

Haifang Ni
Utrecht University

Bayesian Methods in Veterinary Clinical Epidemiology

Cover Design: Haifang Ni & Sandra Tukker

Printing: Ridderprint | www.ridderprint.nl

ISBN 978-94-6458-187-4

© 2022 Haifang Ni

Bayesian Methods in Veterinary Clinical Epidemiology

Bayesiaanse Methoden in Veterinaire Klinische Epidemiologie

(met een samenvatting in het Nederlands)

贝叶斯方法在兽医临床流行病学中的应用

(中文概要)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

donderdag 2 juni 2022 des ochtends te 10.15 uur

door

Haifang Ni

geboren op 21 november 1984

te Hangzhou, China

Promotoren:

Prof. dr. M. Nielen

Prof. dr. I.G. Klugkist

Contents

1. General introduction

Part I. Diagnostic test evaluation with covariates

2. Evaluation of Bayesian Hui-Walter and logistic regression latent class models to estimate diagnostic test characteristics with simulated data

3. Test characteristics of the tuberculin skin test and post-mortem examination for bovine tuberculosis diagnosis in cattle in Northern Ireland estimated by Bayesian latent class analysis with adjustments for covariates

Part II. Informative priors for random effects

4. Prediction models for clustered data with informative priors for the random effects: A simulation study

5. Expert opinion as priors for random effects in Bayesian prediction models: Subclinical ketosis in dairy cows as an example

Part III. Cross-species evidence

6. The effect of oral Glucosamine and Chondroitin in aged horses: A clinical trial re-analyzed with external data as prior information

7. Summarizing discussion

Reference list

English summary

Nederlandse samenvatting

中文概要

Acknowledgements

Curriculum vitae

Chapter 1

General Introduction

There is no Algebraist nor Mathematician so expert in his science, as to place entire confidence in any truth immediately upon his discovery of it, or regard it as any thing, but a mere probability. Every time he runs over his proofs, his confidence encreases; but still more by the approbation of his friends; and is rais'd to its utmost perfection by the universal assent and applauses of the learned world. Now 'tis evident, that this gradual encrease of assurance is nothing but the addition of new probabilities, and is derived from the constant union of causes and effects, according to past experience and observation.

- David Hume, 1738 <A Treatise of Human Nature>

1. Introduction

Bayesian inference is increasingly used in many research disciplines, including veterinary epidemiology (Dohoo et al., 2009). A few examples in this field are applications for diagnostic test evaluation, disease mapping, and risk assessment (e.g., Toft et al., 2005; Best et al., 2005; Ranta et al., 2005). The growing popularity of Bayesian analysis is often motivated by its advantage in handling complicated empirical problems which are difficult within the classical frequentist approach (Dohoo et al., 2009). In addition to this more pragmatic reason to adopt the Bayesian framework, others favor the Bayesian approach for its ability to include prior information when available.

Unlike the frequentist approach that assigns probabilities to random events and to the long-run frequencies, within the Bayesian framework uncertainty about the parameters is modelled using probability distributions (Gelman et al., 2008). Any Bayesian analysis starts with the specification of a prior distribution for the model parameters, that is, before observing the data. Knowledge or uncertainty about the parameters before data collection is captured in the prior. After collecting data, the prior distribution is combined with the data, resulting in the posterior distribution reflecting an updated state of knowledge.

The need to specify a prior distribution is often seen as the bottleneck of the Bayesian approach. It is considered to be subjective and the potential impact of the prior distribution on parameter estimates is seen as undesirable by opponents of the Bayesian approach. A common choice, also seen in veterinary epidemiology, is a non-informative prior distribution, also

denoted by the terms: uninformative, weakly-informative, diffuse, flat, or vague. With a diffuse prior, results of a Bayesian analysis will be dominated by the observed data and lead to estimates that are highly similar to the results of non-Bayesian estimation.

However, it is also argued that relevant background information is often available in scientific and clinical learning processes (Spiegelhalter et al., 2004). The Bayesian approach formalizes this process by offering the opportunity to include background information in informative prior distributions. By doing so, the focus is not on the results of a new study in isolation, but instead, on accumulating knowledge by combining existing information with new data. The existing background information can consist of previous research results or expert knowledge that can be elicited and expressed in the form of informative priors.

The aim of collecting and combining relevant evidence from multiple sources fits well in the practice of Evidence-based Veterinary Medicine (EBVM) and could provide practitioners with more coherent and reliable information to support clinical decisions (Badenoch & Heneghan, 2002; Cockcroft & Holmes, 2003). In this thesis, Bayesian approaches are examined and applied in veterinary epidemiological studies, both for the reason of including external evidence through prior distributions, as well as for its modelling flexibility. Before presenting the research projects, a brief introduction of Bayesian statistics is provided in Section 1.1, followed by a short outline of evidence based research and the relation to EBVM in Section 1.2.

1.1. Bayesian approach

The essence of Bayesian inference is the use of probability distributions to express uncertainty, and the updating of current probabilities (i.e., knowledge) in the light of new evidence (Spiegelhalter et al., 2004). In a Bayesian statistical analysis, uncertainty is attributed to the parameters, while the data is regarded as a fixed quantity once sampled.

Let θ denote the parameter(s) of interest and D the data. Then, the posterior distribution for θ , $p(\theta|D)$, is obtained by multiplying the prior distribution for θ , $p(\theta)$, with the likelihood of the data, $p(D|\theta)$:

$$p(\theta|D) \propto p(\theta) \times p(D|\theta). \quad (1.1)$$

If we take clinical treatment effect of an intervention as the parameter of interest and a randomized controlled trial (RCT) as the data, the posterior distribution for the treatment effect given the RCT is then proportional to the prior information for the treatment effect and the

likelihood for the RCT data (i.e., the probability of observing the RCT data given the treatment effect). The posterior distribution provides the most likely values for the treatment effect by updating the prior knowledge for the treatment effect in the light of the new RCT data. Point estimates for the treatment effect such as the posterior mean, median or mode, as well as 95% posterior central credible intervals (CCIs) can easily be calculated from the posterior distribution.

The Bayesian approach is flexible in dealing with complex models. For instance, model parameters for data with a hierarchical structure (e.g., cows clustered in herds and herds in regions, providing three levels of data) can be difficult to estimate using frequentist approaches, due to the large number of (correlated) parameters. Also, within the Bayesian framework deriving the posterior distribution and posterior estimates for complex models is often not feasible in algebraic form (Spiegelhalter et al., 2004). However, posterior distributions are relatively easily obtained using Markov chain Monte Carlo (MCMC) methods. MCMC is a collection of sampling based techniques that are easily executed given the current computational power and available Bayesian software. With MCMC sampling, analytic integration is replaced by empirical summaries of sampled values. The invention and development of MCMC methods has considerably widened the scope of Bayesian inference in real-world applications the last few decades (Dohoo et al., 2009).

1.2. Evidence-based veterinary medicine

The term evidence-based veterinary medicine (EBVM), followed from evidence-based medicine (EBM), is defined as “the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients” (Badenoch & Heneghan, 2002). While being very closely related to EBM, EBVM differs from EBM in the availability and the strength of evidence (Cockcroft & Holmes, 2003). In human medicine, there is a large and steadily growing body of evidence and clinical trials with sufficient sample sizes are carried out when needed. Veterinary medicine, however, is often faced with a relative scarcity of available evidence, and the strength of evidence that is available is often considered weaker (Cockcroft & Holmes, 2003).

Strength of evidence can be depicted in a pyramid, such as the one presented in Fig. 1.1. In this pyramid, levels of evidence are positioned regarding their reliability and bias (Petrie & Watson, 2013), with the most reliable and the least biased evidence at the top. The strongest level of evidence is obtained by using formal methods for the collection and aggregation of multiple studies, i.e., systematic reviews and meta-analyses. With systematic reviews, one

attempts to find and evaluate all high quality research results for a specific research question. Systematic reviews are often followed by meta-analyses where results are synthesized quantitatively. At the second level of the pyramid, one finds RCTs, in which the effectiveness of an intervention in a controlled experimental context is assessed. Compared to other research designs, RCTs are considered to provide the strongest evidence as single studies. Next in the hierarchy, one finds observational studies (cohort, case control, and cross sectional) which form a fundamental part of veterinary epidemiological research. Within observational studies, the effect of risk factors or the performance of diagnostic tests, treatments or other interventions can be observed and investigated. One level below, descriptive studies, such as case reports and case series, can be found. These studies are empirical inquiries that illustrate features of a specific patient or a group of patients in the real-world clinical setting. At the bottom of the pyramid, we find subjective information such as expert opinions.



Fig. 1.1. The hierarchy of evidence illustrated as a pyramid (based on the book ‘Statistics for Veterinary and Animal Science’, page 247)

To be able to reach the strongest level of evidence, multiple high quality patient-centered studies are required. Only then, evidence synthesis tools such as systematic reviews and meta-analysis can be applied. However, in the veterinary context, available studies on the same research question are, more often than not, from diverse and weaker levels of the evidence pyramid. Still, in veterinary practice, there is a continuously growing demand for clinical decisions to be based on the combination of the best available evidence. This requires other

flexible synthesis tools. The Bayesian approach offers a promising framework for this goal. For instance, one could inform prior distributions using weaker levels of evidence, like case studies or expert knowledge, and update these priors with data from an RCT. This thesis will provide a few examples of using diverse sources of information to obtain more reliable results or better predictions, by using Bayesian models.

1.3. Aim and outline of the thesis

The aim of this thesis is to explore the use of Bayesian methods in the context of veterinary clinical epidemiology for several applications that are common in this field. The methodological research is done using simulation studies and by the (re-)analysis of empirical examples. Challenges and questions encountered at development and application of the methods are discussed.

The first part of the thesis investigates Bayesian latent class models for estimation of diagnostic test characteristics when there is no gold standard reference test with known test accuracy. Diagnostic tests are of great importance for disease detection, disease intervention and disease prevention. When there is no gold standard, the latent class model proposed by Hui and Walter (1980) is routinely used. However, this approach aggregates data to the (sub)population level and ignores individual level information. As an alternative, a logistic regression latent class model can be applied to keep the hierarchical structure of the data in the analysis (Magder & Hughes, 1997). In addition to the targeted estimates for test characteristics, with logistic regression models estimates for the odds ratios of the covariates (i.e., risk factors) can be obtained. In this part, we explore the classical Hui-Walter and the logistic regression latent class modelling methods within a Bayesian framework. In Chapter 2, both methods were investigated in simulated data scenarios with varying population characteristics. In Chapter 3, they were applied to evaluate the diagnostic performance of the tuberculin skin test and post-mortem examination for bovine tuberculosis (bTB) in abattoirs from Northern Ireland.

The second part of the thesis is about building diagnostic models for clustered animals with informative priors for the random cluster effects, commonly herds in which animals reside. When data are clustered, random effects models are routinely used in veterinary epidemiology for parameter estimation (Dohoo et al., 2009). For diagnostic and prediction models, however, random effects are often removed from prediction. In this part, we introduce a prediction modelling approach for clustered data under the Bayesian framework, where random effects are attained in the model and cluster level expert opinion for each new cluster is incorporated.

In Chapter 4, we investigated this new approach in the context of simulated data. In Chapter 5, we applied this method to an observational study for subclinical ketosis in Dutch dairy cows.

The third part of the thesis is about adding cross-species data as informative priors to an RCT to assess the treatment effect of an intervention. In veterinary medicine, for minor species particularly, the number of available studies that provide evidence on specific research questions can be scarce. Consequently, clinical conclusions must be based on limited information. Veterinarians are hence by default trained to be cross-species thinkers. In clinical practice, borrowing information from other species may happen on a daily basis in an implicit manner. In this part, we adopt the Bayesian power prior method (Ibrahim & Chen, 2000) to aggregate relevant cross-species studies after being examined and weighted by a clinical expert regarding the relevance of each study to the research question. In Chapter 6, we used this approach to re-analyze a Dutch equine clinical trial to evaluate the effect of oral Glucosamine and Chondroitin in aged horses.

Chapter 7 summarizes the findings from the studies and discusses the use of Bayesian methods in veterinary clinical epidemiology.

Part I

Diagnostic Test Evaluation with Covariates

Chapter 2

Evaluation of Bayesian Hui-Walter and logistic regression latent class models to estimate diagnostic test characteristics with simulated data

Abstract

Estimation of the accuracy of diagnostic tests in the absence of a gold standard is an important research subject in epidemiology (Dohoo et al., 2009). One of the most used methods the last few decades is the Bayesian Hui-Walter (HW) latent class model (Hui & Walter, 1980). However, HW models aggregate the observed individual test results to the population level, and as a result, potentially valuable information from the lower level(s) is not fully incorporated. An alternative approach is the logistic regression (LR) latent class model (Madgar & Hughes, 1997). In contrast to the HW model, the LR latent class model allows inclusion of multilevel data such as individual level covariates and cluster level effects.

In this study, we explored both Bayesian HW and LR latent class models within a simulation context where true disease status and true test properties were predefined. Population prevalences and test characteristics that were realistic for paratuberculosis in cattle (Toft et al., 2005) were used for the simulation. Individual animals were generated to be clustered within herds in two regions. Two tests with binary outcomes were simulated with constant test characteristics across the two regions. On top of the prevalence properties and test characteristics, one animal level binary risk factor was added to the data.

The main objective was to evaluate the performance of Bayesian HW and LR approaches in estimating test sensitivity and specificity with respect to bias and precision in simulated datasets with different population characteristics. Results from various settings showed that LR models provided more precise posterior estimates. The LR models that incorporated herd level clustering effects provided the most precise and the least biased estimates. This work illustrates that LR models are in many situations preferable over HW models for estimating test characteristics in the absence of a gold standard.

Keywords: Bayesian latent class model; sensitivity; specificity; multilevel; simulation

2.1. Introduction

The detection of disease is essential for disease control and disease intervention. An ideal situation is to use a perfect diagnostic test with both sensitivity (Se) and specificity (Sp) of 100%. However for most diseases, there are only imperfect tests available (e.g., Collins & Huynh, 2014; Johnson et al., 2019). In the absence of a perfect (gold standard) reference test, it is challenging to evaluate diagnostic accuracy of the imperfect tests. One of the methods that has often been applied the last few decades is Hui-Walter latent class modelling (Hui & Walter, 1980). This approach links the observed test results from the imperfect diagnostic tests to the unobserved (i.e., latent) disease status. Estimates of the test sensitivity, specificity and disease prevalence can be obtained by using maximum likelihood or Bayesian estimation with two or more populations with distinct prevalences (Dohoo et al., 2009). One of the limitations of this latent class method is that it aggregates the observed test results from the individual level at the population level. As a result, potentially valuable information from the lower level(s), such as clustering effects within each population and individual level covariates, is not incorporated in the model.

An alternative approach is the logistic regression (LR) latent class model which incorporates the true disease status based on imperfect test results into a LR model (Magdar & Hughes, 1997). In contrast to the Hui-Walter (HW) model, the LR latent class model allows inclusion of multilevel data. This approach has been applied under a Bayesian framework in different epidemiologic studies (e.g., McInturff et al., 2004; Koop et al., 2013; Hartnack et al., 2013; Paul et al., 2014; O'Hagan et al., 2019; Fernandes et al., 2019) and yields not only estimates for test characteristics but also estimates for the effect of the risk factors. Studies that used both HW and LR showed that LR models tended to provide more precise posterior estimates for test sensitivity and specificity (Koop et al., 2013; O'Hagan et al., 2019). However in empirical examples, evaluation of these two methods with various settings of population characteristics can be difficult concerning the amount of data collection. Furthermore, assessments of bias and precision of parameter estimates is difficult in real-world research, as the true disease status and true test characteristics are unknown.

In this study, we therefore explored Bayesian HW and LR latent class models with simulated data where true disease status and true test properties were known. The main objective was to evaluate the performance of these two approaches in estimating test characteristics with respect to bias and precision. Diverse population settings were simulated where population prevalence, the herd level clustering structure and strength of the risk factor

were varied. In addition, we examined the performance of LR models in estimating the association between the risk factor and the disease.

2.2. Materials & Methods

2.2.1. Data simulation

In order to evaluate model performance in a realistic context, we use the prevalence properties and test characteristics comparable to paratuberculosis in cattle (Toft et al., 2005). Data from two regions were artificially created with an overall animal disease prevalence of 10% for region 1 and 30% for region 2. Both regions contained 20 equal-sized herds. Within each herd, there were 100 cows which resulted in 4,000 cattle in total. Two tests with binary outcomes were generated with constant test characteristics across regions. Similar to the study by Toft et al. (2005), the two tests were conditionally independent given the true disease status, with test 1 having a 70% *Se* and a 99% *Sp*, and test 2 a 75% *Se* and a 95% *Sp*.

On top of the prevalence properties and test characteristics, one animal level covariate was added to the data. We chose a binary risk factor generated from the Bernoulli distribution with success/one probability 0.30. The true value of the odds ratio (OR) for the risk factor was set to approximately 1.5. We assumed moderate herd level clustering effects, with an intraclass correlation coefficient (ICC) of 0.20. The random herd effects were sampled from a normal distribution with a herd variance of 0.822 computed from the ICC value. The true disease probability for each animal was subsequently calculated using the LR model that included the risk factor with known OR (expressed as regression coefficient) and the random herd effects. The true binary disease outcome of each animal was then sampled from a Bernoulli distribution with its true disease probability. By adjusting the value of the fixed intercept of the logistic regression, we set the overall animal prevalence for region 1 and region 2 approximately at 10% and 30% respectively. Fig. 2.1 presents the distributions of within herd prevalences in the two regions for the default data setting.

2.2.2. Modelling approach

For the HW approach, crosstabulations based on the combinations of individual animal test results were used as input for the model. Within each population, under the assumptions of conditional independence and constant test properties across populations, the test result combinations of the two tests could be presented in a 2×2 contingency table. The populations for the HW models in our study were defined on the basis of the region ID, the herd ID or the binary risk factor. Table 2.1 presents an example within one of the two populations defined by

the region ID. Sensitivity and specificity for test 1 were denoted as Se_1 and Sp_1 , and for test 2 as Se_2 and Sp_2 . The overall animal prevalence of the population in region 1 was denoted p_1 . The probability of each of the four test result combinations was expressed as a function of sensitivity, specificity and prevalence of the population.

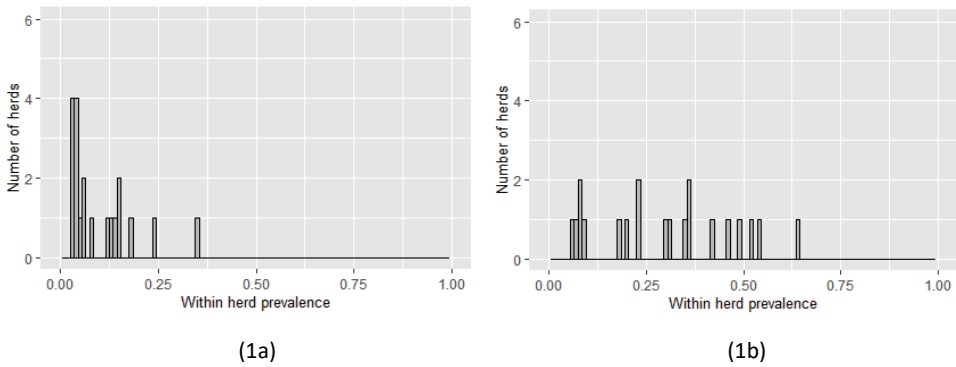


Fig. 2.1. Within herd prevalences of the 20 herds per region under an intraclass correlation coefficient (ICC) of 0.20 at the herd level: (1a) region 1 with an overall animal prevalence of 10%; (1b) region 2 with an overall animal prevalence of 30%.

Table 2.1. Probability of the 4 test result combinations within one population. The probabilities were formulated under the assumptions of the HW latent class model. Sensitivity of test 1 was denoted Se_1 and specificity Sp_1 , likewise sensitivity of test 2 was denoted Se_2 and specificity Sp_2 . The overall animal prevalence of the population in region 1 was denoted p_1 .

Population 1			
		Test 1	
		Positive	Negative
Test 2	Positive	$Se_1 Se_2 p_1 + (1 - Sp_1)(1 - Sp_2)(1 - p_1)$	$(1 - Se_1)Se_2 p_1 + Sp_1(1 - Sp_2)(1 - p_1)$
	Negative	$Se_1(1 - Se_2)p_1 + (1 - Sp_1)Sp_2(1 - p_1)$	$(1 - Se_1)(1 - Se_2)p_1 + Sp_1Sp_2(1 - p_1)$

An LR mixed model was specified for the multilevel data. When all levels of data were incorporated, i.e., region level, herd level and animal level, the regression model was expressed as follows:

$$\begin{aligned}
 \text{logit}(p_{ihr}) &= \beta_0 + \beta_1 RF_{ihr} + u_r + u_h \\
 u_r &\sim \pi(0, \sigma_r^2)
 \end{aligned} \tag{2.1}$$

$$u_h \sim \pi(0, \sigma_h^2).$$

The risk factor at individual level was denoted RF_{ihr} and the latent underlying disease probability for the observed binary outcome was denoted p_{ihr} for individual i ($i = 1, \dots, n_{ihr}$) in herd h ($h = 1, \dots, H$) from region r ($r = 1, \dots, R$). The random region effects and the random herd effects were assumed to have a normal distribution with mean zero and variance σ_r^2 for the regions and variance σ_h^2 for the herds. The disease probability of each animal p_{ihr} was estimated by the LR mixed model. Instead of population level crosstabulations as in HW models, with LR models, a crosstabulation was constructed at the individual level. In our example with two imperfect diagnostic tests, probabilities of the four test result combinations as shown in Table 2.1 could be expressed with the (latent) disease probability of each animal p_{ihr} , the sensitivity and specificity of the two tests.

2.2.3. Analysis of simulated data

Ten latent class models were specified under the Bayesian framework for estimation of the test characteristics. Table 2.2 presents the specification for these models. Three HW models were applied, stratifying on region ID (HW_r), herd ID (HW_h) and the binary risk factor (HW_RF). The crosstabulations for the test result combinations can be found in Table A1 of Appendix A. These crosstabulations were used as input data for the corresponding HW models. Seven LR models were specified with one or more levels of data (i.e., individual, herd, and population level) incorporated. Comparisons were made between the HW models and their corresponding LR models (e.g., HW_r and LR_r) as well as between the seven LR models.

For parameter estimation, non-informative beta prior distributions $beta(1,1)$ were assigned to all four sensitivity and specificity parameters and the region prevalences in both modelling approaches. For the LR models, non-informative normal prior distributions with mean 0 and a large variance $N(0, 1000)$ were specified for the regression coefficients (β_0, β_1), and non-informative inverse-gamma prior distributions $inverse-gamma(0.001, 0.001)$ were specified for the variance of the random region effects (u_r) and random herd effects (u_h). Four Markov chain Monte Carlo (MCMC) posterior chains were sampled for each model using JAGS (Plummer, 2003) called from R (R Core Team, 2016) by using the 'runjags' package (Denwood, 2016). Within each chain, the first 5,000 iterations were discarded as the burn-in phase and the subsequent 10,000 iterations were saved for parameter inferences. The convergence was visually inspected using trace plots.

Table 2.2. Ten Bayesian latent class models to estimate the sensitivity and specificity of two imperfect tests. The region level is subscripted as r , the herd level as h and RF represents the risk factor.

Model specification	
HW model	
HW_r	two equal-sized populations defined by region ID
HW_h	40 equal-sized populations defined by herd ID
HW_RF	Two unequal-sized populations defined by the binary RF
LR model	
LR_r	$\text{logit}(p_{ihr}) = \beta_0 + u_r$
LR_h	$\text{logit}(p_{ihr}) = \beta_0 + u_h$
LR_RF	$\text{logit}(p_{ihr}) = \beta_0 + \beta_1 RF_{ihr}$
LR_r_h	$\text{logit}(p_{ihr}) = \beta_0 + u_r + u_h$
LR_RF_r	$\text{logit}(p_{ihr}) = \beta_0 + \beta_1 RF_{ihr} + u_r$
LR_RF_h	$\text{logit}(p_{ihr}) = \beta_0 + \beta_1 RF_{ihr} + u_h$
$LR_RF_r_h$	$\text{logit}(p_{ihr}) = \beta_0 + \beta_1 RF_{ihr} + u_r + u_h$

2.2.4. Sensitivity analysis with varying population characteristics

Several sensitivity analyses were performed by varying the population characteristics of the simulated datasets. In order to investigate the impact of the risk factor, animals from the default data setting were permuted between the two categories of the risk factor within each region to model a higher OR (i.e., 2.7, 7.4). To examine the effect of herd level clustering, animals from the default data setting were permuted among the herds within each region resulting in a lower (0.10) or a higher ICC (0.30). Further investigations on the size of the region prevalences and the difference between the region prevalences were not done by permuting the original dataset, but based on new simulated datasets, as the overall animal prevalence changed in these settings in comparison to the default setting. Impact of the region prevalences was evaluated in a lower prevalence range (5%, 25%) and in a higher range (30%, 50%) while keeping the difference between the region prevalences 20% as in the default setting. The effect of the difference between region prevalences was evaluated as well by changing the difference to 10% (10%, 20%) and 40% (10%, 50%).

2.2.5. Model comparisons

Within each data setting, posterior estimates for the test sensitivity and specificity of the two tests were obtained for the ten Bayesian latent class models presented in Table 2.2. The estimation bias was determined by comparing the posterior median to the true value, and the precision was based on the range of the 95% posterior central credible interval (CCI). Furthermore, posterior estimates for the regression coefficient of the risk factor were evaluated in all data settings for the four LR models that included the risk factor.

2.3. Results and Discussion

The HW and LR modelling approaches were first examined in the default setting, followed by sensitivity analyses with varying population characteristics. Posterior estimates for test sensitivity and specificity of the two tests are graphically summarized in separate plots. The vertical lines in the plots represent the true values of the four test properties. The horizontal intervals represent the 95% posterior CCIs, and the squares within the interval represent the posterior medians. The first six models (above the dashed line) incorporated either only region, herd or the risk factor using the HW and LR approach. As it is possible for the LR approach to incorporate multilevel data, the last four LR models included two or all three levels of data.

For the sensitivity analyses, in order to compare between various data settings, results from the default setting are added to the plots, with the grey bars displaying the 95% CCIs and the grey squares representing the medians. Numeric summaries for the results are available in Tables A2-A7 of Appendix A.

2.3.1. Default data setting

Fig. 2.2 presents the posterior estimates for the default setting. This figure clearly shows that all ten models provided less biased and more precise estimates for test specificities than for test sensitivities. This was in line with the fact that there was less data available to estimate sensitivity than there was to estimate specificity as the overall animal prevalences were lower than 50% within the two regions (10%, 30%). In addition, the test specificities were much higher ($Sp_1 = 99\%$, $Sp_2 = 95\%$) than the test sensitivities ($Se_1 = 70\%$, $Se_2 = 75\%$).

Further, for all estimates from the ten models, the true values were located within the 95% CCIs. One can see that for the first six models, which only included region, herd or risk factor, the CCIs for the LR models were narrower than the corresponding HW models, with

the LR model that incorporated only herd level clustering effects (LR_h) showing the best precision. It is notable that the HW and LR models that only incorporated data on the risk factor produced wide CCIs. The strikingly poorer performance of these two models relative to the other eight models led to further investigation on the risk factor regarding the sample sizes and prevalence characteristics of the populations defined by the risk factor (see 2.3.6).

The last four LR models that included two or all three levels showed smaller CCIs in comparison to the first six models except for the *LR_h* model. Posterior estimates from LR models that incorporated herd level clustering effects (i.e., *LR_h*, *LR_r_h*, *LR_RF_h*, *LR_RF_r_h*) all showed similar bias and precision.

The difference in precision from the HW and LR models was also observed in other studies that applied both methods. In the study by Koop et al. (2013) for instance, when evaluating the test performance of bacteriological culture and somatic cell counts for subclinical intramammary infection in Dutch goats, authors observed narrower posterior credible intervals from the LR models in comparison to the HW model. In addition, the LR model that included most risk factors (i.e., 3) provided the narrowest credible intervals. Likewise O'Hagan et al. (2019) also reported that the LR model with risk factors showed narrower credible intervals than the HW model for sensitivity and specificity estimates of the single intradermal comparative cervical tuberculin test and post-mortem examination for bovine tuberculosis in cattle from Northern Ireland.

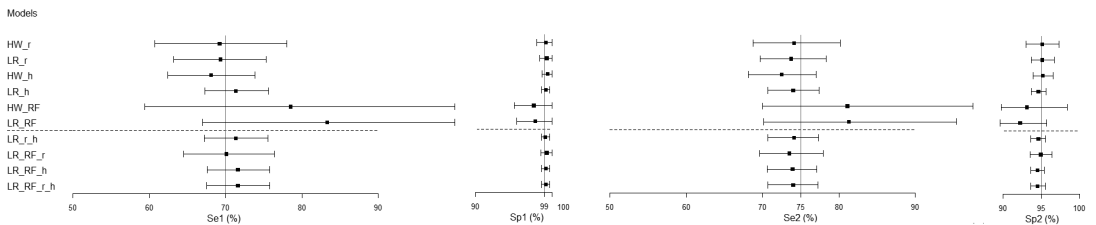


Fig. 2.2. Summary plots for posterior estimates of the test characteristics under the default data setting. Two regions contain in total 4000 cattle, with each region consisting of 20 equal sized herds and each herd consisting of 100 cows. The overall animal prevalences for the two regions are 10% and 30% respectively, and the intraclass correlation coefficient (ICC) at the herd level is 0.20. A binary animal level risk factor is present with success probability 0.30 and is associated to the true disease status with an odds ratio (OR) 1.5. True test characteristics are represented by the grey vertical lines. Ten Bayesian latent class models are specified, with three Hui-Walter (HW) models stratifying on region ID (*r*), herd ID (*h*) and the risk factor (*RF*) and

seven logistic regression (LR) models incorporating data on one or more levels (i.e., region, herd, animal level).

2.3.2. Effect of different associations between the risk factor and disease

We further investigated the impact of a stronger association between the risk factor and disease on the estimates of test characteristics. Bias and precision of the posterior estimates from each model are presented in Fig. 2.3. For models that only incorporated data on region level or herd level, results remained the same as those from the default dataset. This was expected, as for this sensitivity analysis animals simulated under the default setting were permuted between the two categories of the risk factor but remained in the same regions and herds.

For the HW model that defined populations on the basis of the risk factor, estimates were less biased and more precise when the regression coefficient for the risk factor increased from 0.40 (default) to 1, and from 1 to 2. A possible explanation for this finding was that populations stratified by the risk factor had more distinct population prevalences, when the risk factor had a stronger association with the disease status. Similar improvement was seen in the LR model that only included the risk factor. This may be because more variance of the data was explained at the animal level by the risk factor when the regression coefficient is stronger.

Defining populations on the basis of risk factors should be done with caution for HW models. Results of this sensitivity analysis indicated the necessity of checking the strength of the association between the risk factor and the disease status when using the HW approach. Posterior estimates of the HW model were less biased and more precise when the individual level risk factor had a stronger association with the disease. In veterinary epidemiology, individual level risk factors such as history of mastitis in previous lactations for bovine mastitis (Jamali et al., 2018) and body condition score for ketosis in cows (Vanholder et al., 2014), herd level risk factors such as direct cattle importation for paratuberculosis (Rangel et al., 2015) and herd size for bovine tuberculosis (Bessell et al., 2012) are found to have relatively strong association with the respective diseases within the target populations (ORs ranging from 2.06 to 19.22). However, Toft et al. (2005) pointed out that defining populations based on individual level biological risk factors such as age may violate the HW model assumption of constant test characteristics across the stratified populations due to for instance cross reactions. Higher level geographic risk factors such as zip-code and veterinary practices that result in populations with distinct prevalences are often preferred as stratifiers. Based on results of this sensitivity analysis, we recommend researchers to choose the LR approach when risk factors are available.

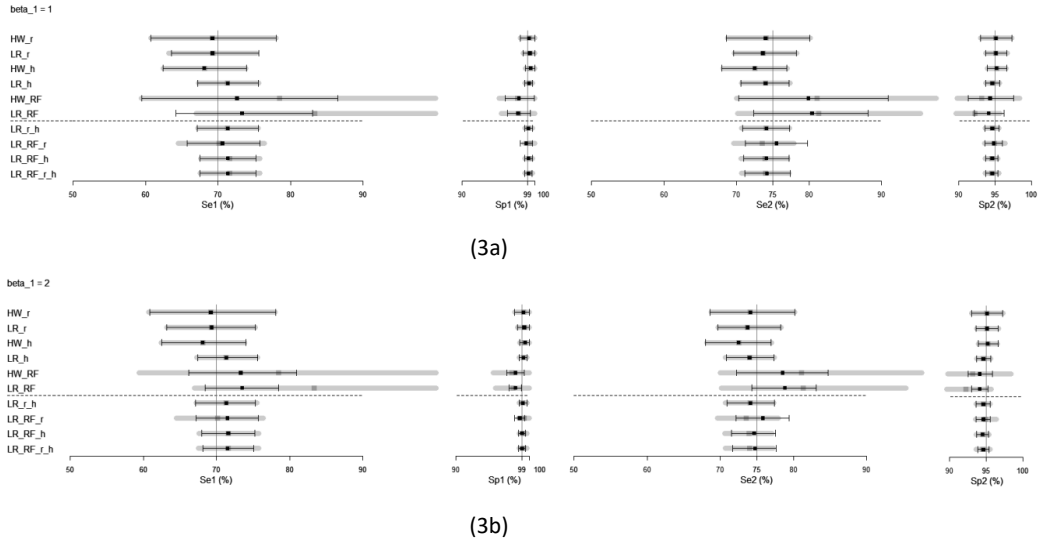


Fig. 2.3. Summary plots for posterior estimates of the evaluation of effect of strength of the association between the animal level risk factor and the disease on test sensitivity and specificity estimation. The animal level risk factor is binary and has success probability 0.30. The upper panel (3a) presents odds ratio (OR) = 2.7 ($\beta_1 = 1$) and the bottom panel (3b) presents OR = 7.4 ($\beta_1 = 2$). The grey squares and bars represent results from the default data setting OR = 1.5 ($\beta_1 = 0.4$). See Fig. 2.2 for further details of the population characteristics.

2.3.3. Effect of strength of herd level clustering (ICC)

In Fig. 2.4, results from datasets with varying strength of herd level clustering effects are presented. Animals simulated under the default setting were permuted between herds but remained in the same regions and risk factor categories. Therefore, the models that did not incorporate herd level clustering effects produced the same results as the default dataset. For models that incorporated herd level effects, when the ICC was reduced from the default 0.20 to 0.10 (4a), posterior estimates were slightly more biased and less precise. However, LR models that incorporated herd level effects as well as the risk factor and/or the region effects showed still reasonable estimates. When the ICC increased from the default 0.20 to 0.30 (4b), the bias of the posterior estimates was similar to the default setting but the precision increased slightly.

In veterinary epidemiology, herd level clustering effects have been computed for various infectious diseases and ICC values are found to vary from 0.04 (*Anaplasma marginale* in cattle) to 0.42 (*bovine viral diarrhea* in cattle), and most diseases have an ICC below 0.20 (Dohoo et

al., 2009). Results in this sensitivity analysis suggest that it might still be useful to include herd level clustering effects in the latent class models for the estimation of diagnostic test characteristics even when the ICC value is relatively low.

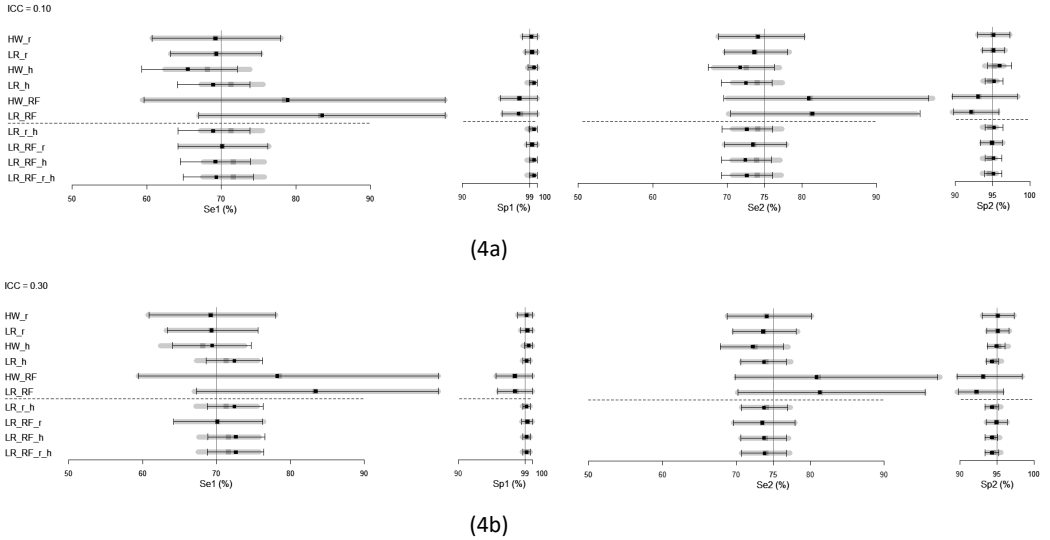


Fig. 2.4. Summary plots for posterior estimates of the evaluation of effect of strength of intraclass correlation coefficient (ICC) at the herd level on test sensitivity and specificity estimation. The upper panel (4a) presents ICC = 0.10 and the bottom panel (4b) presents ICC = 0.30. The grey squares and bars represent results from the default data setting ICC = 0.20. See Fig. 2.2 for further details of the population characteristics.

2.3.4. Effect of different values for region prevalences

Fig. 2.5 summarizes the effect of varying the overall animal prevalences of the two regions while keeping the difference constant. For this analysis, two new datasets were generated as the region prevalences were changed in comparison to the default dataset. When the region prevalences were reduced from the default 10% and 30% to 5% and 25% (5a), the precision for test sensitivities worsened whereas the precision for test specificities improved. In contrast, when the animal prevalences of the regions were increased from the default setting to 30% and 50% (5b), the precision for test sensitivities improved and for test specificities worsened. This is expected, because when the animal prevalence is lower, there is less information available in the data to estimate test sensitivity. Likewise when the animal prevalence is higher, the amount of information for estimation of test sensitivity increased.

The performance of the models regarding bias and precision from the dataset with region prevalences 5% and 25% was comparable to the default setting. This indicates that HW and LR approaches are robust if one of the populations has a low prevalence, as long as the difference between the population prevalences is distinct. The LR models that incorporated herd level clustering effects showed again the least biased and the most precise posterior estimates in comparison to other HW and LR models.

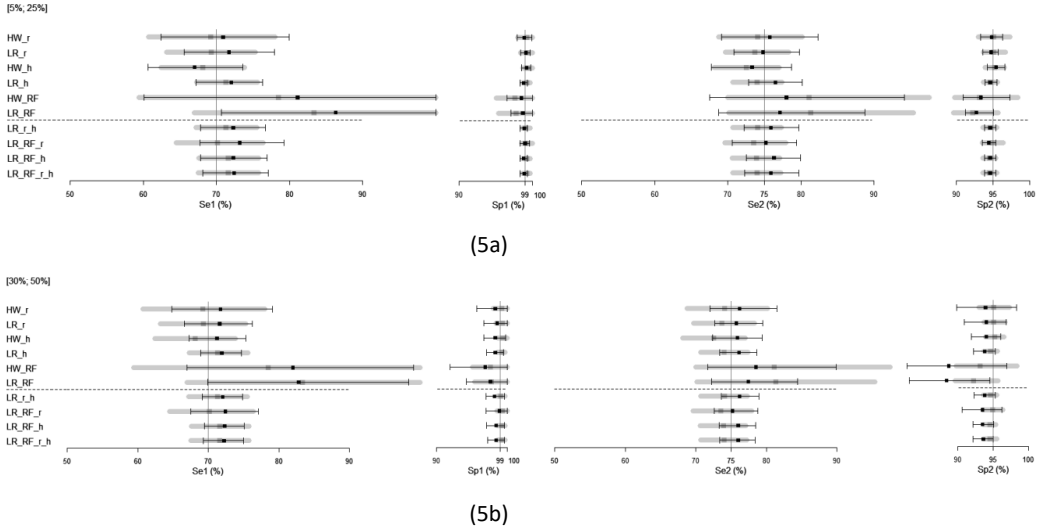


Fig. 2.5. Summary plots for posterior estimates of the evaluation of the size of region prevalences on test sensitivity and specificity estimation. The upper panel (5a) presents lower region prevalences (5%, 25%) and the bottom panel (5b) presents higher region prevalences (30%, 50%). The grey squares and bars represent results from the default data setting (10%, 30%). See Fig. 2.2 for further details of the population characteristics.

2.3.5. Effect of the difference between region prevalences

Fig. 2.6 contains results of the ten models for data settings where difference between the animal prevalences of the two regions was changed from 20% to 10% or 40%. These results were also based on two new datasets as the region prevalences were changed from the default dataset. When the region prevalences were changed from the default (10%, 30%) to (10%, 20%) (6a), precision of the estimates for test sensitivities from all models worsened, whereas precision of the estimates for test specificities improved. This was due to less data available to estimate sensitivity than to estimate specificity as the overall animal prevalence was smaller than in the

default setting. However, when the region prevalences were changed from the default (10%, 30%) to (10%, 50%) (6b), precision of the estimates for test sensitivities from all models improved, whereas precision of the estimates for test specificities worsened.

It is unclear based on the results of this sensitivity analysis, whether the change in model performance that included region information (HW_r , LR_r , LR_r_h , LR_{RF}_r , $LR_{RF}_r_h$) was fully due to the change in the prevalence difference between the two regions. In the next section we further investigated the effect of different population prevalences and sample sizes on model performance.

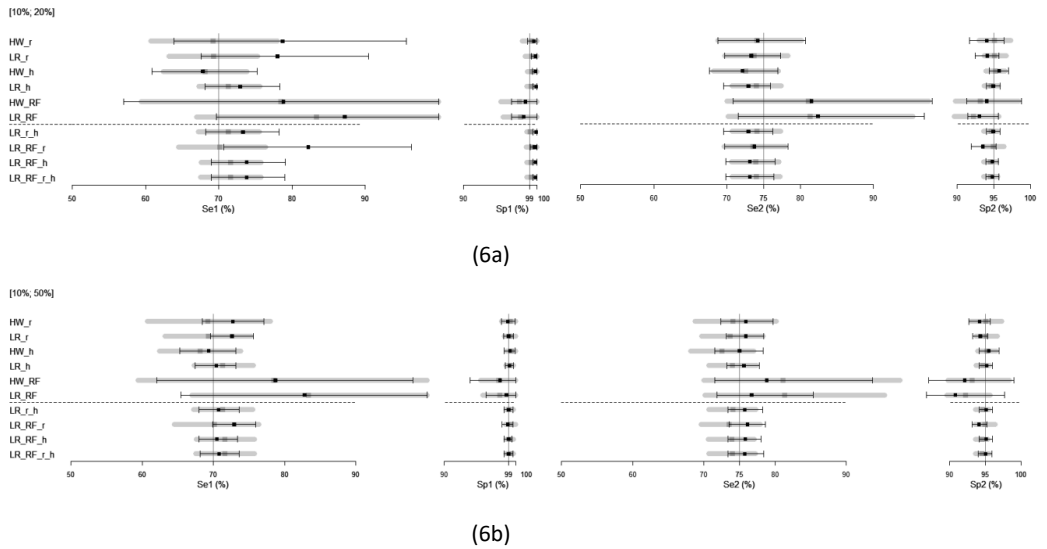


Fig. 2.6. Summary plots for posterior estimates of the evaluation of the difference in region prevalences on test sensitivity and specificity estimation. The upper panel (6a) presents 10% difference (10%, 20%) and the bottom panel (6b) presents 40% difference (10%, 50%). The grey squares and bars represent results from the default data setting (10%, 30%). See Fig. 2.2 for further details of the population characteristics.

2.3.6. Difference in population prevalences and sizes based on the risk factor

In the sensitivity analyses we presented above, it is clear that HW and LR models with only the risk factor showed the largest bias and the worst precision (Figs. 2.2-2.6). In order to grasp whether results from these two models were influenced by unequal population sizes, we simulated one more dataset where the success probability of the binary risk factor was changed

from 0.30 (default) to 0.50. With success probability 0.50, the difference between the sample sizes of the two risk factor categories was minimal.

Estimates from the *HW_RF* and *LR_RF* models became slightly better regarding bias but worse regarding precision compared to the default setting (Fig. 2.7). These results verified that unbalanced sample sizes of the stratified populations were not the main cause of the relatively poor precision of posterior estimates from the models. In order to further understand the impact of population prevalences and population sample sizes on posterior estimates for test characteristics, we listed the values of prevalences and sample sizes of populations stratified by the risk factor from all data settings. Table 2.3 shows that in the setting with the risk factor sampled from the success probability 0.50, the sample sizes were indeed similar, however the population prevalence difference was reduced from the default 6.2% to 5.3%. In fact, for most data settings, when we split the data on the basis of the risk factor, population prevalence differences were below 10%, with the exception of the two settings where the regression coefficient for the individual level risk factor was relatively strong (corresponding to an OR of 2.7 and 7.4 respectively). The effect of a small population prevalence difference on large posterior estimate credible intervals was also reported in Johnson et al. (2019).

However, it is still possible that the unbalanced sample sizes of the stratified populations also played a role in the poor performance of the HW and LR models that only used the risk factor information. Future studies should further investigate the effect of unbalanced population sample sizes on HW and LR modelling approaches.

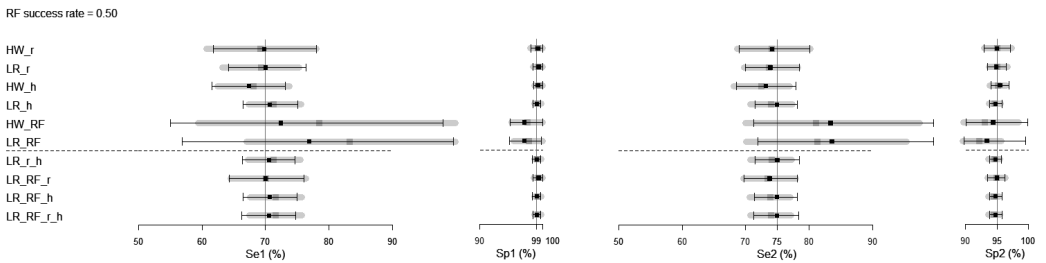


Fig. 2.7. Summary plots for posterior estimates of the test characteristics with the binary risk factor sampled from success probability 0.50 with an odds ratio of 1.5. Two regions contain in total 4000 cattle, with each region consisting of 20 equal sized herds and each herd consisting of 100 cows. The grey squares and bars represent results from the default data setting with success probability 0.30 for the risk factor. See Fig. 2.2 for further details of the population characteristics.

Table 2.3. The prevalences and sample sizes of the two stratified populations based on the individual level risk factor within each data setting.

	Population prevalence (population size)		Difference between prevalences	Figure number
	RF = 0	RF = 1		
Default	18.0% (2760)	24.2% (1240)	6.2%	Fig. 2.2
$\beta_1 = 1$	15.6% (2760)	29.4% (1240)	13.8%	Fig. 2.3a
$\beta_1 = 2$	11.1% (2760)	38.8% (1240)	27.7%	Fig. 2.3b
ICC = 0.10	18.0% (2760)	24.2% (1240)	6.2%	Fig. 2.4a
ICC = 0.30	18.0% (2760)	24.2% (1240)	6.2%	Fig. 2.4b
Region prevalences = (5%, 25%)	13.7% (2798)	18.5% (1202)	4.8%	Fig. 2.5a
Region prevalences = (30%, 50%)	37.5% (2826)	46.0% (1174)	8.5%	Fig. 2.5b
Region prevalences = (10%, 20%)	13.9% (2813)	17.9% (1187)	4.0%	Fig. 2.6a
Region prevalences = (10%, 50%)	28.3% (2761)	34.1% (1239)	5.8%	Fig. 2.6b
RF success probability = 0.50	17.2% (1933)	22.5% (2067)	5.3%	Fig. 2.7

2.3.7. Estimates of the regression coefficient

In order to obtain the estimate of the association between the individual level risk factor and disease, we examined the posterior results of the regression coefficient from the LR models that included the risk factor. In Fig. 2.8, one can see at the upper panel (8a), the true regression coefficient value was changed from 0.40 to 1 and 2, with 0.40 presenting the default setting. Results showed that the LR model without herd level clustering effects incorporated (*LR_RF*, *LR_RF_r*) tended to underestimate the association between the risk factor and the disease. This phenomenon has been shown before and was explained by Hedeker and Gibbons (2006, Chapter 9). Estimates for the regression coefficients of covariates from a fixed-effects LR model tend to be closer to zero than those resulted from a mixed-effects (i.e., random effects) LR model. Estimates from the fixed-effects are considered “population-averaged” which indicate the effect of covariates averaging over the population (in our example, over the herds), whereas estimates from the mixed-effects LR model are “cluster-specific” since they are conditional on the random clustering effects. Results in this study showed that in datasets with moderate clustering effects (ICC = 0.20), when the association between the risk factor and the disease status was weak, population averaged and cluster specific posterior estimates for the

regression coefficient of the risk factor were similar regarding bias and precision. However, when the association was stronger, the cluster specific estimates were much less biased in comparison to the population averaged estimates.

The panels 8b to 8d showed six data settings with other population characteristics, but always with a default regression coefficient of 0.40. Performance of the four LR models on estimation of the association between the risk factor and the disease was similar across various settings, with the exception of the one with low overall animal prevalence (i.e., 5%, 25%). The deviation of the estimates in this setting might be caused by a lack of information on the association between the risk factor and the disease status from region 1 as the prevalence was only 5%.

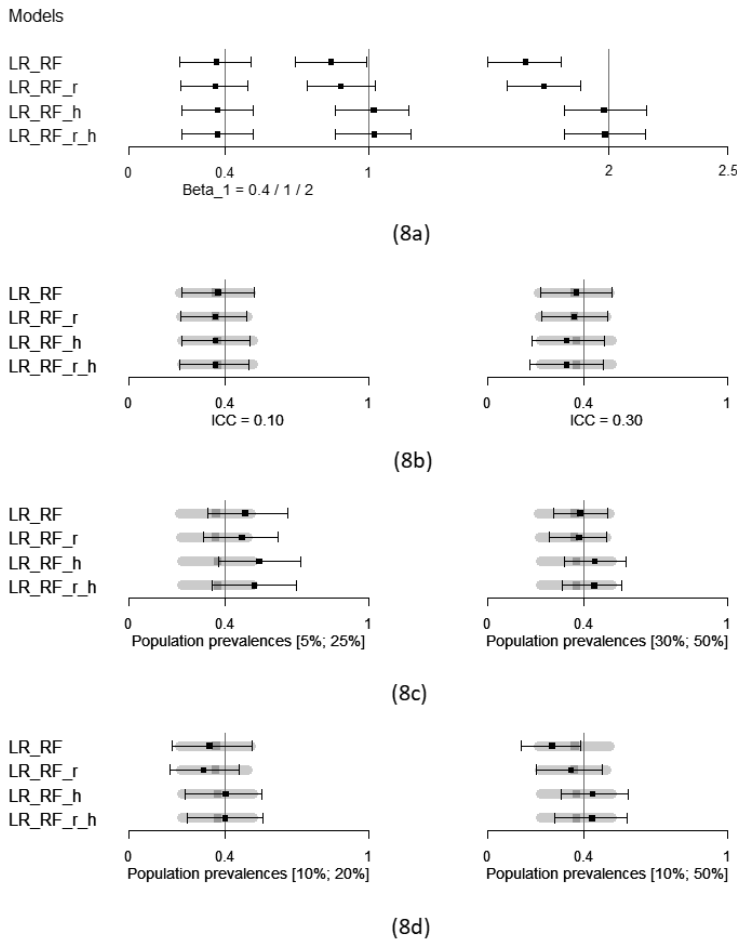


Fig. 2.8. Summary plots for posterior estimates of the regression coefficient β_1 of the risk factor in different data settings from four Bayesian logistic regression latent class models that either

include risk factor only, or include risk factor and region and/or herd level effects. The binary risk factor at the animal level has success probability 0.30. The first panel (8a) presents estimates from data settings that have true β_1 values of 0.4 (default setting), 1 and 2, corresponding to odd ratios (OR) of 1.4, 2.7 and 7.4 respectively. The following panels 8b to 8d present results from data settings that have true β_1 value of 0.4 but with either different intraclass correlation coefficient (ICC) at the herd level or different region prevalences. The grey squares and bars represent results from the default setting.

2.4. Conclusion

HW and LR latent class models are two approaches to estimate test characteristics when the true disease status is unknown and when there is no gold standard. In this simulation study, we explored these two methods under data settings with different population characteristics. We show that LR models are more precise in posterior estimates across various settings, with the LR models that incorporated herd level clustering effects presenting the least biased and most precise estimates. Results also revealed that stratifying data on the basis of an individual level risk factor for the HW modelling approach can be problematic, unless one is certain that the association between the risk factor and the outcome is strong. Altogether, this work shows that LR models are in many situations the preferable alternative to HW models to estimate test characteristics in the absence of a perfect reference test.

Appendix A: Supplementary tables**Table A1.** Three different crosstabulations for test results of Test 1 (T1) and Test 2 (T2) as input for Hui-Walter models (HW_r , HW_h , HW_{RF}).

		T1+T2+	T1+T2-	T1-T2+	T1-T2-
Region ID based (HW_r)					
Region 1		114	52	136	1698
Region 2		304	119	203	1374
Herd ID based (HW_h)					
Herd 1	Region 1	8	4	4	84
Herd 2		5	1	8	86
Herd 3		2	2	4	92
Herd 4		16	4	14	66
Herd 5		11	2	9	78
Herd 6		2	2	7	89
Herd 7		2	1	7	90
Herd 8		1	2	11	86
Herd 9		13	2	4	81
Herd 10		2	3	6	89
Herd 11		1	4	10	85
Herd 12		3	1	7	89
Herd 13		0	1	6	93
Herd 14		14	9	7	70
Herd 15		1	2	4	93
Herd 16		10	3	6	81
Herd 17		9	2	6	83
Herd 18		4	1	2	93
Herd 19		3	4	6	87
Herd 20		7	2	8	83
Herd 21	Region 2	3	2	7	88
Herd 22		20	15	13	52
Herd 23		22	8	13	57
Herd 24		25	9	15	51

Herd 25	8	8	12	72
Herd 26	3	1	4	92
Herd 27	4	2	4	90
Herd 28	14	3	6	77
Herd 29	22	6	7	65
Herd 30	22	12	13	53
Herd 31	11	2	12	75
Herd 32	3	1	6	90
Herd 33	16	2	17	65
Herd 34	6	4	10	80
Herd 35	36	12	13	39
Herd 36	3	0	5	92
Herd 37	31	13	12	44
Herd 38	20	4	11	65
Herd 39	19	9	10	62
Herd 40	16	6	13	65
Risk factor based				
(<i>HW_RF</i>)				
Category 0	259	111	232	2158
Category 1	159	60	107	914

Table A2. Default data setting ($ICC = 0.20$, $\beta_1 = 0.40$, herd size = 100, region prevalences = (10%, 30%)). Estimates of the Se and Sp of the two tests are presented with posterior median and 95% central credible interval (CCI).

Model	Estimates test characteristics (%)			
	Se_1	Sp_1	Se_2	Sp_2
True value	70	99	75	95
<i>HW_r</i>	69.2 [60.7, 78.0]	99.2 [98.0, 100]	74.1 [68.7, 80.2]	95.1 [93.0, 97.3]
<i>LR_r</i>	69.3 [63.2, 75.3]	99.3 [98.4, 100]	73.7 [69.7, 78.3]	95.1 [93.7, 96.7]
<i>HW_h</i>	68.1 [62.4, 73.8]	99.4 [98.7, 100]	72.5 [68.1, 77.0]	95.2 [93.9, 96.5]
<i>LR_h</i>	71.3 [67.3, 75.6]	99.2 [98.6, 99.7]	74.0 [70.7, 77.4]	94.6 [93.7, 95.6]
<i>HW_RF</i>	78.5 [59.4, 100]	97.6 [95.1, 100]	81.1 [70.0, 97.5]	93.1 [89.7, 98.4]
<i>LR_RF</i>	83.3 [67.0, 100]	97.8 [95.4, 100]	81.3 [70.1, 95.4]	92.2 [89.6, 95.7]
<i>LR_r_h</i>	71.3 [67.2, 75.5]	99.1 [98.6, 99.7]	74.1 [70.7, 77.3]	94.6 [93.6, 95.5]
<i>LR_RF_r</i>	70.1 [64.5, 76.4]	99.3 [98.5, 100]	73.5 [69.6, 77.9]	94.9 [93.5, 96.4]
<i>LR_RF_h</i>	71.6 [67.6, 75.8]	99.2 [98.6, 99.7]	73.9 [70.6, 77.1]	94.5 [93.6, 95.4]
<i>LR_RF_r_h</i>	71.6 [67.5, 75.8]	99.2 [98.6, 99.7]	74.0 [70.7, 77.2]	94.5 [93.6, 95.5]

Table A3. Two data settings (ICC = 0.20, herd size = 100, region prevalences = (10%, 30%)) with varying β_1 values. Estimates of the two tests (%) are presented with median and 95% central credible interval (CCI).

Model	$\beta_1 = 1$					$\beta_1 = 2$				
	Se_1	Sp_1	Se_2	Sp_2	True value	Se_1	Sp_1	Se_2	Sp_2	True value
<i>HW_r</i>	69.2 [60.7, 78.0]	99.2 [98.0, 100]	74.0 [68.6, 80.1]	95.1 [93.0, 97.3]	69.2 [60.9, 78.1]	69.2 [60.9, 78.1]	99.2 [98.0, 100]	74.1 [68.6, 80.2]	95.1 [93.0, 97.2]	69.2 [60.9, 78.1]
<i>LR_r</i>	69.2 [63.6, 75.6]	99.3 [98.4, 100]	73.6 [69.6, 78.2]	95.1 [93.7, 96.6]	69.3 [63.2, 75.3]	69.3 [63.2, 75.3]	99.3 [98.4, 100]	73.7 [69.7, 78.2]	95.1 [93.6, 96.6]	69.3 [63.2, 75.3]
<i>HW_h</i>	68.1 [62.4, 73.9]	99.4 [98.7, 100]	72.5 [68.0, 76.9]	95.2 [93.9, 96.6]	68.1 [62.5, 74.0]	68.1 [62.5, 74.0]	99.4 [98.7, 100]	72.5 [68.0, 76.9]	95.2 [93.9, 96.6]	68.1 [62.5, 74.0]
<i>LR_h</i>	71.3 [67.2, 75.5]	99.2 [98.6, 99.7]	74.0 [70.6, 77.2]	94.6 [93.7, 95.6]	71.3 [67.4, 75.6]	71.3 [67.4, 75.6]	99.2 [98.6, 99.7]	74.0 [70.8, 77.3]	94.6 [93.7, 95.6]	71.3 [67.4, 75.6]
<i>HW_RF</i>	72.6 [59.5, 86.5]	97.8 [95.9, 99.9]	79.9 [70.3, 90.9]	94.3 [91.3, 97.5]	73.3 [66.2, 80.9]	73.3 [66.2, 80.9]	98.1 [96.9, 99.3]	78.5 [72.2, 84.7]	94.1 [92.5, 95.8]	73.3 [66.2, 80.9]
<i>LR_RF</i>	73.3 [64.2, 83.0]	97.7 [96.2, 99.4]	80.4 [72.4, 88.1]	94.1 [92.1, 96.2]	73.5 [68.4, 78.5]	73.5 [68.4, 78.5]	98.1 [97.2, 98.9]	78.8 [74.3, 83.1]	94.1 [93.0, 95.2]	73.5 [68.4, 78.5]
<i>LR_r_h</i>	71.3 [67.1, 75.5]	99.1 [98.6, 99.7]	74.1 [70.8, 77.3]	94.6 [93.6, 95.5]	71.3 [67.1, 75.3]	71.3 [67.1, 75.3]	99.1 [98.6, 99.7]	74.1 [70.9, 77.4]	94.6 [93.6, 95.5]	71.3 [67.1, 75.3]
<i>LR_RF_r</i>	70.6 [65.7, 75.8]	98.8 [98.0, 99.7]	75.5 [71.2, 79.8]	94.8 [93.6, 96.0]	71.5 [67.2, 75.7]	71.5 [67.2, 75.7]	98.7 [98.0, 99.4]	75.8 [72.1, 79.4]	94.6 [93.6, 95.5]	71.5 [67.2, 75.7]
<i>LR_RF_h</i>	71.3 [67.5, 75.2]	99.1 [98.6, 99.7]	74.1 [71.0, 77.2]	94.6 [93.7, 95.4]	71.6 [68.0, 75.2]	71.6 [68.0, 75.2]	99.0 [98.5, 99.5]	74.6 [71.5, 77.5]	94.5 [93.7, 95.3]	71.6 [68.0, 75.2]
<i>LR_RF_r_h</i>	71.3 [67.5, 75.2]	99.1 [98.6, 99.6]	74.2 [71.1, 77.4]	94.6 [93.7, 95.4]	71.5 [68.1, 75.1]	71.5 [68.1, 75.1]	99.0 [98.5, 99.5]	74.7 [71.7, 77.6]	94.6 [93.8, 95.4]	71.5 [68.1, 75.1]

Table A4. Two data settings ($\beta_1 = 0.40$, herd size = 100, region prevalences = (10%, 30%)) with varying ICC values. Estimates of the two tests (%) are presented with median and 95% central credible interval (CCI).

Model	ICC = 0.10				ICC = 0.30			
	Se_1	Sp_1	Se_2	Sp_2	Se_1	Sp_1	Se_2	Sp_2
True value	70	99	75	95	70	99	75	95
<i>HW_r</i>	69.2 [60.7, 77.9]	99.2 [98.0, 100]	74.1 [68.8, 80.3]	95.1 [93.0, 97.3]	69.2 [60.9, 78.0]	99.2 [98.0, 100]	74.1 [68.7, 80.2]	95.1 [93.0, 97.3]
<i>LR_r</i>	69.3 [63.2, 75.4]	99.3 [98.4, 100]	73.6 [69.6, 78.1]	95.1 [93.6, 96.6]	69.3 [63.3, 75.6]	99.3 [98.4, 100]	73.6 [69.5, 78.1]	95.1 [93.6, 96.6]
<i>HW_h</i>	65.5 [59.3, 72.1]	99.6 [98.8, 100]	71.7 [67.4, 76.3]	95.9 [94.3, 97.5]	69.4 [64.0, 74.7]	99.5 [98.9, 100]	72.2 [67.8, 76.4]	94.9 [93.7, 96.1]
<i>LR_h</i>	68.9 [64.1, 73.8]	99.6 [99.0, 100]	72.5 [69.2, 76.0]	95.2 [94.0, 96.4]	72.4 [68.6, 76.2]	99.2 [98.7, 99.7]	73.7 [70.6, 76.8]	94.3 [93.5, 95.2]
<i>HW_RF</i>	78.9 [59.6, 100]	97.6 [95.1, 100]	80.9 [69.5, 97.0]	93.0 [89.6, 98.3]	78.2 [59.4, 100]	97.6 [95.1, 100]	80.8 [69.8, 97.2]	93.1 [89.6, 98.4]
<i>LR_RF</i>	83.5 [67.0, 100]	97.5 [95.3, 100]	81.4 [70.4, 95.8]	92.1 [89.7, 95.8]	83.4 [67.3, 100]	97.6 [95.3, 100]	81.3 [70.2, 95.5]	92.2 [89.7, 95.8]
<i>LR_r_h</i>	68.9 [64.2, 73.8]	99.6 [98.9, 100]	72.6 [69.3, 76.1]	95.2 [94.0, 96.4]	72.4 [68.7, 76.3]	99.2 [98.7, 99.7]	73.7 [70.7, 76.9]	94.3 [93.4, 95.2]
<i>LR_RF_r</i>	70.1 [64.2, 76.2]	99.3 [98.5, 100]	73.4 [69.6, 77.9]	94.9 [93.4, 96.3]	70.1 [64.2, 76.2]	99.3 [98.5, 100]	73.5 [69.6, 77.9]	94.9 [93.5, 96.4]
<i>LR_RF_h</i>	69.2 [64.5, 73.9]	99.6 [99.0, 100]	72.4 [69.2, 75.9]	95.1 [94.0, 96.2]	72.6 [68.8, 76.5]	99.2 [98.7, 99.7]	73.7 [70.6, 76.8]	94.3 [93.4, 95.1]
<i>LR_RF_r_h</i>	69.3 [64.9, 74.3]	99.6 [99.0, 100]	72.6 [69.2, 76.1]	95.1 [93.9, 96.2]	72.6 [68.7, 76.4]	99.2 [98.7, 99.7]	73.8 [70.7, 76.8]	94.3 [93.4, 95.2]

Table A5. Two data settings ($ICC = 0.20$, $\beta_1 = 0.40$, herd size = 100) with varying region prevalence values. Estimates of the two tests (%) are presented with median and 95% central credible interval (CCI).

Model	Region prevalences = (5%, 25%)				Region prevalences = (30%, 50%)			
	Se_1	Sp_1	Se_2	Sp_2	Se_1	Sp_1	Se_2	Sp_2
True value	70	99	75	95	70	99	75	95
<i>HW_r</i>	70.9 [62.4, 79.9]	98.9 [97.9, 99.9]	75.7 [69.1, 82.3]	94.8 [93.3, 96.3]	71.7 [64.8, 79.1]	98.3 [95.7, 100]	76.2 [72.0, 81.5]	93.9 [89.8, 98.3]
<i>LR_r</i>	71.7 [65.6, 77.9]	99.1 [98.4, 99.7]	74.8 [70.8, 79.8]	94.7 [93.6, 95.7]	71.6 [66.6, 76.2]	98.6 [96.7, 100]	75.7 [72.7, 79.5]	94.0 [90.9, 96.8]
<i>HW_h</i>	67.0 [60.6, 73.6]	99.2 [98.5, 99.8]	73.3 [67.7, 78.7]	95.4 [94.2, 96.6]	71.2 [67.3, 75.3]	98.3 [96.7, 99.8]	75.9 [72.4, 79.4]	94.0 [91.9, 96.1]
<i>LR_h</i>	72.0 [67.2, 76.3]	98.8 [98.3, 99.3]	76.5 [72.8, 80.2]	94.6 [93.9, 95.5]	71.9 [68.9, 74.7]	98.3 [97.1, 99.5]	76.1 [73.4, 78.6]	93.8 [92.2, 95.3]
<i>HW_RF</i>	81.1 [60.1, 100]	98.5 [96.5, 100]	78.0 [67.5, 94.1]	93.3 [90.9, 97.3]	82.0 [67.0, 99.1]	96.9 [91.9, 100]	78.5 [71.7, 89.9]	88.7 [82.8, 96.9]
<i>LR_RF</i>	86.3 [70.7, 100]	98.7 [97.1, 100]	77.1 [68.7, 88.7]	92.7 [91.2, 95.1]	82.8 [69.9, 98.4]	97.6 [94.2, 100]	77.4 [72.2, 84.4]	88.4 [83.1, 94.5]
<i>LR_r_h</i>	72.3 [67.8, 76.7]	98.9 [98.3, 99.3]	75.9 [72.2, 79.7]	94.6 [93.8, 95.4]	72.0 [69.1, 74.8]	98.2 [97.0, 99.6]	76.2 [73.6, 78.9]	93.8 [92.3, 95.3]
<i>LR_RF_r</i>	73.2 [67.7, 79.2]	99.0 [98.3, 99.6]	75.2 [70.6, 79.4]	94.4 [93.5, 95.4]	72.4 [67.5, 77.1]	98.9 [97.0, 100]	75.2 [72.6, 78.8]	93.5 [90.6, 96.2]
<i>LR_RF_h</i>	72.3 [67.8, 76.9]	98.8 [98.3, 99.3]	76.3 [72.5, 79.9]	94.6 [93.8, 95.4]	72.3 [69.4, 75.1]	98.4 [97.1, 99.6]	76.0 [73.3, 78.5]	93.5 [92.1, 95.1]
<i>LR_RF_r_h</i>	72.4 [68.1, 77.1]	98.9 [98.3, 99.3]	75.9 [72.3, 79.7]	94.6 [93.8, 95.4]	72.2 [69.3, 75.0]	98.4 [97.2, 99.6]	76.0 [73.4, 78.4]	93.6 [92.1, 95.0]

Table A6. Two data settings ($ICC = 0.20$, $\beta_1 = 0.40$, herd size = 100) with varying difference in region prevalences. Estimates of the two tests (%) are presented with median and 95% central credible interval (CCI).

Model	Region prevalences = (10%, 20%)				Region prevalences = (10%, 50%)			
	Se_1	Sp_1	Se_2	Sp_2	Se_1	Sp_1	Se_2	Sp_2
True value	70	99	75	95	70	99	75	95
<i>HW_r</i>	78.7 [63.9, 95.6]	99.6 [98.7, 100]	74.2 [68.7, 80.7]	94.0 [91.7, 96.4]	72.7 [68.4, 77.1]	98.9 [98.0, 99.9]	75.9 [72.4, 79.7]	94.2 [92.7, 95.7]
<i>LR_r</i>	78.0 [67.6, 90.4]	99.8 [99.2, 100]	73.3 [69.7, 77.3]	94.1 [92.4, 95.7]	72.6 [69.6, 75.6]	99.0 [98.3, 99.7]	75.9 [73.1, 78.4]	94.3 [93.2, 95.3]
<i>HW_h</i>	67.8 [60.9, 75.2]	99.8 [99.4, 100]	72.1 [67.6, 76.9]	95.7 [94.4, 97.0]	69.3 [65.3, 73.1]	99.2 [98.4, 99.9]	75.0 [71.5, 78.3]	95.5 [94.1, 96.9]
<i>LR_h</i>	72.9 [68.1, 78.3]	99.9 [99.5, 100]	72.9 [69.5, 75.9]	94.9 [94.0, 95.8]	70.4 [67.4, 73.1]	99.1 [98.5, 99.7]	75.6 [73.2, 77.8]	95.2 [94.1, 96.0]
<i>HW_RF</i>	78.8 [57.0, 100]	98.4 [96.5, 100]	81.5 [70.8, 98.0]	94.0 [91.3, 98.8]	78.7 [62.0, 98.0]	97.8 [93.6, 100]	78.8 [71.5, 93.7]	92.1 [87.0, 99.0]
<i>LR_RF</i>	87.2 [69.7, 100]	98.2 [96.5, 100]	82.4 [71.5, 96.9]	93.0 [91.4, 95.6]	82.8 [65.4, 100]	98.7 [95.8, 100]	76.7 [71.8, 85.3]	90.8 [86.7, 97.7]
<i>LR_r_h</i>	73.3 [68.2, 78.2]	99.9 [99.5, 100]	72.9 [69.5, 76.2]	94.9 [94.0, 95.8]	70.7 [68.0, 73.6]	99.0 [98.4, 99.6]	75.7 [73.4, 78.2]	95.1 [94.1, 96.0]
<i>LR_RF_r</i>	82.2 [70.7, 96.3]	99.7 [99.1, 100]	73.7 [69.7, 78.3]	93.5 [91.9, 95.3]	72.9 [69.9, 75.9]	98.9 [98.1, 99.6]	76.1 [73.5, 78.6]	94.1 [93.1, 95.2]
<i>LR_RF_h</i>	73.8 [69.0, 79.1]	99.8 [99.5, 100]	73.1 [69.8, 76.5]	94.8 [93.9, 95.6]	70.5 [68.0, 73.4]	99.0 [98.4, 99.5]	75.8 [73.4, 78.0]	95.1 [94.1, 96.0]
<i>LR_RF_r_h</i>	73.8 [69.0, 79.0]	99.8 [99.5, 100]	73.1 [69.8, 76.4]	94.8 [93.9, 95.7]	70.8 [68.1, 73.6]	99.0 [98.4, 99.6]	75.7 [73.4, 78.4]	95.0 [94.0, 95.9]

Table A7. Dataset (ICC = 0.20, $\beta_1 = 0.40$, herd size = 100, region prevalences = (10%, 30%)) with the binary risk factor sampled from success probability 0.50. Estimates of the Se and Sp of the two tests are presented with posterior median and 95% central credible interval (CCI).

Model	RF success probability = 0.50					
	Se_1	Sp_1	Se_2	Sp_2		
True value	70	99	75	95		
<i>HW_r</i>	69.8 [61.8, 78.0]	99.2 [98.1, 100]	74.2 [69.0, 80.1]	95.0 [93.0, 97.1]		
<i>LR_r</i>	70.0 [64.2, 76.4]	99.3 [98.4, 100]	73.9 [70.0, 78.5]	94.9 [93.5, 96.5]		
<i>HW_h</i>	67.4 [61.6, 73.1]	99.2 [98.5, 100]	73.2 [68.6, 77.9]	95.5 [94.1, 96.9]		
<i>LR_h</i>	70.7 [66.5, 75.1]	99.0 [98.4, 99.6]	74.9 [71.5, 78.2]	94.8 [93.8, 95.8]		
<i>HW_RF</i>	72.4 [55.0, 98.0]	97.1 [94.8, 100]	83.4 [71.3, 99.6]	94.4 [90.1, 99.8]		
<i>LR_RF</i>	76.9 [56.9, 99.6]	97.1 [94.7, 99.7]	83.6 [71.9, 99.6]	93.4 [89.8, 99.5]		
<i>LR_r_h</i>	70.6 [66.4, 74.7]	99.0 [98.3, 99.6]	75.0 [71.5, 78.5]	94.8 [93.8, 95.7]		
<i>LR_RF_r</i>	70.0 [64.3, 76.1]	99.3 [98.4, 100]	73.8 [69.7, 78.1]	95.0 [93.5, 96.3]		
<i>LR_RF_h</i>	70.7 [66.5, 75.0]	99.0 [98.3, 99.6]	74.9 [71.4, 78.1]	94.8 [93.8, 95.8]		
<i>LR_RF_r_h</i>	70.6 [66.3, 74.8]	99.0 [98.4, 99.6]	74.9 [71.3, 78.4]	94.8 [93.8, 95.8]		

Chapter 3

Test Characteristics of the Tuberculin Skin Test and Post-mortem Examination for Bovine Tuberculosis Diagnosis in Cattle in Northern Ireland Estimated by Bayesian Latent Class Analysis with Adjustments for Covariates

Abstract

The single intradermal comparative cervical tuberculin (SICCT) test and post-mortem examination are the main diagnostic tools for bovine tuberculosis (bTB) in cattle in the British Isles. Latent class modelling is often used to estimate the bTB test characteristics due to the absence of a gold standard. However, the reported sensitivity of especially the SICCT test has shown a lot of variation. We applied both the Hui-Walter latent class model under the Bayesian framework and the Bayesian model specified at the animal level, including various risk factors as predictors, to estimate the SICCT test and post-mortem test characteristics. Data were collected from all cattle slaughtered in abattoirs in Northern Ireland in 2015. Both models showed comparable posterior median estimation for the sensitivity of the SICCT test (88.61% and 90.56%, respectively) using standard interpretation and for post-mortem examination (53.65% and 53.79%, respectively). Both models showed almost identical posterior median estimates for the specificity (99.99% vs. 99.80% for SICCT test at standard interpretation and 99.66% vs. 99.86% for post-mortem examination). The animal-level model showed slightly narrower 95% posterior central credible intervals (CCIs). Notably, this study was carried out in slaughtered cattle which may not be representative for the general cattle population.

Keywords: Bayesian analysis; cattle; latent class model; *Mycobacterium bovis*; post-mortem examination; skin test; test characteristics

3.1. Introduction

This chapter is published with shared first authorship as: O'Hagan MJH, Ni H, Menzies FD, Pascual-Linaza AV, Georgaki A, Stegeman JA. Test characteristics of the tuberculin skin test and post-mortem examination for bovine tuberculosis diagnosis in cattle in Northern Ireland estimated by Bayesian latent class analysis with adjustments for covariates. *Epidemiology and Infection* 2019;147:1-8. doi: 10.1017/s0950268819000888.

Bovine tuberculosis (bTB) is a chronic, infectious disease caused by *Mycobacterium bovis* that affects cattle and many other mammals including humans worldwide. Infection with this bacterium often remains subclinical for a long period whilst cattle can be infectious. Diagnostics therefore must focus on effective detection of cattle at an early stage of infection (De la Rua-Domenech et al., 2006).

The single intradermal comparative cervical tuberculin test (SICCT test), based on detection of a cell-mediated immune response, is the main ante-mortem diagnostic tool for bTB in cattle in the British Isles (Pollock et al., 2003). Animals can be classified as reactors to the SICCT test on standard, severe or super-severe interpretation based on the cut-off point used of the measured response to the injected bovine and avian tuberculins into the skin of the neck (the test is carried out as defined within the EU Council Directive 64/432/EEC, Annex B). In 2015, standard interpretation was where the thickness at the site of injection of the bovine antigen was generally greater than the site of injection of the avian antigen by more than 4 mm. A severe interpretation was generally one in which the bovine bias was 3-4 mm. Super severe interpretation refers to animals considered positive to the SICCT test having a bovine bias <3 mm. Lowering the cut-off point will increase the sensitivity but in return decrease the specificity of the SICCT test and vice versa (Goodchild et al., 2015).

In Northern Ireland, all cattle over 6 weeks old are tested on at least an annual basis and positive cattle (reactors) are slaughtered followed by post-mortem examination and laboratory test (Abernethy et al., 2006). In order to confirm bTB by laboratory tests, most SICCT test reactors with visible lesions (43-60% of reactors animals in Northern Ireland) are subjected to histological examination. Furthermore, those samples that show no histological evidence of bTB are subjected to bacteriological culture as are samples from a proportion of SICCT test reactors without visible lesion (O'Hagan et al., 2015). The SICCT test is supplemented by routine abattoir surveillance of cattle slaughtered aiming to find visible bTB lesions. Due to factors such as the microscopic size of early lesions and the time required to develop a detectable immune response, neither the post-mortem examination nor the SICCT test can be expected to detect every bTB-infected animal. Furthermore, false-negative and false-positive reactions to the SICCT test can occur due to a variety of reasons relating to both animal- and test-related factors including desensitization, drugs, physiological status, tuberculin used, incorrect testing technique (De la Rua-Domenech et al., 2006) and concurrent infection (Godfray et al., 2013).

The sensitivity of the SICCT test reported in the literature shows a lot of variation and was reported in previous research based on summary values of field trials (De la Rua-Domenech et

al., 2006) to be between 52.0% and 100% with median values of 80.0% and 93.5% for standard and severe interpretations, respectively. Research based on meta-analyses in a systematic review of the scientific literature using Bayesian logistic regression models concluded the median sensitivity for the SICCT test (standard interpretation) to be 50% with wide 95% posterior central credible intervals (CCI) ([26%, 78%]) (median sensitivity of 63% (95% CCI [40%, 84%]) at severe interpretation) (Nuñez-García et al., 2018). The same study stated the median sensitivity of routine post-mortem examination at meat inspection to be 71% (95% CCI [37%, 92%]).

The specificity of the SICCT test has previously been estimated at over 99.9% (De la Rua-Domenech et al., 2006). Similar figures were quoted (specificity of 99.98% (95% confidence interval (CI) $\pm 0.004\%$)) for standard interpretation and for severe interpretation (99.91% (95% CI $\pm 0.013\%$)) (Goodchild et al., 2015). The previously mentioned study using meta-analyses (Nuñez-García et al., 2018) found a median specificity for the SICCT test of 100% (95% CCI [99%, 100%]) and a similar figure for the median specificity of routine post-mortem examination (100%; 95% CCI [99%, 100%]).

One of the main problems in relation to determining test characteristics and true disease status is the absence of a gold standard test for bTB. Sensitivity and specificity can be estimated in such cases by using latent class models applying two or more tests to two or more populations with distinct prevalences (Hui & Walter, 1980). However, this approach summarizes the test results to the (sub)population level, and it is difficult to include additional evidence available from data in the analysis. The Bayesian latent class model specified at the animal level offers the possibility of including animal-level information such as disease risk factors for the estimation of test characteristics (Koop et al., 2013).

Therefore, although latent class analyses for test characteristic estimation has been conducted previously for bTB diagnostics (Clegg et al., 2011; Karolemeas et al., 2012; Bermingham et al., 2015; Lahuerta-Marin et al., 2018), the current study is novel as it aims to address the variation in test characteristic estimates by adding a range of animal-level covariates to a Bayesian model in order to provide more precise estimates of the test characteristics for the SICCT test and bTB post-mortem surveillance.

3.2. Material and methods

An observational study encompassing all cattle slaughtered in abattoirs in Northern Ireland in 2015 was conducted. Cattle that were slaughtered but had a presenting herd outside Northern Ireland were excluded from the analyses.

3.2.1. Data collection and definition of variables

All data were extracted from the Animal and Public Health Information System (APHIS) of the Department of Agriculture, Environment and Rural Affairs (DAERA). Details on all individual cattle, cattle movements and bTB tests conducted since 1988 are stored in this database (Houston, 2001). Datasets were manipulated using Microsoft AccessTM (Microsoft Corporation, USA) and subsequently analyzed using R version 3.2.3 (The R Foundation for Statistical Computing) and JAGS version 4.1.0.

Data included in the analyses were based on information at animal level and test level. The data presented at animal level included individual measures on breed, sex, age at death, days from last SICCT test to slaughter and last SICCT test reason. The days from the last SICCT test was included in order to take account of animals with their last SICCT test being negative becoming infected before being slaughtered (i.e., to account for the fact that the two tests (SICCT and post-mortem) are non-contemporaneous). Breeds were categorized as breeds mainly kept for milk production (dairy) and non-dairy breeds. In relation to sex, three categories were constructed: female, non-castrated male (bull) and castrated male (bullock). Age at death was entered into the model as a continuous variable. The last SICCT test reason (i.e., the reason for the last SICCT test being conducted prior to slaughter) was divided into three categories; i.e., routine (in situations where no risk of bTB infection was suspected to be in the herd), at risk (in situations where the herd/animal was at increased risk of having bTB infection) and restricted (in situations where SICCT test reactors or animals with lesions at routine slaughter were found or the herd was at high risk of having bTB infection) (O'Hagan et al., 2015). The duration in days from the last SICCT test to slaughter was not included in the final animal-level model as reasoning from a biological perspective suggests that the fact that the two tests are non-contemporaneous should not matter in the case of bTB. It is estimated that the time period from the point of infection to reactivity to the SICCT test is approximately 2-3 weeks (OIE Terrestrial Manual, 2009). Thereafter bTB develops into a chronic infection with the formation of granulomata with variation in the immune responses over time (based on intermittent flare ups caused by the dynamics between the infection and body's immune system) followed by the animal having a lifelong infection compared to a very small window of the incubation period where detection would be missed. Furthermore, the minimum SICCT testing interval in Northern Ireland is 2 months with the median being much higher over the population being monitored. Animals also have to be at least 18 months before they are slaughtered (unless they are found to be SICTT reactor before that). In order to check the validity of this reasoning, we ran the animal-level model in two ways: (1) based only on cattle

that had ≤ 45 days from the last SICCT test to slaughter; (2) based only on cattle that had ≤ 23 days from the last SICCT test to slaughter.

The data presented at test level were based on the test-related information of the last SICCT test the animal was subjected to prior to slaughter and the tests after slaughter (including the post-mortem inspection result in the abattoir, the histology test and the bacteriological culture test) (OIE Terrestrial Manual, 2009). The interpretation of the SICCT test was based on recorded measurements of the net bovine rise (NBR), calculated as the increase (in millimetres) at the bovine tuberculin (Lelystad) injection site greater than any increase at the avian tuberculin injection site when measured after 72 hours (as per EU Council Directive 64/432/EEC, Annex B). A standard interpretation is read where the thickness at the site of injection of the bovine antigen is generally greater than the site of injection of the avian antigen by more than 4 mm. A severe interpretation is generally one in which the bovine bias is 3-4 mm (Lahuerta-Marin et al., 2018).

Cattle in the dataset were slaughtered in one of 10 abattoirs in 2015. However, as practically all SICCT test reactors were slaughtered in one slaughter house (abattoir E), posterior estimates of test characteristics were obtained on both the entire dataset and data from animals slaughtered in abattoir E only. Background analysis of lesion distributions between abattoir E and all other abattoirs were conducted in order to assess bias in relation to post-mortem examination techniques between abattoir E and the other abattoirs.

3.2.2. Data analysis

Hui-Walter model

The Bayesian Hui-Walter latent class model (Hui & Walter, 1980) was constructed to estimate the test characteristics of the SICCT test and post-mortem inspection for bTB. The 10 Divisional Veterinary Office (DVO) areas were treated as 10 subpopulations in Northern Ireland (see Fig. 3.1). Animals were allocated to a DVO area based on the location of the last herd they resided in before slaughter. We assumed that the 10 subpopulations submitted to slaughter had distinct proportions of bTB-infected cattle and sensitivity and specificity of the two tests were constant across populations. Based on a previous similar study (Lahuerta-Marin et al., 2018), the two tests were assumed to be independent, conditional on the true disease status of bTB.

To check whether the risk factors had an impact on the posterior estimation of sensitivity and specificity for both tests, the Bayesian Hui-Walter model was further applied to stratified samples. Each time the entire dataset (i.e., all cattle slaughtered) was stratified into two or three

samples by one of the five risk factors. The continuous covariates, age at death and days from last SICCT test to slaughter were categorized for the purpose of data stratification. Age at death was divided into two categories, i.e., ≤ 2 and > 2 years. This cut-off point was used as the majority of cattle bred for meat production are slaughtered by 2 years of age. The duration in days from last SICCT test to slaughter was divided into two categories, i.e., ≤ 45 and > 45 days. This cut-off point was chosen in line with previous research (Bermingham et al., 2015).

Non-informative beta prior distributions were specified for the test characteristics (i.e., sensitivity, specificity) and the true proportion of the diseased in each subpopulation. The analyses were repeated using informative priors based on the finding of previous research (Nuñez-Garcia et al., 2018) to see whether this significantly changed the results.

The model represented the risk of being bTB positive. Sensitivity and specificity estimates were considered independent of all covariates.

Animal-level model

As can be seen in the Bayesian Hui-Walter approach, stratifying data by a certain risk factor made it possible for us to assess the effect of the risk factor on the estimation of test performance. However, this approach could only investigate one risk factor at a time. In addition, the continuous covariates such as age at death had to be coded into categorical variables prior to data stratification which might cause loss of information.

To estimate the test sensitivity and specificity for SICCT test and post-mortem examination while taking the possible risk factors into account, a Bayesian logistic regression model was constructed at the animal level (Koop et al., 2013). The (latent) true bTB infection status for each animal was linked to the joint test results of the SICCT test and post-mortem inspection of each animal, expressed in the form of test sensitivity and specificity. The probability of an animal being bTB-infected was the dependent variable in the logistic regression model whereas the animal-level covariates were the predictors. The advantage of this modelling method is that the effect of multiple risk factors can be assessed simultaneously, and continuous covariates can be incorporated without categorization. In our analysis, the animal-level measures on breed, sex, age at death and last SICCT test reason were included as risk factors in the logistic regression model for animals that had ≤ 45 days (or ≤ 23 days) from last SICCT test to Abattoir E (see Appendix B for model code).

Non-informative normal prior distributions were specified for the regression coefficients of the risk factors. Only individual records that consisted of no empty cells from any of the variables mentioned above were used for the analysis. Backward model reduction was

performed by comparing the deviance information criterion (DIC) values among the competing models.

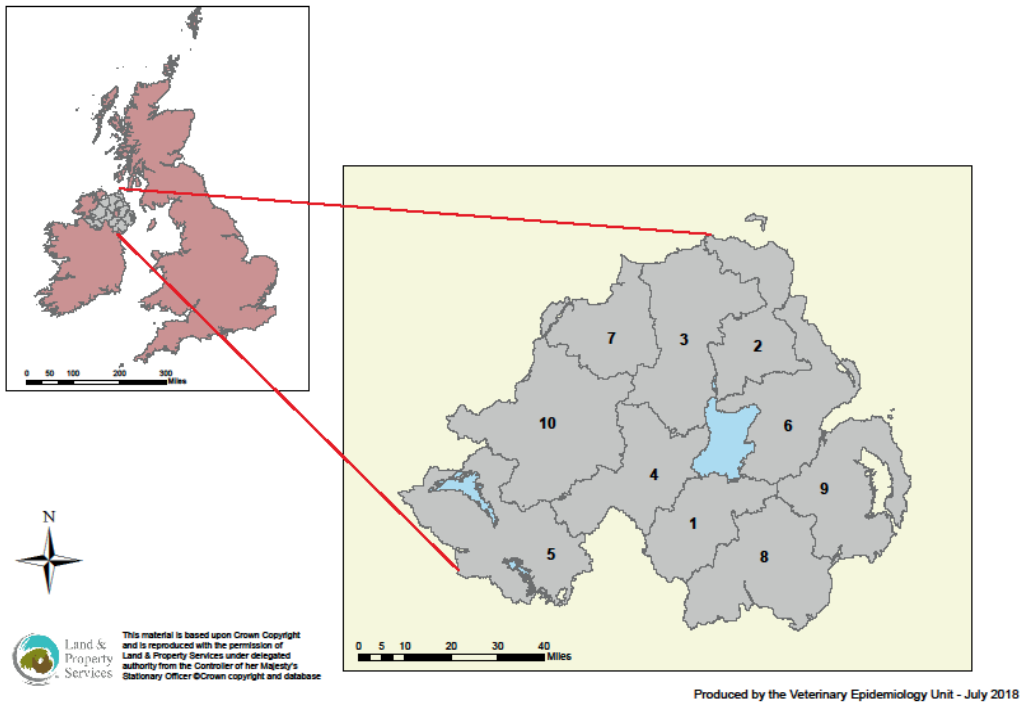


Fig. 3.1. Spatial distribution of the ten Divisional Veterinary Office (DVO) areas in Northern Ireland.

For all analyses, four Markov Chain Monte Carlo (MCMC) chains were sampled. Within each chain, the first 5,000 samples were discarded as the burn-in phase, and the subsequent 10,000 samples were used for posterior parameter estimation. Convergence was visually inspected by using trace plots.

3.3. Results

3.3.1. Descriptive results

In total, 413,383 cattle were slaughtered in abattoirs in Northern Ireland in 2015. A total of 29,839 cattle (7.2%) were dismissed from the analyses due to the fact that their presenting herd was not in Northern Ireland and a further 755 animals (0.2%) had missing values resulting in 382,789 cattle being included in the study.

3.3.2. Bayesian latent class analysis

Hui-Walter model

The posterior medians and 95% CCI obtained for the test sensitivity and specificity of the SICCT test (standard/severe interpretation) and post-mortem examination using non-informative priors are listed in Table 3.1 along with the estimated proportion of disease in the subpopulations (i.e., ten DVO areas). When the standard interpretation was used for the SICCT test, the estimated sensitivity for the SICCT test was 88.61% (95% CCI [85.39%, 92.23%]) and the estimated sensitivity for the post-mortem examination was 53.65% (95% CCI [52.59%, 54.75%]). The estimated specificities for both tests were very high (99.99% for the SICCT test and 99.66% for the post-mortem examination, respectively). Further, as expected, when the cut-off point was changed from the standard to severe interpretation, the sensitivity for the SICCT became higher (93.27%; 95% CCI [90.15%, 96.55%]) while that for post-mortem examination fell slightly (50.87%; 95% CCI [49.88%, 51.92%]). However, the specificity remained very high for both tests. The difference between the estimates using non-informative and informative priors was minimal (see Appendix A; Table A1).

Posterior parameter estimates based on stratified samples are presented in Table 3.2. For the stratified sample that contained only animals that were sent to slaughter after 45 days from their last SICCT test, due to very few SICCT test reactors, the test sensitivity and specificity of the tests could not be estimated.

Table 3.1. Posterior estimates (median and 95% central credible interval (CCI)) for SICCT test (standard/severe interpretation of the variable *net bovine rise*), post-mortem examination characteristics and proportions of the diseased in the subpopulations (DVO areas) based on the entire dataset.

	Standard interpretation (based on <i>net bovine rise</i>)	Severe interpretation (based on <i>net bovine rise</i>)
Sensitivity SICCT test (%)	88.61 [85.39, 92.23]	93.27 [90.15, 96.55]
Specificity SICCT test (%)	99.99 [99.97, 100.00]	99.99 [99.96, 100.00]
Sensitivity post-mortem (%)	53.65 [52.59, 54.75]	50.87 [49.88, 51.92]
Specificity post-mortem (%)	99.66 [99.60, 99.71]	99.68 [99.62, 99.73]

DVO 1 (%)	1.55 [1.43, 1.68]	1.68 [1.56, 1.81]
DVO 2 (%)	1.51 [1.34, 1.69]	1.63 [1.45, 1.81]
DVO 3 (%)	2.47 [2.29, 2.66]	2.63 [2.45, 2.81]
DVO 4 (%)	1.76 [1.63, 1.89]	1.89 [1.76, 2.03]
DVO 5 (%)	7.30 [6.83, 7.78]	7.85 [7.38, 8.34]
DVO 6 (%)	1.12 [0.99, 1.27]	1.33 [1.18, 1.48]
DVO 7 (%)	2.47 [2.20, 2.76]	2.59 [2.33, 2.88]
DVO 8 (%)	2.98 [2.79, 3.16]	3.16 [2.97, 3.34]
DVO 9 (%)	3.93 [3.66, 4.18]	4.22 [3.97, 4.49]
DVO 10 (%)	3.56 [3.34, 3.77]	3.72 [3.51, 3.93]

Animal-level model

As the main interest is based on cattle that went from bTB-negative to bTB-positive status, the final model and the vast majority of SICCT test reactors (8,956 out of in total 8,963 (99.9%)) was sent to Abattoir E. Therefore, the animal-level model with risk factors was performed only on Abattoir E cattle that had ≤ 45 days from last SICCT test to slaughter. The standard interpretation of the NBR was used for all subsequent analyses. No significant difference was found in post-mortem techniques between SICCT test reactors (abattoir E) and non-reactors (all abattoirs) regarding the number, nature and size of the lesions (see Appendix A; Table A2). Results from cattle that had ≤ 23 days from the last SICCT test to slaughter showed similar posterior estimates (Appendix A; Table A3).

Table 3.3 presents the distribution of test results from the SICCT test and post-mortem examination from the samples stratified by the risk factors within Abattoir E. The risk factors age at death and days from last SICCT test to slaughter are shown as categorical variables to provide an overview of the test-positive and negative counts from both tests (Table 3.3). In the model where risk factors were incorporated, age at death remained continuous and was not coded into a categorical variable.

Table 3.4 therefore presents the final posterior parameter estimates from the best-fitting (i.e., lowest DIC) animal-level model and the effect of the risk factors on the odd ratios calculated from the regression coefficients for the risk factors. Posterior estimates from the Hui-Walter model that aggregated both test results to a cross tabulation at the DVO level for Abattoir E are also listed.

Table 3.2. Posterior estimates (median and 95% central credible interval (CCI)) for the sensitivity and specificity for SICCT test (standard interpretation of the variable *net bovine rise*) and post-mortem examination derived from the stratified population based on risk factors.

Stratified population	SICCT test		Post-mortem	
	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
Age at death				
≤2 years	91.95 [85.92, 98.57]	99.97 [99.88, 100.00]	60.42 [58.40, 63.07]	99.71 [99.60, 99.82]
>2 years	86.75 [82.80, 90.74]	99.99 [99.97, 100.00]	50.29 [48.98, 51.67]	99.63 [99.56, 99.69]
Days from last SICCT test to slaughter				
≤45 days	90.30 [87.84, 92.84]	99.98 [99.89, 100.00]	53.68 [52.64, 54.82]	99.69 [99.53, 99.84]
>45 days	--	--	--	--
Breed				
Dairy	89.85 [86.10, 93.64]	99.99 [99.94, 100.00]	47.32 [45.79, 48.96]	99.70 [99.61, 99.79]
Non-dairy	89.85 [84.80, 94.98]	99.99 [99.94, 100.00]	59.25 [57.67, 61.11]	99.61 [99.55, 99.68]
Sex				
Bull	93.33 [80.03, 99.66]	99.95 [99.76, 100.00]	30.94 [25.99, 36.15]	99.66 [99.47, 99.86]
Bullock	72.23 [62.39, 83.65]	99.98 [99.89, 100.00]	63.75 [60.31, 70.89]	99.82 [99.72, 99.93]
Female	90.77 [87.83, 93.70]	99.99 [99.96, 100.00]	52.92 [51.71, 54.16]	99.57 [99.50, 99.65]
Last SICCT test reason				
Routine	93.85 [84.75, 99.55]	99.98 [99.91, 100.00]	56.56 [53.11, 61.19]	99.67 [99.62, 99.74]
Restricted	84.51 [80.43, 88.65]	99.99 [99.96, 100.00]	49.50 [48.05, 50.98]	99.67 [99.58, 99.76]
Risk	93.27 [88.99, 97.41]	99.96 [99.87, 100.00]	59.73 [57.71, 62.29]	99.60 [99.51, 99.69]

Table 3.3. Descriptive statistics of the relationship between risk factors and SICCT test and post-mortem examination results from Abattoir E.

Covariate	SICCT test positive			Post-mortem positive		
	Yes	No	% of positives	Yes	No	% of positives
Age at death						
≤ 2 years	3174	12446	20.3	2048	13572	13.1
> 2 years	5777	36438	13.7	3400	38815	8.1
Days from last SICCT test to slaughter						
≤ 45 days	8944	14573	38.0	5312	18205	22.6
> 45 days	7	34311	0.02	136	34182	0.40
Breed						
Dairy	4063	17736	18.6	2205	19594	10.1
Non-dairy	4888	31148	13.6	3243	32793	9.0
Sex						
Bull	331	903	26.8	109	1125	8.8
Bullock	1506	18934	7.4	1029	19411	5.0
Female	7114	29047	19.7	4310	31851	11.9
Last SICCT test reason						
Routine	1219	14300	7.9	747	14772	4.8
Restricted	4601	20323	18.5	2669	22255	10.7
Risk	3131	14261	18.0	2032	15360	11.7

Results showed that increasing age at death was very slightly related to the decrease of the odds of bTB infection. Females and bullocks had a smaller odds of bTB infection compared to bulls. Furthermore, the odds of being disclosed with bTB was higher for animals whose last SICCT test prior to slaughter was a ‘routine’ test or a ‘risk’ test compared to animals being subjected to a ‘restricted’ test.

Table 3.4. Posterior estimates (median and 95% central credible interval (CCI)) from the Hui-Walter model and the best fitting animal-level model with risk factors using cattle that had ≤ 45 days from the last SICCT test to Abattoir E.

	Hui-Walter model	
	SICCT test	Post-mortem
Sensitivity (%)	92.12 [90.90, 93.33]	53.60 [52.54, 54.65]
Specificity (%)	99.95 [99.71, 100.00]	99.18 [98.74, 99.57]
	Animal-level model with risk factors	
	SICCT test	Post-mortem
Sensitivity (%)	90.56 [89.80, 91.44]	53.79 [53.00, 54.69]
Specificity (%)	99.80 [99.51, 100.00]	99.86 [99.55, 100.00]
	Effect of risk factors (odds ratio)	
Age at death (per day increase)	0.99996 [0.99994, 0.99998]	
Sex		
Bull	Reference category	
Bullock	0.206 [0.186, 0.230]	
Female	0.596 [0.543, 0.667]	
Last SICCT test reason		
Restricted	Reference category	
Routine	1.430 [1.342, 1.525]	
Risk	1.804 [1.715, 1.882]	

3.4. Discussion

Estimation of test characteristics for diagnostic tests in the absence of a gold standard is notoriously difficult and this has been reflected by the increased use of Bayesian latent class analyses (Van Smeden et al., 2014). In the case of bTB diagnostics, in the absence of an accurate reference standard, these analyses have been used previously in order to calculate diagnostic test characteristics (Clegg et al., 2011; Karolemeas et al., 2012; Bermingham et al., 2015; Lahuerta-Marin et al., 2018).

Even with the use of Bayesian latent class modelling, a lot of variation especially in relation to sensitivity estimates of ante- and post-mortem tests for bTB has been reported (Nuñez-Garcia

et al., 2018). The current study aimed to apply a method to obtain more accurate estimates especially in relation to the test sensitivity by adding risk factors measured at the animal level.

By choosing two populations in order to conduct the latent class analyses, one of the issues in relation to bTB is that only cattle positive for the SICCT test will have an ‘immediate’ post-mortem result available. In order to have the availability of post-mortem results in the entire study population, it was decided to choose the study population as ‘all cattle slaughtered in 2015’. This approach means that all cattle that were recorded as reactors to the SICCT test and that were slaughtered in 2015 were in the study population but not all cattle that were negative to the SICCT test that were slaughtered in 2015. One of the consequences of this is that the prevalence estimated in the subpopulations (DVO areas; Table 3.1) do not reflect the prevalence in the actual DVO areas as a whole; it merely represents the bTB prevalence of all the cattle slaughtered presented by herds within these DVO areas. However, these prevalences indicate that the 10 DVO areas have distinct bTB proportions which is one of the prerequisites for Hui-Walter latent class modelling (Hui & Walter, 1980). For the Bayesian logistic regression model, the distinction between populations is not required as the model is constructed on the individual animal level.

The specificity estimated for both the SICCT test and post-mortem examination were very high (>99.4%) in all analyses (i.e., all cattle slaughtered, abattoir E only, standard and severe interpretation of the SICCT test, with and without addition of animal level risk factors) with a narrow 95% posterior CCI. These estimates are similar to previous estimates (Nuñez-Garcia et al., 2018) and show that both bTB diagnostic tests are very unlikely to report a false-positive animal.

The sensitivity estimates for the SICCT test (varying from 88.61% (standard interpretation) to 93.27% (severe interpretation)) were on the high end of figures reported by previous studies. The chosen population (i.e., all animals slaughtered in 2015) is potentially creating a bias for SICCT test reactors compared to SICCT test-negative animals as SICCT test reactors are always slaughtered whereas SICCT test-negative animals are not. Furthermore, previous research conducted in Northern Ireland (Lahuerta-Marin et al., 2018) (reported sensitivity at standard interpretation 40.5-57.7%) was based on chronic bTB breakdown herds only suggesting that herds that are tested on short intervals for prolonged periods of time may lower the sensitivity of the SICCT test in such circumstances. This is confirmed to a certain degree with this study as restricted tests had a significantly lower sensitivity than risk and routine tests. Moreover, estimates for chronic bTB breakdown herds used γ interferon test results for their analyses creating a bias towards herds that have already been censored through the removal of

SICCT test reactors during at least one previous recent herd SICCT test. The sensitivity estimates reported previously in the Republic of Ireland (Clegg et al., 2011) were lower as well, whereas our estimates were more in line with previously reported figures from England (Karolemeas et al., 2012). This is potentially due to the difference in bTB prevalence in the cattle population (Brenner & Gefeller, 1997) and the stage and details of the bTB eradication programmes showing more similarities between England and Northern Ireland than between Ireland and Northern Ireland (Abernethy et al., 2013). However, it has to be noted that the current study was carried out in slaughtered cattle in Northern Ireland which do not accurately represent the general cattle population.

The estimated sensitivity of post-mortem examination was similar to previous reported figures (Nuñez-Garcia et al., 2018). It was noteworthy that when severe interpretation was applied, lower sensitivity for post-mortem examination was obtained (Table 3.1). Severe interpretation is usually applied in herds when there is already infection in the herd and therefore it has been tested recently. It follows that lesions would not have had the time to develop to the ‘visible’ stage in terms of post-mortem inspection. Conditional dependence between the SICCT test and post-mortem inspection was therefore suspected. However, the Bayesian latent class models with covariance between the two tests added were not identifiable without informative priors for the test covariance. Only when informative priors were added to the covariance parameters of the tests, convergence was reached. Furthermore, posterior estimates were sensitive to the changes of informative priors for the test covariance (results not shown). This is not in agreement with a previous similar study in Northern Ireland where authors found a minimal difference in the parameter estimates between the model where conditional dependence was incorporated and the model where conditional independence was assumed among the tests (Lahuerta-Marin et al., 2018). More studies are needed for the investigation of the covariance between the SICCT test and post-mortem examination. Analyses based on data for abattoir E only were conducted in order to account for potential differences in the post-mortem examination conducted between abattoir E and the other abattoirs based on the fact that abattoir E was the destination for practically all (99.9%) SICCT test reactors. Distributions of post-mortem lesions by number, nature and size showed no differences between SICCT test reactors and those detected by post-mortem inspection only. This was backed up by the current study finding no significant differences in estimated test characteristics between abattoir E and the rest of the abattoirs (results not shown).

The estimates for the sensitivity and specificity of the tests varied among the stratified samples (Table 3.2), indicating that the parameter estimation was affected by the risk factors.

Comparing the test characteristic results between the Hui-Walter model and the animal-level model including the risk factors age at death, days from last test to slaughter, sex and last SICCT test reason showed that adding the risk factors created smaller 95% posterior CCI (Table 3.1 vs. Table 3.4). This suggests that by adding animal-level risk factors, we included more information to estimate the sensitivity and specificity for the SICCT test and post-mortem examination.

Furthermore, the relationship between each risk factor and the true disease status was assessed and quantified. The model evaluated possible risk factors for the bTB infection status of the animal as detected by these two diagnostic tests within the population of cattle slaughtered in 2015. The best-fitting model indicated that the risk of having a positive bTB status was significantly influenced by the animal-level characteristics age at death, days from last SICCT test to slaughter, sex and last SICCT test reason. Age at death was negatively correlated to the odds of bTB infection, namely animals with an older age were indicated to have slightly lower odds (0.99996). Increasing age is a risk factor for bTB breakdown (Skuce et al., 2011), but similarly it is protective in relation to the development of visible lesions (O'Hagan et al., 2015; Byrne et al., 2017). Furthermore, compared to bulls, female and castrated male animals (bullocks) tended to have smaller odds of bTB infection detection (0.60 and 0.21, respectively). Sex was not shown to be a risk factor for bTB infection, once adjusted for age, in studies previously conducted in the Republic of Ireland (Clegg et al., 2008), but SICCT test-positive bulls were shown to be less likely to develop visible lesions (O'Hagan et al., 2015). Relative differences in bTB disclosure in relation to sex may be masked by differences in longevity of beef and dairy cattle and different 'between and within' herd movements and contacts experienced (Skuce et al., 2011).

The odds of bTB infection was estimated to be 1.43 times higher for animals that were subjected to a 'routine' test prior to slaughter than for animals that were subjected to a 'restricted' test (reference category), whereas the odds of bTB infection for animals subjected to a 'risk' test prior to slaughter was estimated to be 1.80 times higher than the animals subjected to a 'restricted' test. It is worth noting that 58.8% slaughtered cattle that had ≤ 45 days from last test to slaughter at Abattoir E were from the 'restricted' farms, and 61.0% when all abattoirs were included. Furthermore, no significant relationship was found between the odds of bTB infection and the animal-level characteristic breed.

It should be noted that as our study was carried out in slaughtered cattle in Northern Ireland, the estimated effect of the risk factors on the odds of bTB infection may not be representative

for the general cattle population (1.75 million cattle). Further research may adapt this model to the general population.

3.5. Acknowledgements

Many thanks to Mark Woodside (Department of Agriculture, Environment and Rural Affairs (DAERA)) for his assistance in relation to the data extraction, and to Gerrit Koop (Department of Farm Animal Health, Utrecht University) for his valuable feedback. Also acknowledgements to Roly Harwood, Paddy McGuckian, Raymond Kirke, John Buchanan and David Kyle (DAERA) for their constructive comments. The authors would like to thank two anonymous reviewers. Their comments greatly improved the manuscript.

Appendix A: Supplementary tables**Table A1.** Posterior estimates (median and 95% central credible interval (CCI)) for the test sensitivity and specificity for SICCT and post-mortem with non-informative priors or informative priors.

	SICCT test		Post-mortem	
	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
Non-informative	88.61	99.99	53.65	99.66
prior for sensitivity for SICCT	[85.39, 92.23]	[99.97, 100.00]	[52.59, 54.75]	[99.60, 99.71]
Informative	87.81	100.00	53.27	99.65
Prior for sensitivity for SICCT	[84.76, 90.86]	[99.97, 100.00]	[52.56, 54.74]	[99.62, 99.72]

Table A2. Description of number (%) of animals with lesions in lymph nodes (LN) or close proximity of a particular location, median number of sites per animal having lesions (1-9 with 9 referring to ≥ 9), nature (mode; 1 = calcified; 2 = purulent; 3 = caseous) and size (median; 1-10 mm with 10 referring to ≥ 10 mm) of the lesions for SICCT test reactors and non-reactors with positive post-mortem results. No lesions were found in the following locations: submaxillary LN, all pluck, diaphragm, neck, sternum, all offal, stomach, inguinal LN, skin, hindleg, tail.

Location	SICCT positive				SICCT negative			
	Number (%) of animals with lesions	Number of sites (median)	Nature (mode)	Size (median)	Number (%) of animals with lesions	Number of sites (median)	Nature (mode)	Size (median)
All sites	5214 (100%)				288 (100%)			
Bronchio mediastinal LN	3653 (70.06%)	9	1	10	205 (71.18%)	9		9
Head	1697 (32.55%)	9	1	10	86 (29.86%)	9		10
Mesenteric LN	370 (7.10%)	2	1	10	5 (1.74%)	2	3	3
Lungs	214 (4.10%)	9	1	10	34 (11.81%)	9		2
Prescapular LN	81 (1.55%)	9	1	10	2 (0.69%)	5.50	2	10
Liver	45 (0.86%)	4	1	10	14 (4.86%)	9		10
Pleura	43 (0.82%)	9	1	10	5 (1.74%)	9		10
Precrural LN	21 (0.40%)	2	1	10	0 (0.00%)	-	-	-
Peritoneum	9 (0.17%)	9	1	10	3 (1.04%)	9	1	10
Kidney	8 (0.15%)	5.50	1	6.50	0 (0.00%)	-	-	-
Ham	7 (0.13%)	2	3	10	0 (0.00%)	-	-	-
Supramammary LN	4 (0.08%)	5.50	3	7.5	0 (0.00%)	-	-	-
Popliteal LN	3 (0.08%)	1	3	10	1 (0.35%)	3	3	3
Forequarter	3 (0.06%)	6	3	10	0 (0.00%)	-	-	-
All gut	1 (0.02%)	9	1	10	1 (0.35%)	9	1	10
Spleen	1 (0.02%)	1	1	4	0 (0.00%)	-	-	-
Foreleg	1 (0.02%)	1	1	10	0 (0.00%)	-	-	-
Pelvis	1 (0.02%)	9	1	10	0 (0.00%)	-	-	-
Retropharyngeal LN	0 (0.00%)	-	-	-	5 (1.74%)	10		10

Table A3. Posterior estimates (median and 95% central credible interval (CCI)) from the animal-level model with risk factors using cattle that had ≤ 23 days from last SICCT test to Abattoir E.

	Animal level model with risk factors	
	SICCT test	Post-mortem
Sensitivity (%)	93.26 [91.83, 94.71]	54.23 [53.33, 55.32]
Specificity (%)	99.47 [97.97, 100.00]	98.47 [97.42, 99.37]
Effect of risk factors (odds ratio)		
Age at death (per day increase)	0.99991 [0.99989, 0.99993]	
Sex		
Bull	Reference category	
Bullock	0.210 [0.185, 0.248]	
Female	0.588 [0.522, 0.690]	
Last SICCT test reason		
Restricted	Reference category	
Routine	1.794 [1.671, 1.934]	
Risk	1.979 [1.877, 2.093]	

Appendix B: model code**### Population level model ###**

```

model{
for (j in 1:nr.pops) {

p[j] ~ dbeta(1,1)
pop[j, 1:4] ~ dmulti(par[j,1:4], n[j])
par[j, 1] <- p[j]*Se1*Se2 + (1-p[j])*(1-Sp1)*(1-Sp2)
par[j, 2] <- p[j]*Se1*(1-Se2) + (1-p[j])*(1-Sp1)*Sp2
par[j, 3] <- p[j]*(1-Se1)*Se2 + (1-p[j])*Sp1*(1-Sp2)
par[j, 4] <- p[j]*(1-Se1)*(1-Se2) + (1-p[j])*Sp1*Sp2
}

```

priors

```

Se1 ~ dbeta(1, 1) T(1-Sp1, )
Sp1 ~ dbeta(1, 1)
Se2 ~ dbeta(1, 1) T(1-Sp2, )
Sp2 ~ dbeta(1, 1)
}

```

Best fitting model with animal level covariates

```

data {
for (i in 1:cum.dvo10) {
ones[i] <- 1
}
}

model{
## modelling the data on animal level
for (i in 1:cum.dvo10) {
pop[i, 1:4] ~ dmulti(par[i, 1:4], 1)

```

```

par[i, 1] <- pi[i]*Se1*Se2 + (1-pi[i])*(1-Sp1)*(1-Sp2)
par[i, 2] <- pi[i]*Se1*(1-Se2) + (1-pi[i])*(1-Sp1)*Sp2
par[i, 3] <- pi[i]*(1-Se1)*Se2 + (1-pi[i])*Sp1*(1-Sp2)
par[i, 4] <- pi[i]*(1-Se1)*(1-Se2) + (1-pi[i])*Sp1*Sp2

## Define/compute the contribution to the likelihood from the ith observation
L[i]<- equals(tests[i, 1], 1)*equals(tests[i, 2], 1)*par[i, 1]
+ equals(tests[i, 1], 1)*equals(tests[i, 2], 0)*par[i, 2]
+ equals(tests[i, 1], 0)*equals(tests[i, 2], 1)*par[i, 3]
+ equals(tests[i, 1], 0)*equals(tests[i, 2], 0)*par[i, 4]

## incorporate animal level covariates
logit(pi[i]) <- beta_0 + beta_1 * death.age[i] + beta_2 * Female[i] + beta_3 * Male[i] + beta_4 * Risk[i]
+ beta_5 * Routine[i]

ones[i] ~ dbern(L[i])
}

## prior
Se1 ~ dbeta(1, 1)T(1-Sp1, )
Sp1 ~ dbeta(1, 1)
Se2 ~ dbeta(1, 1)T(1-Sp2, )
Sp2 ~ dbeta(1, 1)

beta_0 ~ dnorm(0, 0.001)
beta_1 ~ dnorm(0, 0.001)
beta_2 ~ dnorm(0, 0.001)
beta_3 ~ dnorm(0, 0.001)
beta_4 ~ dnorm(0, 0.001)
beta_5 ~ dnorm(0, 0.001)

## mean of the animal level predicted probabilities for bTB per DVO
mean.pi1<-mean(pi[1:cum.dvo1])
mean.pi2<-mean(pi[(cum.dvo1+1):cum.dvo2])

```

```
mean.pi3<-mean(pi[(cum.dvo2+1):cum.dvo3])
mean.pi4<-mean(pi[(cum.dvo3+1):cum.dvo4])
mean.pi5<-mean(pi[(cum.dvo4+1):cum.dvo5])
mean.pi6<-mean(pi[(cum.dvo5+1):cum.dvo6])
mean.pi7<-mean(pi[(cum.dvo6+1):cum.dvo7])
mean.pi8<-mean(pi[(cum.dvo7+1):cum.dvo8])
mean.pi9<-mean(pi[(cum.dvo8+1):cum.dvo9])
mean.pi10<-mean(pi[(cum.dvo9+1):cum.dvo10])

}
```

Part II

Informative Priors for Random Effects

Chapter 4

Prediction models for clustered data with informative priors for the random effects: A simulation study

Abstract

Background: Random effects modelling is routinely used in clustered data, but for prediction models, random effects are commonly substituted with the mean zero after model development. In this study, we proposed a novel approach of including prior knowledge through the random effects distribution and investigated to what extent this could improve the predictive performance.

Methods: Data were simulated on the basis of a random effects logistic regression model. Five prediction models were specified: a frequentist model that set the random effects to zero for all new clusters, a Bayesian model with weakly informative priors for the random effects of new clusters, Bayesian models with expert opinion incorporated into low informative, medium informative and highly informative priors for the random effects. Expert opinion at the cluster level was elicited in the form of a truncated area of the random effects distribution. The predictive performance of the five models was assessed. In addition, impact of suboptimal expert opinion that deviated from the true quantity as well as including expert opinion by means of a categorical variable in the frequentist approach were explored. The five models were further investigated in various sensitivity analyses.

Results: The Bayesian prediction model using weakly informative priors for the random effects showed similar results to the frequentist model. Bayesian prediction models using expert opinion as informative priors showed smaller Brier scores, better overall discrimination and calibration, as well as better within cluster calibration. Results also indicated that incorporation of more precise expert opinion led to better predictions. Predictive performance from the frequentist models with expert opinion incorporated as categorical variable showed similar patterns as the Bayesian models with informative priors. When suboptimal expert opinion was used as prior information, results indicated that prediction still improved in certain settings.

Conclusions: The prediction models that incorporated cluster level information showed better performance than the models that did not. The Bayesian prediction models we proposed, with cluster specific expert opinion incorporated as priors for the random effects showed better predictive ability in new data, compared to the frequentist method that replaced random effects with zero after model development.

Keywords: random effects prediction model; clustered data; informative priors for the random effects; expert knowledge; truncated distribution

4.1. Background

In many medical areas, prediction models are used to support clinical practice (Steyerberg, 2009).

When study data collected for the development of a prediction model are clustered e.g., patients are registered with the same general practitioner or farm animals live in the same herd, there is often within cluster dependency. It is suggested that the clustering structure should be taken into account in the development of a prediction model, in order to produce unbiased model parameter estimates (Bouwmeester et al., 2013), whereas regression methods that assume independence between subjects are inappropriate. In such situations, random effects regression analysis can be a viable alternative, as it parameterizes the cluster level heterogeneity by means of random effects, and allows predictions to be made at the subject level (Hox, 2002).

Surprisingly, despite the routine use of random effects regression modelling in etiological or intervention research, this approach is hardly seen in prediction research (Bouwmeester et al., 2012). This is probably because generalization of the random effects model is not straightforward (Finkelman et al., 2016), as the latent random coefficient of a new cluster is considered unknown. In existing clinical applications, e.g., Van der Drift et al. (2012), the random effects were removed from the model after selection of predictors at the model development phase. This is equivalent to setting the random effects for all new clusters to zero. By doing so, the prediction model simply ignores the clustering structure in new data, which may lead to a loss of prediction accuracy (Bouwmeester et al., 2013).

Alternatively, one could maintain and estimate the random effects for new clusters by incorporating external cluster level information into the prediction model. In medical practice for instance, some hospitals are better at treating a certain disease than other hospitals due to hospital specific characteristics. An expert may be able to provide such additional information

about the hospitals. Methods for eliciting information from experts can be found in literature such as Spiegelhalter et al. (2004) and O'Hagan et al. (2006). Incorporation of expert knowledge into the data analysis can easily be done under the Bayesian framework. In this paper, we propose a new approach that includes cluster level expert knowledge as prior evidence for the random effects in a prediction model and investigate the benefit of this approach in the setting of new clustered data.

The paper is organized as follows: in the Methods section, we first review how one can develop and apply a prediction model either in a frequentist or in a Bayesian way. We then propose our approach of incorporating expert opinion into the prediction model. Description of the simulation studies is provided afterwards, followed by the Results section. The paper concludes with a discussion of results and implications for future research.

4.2. Methods

4.2.1. Estimation at model development phase

At the model development phase, data containing measures of the predictor(s) and outcome of interest are collected for the purpose of estimating the model parameters. In empirical applications, selection of relevant predictor(s) is often performed first. In this study, we assume that relevant predictors were selected already and directly focus on parameter estimation.

We consider a simple logistic regression model with one predictor measured on the subject level and random effects at the cluster level. Let x_{ij} be the predictor and y_{ij} be the observed binary outcome of subject i ($i = 1, \dots, n_j$) from cluster j ($j = 1, \dots, J$), and $p_{ij} = p(y_{ij} = 1)$ be the latent underlying risk for the observed binary outcome. A random effects logistic regression model can be expressed as:

$$\begin{aligned} \text{logit}(p_{ij}) &= \beta_0 + \beta_1 x_{ij} + u_j \\ u_j &\sim N(0, \sigma_u^2), \end{aligned} \quad (4.1)$$

where the logit (i.e., log-odds) of the latent underlying risk of the outcome $\text{logit}(p_{ij})$ is equivalent to the linear predictor $LP_{ij} = \beta_0 + \beta_1 x_{ij} + u_j$. The model can alternatively be written in the form of:

$$p_{ij} = \frac{1}{1 + \exp(-LP_{ij})}. \quad (4.2)$$

The linear predictor consists of the regression parameter β_1 for the predictor, the average (fixed) intercept β_0 and the cluster specific random effect u_j . The random effects are assumed to have a normal distribution with mean zero and variance σ_u^2 .

4.2.2. Frequentist estimation

Within the frequentist approach, parameters of a logistic regression prediction model are estimated via maximum likelihood (ML). In our study, functions from the R package ‘lme4’ were used (Bates et al., 2017).

4.2.3. Bayesian estimation

Within the Bayesian framework, parameters are expressed in the form of distributions rather than fixed values. Before observing the data, prior distributions that contain the plausible values for the model parameters need to be specified (Spiegelhalter et al., 2004). The prior distributions are subsequently updated with observed data, resulting in posterior distributions. The posterior can be derived either analytically or by sampling methods. In this study, Markov chain Monte Carlo (MCMC) sampling was used.

Prior distributions needed to be specified for parameters β_0 , β_1 and σ_u^2 in model (4.1). Priors for the regression parameters β_0 and β_1 were assumed to be normally distributed, and prior for the variance σ_u^2 had an inverse gamma distribution. When there is no *a priori* evidence available, one often uses weakly informative priors. A common choice for a normal distribution is to fix the mean at zero and take a large variance (here we used 1,000 for the variance). For an inverse gamma distribution, small values are often assigned to the hyperparameters (here we used 0.001 for both hyperparameters).

$$\begin{aligned}\beta_a &\sim N(0, 1000) \text{ (for } a = 0, 1), \\ \sigma_u^2 &\sim \text{Inv-gamma}(0.001, 0.001).\end{aligned}\tag{4.3}$$

Estimates for the parameters of interest are provided by the MCMC samples from the posterior distribution after convergence. In this study, we used OpenBUGS (via R package ‘BRugs’ (Ligges et al., 2017)) to carry out the Bayesian analyses.

4.2.4. Prediction in new clusters

Let x_{sc} be the predictor, y_{sc} be the observed binary outcome and p_{sc} be the latent underlying risk of the observed outcome for subject s ($s = 1, \dots, n_c$) from new cluster c ($c = 1, \dots, C$). The prediction model developed and estimated from model development data is applied to calculate the risk of outcome for each subject in new clusters. The predicted risk of outcome \hat{p}_{sc} is compared to the true latent underlying risk p_{sc} and the observed outcome y_{sc} for the evaluation of the predictive performance.

4.2.5. Frequentist prediction

In the frequentist approach, prediction for new clusters is usually based on a model where point estimates for the regression parameters are incorporated and the random effect term is substituted with mean 0 (i.e., removed). This leads to the predicted linear predictor

$$\widehat{LP}_{sc}^{ML} = \hat{\beta}_0^{ML} + \hat{\beta}_1^{ML} x_{sc}, \quad (4.4)$$

where $\hat{\beta}_0^{ML}$ and $\hat{\beta}_1^{ML}$ are the estimated regression coefficients using maximum likelihood estimation and x_{sc} contains values of the predictor from subjects in new clusters. Accordingly, the predicted risk for the binary outcome in new clusters can be written as

$$\hat{p}_{sc} = \frac{1}{1 + \exp(-\widehat{LP}_{sc}^{ML})}. \quad (4.5)$$

4.2.6. Bayesian prediction

In the Bayesian approach, by MCMC sampling, we obtain the posterior distribution for the parameters β_0 , β_1 and σ_u^2 . In this study, instead of using summarized point estimates, the full posterior for the parameters is exploited for prediction. To explain the Bayesian prediction, consider Table 4.1 in which a small part of the MCMC output is listed.

The sampled values for parameters from iteration k , denoted by $\tilde{\beta}_0^{(k)}$, $\tilde{\beta}_1^{(k)}$ and $\tilde{\sigma}_u^{2(k)}$, are presented in the left hand part of the table. Predictions made for subjects in a new cluster can be found in the right hand part of the table. As the model development clusters and the new clusters are assumed to be exchangeable, i.e., originated from the same metapopulation, random effects for all clusters are assumed to have the same normal distribution $N(0, \sigma_u^2)$. For new cluster c , in each iteration, the predicted random effect $\hat{u}_c^{(k)}$ can be drawn from distribution $N(0, \tilde{\sigma}_u^{2(k)})$. For each subject s ($s = 1, \dots, n_c$) from cluster c , the estimated linear predictor is hence

$$\widehat{LP}_{sc}^{(k)} = \tilde{\beta}_0^{(k)} + \tilde{\beta}_1^{(k)} x_{sc} + \hat{u}_c^{(k)}, \quad (4.6)$$

and the predicted risk can be obtained by

$$\hat{p}_{sc}^{(k)} = \frac{1}{1 + \exp(-\widehat{LP}_{sc}^{(k)})}. \quad (4.7)$$

Eventually, a predicted risk distribution based on all K iterations is available for each subject. In this paper, in order to compare the results between the Bayesian models and the frequentist model, we used the median of the predicted risk distribution as the summarized predicted risk \hat{p}_{sc} , resulting in a single estimate per subject.

Table 4.1. An example of using posterior samples from model development data analysis for prediction in a new cluster.

Posterior from model development data					Prediction for new cluster c				
Iteration					Subject 1		...	Subject n_c	
K	$\tilde{\beta}_0^{(k)}$	$\tilde{\beta}_1^{(k)}$	$\tilde{\sigma}_u^{2(k)}$	$\hat{u}_c^{(k)\dagger}$	x_{1c}	$\hat{p}_{1c}^{(k)\ddagger}$...	x_{n_cc}	$\hat{p}_{n_cc}^{(k)\ddagger}$
.
.
5001	-1.35	1.07	1.17	0.50	1.11	0.58	...	-0.46	0.21
5011	-1.24	1.08	0.88	-1.89	1.11	0.13	...	-0.46	0.03
5021	-1.36	1.18	1.28	-0.06	1.11	0.47	...	-0.46	0.12
5031	-1.31	1.05	0.98	-0.64	1.11	0.31	...	-0.46	0.08
5041	-0.94	0.98	1.37	0.26	1.11	0.60	...	-0.46	0.24
.
.
Median						0.52			0.15

\dagger random effect sampled from the normal distribution $N(0, \tilde{\sigma}_u^{2(k)})$.

\ddagger predicted risk calculated by $\hat{p}_{sc}^{(k)} = \frac{1}{1 + \exp(-\tilde{\beta}_0^{(k)} + \tilde{\beta}_1^{(k)} x_{sc} + \hat{u}_c^{(k)})}$.

By applying the Bayesian approach, we can maintain the random effect term in the prediction model, which takes uncertainty due to variance between clusters into account. To improve the predictive ability of the Bayesian random effects model, one may include cluster specific expert knowledge as prior for the random effects on new clusters. Demonstration of how this prior knowledge was incorporated into a Bayesian prediction model can be found in the following section.

4.2.7. Bayesian approach with informative priors

In the previous section, for each new cluster c in each iteration k , the predicted random effect $\hat{u}_c^{(k)}$ was sampled from the entire random effects distribution $N(0, \hat{\sigma}_u^2)^{(k)}$. This would add more uncertainty to the prediction compared to the frequentist model that substituted the random effects with the mean 0. However, if there is information available about the position of a new cluster relative to other clusters, we may sample a value for $\hat{u}_c^{(k)}$ from part of the distribution rather than the whole distribution.

Consider again the example of hospitals where some hospitals are known to be better at treating a particular disease than others. When we have no clue about the relative risk of death for a disease from a particular hospital regarding other hospitals, we sample a random effect for the hospital from the entire random effects distribution. However, if an expert is capable of using hospital level information to judge whether a hospital would provide below or above average chance of survival, we could sample the random effect from only the lower or upper half of the distribution (Robert, 1995). Suppose the expert says the hospital will provide a below average probability of survival, the random effect will accordingly be sampled from the lower half of the distribution (see the first plot in Fig. 4.1). Further, if the expert is more precise about the relative position of the hospital with regard to other hospitals, the random effect can also be drawn from a smaller area, such as one third or one fifth of the distribution (see the second and third plots in Fig. 4.1).

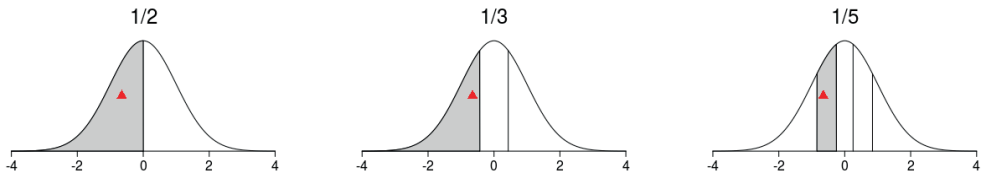


Fig. 4.1. The random effects distribution divided into multiple truncated areas of equal proportions in 3 different scales. A truncated area contains either half, one third, or one fifth of the distribution. Based on elicited expert knowledge for each cluster, a particular truncated area from each scale is chosen and used as prior distribution for the random effect of the cluster. We considered a prior distribution that contains $1/2$ of the distribution as low informative, $1/3$ as medium informative, and $1/5$ as highly informative.

It is expected that more precise prior knowledge leads to better predictions, under the assumption that an expert provides information that matches the true value. It is however likely that experts sometimes provide suboptimal judgments that deviate from the true quantity. Incorporation of discrepant expert opinion into the prediction model was therefore investigated as well. All investigations were done through simulations. Details of the simulation studies are presented in the next section.

4.3. The simulation studies

4.3.1. Data generation

One set of data containing 5,000 subjects ($n=5,000$) was generated for model development. The number of clusters was set to 50, and the number of subjects per cluster was 100. Each cluster consisted of equal numbers of subjects. For each subject, one continuous predictor was sampled from the normal distribution $N(0, 1)$ and allocated to the subjects in all further analyses. The true value for the regression parameter β_1 was set to 1.5. The random effects for clusters were sampled from a normal distribution with cluster variance 0.822. This value corresponds to an intraclass correlation coefficient (ICC) of 0.20, calculated from the formula $\sigma_u^2 / (\sigma_u^2 + \pi^2/3)$ where the error variance has a fixed value $\pi^2/3$ in a logistic regression model (Hox, 2002). The latent underlying risk of the outcome for each subject was calculated by computing the linear predictor. The observed outcome was subsequently sampled from a Bernoulli distribution using its underlying risk. By adjusting the value of the fixed intercept β_0 , we set the overall prevalence of the dataset approximately to 50%. Finally, one new dataset was simulated by the exact same setting and used to evaluate the different prediction models. The simulation study thus contained one dataset for model development and one dataset for prediction. The R code for data simulation is provided in Appendix B. It is worth noting that each time if we replicate a simulation study using the same settings with a different seed, we get different samples for the model development and prediction datasets, as there is randomness

involved in the sampling process. Comparisons of the relative performance between models were hence carried out only within the same sets of development and prediction data.

4.3.2. Analysis of simulated data

Five models were estimated using model development data and applied to predict in new data. In the frequentist model (denoted *FREQ*), ML estimates of the regression parameters $\hat{\beta}_0^{ML}$ and $\hat{\beta}_1^{ML}$ were incorporated, and the random effect term was replaced with 0. Within the Bayesian approach, a prediction model using weakly informative priors for the random effects was first specified and denoted *BAYES.WI*. Three more prediction models were subsequently constructed where cluster specific expert judgments were incorporated as prior information for the random effects. Optimal expert judgment was defined as choosing the truncated area from the random effects distribution which contained the true value of the random effect. Three scales were specified for the random effects distribution on the basis of varying degrees of precision of the expert opinion. In each scale, the distribution was divided into equal sized truncated areas (i.e., equal proportions). The model that used low informative priors which contained half of the distribution was denoted *BAYES.LI*. Similarly, the model that used medium informative priors which contained one third of the distribution was denoted *BAYES.MI*, and the model that used highly informative priors which contained one fifth of the distribution was denoted *BAYES.HI*. For each Bayesian prediction model, two posterior chains were sampled and thinned by the interval 10 (i.e., taking every 10th observation). Within each chain 100 samples were saved after convergence was reached, leading to 200 posterior samples in total for prediction in each subject from the new clusters.

4.3.3. Discrepant expert opinion

Impact of including expert opinion that deviated from the true random effect value as prior information was explored in models *BAYES.LI*, *BAYES.MI* and *BAYES.HI*. We defined discrepant expert judgment as selecting a truncated area that was next to the ‘correct’ area that contained the true value of the random effect. In the *BAYES.LI* condition, it was straightforward, since the random effects distribution was divided into two equal sized areas. When the true value of the random effect was located in one area, the other area would be chosen. However, in the *BAYES.MI* and *BAYES.HI* conditions where the random effects distribution was divided into more than two equal sized areas, choosing a truncated area that was next to the ‘correct’ area indicated that the discrepant expert opinion was still relatively

close to the true value (see Fig. 4.1). Suppose the true value was at the middle 1/5 truncated area of the distribution, the discrepant expert opinion would be selecting the second 1/5 or the fourth 1/5 truncated area from the left hand side. In other words, discrepant expert opinion with more precision (i.e., 1/3, 1/5) was less off from the true value in comparison to the discrepant expert opinion with least precision (1/2). The percentage of new clusters that incorporated discrepant expert opinion was set to 10%, 30% or 50%.

4.3.4. Expert opinion as categorical variable in the frequentist approach

In principle, one could also include prior knowledge in the frequentist model. We performed simulations where optimal expert opinion was incorporated as a fixed effect in the frequentist model in the default scenario. First, in the phase of model development, the true random effects were used to create the additional predictor representing the prior knowledge. For instance, when the true random effect for a specific cluster was among the upper half of the true random effects distribution, expert opinion for this cluster was then coded into 1 when the lower half was the reference category (coded 0). Likewise, expert opinion was coded using a categorical variable with 3 or 5 levels by placing the true random effects for the model development clusters in the correct tertiles and quintiles, with the second tertile and the third quintile as the reference category respectively. The frequentist models with inclusion of expert opinion coded into 2, 3, and 5 categories were denoted *FREQ.2*, *FREQ.3* and *FREQ.5* respectively.

The resulting prediction models with expert opinion were used to predict the outcomes in the new clusters. Again, the expert knowledge (i.e., the scores on the categorical variable) for each new cluster was obtained by placing the true random effect for the new cluster in the correct quantiles of the true random effects for the model development clusters.

4.3.5. Assessment and comparison of model performance

The predictive ability of the models was assessed by Brier scores, model discrimination and calibration. Brier score was computed as the mean of the squared difference between the observed binary outcomes and the predicted risks. Discriminative ability was assessed by the concordance index (C-index) which equals the area under the ROC curve. Model calibration was evaluated using the calibration slope. The ideal value for the calibration slope is 1, which represents perfect prediction. The further the calibration slope deviates from 1, the worse the model is calibrated (Steyerberg, 2009). Since in the simulation research, true latent underlying risks of outcome are available, the calibration was computed as the agreement between predicted risks and true risks. The calibration slope was the linear regression coefficient for the

predicted risk as the independent variable, and the true risk as the dependent variable. It is worth noting that in empirical data, true risks are not available. Model calibration can hence be assessed by using the observed binary outcome as the dependent variable and the estimated linear predictor as the independent variable in a logistic regression analysis (Steyerberg, 2009; Van Calster et al., 2016). Brier scores were computed at the subject level (overall) for all models. Model discrimination and calibration were measured both at subject level (overall) and cluster level (within cluster). The within cluster measures were summarized in mean and standard deviation over all clusters. Calibration of the models was further visualized in calibration plots where the predicted risks of outcome were plotted against the true risks.

4.3.6. Sensitivity analyses

In order to check the influence of prevalence, ICC, sample size and strength of the predictor on prediction, eight sensitivity analyses were carried out. Each sensitivity analysis consisted of one new simulation study that had default simulation settings except for the specific feature that was investigated. Impact of the between cluster variance was evaluated by comparing the default ICC value 0.20 to 0.05 and 0.50. Impact of the prevalence was evaluated by comparing the default prevalence value 50% to 10% and 25%. To examine the effect of smaller sample sizes on prediction, we reduced in one sensitivity analysis the number of clusters, resulting in $n = 2,000$ subjects in total ($J = S = 20$, $n_j = n_c = 100$), and in another sensitivity analysis the number of subjects per cluster, resulting in $n = 1,000$ subjects in total ($J = S = 50$, $n_j = n_c = 20$). Finally, by changing the value for the model parameter β_1 from 1.5 (default) to 0.5 and 3.0, we explored the influence of a weaker or a stronger subject level predictor.

4.4. Results

As can be seen in Table 4.2, the frequentist model and the Bayesian model without prior information showed, as expected, approximately the same Brier scores, similar discrimination and calibration at the overall as well as the cluster level. The Bayesian models with informative priors showed smaller Brier scores and larger overall C-indexes. The increasing C-indexes also revealed a positive relation between the precision of expert opinion and the overall discrimination. Further, difference in overall calibration slopes suggested that the Bayesian models with informative priors had better overall calibration. This can also be inspected in the calibration plots in Fig. 4.2, where the predicted risks were plotted against the true latent

underlying risks. Smaller difference between the predicted and true risks can be seen for the Bayesian models with informative priors, as the calibration plots from these models were more closely around the diagonal line which indicated equality between the predicted and the true risks. It is noteworthy that, for the overall measures, difference between the frequentist model and the Bayesian model with informative priors was larger than difference among the Bayesian models with informative priors. Particularly between the Bayesian models with medium and highly informative priors, there is much less difference in the overall measures. Further, the frequentist models with expert opinion incorporated showed fairly similar patterns in results as the Bayesian models with informative priors.

When it comes to cluster specific predictive performance, five models showed the same within cluster C-index means and variances, suggesting the same discriminative ability at the cluster level. This is because the random cluster effects only contribute to discrimination of subjects from different clusters. However, the Bayesian models with informative priors showed better within cluster calibration, as their within cluster calibration slopes were closer to 1 compared to the frequentist model. In addition, inclusion of more precise cluster level expert evidence led to smaller standard deviation for the within cluster calibration slopes.

Further, as shown in Table 4.3, when the percentage of new clusters that incorporated discrepant expert opinion was 10%, all Bayesian models with informative priors still outperformed the frequentist model concerning the Brier score, the overall discrimination and the overall and within cluster calibration. When the number of clusters that incorporated discrepant expert opinion increased to 30%, the model with low informative priors performed similarly to the frequentist model, whereas the models with medium informative and highly informative priors still performed better. When the percentage was increased to 50%, the Bayesian model with low informative priors showed worse predictive performance than the frequentist model. However, the Bayesian models with medium and highly informative priors remained better in predictive performance. This phenomenon can also be seen in the 9 calibration plots in Fig. 4.3.

Multiple datasets for model development and prediction were generated using different seeds and results (not reported in the paper but available from the first author) showed the same structure among the prediction models. Furthermore, the eight sensitivity analyses showed the same patterns for the five models (see tables in Appendix A), suggesting robustness of the Bayesian prediction models using informative priors. In addition, by varying the prevalence of the true binary outcomes, we noticed that it was more beneficial to add cluster specific priors to the random effects when the prevalence was closer to 50%. Results from different ICC values

implied that in data with higher between cluster variance, it was more useful to include cluster specific expert opinion. In data with small cluster size and in data with small amount of clusters, the Bayesian models with cluster specific expert opinion still showed better performance than the frequentist model. Further, a weak predictor had negative impact on the frequentist model, adding cluster level informative priors in the Bayesian models showed clear improvement.

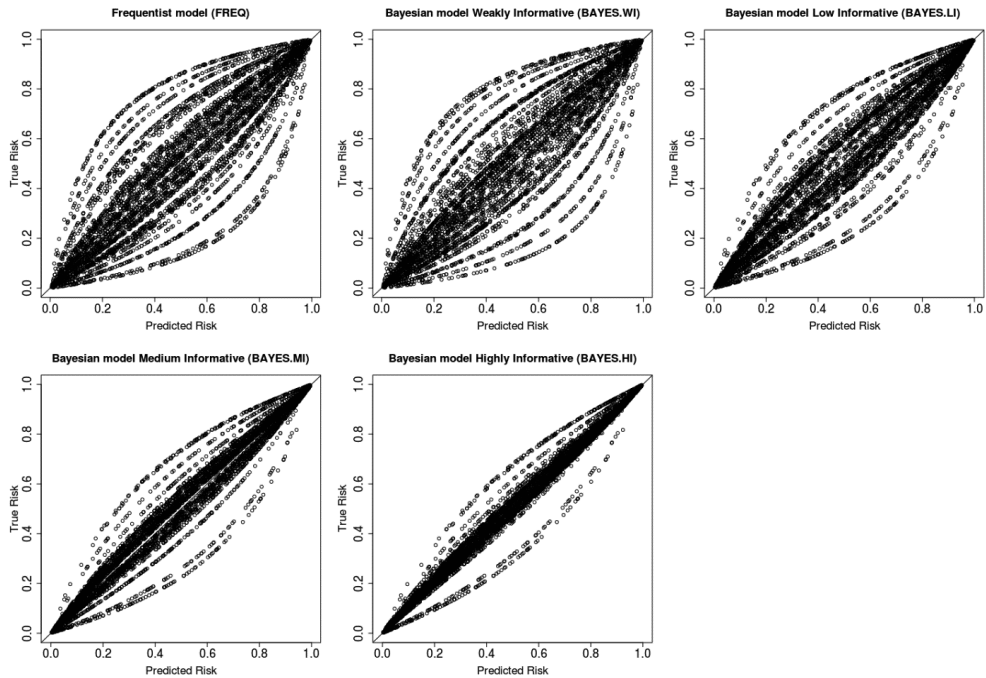


Fig. 4.2. Calibration plots for the five prediction models. Predicted risks are plotted against the true latent underlying risks for 5,000 subjects from 50 equal sized clusters. The diagonal indicates the line of identity (i.e., predicted risks are equivalent to the true risks). Each dot represents a subject, and each line formed by the dots represents a cluster.

4.5. Discussion

The simulation study showed that the Bayesian model with weakly informative priors performed similarly to the frequentist model where the random effect term was replaced with 0. It is hence possible to take the clustering structure from the new cluster(s) into account by means of keeping the random effects in the prediction model, without the loss of predictive ability. An additional benefit of using the Bayesian prediction models may be that they provide

for each subject a distribution for the predicted risk. In many real world situations, such information may be preferred in comparison to point estimates.

Improvement was detected in results from models that had optimal cluster level expert opinion incorporated as informative priors for the random effects in the Bayesian approach as well as in the frequentist approach where expert opinion was incorporated as categorical variable. More specifically, these models showed better discrimination and calibration at the subject level, and better calibration within individual clusters. Comparison between these models also revealed that incorporation of more precisely specified expert knowledge would lead to better predictions. In addition, difference between the models without expert opinion and the models with low informative expert opinion (i.e., low informative prior for the Bayesian approach and a categorical variable with two levels for the frequentist approach) was larger than the difference among the models with cluster level expert opinion.

Results further revealed that the prediction model with low informative priors suffered the most from expert judgments that deviated from the true values of the random effects. When the percentage of clusters that include discrepant expert opinion as prior information exceeds 30%, the Bayesian model with low informative priors is not recommended. However, the other two Bayesian models with medium and highly informative priors seemed less influenced by incorporation of discrepant expert opinion and still gave better predictions compared to the frequentist model. This conclusion is however conditional on how the random effects distribution is divided and how discrepant expert opinion is defined. In this simulation study, we divided the random effects normal distribution into areas with equal proportions, hence intervals for the tail areas were much wider than the intervals at the center zone. We also assumed that when an expert provided information that deviated from the true value, she would select an adjacent area rather than select at random. Although it may perhaps be realistic to make such assumptions, we could as a consequence not fully investigate the impact of incorporating expert opinion that deviates from true values as prior information in prediction. Future studies may look further into this topic. Instead of only using the neighboring truncated areas, all other discrepant possibilities could be considered. In addition, the random effects distribution may be divided into equal intervals rather than equal areas.

Table 4.2. Results from the prediction models for data simulated with prevalence = 50%, ICC = 0.20, $n = 5,000$ ($J = S = 50$, $n_j = n_c = 100$), $\beta_1 = 1.5$ (the default setting).

	Optimal score	FREQ	BAYES,WI	BAYES,LI	BAYES,MI	BAYES,HI	FREQ.2	FREQ.3	FREQ.5
Overall Brier score	0	0.191	0.192	0.179	0.174	0.170	0.173	0.170	0.167
Overall C-index / AUC	1	0.782	0.781	0.808	0.818	0.826	0.822	0.827	0.833
Overall calibration slope	1	0.911	0.907	0.957	0.982	0.989	0.965	0.972	0.994
Within cluster C-index / AUC*	1	0.805 [0.037]	0.805 [0.037]	0.805 [0.037]	0.805 [0.037]	0.805 [0.037]	0.805 [0.037]	0.805 [0.037]	0.805 [0.037]
Within cluster calibration slope*	1	0.914 [0.102]	0.914 [0.102]	0.947 [0.091]	0.956 [0.078]	0.963 [0.058]	0.954 [0.092]	0.973 [0.080]	0.977 [0.062]

*mean[sd]

Table 4.3. Results from the Bayesian models with informative priors including different percentages of discrepant expert opinion.

	FREQ	BAYES.WI		BAYES.LI			BAYES.MI			BAYES.HI		
		--	--	10%	30%	50%	10%	30%	50%	10%	30%	50%
Percentage wrong expert opinion												
Overall Brier score	0.191	0.192		0.180	0.192	0.201	0.174	0.179	0.182	0.170	0.173	0.174
Overall C-index / AUC	0.782	0.781		0.806	0.781	0.764	0.818	0.808	0.801	0.826	0.821	0.818
Overall calibration slope	0.911	0.907		0.946	0.874	0.824	0.982	0.964	0.950	0.989	0.988	0.987
Within cluster C-index / AUC*	0.805 [0.037]	0.805 [0.037]		0.805 [0.037]	0.805 [0.037]	0.805 [0.037]	0.805 [0.037]	0.805 [0.037]	0.805 [0.037]	0.805 [0.037]	0.805 [0.037]	0.805 [0.037]
Within cluster calibration slope*	0.914 [0.102]	0.914 [0.102]		0.946 [0.091]	0.939 [0.100]	0.935 [0.100]	0.953 [0.077]	0.939 [0.084]	0.935 [0.085]	0.962 [0.059]	0.953 [0.068]	0.951 [0.070]

*mean[sd]

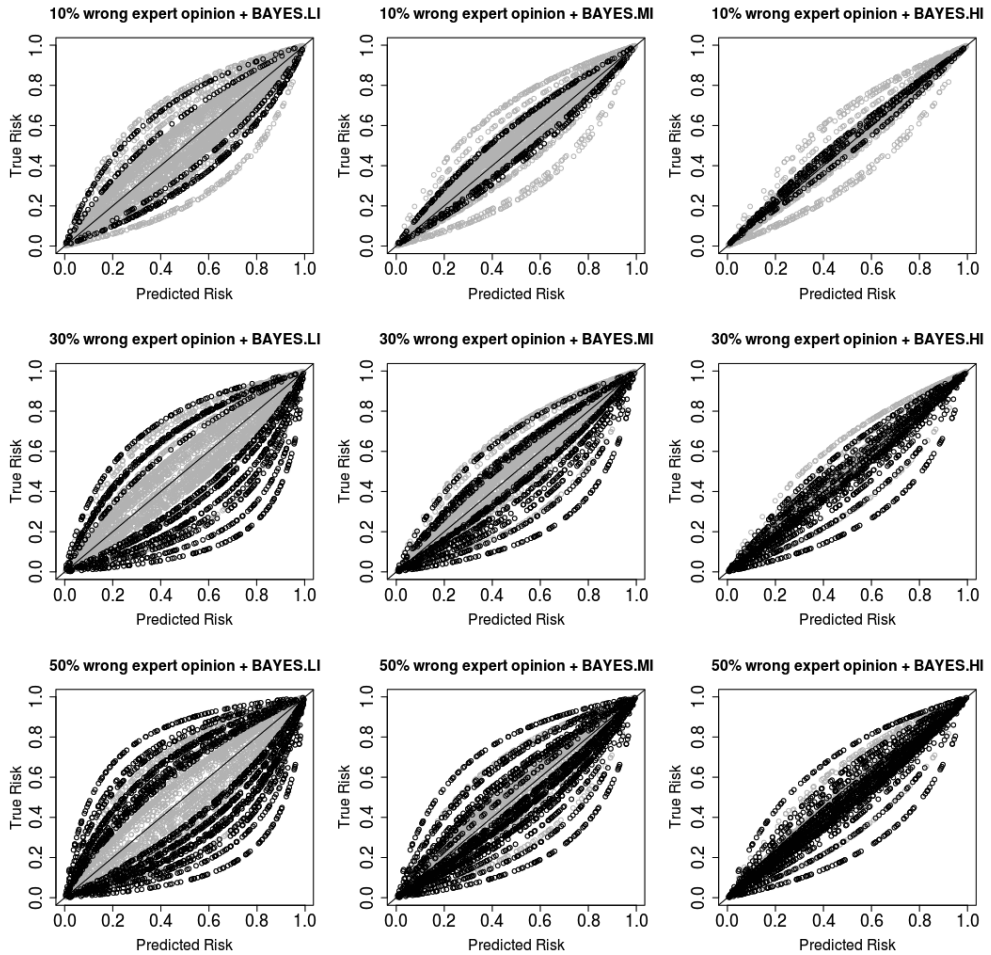


Fig. 4.3. Calibration plots for Bayesian models using discrepant expert opinion as prior information for the random effects. Predicted risks are plotted against the true latent underlying risks for 5,000 subjects from 50 equal sized clusters. Clusters using optimal expert opinion are displayed in grey color, whereas clusters using discrepant expert opinion are addressed in black color. The diagonal line is the line of identity (i.e., predicted risks are equal to the true risks). Each dot represents a subject, and each line formed by the dots represents a cluster.

To our knowledge, this study is the first attempt to combine model development data with expert opinion as prior information for random effects in prediction for new clusters. The simulated expert elicitation method is relatively novel as well. This method was also adapted and used in the frequentist models where expert opinion was incorporated as categorical

variable in this study. It is nevertheless a simulation research, and only limited scenarios have been investigated.

Further, from a practical perspective, it may be a disadvantage and should be taken into account that the Bayesian prediction models proposed in this study are more time consuming than the frequentist prediction models.

Model evaluation was performed both at the subject and the cluster level. It is debatable which measures are more informative for prediction models (Bouwmeester et al., 2013). In clinical practice, the within cluster measures might be most relevant. Other measures such as sensitivity and specificity which were not computed in this study could be used for model evaluation as well in real world applications.

4.6. Conclusion

In the context of simulated data, we investigated prediction models which incorporated cluster level expert opinion in new clusters. Results showed that the prediction models with cluster level information were better at predicting the risks for the outcome in a new cluster than the commonly used frequentist model that replaced random effects with zero after model development. We focused on the Bayesian models we proposed, as it is more intuitive to use the Bayesian approach when incorporating prior knowledge into the analysis. Future research may focus on validation in real world data and evaluation of clinical benefits.

Appendix A: Results for the sensitivity analyses

Table A1. Results from the five prediction models for data simulated with ICC = 0.20, $n = 5,000$ ($J = S = 50$, $n_j = n_c = 100$), $\beta_1 = 1.5$ and varying prevalences.

	Prevalence = 10%					Prevalence = 25%				
	FREQ					FREQ				
	BAYES.WI	BAYES.LI	BAYES.MI	BAYES.HI		BAYES.WI	BAYES.LI	BAYES.MI	BAYES.HI	
Overall Brier score	0.074	0.074	0.073	0.072	0.071	0.148	0.149	0.134	0.130	0.128
Overall C-index / AUC	0.810	0.812	0.825	0.833	0.841	0.789	0.788	0.836	0.845	0.851
Overall calibration slope	0.946	0.957	0.959	0.999	1.017	0.960	0.955	1.035	1.031	1.023
Within cluster C-index / AUC *	0.832 [0.078]	0.832 [0.078]	0.832 [0.078]	0.832 [0.078]	0.832 [0.078]	0.819 [0.040]	0.819 [0.040]	0.819 [0.040]	0.819 [0.040]	0.819 [0.040]
Within cluster calibration slope *	0.972 [0.328]	0.974 [0.321]	0.984 [0.236]	0.989 [0.181]	0.994 [0.124]	0.952 [0.204]	0.952 [0.207]	1.028 [0.148]	1.015 [0.087]	1.004 [0.054]

*mean[sd]

Table A2. Results from the five prediction models for data simulated with prevalence = 50%, $n = 5,000$ ($J = S = 50$, $n_j = n_c = 100$), $\beta_1 = 1.5$ and varying ICC values.

	ICC = 0.05					ICC = 0.50				
	FREQ	BAYES.WI	BAYES.LI	BAYES.MI	BAYES.HI	FREQ	BAYES.WI	BAYES.LI	BAYES.MI	BAYES.HI
Overall Brier score	0.178	0.178	0.175	0.174	0.173	0.199	0.198	0.167	0.157	0.153
Overall C-index / AUC	0.809	0.809	0.817	0.818	0.819	0.765	0.766	0.835	0.851	0.860
Overall calibration slope	0.970	0.970	0.983	0.986	0.987	0.834	0.839	0.946	0.985	1.006
Within cluster C-index / AUC *	0.815 [0.045]	0.815 [0.045]	0.815 [0.045]	0.815 [0.045]	0.815 [0.045]	0.814 [0.044]	0.814 [0.044]	0.814 [0.044]	0.814 [0.044]	0.814 [0.044]
Within cluster calibration slope *	0.970 [.022]	0.970 [0.022]	0.979 [0.019]	0.981 [0.017]	0.982 [0.013]	0.815 [0.208]	0.819 [0.208]	0.934 [0.213]	0.952 [0.182]	0.961 [0.144]

*mean[sd]

Table A4. Results from the five prediction models for data simulated with prevalence = 50%, ICC = 0.20, $n = 5,000$ ($J = S = 50, n_j = n_c = 100$) and varying regression parameter values.

	$\beta_1 = 0.5$, Nagelkerke's $R^2 = 0.053$					$\beta_1 = 3.0$, Nagelkerke's $R^2 = 0.596$				
	FREQ	BAYES.WI	BAYES.LI	BAYES.MI	BAYES.HI	FREQ	BAYES.WI	BAYES.LI	BAYES.MI	BAYES.HI
Overall Brier score	0.241	0.240	0.212	0.205	0.202	0.128	0.129	0.116	0.113	0.111
Overall C-index / AUC	0.613	0.618	0.725	0.747	0.753	0.900	0.899	0.917	0.921	0.923
Overall calibration slope	0.817	0.832	1.112	1.083	1.042	0.959	0.955	0.991	0.995	1.004
Within cluster C-index / AUC *	0.635 [0.063]	0.635 [0.063]	0.635 [0.063]	0.635 [0.063]	0.635 [0.063]	0.921 [0.022]	0.921 [0.022]	0.921 [0.022]	0.921 [0.022]	0.921 [0.022]
Within cluster calibration slope *	0.809 [0.150]	0.813 [0.151]	0.867 [0.155]	0.890 [0.140]	0.914 [0.117]	0.960 [0.065]	0.958 [0.065]	0.987 [0.054]	0.995 [0.046]	0.998 [0.034]

*mean[sd]

Appendix B: R code for simulating the default data setting

```

while(TRUE){
##### determine the sample sizes #####
J=50 # number of clusters per dataset
k=100 #number of subjects per cluster (equal size per cluster)

n_j <- rep(k, J) # number of individuals in jth group
id <-seq(1,J)
group <- rep(id,n_j) # group identifier per subject
N <- sum(n_j)

##### simulate datapoints for predictor X #####
x1<-rnorm(N, 0,1) # development dataset
x<-rnorm(N,0,1)    # prediction dataset

##### assign values to the parameters #####
##### change beta_0 to adjust the prevalence #####
## prevalence=50%##
beta_0 <- 0.05

## prevalence= 25% ##
#beta_0 <- 1.5
## prevalence= 10% ##
#beta_0 <- -3

beta_1 <- 1.5

# intraclass correlation coefficient
icc <-0.20 # var(u_j) is about 0.822

# variance of the random intercept can be calculated based on ICC
var_u <- (icc*(pi^2)/3)/(1-icc)

```

```

u_j1 <- rnorm(J,0,sqrt(var_u)) # true random effects for development set
u_j <- rnorm(J,0,sqrt(var_u)) # true random effects for prediction set
#####

# Now form the linear predictor
eta1 <- beta_0 + beta_1 * x1 + u_j1[group]
eta <- beta_0 + beta_1 * x + u_j[group]

# Transform back to the probability scale
p1 <- exp(eta1)/(1 + exp(eta1)) # true latent underlying risks for development set
p <- exp(eta)/(1+ exp(eta))      # true latent underlying risks for prediction set

# Binary outcomes sampled from Bernoulli distribution
y1 <- rbinom(N, 1, p1) # observed y's for development set
y <- rbinom(N, 1, p)   # observed y's for prediction set

#####
#### model development set and model prediction set
development.set <- data.frame(group,x1,y1)
prediction.set <- data.frame(group,x,y)

##### control the prevalence of the 2 datasets #####
prev1<-sum(y1)/(J*k)
prev<-sum(y)/(J*k)
#### sample data until both development and prediction sets have prevalence close to 50%
if ((prev1>0.498 & prev1<0.502)&(prev>0.498 & prev<0.502)) break()
}

```


Chapter 5

Expert opinion as priors for random effects in Bayesian prediction models: Subclinical ketosis in dairy cows as an example

Abstract

Random effects regression models are routinely used for clustered data in etiological and intervention research. However, in prediction models, the random effects are either neglected or conventionally substituted with zero for new clusters after model development.

In this study, we applied a Bayesian prediction modelling method to the subclinical ketosis data previously collected by Van der Drift et al. (2012). Using a dataset of 118 randomly selected Dutch dairy farms participating in a regular milk recording system, the authors proposed a prediction model with milk measures as well as available test-day information as predictors for the diagnosis of subclinical ketosis in dairy cows. While their original model included random effects to correct for the clustering, the random effect term was removed for their final prediction model. With the Bayesian prediction modelling approach, we first used non-informative priors for the random effects for model development as well as for prediction. This approach was evaluated by comparing it to the original frequentist model. In addition, herd level expert opinion was elicited from a bovine health specialist using three different scales of precision and incorporated in the prediction as informative priors for the random effects, resulting in three more Bayesian prediction models.

Results showed that the Bayesian approach could naturally take the clustering structure of clusters into account by keeping the random effects in the prediction model. Expert opinion could be explicitly combined with individual level data for prediction. However in this dataset, when elicited expert opinion was incorporated, little improvement was seen at the individual level as well as at the herd level. When the prediction models were applied to the 118 herds, at the individual cow level, with the original frequentist approach we obtained a sensitivity of 82.4% and a specificity of 83.8% at the optimal cutoff, while with the three Bayesian models with elicited expert opinion, we obtained sensitivities ranged from 78.7% to 84.6% and specificities ranged from 75.0% to 83.6%. At the herd level, 30 out of 118 within herd prevalences were correctly predicted by the original frequentist approach, and 31 to 44 herds

were correctly predicted by the three Bayesian models with elicited expert opinion. Further investigation in expert opinion and distributional assumption for the random effects was carried out and discussed.

Keywords: clustered data; dairy cattle; subclinical ketosis; random effects prediction model; Bayesian informative priors; expert opinion

5.1. Introduction

Random effects regression models are routinely used in etiological and intervention research. By including the random effect coefficient in the model, variance between clusters can be taken into account. However, this approach is hardly seen in prediction for clustered data. In traditional prediction models, the random effects are either neglected or conventionally substituted with zero for all new clusters after model development (Bouwmeester et al., 2012).

In a recent simulation study from Ni et al. (2018), the authors discussed this neglect in traditional prediction models and developed a Bayesian approach where the random effects could remain in the model for development as well as for prediction. Within the Bayesian framework, a prior distribution for the model parameters reflects the knowledge or uncertainty about the parameters before observing the data. By updating the prior distributions with data, posterior distributions are obtained. One of the advantages of using Bayesian modelling, therefore, is that it can naturally combine multiple sources of evidence (Spiegelhalter et al., 2004). In the study from Ni et al. (2018) for instance, (simulated) cluster level expert opinion was used as informative priors for the random effects of new clusters. The simulations showed that the Bayesian models incorporating cluster level expert opinion outperformed the traditional frequentist model, under the assumption that the expert was able to correctly predict in which part of the random effects distribution each cluster was located. A more detailed explanation of this approach will be provided in the methods section.

Prediction modelling is an explicit and empirical approach to estimate disease risk in medicine and veterinary medicine. It follows the evidence based (veterinary) medicine discipline and aims at using the current best evidence in diagnosis and making decisions for the care of individual subjects (Steyerberg, 2009). In dairy science for instance, attempts have been made to develop diagnostic methods to detect subclinical ketosis (SCK) in dairy cows based on routine milk recording data (e.g., Jorritsma et al., 1998; Krogh et al., 2011). SCK is considered one of the main metabolic disorders in early lactation dairy cows, which is defined by an increased concentration of ketone bodies in body fluids in absence of clinical signs

(Tremblay et al., 2018). Analysis of the concentration of acetone and β -hydroxybutyrate (BHBA) in blood is considered the reference test (e.g., Oetzel, 2004). A recent example was published by Van der Drift et al. (2012), who proposed a prediction model consisting of routine milk measures as well as available test-day information as predictors. The between herd variance was accounted for by random herd effects when selecting the predictors. The current SCK monitor system in the Netherlands is based on this developed prediction model. While their original model included random herd effects to correct for the clustering of cows, the random effect term was removed for their final prediction model.

In this study, we applied a Bayesian approach to the SCK data collected by Van der Drift et al. (2012). Four Bayesian prediction models were investigated. First, a Bayesian prediction model with non-informative priors for the random effects were evaluated by comparing it to the original model. Three more Bayesian models were explored with herd level expert opinion elicited from a bovine health specialist incorporated in the prediction through informative priors for the random herd effects. The main aim of this study is to explore whether the proposed Bayesian prediction modelling approach is feasible for empirical data, and whether in this dataset, it would outperform the original prediction model without random effects and improve the diagnostic accuracy for SCK.

5.2. Materials and Methods

5.2.1. Data

Van der Drift et al. collected both blood and milk samples at the individual cow level for the development of the prediction model. Throughout the paper, this model will be labeled as the 2012SCK model. In short, a total of 123 Dutch farms were randomly selected from the milk recording organization the Dutch-Flemish Cattle Improvement Cooperative (CRV), which includes 83.8% of all Dutch dairy farmers. The 123 farms were visited on a planned milk recording test day between November 2009 and November 2010 for data collection. On each farm, all cows between 5 and 60 days in milk (DIM) were blood sampled, which is the risk group for SCK. As a consequence, there were not many eligible cows present on smaller farms. Five farms were excluded due to incompleteness of cow level measures. The final dataset consisted of 1,678 cows from 118 farms. On average, 14 cows per farm were sampled, varying from 3 to 47 with a median of 13. The overall animal prevalence of SCK for the 1,678 cows was 11.2% based on the reference test results in the blood samples. Within herd animal prevalences ranged from 0% to 80% and was not symmetric with a peak at zero (39 herds).

Additional herd information was collected during the farm visit at the test day, including feeding management. Information on milk production for each herd was provided by CRV. Characterization of the feeding management for each herd was collected by means of standard questionnaire for the farmer.

5.2.2. The 2012SCK Model

The cow level measures *milk acetone*, *milk BHBA*, *milk fat-to-protein ratio*, *parity* as well as the herd level measure *season* were selected as predictors in the 2012SCK logistic regression random effects model. Milk acetone (97.41 ± 116.76 $\mu\text{mol/L}$) and milk BHBA (74.95 ± 77.82 $\mu\text{mol/L}$) measures at individual cow level were obtained from routine milk analysis by Fourier transform infrared (FTIR) spectroscopy. Milk fat-to-protein ratio (1.33 ± 0.23), parity of each animal and season during the farm visit at the test day were included as well. Observed binary outcomes were obtained by applying the plasma BHBA threshold of 1,200 $\mu\text{mol/L}$, above which an animal was considered SCK positive. The random herd effects were assumed to be normally distributed with mean 0 and variance σ_u^2 .

Parameters of the diagnostic model were estimated with maximum likelihood. Prediction of the presence of SCK in cows from new herds was based on point estimates for the regression parameters of the model where the random effect term was removed. All 118 herds were used for model development as well for model prediction. For proper comparisons, the Bayesian prediction models took the same approach.

5.2.3. Bayesian Approach

To obtain Bayesian estimates, we used non-informative priors for all the regression parameters of the predictors (normal distributions with mean 0 and variance 10,000) and for the variance of the random herd effects (inverse gamma distribution with both hyperparameters equivalent to 0.001). Three Markov chain Monte Carlo (MCMC) posterior chains were sampled. Within each chain, the first 5,000 iterations were discarded as the burn-in phase. The convergence was visually inspected using trace plots. Proper convergence was observed for all chains and the subsequent 20,000 iterations were used for parameter estimates. For the purpose of reducing computational effort in the prediction phase, the 20,000 iterations were thinned by 100, resulting in 200 per chain and 600 in total. Each of the 600 iterations consisted of a sampled value for the regression coefficients and the variance of the random effects respectively.

For the prediction without incorporation of expert knowledge, in each iteration and for each cluster a value was drawn from the random effects normal distribution with mean zero and the sampled variance. For each cow within each iteration, a predicted risk on SCK was computed by the prediction model. As a result, a distribution of predicted risk was available based on all iterations for each individual. In this study, in order to compare the results between the Bayesian models and the 2012SCK model, the median of the predicted risk distribution was used as the summarized predicted risk, resulting in a single estimate per individual cow. The R code for the Bayesian prediction model without the incorporation of expert prior knowledge can be found in Appendix A.

Herd level prior information could be incorporated by sampling the random effects from a specific part of the random effects normal distribution. For instance, when the prior information would indicate a herd to have a below average risk for SCK, the random effect for this herd would be sampled from the lower half of the distribution as it is displayed in Fig. 5.1a. By incorporating herd level prior information, we could thus restrict the parameter space for each random herd effect hence resulting in more precise estimation for the random effects.

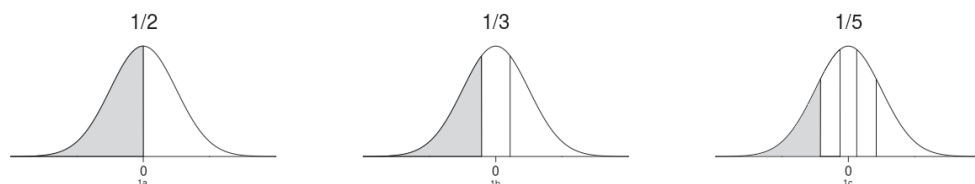


Fig. 5.1. The random effects distribution divided into multiple parts of equal proportions in three different scales. Per part contains either half, or one third, or one fifth of the distribution.

5.2.4. Expert Opinion

An experienced ruminant health specialist (GH, dipl. ECBHM, i.e., European College of Bovine Health Management) was asked to give his personal opinion for each herd on the SCK risk in early lactation cows in relation to the total Dutch dairy population.

As previous research indicated that the risk for SCK is related to routine feeding practices (e.g., Goldhawk et al., 2009) and herd level milk yield (e.g., Raboisson et al., 2014), herd level feeding management and herd level milk production documents that were collected during the farm visit (Van der Drift et al., 2012) were provided to the expert as a proxy for a farm visit. This herd level information on feeding and milk production was hence incorporated in the prediction model through elicited expert opinion.

We adopted the three scales specified for the expert elicitation from the simulation study (Ni et al., 2018). The 2-level scale divided the Dutch dairy population into two groups: the lower 50% risk group and the upper 50% risk group. The 3-level scale divided the population into three equal probability groups, and the 5-level scale divided the population into five groups (see Fig. 5.1). As all herd level information was available for the 5 herds that were not included in the analysis, these herds were used as test herds to pilot and evaluate the instruction and the scoring form for the expert elicitation. The original instruction provided for the expert can be found in Appendix B.

5.2.5. Optimal expert opinion and distributional assumption

Optimal expert opinion

In order to provide a benchmark for the best possible results using the proposed approach for this particular dataset, we also explored the predictive performance in case the elicited expert opinion would always be correct. This from here on called 'optimal expert opinion' was defined as placing all clusters in the correct part of the random effects distribution. Determination of the correct part was based on the percentile/ranking of each herd among all 118 observed within herd animal prevalences which were computed using the true disease status of the cows determined by plasma BHBA values. Values for the random intercepts were still randomly drawn from the assigned part of the random effects distribution. Predictions were subsequently made in the same way as for the real expert model.

Distributional assumption

The assumed distribution for random cluster effects is almost always the normal distribution and it is usually chosen for computational convenience (Agresti et al., 2004). Some researchers argued that misspecification of the random effects distribution had little impact on parameter estimates (Butler & Louis, 1992; Heagerty & Kurland, 2001), while others pointed out that regression parameter estimates were very sensitive to the random effects distribution and suggested more flexible distributional assumptions (Litière et al., 2007; McCulloch & Neuhaus, 2011). Given the asymmetric nature of the SCK within herd prevalences, we also investigated the skew-normal distribution with optimal expert opinion for all Bayesian models (Azzalini & Capitanio, 2014). The random effects under the skew-normal assumption were sampled from the asymmetric distribution with mean zero, variance σ_u^2 , and skewness α_u . For parameter estimation, the same non-informative priors were specified as for the normal random

effects models. A normal prior distribution with mean zero and standard deviation one was specified for the skewness parameter α_u .

5.2.6. Model Notations and Computation

We denote our reproduced 2012SCK model as FREQ (i.e., the frequentist model). Further, four Bayesian models were specified. We denote the Bayesian prediction model without herd specific information as Bayes0. The Bayesian models with herd level prior information are denoted as Bayes2 (2-level scale), Bayes3 (3-level scale), Bayes5 (5-level scale) respectively. The model assessment measures used were the same as in the simulation study (Ni et al., 2018), that is the area under the curve (AUC), Brier score (i.e., mean squared error), calibration slope, sensitivity and specificity at the optimal cutoff, sensitivity at 95% and 90% specificity cutoff. All analyses were carried out in R (R core team, 2013). The frequentist parameter estimates were obtained by using the package ‘lme4’ (Bates et al., 2015) and the Bayesian results were obtained by calling Stan from R using the package ‘rstan’ (Stan development team, 2018).

5.3. Results

5.3.1. Estimation

The 2012SCK model was reproduced using the generalized linear model function in ‘lme4’. After adjusting the optimizer to ‘bobyqa’, parameter estimates from our analysis were identical to the originally reported results in Van der Drift et al. (2012) up to the third decimal, but the standard errors differed slightly. The estimated intraclass correlation coefficient (ICC) was 0.35, and the Nagelkerke’s R^2 was 41.2%.

Parameters were further estimated in a Bayesian approach using non-informative priors. The point estimates for the regression coefficients from the posterior were similar to the results from the reproduced 2012SCK model (see Appendix C).

5.3.2. Prediction and Model Comparisons

Animal level

The reproduced 2012SCK model (FREQ) showed identical diagnostic accuracy as the originally reported results. Table 5.1 presents the diagnostic performance of the models at the individual cow level. The effects of including the elicited expert opinion, as well as the simulated optimal expert opinion were assessed within the Bayesian approach. As can be seen, the Bayesian model without herd level information (Bayes0) performed approximately the same as the frequentist model. The Bayesian models with elicited expert opinion showed no

improvement in comparison to the frequentist model. However, the Bayesian models with optimal expert opinion slightly outperformed the frequentist model, with Bayes5 showing the best prediction.

Herd level

Predicted within herd animal prevalences were compared to the observed prevalences based on the reference test results in blood samples. The predicted prevalence of each herd was calculated from the predicted binary outcome (with disease probability >0.50 as positive) from each cow within the herd using the optimal cutoff (defined as maximum sum of sensitivity and specificity). As expected, the frequentist model and the Bayesian model without herd specific information resulted in similar predictive accuracy. The Bayesian models with elicited expert opinion from the 2-level and 3-level scales had more herds correctly estimated (44 and 43 respectively) compared to the frequentist model (30), and less false positives (19 and 20 respectively) compared to the frequentist model (33). Further, Bayesian models with optimal expert opinion outperformed the Bayesian models with elicited expert opinion regarding the number of false positives (see Appendix C).

5.3.3. Expert Opinion

Table 5.2 presents a summary of the number of herds assigned to each risk group within each scale by the expert. More herds were assigned to the lower risk group(s) by the expert than to the higher risk group(s) in all three scales. When the number of risk groups increased (i.e., from 2 to 5 levels), the frequency of disagreement between the elicited expert opinion and the observed within herd animal prevalences increased accordingly. This can also be seen in the three plots (Appendix C). It should be noted that in the 5-level scale, as 39 out of 118 herds have zero diseased cows which exceeds 20% that cannot be ranked, the lowest 40% herds are combined into one risk group.

Results also reveal that the degree of agreement between the elicited expert opinion and the observed within herd animal prevalence is affected by the herd sample sizes. Table 5.3 shows that for herds with at least 12 sampled cows, there is more agreement between expert opinion and observed within herd prevalence than herds with less than 12 sampled cows.

Table 5.1. Animal level measures ($n = 1,678$) area under the curve (AUC), Brier score, calibration slope, sensitivity (Se), specificity (Sp) using the optimal cutoff, sensitivity using the 95% and 90% specificity cutoffs for the predicted outcomes.

	Optimal Value	Elicited expert opinion					Optimal expert opinion				
		FREQ	Bayes0	Bayes2 (2 levels)	Bayes3 (3 levels)	Bayes5 (5 levels)	Bayes2 (2 levels)	Bayes3 (3 levels)	Bayes5 (5 levels)		
AUC (%)	100	88.5	88.3	88.2	87.5	88.5	91.1	92.4	92.5		
Brier score	0	0.069	0.069	0.071	0.077	0.084	0.062	0.059	0.058		
Calibration slope	1	0.809	0.787	0.674	0.605	0.784	0.796	0.832	0.821		
Se (optimal cutoff) (%)	100	82.4	81.4	84.6	78.7	80.9	81.4	81.4	88.3		
Sp (optimal cutoff) (%)	100	83.8	83.5	75.0	82.8	83.6	85.8	86.7	80.3		
Se (95% Sp cutoff) (%)	100	51.1	51.6	49.5	44.7	52.1	56.9	63.3	64.4		
Se (90% Sp cutoff) (%)	100	69.7	69.1	66.0	61.7	70.2	74.5	76.1	75.0		

Table 5.2. Summary of the elicited expert opinion on 118 herds. Each column presents the number of herds assigned to each risk group within each scale (from the lowest risk to the highest risk).

	2-level scale	3-level scale	5-level scale
Low risk	72	51	33
↓			33
		48	26
			18
High risk	46	19	8
Total	118	118	118

Table 5.3. The number of herds agreed between the elicited expert opinion and the observed within herd animal prevalence on the relative position of each herd among the 118 herds.

	Agreement (%)	
	Herd sample size <12	Herd sample size ≥12
	(<i>n</i> = 48)	(<i>n</i> = 70)
2-level scale	30 (62.5)	45 (64.3)
3-level scale	18 (37.5)	38 (54.3)
5-level scale	19 (39.6)	31 (44.3)

5.3.4. Distributional Assumption

At the individual cow level, the four Bayesian models with a skew-normal distribution for the random effects performed similar to the Bayesian models with the normal distribution. At the herd level, the skew-normal models had more herds correctly estimated and less false positives at the alarm level of 10% (see Appendix C).

5.3.5. Splitting Data into a Training and a Test Set

In order to compare objectively between results from the frequentist model in the original study Van der Drift et al. (2012) and the Bayesian prediction models in this study, we used the same dataset for model development as well as for model prediction as it was done in the original study. However, we additionally performed model comparisons based on a 80/20 training/test set approach as follows. About 80% of the 118 herds were randomly selected for model development, resulting in 94 herds with 1,331 cows. The number of cows per herd was

approximately 14 on average and 32 out of the 94 herds had zero diseased animals. The parameter estimates from the frequentist as well as the Bayesian approach of the training set are shown in Table C5 in Appendix C, while the prediction results on the test set are found in Table C6.

5.4. Discussion

This study demonstrates an application of a Bayesian prediction modelling approach (Ni et al., 2018) that incorporates the clustering structure by keeping the random effects in the prediction model. In addition, this approach provides a natural framework to combine evidence from various sources, such as expert in the field in our SCK example. Herd level expert opinion provided by the bovine health specialist enabled us to combine herd specific information with individual level milk measures and available test-day data in the prediction. However in this dataset, little improvement was seen in prediction resulting from the Bayesian models with elicited expert opinion incorporated in comparison to the prediction from the original frequentist model. The predictive performance from the three Bayesian models with different levels of precision in expert opinion remained poor at the individual cow level. At the herd level, the Bayesian prediction models showed slightly higher diagnostic accuracy, with more within herd prevalences being correctly estimated and less false positives at the alarm level of 10%. We therefore conclude in agreement with Van der Drift et al. (2012) that the prediction models, both without and with the addition of herd level information, are not suited for the cow level SCK diagnosis.

A reason to include the simulated optimal expert was to rule out the possibility that predictive performance did not improve because the elicited expert opinion was of suboptimal quality. In the current study, we observed that the expert tended to underestimate the herd risk for SCK, as more herds were assigned to the lower risk group(s). In practice, one could try to improve the quality of the expert knowledge by eliciting multiple experts and the application of approved methods to reach agreement between the experts (O'Hagan et al., 2006). In this study, we decided to simulate expert information under the assumption that the expert was always correct. Note that this was a methodological exercise to further investigate the potential of the proposed approach for this particular dataset, and not an approach that should be applied in practice to reach better predictive performance. Although the simulation study from Ni et al. (2018) demonstrated that the Bayesian models including correct cluster level expert opinion was able to provide better predictions, the improvement in predictive performance in this study

was limited. Therefore, we investigated several other possible explanations for lack of (substantial) improvement as well.

Another explanation for little improvement in prediction both at the individual cow level and at the herd level might be the small sample sized herds in this study. In herds that consisted of at least 12 sampled cows, the degree of agreement was higher between the observed within herd prevalences resulted from the reference test with blood samples and the elicited expert opinion than in herds with less than 12 sampled cows in all three scales. As Oetzel (2004) concluded in his study, the minimum sample size for herd-based tests that gave moderate confidence (75% or more) was 12. In our dataset, the observed within herd prevalences from the small sample sized herds ($n < 12$) may therefore not represent the true prevalence for these herds. However, a sensitivity analysis in including only the 70 herds that had at least 12 cows ($n \geq 12$) showed similar results compared to 118 herds (results not shown).

Also, the relatively low ICC in this dataset may limit the benefit of incorporating cluster level prior information. Using the data from all 118 herds, the ICC was 0.35. However, 39 out of 118 herds had zero SCK animals which influenced the ICC and the subsequent random effects estimation. The estimated variance of the random effects was 1.792 when all 118 herds were included, but reduced to 0.539 when 39 herds with zero diseased cows were removed. This removal also reduced the ICC substantially, to the value 0.14. A lower ICC leaves less potential influence of the clustering effect, hence less benefit from adding herd level prior information to a prediction model.

Finally, the distributional assumption for the random effects may have influenced the estimation as well. The normal distribution was examined by comparing it with skew-normal distribution within the Bayesian models using the optimal expert opinion. The model with random herd effects under skew-normal distributional assumption did not show better prediction at the individual cow level than the model with random effects under normality assumption. However, the skew-normal random effects Bayesian models provided better predictions at the herd level than the respective Bayesian models with normal random effects, which indicated that the skew-normal random effects distribution may be better suited for zero-inflated data.

The regression models based on the training set showed very similar parameter estimates to the full dataset, albeit with larger standard errors for the regression coefficients and lower variance for the random effects. The model assessments on the test set showed less favorable prediction results for all methods, as was to be expected, but did not alter our conclusions about the comparisons between the methods.

5.5. Conclusions

This study illustrates how the Bayesian prediction modelling approach can take the clustering effect into account and how cluster level expert opinion can be combined with individual level data. However in this dataset, incorporation of elicited expert opinion did not improve prediction at the individual level nor at the herd level. Therefore, further investigation of the potential gain of using this approach requires applications in studies where the between cluster variance is relatively large and where all clusters harbor individuals with the outcome under study.

5.6. Acknowledgments

The authors thank Hiemke M. Knijn from the Dutch-Flemish Cattle Improvement Cooperative (CRV) for providing the milk production data. Dr. Tine van Werven and Joost de Veer are gratefully acknowledged for their help in pre-testing the expert elicitation form and giving valuable feedback for improvement in the instruction.

Appendix A: R code for Bayesian prediction model without expert prior knowledge

Three steps:

1. Use 'rstan' package to get posterior estimates for the parameters
2. Save the posterior iterations from all chains per parameter
3. Predict the SCK risk per animal

Step 1: use 'rstan' to get the posterior estimates for the parameters

First specify the model and the priors in "model development.stan"

```
data {
  int<lower=0> Nk;
  int<lower=0> Nj;

  int<lower=0,upper=1> y[Nk];
  int<lower=1> herds[Nk];
  vector[Nk] parity2;
  vector[Nk] parity3;
  vector[Nk] parity4;
  vector[Nj] spring;
  vector[Nj] winter;
  vector[Nj] summer;
  vector[Nk] bhbz_f;
  vector[Nk] acet_f;
  vector[Nk] ve_rat;
}
```

```
parameters {
  real beta_0;
  real beta_1;
  real beta_2;
  real beta_3;
  real beta_4;
  real beta_5;
  real beta_6;
```

```

real beta_7;
real beta_8;
real beta_9;

// Level-2 random effect
real u_j[Nj];
real sigma_u;
}

transformed parameters {
real herd_level[Nj];
real mu_herd[Nk];

for (j in 1:Nj) {
herd_level[j] <- beta_4 * spring[j] + beta_5 * winter[j] + beta_6 * summer[j] + u_j[j];
}

for (i in 1:Nk) {
mu_herd[i] <- herd_level[herds[i]];
}
}

model {
// non-informative priors
beta_0 ~ normal(0,100);
beta_1 ~ normal(0,100);
beta_2 ~ normal(0,100);
beta_3 ~ normal(0,100);
beta_4 ~ normal(0,100);
beta_5 ~ normal(0,100);
beta_6 ~ normal(0,100);
beta_7 ~ normal(0,100);
beta_8 ~ normal(0,100);
beta_9 ~ normal(0,100);

```

```
// Random effects u_j
u_j ~ normal(0, sigma_u);
pow(sigma_u, 2) ~ inv_gamma(0.001, 0.001);

// Likelihood
for (i in 1:Nk) {
y[i] ~ bernoulli_logit(beta_0 + beta_1 * parity2[i] + beta_2 * parity3[i] + beta_3 * parity4[i] + beta_7 * bhbz_f[i]
+ beta_8 * acet_f[i] + beta_9 * ve_rat[i] + mu_herd[i]);
}
}

## then run the model with 3 chains, with each chain having 5000 burn-in iterations and the saved 20000
iterations thinned by 100
fileName <- './model development.stan'
stan_code <- readChar(fileName, file.info(fileName)$size)
cat(stan_code)

development <- stan(model_code = stan_code, data = data, chains = 3, iter = 25000, warmup = 5000, thin =
100)

# saved posteriors for parameters
print(development, pars = c('beta_0', 'beta_1', 'beta_2', 'beta_3', 'beta_4', 'beta_5', 'beta_6', 'beta_7', 'beta_8',
'beta_9', 'sigma_u'), digits=4)

# traceplots
rstan::traceplot(development, pars = c('beta_0', 'beta_1', 'beta_2', 'beta_3', 'beta_4', 'beta_5',
'beta_6', 'beta_7', 'beta_8', 'beta_9', 'sigma_u'), inc_warmup = FALSE)

## Step 2: save the posterior iterations per parameter for prediction ##
stan.output <- extract(development, permuted = TRUE, inc_warmup = FALSE)
post.beta0<- stan.output[[1]]
post.beta1<- stan.output[[2]]
post.beta2<- stan.output[[3]]
post.beta3<- stan.output[[4]]
post.beta4<- stan.output[[5]]
post.beta5<- stan.output[[6]]
post.beta6<- stan.output[[7]]
```



```

post.beta7<- stan.output[[8]]
post.beta8<- stan.output[[9]]
post.beta9<- stan.output[[10]]
post.sdU <- stan.output[[12]]

}

## Step 3: predict for each animal the SCK risk based on the posteriors ##
pPred.all<-c()
yPred.all<-c()
pPred <-c()
yPred <-c()

nr.iter<- length(post.beta0)
nr.datapoints<-length(ket_1.2)
nr.clusters <- length(unique(herds))
post.uj<-matrix(, nrow = nr.iter, ncol = nr.clusters)

# in each iteration and for each herd, draw a value from the normal distribution with mean 0 and sampled
posterior standard deviation for the random effects
for (j in 1:nr.clusters) {
  uj.hatall <- c()

  for (i in 1:nr.iter) {
    uj.hat <- rnorm(1, 0, sd=post.sdU[i])
    uj.hatall <- c(uj.hatall, uj.hat)
  }
  post.uj[,j] <-uj.hatall
}

cum.herd <- ave(herd.size, FUN=cumsum)

# predict the risk on SCK for all animals (pPred.all) and use cut-off 0.5 to determine the predicted binary
outcomes (yPred.all)
for (d in 1:nr.clusters) {
  for (j in (cum.herd[d]-herd.size[d]+1):(cum.herd[d])) {
    for (i in 1:nr.iter) {

```

```
lp <- post.beta0[i] + post.beta1[i] * parity2[j] + post.beta2[i] * parity3[j] + post.beta3[i] * parity4[j] +
post.beta4[i] * spring[j] + post.beta5[i] * winter[j] + post.beta6[i] * summer[j] + post.beta7[i] * bhz_f[j] +
post.beta8[i] * acet_f[j] + post.beta9[i] * ve_rat[j] + post.uj[,d][i]
```

```
pPred<- exp(lp)/(1+exp(lp))
```

```
pPred.all <- c(pPred.all, pPred)
```

```
yPred <- ifelse(pPred<0.5, 0, 1)
```

```
yPred.all <-c(yPred.all, yPred)
```

```
}
```

```
}
```

```
}
```

Appendix B: Instruction for expert elicitation

The Instruction

Background Information This study is an addition to the earlier study by Van der Drift et al.¹ (2012) on subclinical ketosis in Dutch dairy farms. In the original study, data were collected from 123 randomly selected Dutch farms between November 2009 and November 2010. A diagnostic model for subclinical ketosis in early lactation dairy cows (5-60 DIM) in the Netherlands was developed on the basis of the data. The model contained predictors *parity*, *season*, *milk fat-to-protein ratio*, *milk acetone* and *milk β -hydroxybutyrate (BHBA)*. The final analysis was performed with 1,678 cows from 118 farms. For your information, we add the paper (Van der Drift et al., 2012), however, it is not necessary to read the original paper or search for other relevant literature. We would like to ask you to provide your personal opinion based on your existing knowledge and experience.

This Study In addition to the animal level predictors, the researcher also recorded feeding management and ration during her visit at the farms. For some farms, proposals from feed advisors were also provided by the farm owners. Milk production registration (MPR) summaries at the herd level were obtained from the organization Cattle Improvement Cooperative (CRV). The MPR reports were based on the test day that the researcher visited the farms.

In this study, we would like to include the above mentioned farm level information in the diagnostic model. To do so, we would like to ask you to examine the information from each farm and make your personal opinion about the risk level for subclinical ketosis of the farm relative to other farms. Please note, we are *not* asking for absolute numbers, such as the actual risk or prevalence, but for the position of the farm in the Dutch dairy farm population. We hence assume that you have in your mind an idea of how an ‘average’ farm looks like in the population of the Dutch dairy farms. Then for each farm, we ask you to label the farm as, for instance, average, below average or above average. With below average we mean a lower risk

¹ Van der Drift SGA, Jorritsma R, Schonewille JT, Knijn HM, Stegeman JA. Routine detection of hyperketonemia in dairy cows using Fourier transform infrared spectroscopy analysis of β -hydroxybutyrate and acetone in milk in combination with test-day information. J Dairy Sci. 2012;95:4886–4898.

and with above average a higher risk for subclinical ketosis for dairy cows in early lactation. Follow your gut feeling as an expert that knows Dutch dairy farms well.

We provide three different scales and ask you to fill in each scale for each farm. We will first define the three scales and then give two examples.

1. The 2-level scale divides the Dutch dairy farm population into two equal groups. A herd can be placed either to the lower 50% risk group or to the upper 50% risk group of the population (i.e., below or above ‘average’).
2. The 3-level scale divides the population into three equal groups. A herd can be placed either to the lowest 33.3% risk group, or to the middle 33.3% risk group, or to the highest 33.3% risk group of the population.
3. The 5-level scale divides the population into five equal groups. A herd can be placed either to the lowest 20%, or to the highest 20%, or to the three risk groups in between, regarding its position in the population.

As an example, consider that you judge a farm to be of very low risk. On the 2-level and 3-level scale, you will then probably choose the very left box (as illustrated in the picture below). The 5-level scale may be more demanding: is the risk compared to other farms extremely low (first box) or below average but maybe not at the lowest 20% (second box; as illustrated in the picture below). Even when it is difficult to decide between the boxes, please always make one choice. There is no right or wrong answer, what we would like to know is your personal (subjective) guess.

2-level scale	3-level scale	5-level scale
50% 50%	33.3% 33.3% 33.3%	20% 20% 20% 20% 20%
<input checked="" type="checkbox"/> <input type="checkbox"/>	<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

A second example shows the boxes when you consider a farm to be very average. Then the 3-level and 5-level scales are perhaps easy (the middle box; as illustrated in the picture below), but in the 2-level scale you are still asked to make a choice between below average and above average. Give your best guess even if it is difficult to decide.

2-level scale	3-level scale	5-level scale
50% 50%	33.3% 33.3% 33.3%	20% 20% 20% 20% 20%
<input checked="" type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

The scoring form

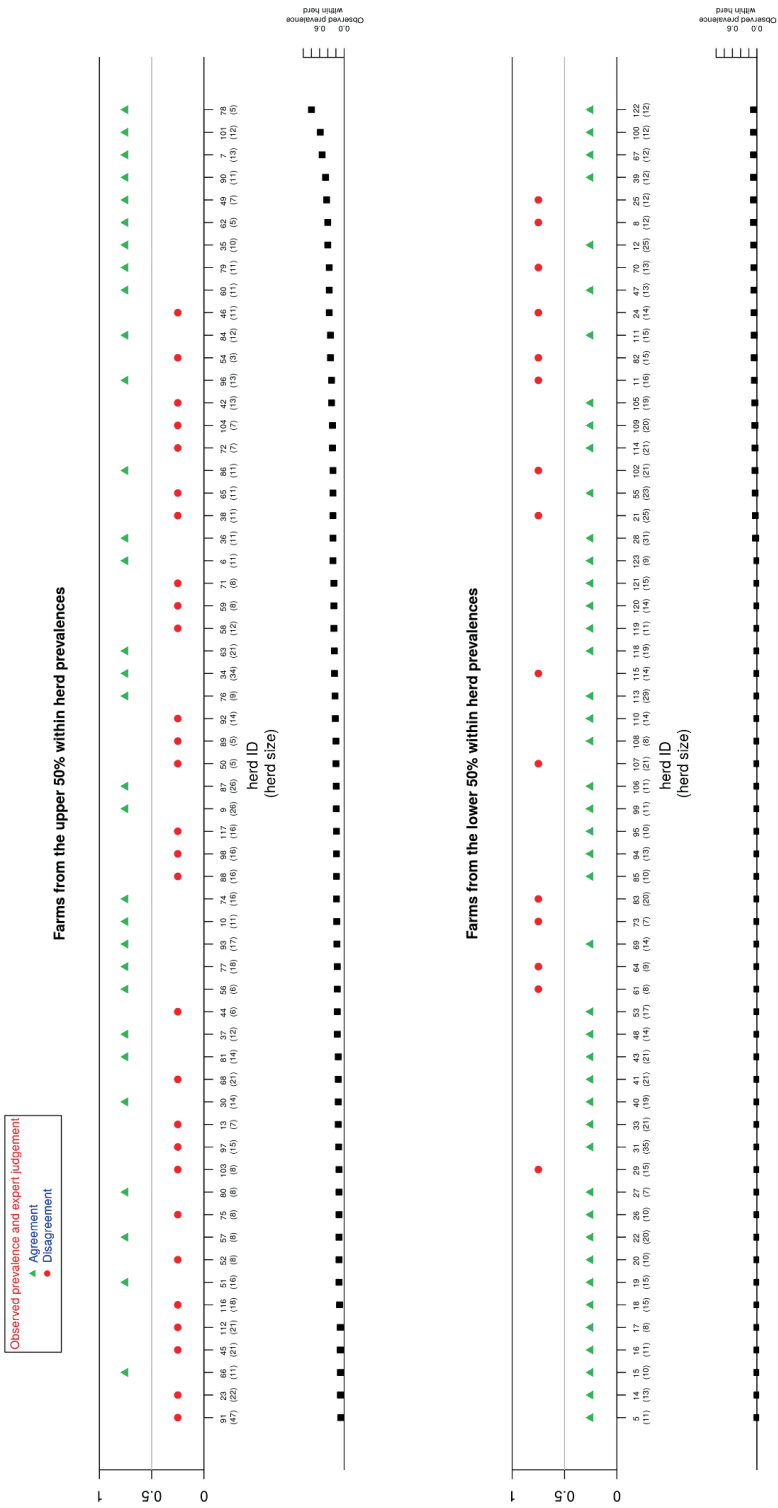
The scoring form can be found in the subsequent pages. Each page contains 10 farms (except for the first page), with each row representing one farm. The label in the front indicates the farm ID. For instance, B1 is for farm number one. Please always cross 1 box in each scale for each farm.

The 5 farms B1, B2, B3, B4, B32 that were collected but excluded from the final analysis will be used as practice farms. This is presented on the first page, and you may use the 5 farms to familiarize yourself with the three scales prior to the 118 farms that will be included in the analysis. Please contact me when you finish the 5 practice farms. We can then make an appointment and I will bring you the materials for the 118 farms.

Appendix C: Supplementary tables and figures

Table C1. Frequentist and Bayesian estimates for the regression coefficients and variance of the random effects.

	2012SCK model reproduced		Bayesian approach	
	Mean	(SE)	Mean	(SE)
Intercept	-9.097	0.940	-9.538	0.973
Parity 1	Referent		Referent	
Parity 2	-0.055	0.373	-0.074	0.407
Parity 3	0.690	0.348	0.699	0.380
Parity >= 4	1.362	0.313	1.391	0.335
Fall	Referent		Referent	
Winter	0.179	0.531	0.214	0.566
Spring	1.514	0.520	1.621	0.534
Summer	1.184	0.512	1.256	0.542
Milk fat-to-protein ratio	2.534	0.535	2.677	0.538
Milk acetone ($\mu\text{mol/L}$)	0.0100	0.002	0.0105	0.002
Milk BHBA ($\mu\text{mol/L}$)	0.0029	0.002	0.0029	0.002
Variance of the random effects	1.792	1.339	2.230	0.714



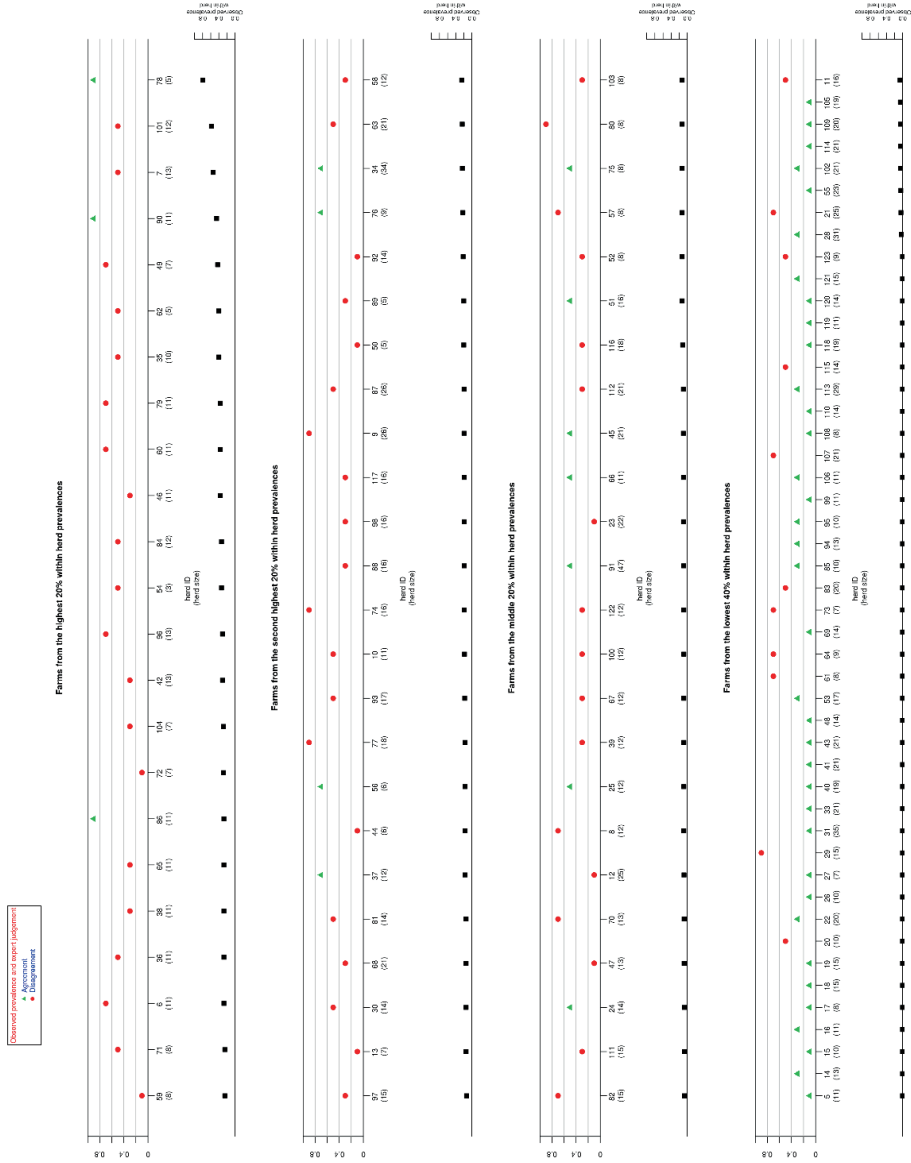


Fig. C3. The agreement between the observed within herd animal prevalence and the elicited expert opinion in the 5-level scale.

Table C2. The number of correctly/under-/over-estimated within herd animal prevalences ($n = 118$) from the frequentist model and the Bayesian models with herd level prior information incorporated using the optimal cutoff.

	Correct Prevalence (%)	Underestimated Prevalence (%)	Overestimated Prevalence (%)	Alarm level of prevalence 10% ($n = 54$)	
				False positives (%)	False negatives (%)
FREQ	30 (25.4)	11 (9.3)	77 (65.3)	33 (51.6)	2 (3.7)
Bayes0	30 (25.4)	8 (6.8)	80 (67.8)	36 (56.3)	1 (1.9)
Bayes2	44 (37.3)	20 (16.9)	54 (45.8)	19 (29.7)	9 (16.7)
(2 levels)	42 (35.6)	15 (12.7)	61 (51.7)	10 (15.6)	1 (1.9)
Bayes3	43 (36.5)	13 (11.0)	62 (52.5)	20 (31.2)	4 (7.4)
(3 levels)	48 (40.7)	12 (10.1)	58 (49.2)	13 (20.3)	5 (9.3)
Bayes5	31 (26.3)	13 (11.0)	74 (62.7)	31 (48.4)	3 (5.6)
(5 levels)	38 (32.2)	8 (6.8)	72 (61.0)	15 (23.4)	2 (3.7)

Table C3. Comparison between the Bayesian models with optimal expert opinion under the normal and skew-normal assumption for the random herd effects on animal level measures: area under the curve (AUC), Brier score, calibration slope, sensitivity (*Se*) and specificity (*Sp*) using the optimal cutoff, sensitivity using the 95% and 90% specificity cutoffs for the predicted outcomes.

	Optimal value	FREQ	Optimal expert opinion (Normal distribution)					Optimal expert opinion (Skew-normal distribution)				
			Bayes0	Bayes2 (2 levels)	Bayes3 (3 levels)	Bayes5 (5 levels)	Bayes0	Bayes2 (2 levels)	Bayes3 (3 levels)	Bayes5 (5 levels)		
AUC (%)	100	88.5	88.3	91.1	92.4	92.5	88.3	90.8	92.8	92.9		
Brier score	0	0.069	0.069	0.062	0.059	0.058	0.070	0.069	0.064	0.060		
Calibration slope	1	0.809	0.787	0.796	0.832	0.821	0.782	0.646	0.743	0.755		
<i>Se</i> (optimal cutoff) (%)	100	82.4	81.4	81.4	81.4	88.3	80.9	80.9	80.9	89.4		
<i>Sp</i> (optimal cutoff) (%)	100	83.8	83.5	85.9	86.7	80.3	84.0	86.0	87.8	80.7		
<i>Se</i> (95% <i>Sp</i> cutoff) (%)	100	51.1	51.6	56.9	63.3	64.4	51.6	61.7	64.9	64.4		
<i>Se</i> (90% <i>Sp</i> cutoff) (%)	100	69.7	69.1	74.5	76.1	75.0	69.1	75.0	76.6	75.5		

Table C4. The number of correctly/under-/over-estimated within herd animal prevalences ($n = 118$) from the frequentist model and the Bayesian models with optimal expert opinion incorporated and under either normal or skew-normal distribution for the random herd effects using the optimal cutoff.

	Correct	Underestimated		Overestimated	Alarm level of prevalence 10% ($n = 54$)	
	Prevalence (%)	Prevalence (%)	Prevalence (%)		False positives (%)	False negatives (%)
FREQ	30 (25.4)	11 (9.3)	77 (65.3)	33 (51.6)	2 (3.7)	
Bayes0	Normal	8 (6.8)	80 (67.8)	36 (56.3)	1 (1.9)	
	Skew-normal	12 (10.2)	76 (64.4)	33 (51.6)	3 (5.6)	
Bayes2 (2 levels)	Normal	15 (12.7)	61 (51.7)	10 (15.6)	1 (1.9)	
	Skew-normal	19 (16.1)	50 (42.4)	5 (7.8)	1 (1.9)	
Bayes3 (3 levels)	Normal	12 (10.1)	58 (49.2)	13 (20.3)	5 (9.3)	
	Skew-normal	18 (15.3)	45 (38.1)	3 (4.7)	6 (11.1)	
Bayes5 (5 levels)	Normal	8 (6.8)	72 (61.0)	15 (23.4)	2 (3.7)	
	Skew-normal	7 (5.9)	65 (55.1)	12 (18.7)	1 (1.9)	

Table C5. Frequentist and Bayesian parameter estimates based on the training set (94 herds with 1,331 cows).

	2012SCK model reproduced		Bayesian approach	
	Mean	(SE)	Mean	(SE)
Intercept	-9.255	1.060	-9.607	1.107
Parity 1	Referent		Referent	
Parity 2	-0.055	0.427	-0.028	0.467
Parity 3	0.668	0.391	0.710	0.394
Parity >= 4	1.277	0.345	1.340	0.355
Fall	Referent		Referent	
Winter	0.046	0.589	0.057	0.627
Spring	1.549	0.550	1.642	0.582
Summer	1.256	0.538	1.354	0.589
Milk fat-to-protein ratio	2.703	0.611	2.804	0.639
Milk acetone ($\mu\text{mol/L}$)	0.0100	0.002	0.0099	0.002
Milk BHBA ($\mu\text{mol/L}$)	0.0028	0.002	0.0028	0.002
Variance of the random effects	1.491	1.221	1.727	0.710

Table C6. Animal level measures ($n = 347^*$) area under the curve (AUC), Brier score, calibration slope, sensitivity (Se), specificity (Sp) using the optimal cutoff and sensitivity using the 95% and 90% specificity cutoffs for the predicted outcomes.

		Elicited expert opinion				
	Optimal	FREQ	Bayes0	Bayes2 (2 levels)	Bayes3 (3 levels)	Bayes5 (5 levels)
	Value					
AUC (%)	100	83.1	83.0	85.7	88.1	87.5
Brier score	0	0.082	0.082	0.076	0.073	0.073
Calibration slope	1	0.719	0.701	0.667	0.735	0.674
Se (optimal cutoff) (%)	100	72.7	72.7	68.2	88.6	81.8
Sp (optimal cutoff) (%)	100	85.8	83.8	85.8	71.3	76.6
Se (95% Sp cutoff) (%)	100	50.0	50.0	50.0	52.3	52.3
Se (90% Sp cutoff) (%)	100	63.6	63.6	63.6	61.4	65.9

*94 herds with 1,331 cows were randomly selected as the training set, and the remaining 24 herds with 347 cows were subsequently predicted.

Part III

Cross-Species Evidence

Chapter 6

The effect of oral Glucosamine and Chondroitin in aged horses: A clinical trial re-analyzed with external data as prior information

Abstract

Joint stiffness and pain caused by osteoarthritis (OA) is one of the major health issues in aged horses. In recent years, dietary supplements have gained popularity with expected effect to alleviate OA symptoms. The combination of nutraceuticals Glucosamine (GS) plus Chondroitin Sulfate (CS) is one of the common options. However, research studies that assess the clinical effectiveness of GS and/or CS in equine OA patients are scarce and yield conflicting results.

In this study, by using the Bayesian power prior approach, we re-analyzed the data from a Dutch equine trial where treatment effect from oral supplement with GS and CS on stiffness in aged horses was evaluated. The Bayesian power prior method enabled us to borrow information from comparable biosimilar studies into an informative prior for the actual analysis of the equine trial.

Systematic searches were done in the species horse, dog and human. Cross-species search terms were specified on the basis of the PICO components extracted from the equine trial. A total group of 20 studies remained after screening and eligibility assessments, from which 9 used the combination of GS and CS as treatment and were included in the main power prior analysis. An equine specialist weighted each study in relation to the clinical research question of the equine trial. Before data analysis, in order to have all outcomes on a same scale, standardized values were derived from the raw outcome measures of the equine trial and the historical studies.

Posterior inference of the standardized treatment effect showed no difference between the treatment and control arm (posterior mean = 0.179 with a 95% central credible interval (CCI) [-0.105, 0.464]). This supports the results from the original equine trial. However, further sensitivity analyses showed heterogenous results which was in agreement with the current state of knowledge on this clinical intervention. This study shows a method to combine cross-species interventions towards clinical interpretation.

Keywords: horse, nutraceuticals, stiff joint, systematic search, cross-species evidence, Bayesian analysis, power prior approach

6.1. Introduction

Stiff joints and lack of joint flexibility are one of the main health problems in elderly horses, most likely caused by inflammation and pain due to osteoarthritis (OA) (Ireland et al., 2012). Dietary supplements, especially nutraceuticals, to reduce OA symptoms have gained popularity during the last two decades for their reputation of being well tolerated and safe regarding adverse effects. The most widely used nutraceuticals in horses are those related to chondroprotection such as glucosamine (GS) and chondroitin sulphate (CS; Higler et al., 2014).

However, studies that assess the clinical efficacy of GS and/or CS in equine OA patients are scarce and yield conflicting results (e.g., Hanson et al., 2001; Clayton et al., 2002; White et al., 2003; Gupta et al., 2009). For example, one trial by Forsyth et al. (2006) showed a large clinical effect of the oral GS and CS supplement on stride parameters of old horses, trotted in-hand by an experienced handler. A subsequent trial by Higler et al. (2014), where the stride parameters were objectively quantified at a standardized treadmill walk and trot, found no treatment effect. Therefore, it is difficult to draw conclusions and base clinical advice for equine patients when only a few relatively small trials are available.

In fact, OA is a common disease in other species as well. One veterinary systematic review on this topic conducted by Vandeweerd et al. (2012), for instance, included 16 studies in dogs, 5 in horses and 1 in cats. The majority of OA treatment studies is, however, conducted in elderly humans. There is ample research done on evaluation of the clinical effectiveness of various therapies for reduction of OA symptoms. Among other alternatives, the oral supplement with GS and/or CS is one of the most popular, well-known and widely used options in humans (Clegg et al., 2006). Similar to the equine field, in the clinical guidelines published by ACR, EULAR and OARSI (Hochberg et al., 2012; Pendleton et al., 2000; McAlindon et al., 2014) and the review published by Cochrane (Towheed et al., 2005) on treating OA, the efficacy of these two agents was concluded as ‘uncertain’, since significant heterogeneity in effect sizes was found among studies.

In human medicine, there has been an increased interest in statistical methods that incorporate existing evidence into a Bayesian analysis (e.g., Spiegelhalter et al., 2004; Rietbergen et al., 2011; Hobbs et al., 2012; Viele et al., 2014). Existing evidence is traditionally used to help in the design of an experiment or when pooling data in a meta-analysis (Spiegelhalter et al., 2004). With Bayesian inference, however, we can aggregate the existing

external data into an informative prior for the analysis of current data, by assuming that their underlying probability models (i.e., the likelihoods) share parameters of interest (Spiegelhalter et al., 2004). The final ‘updated’ state of knowledge is thus based on the external and current evidence combined in posterior estimates.

However, external evidence may often not be directly related to the current research question, and we may consider to discount its influence. Ibrahim and Chen (2000) therefore suggest the power prior approach, in which the parameter of interest is assumed to be the same for the external and the current study, but impact of each external study will be downweighted with regard to its relevance to the current research question.

In veterinary medicine, the power prior approach can be a valuable method, especially for species such as horses, in which studies with small sample sizes are common. Clinical conclusions must be based on relatively scarce information. Veterinarians are hence by default trained to be cross-species thinkers. In clinical practice, borrowing information from other species may happen on a daily base in an implicit manner. The power prior method offers the opportunity to utilize multiple relevant cross-species sources in a more transparent and explicit way.

In the present paper, we are interested in the clinical research question from the small equine trial conducted by Higler et al. (2014). In their trial, the authors evaluated the clinical efficacy of oral supplement with GS, CS and methylsulfonylmethane (MSM) on locomotion in aged horses. In our study, we focused on the effect of GS and CS. An equine expert suggested that it is clinically applicable to borrow cross-species information on this research question. The objective of our study is to re-assess the clinical effectiveness of GS and CS on stiff gaits in elderly horses, by aggregating cross-species quantitative information as prior into the analysis of the Higler et al. trial (2014). Throughout the paper, this equine trial is recognized as ‘the current trial’.

6.2. Materials and Methods

6.2.1. Literature search and study selection

A systematic search was conducted in order to capture relevant studies from species horse, dog and human. The PRISMA checklist is provided in Appendix A. PICO components were identified with respect to the current trial (see Appendix B). Search terms were specified based on the PICO keywords. Four electronic databases were searched through either in ‘Topic’ (Web of Science) or in ‘All Fields’ (PubMed, CAB Abstracts) or in ‘Title, Abstract and Keywords’

(Scopus). No language and document type restrictions were specified, but the timespan was restricted between 1980 and 2018 following from a systematic review in horses (Pearson & Lindinger, 2009) and a meta-analysis for humans (Wandel et al., 2010). Within the electronic databases Web of Science, PubMed and Scopus, four searches were performed: a joint search without specifying search terms for species, and three separate searches specifying the species as horse, dog or human. Within the database CAB Abstracts, three searches were done: a joint search and two separate searches specifying the species as horse or dog. The exact search terms can be found in Appendix B. Results of the systematic search were exported to RefWorks for further management.

Study selection was carried out in the following way. After removing duplicates, titles and abstracts were screened by the first author with the inclusion criteria (1) written in English; (2) intervention containing oral supplement with GS and/or CS; (3) species horse/dog/human; (4) only primary research. Biochemical studies, designs of new studies, health news, book chapters, conference proceedings, guidelines and tutorial articles were excluded. The current trial (Higler et al., 2014) was excluded as well. The remaining studies were next screened by an equine specialist (co-author: WB) again through titles and abstracts. Studies that investigated the clinical efficacy, and that were clinically relevant to the current research question were kept by the equine specialist. Full-text assessment for eligibility was subsequently done by the first author. Studies were excluded if they contained (1) synthesized evidence, i.e., systematic review and meta-analysis; (2) no data; (3) no placebo controlled arm; (4) no reported mean and variance for treatment and control arms.

6.2.2. Data extraction

Information extraction included general article characteristics such as year of publication and name of the first author. Data were extracted on study species, study design, nutraceuticals used for the treatment arm, duration of the intervention, and the outcome variable with its respective scale. Relevant extracted statistical properties for the power prior specification were (1) the total sample size and the sample size of each arm that was of interest (i.e., a treatment arm that used GS and/or CS and a placebo controlled arm) and (2) the pre- to post-treatment mean change in each arm and its standard deviation. For studies that only reported pre-post change in the form of median with its range, we approximated the mean and the standard deviation (Hozo et al., 2005). For studies that did not report the pre-post change in a numeric form, values were approximated based on relevant figures. Raw individual data points were available from the current trial but from none of the external studies.

6.2.3. The power prior

When results are available from similar existing studies, it is reasonable to use them as the basis for the prior distribution of the Bayesian analysis of a new study. However, often these studies are not exactly similar to the clinical research question at hand. The power prior approach (Ibrahim & Chen, 2000) proposes to discount the impact of the external studies by taking the likelihood of the data to a power a . This method assumes that the external studies and the current study share the same parameters of interest θ , but the informativeness of an external dataset D_e is weighted by the power a relative to the current data D . Let n_E denote the effective sample size of the external dataset, computed as $n_E = n_e \times a$ where n_e refers to the original sample size of the external dataset and a refers to the weight at hand.

The power prior for a single external study is expressed as:

$$\pi(\theta|D_e, a) \propto L(\theta|D_e)^a \times \pi_0(\theta), \quad (6.1)$$

where $\pi_0(\theta)$ is the initial prior which is often chosen to be low-informative, and $L(\theta|D_e)^a$ is the powered likelihood of the external data. The resulting posterior $\pi(\theta|D_e, a)$ can then be used as the prior distribution, that is, the power prior, for the Bayesian analysis of the current data D . The posterior distribution of θ is:

$$\pi(\theta|D, D_e, a) \propto L(\theta|D) \times \pi(\theta|D_e, a). \quad (6.2)$$

Posterior estimates of θ can be obtained through sampling by Markov chain Monte Carlo (MCMC) methods. More details about MCMC methods can be found in, for instance, Spiegelhalter et al. (2004) and Gelman et al., (2008).

The power a is also known as the weight parameter, as it weights the external study with regard to the data from the current study. Ibrahim and Chen (2000) suggested to restrict the weight parameter to be between 0 and 1, with $a = 0$ equivalent to no incorporation of external evidence at all, and $a = 1$ equivalent to full inclusion of external evidence. In (6.1), the weight parameter is specified as a fixed value, leading to what is called a conditional power prior. To express uncertainty about the weight parameter, a prior for a can be specified, leading to a so-called hierarchical or joint power prior approach (e.g., Chen & Ibrahim, 2006). In the present paper, we applied the power prior approach with multiple external studies. Since the joint power prior approach is not yet fully developed for multiple studies (Hobbs et al., 2012; Chen & Ibrahim, 2006; Neuenschwander et al., 2009), we will apply the conditional power prior approach, that is, with study-specific fixed weights.

For multiple external studies, let $\mathbf{D}_e = (D_{e1}, \dots, D_{eK})$ where D_{ek} is the data from the k th study for $k = 1, \dots, K$. Further, let $\mathbf{a} = (a_1, \dots, a_K)$, where a_k is the study-specific weight for the data of the k th study D_{ek} . The effective sample size n_E is additive and the total effective sample size of K external studies is thus $\sum_{k=1}^K n_{Ek}$ and the power prior for multiple external studies is defined as (Ibrahim & Chen, 2000):

$$\pi(\theta|\mathbf{D}_e, \mathbf{a}) \propto \prod_{k=1}^K [L(\theta|D_{ek})^{a_k}] \times \pi_0(\theta). \quad (6.3)$$

6.2.4. The statistical model

A Bayesian analysis was performed on the current trial for assessing the effect of oral supplementation of GS and CS on stiffness in old horses. This trial was well-powered on the basis of the study conducted by Forsyth et al. (2006) where clinical treatment effect was found. In the current trial, a total of 24 horses and ponies with a mean age of 29 years were involved over a 3-month treatment period, during which 12 were randomized into the treatment arm and 12 into the placebo control arm. Change in stride length from pre- to post-treatment was the primary outcome variable. Assessment of treatment effect (denoted T) was obtained by comparing the mean difference of the primary outcome variable (i.e., change in stride length) between the treatment and the control arm.

As different studies used different outcome variables to evaluate the clinical treatment effect, data from all studies needed to be standardized before being entered in the analysis. For the current trial, available primary data were used. For the external studies only summary sufficient statistics were available instead of primary data. To be able to standardize the results of the external studies, individual data points were simulated such that they summarized into the exact sufficient statistics as reported.

The standardized data within each study, the current as well as the external, were assumed to be normally distributed in both the treatment and the control arm:

$$\begin{aligned} y_{ti} &\sim N(\mu_t, \sigma_t^2) & \text{for } i = 1, \dots, n_t, \\ y_{cj} &\sim N(\mu_c, \sigma_c^2) & \text{for } j = 1, \dots, n_c. \end{aligned} \quad (6.4)$$

Here y_{ti} and y_{cj} refer to the standardized change in the outcome variable before and after treatment of the i th subject of the treatment and the j th subject of the control arm respectively; and n_t and n_c represent the sample sizes of the treatment and the control arm respectively.

The parameter of main interest is the standardized treatment effect τ expressed as the mean difference between the two intervention arms, that is, $\tau = \mu_t - \mu_c$.

6.2.5. Weight elicitation

The equine specialist who screened the systematic search results was asked to assign and motivate a study-specific weight for each external study, regarding its clinical relevance to the research question of the current trial. He was provided with the studies in full-text along with an instruction and a form for the weight elicitation (Appendix C).

The visual analogue scale (VAS) was employed for the weight elicitation (Crichton, 2001). For each external study, a VAS was used representing the degree of relevance, with 0 indicating 0% and 100 indicating 100%. The equine expert was asked to pinpoint a value from the visual scale that expressed the relevance of an external study to the current research question. For K external studies, the VAS value for the k th study is therefore the weight a_k . When incorporating multiple studies, the weight \mathbf{a} is a vector of weights assigned to all K studies.

The equine expert was asked to explicitly ignore the results found in each study, such as the effect size or the significance value. Sample size of each study or of each intervention arm should be ignored as well since sample size of an external study is included in the specification of the power prior through the likelihood function for the data. In addition, he was explicitly asked to consider the relative importance of each external study for the current research question when assigning a weight, in relation to the other external studies. The expert was further free to decide which criteria to use as motivation.

6.2.6. Sampling of the posterior

The discounted external evidence contained in the power prior is updated in the light of new evidence from the current trial, providing posterior estimates for the model parameters. Posterior estimates for the standardized treatment effect τ can easily be obtained with the MCMC sampling approach. In each iteration, a value for μ_t as well as μ_c is sampled and therefore also a value for $\tau = \mu_t - \mu_c$. With the resulting sample from the posterior of τ , all estimates of interest are available. Statistical inference on the clinical treatment effect was made on the basis of the posterior mean and variance of τ , as well as the 95% central credible interval (CCI).

The power prior was calculated analytically using the program R, while the Bayesian analysis of the current trial combined with the resulting power prior was carried out in OpenBUGS version 3.2.2 (model code is provided in Appendix D). The first 10,000 iterations were discarded as the burn-in phase, and the following 10,000 iterations were used for posterior

inference. Convergence of the MCMC chains was monitored by visual inspection of the trace plots of the parameters of interest.

6.3. Results

6.3.1. Systematic search

There were 2,628 documents identified by the systematic search through the four electronic databases. After removing the duplicates automatically by RefWorks, 998 were left for the titles and abstracts screening. The first author removed 755 documents based on the inclusion/exclusion criteria. The equine specialist subsequently reviewed the remaining studies and identified 84 articles as clinically relevant to the current research question. Assessment for eligibility was done by the first author and 64 articles were considered ineligible due to the lack of quantitative information required for the power prior specification. The remaining 20 studies (see Appendix E) were sent to the equine expert for weight elicitation. An information flow diagram can be seen in Fig. 6.1.

6.3.2. Data

Characteristics and findings of the 20 external studies as well as the current trial are displayed in Table 6.1. The year of publication ranged from 1994 to 2017, with 18 out of 20 published after 2000. The number of studies identified from different species varied with 3 studies conducted in horses, 2 in dogs, and the rest in humans.

Although no criteria were specified about the study design for the systematic search, all 20 studies were trials. This may be due to the requirement of including the pre-post treatment mean changes from two arms, including a placebo controlled arm. Further, nutraceuticals in the treated participants varied between studies: some studies used only GS or only CS, others used the combination (i.e., GS+CS). The treatment effect was evaluated by various outcomes with their own scales, however, most human studies used either AFI or WOMAC (pain) scores. Note that in Table 6.1, for outcomes such as stride length, clinical improvement corresponds to a positive T value, while for outcomes such as AFI or WOMAC scores, clinical improvement corresponds to a negative T value since an effective treatment would reduce, for example, pain intensity and lameness. Before inclusion in the power prior approach, all scores were recoded such that clinical improvement always corresponds to a positive mean difference between the treatment and the control arm.

The standardized statistics used for specification of the power prior are listed in Table 6.2. The last column contains the weights elicited by the equine expert. The equine expert did not

use the VAS elicitation method explicitly, instead he was inspired by this method and weighted each external study by using a 100-point scale. For studies from species other than horses, 20 points were subtracted. Further, for each aspect that he considered not relevant to answering the current research question, 10 points were subtracted. As can be seen from the last column, the weights were not equally distributed among the external studies, reflecting variable degrees of informativeness of these studies on the current trial.

Table 6.1. Relevant study characteristics of the current trial and the 20 external studies. The following information is provided: year published (Year), first author (Author), sample size (Size), animal species (Species), nutraceuticals used for treatment (Nutraceuticals), Outcome measure (Outcome), reported raw effect size T and its scale (T (scale)), 95% confidence interval of the reported effect size (95% CI).

Year	Author	Size	Species	Nutraceuticals	Outcome	T (scale)	95% CI
2014	Higler	24	Horse	GS+CS	stride length	0.5 cm	[-1.3, 2.3]
1994	Noack	252	Human	GS	AFI score ^a	-1[0-26]	- ^d
1998	Bourgeois	127*	Human	CS	AFI score	-3[0-26]	-
2001	Mazieres	130	Human	CS	AFI score	-0.8[0-26]	-
2001	Hanson	14	Horse	GS+CS	AFI lameness	-17.2[0-58]	-
2002	Pavelka	202	Human	GS	AFI score	-0.91[0-26]	[-0.34, -1.5]
2004	McAlindon	205	Human	GS	WOMAC score ^b	0.6[0-96]	[-4.0, 5.2]
2004	Usha	118*	Human	GS	AFI score	-4.14[0-26]	-
2006	Forsyth	20	Horse	GS+CS	stride length	12.5 cm	-
2007	Mazieres	307	Human	CS	AFI score	-0.7[0-24]	-
2007	D'Altilio	20	Dog	GS+CS	overall pain	-0.4[0-10]	-
2008	Frestedt	70*	Human	GS	WOMAC pain	-9.7[0-100]	-
2009	Rozendaal	222	Human	GS	WOMAC pain	-1.5[0-100]	[-5.4, 2.4]
2009	Gupta	25*	Horse	GS+CS	overall pain	-2.8[0-10]	-
2010	Petersen	36*	Human	GS	knee strength	-5 kg	-
2013	Nieman	100*	Human	GS	WOMAC score	-5[0-2400]	-
2015	Tsuji	50	Human	GS+CS	VAS pain score	7.5[0-100]	-
2016	Lugo	191*	Human	GS+CS	WOMAC score	-40[0-2400]	-
2016	Sterzi	53	Human	GS+CS	AFI score	-0.9[0-26]	-
2017	Roman-Blas	158	Human	GS+CS	VAS pain score	8.3[0-100]	-
2017	Scott	60	Dog	GS+CS	CBPI pain score ^c	0.2[0-25]	-

GS = Glucosamine; CS= Chondroitin Sulphate.

^a Lequesne's algofunctional Index; ^b Western Ontario and McMaster Universities Arthritis Index.

^c Canine Brief Pain Inventory questionnaire ; ^d When a certain statistic is not reported, it is denoted as "-";

* More than two intervention arms.

Table 6.2. Relevant statistical properties of the current trial and the 20 external studies. The following information is provided: the sample size (n), the sample mean (\bar{y}) and standard deviation (s) after standardization for each intervention arm (t = treatment, c = control), the weight a_k for clinical relevance.

Author	Outcome	Species	n_t	n_c	\bar{y}_t	\bar{y}_c	s_t	s_c	a_k
Higler (current)	stride length	horse	12	12	0.029	-0.029	0.980	1.062	1
Noack	AFI score	human	120	121	0.641	-0.635	0.817	0.722	0.60
Bourgeois	AFI score	human	43	44	0.313	-0.305	0.874	1.030	0.40
Mazieres	AFI score	human	63	67	0.132	-0.125	0.996	0.996	0.40
Hanson	AFI lameness	horse	8	6	0.645	-0.860	0.778	0.448	0.60
Pavelka	AFI score	human	101	101	0.899	-0.899	0.521	0.323	0.10
McAlindon	WOMAC score	human	101	104	0.000	0.000	0.987	1.017	0.40
Usha	AFI score	human	30	28	0.646	-0.692	1.035	0.003	0.30
Forsyth	stride length	horse	15	5	0.511	-1.534	0.409	0.491	0.80
Mazieres	AFI score	human	153	154	0.104	-0.104	1.011	0.981	0.70
D’Altilio	Overall pain	dog	5	5	0.541	-0.541	0.255	1.206	0.50
Frestedt	WOMAC pain	human	19	16	0.241	-0.286	0.885	1.080	0.60
Rozendaal	Pain score	human	111	111	0.448	-0.448	0.896	0.896	0.70
Gupta	Overall pain	horse	5	5	0.854	-0.854	0.390	0.524	0.60
Petersen	knee strength	human	12	12	-0.183	0.183	1.163	0.816	0.60
Nieman	WOMAC score	human	36	36	0.733	-0.733	0.756	0.593	0.60
Tsuji	VAS pain score	human	25	25	-0.478	0.478	0.883	0.887	0.40
Lugo	WOMAC score	human	57	53	0.561	-0.604	0.801	0.830	0.40
Sterzi	AFI score	human	23	27	0.102	-0.087	1.028	0.987	0.40
Roman-Blas	VAS pain score	human	80	78	-0.863	0.885	0.482	0.482	0.20
Scott	CBPI pain score	dog	30	30	-0.000	0.000	1.187	0.791	0.30

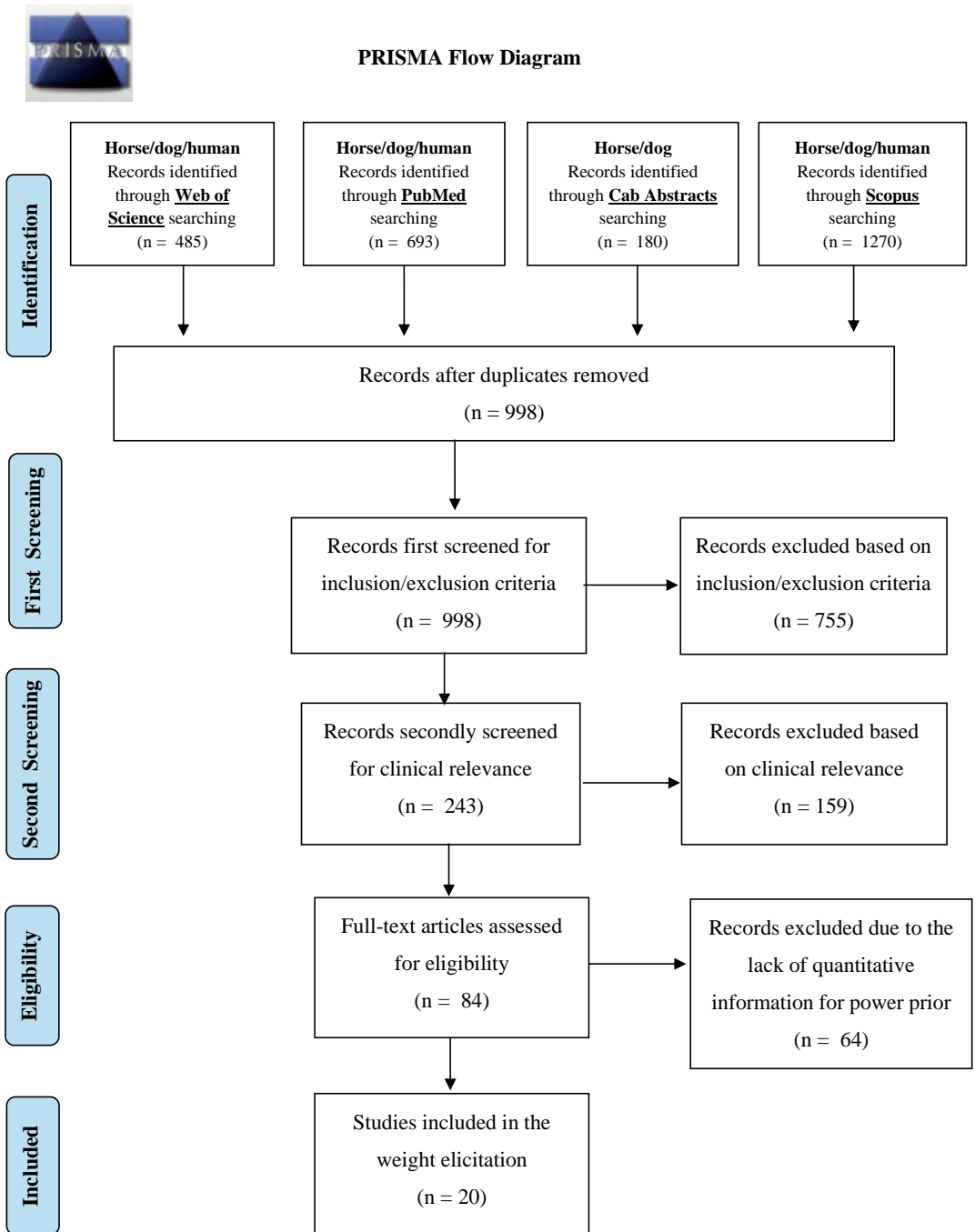


Fig. 6.1. An information flow diagram for the systematic searches with regard to the clinical question of Higler et al. (2014) trial.

6.3.3. Posterior inference

Proper convergence of the MCMC chains was observed for all chains, allowing assessment of the posterior estimates.

Table 6.3 reveals the posterior estimates of the standardized treatment effect conditional on (1) using a low-informative prior (with no external evidence included), (2) using the power prior with 9 weighted GS+CS studies included.

Table 6.3. Posterior estimates for the standardized treatment effect.

	Standardized treatment effect τ			
	Posterior estimates			
	n_E total	Mean	Variance	95% CCI
Current trial + low informative prior	0.2	0.056	0.189	[-0.799, 0.911]
Current trial + power prior with GS+CS studies	169.2	0.179	0.020	[-0.105, 0.464]

The posterior standardized results suggested that there was no standardized treatment effect when 9 studies that investigated GS+CS treatment were incorporated as prior information into the analysis. A graphical illustration is provided in Fig. 6.2, where the power prior based on the 9 GS+CS treatment studies is plotted with the likelihood of the current data and the resulting posterior.

It is clear in Fig. 6.2, that the power prior using 9 external studies (n_E total = 169.2) has a much larger impact on the posterior distribution of the standardized treatment effect, in comparison to the likelihood of the current data ($n = 24$). To further investigate the impact of external information on the posterior estimates and subsequent conclusions, we performed sensitivity analyses based on different specifications of the power prior. In the first sensitivity analysis, we downweighted all 9 GS+CS studies such that the total effective sample size (i.e., n_E total) of the power prior was approximately equivalent to the sample size of the current trial, while keeping the relative importance of the studies unchanged. In other words, the external evidence (with n_E total ≈ 24) contributed the same amount of information to the posterior distribution as the current trial. The second sensitivity analysis was done with the 3 equine studies using the original elicited weights. Further, as one might argue that actual physical measurements could differ from self-reported scores, in the third sensitivity analysis, we included studies that used physical outcome measures (Forsyth et al., 2006; Petersen et al., 2010) instead of questionnaire scores in, e.g., pain. In the last two sensitivity analyses, we included 8 studies that used only GS and 3 studies that used only CS nutraceuticals as treatment

respectively.

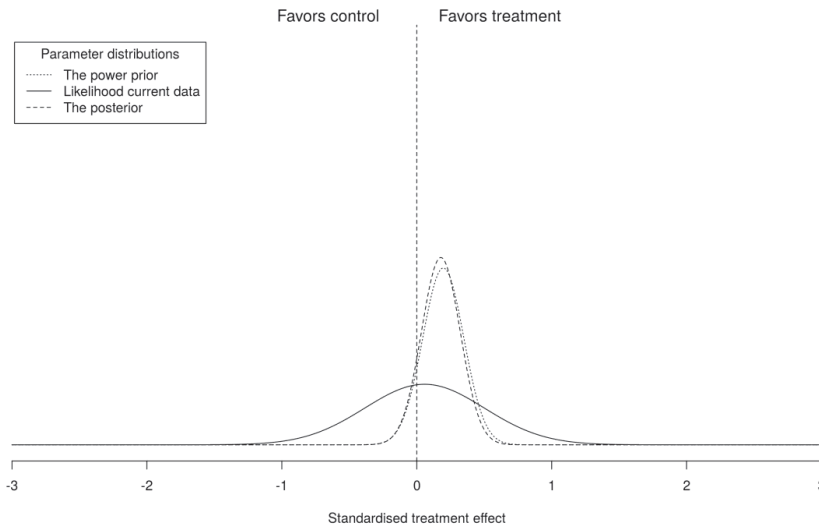


Fig. 6.2. The power prior with total effective sample size of the 9 GS+CS studies (n_E total = 169.2) using original assigned weights by the equine expert, the likelihood of the current data ($n = 24$) and the resulting posterior.

Table 6.4 lists the posterior estimates from the 5 sensitivity analyses. We notice that these analyses yield different results: the 95% posterior CIs of the first and the third sensitivity analysis contain zero, indicating no effect; while the 95% posterior CIs of the other three sensitivity analyses are larger than zero, indicating a positive standardized effect. The first sensitivity analysis with all 9 GS+CS studies downweighted to obtain a total effective sample size of 24, shows no effect. This result supports the results from the main analysis (presented in Table 6.3 and Fig. 6.2) where these 9 studies were included in the prior with their original weights (i.e., n_E total = 169.2). By downweighting the relative importance of the external studies, the effective sample size of the power prior becomes smaller and the posterior variance and 95% CCI become larger, representing more uncertainty about the effect size.

It is further worth noting that the third sensitivity analysis including only studies with physical outcome measures (studies from Forsyth and Petersen), shows no treatment effect as well. However, the sensitivity analyses including the equine studies, only GS treatment and only CS treatment studies, all show a positive effect. The impact of different specifications of the power prior is also displayed in Fig. 6.3.

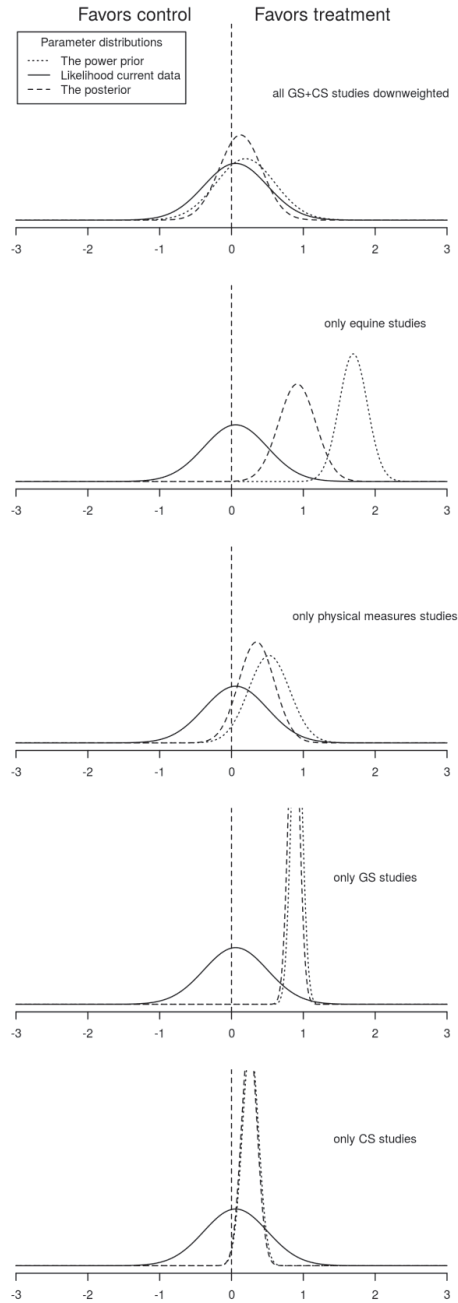


Fig. 6.3. Sensitivity analyses on different power priors. The power prior included (1) all 9 GS+CS studies equally downweighted such that n_E total ≈ 24 ; (2) the 3 equine studies; (3) the 2 studies with physical measure outcomes; (4) the 8 studies that used only GS as treatment; (5) the 3 studies that used only CS as treatment.

Table 6.4. Results of the sensitivity analyses based on 5 different power priors.

The power prior Specification	Standardized treatment effect τ Posterior estimates			
	n_E total	Mean	Variance	95% CCI
All GS+CS studies downweighted	24.2	0.125	0.087	[-0.450, 0.703]
Only equine studies	30.6	0.915	0.066	[0.412, 1.425]
Only physical measures studies	44.2	0.349	0.062	[-0.135, 0.846]
Only GS studies	498.4	0.868	0.006	[0.715, 1.023]
Only CS studies	301.9	0.248	0.012	[0.031, 0.464]

6.4. Discussion

In the present study, clinical trial data on treatment effect of oral supplement with GS and CS on stiffness in aged horses (Higler et al., 2014) were re-analyzed with the inclusion of relevant cross-species existing evidence using the Bayesian power prior method. When incorporating 9 weighted external studies into the current trial analysis, the posterior results showed no difference between the treatment and control arm. This supports the original conclusion of the current trial where no treatment effect was observed. However, sensitivity analyses performed afterwards with differently specified power priors gave heterogenous results: three analyses showed a positive standardized treatment effect while in the other two there was no effect. These mixed results, nevertheless, reflect the current state of evidence on this clinical intervention. In the literature, substantial heterogeneity of findings across studies is seen both in humans and horses. This might be due to differences in study quality, duration of the intervention, preparation of the supplements, industry involvement, publication bias, selection bias and various study sample sizes (Towheed et al., 2005; Pearson & Lindinger, 2009; Singh et al., 2015).

In our analysis, the posterior results are affected both by the current trial data and the power prior specification. Since the current trial automatically gets the weight one, the impact of the current trial to the posterior results will be bigger than the impact of a downweighted external study with an equal original sample size. Authors of the current trial have based their power calculation on the equine study by Forsyth et al. (2006) in which a large clinical effect was detected. If there is a true existing but small treatment effect, research with a small dataset might fail to detect it.

When it comes to the included external studies, there are several aspects that have influenced the results.

Firstly, the studies resulting from the systematic searches were dependent on the search terms. Different search terms would probably result in different sets of studies. For example, a large randomized and placebo-controlled Glucosamine/chondroitin Arthritis Intervention Trial (GAIT) conducted in humans was not detected by our search terms, while their summarizing report which contained no data was included in the search. In the GAIT trial, nearly 1,600 human participants with documented OA of the knee were enrolled, and results showed no significant treatment effect (Clegg et al., 2006).

Secondly, the quality of the studies was not explicitly controlled. No methodological screening was performed before the weight elicitation. Among the 20 trials, there might be some that had low study quality. Additionally, publication bias of positive effect sizes might have played a role in this field (i.e., nutraceuticals) as well. In veterinary medicine, clinical trials are not common and not obligatory to conduct. As a consequence, published results may be overly optimistic and not representative of the evidence.

Thirdly, the use of cross-species data in an analysis might yield challenges in practice. In our case, studies from different species were first screened by the equine expert, such that the pathophysiology, the dose, the duration and the treatment of the selected studies would be relevant to the current research question. Nevertheless, it was difficult to define whether different outcome measures were clinically equivalent. Stride length from the current trial and, for example, pain scores from several human studies were treated as measurement of the same underlying dysfunction(s) and were standardized into a theoretical single outcome parameter. This resulted in standardized posterior estimates for the treatment effect. For clinical interpretation, ideally, the standardized effect estimates should be translated back to the scale of the original outcome of the current study, in order to evaluate the actual clinical effect. It is however unclear how we should convert such standardized effect back to a clinical quantity. Further research may look into the back transformation of a standardized effect size to a clinical one. In our application, the standardization was useful and unavoidable. Only if all incorporated external studies use the same outcome measure as the current study, standardization would not be needed. However, if the studies are from different species, even with the same outcome measure, such as stride/step length, standardization is still necessary, because a change of two centimeter is clinically not the same between dogs and horses.

In our study, due to the required standardization, we needed to model the treatment and control arm separately. As a consequence, we also needed the means and standard deviations

of pre-post treatment change for both arms. The requirement of information from both arms may have restricted the amount of studies that could be used. In our application, however, among the 64 studies that were filtered out during the full-text eligibility assessment, only one study was excluded as a consequence of needing specific information from both arms. The difference between modelling on both arms other than directly on the treatment effect was therefore minimal.

A crucial but challenging part of the power prior approach is the specification of the weight parameter. The weight, or power, of the power prior controls the impact of the external evidence on the current data analysis and might have influential effect on the conclusions. There are a few aspects regarding the weight parameter that need to be highlighted.

The VAS elicitation method was provided to the equine expert, but the expert found it difficult to assign weights using a visual scale. He was instead inspired by the scaling, and used maximum 100 points for each study and subtracted points for aspects of the study that made it less relevant for the current research question. This might indicate that, in comparison to pinpointing a value on a visual scale, subtracting points from a pre-defined scale is practically easier and thus more useful when weighting a study.

Further, we only asked one clinical expert to assign clinical weights. If multiple experts provide weights, alternative elicitation methods might be explored, such as methods where consensus is met between experts (Rietbergen et al., 2016) or elicitation methods where experts also express their uncertainty to each assigned weight (Higgins et al., 2014).

For two studies with equal weight, the largest study will have a bigger effective sample size, hence contribute more information to the posterior distribution. Special attention may be required when large observational studies are included. Even with moderate or small weights, the effective sample size might be unreasonably large. A solution could be to pre-define a maximum value for the effective sample size, for instance, such that it never exceeds the sample size of the current study. On the other hand, if huge human trials are deemed clinically very relevant to the current veterinary research question, it might be unwise to severely downweight such evidence.

In the present study, we adopted the standard conditional power prior approach. The current study was therefore not downweighted, or stated differently, received weight one. However, the current study could also be methodologically suboptimal. In such circumstances, one might want to downweight the current study as well. Furthermore, we asked the expert to give a single weight for each relevant study, while in fact information from the control arm could differ in relevance from the treatment arm. It might then be appealing to pool, for instance, more data

from the control arm than from the treatment arm into the power prior. In fact, aggregating external evidence only from the control arm is more often applied in human medicine than aggregating from both control and treatment arms (Pocock, 1976; Gould, 1991; Ryan, 1993). Investigation on the use of unequal weights for each intervention arm seems worthwhile.

We have applied the power prior method to a simple statistical model for the primary outcome, and used it retrospectively. That is to say, the data of the current trial was collected before our application. Further studies might expand the use of this method to more complex models, and explore the prospective use of the power prior approach to design and power a veterinary clinical trial. Using Monte Carlo methods, values of the parameters for the predictive distribution can be simulated from the power prior (Spiegelhalter et al., 2004). From the predictive values of the parameters, the necessary sample size to obtain the aspired power (e.g., 80%) can be calculated.

The power prior can be considered an evidence-based prior based on available external sources while taking the expert knowledge about relevance of these sources into account. In veterinary medicine, due to the scarcity of clinical trials and the variety of species, the pragmatic use of pooling relevant evidence into the prior seems even more beneficial than it is for human medicine. Particularly the potential use of inter-species data could be promising for 'minor species' therapies.

6.5. Conclusions

This study attempts to grasp the best picture of the current knowledge on the clinical intervention of oral supplement with GS and CS on stiffness in aged horses. The posterior estimates resulting from the existing relevant cross-species evidence and the current trial (Higler et al., 2014) indicated no standardized treatment effect which supported the conclusion from the original equine trial. However, the sensitivity analyses showed mixed results. This heterogeneity in results is in agreement with the current uncertainty of effect size on this clinical intervention.

To our knowledge, this is the first study that applied the power prior method in a veterinary context. It is also the first attempt to systematically and transparently use cross-species information for a clinical trial analysis.

6.6. Acknowledgements

We gratefully thank Dr. Roos Goverde (Utrecht University Library) for her help on the systematic search.

Appendix A: PRISMA 2009 checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	N/A
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	N/A
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	N/A
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	6, 7
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	6, 7
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Appendix B
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	6, 7
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	7
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	7

Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	7
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	7
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	9
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	N/A
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	N/A
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	11, 12
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	12, 13
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	13, 14
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	13, 14
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency. (Bayesian power prior analysis instead of meta-analysis in this study, and credible intervals instead of confidence intervals)	15, 16
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see item 15).	N/A
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see item 16]). (Sensitivity analyses using the Bayesian power prior method)	17, 18
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	18
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	19, 20
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	22
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	N/A

Appendix B: Systematic literature search PICO components and search terms

The systematic search was carried out in four electronic databases, specifically Web of Science, PubMed, Scopus and CAB Abstracts. The search terms were selected based on the PICO components extracted from the current trial (see Table B1). The exact search terms are shown below for the database Web of Science. For the other three databases, the search terms varied slightly and are available from the first author. systematic search was carried out in four electronic databases, specifically Web of Science, PubMed, Scopus and CAB Abstracts. The search terms were selected based on the PICO components extracted from the current trial (see Table B1). The exact search terms are shown below for the database Web of Science. For the other three databases, the search terms varied slightly and are available from the first author.

Table B1. PICO elements based on the current trial (Higler et al., 2014).

Question	PICO elements				
	Animal	Patients	Intervention	Comparison	Outcome
Central question	Horse(s), equine	geriatric, old, older, aged, stiff(ness), lame(ness), inflammation	oral, supplement(s), oral dosing, oral supplementation, oral administration, oral treatment, dietary product, nutraceutical, glucosamine, chondroitin	control, controlled, placebo	joint pain, pain intensity, change in width of joint space, joint space width, clinically relevant effect, change from baseline
..	Dog(s) Canine	joint pain, clinical scoring, joint mobility, weight bearing, clinically relevant effect, change from baseline
..	Man, woman, human	joint pain, pain intensity, change in width of joint space, joint space width, clinically relevant effect, change from baseline

Search terms for Web of Science**(1) The joint search without specifying the species**

Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI, Timespan=1980-2018

(TS=(stiff) OR TS=(stiffness) OR TS=(lame) OR TS=(lameness) OR TS=(inflammatory) OR TS=(inflammation))

AND

(TS=(oral) OR TS=(supplement) OR TS=(oral supplementation) OR TS=(oral dosing) OR TS=(oral administration) OR TS=(oral treatment) OR TS=(dietary product) OR TS=(nutraceutical))

AND

(TS=(glucosamine) OR TS=(chondroitin))

(2) Search in the species horse

Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI, Timespan=1980-2018

(TS=(oral) OR TS=(supplement) OR TS=(oral supplementation) OR TS=(oral dosing) OR TS=(oral administration) OR TS=(oral treatment) OR TS=(dietary product) OR TS=(nutraceutical))

AND

(TS=(glucosamine) OR TS=(chondroitin))

AND

(TS=(horse) OR TS=(equine) OR TS=(equidae))

(3) Search in the species dog

Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI, Timespan=1980-2018

(TS=(oral) OR TS=(supplement) OR TS=(oral supplementation) OR TS=(oral dosing) OR TS=(oral administration) OR TS=(oral treatment) OR TS=(dietary product) OR TS=(nutraceutical))

AND

(TS=(glucosamine) OR TS=(chondroitin))

AND

(TS=(dog) OR TS=(canine) OR TS=(canidae))

(4) Search in the species human

Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI, Timespan=1980-2018

(TS=(stiff) OR TS=(stiffness) OR TS=(lame) OR TS=(lameness) OR TS=(inflammatory) OR TS=(inflammation))

AND

(TS=(oral) OR TS=(supplement) OR TS=(oral supplementation) OR TS=(oral dosing) OR TS=(oral administration) OR TS=(oral treatment) OR TS=(dietary product) OR TS=(nutraceutical))

AND

(TS=(glucosamine) OR TS=(chondroitin))

AND

(TS=(man) OR TS=(woman) OR TS=(human))

Appendix C: Weight elicitation instruction and form

Background information for the power prior weight elicitation

Brief Summary of The Project This project is a collaboration between the department Methodology and Statistics and the department Farm Animal Health at Utrecht University. The main objective is to support clinical veterinary decision making by investigating how prior information can be incorporated into the analysis of a clinical trial using Bayesian modelling. Two necessary ingredients are a current study and one or more external studies to provide prior information.

In this project, the current study of interest is the clinical trial from Higler et al. (2014). Their research goal was to assess the clinical efficacy of oral supplementation with glucosamine (GS) and chondroitin sulphate (CS) on the improvement of stiff gait in old horses. Twenty published studies with various degrees of relevance to the research question of the current trial will be incorporated as prior information.

The degree of relevance of each selected study to the current research question is expressed by a weight. In other words, the weight represents how much information a study should contribute to the prior for the analysis of the current study.

The Elicitation Method As an experienced equine specialist, you are asked to weight the *clinical relevance* of each study with respect to the research question of the current trial.

It is important for you to bear in mind that results found from each study (e.g., effect sizes, significance values) should be ignored. That is to say, study results should not influence the values of the weights. Furthermore, the sample size of each study or each intervention arm should be ignored as well, because it is automatically included in the analysis.

The visual analogue scale (VAS) method is used for eliciting the weights. It is a 100mm horizontal line treated as a continuous scale with two end points: on the left hand 0 and on the right hand 100. You may interpret them as 0% (not relevant) and 100% (completely relevant) respectively. For each study, you are asked to pinpoint a percentage that represents the degree of relevance this study shows to the current research question.

The more a study is related to the research question of the current trial, the higher percentage it should get, and vice versa. As a result, the current trial (Higler) will automatically get 100%, indicating a weight valued 1. In the case that you consider one study as not qualified for

inclusion, a 0 can be assigned, which means this study will be excluded from the subsequent analyses.

As it is rather difficult to summarize your opinion into one number, please use the pencil and eraser provided. You may change a weight until you feel it best reflects your true opinion.

Furthermore, as you will notice in the elicitation form, all 20 studies are presented on one page, you can therefore have a clear overview of the whole body of evidence. You are kindly asked to take the relative importance of a study regarding other 19 studies into account while assigning a weight.

Other Related Issues It is possible to discuss the weights with your colleagues as long as you remain responsible for the total elicitation process of all 20 studies (i.e., determine the final weights). You are also kindly asked to provide a brief motivation for each elicited weight.

Form for the power prior weight elicitation

Please assign a weight with a short motivation for each study after assessing the full-text article. Note that: (1) the weight should represent the clinical relevance for the current research question; (2) study results (e.g., effect sizes or significance values) should *not* be taken into account; (3) sample sizes should *not* be taken into account.

Horse

Hanson et al. (2001)	0 ————— 100
Forsyth et al. (2006)	0 ————— 100
Gupta et al. (2009)	0 ————— 100

Dog

D'Altilio (2007)	0 ————— 100
Scott et al. (2017)	0 ————— 100

Human

Noack et al. (1994)	0 ————— 100
Bourgeois et al. (1998)	0 ————— 100
Mazieres et al. (2001)	0 ————— 100
Pavelka et al. (2002)	0 ————— 100
McAlindon et al. (2004)	0 ————— 100
Usha et al. (2004)	0 ————— 100
Mazieres et al. (2007)	0 ————— 100
Frestedt et al. (2008)	0 ————— 100
Rozendaal et al. (2009)	0 ————— 100
Petersen et al. (2010)	0 ————— 100
Nieman et al. (2013)	0 ————— 100
Tsuji et al. (2015)	0 ————— 100
Lugo et al. (2016)	0 ————— 100
Sterzi et al. (2016)	0 ————— 100
Roman-Blas (2017)	0 ————— 100

Appendix D: OPENBUGS model code

The model is specified in OpenBUGS as shown below, saved in a text file 'model-2mu.txt'. Data from both intervention arms were assumed to be normally distributed with mean `mean.t` and precision `tau.t` for the treatment arm, and `mean.c` and `tau.c` for the control arm. Note that `tau` is equivalent to $1/\sigma^2$, where σ^2 refers to the variance parameter. The parameter estimates for standardized treatment effect can be provided by `TE= mean.t-mean.c`.

```
model{
  # Likelihood current data
  for (i in 1:12) {
    yt[i] ~ dnorm(mean.t, tau.t)    # treatment arm
    yc[i] ~ dnorm(mean.c, tau.c)    # control arm
  }

  # The power prior
  a <- vn.t/2                        # vn.t =  $v_{tn}$ , prior sample size
  b <- (vn.t*sig2n.t)/2              # sig2n.t =  $\sigma_{tn}^2$ 
  c <- vn.c/2                        # vn.c =  $v_{cn}$ 
  d <- (vn.c*sig2n.c)/2              # sig2n.c =  $\sigma_{cn}^2$ 

  tau.t ~ dgamma(a, b)              # because  $\sigma_t^2 \sim inv - gamma(a, b)$ 
  tau.c ~ dgamma(c, d)
  sigma2.t <- 1/tau.t
  sigma2.c <- 1/tau.c

  taun.t <- tau.t*vn.t               # Gelman et al. Book(2008) p.78
  taun.c <- tau.c*vn.c
  mean.t ~ dnorm(mun.t, taun.t)     # mun.t =  $\mu_{tn}$ 
  mean.c ~ dnorm(mun.c, taun.c)     # mun.c =  $\mu_{cn}$ 
  TE <- mean.t-mean.c               # the treatment effect
}
```

Appendix E: References for the 20 external studies

- Noack W, Fischer M, Forster KK, Rovati LC, Setnikar I. Glucosamine sulfate in osteoarthritis of knee. *Osteoarthritis and Cartilage* 1994;2:51-59.
- Bourgeois P, Chales G, Dehais J, Delcambre B, Kuntz J, Rozenberg S. Efficacy and tolerability of chondroitin sulfate 1200 mg/day vs chondroitin sulfate 3x400 mg/day vs placebo. *Osteoarthritis and Cartilage* 1998;6:25-30.
- Mazieres B, Combe B, Phan Van A, Tondut J, Grynfeldt M. Chondroitin sulfate in osteoarthritis of the knee: a prospective, double blind, placebo controlled multicenter clinical study. *The Journal of Rheumatology* 2001;28:173-181.
- Hanson RR, Brawner WR, Blaik MA, Hammad TA, Kincaid SA, Pugh DG. Oral treatment with a nutraceutical (Cosequin) for ameliorating signs of Navicular syndrome in horses. *Veterinary Therapeutics* 2001;2:148-159.
- Pavelka K, Gatterova J, Olejarova M, Machacek S, Giacovelli G, Rovati LC. Glucosamine sulfate use and delay of progression of knee osteoarthritis: a 3-year, randomized, placebo-controlled, double-blind study. *Arch Intern Med.* 2002;162:2113-2123.
- McAlindon T, Formica M, LaValley M, Lehmer M, Kabbara K. Effectiveness of Glucosamine for symptoms of Knee Osteoarthritis: results from an internet-based randomized double-blind controlled trial. *The American journal of medicine* 2004;117:643-649.
- Usha PR, Naidu MUR. Randomised, double-blind, parallel, placebo-controlled study of oral Glucosamine, Methylsulfonylmethane and their combination in osteoarthritis. *Clin Drug invest* 2004;24:353-363.
- Forsyth RK, Brigden CV, Northrop AJ. Double blind investigation of the effects of oral supplementation of combined glucosamine hydrochloride (GHCL) and chondroitin sulphate (CS) on stride characteristics of veteran horses. *Equine vet. J. Suppl.* 2006;36:622-625.
- Mazieres B, Hucher M, Zaim M, Garnerio P. Effect of chondroitin sulphate in symptomatic knee osteoarthritis: a multicenter, randomized, double-blind, placebo-controlled study. *Ann Rheum Dis* 2006;66:639-645.
- D'Altio M, Peal A, Alvey M, Simms C, Curtsinger A, Gupta RC, et al. Therapeutic efficacy and safety of undenatured type II collagen singly or in combination with glucosamine and Chondroitin in arthritic dogs. *Toxicology Mechanisms and Methods* 2007;17:189-196.
- Frestedt JL, Walsh M, Kuskowski MA, Zenk JL. A natural mineral supplement provides relief from knee osteoarthritis symptoms: a randomized controlled pilot trial. *Nutrition Journal* 2008; doi: 10.1186/1475-2891-7-9.

- Rozendaal RM, Koes BW, Van Osch JVM, Uitterlinden EJ, Garling EH, Willemsen SP, et al. Effect of Glucosamine sulfate on hip osteoarthritis: a randomized trial. *Ann Intern Med.* 2008;148:268-277.
- Gupta RC, Canerdy TD, Skaggs P, Stocker A, Zyrkowski G, Burke R, et al. Therapeutic efficacy of undenatured type-II collagen (UC-II) in comparison to glucosamine and chondroitin in arthritic horses. *J. vet. Pharmacol. Therap.* 2009;32:577-584.
- Petersen SG, Saxne T, Heinegard D, Hansen M, Holm L, Koskinen S, et al. Glucosamine but not ibuprofen alters cartilage turnover in osteoarthritis patients in response to physical training. *Osteoarthritis and Cartilage* 2010;18:34-40.
- Nieman DC, Shanely RA, Luo B, Dew D, Meaney MP, Sha W. A commercialized dietary supplement alleviates joint pain in community adults: a double-blind, placebo-controlled community trial. *Nutrition Journal* 2013; doi: 10.1186/1475-2891-12-154.
- Tsuji T, Yoon J, Kitano N, Okura T, Tanaka K. Effects of N-acetyl glucosamine and chondroitin sulfate supplementation on knee pain and self-reported knee function in middle-aged and older Japanese adults: a randomized, double-blind, placebo-controlled trial. *Aging clin Exp res* 2015;28:197-205.
- Lugo JM, Saiyed ZM, Lane NE. Efficacy and tolerability of an undenatured type II collagen supplement in modulating knee osteoarthritis symptoms: a multicenter randomized, double-blind, placebo-controlled study. *Nutrition Journal* 2016; doi: 10.1186/s12937-016-0130-8.
- Sterzi S, Giordani L, Morrone M, Lena E, Magrone G, Scarpini C, et al. The efficacy and safety of a combination of glucosamine hydrochloride, chondroitin sulfate and bio-curcumin with exercise in the treatment of knee osteoarthritis: a randomized, double-blind, placebo-controlled study. *Eur J Phys Rehabil Med.* 2016;52:321-30.
- Roman-Blas JA, Castañeda S, Sánchez-Pernaute O, Largo R, Herrero-Beaumont G. Combined Treatment With Chondroitin Sulfate and Glucosamine Sulfate Shows No Superiority Over Placebo for Reduction of Joint Pain and Functional Impairment in Patients With Knee Osteoarthritis: A Six-Month Multicenter, Randomized, Double-Blind, Placebo-Controlled Clinical Trial. *Arthritis Rheumatol.* 2017;69:77-85. doi: 10.1002/art.39819.
- Scott RM, Evans R, Conzemius MG. Efficacy of an oral nutraceutical for the treatment of canine osteoarthritis. A double-blind, randomized, placebo-controlled prospective clinical trial. *Vet Comp Orthop Traumatol.* 2017;30:318-323. doi: 10.3415/VCOT-17-02-0020.

Chapter 7

Summarizing Discussion

In this thesis we applied several Bayesian methods in the context of veterinary clinical epidemiology. The aims were to use available sources of information to improve the precision or reduce the bias of parameter estimates, and to increase the diagnostic ability of prediction models.

In Part I, we explored Bayesian Hui-Walter (HW) and logistic regression (LR) latent class models for estimation of test characteristics when there is no gold standard. The HW approach is conventionally used in veterinary diagnostic test characteristic estimation (Dohoo et al., 2009). In order to capture the hierarchical structure of the data, we also applied multilevel LR latent class models with herd level random effects and individual level covariates. Results of a series of simulations with varying population characteristics showed that LR latent class models provided narrower posterior 95% central credible intervals (CCIs) for the test characteristics in comparison to the respective HW models. This was confirmed in the empirical example for the single intradermal comparative cervical tuberculin test (SICCT, i.e., skin test) and post-mortem examination in bovine tuberculosis (bTB) diagnosis with the abattoir data in Northern Ireland. In addition to the test characteristics, the LR model also provided estimated odds ratios of the risk factors age at death, days from last skin test to slaughter, sex and last skin test reason.

Results from the simulation and the empirical application suggested that the performance of HW and LR latent class models was robust across different data scenarios for estimation of test sensitivity and specificity, with the LR models providing more precise estimates. In particular, LR models that incorporated herd level clustering effects resulted in the least biased and most precise estimates. Furthermore, if researchers are interested in obtaining the odds ratios of risk factors, then the LR modelling approach is a better option. This work shows that LR models are in many situations a valuable alternative to HW models for estimation of test characteristics in the absence of a perfect reference test.

The chapters in part II also concern hierarchical data. In these chapters, we aimed to add cluster specific informative priors for the random effects for individual prediction in new

clusters. As clusters from the model development data and the new clusters are assumed to be exchangeable (i.e., originated from the same population), random effects for all clusters are assumed to have the same normal distribution. We could hence utilize this information from the known clusters to obtain cluster specific effects for the new clusters. The method was first investigated in the context of simulated data. Bayesian prediction models were specified with non-informative priors or informative priors with expert opinion at cluster level. Both optimal and suboptimal expert opinion was explored. Results showed that models with expert opinion produced better overall discrimination and calibration as well as better within cluster calibration. Results also revealed that incorporation of more precise expert opinion led to better individual level prediction, and even with suboptimal expert opinion, the predictive performance still improved in certain settings. This approach was subsequently applied to an observational study where subclinical ketosis was diagnosed by a prediction model in Dutch dairy cows (Van der Drift et al., 2012). Herd specific expert opinion was elicited from a bovine health specialist using three different scales of precision, resulting in three Bayesian prediction models with different degrees of informative priors for the random effects. However results showed that the three Bayesian models with expert opinion remained poor at the individual cow level. At the herd level, Bayesian prediction models showed slightly higher diagnostic accuracy.

As the current research literature in prediction modelling for clustered data shows, random effects are commonly ignored or removed after the model development phase. The prediction models that neglect the clustering structure in future new data may lead to a loss of prediction accuracy (Bouwmeester et al., 2013). We developed a Bayesian approach that could attain random effects in the prediction model when applied to future similar clusters. If there is cluster level information of new clusters available, for instance judgements from clinical experts, this approach offers the opportunity to naturally incorporate such information in the prediction by means of the priors for the random cluster effects. In our case, improvements of adding cluster specific expert opinion seen in the simulation were however not seen in real-world ketosis data. This demonstrates the importance of implementing methods developed in simulations in empirical examples where the data are more complex. It also indicates the difficulty of expert elicitation in practice. The form of expert opinion, as one of the important elements of expert elicitation, for instance, can be expressed as a pinpointed value or area on a visual probability scale, or simply a number with its range of uncertainty given by the expert(s). Results from our studies showed that some experts preferred to extract their opinion using visual scales while others found assigning numbers practically easier. Elicitation of subjective information from

experts is a research field of itself. Interested readers may be referred to research done for instance by O'Hagan et al. (2006).

In part III we further investigated the possibility to include external evidence in the analysis through informative priors. Specifically, we attempted to add cross-species evidence to a randomized controlled trial (RCT) analysis. The use of cross-species data in an analysis might yield challenges in clinical practice, however it is not new. Research studies on this topic can be found in veterinary and laboratory animal species, such as between cattle and swine, and between dogs and human (Huang et al., 2015; Li et al., 2014; Lin et al., 2015). We applied the Bayesian power prior approach to re-analyse a Dutch equine RCT evaluating the effect of oral Glucosamine and Chondroitin in aged horses (Higler et al., 2014). Nine cross-species studies (i.e., horse, dog, human) that used the same supplements as the equine trial were included in the analysis. Posterior results of the standardized treatment effect showed no difference between the treatment and control arm. This supported the results from the original equine trial. However, sensitivity analyses produced heterogeneous results which reflected the current state of knowledge on this clinical intervention, as substantial heterogeneity was found across studies both in humans and horses (Towheed et al., 2005; Pearson et al., 2009; Singh et al., 2015). Possible reasons could be the differences in study quality, duration of the intervention, preparation of the supplements, industry involvement, publication bias, selection bias and various study sample sizes.

In this clinical re-analysis, studies from horses, dogs and human that could potentially be used to specify the power prior distribution were first screened by an equine expert, such that the pathophysiology, the dose, the duration and the treatment of the selected studies were relevant to the equine RCT research question. We assumed that different clinical measurements, such as stride length from the equine RCT and pain scores from human studies, shared the same underlying dysfunction(s) and therefore could be standardized into a single theoretical outcome parameter. However, this is a strong assumption that some researchers may find unacceptable. In addition, converting standardized effect back to a clinical quantity may be challenging. Furthermore, when the sample sizes of the observed clinical data are small, for instance RCTs for the 'minor-species', the influence of the RCT data may be limited and the posterior distribution may be dominated by the prior distribution. Future studies should investigate further on these issues.

We think the synthesis of cross-species evidence in an explicit and transparent way has merits in veterinary clinical practice. It was demonstrated in this part that it is methodologically

possible within the Bayesian framework. Clinical applications of such approach, however, can be difficult to implement.

This thesis presented the development and evaluation of several Bayesian statistical methods with simulated and empirical data. In addition to the well-accepted Bayesian analysis with non-informative priors, our research results showed the feasibility of using informative priors to accumulate relevant evidence in veterinary clinical epidemiology.

Reference list

- Abernethy DA, Denny GO, Menzies FD, McGuckian P, Honhold N, Roberts AR. The Northern Ireland programme for the control and eradication of *Mycobacterium bovis*. *Veterinary Microbiology* 2006;112:231-237. doi: 10.1016/j.vetmic.2005.11.023.
- Abernethy DA, Upton P, Higgins IM, McGrath G, Goodchild AV, Rolfe SJ, Broughan JM, Downs SH, Clifton-Hadley R, Menzies FD, de la Rua-Domenech R, Blissitt MJ, Duignan A, More SJ. Bovine tuberculosis trends in the UK and the republic of Ireland, 1995-2010. *Veterinary Record* 2013;172:312. doi: 10.1136/vr.100969.
- Agresti A, Caffo B, Ohman-Strickland P. Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics and Data Analysis* 2004;47:639-653.
- Azzalini A, Capitanio A. The skew-normal and related families. Cambridge university press, New York; 2014.
- Badenoch D, Heneghan C. Evidence-based Medicine Toolkit. BMJ Books, London; 2002.
- Bates D, Maechler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 2015;67:1-48.
- Bermingham ML, Handel IG, Glass EJ, Woolliams JA, de Clare Bronsvort BM, McBride SH, Skuce RA, Allen AR, McDowell SWJ, Bishop SC. Hui and Walter's latent-class model extended to estimate diagnostic test properties from surveillance data: a latent model for latent data. *Scientific Reports* 2015;5:11861. doi: 10.1038/srep11861.
- Berry S.M, Lee, J.J., Carlin, B.P. Bayesian Adaptive Methods for Clinical Trials. Chapman & Hall/CRC, Boca Raton, FL; 2011.
- Bessell PR, Orton R, White PCL, Hutchings MR, Kao RR. Risk factors for bovine Tuberculosis at the national level in Great Britain. *BMC Vet Res* 2012;8:51. doi: 10.1186/1746-6148-8-51.
- Best N, Richardson S, Thomson A. A comparison of Bayesian spatial models for disease mapping. *Statistical methods in medical research*. 2005;14:35–59.
- Bouwmeester W, Twisk JWR, Kappen TH, Van Klei WL, Moons KGM, Vergouwe Y. Prediction models for clustered data: comparison of a random intercept and standard regression model. *BMC medical research methodology* 2013;13:19. doi: 10.1186/1471-2288-13-19.

- Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, Altman DG, Moons KGM. Reporting and methods in clinical prediction research: A systematic review. *PLoS Medicine* 2012;9:1-12. doi: 10.1371/journal.pmed.1001221.
- Brenner H, Gefeller O Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Statistics in Medicine* 1997;16:981-991.
- Butler SM, Louis TA. Random effects models with non-parametric priors. *Statistics in Medicine* 1992;11:1981-2000. doi: 10.1002/sim.4780111416.
- Byrne AW, Graham J, Brown C, Donaghy A, Guelbenzu-Gonzalo M, McNair J, Skuce R, Allen A, McDowell S. Bovine tuberculosis visible lesions in cattle culled during herd breakdowns: the effects of individual characteristics, trade movement and co-infection. *BMC Veterinary Research* 2017;13:400. doi: 10.1186/s12917-017-1321-z.
- Chen MH, Ibrahim JG. The relationship between the power prior and hierarchical models. *Bayesian Analysis* 2006;1:551-574.
- Clayton HM, Almeida PE, Prades M, Brown J, Tessier C, Lanovaz JL. Double-Blind Study of the Effects of an Oral Supplement Intended to Support Joint Health in Horses with Tarsal Degenerative Joint Disease. *Proceedings of the Annual Convention of the AAEP* 2002;48:314-317.
- Clegg DO, Reda DJ, Harris CL, Klein MA, O'Dell JR, Hooper MM, Bradley JD, Bingham CO 3rd, Weisman MH, Jackson CG, Lane NE, Cush JJ, Moreland LW, Schumacher HR Jr, Oddis CV, Wolfe F, Molitor JA, Yocum DE, Schnitzer TJ, Furst DE, Sawitzke AD, Shi H, Brandt KD, Moskowitz RW, Williams HJ. Glucosamine, chondroitin sulfate, and the two in combination for painful knee osteoarthritis. *N Engl J Med.* 2006;354:795-808. doi: 10.1056/NEJMoa052771.
- Clegg TA, Duignan A, Whelan C, Gormley E, Good M, Clarke J, Toft N, More SJ. Using latent class analyses to estimate the test characteristics of the γ -interferon test, the single intradermal comparative tuberculin test and a multiplex immunoassay under Irish conditions. *Veterinary Microbiology* 2011;151:86-76. doi: 10.1016/j.vetmic.2011.02.027.
- Clegg TA, More SJ, Higgins IM, Good M, Blake M, Williams DH. Potential infection-control benefit for Ireland from pre-movement testing of cattle for tuberculosis. *Preventive Veterinary Medicine* 2008;84:94-111. doi: 10.1016/j.prevetmed.2007.11.004.
- Cockcroft PD, Holmes M.A. *Handbook of Evidence-based Veterinary Medicine*. Blackwell Publishing, Oxford; 2003.

- Collins J, Huynh M. Estimation of diagnostic test accuracy without full verification: a review of latent class methods. *Stat. med.* 2014;33:4141-4169.
- Crichton N. Visual Analogue Scale (VAS). *J CLIN NURS* 2001;10:706-706.
- DAERA. Department of Agriculture, Environment and Rural Affairs, Bovine Tuberculosis statistics 2015. Online available on: (<https://www.daera-ni.gov.uk/publications/tuberculosis-disease-statistics-northern-ireland-2015>).
- Denwood MJ. Runjags: an R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *J. Stat. Software*, 2016;71. doi: 10.18637/jss.v071.i09.
- de la Rua-Domenech R, Goodchild AT, Vordermeier HM, Hewinson RG, Christiansen KH, Clifton-Hadley RS. Ante-mortem diagnosis of tuberculosis in cattle: a review of the tuberculin tests, gamma-interferon assay and other ancillary diagnostic techniques. *Research in Veterinary Science* 2006;81:190-210. doi: 10.1016/j.rvsc.2005.11.005.
- Dohoo IR, Martin W, Stryhn HE. *Veterinary epidemiologic research* 2nd edition. VRC Inc; 2009.
- Fernandes LG, Denwood MJ, Santos CSAB, Alves CJ, Pituco EM, Romaldini AHCN, De Stefano E, Nielsen SS, De Azevedo SS. Bayesian estimation of herd-level prevalence and risk factors associated with BoHV-1 infection in cattle herds in the State of Paraíba, Brazil. *Prev Vet Med.* 2019;169:104705. doi:10.1016/j.prevetmed.2019.104705.
- Finkelstein BS, French B, Kimmel SE. The prediction accuracy of dynamic mixed-effects models in clustered data. *BioData Min.* 2016;9:5. doi: 10.1186/s13040-016-0084-6.
- Forsyth RK, Brigden CV, Northrop AJ. Double blind investigation of the effects of oral supplementation of combined glucosamine hydrochloride (GHCL) and chondroitin sulphate (CS) on stride characteristics of veteran horses. *Equine vet. J. Suppl.* 2006;36:622-625.
- Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian data analysis*, 2nd edition. Chapman & Hall/CRC, London; 2008.
- Godfray HCJ, Donnelly CA, Kao RR, Macdonald DW, McDonald RA, Petrokofsky G, Wood JLN, Woodroffe R, Young DB, McLean AR. A restatement of the natural science evidence base relevant to the control of bovine tuberculosis in Great Britain. *Proc. R. Soc. B. Biol. Sci.* 2013;280:20131634. doi: 10.1098/rspb.2013.1634.
- Goldhawk C, Chapinal N, Veira DM, Weary DM, Von Keyserlingk MA. Parturition feeding behavior is an early indicator of subclinical ketosis. *J Dairy Sci.* 2009;92:4971-4977. doi: 10.3168/jds.2009-2242.

- Goodchild AV, Downs SH, Upton P, Wood JL, de la Rua-Domenech R. Specificity of the comparative skin test for bovine tuberculosis in Great Britain. *Veterinary Record* 2015;177:258. doi: 10.1136/vr.102961.
- Gould AL. Using prior findings to augment active-controlled trials and trials with small placebo groups. *Drug Information Journal* 1991;25:369–80.
- Gupta RC, Canerdy TD, Skaggs P, Stocker A, Zyrkowski G, Burke R, Wegford K, Goad JT, Rohde K, Barnett D, DeWees W, Bagchi M, Bagchi D. Therapeutic efficacy of undenatured type-II collagen (UC-II) in comparison to glucosamine and chondroitin in arthritic horses. *J. vet. Pharmacol. Therap.* 2009;32:577-584. doi: 10.1111/j.1365-2885.2009.01079.x.
- Hanson RR, Brawner WR, Blaik MA, Hammad TA, Kincaid SA, Pugh DG. Oral treatment with a nutraceutical (Cosequin) for ameliorating signs of Navicular syndrome in horses. *Veterinary Therapeutics* 2001;2:148-159.
- Hartnack S, Budke CM, Craig PS, Jiamin Q, Boufana B, Campos-Ponce M, Torgerson PR. Latent-class methods to evaluate diagnostics tests for *Echinococcus* infections in dogs. *PLoS Negl Trop Dis.* 2013;7:e2068. doi: 10.1371/journal.pntd.0002068.
- Heagerty PJ, Kurland BF. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* 2001;88:973-985.
- Hedeker D, Gibbons RD. *Longitudinal data analysis*. New Jersey: John Wiley & Sons; 2006.
- Higgins HM, Huxley JN, Wapenaar W, Green MJ. Quantifying veterinarians' beliefs on disease control and exploring the effect of new evidence: a Bayesian approach. *J Dairy Sci* 2014;97:3394-3408.
- Higler MH, Brommer H, L'Ami JJ, De Grauw JC, Nielen M, Van Weeren PR, Laverty S, Barneveld A, Back W. The effects of three-month oral supplementation with a nutraceutical and exercise on the locomotor pattern of aged horses. *Equine Vet J.* 2014;46:611-617.
- Hobbs BP, Carlin BP, Sargent DJ. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Analysis* 2012;7:1-36.
- Hochberg MC, Altman RD, April KT, Benkhalti M, Guyatt G, McGowan J, Towheed T, Welch V, Wells G, Tugwell P; American College of Rheumatology. American College of Rheumatology 2012 recommendations for the use of nonpharmacologic and pharmacologic therapies in osteoarthritis of the hand, hip, and knee. *Arthritis Care Res (Hoboken).* 2012;64:465-74. doi: 10.1002/acr.21596.

- Houston R. A computerised database system for bovine traceability. *Rev Sci Tech* 2001;20:652-661. doi: 10.20506/rst.20.2.1293.
- Hox JJ. Multilevel analysis: techniques and applications. Lawrence Erlbaum associations, New Jersey; 2002.
- Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range and the size of a sample. *BMC medical research methodology* 2005;5:13. doi: 10.1186/1471-2288-5-13.
- Huang Q, Gehring R, Tell LA, Li M, Riviere JE. Interspecies allometric meta-analysis of the comparative pharmacokinetics of 85 drugs across veterinary and laboratory animal species. *J Vet Pharmacol Ther.* 2015;38:214-226. doi: 10.1111/jvp.12174.
- Hui SL, Walter SD. Estimating the error rates of diagnostic tests. *Biometrics* 1980;36:167-171.
- Ibrahim JG, Chen MH. Power prior distributions for regression models. *Statistical Science* 2000;15:46-60.
- Ireland JL, Clegg PD, McGowan CM, McKane SA, Chandler KJ, Pinchbeck GL. Disease prevalence in geriatric horses in the United Kingdom: Veterinary clinical assessment of 200 cases. *Equine Veterinary Journal* 2012;44:101-106.
- Jamali H, Barkema HW, Jacques M, Lavallée-Bourget EM, Malouin F, Saini V, Stryhn H, Dufour S. Invited review: Incidence, risk factors, and effects of clinical mastitis recurrence in dairy cows. *J Dairy Sci.* 2018;101:4729-4746. doi: 10.3168/jds.2017-13730.
- Johnson WO, Jones G, Gardner IA. Gold standards are out and Bayes is in: Implementing the cure for imperfect reference tests in diagnostic accuracy studies. *Preventive Veterinary Medicine* 2019;167:113-127. doi: 10.1016/j.prevetmed.2019.01.010.
- Jorritsma R, Baldée SJC, Schukken YH, Wensing T, Wentink GH. Evaluation of a milk test for detection of subclinical ketosis. *Vet. Q.* 1998;20:108–110.
- Karolemeas K, de la Rua-Domenech R, Cooper R, Goodchild AV, Clifton-Hadley RS, Conlan AJ, Mitchell AP, Hewinson RG, Donnelly CA, Wood JL, McKinley TJ. Estimation of relative sensitivity of the comparative tuberculin skin test in tuberculous cattle herds subjected to depopulation. *PLoS ONE* 2012;7:e43217. doi: 10.1371/journal.pone.0043217.
- Koop G, Collar CA, Toft N, Nielsen M, Van Werven T, Bacon DAC, Gardner IA. Risk factors for subclinical intramammary infection in dairy goats in two longitudinal field studies evaluated by Bayesian logistic regression. *Preventive Veterinary Medicine* 2013;108:304-312.

- Krogh MA, Toft N, Enevoldsen C. Latent class evaluation of a milk test, a urine test, and the fat-to-protein percentage ratio in milk to diagnose ketosis in dairy cows. *Journal of Dairy Science* 2011;94:2360-2367.
- Lahuerta-Marin A, Milne MG, McNair J, Skuce RA, McBride SH, Menzies FD, McDowell SJW, Byrne AW, Handel IG, de C Bronsvort BM. Bayesian latent class estimation of sensitivity and specificity parameters of diagnostic tests for bovine tuberculosis in chronically infected herds in Northern Ireland. *Veterinary Journal* 2018;238:15-21. doi: 10.1016/j.tvjl.2018.04.019.
- Li M, Gehring R, Tell L, Baynes R, Huang Q, Riviere JE. Interspecies mixed-effect pharmacokinetic modeling of penicillin G in cattle and swine. *Antimicrob Agents Chemother.* 2014;58:4495-4503.
- Ligges U, Sturtz S, Gelman A, Gorjanc G, Jackson C. Package 'BRugs': Interface to the 'OpenBUGS' MCMC Software. 2017; R CRAN Project.
- Lin Z, Li M, Gehring R, Riviere JE. Development and application of a multiroute physiologically based pharmacokinetic model for oxytetracycline in dogs and humans. *J Pharm Sci.* 2015;104:233-243. doi: 10.1002/jps.24244.
- Litière S, Alonso A, Molenberghs G. Type I and type II error under random-effects misspecification in generalized linear mixed model. *Biometrics* 2007;63:1038-1044.
- Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainty. *Am. J. Epidemiol.* 1997;146:195-203.
- McAlindon TE, Bannuru RR, Sullivan MC, Arden NK, Berenbaum F, Bierma-Zeinstra SM, Hawker GA, Henrotin Y, Hunter DJ, Kawaguchi H, Kwoh K, Lohmander S, Rannou F, Roos EM, Underwood M. OARSI guidelines for the non-surgical management of knee osteoarthritis. *Osteoarthritis Cartilage.* 2014;22:363-88. doi: 10.1016/j.joca.2014.01.003.
- McCulloch CE, Neuhaus JM. Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical science* 2011;26:388-402.
- McInturff P, Johnson WO, Cowling D, Gardner IA. Modelling risk when binary outcomes are subject to error. *Stat. Med.* 2004;23:1095-1109.
- Neuenschwander B, Branson M, Spiegelhalter DJ. A note on the power prior. *Statistics in Medicine* 2009;28:3562-3566.
- Ni H, Groenwold RHH, Nielen M, Klugkist I. Prediction models for clustered data with informative priors for the random effects: A simulation study. *BMC Medical Research Methodology* 2018;18:83. doi: 10.1186/s12874-018-0543-5.

- Nuñez-García J, Downs SH, Parry JE, Abernethy DA, Broughan JM, Cameron AR, Cook AJ, de la Rua-Domenech R, Goodchild AV, Gunn J, More SJ, Rhodes S, Rolfe S, Sharp M, Upton PA, Vordermeier HM, Watson E, Welsh M, Whelan AO, Woolliams JA, Clifton-Hadley RS, Greiner M. Meta-analyses of the sensitivity and specificity of ante-mortem and post-mortem diagnostic tests for bovine tuberculosis in the UK and Ireland. *Preventive Veterinary Medicine* 2017;153:94-107. doi: 10.1016/j.prevetmed.2017.02.017.
- O'Hagan A, Buck CE, Daneshkhah A, Eiser R, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T. *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons, Chichester; 2006.
- O'Hagan MJH, Ni H, Menzies FD, Pascual-Linaza AV, Georgaki A, Stegeman JA. Test Characteristics of the Tuberculin Skin Test and Post-mortem Examination for Bovine Tuberculosis Diagnosis in Cattle in Northern Ireland Estimated by Bayesian Latent Class Analysis with Adjustments for Covariates. *Epidemiology and Infection* 2019;147:e209. doi: 10.1017/S0950268819000888.
- O'Hagan MJ, Courcier EA, Drewe JA, Gordon AW, McNair J, Abernethy DA. Risk factors for visible lesions or positive laboratory tests in bovine tuberculosis reactor cattle in Northern Ireland. *Preventive Veterinary Medicine* 2015;120:283-290. doi: 10.1016/j.prevetmed.2015.04.005.
- Oetzel GR. Monitoring and testing dairy herds for metabolic disease. *Vet. Clin. North Am. Food Anim. Pract.* 2004;20:651-674.
- OIE Terrestrial Manual 2009. Chapter 2,4,7. Online available on: (http://www.oie.int/fileadmin/Home/eng/Health_standards/tahm/2008/pdf/2.04.07_BOVINE_TB.pdf).
- Paul S, Agger JF, Agerholm JS, Markussen B. Prevalence and risk factors of *Coxiella burnetii* seropositivity in Danish beef and dairy cattle at slaughter adjusted for test uncertainty. *Prev. Med. Vet.* 2014;113:504–511. doi: 10.1016/j.prevetmed.2014.01.018.
- Pearson W, Lindinger M. Low quality of evidence for glucosamine-based nutraceuticals in equine joint disease: review of in vivo studies. *Equine vet J.* 2009;41:706-712.
- Pendleton A, Arden N, Dougados M, Doherty M, Bannwarth B, Bijlsma JW, Cluzeau F, Cooper C, Dieppe PA, Günther KP, Hauselmann HJ, Herrero-Beaumont G, Kaklamanis PM, Leeb B, Lequesne M, Lohmander S, Mazieres B, Mola EM, Pavelka K, Serni U, Swoboda B, Verbruggen AA, Weseloh G, Zimmermann-Gorska I. EULAR

- recommendations for the management of knee osteoarthritis: report of a task force of the Standing Committee for International Clinical Studies Including Therapeutic Trials (ESCISIT). *Ann Rheum Dis.* 2000;59:936-44. doi: 10.1136/ard.59.12.936.
- Petersen SG, Saxne T, Heinegard D, Hansen M, Holm L, Koskinen S, Stordal C, Christensen H, Aagaard P, Kjaer M. Glucosamine but not ibuprofen alters cartilage turnover in osteoarthritis patients in response to physical training. *Osteoarthritis and Cartilage* 2010;18:34-40. doi: 10.1016/j.joca.2009.07.004.
- Petrie A, Watson P. *Statistics for Veterinary and Animal Science*, 3rd edition. Wiley-Blackwell, Oxford; 2013.
- Plummer M. JAGS: a program for analysis of bayesian graphical models using gibbs sampling JAGS: just another gibbs sampler. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, Pp. March 20–22 ISSN 1609–395X.
- Pocock S. The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases* 1976;29:175–88.
- Pollock JM, McNair J, Bassett H, Cassidy JP, Costello E, Aggerbeck H, Rosenkrands I, Andersen P. Specific delayed-type hypersensitivity responses to ESAT-6 identify tuberculosis-infected cattle. *Journal of Clinical Microbiology* 2003;41:1856-1860.
- R Core Team. R: a Language and Environment for Statistical Computing. URL. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>. 2016.
- Raboisson D, Mounié M, Maigné E. Diseases, reproductive performance, and changes in milk production associated with subclinical ketosis in dairy cows: a meta-analysis and review. *J Dairy Sci.* 2014;97:7547-63. doi: 10.3168/jds.2014-8237.
- Rangel SJ, Paré J, Doré E, Arango JC, Côté G, Buczinski S, Labrecque O, Fairbrother JH, Roy JP, Wellemans V, Fecteau G. A systematic review of risk factors associated with the introduction of *Mycobacterium avium* spp. paratuberculosis (MAP) into dairy herds. *Can Vet J.* 2015;56:169-77.
- Ranta J, Tuominen P, Maijala R. Estimation of true *Salmonella* prevalence jointly in cattle herd and animal population using Bayesian hierarchical modelling. *Risk Anal.* 2005;25:23-37.
- Rietbergen C, Groenwold RHH, Hooijink HJA, Moons KGM, Klugkist I. Expert Elicitation of Study Weights for Bayesian Analysis and Meta-Analysis. *Journal of Mixed Methods Research* 2016;10:168-181. doi: 10.1177/1558689814553850.

- Rietbergen C, Klugkist I, Janssen KJM, Moons KGM, Hoijtink HJA. Incorporating of historical data in the analysis of randomized therapeutic trials. *Contemporary Clinical Trials* 2011;32:848-855.
- Robert CP. Simulation of truncated normal variables. *Statistics and computing* 1995;5:121-125.
- Ryan L. Using historical controls in the analysis of developmental toxicity data. *Biometrics* 1993;49:1126-35.
- Singh JA, Noorbaloochi S, MacDonald R, Maxwell LJ. Chondroitin for osteoarthritis. *Cochrane Database Syst Rev.* 2015;1:CD005614. doi: 10.1002/14651858.CD005614.pub2.
- Skuce RA, Allen AR, McDowell SWJ. Bovine Tuberculosis (TB): A review of cattle-to-cattle transmission, risk factors and susceptibility, 2011. Online available on: (<https://www.daera-ni.gov.uk/sites/default/files/publications/dard/afbi-literature-review-tb-review-cattle-to-cattle-transmission.pdf>).
- Spiegelhalter DJ, Abram KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation. John Wiley & Sons, Chichester; 2004.
- Stan Development Team. RStan: the R interface to Stan. R package version 2.17.3. <http://mc-stan.org/>. 2018.
- Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. Springer, New York; 2009.
- Toft N, Jørgensen E, Højsgaard S. Diagnosing diagnostic tests: evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard. *Prev Vet Med.* 2005;68:19-33.
- Towheed TE, Maxwell L, Anastassiades TP, Shea B, Houpt J, Robinson V, Hochberg MC, Wells G. Glucosamine therapy for treating osteoarthritis. *Cochrane Database of Systematic Reviews* 2005;2005:CD002946. doi: 10.1002/14651858.CD002946.pub2.
- Tremblay M, Kammer M, Lange H, Plattner S, Baumgartner C, Stegeman JA, Duda J, Mansfeld R, Döpfer D. Identifying poor metabolic adaptation during early lactation in dairy cows using cluster analysis. *Dairy Sci.* 2018;101:7311-7321. doi: 10.3168/jds.2017-13582.
- Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol* 2016;74:167-176.

- Van der Drift SGA, Jorritsma R, Schonewille JT, Knijn HM, Stegeman JA. Routine detection of hyperketonemia in dairy cows using Fourier transform infrared spectroscopy analysis of β -hydroxybutyrate and acetone in milk in combination with test-day information. *J Dairy Sci.* 2012;95:4886-4898.
- Vandeweerd JM, Coisson C, Clegg P, Cambier C, Pierson A, Hontoir F, Saegerman C, Gustin P, Buczinski S. Systematic review of efficacy of nutraceuticals to alleviate clinical signs of osteoarthritis. *J Vet Intern Med.* 2012;26:448-56. doi: 10.1111/j.1939-1676.2012.00901.x.
- Vanholder T, Papen J, Bemers R, Vertenten G, Berge AC. Risk factors for subclinical and clinical ketosis and association with production parameters in dairy cows in the Netherlands. *J Dairy Sci.* 2015;98:880-888. doi: 10.3168/jds.2014-8362.
- Van Smeden M, Naaktgeboren CA, Reitsma JB, Moons KG, de Groot JA. Latent class models in diagnostic studies when there is no reference standard – a systematic review. *American Journal of Epidemiology* 2014;179:423-431. doi: 10.1093/aje/kwt286.
- Viele K, Berry S, Neuenschwander B, Amzal B, Chen F, Enas N, Hobbs B, Ibrahim JG, Kinnnersley N, Lindborg S, Micallef S, Roychoudhury S, Thompson L. Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat.* 2014;13:41-54. doi: 10.1002/pst.1589.
- Wandel S, Jüni P, Tendal B, Nüesch E, Villiger PM, Welton NJ, Reichenbach S, Trelle S. Effects of glucosamine, chondroitin, or placebo in patients with osteoarthritis of hip or knee: network meta-analysis. *BMJ.* 2010;341:c4675. doi: 10.1136/bmj.c4675.
- White GW, Stites T, Jones EW, Jordan S. Efficacy of Intramuscular Chondroitin Sulfate and Compounded Acetyl-d-Glucosamine in a Positive Controlled Study of Equine Carpal. *Journal of Equine Veterinary Science* 2003;23:295-300.

English summary

In this thesis, we examined and applied several Bayesian methods in veterinary clinical epidemiological studies, both for their modelling flexibility in handling complex empirical problems and for the possibility to include external evidence through prior distributions. The majority of current Bayesian applications seen in veterinary epidemiology is largely motivated by the simulation based Markov chain Monte Carlo (MCMC) parameter estimation. However, more and more researchers choose the Bayesian framework for its ability to include relevant background information in the clinical analyses by means of the prior distribution. Bayesian updating with informative priors is also in line with the practice of Evidence-based Veterinary Medicine (EBVM) which requires “the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients”. We explored three Bayesian methods with simulated or empirical data. Part I examined the Bayesian latent class modelling approach that estimated the diagnostic test characteristics and odds ratios of the risk factors. Part II illustrated the Bayesian diagnostic prediction model that incorporated cluster level expert opinion via informative priors of the random effects. And part III demonstrated the Bayesian power prior method that combined cross-species historical data in a clinical trial analysis.

We started our research in the context of diagnostic test characteristics estimation where the Bayesian framework has often been used for its modelling flexibility compared to the traditional frequentist framework. Bayesian latent class models are routinely applied to evaluate the diagnostic performance of imperfect tests when there is no reliable reference test with known measures of test accuracy (i.e., gold standard). We investigated two latent class methods. One was the classical Bayesian Hui-Walter (HW) model which aggregated the observed individual test results to the (sub)population level. As a result, potentially valuable information from the lower level(s) could not be fully incorporated. The other was the logistic regression (LR) latent class model which allowed inclusion of hierarchical data such as animal level covariates and herd level clustering effects. In chapter 2 we examined both approaches within simulations where true disease status and true test properties were predefined. Prevalence properties and test characteristics realistic for paratuberculosis in cattle were used. Individual cows were generated to be clustered within herds from two regions. Two tests with binary outcomes were simulated with constant test characteristics across the two regions. On top of the prevalence properties and test characteristics, one animal level binary risk factor was

added to the data. Results from various simulated scenarios showed that LR models were in many situations preferable over HW models to estimate test characteristics in the absence of a gold standard. The LR models that incorporated herd level clustering effects provided the most precise and the least biased estimates. In addition, LR models with the animal level risk factor provided robust estimate for the odds ratio of the risk factor across different data scenarios. Results also revealed that stratifying data on the basis of an animal level risk factor for the HW approach can be problematic, the LR approach is recommended when there are risk factors available.

To verify the conclusions obtained from the simulation study, we applied both Bayesian HW and LR latent class models to real-world data in chapter 3 where test characteristics of the ante-mortem single intradermal comparative cervical tuberculin (SICCT) test and post-mortem test for the diagnosis of bovine tuberculosis (bTB) in Northern Ireland were evaluated. The main motivation of this study was to gain more insight in the sensitivity of the SICCT test, for the reported sensitivity of the SICCT test using HW latent class models in the current literature has shown a lot of variation. Impact of various risk factors on bTB infection was also of research interest. Data were collected from all cattle slaughtered in abattoirs in Northern Ireland in 2015. For the HW model, ten divisional veterinary office (DVO) areas in Northern Ireland were treated as subpopulations with distinct prevalences. For the LR model, animal level risk factors (age at death, sex, days from last SICCT test to slaughter and last SICCT test reason) were incorporated in the final analysis. Both methods showed comparable posterior medians for the SICCT and post-mortem test sensitivity and specificity, but the LR model provided narrower posterior 95% credible intervals. The association between the bTB infection and each risk factor was evaluated by the LR model which was valuable information for the clinical practice of bTB disease control in Northern Ireland.

We further explored the use of Bayesian methods in the context of multilevel diagnostic prediction models. In veterinary epidemiology, random effects models are routinely used to take the clustering structure into account. For prediction models however, random effects are removed after model development as no information is considered available for future clusters. In chapter 4 we developed a new method to attain the random effects and include prior knowledge through the random effects distribution for prediction in future clusters. Study data were simulated and five prediction models were specified: a frequentist model that set the random effects to zero for all new clusters, a Bayesian model with non-informative priors for the random effects of new clusters, Bayesian models with expert opinion incorporated into low informative, medium informative and highly informative priors for the random effects of new

clusters. Simulated expert opinion at the cluster level was expressed in the form of a truncated area of the random effects distribution. The predictive performance of the five models was assessed within simulated populations having different characteristics. Impact of suboptimal expert opinion that deviated from the true quantity was explored as well. Results showed that the Bayesian prediction model using non-informative priors for the random effects performed similarly as the traditional frequentist model that removed the random effects. Bayesian prediction models using expert opinion as informative priors showed smaller Brier scores, better overall discrimination and calibration, as well as better within cluster calibration than the frequentist model. Results also indicated that incorporation of more precise expert opinion led to better predictions at the individual as well as at the cluster level. When suboptimal expert opinion was used as prior information, results indicated that prediction still improved in certain settings.

This prediction approach was subsequently applied in a clinical study in chapter 5. We were interested if by adding herd specific prior information, the prediction of subclinical ketosis (SCK) in early lactation dairy cows would improve. To answer this question, we used the data collected by Van der Drift et al. (2012) and the prediction model the authors developed where available test-day measures were used as predictors. In our study, a bovine specialist was asked to use available information on the feeding management and production level of the farms and provide his personal opinion on the SCK risk for each herd. The elicited expert opinion was incorporated as informative priors for the random effects in the prediction model. Results showed no improvement in prediction at the individual cow level when expert opinion was included. There was some improvement of the prediction at the herd level, but this was not considered clinically important. This clinical study showed that it was methodologically possible in practice to include herd level information as priors for the random effects in a prediction model. The Bayesian prediction modelling approach can be promising for diagnosis in herds when the between herd variance is relatively large and where all herds contain diseased animals under study.

The possibility to incorporate external evidence as informative priors in clinical analyses was further investigated in minor animal species. In veterinary epidemiology, veterinarians are by default trained to be cross-species thinkers as evidence on specific research questions can be scarce. In clinical practice, borrowing information from other species may happen on a daily basis in an implicit manner. In chapter 6, we adopted the Bayesian power prior method to aggregate relevant cross-species studies after being examined and weighted by a clinical expert regarding the relevance of each study to the research question at hand. A Dutch equine clinical

trial was re-analyzed to evaluate the effect of oral Glucosamine (GS) and Chondroitin Sulfate (CS) on stiff joints in elderly horses. Research studies that assess the clinical effectiveness of these two nutraceuticals on stiff joints are limited and yield conflicting results. In our study, systematic literature searches were therefore done to detect existing relevant information across species. Twenty studies conducted in humans, dogs and horses were found eligible and were subsequently weighted by an equine expert for their clinical relevance on the equine research question. Nine clinical studies that used the combination of GS and CS as treatment were included to specify the power prior and incorporated in the main analysis. In order to have all outcomes on a same scale, standardized values were derived from the raw outcome measures of the equine trial and the nine cross-species studies for the power prior. Results from the main analysis showed no effect of the nutraceuticals which supported the original conclusion from the equine trial. However, sensitivity analyses showed mixed results, which reflected the current heterogeneity and uncertainty on the clinical effect size of GS and CS for stiff joints in elderly individuals. The Bayesian power prior approach showed merits in incorporating historical relevant data into a clinical trial analysis, however, cross-species evidence synthesis yielded challenges and needed further investigations.

This thesis presents the development and evaluation of several Bayesian methods in veterinary clinical epidemiology. We showed that in addition to the well-accepted Bayesian analysis with non-informative priors, informative priors could be a useful tool to accumulate relevant evidence in an explicit and transparent manner in the clinical practice of veterinary epidemiology.

Nederlandse Samenvatting

In dit proefschrift hebben we verschillende Bayesiaanse methoden onderzocht en toegepast binnen de context van veterinaire klinische epidemiologie. Hierbij hebben we gekeken naar zowel hun modelleringsflexibiliteit voor het oplossen van complexe empirische vraagstukken als naar de mogelijkheid om extern bewijs op te nemen via prior verdelingen. Het merendeel van de huidige Bayesiaanse toepassingen in de veterinaire epidemiologie is gemotiveerd door de op simulatie gebaseerde Markov Chain Monte Carlo (MCMC) schattingsmethoden. Steeds meer onderzoekers gebruiken echter het Bayesiaanse raamwerk om relevante achtergrondinformatie via de prior verdeling op te nemen in de klinische analyses. Bayesiaanse updating met informatieve priors is ook in lijn met de praktijk van Evidence-based Veterinary Medicine (EBVM), die "consciëntieus, expliciet en oordeelkundig het meest actuele en juiste bewijs gebruikt bij het nemen van beslissingen over de zorg voor individuele patiënten in het veterinaire beroep". We hebben drie Bayesiaanse methoden onderzocht met gesimuleerde of empirische data. Deel I onderzocht de Bayesiaanse latente klasse modelleringsbenadering die de diagnostische testkenmerken en odds ratio's van de risicofactoren schatte. Deel II illustreerde het Bayesiaanse diagnostische voorspellingsmodel dat expert meningen op clusterniveau omvatte via informatieve priors voor de random effects. En deel III demonstreerde de Bayesiaanse power prior methode die cross-species historische data combineerde in een klinische analyse.

We zijn begonnen in de context van het schatten van diagnostische testkenmerken, waarbij het Bayesiaanse raamwerk vaak is gebruikt vanwege zijn modelleringsflexibiliteit in vergelijking met het traditionele frequentistische raamwerk. Bayesiaanse latente klasse modellen worden routinematig toegepast om de diagnostische kwaliteit van imperfecte tests te evalueren wanneer er geen betrouwbare referentietest is met bekende waarden van testnauwkeurigheid (de zogenaamde gouden standaard). We onderzochten twee latente klasse methoden. Een daarvan was het klassieke Bayesiaanse Hui-Walter (HW) model dat de waargenomen individuele testresultaten aggregateerde tot op het (sub)populatie niveau. Als gevolg hiervan kon potentieel waardevolle informatie van de lagere niveaus niet volledig worden meegenomen. De andere was het logistische regressie (LR) latente klasse model, waarmee hiërarchische data kon worden opgenomen, zoals covariaten op dierniveau en clustereffecten op koppelniveau. In hoofdstuk 2 onderzochten we beide modelleringsmethoden binnen simulaties waarbij de ware ziektestatus en testeigenschappen vooraf waren

gedefinieerd. Prevalentie eigenschappen en testkenmerken die realistisch zijn voor paratuberculose bij runderen werden gebruikt. Individuele koeien geclusterd binnen koppels uit twee regio's en twee imperfecte tests met constante testkenmerken in de regio's werden gesimuleerd. Naast de prevalentie eigenschappen en testkenmerken werd één risicofactor op dierniveau aan de data toegevoegd. Resultaten van verschillende scenario's toonden aan dat LR-modellen in veel situaties de voorkeur hadden boven HW-modellen om testkenmerken te schatten bij afwezigheid van een gouden standaard. De LR-modellen die clustereffecten op koppelniveau bevatten, leverden de meest nauwkeurige en minst vertekende schattingen op. Bovendien leverden LR-modellen met de risicofactor op dierniveau een robuuste schatting voor de odds ratio van de risicofactor. De resultaten lieten ook zien dat het splitsen van (sub)populaties op basis van een risicofactor voor de HW-aanpak problematisch kan zijn. De LR-aanpak wordt daarom aanbevolen wanneer er risicofactoren beschikbaar zijn.

Om de conclusies van het simulatieonderzoek te verifiëren, hebben we in hoofdstuk 3 de Bayesiaanse HW en LR latente klasse modellen toegepast op real-world data, waar testkenmerken van de ante-mortem single intradermal comparative cervical tuberculin (SICCT) test en post-mortem test voor de diagnose van runder tuberculose (bTB) in Noord-Ierland werden geëvalueerd. De motivatie van deze studie was om meer inzicht te krijgen in de sensitiviteit van de SICCT test, aangezien de gerapporteerde sensitiviteit van de SICCT test met behulp van HW latente klasse modellen in de huidige literatuur veel variatie laat zien. De impact van verschillende risicofactoren op bTB infectie was ook van onderzoeksbelang. Data was verzameld van alle runderen die in 2015 zijn geslacht in slachthuizen in Noord-Ierland. Voor het HW-model werden veterinaire kantoren (DVO) van tien gebieden in Noord-Ierland behandeld als subpopulaties met verschillende prevalenties. Voor het LR-model werden risicofactoren op dierniveau (leeftijd bij overlijden, geslacht, dagen van laatste SICCT test tot slachting en laatste SICCT testreden) meegenomen in de analyse. Beide methoden lieten vergelijkbare posterior medianen zien voor de sensitiviteit en specificiteit van de SICCT en de post-mortem tests, maar het LR-model leverde smallere posterior 95% schattingsintervallen. De associatie tussen de bTB infectie en elke risicofactor werd geëvalueerd door het LR-model, wat waardevolle informatie was voor het bTB controleprogramma in Noord-Ierland.

We hebben het gebruik van Bayesiaanse methoden verder onderzocht in de context van multilevel diagnostische voorspellingsmodellen. In veterinaire epidemiologie worden random effects modellen vaak gebruikt om de clusterstructuur te includeren. Voor voorspellingsmodellen worden random effects echter verwijderd na modelontwikkeling, omdat er geen informatie beschikbaar wordt geacht voor toekomstige clusters. In hoofdstuk 4 hebben

we een nieuwe methode ontwikkeld om random effects in het model te houden en prior informatie op te nemen via de random effects verdeling voor voorspelling in toekomstige clusters. Data was gesimuleerd en vijf voorspellingsmodellen waren gespecificeerd: een frequentistisch model dat random effects op nul zette voor alle nieuwe clusters, een Bayesiaans model met non-informatieve priors voor random effects van nieuwe clusters, Bayesiaanse modellen met expert meningen opgenomen in laag informatieve, medium informatieve en hoog informatieve priors voor random effects van nieuwe clusters. Gesimuleerde expert mening voor elke cluster werd gespecificeerd in de vorm van een getrunceerde verdeling van de random effects. De vijf modellen zijn onderzocht in gesimuleerde populaties met verschillende kernmerken. Ook werd de impact onderzocht van suboptimale expert schattingen die afweken van de werkelijke waarden. De resultaten toonden aan dat het Bayesiaanse voorspellingsmodel dat non-informatieve priors voor random effects gebruikte, vergelijkbare resultaten opleverde als het traditionele frequentistische model dat random effects verwijderde. Bayesiaanse voorspellingsmodellen die expert meningen als informatieve priors gebruikten toonden kleinere Brier scores, betere algemene discriminatie en kalibratie, evenals betere kalibratie binnen clusters dan het frequentistische model. De resultaten gaven ook aan dat het meenemen van een meer nauwkeurige expert meningen tot betere voorspellingen leidde zowel op individu als op clusterniveau. Wanneer suboptimale expert meningen werden gebruikt als prior informatie, gaven de resultaten aan dat de voorspelling in bepaalde situaties nog steeds verbeterde.

Deze voorspellingsmethode werd vervolgens toegepast in een klinische studie in hoofdstuk 5. We waren geïnteresseerd of door het toevoegen van bedrijfsspecifieke prior informatie de voorspelling van subklinische ketose (SCK) bij melkkoeien in de vroege lactatie zou verbeteren. Om deze vraag te beantwoorden hebben we gebruik gemaakt van de data verzameld door Van der Drift et al. (2012) en het voorspellingsmodel ontwikkeld door de auteurs waar beschikbare testdagmetingen waren gebruikt als voorspellers. In onze studie, heeft een runderspecialist zijn persoonlijke schatting gegeven over het SCK-risico voor elk koppel ten opzichte van andere koppels in Nederland op basis van beschikbare informatie over het voermanagement en het productieniveau. Deze informatie over het SCK-risico voor elk koppel was vervolgens geïnccludeerd door trunceren van de prior verdeling voor de random effects in het voorspellingmodel. De resultaten lieten geen verbetering zien in de voorspelling op individueel niveau wanneer expert meningen werd meegenomen. Er was enige verbetering van de voorspelling op koppelniveau, maar dit werd niet als klinisch belangrijk beschouwd. Deze klinische studie toonde aan dat het in de praktijk methodologisch mogelijk was om

bedrijfsinformatie als prior op te nemen voor de random effects in een voorspellingsmodel. De Bayesiaanse benadering van voorspellingsmodellering kan veelbelovend zijn voor diagnose in koppels wanneer de variantie tussen koppels relatief groot is en waar alle koppels zieke dieren bevatten die worden bestudeerd.

De mogelijkheid om extern bewijs als informatieve priors op te nemen in klinische analyses werd verder onderzocht bij minor species. In de veterinaire epidemiologie worden dierenartsen standaard opgeleid tot cross-species denkers, aangezien bewijs voor specifieke onderzoeksvragen schaars kan zijn. In de klinische praktijk kan het ontnemen van informatie van andere diersoorten op een dagelijkse basis impliciet gebeuren. In hoofdstuk 6 hebben we de Bayesiaanse power prior methode toegepast om relevante cross-species studies samen te voegen na te zijn onderzocht en gewogen door een klinische expert met betrekking tot de relevantie van elke studie voor de betreffende onderzoeksvraag. Een Nederlandse klinische studie bij paarden werd opnieuw geanalyseerd om het effect van orale Glucosamine (GS) en Chondroitin sulphate (CS) op stijve gewrichten bij oudere paarden te evalueren. Onderzoeken die de klinische effectiviteit van deze twee nutraceuticals op stijve gewrichten beoordelen, zijn beperkt en leveren tegenstrijdige resultaten op. In onze studie werd daarom een cross-species systematisch literatuuronderzoek gedaan om bestaande relevante informatie te detecteren. Twintig onderzoeken uitgevoerd bij mensen, honden en paarden kwamen in aanmerking en werden vervolgens gewogen door een paardensdeskundige op hun klinische relevantie voor de onderzoeksvraag van paarden. Negen klinische onderzoeken die de combinatie van GS en CS als behandeling gebruikten, werden in de analyse opgenomen via de power prior. Om alle uitkomsten op dezelfde schaal te krijgen, werden alle ruwe uitkomsten van de paarden en de negen cross-species studies voor de power prior gestandaardiseerd voor de analyse. De resultaten van de analyse lieten geen effect zien van de nutraceuticals. Dit ondersteunde de oorspronkelijke conclusie van de klinische trial bij paarden. Sensitiviteitsanalyses lieten echter gemengde resultaten zien, die de huidige heterogeniteit en onzekerheid over de klinische effectgrootte van GS en CS voor stijve gewrichten bij oudere individuen weerspiegelden. De Bayesiaanse power prior methode toonde de voordelen bij het opnemen van historische relevante studies in de klinische analyse, maar het synthetiseren van cross-species bewijs leverde ook uitdagingen op en heeft verder onderzoek nodig.

Dit proefschrift presenteert de ontwikkeling en evaluatie van verschillende Bayesiaanse methoden in de veterinaire klinische epidemiologie. We toonden aan dat naast de algemeen aanvaarde Bayesiaanse analyse met non-informatieve priors, informatieve priors een nuttig

hulpmiddel kunnen zijn om relevant bewijs op een expliciete en transparante manier te accumuleren in de klinische praktijk van veterinaire epidemiologie.

中文概要

此本论文汇总记录了我们在兽医临床流行病学领域中开发或应用的几种贝叶斯建模方法。在兽医流行病学中,目前大多数贝叶斯应用主要是基于模拟马尔可夫链蒙特卡罗(Markov Chain Monte Carlo; MCMC)在处理复杂问题时的灵活性。这也是我们选择贝叶斯方法的原因之一。再者通过贝叶斯先验分布,我们可以将相关信息和收集的数据结合从而为研究问题提供更多证据。贝叶斯更新也符合循证兽医学(Evidence Based Veterinary Medicine; EBVM)的实践。我们用模拟或者真实世界数据研究了三种贝叶斯方法:第一部分阐述了贝叶斯潜在类别模型在诊断测试特征和识别风险因素中的运用;第二部分介绍了通过先验分布结合专家意见的贝叶斯预测模型;第三部分展示了通过贝叶斯模型在兽医临床分析中结合跨物种数据。

当缺乏可靠参考测试(即黄金标准)时,贝叶斯潜在类别模型通常用于评估不完善测试的诊断性能。我们研究了两种潜在类别方法:经典的 Hui-Walter (HW)模型和逻辑回归(logistic regression; LR)模型。通常 HW 模型会将观察到的个体测试结果整合到总体水平,因此无法完全包括来自较低层的信息。而 LR 模型允许包含分层数据,例如个体水平协变量和群体水平聚类效应。在第二章中,我们用已知真实疾病状态和真实测试属性的模拟研究了这两种方法。模拟的数据参考了牛副结核病的流行特性和测试特征:模拟的个体奶牛来自相同或不同的牛群并聚集在两个地区;模拟的两种测试都具有二元结果且在两个地区具有恒定的测试特征。除了流行特性和测试特征之外,模拟中还添加了一个个体水平的二元风险因素。各种模拟场景的结果表明在没有黄金标准测试的情况下,LR 模型在许多情况下优于 HW 模型。其中结合了群体水平聚类效应的 LR 模型提供了最精确和最小偏差的估计。此外,含有个体水平风险因素的 LR 模型在各种模拟场景中都提供了可靠的风险因素的优势比估计。结果还显示,基于个体水平风险因素对数据进行分区的 HW 模型可能存在问题。因此当收集的数据中有相关的疾病风险因素时,我们建议使用 LR 方法。

为了验证从模拟研究中获得的结论,我们将贝叶斯 HW 和 LR 潜类模型应用于第三章中北爱尔兰牛结核病(bovine tuberculosis; bTB)的真实世界数据。本研究的主要目的是为了更深入地了解单次皮内结核菌素(single intradermal comparative cervical tuberculin; SICCT)测试的敏感性,因为到目前为止文献中使用 HW 模型报告

的 SICCT 测试敏感性显示出了很大的跨度。我们的数据收集自 2015 年所有在北爱尔兰屠宰场屠宰的牛。对于 HW 模型, 北爱尔兰的十个部门兽医办公室 (divisional veterinary office; DVO) 地区被视为具有不同流行率的群体。对于 LR 模型, 我们将个体水平的风险因素 (死亡年龄、性别、从最后一次 SICCT 测试到屠宰的天数和最后一次 SICCT 测试原因) 纳入了最终分析。两种方法得出了相似的 SICCT 测试敏感性和特异性, 但 LR 模型提供了更窄的后验分布 95% 可信区间。此外我们还利用 LR 模型评估了 bTB 与每个风险因素之间的关联, 这对于北爱尔兰 bTB 疾病控制的临床实践具有重要价值。

在兽医流行病学中我们通常使用随机效应模型来考虑聚类结构。然而对于预测模型, 随机效应在模型开发后用以预测未来集群时会被移除。在第四章中我们提供了一种新的贝叶斯方法来保留随机效应并通过随机效应分布包含先验信息以预测未来集群。我们先对研究数据进行了模拟, 然后比较了五个预测模型: 将所有新集群的随机效应设置为零的频率论模型; 将所有新集群的随机效应先验分布设置为不包含信息的贝叶斯模型; 通过随机效应先验分布包含模拟专家意见 (低等信息、中等信息和高度信息) 的贝叶斯模型。我们在不同模拟场景中评估了这五个模型的预测性能, 并探讨了在模型中加入偏离真值的次优专家意见对预测的影响。结果表明, 对随机效应使用非信息先验的贝叶斯模型与去除随机效应的传统频率论模型的表现相似, 而使用专家意见作为信息先验的贝叶斯模型显示出更小的 Brier 值, 更好的整体辨别、整体校准和集群校准。结果还表明, 通过结合更精确的专家意见, 我们可以在个人和集群水平做出更好的预测。当使用次优专家意见作为先验信息时, 在某些情况下预测仍然有所改善。

这种预测方法随后在第五章的临床研究中得到应用。研究前我们猜想通过添加牛群特定的先验信息, 对泌乳早期奶牛亚临床酮症 (subclinical ketosis; SCK) 的预测会得到改善。为了验证这个猜想, 我们使用了 Van der Drift (2012) 等人收集的数据以及他们开发的预测模型。在我们的研究中, 一位牛类疾病专家基于每个农场饲养管理和生产水平的信息提供了他对每个牛群的 SCK 风险评估。他的专家意见随后被纳入了预测模型中随机效应的先验分布。结果显示, 当包括专家意见时, 个体奶牛水平的预测没有改善, 而畜群水平的预测有所改善。这项临床研究表明, 将畜群水平信息加入预测模型中随机效应先验分布在方法上是可行的。当畜群之间的差异相对较大且所有畜群都包含患病动物时, 这种贝叶斯预测模型可用于临床畜群诊断。

我们在少数型动物物种中进一步探讨了将不同证据以贝叶斯信息先验分布的形式纳入临床分析的可能性。在第六章中,我们采用了贝叶斯 power prior 模型来汇总相关的跨物种数据,并重新分析了一项荷兰马临床试验。此临床试验用来评估口服氨基葡萄糖(Glucosamine; GS)和硫酸软骨素(Chondroitin Sulphate; CS)对老年马关节僵硬的作用。现存文献对评估这两种营养保健品对缓解马僵硬关节的临床有效性是有限的。再者有些研究得出了相互矛盾的结论。因此在我们的研究中,我们先进行了系统的跨物种文献搜索。在人类、狗和马中进行的 20 项研究被认为符合条件,随后由一位马类疾病专家将它们与马的临床试验相关性进行筛选和评分。九项使用 GS 和 CS 联合治疗的临床研究被采纳并以 power prior 的形式被加入临床分析。分析结果表明 GS 和 CS 对老年关节僵硬没有缓解作用,这一结果支持了荷兰马临床试验的原始结论。然而敏感度分析得出了不同的结果,这体现了 GS 和 CS 对老年僵硬关节的临床效果大小的异质性和不确定性。贝叶斯 power prior 方法展示了将历史文献相关数据纳入临床试验分析的可行性,但跨物种证据在临床兽医实践的应用需要进一步研究。

本论文记述了几种贝叶斯方法在兽医临床流行病学中的开发应用和评估。我们发现兽医流行病学临床实践中,除了广为接受的非信息先验贝叶斯分析外,信息先验贝叶斯分析也是有效的工具。

Acknowledgements

Every time I look back at the years working at the departments Methodology & Statistics and Farm Animal Health, I am profoundly grateful and feel truly lucky to have had the opportunity to know and learn from wonderful people.

First of all, I would like to express my deep gratitude to my supervisors, without whom this research project would not be possible. Thank you Mirjam and Irene for your endless patience at each stage of the project. Thank you for granting me the freedom to explore different topics and always being there whenever I needed feedback. Thank you for your broad expertise as well as your open-mindedness for new ideas. I am really grateful for your understanding, honesty and compassion regarding my procrastination and thank you for all those positive reminders that helped me get through difficult days.

I would also like to thank my co-authors Gerrit, Rolf, Maria, Charlotte, Ruurd, Arjan, Fraser, Saskia and Wim. It was a great pleasure working with you. Thank you for your valuable input and all the inspiring discussions.

Now and then I still miss the daily coffee/tea breaks at DGK and the Monday morning gatherings at MS. I miss the Epi-meetings and the Bayesian meetings. It was always pleasant talking to colleagues who are passionate about work and life in general. Thank you all for the fun and inspiration I had throughout the years at Utrecht University. And Hans and Jan, thank you also for allowing me to observe and learn how to give consultations to veterinary professionals.

I owe sincere thanks to my family as well. 感谢父母和妹妹的理解和支持。Thank you my mother-in-law, father-in-law, sister-in-law and Margreeth for your continuous emotional support. My dearest Daniel and Noa, I am so proud and grateful to be your mother. Thank you for your unconditional love. And Mark, thank you for being my rock, the fantastic father of my children, a critical yet empathetic listener to all my thoughts and concerns. Life without you is unimaginable.

Curriculum Vitae

Haifang Ni was born on November 21, 1984 in Hangzhou, China. She finished her high school in 2003 and studied subsequently Economics. She came to The Netherlands in 2006 and completed her bachelor in Psychology at Leiden University. In 2012, she gained her MSc in Methodology and Statistics. She started the PhD project on Bayesian methods in veterinary epidemiology in 2013 under the supervision of prof. dr. Mirjam Nielen and prof. dr. Irene Klugkist at Utrecht University.

Currently, Haifang works as a senior data analyst at Nationale-Nederlanden.

