prevalence

	eGFR$<$ 60 ml/min/1.73 m^2		
	Positive	Negative	Total
PCR $\geq$ 15 mg/mmol Positive	4	60	64
Negative	4	336	340
Total	8	396	404

Due to the fact that we only know the mean estimations for sensitivities and specificities, we specify the priors in terms of the mean ($\mu = \frac{\alpha}{\alpha+\beta}$) and precision (the inverse of the variance, and for this case $\phi = \mu(1-\mu)$). Assuming a precision equals to 5 for all parameters (Table 4.6). In this case, we choose a non informative prior for the prevalence due to the fact that we do not have reliable information about the prevalence of this disease in Peru.

Table 4.6: Prior parameters and their quantiles.

		Parameters		Quantiles	
	Mean	α	β	$P_{2.5\%}$	$P_{97.5\%}$
π	0.500	1.00	1.00	0.02	0.98
S_1	0.780	3.90	1.10	0.40	0.99
C_1	0.790	3.95	1.05	0.40	0.99
S_2	0.500	2.50	2.50	0.12	0.88
C_2	0.980	4.90	0.10	0.83	1.00

To obtain a sample from the posterior distribution, it was simulated 1,000 data points from the posterior density. In order to get the 1,000 observations, we run a chain with 20,500 iterations from which we discard the first 500 and select one sample every 10 observations.

Table 4.7 shows the estimation of CKD prevalence by combining the four models using the Bayesian Model Averaging described in the previous section, and the estimation using the FE approach.

The results show that the estimation based on BMA an FE approach of the CKD prevalence is far less than the estimation described by Francis et al. (2015). The BMA estimation

Table 4.7: Estimation using BMA and Fixed Effect Approach

	BMA Approach			Fixed Effects Approach		
	Median	95% C.I.		Median	95% C.I.	
	$P_{50\%}$	$P_{2.5\%}$	$P_{97.5\%}$	$P_{50\%}$	$P_{2.5\%}$	$P_{97.5\%}$
π	0.015	0.001	0.112	0.010	0.000	0.072
S_1	0.695	0.243	0.985	0.667	0.266	0.936
C_1	0.939	0.917	0.985	0.931	0.894	0.965
S_2	0.437	0.066	0.906	0.450	0.108	0.823
C_2	0.985	0.972	1.000	0.975	0.926	0.994
$CovS$	0.000	0.000	0.151	0.046	0.002	0.178
$CovC$	0.000	0.000	0.007	0.004	0.000	0.021

C.I.: Credible Interval

Table 4.8: Details of the estimation of CKD prevalence by using the BMA Approach.

	Model 1			Model 2			Model 3			Model 4		
	Mean	$P_{2.5\%}$	$P_{97.5\%}$	Mean	$P_{2.5\%}$	$P_{97.5\%}$	Mean	$P_{2.5\%}$	$P_{97.5\%}$	Mean	$P_{2.5\%}$	$P_{97.5\%}$
π	0.017	0.002	0.113	0.017	0.002	0.121	0.011	0.000	0.107	0.012	0.000	0.099
S_1	0.771	0.249	0.993	0.615	0.243	0.948	0.788	0.245	0.993	0.638	0.238	0.949
C_1	0.940	0.919	0.980	0.939	0.919	0.990	0.936	0.916	0.982	0.937	0.915	0.981
S_2	0.421	0.087	0.877	0.441	0.064	0.934	0.438	0.060	0.873	0.453	0.045	0.903
C_2	0.986	0.974	1.000	0.986	0.975	1.000	0.982	0.970	0.997	0.982	0.970	0.999
$Covs$	-	-	-	0.027	0.000	0.170	-	-	-	0.026	0.000	0.190
$Covc$	-	-	-	-	-	-	0.002	0.000	0.009	0.002	0.000	0.008
$P(M\|\mathbf{n})$	0.276			0.305			0.211			0.208		

was 1.5% while using the FE approach was 1%. In this case, similar to the first application, the length of the credible interval for the BMA is bigger than the one for the FE approach.

They considered a cutoff of 15 mg/mmol, this cutoff is used to detect patients that are at some risk of having the CKD disease (KDIGO (2012)). As this is a cutoff that detects the risk of having the disease, the estimation of the real prevalence could be overestimated.

They used a worst case scenario to estimate the prevalence, which means that they consider a subject as having the disease if at least it has been positive in one of the two diagnostic tests. In addition, the sensibility and specificity were not considered at any point in the estimation, while on the present study, both information, related to the sensibility and specificity, were taken into account.

Finally, our application considered a more restrictive cutoff point for the diagnostic test: To detect the proteinuria $\geq 150mg/24h$, we are using now PCR $\geq 23mg/mmol$ instead of PCR $\geq 15mg/mmol$ which was used in Francis et al. (2015).

Based on the results of the simulations, we consider that the estimations found in the FE approach are more accurate than with the BMA approach. Therefore, we consider that the prevalence of CKD is around 1% with a credible interval between 0% and 7.2%.

Chapter 5

Conclusions

5.1 Conclusions

- This work describes two Bayesian approaches to estimate the prevalence of a disease in the case of outcome misclassification due to the use of two imperfect diagnostic tests.
- For both of these approaches, there is no need to make assumptions about the association of test results between patients. Therefore, both models can be used even if there is uncertainty about the type of association between the available tests.
- Based on our simulations results, we found that the BMA approach, in some cases, could lead to inadequate credible intervals of the disease prevalence. On the other hand, the FE approach shows robustness under different cases of association (i.e. credible intervals show good coverage under different scenarios).
- In the application of both approaches to the estimation of the prevalence of chronic kidney disease prevalence, we found that the estimated prevalence was around 1% (95% CI: $0\% - 7.2\%$) which differs from the estimations described by Francis et al. (2015).

5.2 Further research

- . Our results on the credible intervals still have wide credible intervals. Dendukuri et al. (2004) showed the importance of sample size to achieve smaller intervals whenever the independence assumption is met. We are testing that hypothesis under different scenarios of association between the test results.
- Gustafson (2005) observed that sample size was important in order to identify the underline association. In our simulations, we consider a sample size of 200 individuals for all scenarios. A step forward can be to compare our results with other scenarios with a larger sample size. For example, in our simulations, we found that the BMA approach had limitations to identify the association between the test results and therefore, it would be interesting to study whether a larger sample size could help us in this problem.
- Gustafson (2005) and Berkvens et al. (2006) agree that prior information is important when considering a Bayesian approach to estimate the prevalence of a disease. In this article, we used the same prior information for the BMA and FE approach in order to make a fair comparison between them. However, each approach has a different

parametrization of the association. We are currently working on the re-parametrization of our models in order to make a proper comparison.

- Our program has been written in R and therefore we are currently working on moving all our code to C in order to drastically reduce the computation time and also to be able to develop an R package.
- Our collaborators at the Centro de Excelencia en Enfermedades Cronicas (CRONICAS) are interesting in extending this problem to the case of combining multiple studies. This new problem has multiple challenges that are we are currently exploring.

Appendix A

Algorithms

A.1 Gibbs Sampling and SIR algorithm for Fixed Effect Model

Given the posterior distribution, the conditional distributions are recognized:

$$
\begin{aligned}
P(S_i|S_{3-i}, covS, \mathbf{y}, u_s) &\propto \prod_{t_1=0}^{1}\prod_{t_2=0}^{1}\left({S_1}^{t_1}(1-S_1)^{1-t_1}{S_2}^{t_2}(1-S_2)^{1-t_2}+(-1)^{t_1+t_2}covS\right)^{y_{t_1t_2}} \\
&\quad S_i^{\alpha_{S_i}-1}(1-S_i)^{\beta_{S_i}-1}(u_s-covS)^{\beta_{covS}-1} \\
P(C_i|C_{3-i}, covC, \mathbf{n}, \mathbf{y}, u_c) &\propto \prod_{t_1=0}^{1}\prod_{t_2=0}^{1}\left({C_1}^{1-t_1}(1-C_1)^{t_1}{C_2}^{1-t_2}(1-C_2)^{t_2}+(-1)^{t_1+t_2}covC\right)^{(n_{t_1t_2}-y_{t_1t_2})} \\
&\quad C_i^{\alpha_{C_i}-1}(1-C_i)^{\beta_{C_i}-1}(u_c-covC)^{\beta_{covC}-1} \\
P(CovS|S_1, S_2, \mathbf{y}, u_s) &\propto \prod_{t_1=0}^{1}\prod_{t_2=0}^{1}\left({S_1}^{t_1}(1-S_1)^{1-t_1}{S_2}^{t_2}(1-S_2)^{1-t_2}+(-1)^{t_1+t_2}covS\right)^{y_{t_1t_2}} \\
&\quad (u_s-covS)^{\beta_{covS}-1} \\
P(covC|C_1, C_2, \mathbf{n}, \mathbf{y}, u_c) &\propto \prod_{t_1=0}^{1}\prod_{t_2=0}^{1}\left({C_1}^{1-t_1}(1-C_1)^{t_1}{C_2}^{1-t_2}(1-C_2)^{t_2}+(-1)^{t_1+t_2}covC\right)^{(n_{t_1t_2}-y_{t_1t_2})} \\
&\quad C_i^{\alpha_{C_i}-1}(1-C_i)^{\beta_{C_i}-1}(u_c-covC)^{\beta_{covC}-1} \\
y_i|\mathbf{n}, \pi, S_1, S_2, C_1, C_2 &\sim Bin(n_i, pos_i)
\end{aligned}
$$

where:

$$
\begin{aligned}
pos_1 &= \frac{\pi P_{00|1}}{\pi P_{00|1}+(1-\pi)P_{00|0}} \quad ; \quad pos_2 = \frac{\pi P_{10|1}}{\pi P_{10|1}+(1-\pi)P_{10|0}} \\
pos_3 &= \frac{\pi P_{01|1}}{\pi P_{01|1}+(1-\pi)P_{01|0}} \quad ; \quad pos_4 = \frac{\pi P_{11|1}}{\pi P_{11|1}+(1-\pi)P_{11|0}}
\end{aligned}
$$

and $P_{ij|d}$ could be obtained from 2.2 and 2.3.

The Gibbs sampling algorithm consists of sampling from each one of the conditional distributions. However, as the distributions are not easy to recognize in this case, it will be used a sapling importance resampling algorithm(SIR) to sample from each one of the unknown distributions.

The algorithm of sapling importance resampling (SIR) consists on:

1. Draw samples from a proposal distribution g(.)
2. Compute the importance weighting for each one of the samples.
3. Re-sample from the new set of samples considering the weights calculated in 2.

A.2 Gibbs Sampling algorithm for Model 1 of BMA approach

First, start by identifying the conditional distribution of each parameter by looking at the posterior distribution. In this case, we could identify the following conditional distributions:

$$\begin{aligned} S_1|\mathbf{y} &\sim Beta(\alpha_{S_1} + y_{11} + y_{10}, \beta_{S_1} + y_{01} + y_{00}) \\ S_2|\mathbf{y} &\sim Beta(\alpha_{S_2} + y_{11} + y_{01}, \beta_{S_1} + y_{10} + y_{00}) \\ C_1|\mathbf{n},\mathbf{y} &\sim Beta(\alpha_{C_1} + n_{01} + n_{00} - y_{01} - y_{00}, \beta_{C_1} + n_{11} + n_{10} - y_{11} - y_{10}) \\ C_2|\mathbf{n},\mathbf{y} &\sim Beta(\alpha_{C_2} + n_{10} + n_{00} - y_{10} - y_{00}, \beta_{C_2} + n_{11} + n_{01} - y_{11} - y_{01}) \\ y_i|\mathbf{n},\pi,S_1,S_2,C_1,C_2 &\sim Bin(n_i, pos_i) \end{aligned} \tag{A.1}$$

where

$$\begin{aligned} pos_1 &= \frac{\pi P_{00|1}}{\pi P_{00|1} + (1-\pi)P_{00|0}} \quad ; \quad pos_2 = \frac{\pi P_{10|1}}{\pi P_{10|1} + (1-\pi)P_{10|0}} \\ pos_3 &= \frac{\pi P_{01|1}}{\pi P_{01|1} + (1-\pi)P_{01|0}} \quad ; \quad pos_4 = \frac{\pi P_{11|1}}{\pi P_{11|1} + (1-\pi)P_{11|0}} \end{aligned} \tag{A.2}$$

Finally, the conditional distribution of the prevalence is :

$$\pi|\mathbf{n}, Y, S_1, S_2, C_1, C_2 \sim Beta(\alpha_\pi + Y, \beta_\pi + N - Y)$$

Having identified the conditional distributions, it is easy to describe the Gibbs Sampling algorithm as follows:

1. Start with arbitrary values for the parameters and latent variables: $y, \pi, S_1, S_2, C_1, C_2$
2. Draw samples from each one of the conditional distribution defined in (A.1).
3. Repeat step 2 a large number of times.

The algorithm creates a chain of simulated values from the posterior distributions. After creating the chain, it is recommended to burn the first ones and just take observations jumping between observations to avoid having an autocorrelated sample.

A.3 Gibbs Sampling algorithm for Model 2

If we try to use the Gibbs Sampling algoritm to draw samples from 2.12, it is easy to define the conditional distribution for π, C_1, and C_2, but not so easy for S_1, S_2 or $P_{11|1}$ alone. They do not have an easy-to-identify conditional distribution function. Then, in this case, we can sample from the conditional distribution of the set $\{S_1, S_2, P_{11|1}\}$ instead of each one alone; and to do so we need to use the Metropolis Hasting algorithm.

The very first step is to identify the conditional distribution of each parameter and the conjoin conditional distribution for $\{S_1, S_2, P_{11|1}\}$ from 2.12. The distributions we identified are shown below.
First, the conditional distribution of the set $\{S_1, S_2, P_{11|1}\}$:

$$\begin{aligned} f(S_1, S_2, P_{11|1}|\mathbf{y}) \quad \propto \quad & \left(P_{11|1}^{y_{11}}(S_1 - P_{11|1})^{y_{10}}(S_2 - P_{11|1})^{y_{01}}(1 - S_1 - S_2 + P_{11|1})^{y_{00}}\right) \\ & \left({S_1}^{\alpha_{S_1}-1}(1-S_1)^{\beta_{S_1}-1}\right)\left({S_2}^{\alpha_{S_2}-1}(1-S_2)^{\beta_{S_1}-1}\right)\left(\frac{1}{min(S_1,S_2) - S_1S_2}\right) \end{aligned}$$

equivalent to the folowing, which is the one we will use for the MH algorithm:

$$\begin{aligned} f(S_1, S_2, P_{11|1}|\mathbf{y}) \quad \propto \quad & P_{11|1}^{y_{11}}P_{10|1}^{y_{10}}P_{01|1}^{y_{01}}P_{00|1}^{y_{00}} \\ & [{S_1}^{\alpha_{S_1}-1}(1-S_1)^{\beta_{S_1}-1}][{S_2}^{\alpha_{S_2}-1}(1-S_2)^{\beta_{S_1}-1}]\frac{1}{min(S_1,S_2) - S_1S_2} \end{aligned}$$

And the conditional distribution of the rest of parameters:

$$\begin{aligned} C_1|\mathbf{n}, \mathbf{y} \quad &\sim \quad Beta(\alpha_{C_1} + n_{01} + n_{00} - y_{01} - y_{00}, \beta_{C_1} + n_{11} + n_{10} - y_{11} - y_{10}) \\ C_2|\mathbf{n}, \mathbf{y} \quad &\sim \quad Beta(\alpha_{C_2} + n_{10} + n_{00} - y_{10} - y_{00}, \beta_{C_2} + n_{11} + n_{01} - y_{11} - y_{01}) \\ a_i|\mathbf{n}, S_1, S_2, C_1, C_2 \quad &\sim \quad Bin(n_i, pos_i) \\ \pi|\mathbf{n}, Y, S_1, S_2, C_1, C_2 \quad &\sim \quad Beta(\alpha_\pi + Y, \beta_\pi + N - Y) \end{aligned}$$

where pos_i was specified in (A.2).

The algorithm to use is a Gibbs Sampling similar to the one described in Model 1; however, this time, we have one of the distribution to be the distribution of a set of variables $S_1, S_2, P_{11|1}$, which needs to be sampled by using a Metropolis Hasting algorithm. The MH algotithm is described in the following steps:

1. Choose a proposal distribution. Conveniently, it is proposed a Dirichlet: $\mathbf{p_1} \sim Dirichlet(1+ y_{11}, 1 + y_{10}, 1 + y_{01}, 1 + y_{00})$. Then, the probability distribution (pd) for $\mathbf{p_1}$ is:

$$g(\mathbf{p_1}) \quad \propto \quad P_{11|1}^{y_{11}}P_{10|1}^{y_{10}}P_{01|1}^{y_{01}}P_{00|1}^{y_{00}}$$

2. Draw a sample from the proposal distribution: $\mathbf{p_1}^*$, and verify if it satisfies the restriction: $p^*_{11|1} > S_1^*S_2^*$, where $S_1^* = (P^*_{11|1} + P^*_{10|1})$ and $S_2^* = (P^*_{11|1} + P^*_{01|1})$.
If $p_1{}^*$ does not satisfy the restriction, we continue drawing samples of $\mathbf{p_1}^*$ until finding one that satisfies the restriction. Once we find a sample $\mathbf{p_1}^*$ that satisfies the restriction, we can continue to the next step.

3. Update $\mathbf{p_1}$ with $\mathbf{p_1}^*$ with a probability of update defined as $PUpd = min\left(1, \frac{f(x^*)g(x|x^*)}{f(x)g(x^*|x)}\right)$; where f is the pd which we want to sample from (pd of the set $S_1, S_2, P_{11|1}$), and g is the pd of the proposal distribution (pd of $\mathbf{p_1}$).

Note that g and f could be written as follows:

$$\begin{aligned} g(x) &\propto P_{11|1}^{y_{11}} P_{10|1}^{y_{10}} P_{01|1}^{y_{01}} P_{00|1}^{y_{00}} \\ f(x) &\propto P_{11|1}^{y_{11}} P_{10|1}^{y_{10}} P_{01|1}^{y_{01}} P_{00|1}^{y_{00}} [{S_1}^{\alpha_{S_1}-1}(1-S_1)^{\beta_{S_1}-1}] \\ &\quad [{S_2}^{\alpha_{S_2}-1}(1-S_2)^{\beta_{S_1}101}] \frac{1}{min(S_1,S_2)-S_1S_2} \end{aligned}$$

Then, replacing f and g in $PUpd$, the probability of updating p_1 with p_1^* is summarized in:

$$PUpd = min\left(1, \prod_{i=1}^{2}\left(\left(\frac{S_i^*}{S_i}\right)^{\alpha_{S_i}-1}\left(\frac{1-S_i^*}{1-S_i}\right)^{\beta_{S_i}-1}\right)\left(\frac{min(S_1,S_2)-S_1S_2}{min(S_1^*,S_2^*)-S_1^*S_2^*}\right)\right) \quad \text{(A.3)}$$

4. Repeat steps 2 and 3 a large number of times.

A.4 Gibbs Sampling algorithm for Model 3

The algorithm is similar as in Appendix A.3 except that now we use the Metropolis Hasting (MH) algorithm to sample the set of variables $(C_1, C_2, P_{00|0})$.
The poposed distribution for the MH algorithm will be

$$g(\mathbf{p_0}) \propto \left(P_{11|0}^{n_{11}-y_{11}} P_{10|0}^{n_{10}-y_{10}} P_{01|0}^{n_{01}-y_{01}} P_{00|0}^{n_{00}-y_{00}}\right)$$

and the probability to update will be:

$$PUpd = min\left(1, \prod_{i=1}^{2}\left(\left(\frac{C_i^*}{C_i}\right)^{\alpha_{C_i}-1}\left(\frac{1-C_i^*}{1-C_i}\right)^{\beta_{C_i}-1}\right)\left(\frac{min(C_1,C_2)-C_1C_2}{min(C_1^*,C_2^*)-C_1^*C_2^*}\right)\right) \tag{A.4}$$

A.5 Gibbs Sampling algorithm for Model 4

This algorithm is also similar as in Appendix A.3 except that we use the Metropolis Hasting (MH) algorithm to sample two set of variables $(S_1, S_2, P_{11|1})$ and $(C_1, C_2, P_{00|0})$.

Their proposals will be:

$$\begin{aligned} g(\mathbf{p_0}) &\propto \left(P_{11|0}^{n_{11}-y_{11}} P_{10|0}^{n_{10}-y_{10}} P_{01|0}^{n_{01}-y_{01}} P_{00|0}^{n_{00}-y_{00}}\right) \\ g(\mathbf{p_1}) &\propto P_{11|1}^{y_{11}} P_{10|1}^{y_{10}} P_{01|1}^{y_{01}} P_{00|1}^{y_{00}} \end{aligned} \tag{A.5}$$

At each of the MH algorithms, the probility of update ($PUpd$) is the same as in Model 2(A.4) and Model 3(A.4).

A.6 Algorithm for Bayesian Model Averaging (BMA)

Starting in a random initial state, say (θ_k, M), the algorithm for BMA is described below:

1. Propose a new model, say M^* with probability $j(M^*|M)$ given the current model.
2. Generate a vector u from a continuous distribution $q(u|\theta_M, M, M^*)$, which conveniently will be chosen to be the posterior distribution of $P(\theta^*_M|M^*)$. The convenience to choose this distribution will be seen in step 4.
3. Set a funcion, $g_{M.M^*}$ that is bijective and converts from (θ_M, M) to (θ_{M^*}, M^*). This is: $g_{M.M^*}(\theta_M, M) = (\theta_{M^*}, M^*)$.
4. Accept the proposal move to (θ_{M^*}, M^*) with probability:

$$\alpha = min\left(1, \frac{P(n|\theta_{M^*}, M^*)P(\theta_{M^*}|M^*)P(M^*)}{P(n|\theta_M, M)P(\theta_M|M)P(M)} \frac{j(M|M^*)}{j(M^*|M)} \frac{q(u^*|\theta_{M^*}, M^*, M)}{q(u|\theta_M, M, M^*)} \left|\frac{\delta g_{M.M^*}(\theta_M, M)}{\delta \theta_M, M}\right|\right)$$

Having chosen the convenient ditribution in step 2. The probability of move can be simplified to the following expression:

$$\alpha = min\left(1, \frac{P(n|M^*)P(M^*)}{P(n|M)P(M)} \frac{j(M|M^*)}{j(M^*|M)}\right)$$

Now, if we specify the prior distribution of each model to be the same for each model, then $P(M) = \frac{1}{4}$. Also, if we take the same probability to jump to other models from whatever model we are, then $j(M^*|M) = j(M|M^*)$. Replacing it to the equation, the probability to move is even more simplified to the following:

$$\alpha = min\left(1, \frac{P(n|M^*)}{P(n|M)}\right) \tag{A.6}$$

where $P(n|M)$ is the likelihood of the current model, and $P(n|M^*)$ the likelihood of proposed model.

For example, if we jump from Model 1 to Model 2. The only parameters that change are $p_1 = (P_{11|1}, P_{10|1}, P_{01|1}, P_{00|1})$. They move from being in terms of (S_1, S_2) in model 1 to being in terms of $(S^*{}_1, S^*{}_2, P_{11|1})$ in model 2. As a result, the probability to move simplifies to:

$$\begin{aligned}
\alpha &= min\left(1, \frac{P(n|M^*)}{P(n|M)}\right) \\
&= min\left(1, \frac{K_{a^*{}_1}[P^{*y_{11}}_{11|1} P^{*y_{10}}_{10|1} P^{*y_{01}}_{01|1} P^{*y_{00}}_{00|1}]}{K_{y_{11}}[P^{y_{11}}_{11|1} P^{y_{10}}_{10|1} P^{y_{01}}_{01|1} P^{y_{00}}_{00|1}]}\right) \\
&= min\left(1, \frac{P(a|S_1, S_2, P_{11|1})}{P(a|S_1, S_2)}\right)
\end{aligned}$$

Bibliography

Albert, P. S. and Dodd, L. E. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard, *Biometrics* **60**(2): 427–435.
URL: *http://dx.doi.org/10.1111/j.0006-341X.2004.00187.x*

Berkvens, D., Speybroeck, N., Praet, N., Adel, A. and Lesaffre, E. (2006). Estimating disease prevalence in a bayesian framework using probability constraints, *Epidemiology* **17**(2): 145–153.

Black, M. A. and Craig, B. A. (2002). Estimating disease prevalence in the absence of a old standard, *Statistics in Medicine* **21**(18): 2653–2669.

Dendukuri, N. and Lawrence, J. (2001). Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests, *Biometrics* **57**: 158–167.

Dendukuri, N., Rahme, E., Belisle, P. and L., J. (2004). Bayesian sample size determination for prevalence and diagnostic studies in the absence of a gold standard test, *Biometrics* **60**: 388–397.

Francis, E. R., Kuo, C.-C., Bernabe-Ortiz, A., Nessel, L., Gilman, R. H., Checkley, W., Miranda, J. J., Feldman, H. l. and cohort Study Group, C. (2015). Burden of cronic disease in resourse limited settings from Peru: a population-based study, *BMC Nephrology* **16**(114).

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (1995). *Bayesian Data Analysis*, Chapman & Hall Texts in Statistical Science.

Gustafson, P. (2005). On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables, *Statist. Sci.* **20**(2): 111–140.
URL: *http://dx.doi.org/10.1214/088342305000000098*

Guy, M., Borzomato, J. K., Newall, R. G., Kalra, P. A. and Price, C. P. (2009). Protein to creatinine ratios in random urines accurately predict 24h protein and albumin loss in patients with the disease, *Annals of Clinical Biochemistry* **46**: 486–476.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial, *Statistical Science* **14**(4): 382–417.

KDIGO 2012, Clinical Practice Guideline for the Evaluation and Managment of Chronic Kidney Disease (2013). *Official Journal of the international society of nephrology* **3**.

Lawrence, J., Gyorkos, T. W. and Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard, *American Journal of Epidemiology* **141**(3): 263–272.

Murata, K., Baumann, N. A., Saenger, A. K., Larson, T. S., Rule, A. D. and Lieske, J. C. (2011). Relative performance of the mdrd and ckd-epi equations for estimating glomerular

filtration rate among patients with varied clinical presentations, *Clinical Journal of the American Society of Nephrology* **6**: 1963–1972.