University of Veterinary Medicine

Doctoral School of Veterinary Science

# Conditional and marginal modelling in animal health using bovine paratuberculosis as an example

PhD thesis

**Katalin Veres**

**2025**

**Supervisors**

....................................................

Dr. Zsolt Lang, Senior research fellow, PhD

Department of Biostatistics, Institute of Economics and Biostatistics, University of Veterinary Medicine Budapest

....................................................

Prof. Dr. László Ózsvári, Full professor, PhD

Department of Veterinary Forensics and Economics, Institute of Economics and Biostatistics, University of Veterinary Medicine Budapest

**Committee members**

Dr. Andrea Harnos, Associate professor, PhD

Department of Biostatistics, Institute of Economics and Biostatistics, University of Veterinary Medicine Budapest

Dr. Jenő Reiczigel, Full professor, DSc

Department of Biostatistics, Institute of Economics and Biostatistics, University of Veterinary Medicine Budapest

....................................................

Katalin Veres

Copy Nr.:…..

# Table of contents

## List of abbreviations

| | |
|---|---|
| AGID | agar gel immunodiffusion assay |
| aHR | adjusted hazard ratio |
| AP | apparent prevalence |
| BF | Bayes factor |
| CFT | complement fixation test |
| CI | confidence interval |
| CrI | credible interval |
| CWHP | conditional within-herd prevalence |
| DMMM | distribution based marginalized multilevel model |
| ELISA | enzyme-linked immunosorbent assay |
| GEE | generalized estimating equations |
| HP | herd-level prevalence |
| HR | hazard ratio |
| HTP | herd-level true prevalence |
| HZ | herpes zoster |
| IPW | inverse probability weighting |
| MAD | median absolute difference |
| MAP | mycobacterium avium subspecies paratuberculosis |
| MCMC | Markov chain Monte Carlo |
| MMM | marginalized multilevel model |
| NLMM | non-linear mixed model |
| OR | odds ratio |
| PCR | polymerase chain reaction |
| PP | posterior probability of infection of the herds |
| PTBC | paratuberculosis |
| SD | standard deviation |
| Se | sensitivity |
| Sp | specificity |
| TP | true prevalence |

# 1 Summary

This research is a collection of methodological studies to model infectious diseases, motivated and illustrated mainly by problems linked to bovine paratuberculosis. Bovine paratuberculosis (PTBC), caused by Mycobacterium avium subsp. paratuberculosis (MAP) negatively impacts milk yield, health and reproductive parameters on both individual and herd level. PTBC is spread worldwide and causes substantial economic losses. Accurate information on disease prevalence is crucial to choose and implement appropriate control measures, to reduce harm and limit spread.

There is no standard method of screening for PTBC, which makes it challenging to detect the infection and to quantify the proportion of infected animals on a farm. Due to false negative and false positive test results, there can be marked differences between apparent prevalence (AP) calculated from diagnostic test results and the true, underlying disease prevalence (TP). Relying solely on diagnostic tests can be misleading and leads to flawed management decisions. To determine more accurately the level of infection, the apparent prevalence needs to be corrected for test sensitivity, specificity and other unit-specific data. Legacy methods use hierarchical models and Bayesian methodology to estimate the mean herd-level true prevalence (HTP) or the conditional within-herd true prevalence (CWHP) in a region but fail to consider the intrinsically different animal subgroups within herds. My aim was to fill this gap by introducing and describing a modeling framework which provides relevant herd- and subgroup-level prevalence estimates.

Based on legacy methodology, I introduced a novel framework for hierarchical data; the Distribution based Marginalized multilevel models (DMMM). DMMMs are a special case of Marginalized multilevel models (MMM). MMMs allow to model conditional and marginal means within the same framework. Additionally, in DMMMs, the link function for the conditional means is chosen so that the probability distribution of these means belongs to a given parametric family. This creates a natural link between the conditional and the marginal components. The model incorporates a Gaussian copula construction to model the joint distribution of subgroup means. This approach allows for simple and flexible control over both the marginal and conditional mean and dispersion measures. Besides allowing for a direct, distributional representation of subgroup level prevalences, this approach is also convenient to use as a basis for the calculation of further measures.

The new framework was applied to estimate the true prevalence of PTBC in dairy herds in Hungary. I used a Bayesian approach to incorporate both current data and prior knowledge

into the final estimates. The DMMM framework allowed modelling groups of primiparous and multiparous cows as correlated subgroups. I estimated both the mean and median conditional within-herd true prevalence (CWHP) and calculated other measures characterizing the interrelation of the two subgroups. The use of the new methodology along with a more recent and uniquely large dataset and historical priors confirmed and also refined the results of the latest PTBC screening in Hungary.

Nowadays, Bayesian methodology is widely accepted in veterinary science to model the prevalence of infectious diseases. I gave the Hungarian veterinary community an overview of how Bayesian methodology works and introduced its key concepts. The concept, method and interpretation of the results were illustrated by modelling the within-herd prevalence of PTBC infection in individual dairy cattle herds. I built our model on the national results incorporated as prior distributions. I provided international context for the analysis by integrating prior information from diverse regional studies together with region specific synthetic data. I also set up an easy-to-run applet providing prevalence estimates and visuals for practitioners.

Understanding the difference between the apparent and true prevalence and estimating the latter is essential in the quantitative analysis of infectious diseases. Our novel methodology, the DMMM, in synergy with a Bayesian approach, provided relevant herd- and subgroup-level prevalence estimates which help practitioners better assess the infection status and shaping adequate control measures.

Survival analysis is a method to analyze time-to event data. It has its roots in human medicine and evolved in parallel with the proliferation of patient registries, which are rich data sources ready to be exploited. With technological advances, tracking data is now also available in veterinary medicine, it is therefore important to familiarize veterinarians with this versatile methodology. I presented the basic concepts and methods of survival analysis to Hungarian veterinary community, along with examples analyzing the impact of PTBC infection on culling times in dairy cows using conditional and marginal models. To pair my doctoral work to my research in human epidemiology, I co-authored a human study where I analyzed the impact of herpes zoster on dementia using survival analysis. The distinguishing feature of this study was the use of a special conditional modeling approach which could be adapted to veterinary problems. As a sequel to this study, I extended the analysis of the impact of PTBC infection on culling times using the above approach, which represents a novel contribution to veterinary research.

# 2 Introduction and aims of the study

## 2.1 Significance of the study

To assess the prevalence of infectious diseases various diagnostic tests are used. These tests are used to calculate the apparent prevalence, i.e. the proportion of test-positive cases, which may differ greatly from the true prevalence due to the possible occurrences of false positive and false negative cases. Using the apparent prevalence (AP) as base for herd management decisions can be highly misleading. However, this measure can be corrected using the method of Rogan and Gladen [1] to obtain the true underlying prevalence.

Animal groups as economic units imply a hierarchical data structure. Both conditional and marginal models are available to accommodate hierarchical data. Marginalized multilevel models (MMM) combine the advantages of these two approaches, while allowing for a likelihood-based estimation [2].

The true prevalence of the investigated disease can be estimated using a Bayesian approach. Hanson et al. [3] proposed a hierarchical model to estimate the true infection prevalence in the herds of a region. Branscum introduced a hierarchical model to estimate the mean herd-level true prevalence (HTP) of a region, which is a proportion of herds with at least one infected cow. To date, a large number of studies estimated the HTP and the conditional within-herd true prevalence (CWHP) of different infectious diseases using the above-mentioned frameworks. To my knowledge there was no method that would have allowed researchers to model animal subgroups as correlated factors providing both conditional and marginal estimates.

Thus, the aim of the research was to develop biostatistical models providing new insight into the interrelation of subgroups. I introduce a new framework based on legacy methods, the DMMM, and illustrate its merits by the application to PTBC prevalence modelling in Hungarian dairy cattle herds. To further broaden the modelling toolbox for veterinary data, I aimed to adapt a survival analysis approach validated in human epidemiological research. The adapted methodology has legitimacy in the veterinary field due to the growing numbers of so far unused electronic data records with time-to-event data.

As illustration for the methodological novelties, I used data on PTBC infection of dairy cows. PTBC, caused by Mycobacterium avium subsp. paratuberculosis (MAP), is a globally widespread cattle disease. The true prevalence of PTBC infection in Hungary was estimated only once by Ózsvári, Lang et al. [4] on a limited dataset. The results indicated a high herd-level (89.1 %) and a moderately high median animal-level true prevalence (primipara: 4.4 %, multipara:10.3 %). The trend of increasing herd sizes globally [5] and the parallel increase in PTBC risk [6–8] highlight the need for further research in the field.

## 2.2 Aims of the study

The aims of this study were:

- to develop biostatistical models providing easily interpretable marginal and conditional estimates for subgroup means and new insight into the interrelation of subgroups,

- to illustrate the use of the new framework by modeling the prevalence of PTBC infection in Hungarian dairy cattle herds, and to consolidate previous findings using a recent, extensive dataset and historical priors,

- to model the PTBC prevalence of individual herds both in Hungary and internationally and providing rule of thumb for the simple estimation of true prevalence.

- to explore a conditional modelling approach in survival analysis from the field of human epidemiology which has not been previously applied in veterinary science and adapt it to analyze the impact of PTBC seropositivity on culling times in dairy cows.

# 3 Literature review

## 3.1 The notion of prevalence

The prevalence of an infectious disease is a fundamental measure in epidemiology. The frequency or proportion of affected animals reflects the burden of the disease in a population, at a given time and place [9]. This measure also serves as a base for estimating the economic impact of the disease and provides context for appropriate decision making to control and monitor the infection.

A device that reduces the uncertainty about the presence of disease is called a diagnostic test [10]. The *sensitivity* ($Se$) of a diagnostic test is the probability that a subject with the disease will test positive, whilst the *specificity ($Sp$)* is the probability that the test will give a negative result for a subject without the condition. A test is *imperfect* if either the sensitivity or the specificity is less than one. In practice, all diagnostic tests are imperfect. Imperfect tests are prone to both false positive and false negative results, with probabilities $1-\mathrm{Sp}$ and $1-\mathrm{Se}$, respectively [11]. The proportion of positive cases measured by such tests is called apparent prevalence ($\mathrm{AP}$) which can differ markedly from the true prevalence ($\mathrm{TP}$).

## 3.2 Estimation of prevalence

When a diagnostic test is applied to a random sample of subjects, the standard way to estimate prevalence is

$$\hat{\pi} = \frac{\text{number of positive cases in the sample}}{\text{number of all cases in the sample}}. \tag{3.1}$$

Due to the imperfection of the test, the estimator (3.1) may be biased [11]. The relationship between the apparent prevalence (expectation of $\hat{\pi}$) and the true prevalence ($\pi$) is linear, and given by the following equation [1]:

$$AP = E(\hat{\pi}) = Se * \pi + (1 - Sp) * (1 - \pi). \tag{3.2}$$

To express the accuracy of the estimate, authors usually provide confidence or credible intervals (CI, CrI). CIs are often provided for the unadjusted measure, the apparent prevalence, implicitly assuming that both the sensitivity and the specificity of the test is 1. The Wald method, which is the simplest asymptotic method, is widely used. Wilson's score method or the exact Clopper–Pearson method are also popular for the calculation of the confidence limits [12]. A naive method is adjusting the apparent prevalence using the

Rogan-Gladen formula [1], and then calculating a CI with the adjusted prevalence as if that were observed directly. However, this procedure provides incorrect CIs, with actual coverage at a nominal 95% level can be as low as 60%, even for large samples [12].

There are correct methods to construct CIs adjusted for sensitivity and specificity. An asymptotic method based on Wald method is presented by Rogan and Gladen [1]. Reiczigel et al. [12] propose a way to calculate exact confidence limits assuming sensitivity and specificity are known. This method is applied for example by Lang et al. [13] for estimating the true prevalence of BHV-1 infection. Lang and Reiczigel [14, 15] present a method of constructing approximate confidence intervals for the true prevalence when sensitivity and specificity are inferred from independent samples. The risk ratio, the risk of disease in a study population relative to a reference population, should also be estimated accounting for the imperfect diagnostic test [16].

Since estimates of diagnostic sensitivity and specificity may vary among populations and subpopulations of animals depending on the distribution of influential covariates [10], complex models are needed to match the target conditions and test methods. For example, in the case of the milk ELISA test for the diagnostic of PTBC, Meyer et al. suggested the use of a model with age-specific specificity and constant sensitivity. The target condition was the MAP-infected cow in which a humoral response would become detectable during the economic life of the cow [17]. They considered a non-linear logistic regression model of Nielsen [18]:

$$logit\big(Se(t)\big) = a - b * e^{-c*t}. \quad (3.3)$$

The parameters a, b and c were estimated from a large number of samples. Here, a is the upper limit of the logit of the sensitivity when age increases, b is a scaling factor, c is the coefficient for the decay of age effect, and $t$ is the age of the animals at the time of diagnosis, measured in years. The age-specific sensitivity takes into account both the probability that an infected cow will excrete detectable levels of MAP antibodies in the milk and the intrinsic characteristics of the ELISA milk test.

## 3.3 Bayesian models for prevalence estimation

Bayesian models have become widespread in veterinary science in recent years [19, 20]. The main appeal of this approach is that both current data and prior knowledge – such as the result of previous studies, experience, expert opinions, or empirical evidence – can be incorporated into the final estimates in the form of priors. Priors are probability distributions formed based on prior knowledge and are updated in the light of new data to produce the final estimates. Moreover, these models are able to accommodate small sample sizes and account for the misclassification of animals arising from the imperfect accuracy of existing diagnostic tests [21–23], making them an appropriate tool for modeling disease prevalence. Bayesian methods are widely used to estimate animal- and herd-level prevalences [24, 25] and to substantiate freedom from infection [26].

First in its field, Hanson et al. [25] proposed two Bayesian models to estimate the infection prevalence in a single herd based on the result of an imperfect diagnostic test. The first model assumes that screened animals are drawn from a finite population, thus the distribution of the number of truly infected subjects, conditional on the number of infected animals in the whole herd, follows a hypergeometric distribution. A simpler approach would be assuming independent infection status; thus, the number of test-positive diseased and test-positive non-diseased subjects would follow binomial distribution regardless of the herd size. The model consists of only three parameters: the $Se$ and $Sp$ of the test and the true prevalence $\pi$. The prior probability that the herd is infected is provided by an expert – e.g. as the proportion of infected herds among herds with similar size and similar herd management practices in the area. Alternatively, one can use a beta distributed prior to reflect a plausible range for the infection probability. The uncertainty of sensitivity and specificity is also modeled using independent beta distributed priors.

Hanson's second model assumes that the herd is infinite but provides an algorithmically more efficient and yet accurate approximation to the first model when the herd size is much larger relative to the number sampled. In this case the number of test-positive animals (*y*) follows an approximately binomial distribution:

$$y|\pi, Se, Sp \sim Bin\big(n, \pi Se + (1-\pi)(1-Sp)\big). \tag{3.4}$$

To set up the Bayesian model, we need to define prior distributions for model parameters. As previously, $Se$ and $Sp$ are modelled using independent beta priors. For infection prevalence, Hanson proposes to use a mixture distribution:

$$\pi \sim Beta(a_\pi, b_\pi), \qquad with\ probaility\ \tau\ ,$$

$$\pi = 0 \qquad with\ probability\ (1-\tau), \tag{3.5}$$

$\tau$ being the probability that the herd is infected, i.e. there is at least one infected cow in the herd. $\tau$ may be modelled using beta distribution or set as constant. Beta distribution is also suitable to model infection prevalence provided that $P(\pi = 0) = 0$ [27].

To estimate the herd-level prevalences ($\mathrm{HTP}$) in a region, Branscum et al. [24] proposed a hierarchical model. This model allows to estimate the prevalence distribution of herds, the probability that the region is infection-free and the predicted probability that a randomly selected herd in the region is infection-free.

Branscum et al. [24] considers a two-stage cluster-sampling design, i.e. $K$ herds are first sampled randomly from the region, and subsequently $n_t$ animals are sampled randomly from herd $\mathrm{t}$.

It is assumed that the number of test-positive animals from each herd $y_1, y_2, \dots, y_K$ are independent and follow a binomial distribution:

$$y_t|\,\pi_t, Se, Sp \sim Bin(n_t, \pi_t Se + (1-\pi_t)(1-Sp)). \tag{3.6}$$

As usual, independent beta priors are used for $\mathrm{Se}$ and $\mathrm{Sp}$. Similarly to the method used for single herd, prevalences are modelled with a mixture distribution:

$$\pi_t|\mu, \psi \sim Beta\big(\mu\psi, \psi(1-\mu)\big) \qquad with\ probaility\ \tau,$$

$$\pi_t = 0 \qquad with\ probability\ (1-\tau). \tag{3.7}$$

$$\tau \sim Beta(a_\tau, b_\tau) \qquad with\ probability\ \gamma,$$

$$\tau = 0 \qquad with\ probability\ 1-\gamma. \tag{3.8}$$

$$\mu \sim Beta\big(a_\mu, b_\mu\big). \tag{3.9}$$

$$\psi \sim Gamma(a_\psi, b_\psi), \tag{3.10}$$

where $\mu$ is the average prevalence in the population and $\psi$ is its variability measure, so that the variance of the beta distribution is $\sigma^2 = \mu(1-\mu)/(\psi+1)$. Larger values of $\psi$ correspond to less heterogeneity in the data. This parameterization was proposed by Hanson et al. [25].

McAloon et al. [28] applied this methodology to estimate $\mathrm{HTP}$ of bovine PTBC in Ireland. They used slightly different notations, introducing the term conditional-within herd prevalence ($\mathrm{CWHP}$) to denote the animal-level prevalence in herd *i* (*i* = 1 to *n*, *n* is the number of herds sampled in the region), conditional on the herd being infected. Their main model is as follows:

$$npos_i \sim Bin(AP_i, nherd_i), \tag{3.11}$$

where $nherd_i$ is the count of animals in herd *i*, $npos_i$ is the count of test-positive animals in herd *i* and $AP_i$ is the probability that an animal tests positive in herd *i*.

$$AP_i = \pi_i \times Se + (1 - \pi_i) \times (1 - Sp), \tag{3.12}$$

$\pi_i$ denotes the animal-level true prevalence in herd *i*, which is represented as the product of herd-level true prevalence ($HTP_i$ with mean $\mu$) and the conditional-within herd prevalence ($CWHP_i$). The latter is the animal-level prevalence in herd *i,* conditional on the herd being infected and

$$\pi_i = HTP_i \times CWHP_i. \tag{3.13}$$

McAloon et al. [28] defined a Bernoulli prior for $HTP_i$, beta priors for $\mathrm{Se}$, $\mathrm{Sp}$, $CWHP_i$ and beta hyperprior on the $\mu$ parameter of the Bernoulli distribution.

Branscum et al. [29] elaborated on Bayesian beta regression models allowing for incorporating covariate information while estimating prevalence. Their model is

$$y_i|\mu_i, \gamma \sim Beta(\mu_i\gamma, \gamma(1 - \mu_i)), \tag{3.14}$$

$$\mu_i \equiv \mu(\boldsymbol{x}_i) = F(\boldsymbol{x}_i^T\boldsymbol{\beta}), \tag{3.15}$$

where $\boldsymbol{y_i}$ are the outcome variables in the model, $\boldsymbol{\beta}$ is the vector of regression coefficients, $\boldsymbol{x}_i$ are the covariates. The prior distributions of $\boldsymbol{\beta}$ and $\gamma$ are independent. The beta distribution is parameterized in terms of its mean $\mu$ and precision $\gamma$. The inverse-link function *F* can be any continuous cumulative distribution function. This model also allows for non-constant variance.

As the Bayesian methodology for prevalence estimation has become more sophisticated, veterinary applications also emerged, with numerous publications estimating the prevalence of a wide range of infectious diseases worldwide including PTBC. A frequentist approach was still used to model the PTBC prevalence in dairy herds in the Netherlands [30] and in a review on prevalence of paratuberculosis in European farm animals [20]. Since 2011, papers implementing the Bayesian latent class framework proposed by Branscum et al. [24] were abundantly published [31–34]. After the publication of McAloon et al. [28], applications took up his methodology [4, 35, 36].

Ózsvári, Lang, et al. [4] were among the first to apply this model to dairy herds in Hungary. As an additional feature, they introduced herd level random effects using a logistic model for CWHP. The model treated primiparous and multiparous cows as separate subgroups.

## 3.4 Marginalized multilevel models

Coupled with Bayesian methods, marginalized multilevel models (MMM) provide an appropriate framework for prevalence modeling. Heagerty and Zeger [2] proposed the marginalized multilevel model (MMM) framework to analyze clustered data. $\boldsymbol{Y}_{ij}$ is the response variable for the *j*-th observation in the *i*-th cluster and $\boldsymbol{X}_i$ is the vector of cluster-specific covariates. In this model, the population mean, which is the marginal expectation $\mu_{ij}^m = E(\boldsymbol{Y}_{ij}|\boldsymbol{X}_i)$ is directly linked to explanatory variables $x_{ij}$ using a generalized linear model (GLM). *g* is the link function for the marginal mean and $\boldsymbol{\alpha^m}$ is the vector of marginal parameters. The so-called "mean model" component is as follows:

$$\mathrm{g}(\mu_{ij}^m) = \boldsymbol{x}_{ij}\boldsymbol{\alpha^m}. \qquad \text{(Mean model, 3.16)}$$

The second component, the "conditional" or "association" model describes the within-cluster associations using a non-linear mixed model (NLMM) conditioning on a cluster-specific vector of covariates $\boldsymbol{X}_i$ and on a cluster-specific vector of latent variables $\boldsymbol{a_i}$ with covariance matrix ***D***. The conditional expectation $\mu_{ij}^c = E(Y_{ij}|\boldsymbol{X}_i, \boldsymbol{a_i})$ is linked to $\Delta_{ij} + z_{ij}\boldsymbol{a_i}$ with the same link function $g$ as in the mean model. Here $z_{ij}$ are known covariates associated with the random effect $\boldsymbol{a_i}$ . $\Delta_{ij}$ depends both on covariates, marginal parameters and random effects and connects the marginal and the conditional models. The association is described by:

$$\mathrm{g}(\mu_{ij}^c) = \Delta_{ij} + z_{ij}\boldsymbol{a_i}. \qquad \text{(Association model, 3.17)}$$

The model assumes that the distribution of $\boldsymbol{a_i}$ is completely specified by the parameter α, and that the distribution of conditional observations $Y_{ij}^c = (Y_{ij}|\boldsymbol{X}_i, \boldsymbol{a_i})$ follows an exponential distribution family $F_{Y^c}$ with mean and dispersion parameters $\mu_{ij}^c$ and $v$.

$\boldsymbol{a_i} \sim F_\alpha(0, \boldsymbol{D})$ (Latent effects distribution, 3.18)

$Y_{ij}^c \sim F_{Y^c}(\mu_{ij}^c, \boldsymbol{v})$ (Conditional Response Distribution, 3.19)

Griswold et al. (2013) relaxed the usual MMM assumption that marginal and conditional link functions are identical. The new framework also allows for nonlinear effects both for the mean and for the conditional components.

$\mu^m = h_m(\boldsymbol{\theta}^{\mathbf{m}})$ (Mean model, 3.20)

$\mu^c = h_c(\boldsymbol{\theta}^{\mathbf{m}}, \boldsymbol{\psi}, \boldsymbol{a})$ (Association model, 3.21)

$\boldsymbol{a} \sim F_{\boldsymbol{a}}(\boldsymbol{\psi})$ (Latent effects distribution, 3.22)

$\boldsymbol{Y}^c \sim F_{\boldsymbol{Y}^c}(\mu^c, \boldsymbol{v})$ (Conditional Response Distribution, 3.23)

In this model $h_m$ and $h_c$ are the inverse-link functions (the inverse of original *g*) for the marginal mean and the conditional mean, respectively, $\boldsymbol{\theta}^{\mathbf{m}}$ is a parameter vector upon which the marginal mean depends, $\boldsymbol{\psi}$ is the parameter vector for latent effect distribution.

This reformulation sheds light on the connection between marginal and conditional models and allows us to construct marginal models through conditional specifications. The computing solutions already in place used to estimate mixed models are also suitable for estimation of the reformulated model.

MMMs incorporate the familiarity, robustness and interpretability of a marginal mean model, while retaining flexible dependence structures and likelihood-based inference from conditional models. However, it is important to note that, as the inverse link of the conditional mean $\mu^c$ is strictly monotone increasing, e.g. as in classical generalized linear mixed models, the rescaled fixed effects $h_c(\boldsymbol{\theta}^{\mathbf{m}}, \boldsymbol{\psi}, (0,0,\dots,0)^T)$ is equal to the marginal median, rather than to the marginal mean.

### 3.5 Survival analysis

Survival analysis is a widely used method, the foundations of which were laid in the mid-1950s. Originally, as the name suggests, it was developed to analyze the risk of mortality, but it can be used to study the time to the occurrence and the risk of any events. Time-to-event data are data series in which each observation unit (e.g. patient) has a follow-up time and a status indicator that indicates whether the event of interest occurred by the end of the follow-up time. A distinct feature of time-to-event data is that in some of the observations the event does not occur during the studied period or during the period when the individual can be observed. In such cases, the exact time of the event is unknown and only a lower bound for the time of occurrence can be given. This feature of survival data is called censoring, and such data are called censored data. Unlike standard regression models, survival analysis accounts for censored data. Two key functions describe time-to-event data: the survival function which represents the probability that an individual survives beyond a given point in time; and the hazard function representing the instantaneous risk of the event occurring at a time point t, given that the individual has survived up to this time [38–40].

The Kaplan-Meier method [41, 42] or life tables [43, 44] can be used to estimate the survival function with a confidence interval. The Kaplan-Meier curve can be thought of as a visual summary of the data. It can be used, for example, to calculate median survival times, probabilities of survival over time, or to visually compare the survival experience of experimental groups. The log-rank and Gehan-Wilcoxon test [38, 42, 45, 46] serve to compare the survival experience of two or more groups.

Besides non-parametric methods, various semi-parametric and parametric methods are used to explore how different factors affect survival times. In parametric analysis, survival times are assumed to follow a parametric distribution, from which the survival function can be derived. The most commonly used parametric models for survival time assume exponential, Weibull or lognormal distribution [47]. The Cox proportional hazard model is a semi-parametric model, that starts from a parametric regression model but makes no assumptions about the distribution of survival times. It is used to estimate the relative risk (hazard ratio) of an event in two or more groups. In contrast to the log-rank test, it not only provides information on the fact that survival differs between groups but also provides an estimate of the magnitude and direction of the effects. This model assumes that the hazard ratio remains constant over time [48].

## 3.6 Bovine paratuberculosis

PTBC, caused by Mycobacterium avium subsp. paratuberculosis (MAP), is a disease that affects ruminants, widespread both worldwide and in Hungary. The infection affects more than 50% of dairy farms globally [49]. The average animal-level prevalence ranges from 2.7% to 15% in farmed animals in Europe [20], while in Hungary, Ózsvári, Lang, et al. [4] reported an animal-level true prevalence of 4.4% and 10.3% in primiparous and multiparous cows, respectively. The average apparent animal-level PTBC prevalence in

Hungary based on serum ELISA was estimated to 8.3% [50]. A recent study showed that on Hungarian farms with no measures in place to control paratuberculosis, the seroprevalence increased by 12.1% per year on average [51].

### 3.6.1 Epidemiology of PTBC

The infection typically progresses slowly, has a long incubation period, with around 95% of infected cattle in subclinical stage and only 5% displaying clinical signs. Both the clinical and subclinical stages of PTBC infection cause substantial economic losses, but the subclinical stage is economically more influential, as at this stage the animals do not show obvious symptoms but the negative impact on production is already present [52]. Infection with MAP negatively affects the slaughter weight of cattle and, through the deterioration of meat quality, the slaughter value [53]. MAP seropositive cattle also have significantly higher culling and mortality rates compared to seronegative cattle [52, 54–56]. The infection has a detrimental effect on milk yield and significantly increases the somatic cell count [52, 55, 57–59]. Additionally, the reproductive performance of MAP seropositive cattle is significantly worse than that of their MAP-negative counterparts, with an increase in time from calving to conception, time between calving, and a decrease in the insemination index [52, 55, 58–60]. In a large-scale dairy farm, the fertility index of MAP seropositive cows increased by 2 compared to MAP-negative cows, a difference that cost the farm millions of euros per year in additional costs due to the increased amount of semen used [52]. The economic loss due to MAP infection is estimated at an annual $198 million in the US dairy industry, $75 million in Germany and $56 million in France [61]. Trade restrictions, aimed at limiting the transmission of the disease, negatively impact international and national animal trade. Furthermore, several studies have shown a possible association between MAP and Crohn´s disease and other autoimmune diseases in humans [62–64]. The trend of increasing herd sizes globally [5] and the parallel increase in PTBC risk [6–8] highlight the need for further research in the field.

The main route of transmission of the infection is fecal-oral, but in-utero infection and infection by semen can also happen [65]. Typically, the infection starts in the early neonatal period via colostrum, bulk colostrum or from the contaminated environment [66, 67]. The resistance to MAP progression increases with age [68]. It is hypothesized that in young calves the openness of the gut is a portal for the bacteria, and the immaturity of the immune system also contributes to the high susceptibility to the disease. However, cattle can also become infected at a later age due to high infection pressure [67]. In addition to the age of the cows, genetics is also likely to influence susceptibility to MAP infection [69]. Evidence suggests that the main portal of entry for the pathogen is the ileum. The intrusion triggers a strong cellular immune reaction. Although macrophages attempt to destroy MAP, the bacteria are capable of surviving intracellularly, so this defense mechanism does not always succeed, resulting in infection [70].

Progression of the disease is often associated with a shift from a predominantly cell-mediated immunity to a predominantly humoral immunity [71]. Cell-mediated immune response may be absent in clinical cases. Serum antibodies may also be present in animals that have recovered from the infection [72]. As the disease progresses, infected animals are also increasingly likely to become infectious via shedding of MAP or via in-utero transmission [73].

Four clinically distinct stages of the infection have been described [67, 70, 74]. Stage I is characterized by the absence of clinical signs, no measurable effect on production and no detectable antibody response. Nevertheless, low-level shedding of bacteria may occur, and the infection is already extensively disseminated in tissues already at this stage. Stage II starts when the concentration of MAP in the intestinal tissues is high enough to trigger inflammatory response and consequential histological changes are already present. Animals in this stage have no clinical signs but may shred MAP intermittently in medium concentration, thus representing a source of infection. Milk production and reproductive efficiency are also affected [54, 75], and dissemination of MAP to other organs may occur. Antibody response can be intermittently detectable. The length of this stage can vary highly, depending on several factors (age at infection, dose, re-infections, infection pressure, stressors, nutrition, overall health). The length of the subclinical stage (I, II) varies from 1-10 years. In stage III, the clinical disease begins, animals lose weight, have loose manure, shred high concentration MAP, the infection is disseminated, all diagnostic tests are likely to become positive. Stage IV is the final stage of the disease and is characterized by chronic, profuse diarrhea. Animals are weak and emaciated and their condition is rapidly deteriorating until death from dehydration and cachexia. Animals in stage IV shed massive amounts of MAP in their manure. Transition from stage II to IV can occur rapidly [67, 70, 74].

As MAP is an intracellular pathogen, the role of the antibodies has long been underestimated. Recent studies showed protective role of antibodies in the infection of macrophages in vitro and in vivo in the cattle. In an ileal short-term loop model, antibodies from infected cows reduced MAP invasion and early histological changes [76].

### 3.6.2 Detection of MAP infection

Direct and indirect methods exist to detect MAP infection. Indirect methods, which detect host antibodies rather than the pathogen itself, are well established for testing milk or serum. Milk and serum ELISA are commonly used, as they are cost-effective and provide rapid results. In addition, milk ELISA is non-invasive and can be performed routinely and conveniently [70, 77]. Numerical ELISA test results (S/P ratios) are classified into negative and positive categories according to predefined cut-offs. To avoid excessive false positive results, the cut-off of MAP ELISA is calibrated to favor specificity over sensitivity [78]. MAP ELISA specificity is high (98%-100%), its main disadvantage is the low sensitivity (15% for milk ELISA) in the early stages of the infection due to the delay in the appearance of antibodies [79]. The sensitivity depends on the stage of the infection, on the stage of lactation and on the age of the animal, reaching 90% in clinically affected cattle [80]. Milk ELISA results are correlated with fecal shedding. In a study from Beaver et al. [81], all milk-ELISA positive cows tested positive on fecal culture. The contrary was not true, suggesting the necessity of a lower S/P threshold to identify infectious animals. Bulk milk ELISA value is associated with season of sampling – with peak in the summer and low concentrations in the winter – and with the absence or presence of protocols for managing MAP-positive cows [82].

Direct methods detect the presence of the organism itself. Fecal culture and fecal PCR are examples of routinely used direct tests. These tests allow earlier detection of subclinical cases as fecal shedding usually starts when antibody response is not yet detectable. The sensitivity of these methods is around 60%. Direct methods are more expensive than ELISA, and culture has a much longer turnaround time [70]. Both methods are prone to false positive results from non-infected cows shedding only microorganisms consumed by the animal. This is called passive shredding. This is a rather rare event, the specificity of both culture and PCR is as high as 99% [83].

In terms of susceptibility and detectability of MAP infection, the age of the animal is a crucial factor. Susceptibility to MAP infection decreases with age, and animals infected during adulthood exhibit lower levels of MAP excretion in the faeces. [68]. However, older cows infected as a calf have a higher detectable infection level due to the progression of the disease toward stages with higher antibody production and more intensive shedding [70]. According to Nielsen [18] and Meyer [17] the sensitivity of the tests is age-dependent, i.e. the accuracy of the tests increases with age.

Although humoral immunity can be detected as early as 10 to 17 months after experimental infection, testing before 2 years old was typically not recommended. However, recent studies have contradicted these findings. With experimental diagnostic methods, researchers were able to detect antibodies as soon as 2 weeks to 4 months from the challenge. Using commercial serum ELISA test, 42% of experimentally infected calves had at least one positive test result before 17 months of age. Dose-dependent antibody response was detected soon after inoculation [84],

Given the delayed antibody response and the limited sensitivity of tests, repeated testing is advised over the productive lifetime of the cows. Still, interpreting serial test results has its own challenges [73, 85].

In a recent study using Hidden Markov Chains from Glover at al. [73], four latent stages of infection were identified based on the S/P ratios of milk ELISA tests, one stage with S/P values under the conventional test-positive cutoff value. Animals at this stage are at high risk to transition to the conventionally positive group. Identification of these animals could be used to make control programs more effective.

The choice of the appropriate testing method depends on the infection prevalence of the herd and on the purpose of testing. To estimate the biological burden, environmental testing and pooled fecal culture are appropriate. The diagnostic test of choice in the control program in herds with known high prevalence should be individual ELISA testing, and in low prevalence herds practitioners should choose direct testing methods given that the positive predictive value of ELISA is low in this setting. If the goal is the quick eradication of infection, individual fecal culture or PCR should be performed. A clinical diagnosis of an individual animal in an uninfected herd can be confirmed by necropsy, fecal culture or PCR, while ELISA testing can also be sufficient if the herd is MAP-positive [67, 70, 77, 83]. The control of paratuberculosis in cattle has three main cornerstones: preventing new infections, managing infected cattle and improving resistance. Decreasing the exposure of young replacement stock to MAP is a basic step in prevention. Biosecurity measures are separated calving area and calf rearing, feeding MAP-free colostrum and eventual chemoprophylaxis of neonatal calves. New purchases and bulls should be repeatedly tested for MAP, but introducing new animals to the herd still represents a risk, as testing cannot rule out infection at early stages [50, 70, 86, 87].

The implementation of test and cull strategy depends on the infection prevalence of the herd and the goals of the management. All positive animals may be culled or only the ones considered the most infectious. If positive animals are held in the herd, appropriate biosecurity measures should be implemented. Vaccination and genetic testing may be used to improve resistance. Vaccination has its drawbacks; it does not fully prevent infection and may interfere with diagnostic testing for bovine paratuberculosis. Thus, the use of the vaccines is regulated, its administration can be limited or even prohibited depending on the legal environment [67, 70, 74].

The latent and slow progression of MAP infection and the low sensitivity of diagnostic tests make the diagnosis challenging, leading to a marked difference between apparent and true prevalence levels. Disease frequency is the base measure serving as a starting point for the choice of appropriate testing method, control measures and the assessment of the efficiency of the control program. Using crude apparent prevalence in herd health management can lead to poor decisions and to the failure of control programs. Appropriate statistical methods are needed to adjust apparent prevalence calculated from test results to obtain a prevalence estimate reflecting the true, underlying prevalence of a herd.

# 4 Materials and methods

## 4.1 Datasets

Two datasets were used in the analyses. The *prevalence dataset* was used in the prevalence estimation studies while the *survival dataset* was used in the study assessing the impact of seropositivity on the risk of culling.

### 4.1.1 Prevalence dataset

A nationwide voluntary PTBC testing program started in February 2018 in Hungary, which was set up for dairy farms to provide help for herd managers to reduce the prevalence and impact of PTBC in their herds. In this program state compensation was available for the dairy farms after taking samples and conducting laboratory PTBC tests up to a maximum amount per individual. I analyzed the PTBC screening test results from 2019, covering 58,196 cows in 117 large (number of cows>=100) intensive dairy cattle herds, which represented 24.3% of all dairy cows in Hungary and 26.8% of the performance tested herds. The target population was the set of intensive large dairy cattle herds. The Hungarian dairy cattle industry consists mainly of intensive farms with no or minimal grazing opportunities. Milking cows are housed in confinement barns and generally fed with total mixed ration.

#### 4.1.1.1 Laboratory analyses

PTBC screening tests were performed on all milking cows in the studied herds between January 1 and December 31, 2019. Milk samples, which were collected from individual cows during official performance testing were treated with bronopol as a preservative, transported to the laboratory of Livestock Performance Testing Ltd. (Gödöllő, Hungary), and examined within 48 h of sample collection. Milk diagnostic tests were carried out by Paratuberculosis Screening Antibody Test (IDEXX Laboratories Inc., Westbrook, ME, USA) as per the manufacturer’s instructions. To classify the test results, I used 0.2 and 0.3 as negativity and positivity thresholds for the S/P ratio, respectively. Results within the two thresholds were classified as inconclusive.

#### 4.1.1.2 Description of variables

The results of the 58,769 ELISA tests were then linked to basic cow-level information. Cows with inconclusive ELISA test result (n = 573, 1%) were excluded. The following data were collected for each studied cow: herd ID, cow ID, birth date, number of lactations, date of calving, sampling date, and ELISA test results. Cows were categorized according to parity (primiparous vs. multiparous), according to the herd size (100-300 cows; 301-600 cows; >600 cows) and according to the seven official regions of the country. Herd size categories were designed to include enough herds for stable statistical modelling.

### 4.1.2 Survival dataset

I analyzed the data from [4] to assess the impact of PTBC positivity on the risk of culling of cows in Hungarian-dairy herds. The inclusion criteria for the farms were as follows: (1) computerized on-farm data recording; (2) continuous participation in milk recording at least since January 1, 2014; (3) herd size of more than 250 cows; (4) implementation of a PTBC screening test for all milking cows in the herd in the voluntary PTBC testing program in February and March 2018; and (5) a willingness to provide data to the authors. Altogether, five large dairy herds were included in the study.

PTBC screening tests were performed on all milking cows (n = 4,347) in the studied herds between February 22 and March 22, 2018. Milk samples collected from individual cows during official performance testing were treated with bronopol as a preservative, transported to the laboratory (Livestock Performance Testing Ltd., Gödöllő, Hungary), and examined ˝ within 48 h of sample collection. Milk ELISA tests were used for the detection of antibodies against MAP (IDEXX Paratuberculosis Screening Ab Test, IDEXX Laboratories, Inc., Westbrook, ME, USA). Milk samples with S/P ratios ≤0.2 were classified as negative for MAP antibodies according to the manufacturer. If the S/P ratio was >0.2, but <0.3, the sample was classified as inconclusive, whereas samples with an S/P ratio of ≥0.3 were classified as positive for MAP antibodies. The outcome of the initial screening test involving all milking cows in the herds was used for the classification of the cows. Cows with inconclusive ELISA test results (n = 6) were excluded from the analyses. Individual data for all tested cows were collected from the official milk recording database (Livestock Performance Testing Ltd., Gödöllő, Hungary) [4].

The dataset used for the analysis included herd ID, cow ID, date of birth, date and result of MAP ELISA testing, parity and date of culling (if applicable) of 4,341 cows.

## 4.2 Bayesian modelling in practice

The probabilistic concept of Bayesian statistics is fundamentally different from the concept of traditionally used frequentist statistics. Frequentist statistics estimate the probability of an event based on the proportion of occurrences, i.e. the relative frequency, calculated from many trials. The more times an experiment is performed, the closer the relative frequency of the event approximates the theoretical probability [88]. For example, the more times a regular coin is flipped, the closer the cumulative proportion of heads (and tails) approximates 50%, the theoretical probability (**Figure 1**).

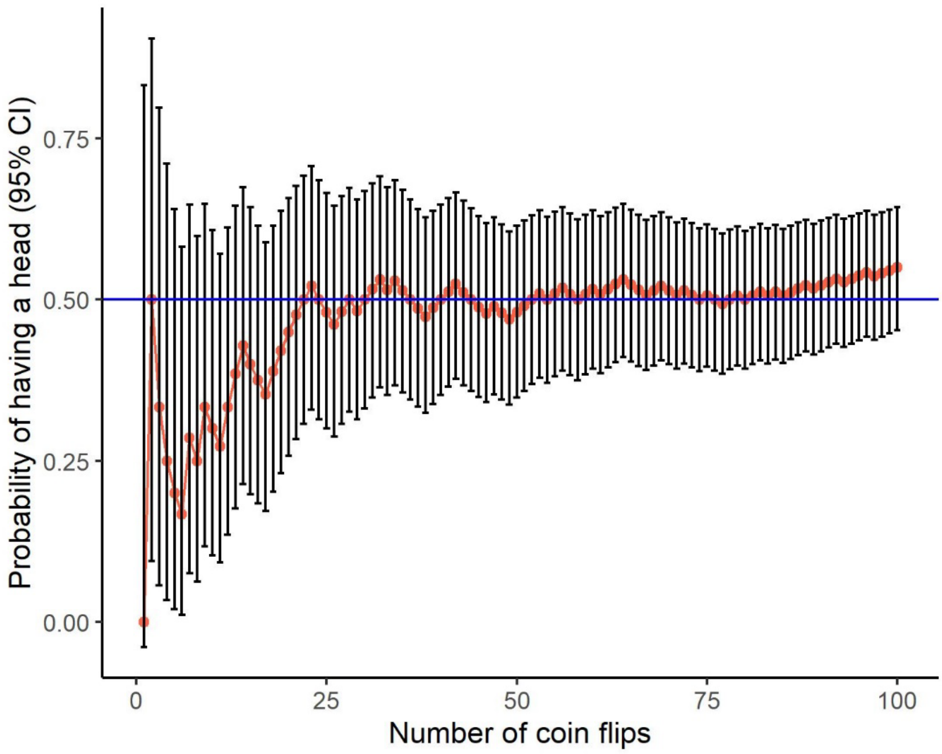



***Figure 1*** *Coin toss with fair coin (author's simulation in R). Red: proportion of head tosses, black: 95% confidence interval.*

In this framework, probability is an objective quantity, independent of any prior experience. After experiencing several consecutive head tosses, we continue to assume that the probability of tossing head on the next trial is 50%. The frequentist estimation of a model parameter aims to calculate this fixed probability.

Besides the data, Bayesian methodology takes into account the prior experience and knowledge when specifying the probability. This process is like the way we think. Sticking with the coin toss example, if we get heads over and over, we start to suspect that the coin is not regular. We update the odds in our minds based on our current experience, and then expect heads with a higher probability [89].

Bayesian modelling requires three basic components: the model, the data and the so-called prior. Prior information condenses prior expectations and knowledge about the process. We distinguish three main types of priors. If we have no useful information to incorporate in the model - for example when exploring a new, unknown field of science - we choose a so-called non-informative prior. This type of prior does not carry information on previous expectations, so the model parameters are estimated solely from available data. If a large amount of data is available from which a sensible estimate can be expected without further assumptions, we may also chose a non-informative prior [22]. The second type of prior is a so-called weakly informative prior. Its role is solely to restrict the range of parameters and to stabilize the model. For example, the sensitivity of a diagnostic procedure can usually be assumed to be at least 50%. When modelling an incurable disease, we can exclude parameter values that would allow a cure. Other restrictions may arise for variance parameters characterizing the heterogeneity of the data. These parameters should generally neither be close to zero nor too large, otherwise the fitted Bayesian model may be unstable. The third type of prior distributions is the informative prior, which allows us to incorporate reliable prior knowledge into the model. For instance, we can use the results of previous studies, our experience, knowledge, or experts' expectations to specify the range of values in which the parameters are likely to fall with high probability [22]. The prior distributions should be constructed so that the probability that a parameter falls within a given interval is equal to the area under the curve of the so-called density function in the specified interval.

The values of this function are called densities, a high density indicating a high probability of a parameter value (**Figure 2**). In the Bayesian model, the prior distribution is refined and updated using the actual data, to obtain the so-called posterior distribution.

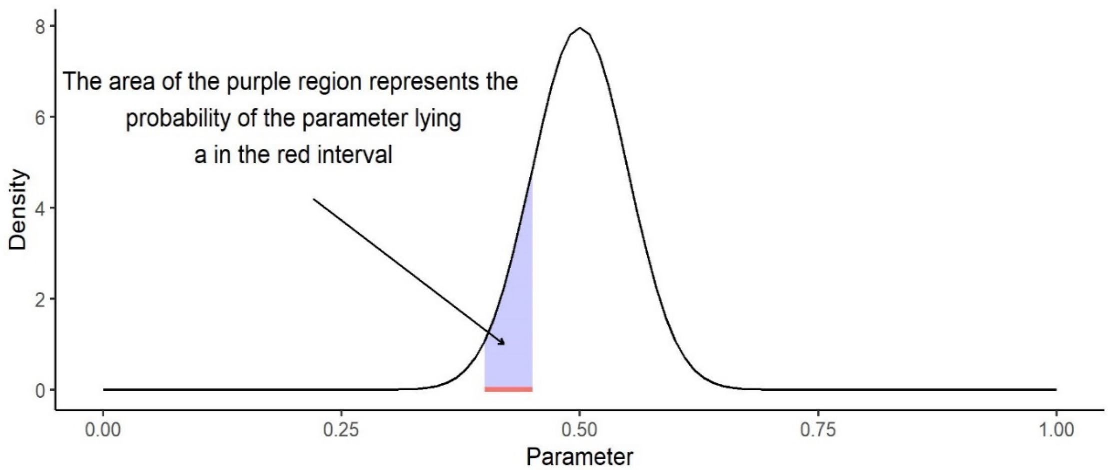



***Figure 2*** *Interpretation of prior and posterior distribution.*

In the case of tossing a regular coin with possible small manufacturing defects or wears, we should choose a prior distribution that gives a chance of about 50%-50% for having head or tail, and sets the chance of a larger deviation from this ratio to small (red curve in **Figure 3**).

The parameter of interest is estimated from the prior distribution using the additional information in the data. The estimate will not be a single number, but a so-called posterior distribution. The probability that the true parameter falls within a given interval is the area under the curve in the specified interval (purple region in **Figure 2**). In the case of a coin toss, the posterior distribution provides information about the probability that the coin is regular and the probability of a given degree of deviation from regularity.

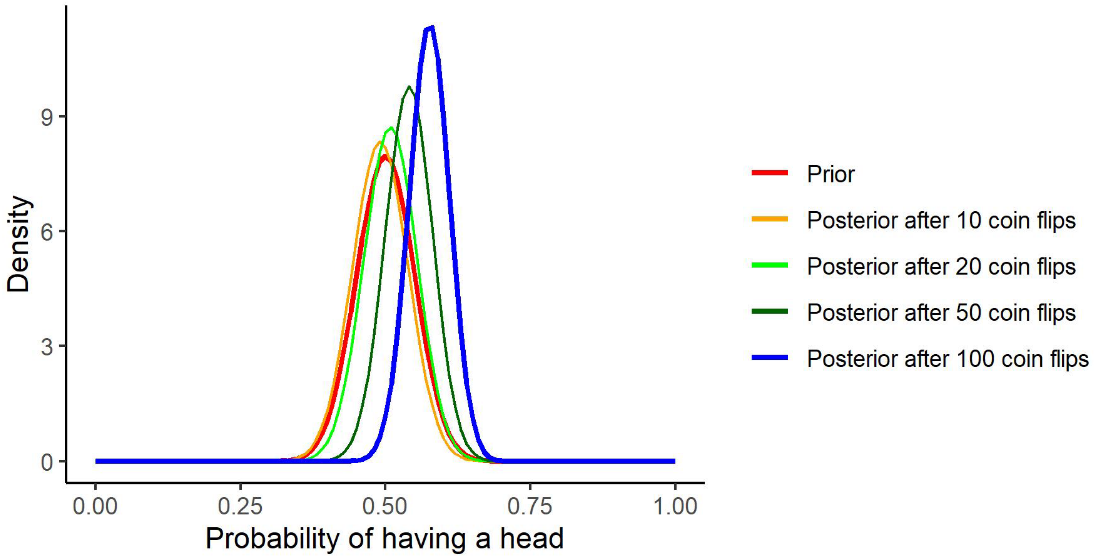



***Figure 3** Bayesian estimation of the probability of head toss in a coin flip - simulation, unfair coin, probability of head: 0.6 (60%). Prior assumption: fair coin.*

**Figure 3** shows that, compared to the prior distribution representing our assumptions, the posterior distribution becomes narrower as experience accumulates (the number of throws increases), which means that our confidence in our estimates increases. At the same time, the posterior curve shifts towards the true parameter value, i.e. the values around the true parameter will have the highest probability based on the posterior distribution.

The posterior distribution of the parameters is calculated using the data and the prior. For ease of interpretation, the posterior distribution can be compressed into a single number by calculating its mean, median or other statistics and the corresponding credible interval (CI) of a given coverage (**Figure 4**). The coverage of the credible interval is the probability that the parameter falls within the given interval.

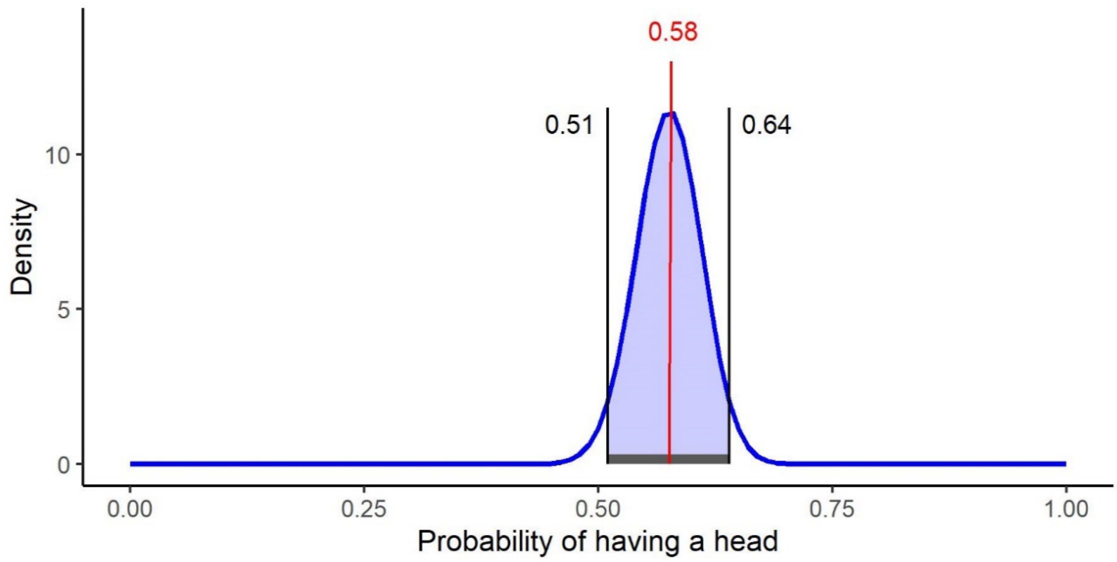



***Figure 4*** *Posterior distribution estimated mean and 95% credible interval for head toss probability after 100 coin flips. Prior assumption: fair coin; in reality: unfair coin, 0.6 (60%) probability of head tosses, mean estimate and 95% credible interval: 0.58 (95% CrI: 0.52-0.64).*

The final estimate reflects both prior knowledge and the additional information in the data, consequently our knowledge increases. The updated posterior distribution can be used as prior distribution in a subsequent model. What happens here fits well with the natural process of scientific cognition: we build a new theory from our prior knowledge and the results of our current research. We revise the state of the art again and again in the light of new data, step by step refining our understanding of the subject.

Note that the thinking of medical diagnostics is also naturally Bayesian. The main task of diagnostic workup is to quantify the likelihood of each possible underlying cause of a disease from a set of symptoms. When pondering the probabilities, we also take into account the patient's characteristics and the environmental factors. The likelihood of an underlying cause may vary depending on the age and the anamnesis of the patients and on environmental factors [90]. For example, in a 6-month-old puppy with fever, diarrhea and lethargy, under regular veterinary care and having all recommended age-appropriate vaccination, we would not suspect a parvovirus infection. However, the same symptoms at a first veterinary consultation of a 3-month-old puppy just purchased from a breeder, with no appropriate vaccination, would warrant a clinician of an increased likelihood of parvovirus infection. Another important application of Bayesian methods in veterinary clinical practice is to revise and update an initial estimate of the probability of disease (pre-test probability or prevalence) in the light of a positive or negative test result. Uncertainty in the estimates is incorporated through the use of prior distributions [19].

## 4.3 Distribution based Marginalized Multilevel Models

To my knowledge, no method has existed that allowed researchers to model disease prevalence in hierarchical subgroups of animals as correlated random quantities, estimating both conditional and marginal means. A new methodology was motivated by the problem of estimating the true, underlying PTBC prevalence in Hungarian dairy herds. My goal was to develop a methodology allowing to appropriately model both the population and the herd level mean true PTBC prevalence of primiparous and multiparous cows.

I developed a special type of marginalized multilevel model called distribution based MMM where the distributions of the conditional, cluster specific means belong to a given parametric family. In DMMMs, a natural connection exists between the conditional and the marginal components, implying that the expected value of the conditional means coincide with the corresponding marginal means (4.6). In the general MMM framework this consistency is achieved by an external constraint on the parameter space [2, 37, 91] DMMMs contain normally distributed random effects that are mapped to conditional cluster specific means (4.2) using a Gaussian copula model [92, 93]. However, the variance structure of the transformed random effects is partially lost, due to standardization. To recover the role of the variance parameters in the distribution of the conditional cluster specific means, the DMMM framework is extended by including a dispersion model (4.7). Consequently, to define a DMMM, we need to specify both the core equations (4.1-4.4) and the dispersion model (4.7). Both are incorporated in the likelihood models.

Let the family $H$ of distribution functions be parameterized by the $r$-dimensional vector $\boldsymbol{\theta}$. We suppose that $H$ has finite expected value $\mu$ depending on $\boldsymbol{\theta}$. The MMM (3.20-3.23) is modified as

$$\mu_{ij} = h(\boldsymbol{\theta}_{ij}), \tag{4.1}$$

$$\mu_{ij}^{c} = H^{-1}\left(\boldsymbol{\theta}_{ij}, \Phi\left(\boldsymbol{Z}_{\boldsymbol{ij}}^{\boldsymbol{T}}\boldsymbol{a}_{\boldsymbol{i}} \times \left(\boldsymbol{Z}_{\boldsymbol{ij}}^{\boldsymbol{T}}\boldsymbol{D}\boldsymbol{Z}_{\boldsymbol{ij}}\right)^{-\frac{1}{2}}\right)\right), \tag{4.2}$$

$$a_i \sim N(0, \boldsymbol{D}), \tag{4.3}$$

$$Y_{ij}|\boldsymbol{a}_i \sim F_Y\left(\mu_{\boldsymbol{ij}}^{\boldsymbol{c}}, v_{ij}\right), \tag{4.4}$$

where $\mu_{ij}, \mu_{ij}^{c}, h, \theta_{ij}, Y_{ij}, v_{ij}, F_Y$ are the same quantities as in (3.20-3.23), $\boldsymbol{a}_{\boldsymbol{i}}$ are independent normally distributed $q$-dimensional vectors of random effects with mean zero and covariance matrix $\boldsymbol{D}$, $\boldsymbol{Z_{ij}}$ are known covariates related to the random effects, and $\Phi$ is the standard normal distribution function. We assume that $\boldsymbol{\theta}_{ij}$ and $\boldsymbol{v}_{ij}$ are $r$ and $t$ dimensional vectors in the open sets $\Theta$ and $\Psi$, respectively. The inverse of the distribution function $H$ is

$$H^{-1}(\boldsymbol{\theta}, y) = \min\{x: H(\boldsymbol{\theta}, x) \geq y\}, \tag{4.5}$$

which is a proper definition for right continuous monotone increasing functions [94]. Note that (4.2) is a Gaussian copula model for the conditional means [92, 93].

In (4.2), the transformed random effect $u = \Phi\left(\boldsymbol{Z_{ij}^T a_i} \times \left(\boldsymbol{Z_{ij}^T D Z_{ij}}\right)^{-\frac{1}{2}}\right)$ has a uniform distribution on [0,1] interval, thus this construction assures that the probability distribution of $\mu_{ij}^c$ is $H$ with parameters $\theta_{ij}$, therefore the required constraint

$$E\left(\mu_{ij}^c\right) = \mu_{ij} \tag{4.6}$$

is satisfied. However, the random effects $\boldsymbol{Z_{ij}^T a_i}$ are standardized in (4.2), hence the framework (4.1-4.4) captures merely the correlation structure of the conditional means $\mu_{ij}^c$.

To restore the role of the variance of the random effects $\boldsymbol{Z_{ij}^T D Z_{ij}}$ in the DMMM model, we introduce a dispersion model

$$\boldsymbol{Z_{ij}^T D Z_{ij}} = w\left(\boldsymbol{\theta}_{ij}, Var\left(t\left(\boldsymbol{\theta}_{ij}, \mu_{ij}^c\right)\right)\right), \tag{4.7}$$

where $w(\boldsymbol{\theta}, x)$ is a strictly monotone increasing nonnegative function of $x$ at each parameter vector $\boldsymbol{\theta}$, and $t(\boldsymbol{\theta}, x)$ is a monotone increasing scale transformation at each parameter vector $\boldsymbol{\theta}$. Suppose that we can choose a monotonically increasing function t such that $Var\left(t\left(\boldsymbol{\theta}_{ij}, \mu_{ij}^c\right)\right)$ is finite and that $Var\left(t\left(\boldsymbol{\theta}_{ij}, \mu_{ij}^c\right)\right)$ and $\boldsymbol{Z_{ij}^T D Z_{ij}}$ are on the same scale. The transformation $w(\boldsymbol{\theta}, x)$ separates the parameter capturing the dispersion of the distribution from the mean, if needed. We suppose that the right-hand side of (4.7) can take any positive value, and that based on (4.1) and (4.7) the vector $\theta_{ij}$ can be reparameterized uniquely with $\mu_{ij}$, $\boldsymbol{Z_{ij}^T D Z_{ij}}$ and possibly some components of $\boldsymbol{\theta}_{ij}$. Hereafter we consider that this reparameterization has been done, and that $\mu_{ij}$ and $\boldsymbol{Z_{ij}^T D Z_{ij}}$ are the first two components in the vector $\boldsymbol{\theta}_{ij}$ . Note that, for specific applications, (4.7) can provide a flexible, simple and easily interpretable dispersion model if $w(\boldsymbol{\theta}, x)$ and $t(\boldsymbol{\theta}, x)$ are chosen such that the right-hand side reduces to a function of a single component of $\boldsymbol{\theta}$. For example, in the case of beta distribution family from the application in 4.4.2., the transformation $t(\boldsymbol{\theta}_{ij}, \mu_{ij}^{c})$ is the identity function as beta distribution has finite variance. The scale of the variance of beta distribution is [0,1] and the scale of $\boldsymbol{Z_{ij}^T D Z_{ij}}$ is [0,∞], thus the function $w(\boldsymbol{\theta}, x)$ is chosen to $w(\boldsymbol{\theta}, x) = 1/\psi_j$ .

The marginal variance of the response $Y_{ij}$ can be decomposed into the mean of the conditional variance plus the variance of the conditional mean, i.e.,

$$Var(Y_{ij}) = E\left(Var(Y_{ij}|\boldsymbol{a_i})\right) + Var(\mu_{ij}^{c}), i = 1, \ldots, N, j = 1, \ldots, n_i. \tag{4.8}$$

DMMMs can be fitted to data using either frequentist maximum likelihood or Bayesian methods. This approach allows for simple and flexible control over the marginal and conditional mean and dispersion measures.

The theoretical properties of the framework are presented in 11.3 of the Appendix.

## 4.4 Bayesian latent class modelling of true prevalence in animal subgroups with application to bovine paratuberculosis infection

The *prevalence dataset* - as any dataset containing diagnostic test results from animals living in separate groups – has some special features. The data is inherently hierarchical. Due to common herd characteristics (management practices, region, age distribution, etc.) animals from the same herds are more similar one to another than compared to an animal from another herd. Besides data hierarchy, the other important feature is related to the imperfection of the diagnostic tests. We do not have direct information on the real infection status of the animals. The inference can only be done based on test results and a probabilistic link between seropositivity and the real infection status. All these characteristics had to be considered when choosing the proper methodology to analyze the data.

The prevalence of infectious diseases in animals living in separate groups (e.g. herds) can be naturally analyzed using a Bayesian hierarchical latent class model. The hierarchical nature of the data is represented by random effects, and the group of infected and not infected animals are latent classes. The Bayesian methodology allows us to incorporate expert knowledge and previous experiences and to calculate various measures of interest.

In this work, I aimed to extend the legacy Bayesian methodology of prevalence estimation considering animal subgroups (parity group, species, etc.) within the herds, a new level in the data hierarchy. Animals within the same subgroup are generally homogeneous with respect to infection-related characteristics. Yet, despite this apparent similarity, substantial differences may emerge between animals in the same subgroup but from different herds. We would like to estimate the infection prevalence of these subgroups in the population, taking into consideration herd and subgroup differences and similarities. The differences are captured by herd- and subgroup-level random effects, which simultaneously contribute to the greater similarity of animals within the same group or subgroup. The model is specified such that the random effects of subgroups within the same herd are correlated. According to the legacy, prevalences will be modelled using beta distribution.

First, a general framework will be presented, then the methodology will be adapted to reassess the prevalence of bovine paratuberculosis (PTBC) infection in Hungarian commercial dairy farms.

### 4.4.1 General modelling framework

We propose a generalization of the models of Hanson et al. [25] and Branscum et al. [24] to estimate the true infection prevalence. McAloon et al. [28] applied Branscum's Bayesian methodology to estimate HTP of bovine PTBC in Ireland. His model is as follows:

$$npos_i \sim Bin(AP_i, nherd_i),$$

$$AP_i = \pi_i \times Se + (1 - \pi_i) \times (1 - Sp), \tag{4.9}$$

$$\pi_i = HTP_i \times CWHP_i,$$

where $nherd_i$ is the number of animals, $npos_i$ is the number of test-positive animals, $AP_i$ is the probability that an animal tests positive in the *i*-th herd. $\pi_i$ denotes the true prevalence in the *i*-th herd, $HTP_i$ is 1 if the *i*-th herd is infected, and it is 0 otherwise. The conditional within-herd prevalence ($CWHP_i$) is a term introduced by McAloon to denote the probability that a randomly selected animal in the herd is infected, given the herd itself is infected. We extend this model to handle within-herd animal subgroups. In the following, we use notions and notations of McAloon et al. [28].

#### 4.4.1.1 The components of the extended statistical model

Let the subgroups of the animals (e.g. parity, breed, species, genetic trait) be indexed by $j=1,2,\ldots,m$. Let $\mathrm{Pos}_{ijk}$ be 1 if the *k*-th animal from the *j*-th subgroup in the *i*-th herd is test positive and 0 if the selected animal is test negative. Further, let $\boldsymbol{Pos}_{ij} = (\ldots, Pos_{ijk}, \ldots)$ be the vector containing all the $n_{ij}$ test results for animals from the *j*-th subgroup in herd *i*. We assume that the components of $\boldsymbol{Pos_{ij}}$ (indicating test results of all animals in herd *i* and subgroup *j*) follow independent Bernoulli distributions. The probability distribution of $Pos_{ij}$ is modelled as the mixture of two products of Bernoulli distributions:

$$\boldsymbol{Pos_{ij}} \sim HTP \times \prod_{k=1}^{n_{ij}} Bernoulli(AP_{ij}) + (1 - HTP) \times \prod_{k=1}^{n_{ij}} Bernoulli(1 - Sp). \tag{4.10}$$

*HTP* stands for the herd-level true prevalence, $AP_{ij}$ represents the conditional within-herd apparent prevalence, which is the probability that a randomly selected individual from the *j*-th subgroup in the *i*-th herd tests positive, given the herd is infected. A positive test result for an animal from the *i*-th herd can be attributed to (i) either the herd being infected with a probability of *HTP* and the test giving a positive result with a probability of $AP_{ij}$, or (ii) the herd not being infected with a probability of *(1-HTP)* and the test incorrectly identifying the animal as positive with a probability of *(1-Sp)*.

The first part of the sum in (4.10) applies when the *i*-th herd is assumed to be infected and the second part describes the case when the herd is free of infections. The factors in the product express the probability of the individual test results being positive or negative and the multiplication accounts for the independence of the test results of the animals.

$AP_{ij}$ can be expressed as the sum of the probability of true positive and false positive test result:

$$AP_{ij} = Se \times CWHP_{ij} + (1 - Sp) \times (1 - CWHP_{ij}), \tag{4.11}$$

where $CWHP_{ij}$ is the conditional within-herd prevalence for the *j*-th subgroup within the *i*-th herd and $Se, Sp$ stands for the sensitivity and specificity of the diagnostic test.

We propose to model the prevalence of infection for the *j*-th subgroup within the *i*-th herd ($CWHP_{ij}, j = 1,2, \ldots, m$) by $m$ beta distributions. The beta distribution is defined on the interval [0, 1] and is widely used to model probabilities and proportions [3, 28, 29, 95]. Parameterized by two positive shape parameters, mean and precision, this is a flexible distribution family including symmetric, skewed, and U-shaped distributions.

The means $\mu_j$ of the beta distributions represent the marginal, population level means of $CWHP_{ij}$. The spread of $CWHP_{ij}$ around the mean $\mu_j$, characterizing the heterogeneity of prevalence of infection between infected herds is captured by the precision parameter $\psi_j$. We assume correlations between $CWHP_{ij_1}$ and $CWHP_{ij_2}$, $j_1 \neq j_2$ to account for the common parent herd. Beta distributions are derived from independent normally distributed random herd effects $b_i$ and additional normal random herd and subgroup interaction effects $b_{ij}, j = 1,2, \dots m$ using Gaussian copulas [96]. More precisely, let

$$b_i \sim Normal(0, \sigma),$$

$$b_{ij} \sim Normal(0, \sigma_j),\ j = 1,2, \dots, m \tag{4.12}$$

and

$$CWHP_{ij} = B^{-1}_{\mu_j,\psi_j}\left(\Phi\left(\frac{b_i + b_{ij}}{(\sigma^2 + \sigma_j^2)^{\frac{1}{2}}}\right)\right), \tag{4.13}$$

where $B^{-1}_{\mu_j,\psi_j}$ are the inverse of the beta distribution functions having mean $\mu_j$ and precision parameter $\psi_j, j = 1,2, \dots, m$, where $\Phi$ is the standard normal distribution function.

The precision parameters were specified in the model to be inversely proportional to the variances of the random effects.

$$\psi_j = \frac{1}{(\sigma^2 + \sigma_j^2)}, \tag{4.14}$$

where $\sigma^2$ is the variance of base random herd effect $b_i$, and $\sigma_j^2$ represent the variances of the additional independent random herd effects $b_{ij}$ related to the animal subgroups.

Note, that the Pearson correlation coefficient $\rho_{ab}$ between two random herd effects $b_i + b_{ia}$ and $b_i + b_{ib}$ is

$$\rho_{ab} = \frac{\sigma^2}{(\sigma^2 + \sigma_a^2)^{\frac{1}{2}} \times (\sigma^2 + \sigma_b^2)^{\frac{1}{2}}}. \tag{4.15}$$

The coefficient $\rho_{ab}$ is a suitable measure of association of $CWHP_{ia}$ and $CWHP_{ib}$, i.e., between $CWHP$ of animals from subgroups *a* and *b* kept in the same herd. We mention that $\rho_{ab}$ is the maximal correlation [97] between $CWHP_{ia}$ and $CWHP_{ib}$ (see 11.2 of the Appendix for details).

The introduced random effect and dispersion model (4.14) describe how the herd-specific conditional means $CWHP_{ij}$ vary around their marginal, population level means $\mu_j, j = 1,2, \dots, m$. This framework belongs to the class of marginalized multilevel models [37].

### 4.4.2 Application: Reassessing the PTBC prevalence in Hungary

The general framework was adapted for estimating PTBC prevalence in Hungary. The aim was to estimate both the proportion of truly infected herds among all herds (HTP) and the proportion of infected animals within infected herds (CWHP), using milk ELISA test results, individual cow-level data, and prior information derived from the latest nationwide voluntary PTBC testing in form of prior distributions [4, 98].

#### 4.4.2.1 The adapted model

To provide meaningful estimates which are easy to interpret for practitioners and compatible with herd management categories, parity was chosen as grouping factor. Based on herd (i), parity (j) and MAP ELISA test result of the cows (i.e., $\boldsymbol{Pos}_{ij}$ = $(\dots, Pos_{ijk}, \dots)$) I aimed to estimate the mean HTP and the mean $CWHP_1$ and $CWHP_2$ for primiparous and multiparous cows, respectively. $CWHP_1$ and $CWHP_2$ are the conditional within-herd prevalences for primiparous and multiparous cows, respectively, in a randomly selected herd. Herd true infection status was handled as latent class, while multiparous and primiparous cows were considered as within-herd subgroups. The age of the cows was only accounted for in Meyer's formula [17] of age-dependent sensitivity. The target condition for this analysis was the MAP-infected cow in which humoral response would become detectable within the economic life of the cow.

The extended Bayesian two-stage hierarchical model was fitted to the animal-level data as input for the first, and herd-level clusters for the second stage (**Figure 5**). According to the legacy of prevalence modelling, the PTBC prevalence of primiparous and multiparous cows in the herds $(CWHP_1$ and $CWHP_2)$ were modelled with two beta distributions. The marginal means were parameterized with $\mu_1$ and $\mu_2$ and the precisions with $\psi_1$ and $\psi_2$, respectively, as defined in (4.13-4.14), where $\sigma$ and $\sigma_j$ are the standard deviations of the random herd effects $b_i$ and the additional random herd and parity interaction effects $b_{ij}, j = 1{,}2$, according to (4.12). In line with (4.15) this model takes into account the correlation between $CWHP_1$ and $CWHP_2$.

| | Infected herds HTP | | | | Not infected herds 1-HTP | | |
|---|---|---|---|---|---|---|---|
| Data<br>ELISA test results | ⊕ ⊖ ⊕ ⊕<br>⊕ ⊖ ⊖ ⊖<br>⊖ | ⊖ ⊕ ⊕ ⊕<br>⊖ ⊖ ⊖ ⊕ | P(⊕) = AP | ⊕ ⊕ ⊖ ⊖<br>⊕ ⊕ | ⊕ ⊖ ⊖<br>⊖ ⊖ ⊖<br>⊕ ⊖ ⊖ | P(⊕) = 1-Sp | ⊖ ⊖ ⊖<br>⊖ ⊕ ⊖<br>⊖ |
| Stage 1<br>True within-herd prevalences | $CWHP_{1j}$ | $CWHP_{2j}$ | $CWHP_{ij}$ follows beta distribution with mean $\mu_j$ and precision $\psi_j$ | $CWHP_{Nj}$ | 0 | no infection | 0 |
| Stage 2<br>True mean within-herd prevalence | $CWHP = \mu_j$ | | | | 0 | | |

**Figure 5** Stages of the Bayesian model. HTP – herd-level true prevalence, AP – apparent prevalence, Sp – specificity, j – parity, $CHWP_{ij}$ – conditional within-herd prevalence for parity j in herd i.

The PTBC infection status of the herds were assumed to be independent of each other having the same probability HTP. In formulating informative prior distributions for HTP I drew on the evidence from the study of Ózsvári, Lang et al. [4] on the prevalence of paratuberculosis in large-scale commercial dairy cattle:

$$HTP \sim Beta(a_{HTP}, b_{HTP}). \qquad (4.16)$$

The procedure epi.betabuster [99] in R 4.1.0 package epiR [100, 101] was used to calculate the parameters $a_{HTP} = 66.19$ and $b_{HTP} = 8.65$ of the beta distribution, which corresponded to a prior mode of 0.895 and a 5th percentile value of 0.819.

Mean CWHP for primiparous and multiparous subpopulations was modelled assuming vague, normally distributed prior distributions:

$$logit(\mu_1) \sim Normal(0,10),$$

$$logit(\mu_2) \sim Normal(0,10). \qquad (4.17)$$

Independent normal prior distributions with mean zero were fitted to the herd-level random effects:

$$b_i \sim Normal(0, \sigma),$$

$$b_{i1} \sim Normal(0, \sigma_1),\ b_{i2} \sim Normal(0, \sigma_2) \qquad (4.18)$$

and flat inverse gamma prior distributions were imposed on the variance components:

$$\sigma^2 \sim Inverse.gamma(0.01,0.01),$$

$\sigma_1^2 \sim Inverse.gamma(0.01,0.01)$,

$\sigma_2^2 \sim Inverse.gamma(0.01,0.01)$. (4.19)

The Bayesian model was implemented using the rstan package of R 4.1.0 statistical software [100, 102]. The model was run using rjags and runjags packages [103]. The R and RStan source codes of the Bayesian model, as well as code for simulating illustrative data, are available in the supplementary material of the related publication.

Along with the estimation of model parameters, the methodology allowed for the estimation of marginal median $CWHP_1$ and $CWHP_2$ with the random effects set to zero in (4.13).

The 20,000 Markov Chain Monte Carlo (MCMC) iterates were used to estimate the $\mu_j, \psi_j$ parameters of the marginal beta distributions of $CWHP_j$, the standard deviations of the random effects $\sigma$ and $\sigma_j, j = 1,2$, and the correlation coefficient $\rho$ in (4.15). The obtained marginal beta distributions and a Gaussian copula with correlation matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ [96] were used to simulate from the posterior joint distribution of $CWHP_1$ and $CWHP_2$. I simulated independently 10,000 pairs as follows. In each simulation step I randomly selected a parameter vector from the 20,000 MCMC iterations. The actual $CWHP_1, CWHP_2$ pair was generated from the joint distribution corresponding to the selected parameter vector. I plotted the simulated $CWHP_1$, $CWHP_2$ pairs to illustrate the correlation pattern of the PTBC prevalence of primiparous and multiparous cows within the same herd. Based on the simulated ratios $CWHP_2/CWHP_1$, I estimated the posterior distribution, the posterior median, and the credible interval (CrI) of the prevalence ratio between multiparous and primiparous cows within a randomly selected herd.

The posterior distribution of the standard deviation (SD) of $CWHP_j, j = 1,2$ was calculated from the MCMC iterates using the variance formula of the beta distribution $\mu_j(1-\mu_j)/(1+\psi_j)$.

The posterior distribution of $CWHP_1$ and $CWHP_2$ can be expressed using the 20,000 MCMC iterates of the $\mu_j$ and $\psi_j, j = 1,2$ parameters of the corresponding beta distributions. By taking the arithmetic averages of the obtained 20,000-20,000 "iterates" of the beta distributions, I obtained the estimated posterior distributions of $CWHP_1$ and $CWHP_2$. From this, I computed the estimated proportions of infected herds at different levels of CWHP.

By modifying the location parameters of the prior beta distribution of HTP, we can assess the sensitivity of the posterior mean and median of HTP. The mode and 5th percentile were decreased to 0.795 and 0.719 and increased to 0.995 0.919, by 10 percentile points, respectively. Beta distributions were fitted using betabuster [99] in R 4.1.0 package epiR [100, 101]. The parameters are $a_{HTP} = 78.09$, $b_{HTP} = 20.88$ and $a_{HTP} = 39.54$, $b_{HTP} = 1.19$.

Specificity and age-dependent sensitivity estimates were based on Meyer et al. [17]. Specificity was set to $Sp = 0.995$ and sensitivity was calculated using the age-dependent nonlinear formula:

$$logit\big(Se(age)\big) = a - b \times \exp\left(-c \times age\right), \tag{4.20}$$

where *a* is the upper limit of the logit of the sensitivity when age increases, *b* is a scaling factor, *c* is the coefficient for the decay of age effect and age is the age of the animals at the time of diagnosis, measured in years. Meyer's estimates for the coefficients of the non-linear regression, derived from a large dataset of repeated milk ELISA test results, are a = 1.2, b = 3.0 and c = 0.3. Substituting the age of 1 year and 10 years into (4.20), we can conclude that the sensitivity of the MAP ELISA test is highly age-dependent, starting from 0.265 for one-year-old and reaching 0.741 for ten-year-old animals. To increase the efficiency and stability of the Bayesian model I defined age groups as follows: 1.75, 2.00, …, 2.75 years (increment=0.25), 3.00, 3.33, …, 4.33 years (increment=0.33), and 4.50, 5.00, 5.50, …, 16 years (increment=0.5). The maximum absolute deviance between sensitivities calculated with the actual and rounded ages was 0.015 [4].

To compute posterior estimates, four chains were run, each of 20,000 MCMC iterations. The runs were started from different sets of initial values of the parameters. The first 10,000 warm-up iterations in each chain were discarded. Convergence was checked by visual inspection of the trace plots of the chains. Additionally, the split R-hat formula [22] was used as convergence measure. Both approaches confirmed convergence.

#### 4.4.2.2 Model checks

Posterior predictive check of model fit was performed using the R package bayesplot [104]. To evaluate the fit of the model to the observed data, replicates of counts of positive tests were simulated from the posterior distributions for each herd and parity and compared to observed counts. Visual inspection of smoothed histograms and a scatterplot of observed counts against mean replicated counts was used to check the discrepancy of the distributions of replicated and observed counts. To compare means, medians, and standard deviations of observed and replicated counts Bayesian p-values were computed and evaluated. All the p-values fell within the range of 0.39-0.61, demonstrating good fit of the model [22].

#### 4.4.2.3 Sensitivity analyses

To assess their influence on HTP and CWHP separate models were set up, allowing for distinct parameters for herd size categories and regions. When investigating the effect on CWHP, the parameters were allowed to differ by parity. Effects were assessed by pairwise comparisons of the parameter estimates of the herd size categories. For regions, the difference between the overall mean and the individual regional estimates was assessed. A potential interaction between herd size and region was also investigated, with different CWHP parameters estimated for each herd size–region combination. The difference between the herd-size related means and the corresponding region estimates was assessed.

A sensitivity analysis was performed to assess the effects of changing the sensitivity and specificity parameters on posterior estimates. The main analysis was repeated with both the lowest and highest possible *Se* for each age groups, adding or subtracting their standard errors appropriately to the coefficients a, b and c in Meyer's formula [17]. For the two analyses, the coefficients were set to a = 1.2 - 0.17, b = 3.0 + 0.23, c = 0.3 - 0.067 and to a = 1.2 + 0.17, b = 3.0 - 0.23, c = 0.3 + 0.067, respectively. The change in estimates due to the uncertainty in specificity was also evaluated by assigning to *Sp* the 95 % confidence limits 0.994 and 0.996, reported in Meyer et al. [17].

I investigated the effect of using a prior distribution for specificity parameter instead of a fixed value. Considering a lower estimate *Sp*=0.986 from Nielsen et al. [18], a uniform prior in the interval 0.986-0.996 was defined. I also quantified the possible between-herd heterogeneity of the specificity parameter allowing for different and independently varying specificity for individual herds, using the same uniform prior in the interval 0.986-0.996.

#### 4.4.2.4 Model validation via simulation

McAloon et al. [105] pointed out that under certain disease characteristics, including low *Se* of diagnosing paratuberculosis Bayesian latent class analyses can exhibit low accuracy of HTP estimates. It was proposed to explore the eventual bias of the Bayesian model via data simulation. Following the guidance, new diagnostic test results were simulated to each cow recorded in our data set, while preserving their initial covariates. Values of HTP, mean $CWHP_1$, mean $CWHP_2$ and standard deviations of the random effects used in the simulation were specified in advance. The HTP range of 0.82, 0.85, 0.88, 0.91 and 0.94 corresponded to the 95% symmetric CrI of the prior used. Mean CWHP for primiparous and multiparous animals were set to 0.08 and 0.16, respectively. Standard deviation of the random effects of the herds was set to σ = 0.32, $\sigma_1 = 0.12$, $\sigma_2$ = 0.12, in accordance with the magnitude of the posterior estimates obtained from the main Bayesian model. Finally, the Bayesian model outlined above was fitted to the simulated data and the discrepancy between the known parameter values and their obtained posterior estimates were assessed.

## 4.5 Estimation of paratuberculosis infection prevalence in subgroups of individual dairy herds

After estimating the population-level infection prevalence of the primiparous and multiparous cows, it is possible to use the information extracted from this large sample to make inferences to single herds. The aim of this research was to provide practitioners with a tool to estimate the true infection prevalence of their herd based on cow level data and on the results of the nationwide voluntary PTBC testing program as prior information. Estimating the prevalence of individual herds is a direct sequel of the hierarchical modeling of nationwide data and relies strongly on the results.

The goal of the modelling was to estimate the probability that a randomly selected animal in the herd was infected. The model results were illustrated by analyzing two simulated large commercial dairy cattle herds. The model is available for use online (https://github.com/VeresKatalin/PTBC). In addition to that, thumb rules were derived from the results. The rules describe how to estimate the true prevalence without using the model, solely from the apparent prevalence, by a simple multiplication.

Additionally, I aimed to provide an internationally relevant context for our analysis by integrating prior information from diverse regional studies together with region-specific synthetic data. I set up a model which can be customized for any region using real local data and prior information. The international model was also enabled to estimate the likelihood that a herd is infected. The model was tested using synthetic data representing six regional scenarios in four countries (Chile, Denmark, Italy, and Hungary). This model is available online at https://github.com/VeresKatalin/PTBCprevalence .

### 4.5.1 The statistical model

In the following, I assumed that the analyzed herd is infected. It is a reasonable assumption, since the true prevalence estimated from the latest national data was 92% (section 5.1). I fitted a Bayesian two-stage hierarchical model to the data. It was assumed that the true prevalence of each herd is well represented by the national data, therefore posterior distributions from historical data were used as informative priors. **Figure 6** shows the prior for $CWHP_1$ (conditional within-herd prevalence among primiparous cows) and $CWHP_2$ (conditional within-herd prevalence among multiparous cows). The priors were chosen to best fit the corresponding posterior distributions from the national data.

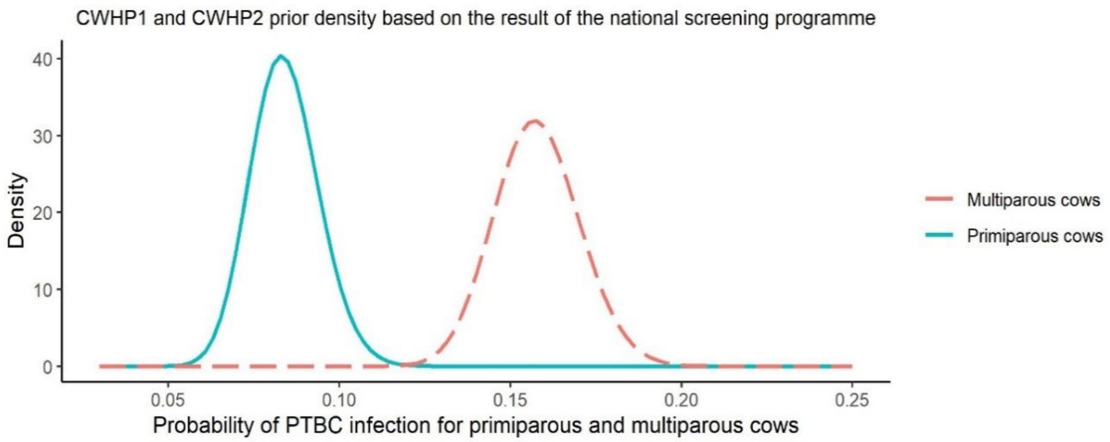



***Figure 6*** *Prior distributions for Bayesian estimation of the true prevalence of paratuberculosis in primiparous and multiparous cows based on the posterior distribution from national data.*

The above priors, the results of the ELISA tests, the parity and the age of the cows were used to calculate the estimate. It is easy to calculate the proportion of positive cases (apparent prevalence, AP) from raw data, but in practice, due to eventual false positive and false negative results, the true underlying prevalence is often different. In this study, the sensitivity (Se) of the ELISA test was assumed to be age-dependent according to the Meyer formula [17]. Applying this formula to the screened herd, approximately 24-67% of the tests will produce false negative results, depending on the age of the animals. The specificity (Sp) of the test was considered to be 99.5% [17]. This means that false positive cases account for only 0.5% of the test results obtained from non-infected animals. The model aims to provide an estimate of the true prevalence of infection, taking into account the limitations of the diagnostic test.

The apparent prevalence is defined as the probability of a positive test of a cow in an infected herd. It can be expressed as the sum of the probability of true positive and false positive test results:

$$AP = Se \times CWHP + (1 - Sp) \times (1 - CWHP). \tag{4.21}$$

The above formula describes that if the proportion of infected animals in the parity group is CWHP, then $Se \times CWHP$ percentage of cows will have a true positive test result, while the $(1 - Sp) \times (1 - CWHP)$ proportion of cows will have a false positive result.

A Bayesian model was fitted to the herd data to calculate the posterior distributions of $CHWP_1$ and $CHWP_2$. Posterior distributions are the Bayesian estimates of the prevalence of PTBC associated with primiparous and multiparous cows, respectively.

The Bayesian model was set up and run using the rstan package of the statistical software R 4.1.3 [100, 102].

### 4.5.2 Estimating the PTBC infection prevalence in an individual herd without the model

In addition to the Bayesian model, a simple thumb rule for estimating the true herd-level prevalence has been developed. The model was run on the individual PTBC ELISA milk test results from 116 dairy cattle farms from the *prevalence dataset*. Next, a linear regression was fitted considering the estimated true prevalence as a dependent variable and the apparent prevalence as an explanatory variable. In the equation of the regression line the estimate of the intercept was close to zero, therefore the true herd-level prevalence can be estimated by simply multiplying the apparent prevalence by a constant.

### 4.5.3 Simulation study

The accuracy of our Bayesian model was tested by simulation. Data from 500 herds were simulated, taking into account the herd sizes, the proportion of primiparous and multiparous cows, and the PTBC infection patterns of the real data. After running the model on the simulated data, the proportion of cases in which the true prevalence fell within the 95% credible interval provided by the model was examined. I then looked at how the accuracy of the estimate would change if only a fixed proportion of primiparous or multiparous cows were tested.

The simulation was carried out as follows. First, the basic data for the herds were generated. The size of the herd was obtained from a negative binomial distribution with mean 450 and standard deviation 500. To this, 50 was added, thus defining a minimum number of 50 cows and an average number of 500 cows. Taking into account the national characteristics, "primiparous" status was assigned with 40% probability to an individual cow. Cow age was sampled from a gamma distribution, with the minimum age observed in the corresponding parity group of the national data added. The average age and its standard deviation were set to match the national data.

For each virtual cow test result was generated as follows. The true prevalences for the parity groups (primiparous and multiparous) are positively correlated since in a herd where primiparous cows have a higher-than-average infection level, multiparous cows also tend to be more infected than average. Thus, the underlying prevalences were determined according to the joint posterior distribution. This method takes the correlation into account. Based on CWHP, Se and Sp corresponding to cow age, the AP was calculated using formula (4.21). Test results were generated as positive with probability AP and as negative with probability 1-AP.

### 4.5.4 International model

The international model builds on expert opinions or insights from the historical data of the region, in the form of prior distributions. It uses priors for the proportion of infected herds in the region (HTP), the mean regional within-herd infection prevalence of primiparous and multiparous cows ($\mu_1$ and $\mu_2$), and for the variances of the herd- and parity-level random effects. HTP, $\mu_1$ and $\mu_2$ are modeled by beta distributions. The priors of the variances of the random effects ($\sigma^2$, $\sigma_1^2$ and $\sigma_2^2$) are described using inverse gamma distribution.

Priors for Hungary were defined to best fit the posterior distributions from the analyses of the national PTBC screening program. For Denmark, region-specific $CWHP_1$ and $CWHP_2$ prior distributions were derived from posterior distributions reported in the literature (Appendix Table A4), using the procedure epi.betabuster [99] in R 4.1.0 package epiR [101] to specify the parameters of the beta distributions. For the other non-Hungarian regions, region-specific prior distributions were derived from prior distributions found in the literature [34, 35, 106, 107]. For each region, the prior distributions for the variances of the random effects were set to be the same as those used for Hungary. The details of the derivations of the priors and the priors used in the analysis are shown in the Appendix (Table A2,Table A3, Section 11.1.2.).

To infer the infection probability of the herd, Bayes factors [108] were used, defined as the ratio of the probability of the observed data in one statistical model to the probability of the same data in another model. The Bayes factor quantifies how much more (or less) likely the data are in one model than in another, thus indicating which model provides a better fit to the observed data. The Bayes factor was calculated using the R package bridgesampling

[109], version 1.1.2. Specifically, a model assuming the herd was infected was compared with a model assuming the herd was not infected. The posterior probability of the model given the data is $\frac{BF \times \mu_{HTP}}{BF \times \mu_{HTP} + (1-\mu_{HTP})}$, where BF is the Bayes factor, and $\mu_{HTP}$ is the expected value of the prior distribution of HTP.

Within-herd animal-level prevalence in the subgroups of primiparous and multiparous cows was estimated using the method from section 4.5.1. As real-world data were available only for Hungary, synthetic datasets for the other regions were generated using prior information from the literature. A herd was considered infected if it had at least one truly infected cow. Throughout the study, the use of the IDEXX milk ELISA test was assumed. The sensitivity (Se) of the test was assumed to be age-dependent and calculated according to (4.20).The specificity (Sp) of the test was considered to be 98.6% [18].

I analyzed synthetic datasets mimicking herds in Hungary, Denmark, Southern Italy, Lombardy (Northern Italy), Veneto (Northern Italy), and Chile. To imitate real-world data, the Hungarian data were manipulated to reflect the prevalence and mean herd size of the regions analyzed (Appendix Table A1). Twenty synthetic datasets were generated for each region as follows: from the data of a sufficiently large Hungarian herd, samples were drawn with replacement, in accordance with the mean herd size reported for the given region. The original ELISA test results were replaced with independently simulated values. The MAP infection status of the herd was randomly generated from a Bernoulli distribution, with the parameter taken from the prior distribution of the region-specific herd true prevalence. If the herd was randomized to be infected, the infection probabilities of individual cows in parity subgroups were generated according to a Gaussian copula model (Xue-Kun Song, 2000) using region- and parity-specific beta prior distributions (Appendix 11.1.3).

## 4.6 The relationship between PTBC infection and survival in Hungarian dairy herds

Time-to-event data are data series in which each observation unit (e.g. patient or animal) has a follow-up time and a status indicator that indicates whether the event under study occurred at the end of the follow-up time. Statistical processing of time-to-event data is called survival analysis. In human studies survival analysis is already an established method. Veterinary medicine should follow the trend and make use of the time-to-event data from individual digital records.

The use of survival analysis techniques is illustrated through a public health case study (see section 6.2.2 for details) and through the investigation of the relationship between paratuberculosis serological status and the probability of culling in dairy cows.

The hazard is the instantaneous rate of occurrence of events as a function of time. The hazard function is a key concept in survival analysis. In a homogeneous population the hazard function is the same for all individuals. Heterogeneity can be accounted for by including observed covariates in the model. This way, individuals with the same values of covariates will have the same hazard function and time-to-event distribution. But there can be latent variables that are not included in the model, causing that clusters with apparently the same covariates may have different distributions. [110]. In the *survival dataset* the hierarchy of the data implies the existence of latent variables at herd level thus we need to control for unobserved heterogeneity in the survival model by using adequate marginal or conditional modelling techniques.

I analyzed the *survival dataset* to assess the impact of PTBC positivity on the risk of culling in Hungarian dairy herds. Cows were followed from the day of the ELISA test until culling or death, but at the latest until the end of the study on 31 July 2019, when administrative censoring was performed. I estimated the relative culling risk (hazard ratio) of the ELISA-positive group compared to the ELISA-negative group.

I checked the proportional hazards assumption by visual inspection of the log-log survival curves (**Figure 7**) [111], which is a prerequisite for fitting a proportional hazards model.

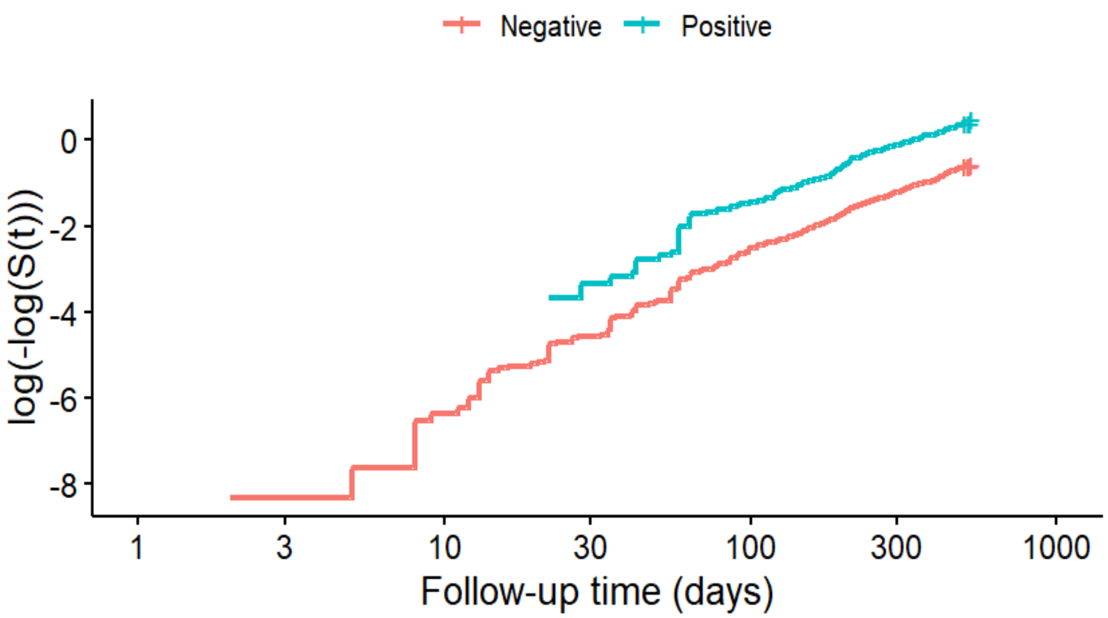



***Figure 7*** *Log-log survival curves of seropositive and seronegative cows.*

The proportional hazard assumption was acceptable, meaning that the risk of culling in the ELISA positive group was reasonably modelled as a constant multiple of the risk of the ELISA negative group over the follow-up period. This constant, the hazard ratio, was determined using the Cox regression model. The age of the cows at the time of the test, measured in days, was included as a non-linear term in the model ($AGE + AGE^2 + AGE^3$) in the marginal (4.6.1) and the frailty model (4.6.2), while in the stratified regression (4.6.3) age was discretized. The herd was handled as a cluster. I intended to account for within-cluster correlation reflecting the influence of herd-level latent variables, including biosecurity measures, management practices such as the separation of younger and more susceptible animals, and financial conditions. Outlier diagnostics were conducted using martingale and deviance residuals, confirming the adequacy of the models.

### 4.6.1 Marginal approach

I fitted a marginal Cox proportional hazards regression model to the survival data to estimate the hazard ratio (HR) of the two groups. In this marginal model, the hazard is a function of the covariates, and the latent effect related to herd is averaged out. The variance-covariance matrix of the estimated parameters is calculated using a robust variance estimator [111].

### 4.6.2 Conditional approach, frailty model

I fitted a frailty (conditional) Cox proportional hazards regression model to the previously analyzed clustered survival data to estimate the HR of the ELISA positive and negative groups. In conditional models, the hazard is modelled as a function of the covariates and of the latent frailty terms. The within-cluster correlations are modelled with the frailties. I assumed Gaussian distribution for the logarithmically scaled frailties.

### 4.6.3 Conditional approach, stratified cox regression with individual level strata

Based on expert knowledge, herd affiliation and age group are the most influential sources of heterogeneity with respect to MAP infection status. A 1:5 matching was performed on these two factors. This means that each ELISA positive cow was matched with 5 “similar” ELISA negative control cows to define a stratum. The 5 control animals were chosen randomly, without replacement, from a pool of ELISA negative cows from the same herd and from the same age group at the time of testing. To avoid statistical dependency, different positive cows were assigned to different, non-overlapping control groups selected without replacement. Age groups were defined as follows: Under 2 years, 2-<3 years, 3-<4 years, 4-<5 years, 5-<6 years, 6+ years. A stratified Cox proportional hazard regression was then fitted with ELISA positivity as the sole explanatory variable. The model allowed each stratum to have a different baseline hazard function while sharing the same HR estimate.

# 5 Results

## 5.1 Bayesian latent class modelling of true prevalence in animal subgroups with application to bovine paratuberculosis infection

I aimed to provide easily interpretable subgroup-level estimates for marginal and population means, and to gain new insights into the interrelation of parity groups. Since the true prevalence of PTBC infection in Hungary was estimated only once by Ózsvári, Lang et al. [4] on a limited dataset, my secondary goal was to consolidate previous findings using a large body of recent data and prior distributions derived from historical results.

### 5.1.1 Descriptive statistics

The data used in the study consisted of 58,196 cows from 117 large-scale intensive dairy cattle herds in Hungary. Excluding records with missing or invalid identifier, ELISA test result or birth date (n=2,602, 4.5%), data on 55,594 cows from 116 herd were eligible for analysis. The median herd size was 479 cows (interquartile range: 274 - 602). Distribution of herd size and age of cows are shown in **Figure 8-9**.

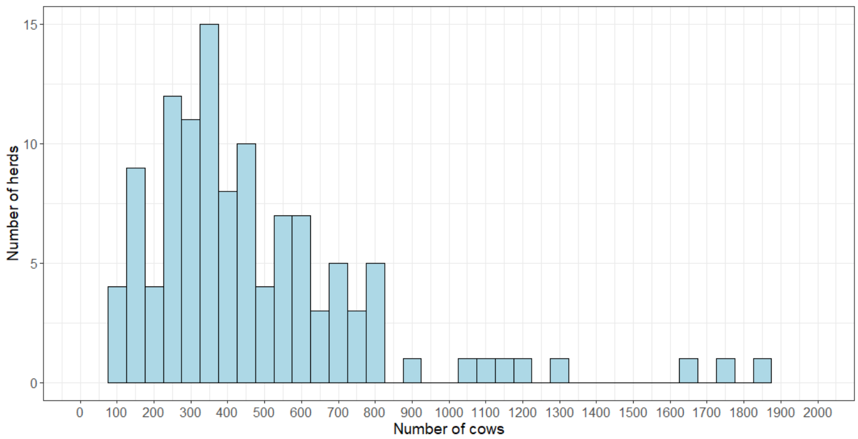



***Figure 8*** *Histogram of herd sizes of 116 large-scale intensive dairy cattle herds in Hungary, 2019.*

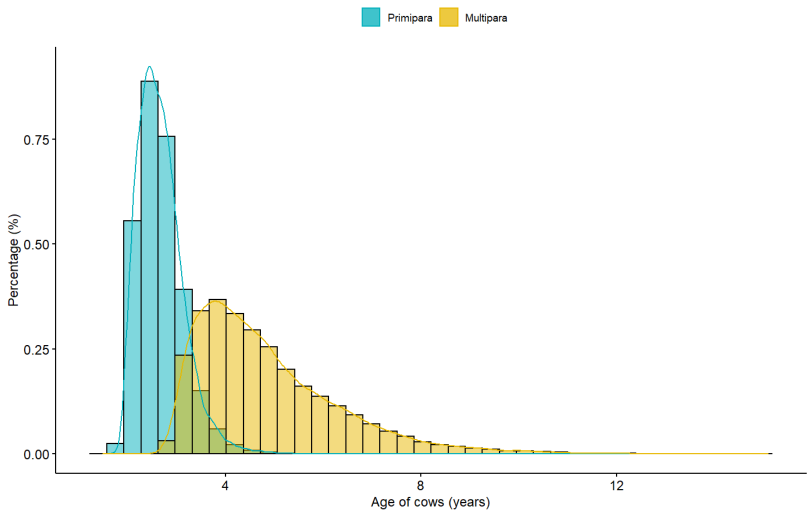



***Figure 9*** *Age distribution by parity of cows from 116 large-scale intensive dairy cattle herds, Hungary, 2019 (n=55,594).*

In total, 3,343 cows were classified as ELISA positive, resulting in an overall AP of 6.0%. The number of cows by parity and ELISA status split by herd size category is shown in **Table 1**. There were 4 farms with no ELISA positive cow. This corresponded to an apparent HTP of 96.6%. The percentage of farms with an apparent prevalence of >1%, >2%, >5%, and >10% was 92.2, 78.4, 53.4, and 19%, respectively.

**Table 1** ELISA test results and apparent prevalence (AP) of paratuberculosis in the collected sample of dairy cows in Hungary.

| | **Primiparous cows** | | | **Multiparous cows** | | | **Overall** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Herd size category** | **N** | **+** | **AP** | **N** | **+** | **AP** | **N** | **+** | **AP** |
| 100-300 cows | 2 590 | 138 | 5.3 | 4 073 | 444 | 10.9 | 6 663 | 582 | 8.7 |
| 301-600 cows | 9 361 | 271 | 2.9 | 13 472 | 1 081 | 8.0 | 22 833 | 1 352 | 5.9 |
| >600 cows | 11 457 | 357 | 3.1 | 14 641 | 1 052 | 7.2 | 26 098 | 1 409 | 5.4 |
| Overall | 23 408 | 766 | 3.3 | 32 186 | 2 577 | 8.0 | 55 594 | 3 343 | 6.0 |

ELISA: Paratuberculosis Screening Antibody Test (IDEXX Laboratories Inc., Westbrook, ME, USA), AP: apparent prevalence, N: number of tests, +: number of positive tests

ELISA test results by lactation number are depicted in **Figure 10.**

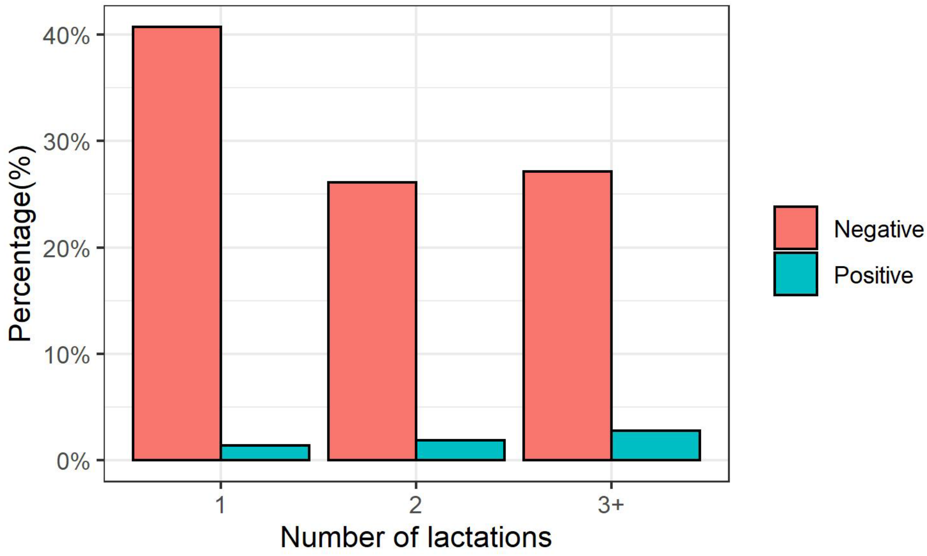



***Figure 10*** *Paratuberculosis milk ELISA test results by lactation number of all cows from 116 large-scale intensive dairy cattle herds in Hungary, 2019 (n=55,594). ELISA: Paratuberculosis Screening Antibody Test (IDEXX Laboratories Inc., Westbrook, ME, USA.*

### 5.1.2 Bayesian model

The prior and the posterior probability density functions for HTP are shown in **Figure 11**.

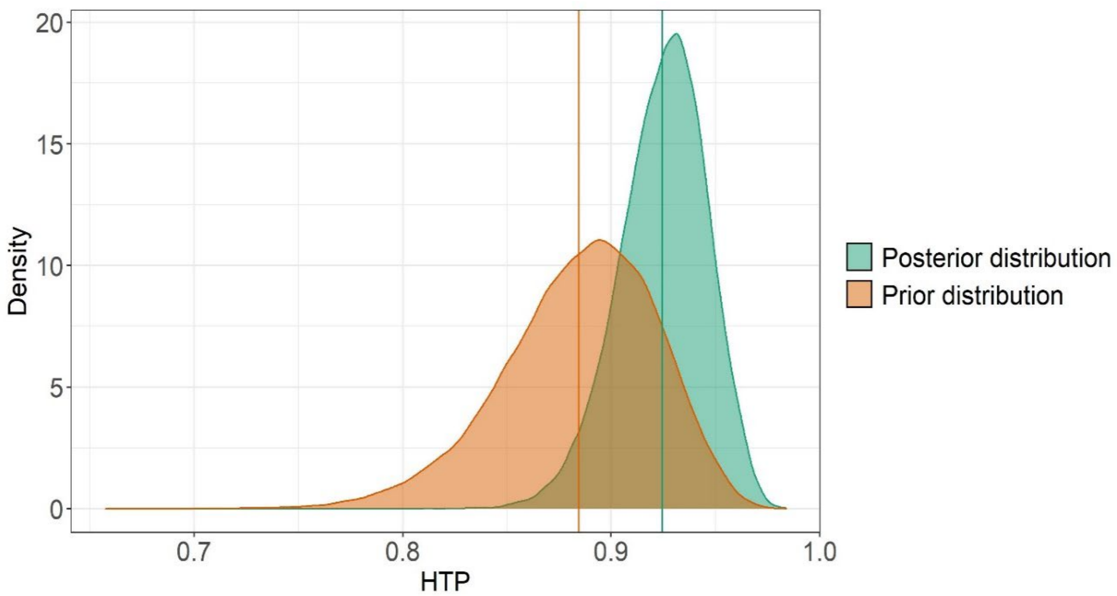



***Figure 11*** *Prior and posterior probability density functions and mean values for herd-level true prevalence (HTP) of paratuberculosis estimated from the Bayesian model fitted to the collected sample of dairy cows, Hungary, 2019.*

The posterior parameter estimates of the main Bayesian model are shown in **Table 2**. While the apparent HP was 96.6% based on ELISA test result, mean HTP was estimated to a lower level of 92.5% (95% CrI: 88.0%; 96.1%). According to the high specificity of the test, the cut-off for a herd to be considered infected when calculating apparent HP is at least one seropositive cow. The CWHP was estimated to 8.4% (6.6%; 10.4%) for primiparous and 15.8% (13.5%; 18.4%) for multiparous cows, in contrast with 3.3% and 8% of AP, respectively.

**Table 2** Posterior parameter estimates and additional derived measures from the Bayesian model fitted to the collected sample of dairy cows, Hungary, 2019.

| **Estimate** | **mean (95% CrI)** |
|---|---|
| HTP | 0.924 (0.880; 0.960) |
| Mean $CWHP_1$ | 0.084 (0.066; 0.104) |
| Mean $CWHP_2$ | 0.158 (0.134; 0.184) |
| Median $CWHP_1$ | 0.047 (0.032; 0.065) |
| Median $CWHP_2$ | 0.125 (0.100; 0.150) |
| SD of $CWHP_1$ | 0.098 (0.079; 0.120) |
| SD of $CWHP_2$ | 0.130 (0.111; 0.153) |
| Maximal correlation between $CWHP_1$ and $CWHP_2$ | 0.886 (0.806; 0.942) |

HTP: Herd-level true prevalence, beta prior based on historical data.
$CWHP_1$ and $CWHP_2$: Conditional within-herd prevalence for primiparous cows and multiparous cows from a randomly selected herd, modelled with beta distributions.
Mean $CWHP_1$ and $CWHP_2$: Population mean of disease prevalence by parity.
Median $CWHP_1$ and $CWHP_2$: Population median of disease prevalence by parity.
SD of $CWHP_1$ and $CWHP_2$: standard deviation of $CWHP_1$ and $CWHP_2$

Posterior probability density functions of mean CWHP by parity are presented in **Figure 12**. The estimates for median CWHP within infected herds are 4.7% (3.2%; 6.4%) for primiparous and 12.4% (10%; 15%) for multiparous cows.

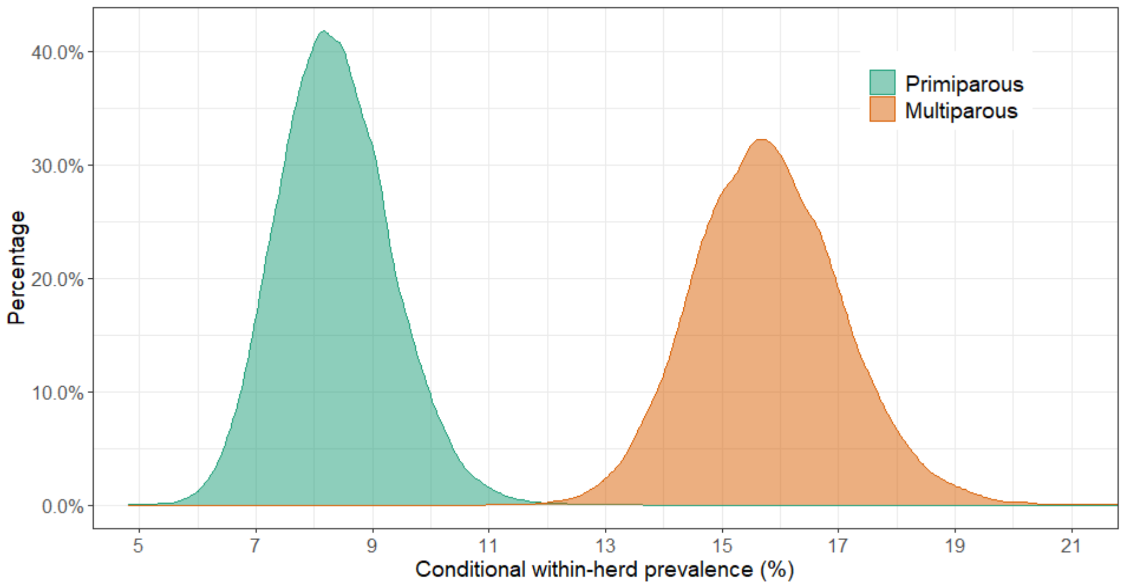



***Figure 12*** *Posterior probability density functions of mean conditional within-herd prevalence (CWHP) of paratuberculosis by parity from the Bayesian model fitted to the collected sample of dairy cows, Hungary, 2019.*

**Figure 13** demonstrates positive association between $CWHP_1$ and $CWHP_2$. The maximal correlation of $CWHP_1$ and $CWHP_2$ is as strong as 0.89.

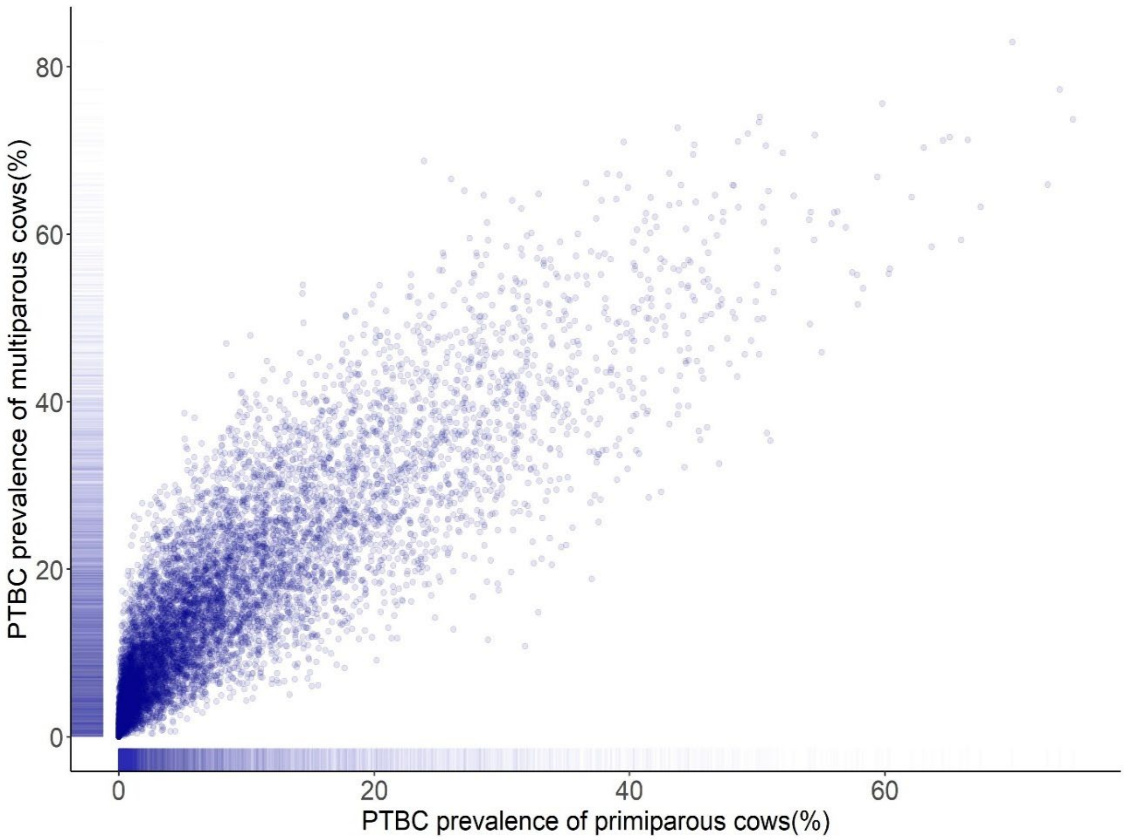



***Figure 13*** *Correlation of the PTBC prevalence of primiparous and multiparous cows within the same herd, from the Bayesian model fitted to the collected sample of dairy cows, Hungary, 2019.*

The proportion (i.e., posterior probability) of infected herds having at least 5% or 15% true prevalence among primiparous cows was 49% and 19.3%, respectively. The proportion of infected herds having at least 5% or 15% true prevalence among multiparous cows was 77.9% and 43.3%, respectively. The proportion of infected herds having at least 50% true prevalence among primiparous cows was 0.5%, and among multiparous cows it was 2.2% (**Figure 14**).

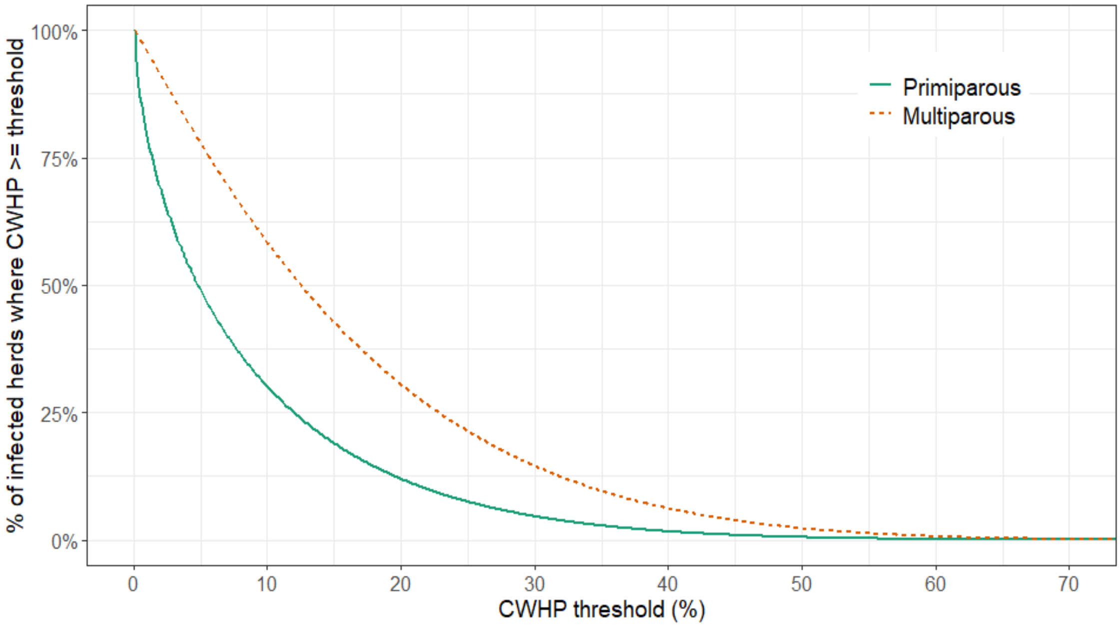



***Figure 14*** *Posterior probability of conditional within-herd prevalence (CWHP) of paratuberculosis being at least a given threshold by parity from the Bayesian model fitted to the collected sample of dairy cows, Hungary, 2019.*

The estimated proportions of infected herds with different levels of CWHP are shown in **Table 3**. The percentage of infected herds having less than 5% infected primiparous cows was estimated at 51.5% (43; 60%). The proportion of infected primiparous cows exceeded 10% on 29.7% (23; 37%) of the infected herds. In multiparous cows, 58% (49.9; 65.9%) of the infected farms had higher than 10% infected cases, and 22.5% (15.6; 30.4%) of the infected herds had less than 5% of infected animals.

**Table 3** Estimated proportions of infected herds with different levels of conditional within-herd prevalence (CWHP) from the Bayesian model fitted to the collected sample of dairy cows, Hungary, 2019.

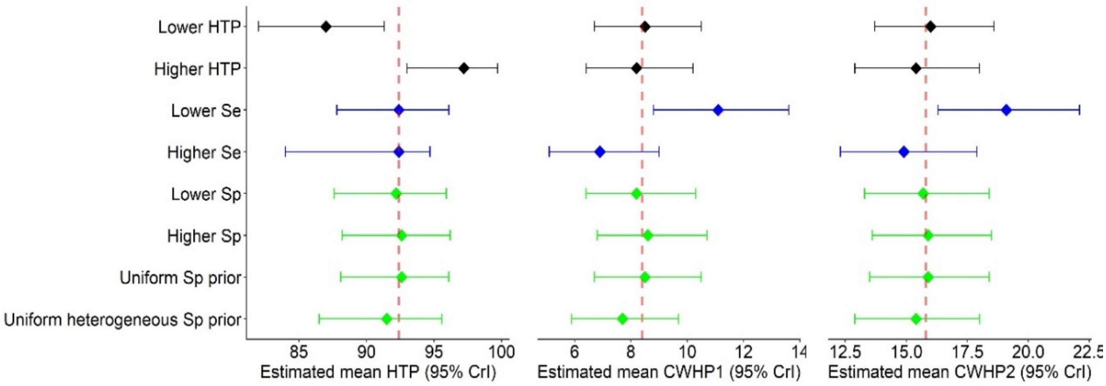
| Estimated proportion of infected herds in the given CWHP range (%) | Primiparous cows | Multiparous cows |
|---|---|---|
| | Mean (95% CrI) | |
| <=5 | 51.5 (43.0; 60.0) | 22.5 (15.6; 30.4) |
| 5<CWHP<=10 | 18.6 (15.6; 22.1) | 19.4 (17.0; 21.9) |
| >10 | 29.7 (23.0; 37.0) | 58.0 (49.9; 65.9) |

The joint distribution of CWHP of primiparous and multiparous cows demonstrates that in 22.2% of infected herds the prevalence in both the primiparous and the multiparous subgroups is lower or equal than 5%, and in 29.3% of infected herds the prevalence in both subgroups is higher or equal than 10%.

The median and 95% CrI of within-herd prevalence ratio between primiparous and multiparous cows being infected is 2.6 (0.8; 56.7). In 66% of the herds the within-herd prevalence ratio is larger than the ratio of mean prevalence of 1.88 of primiparous and multiparous cows calculated from **Table 2.**

No substantial effect of region and herd size on the HTP and CWHP and no interaction between herd size and region were found.

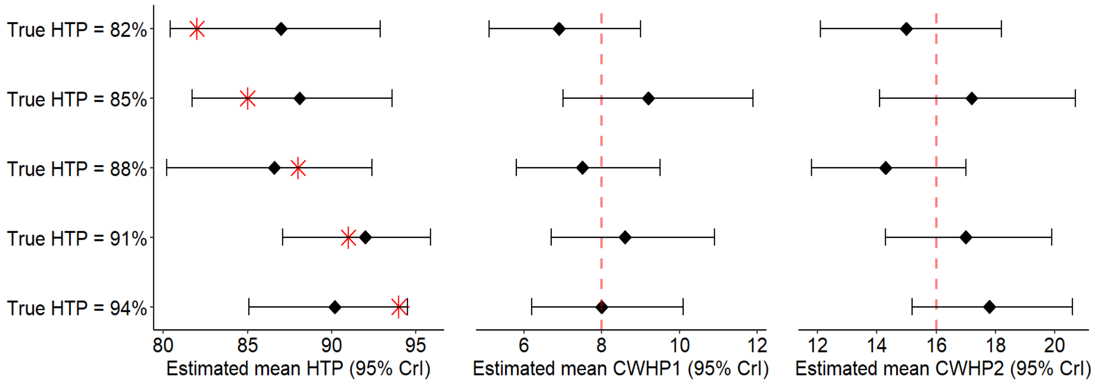
The results of the sensitivity analyses are presented in **Figure 15**. The 10-percentage-point change of the mode and 5th percentile of the HTP prior altered the posterior estimate of HTP by -5.4 and 4.8. The estimates of the mean $CWHP_1$ and $CWHP_2$ were only marginally affected. The changes in sensitivity parameter did not affect the posterior estimates of HTP but had an inverse impact on $CWHP_1$ and $CWHP_2$ estimates. Varying the specificity parameter had only a marginal impact on posterior estimates of HTP and CWHP.




HTP: herd-level true prevalence, $CWHP_1$: conditional within-herd prevalence for primiparous cows, $CWHP_2$: conditional within-herd prevalence for multiparous cows, Se: sensitivity, Sp: specificity, Change in HTP: 10-percentage-point change of the mode and 5th percentile of the beta distributions, Change in Se: parameters in Meyer's formula adding and subtracting their standard error: a = 1.2 - 0.17, b = 3.0 + 0.23, c = 0.3 - 0.067 and a = 1.2 + 0.17, b = 3.0 - 0.23, c = 0.3 + 0.067. Lower Sp: 0.994, Higher Sp: 0.996, Uniform Sp prior: uniform prior for specificity parameter on the interval 0.986-0.996, Uniform heterogeneous Sp prior: individually varying uniform prior for specificity parameter on the interval 0.986-0.996 for each herd, red dashed lines: estimated parameter values from the original model

***Figure 15*** *Posterior estimates of herd-level true prevalence (HTP) and mean conditional within-herd prevalence of primiparous ($CWHP_1$) and multiparous cows ($CWHP_2$) in the sensitivity analyses.*

The parameter estimates from the simulation models were in the same range as the posterior estimates of the main Bayesian model. As expected, the HTP estimates were biased toward the prior. True values for $CWHP_1$ and $CWHP_2$ lied always in the credible interval for the simulation estimates. **Figure 16** presents the parameter estimates from the simulation.




HTP: herd-level true prevalence (red star) , $CWHP_1$: conditional within-herd prevalence for primiparous cows, $CWHP_2$: conditional within-herd prevalence for multiparous cows, the true value of $CWHP_1$= 8% (red dashed line), the true value of $CWHP_2$= 16% (red dashed line), the standard deviations of the random effects σ = 0.32, σ1= 0.12, σ2 = 0.12.

***Figure 16*** *Bayesian estimation of the mean vs. the true value of herd-level true prevalence (HTP) and mean conditional within-herd true prevalence of primiparous ($CWHP_1$) and multiparous cows ($CWHP_2$) based on simulation.*

## 5.2 Estimation of paratuberculosis infection in subgroups of individual dairy herds

### 5.2.1 Model results

As the data used is confidential, model results are illustrated by analyzing two simulated herds. In the first herd the actual level of infection is low, while in the second herd the infection level is relatively high (**Table 4 -**

**Table 5**).

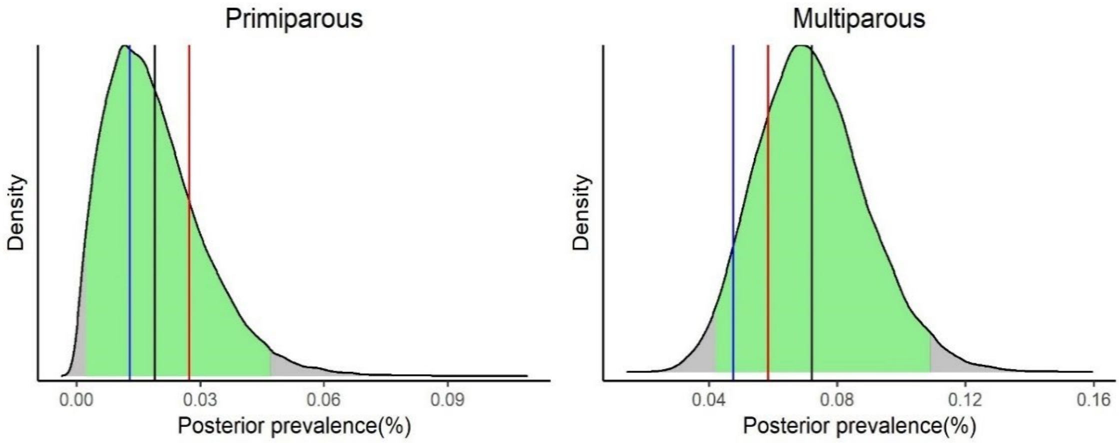
**Table 4** Simulated herd data - Herd 1.

| Herd 1 | Herd size | Proportion (%) | Apparent prevalence (%) | True prevalence (%) |
|---|---|---|---|---|
| **Primiparous cows** | 310 | 44 | 1,3 | 2,7 |
| **Multiparous cows** | 400 | 56 | 4,8 | 5,8 |
| **Overall** | 710 | 100 | 3,3 | 4,4 |

**Table 5** Simulated herd data – Herd 2.

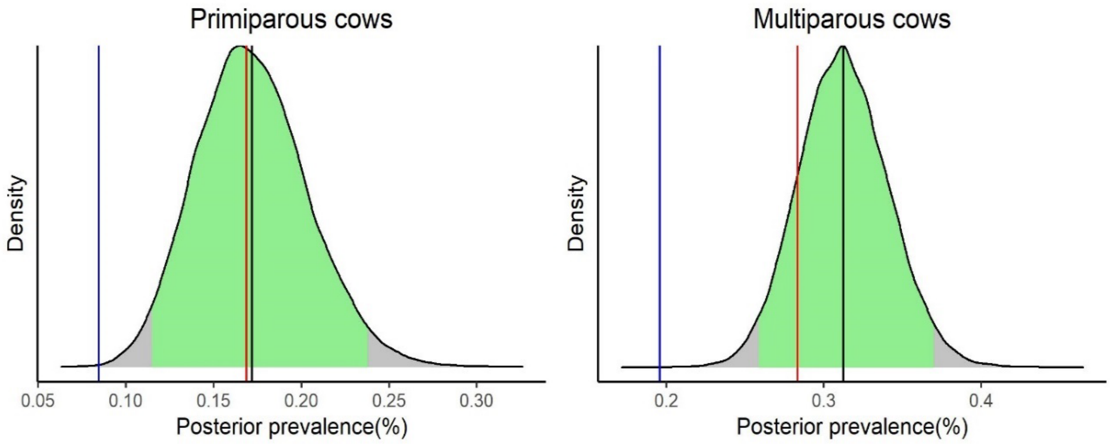
| Herd 2 | Herd size | Proportion (%) | Apparent prevalence (%) | True prevalence (%) |
|---|---|---|---|---|
| **Primiparous cows** | 307 | 39 | 8,5 | 16,9 |
| **Multiparous cows** | 475 | 61 | 19,6 | 28,3 |
| **Overall** | 782 | 100 | 15,2 | 23,8 |

The posterior estimates for the two sites are shown in **Figure 17** and **Figure 18**.




Red vertical line: true prevalence, black vertical line: posterior mean, blue vertical line: apparent prevalence, black curve: posterior distribution, green range of values: 95% credible interval for true prevalence.

***Figure 17*** *Posterior prevalence estimation for simulated Herd 1 (herd characteristics shown in Table 4).*




Red vertical line: true prevalence, black vertical line: posterior mean, blue vertical line: apparent prevalence, black curve: posterior distribution, green range of values: 95% credible interval for true prevalence.

***Figure 18*** *Posterior prevalence estimate for simulated Herd 2 (herd characteristics shown in*

*Table* ***5****).*

The horizontal axis shows the disease prevalence; the vertical axis shows the density associated with the parameter value. The probability that the true prevalence falls within a given interval is equal to the area under the curve in that interval. The black vertical line indicates the mean of the posterior distribution; the green range is the corresponding 95% credible interval. In the figure, the blue line shows the apparent prevalence and the red line the true prevalence. The true prevalence for both herds and parity groups falls within the KI of the posterior mean estimate from the model. In **Figure 17** for herd 1, where the true prevalence is relatively low, the estimate of the prevalence is higher than the apparent prevalence, but for primiparous cows, the true prevalence is slightly underestimated while for multiparous cows, it is slightly overestimated. In **Figure 18** for herd 2, where the true prevalence is relatively high, the estimate of the prevalence is considerably higher than the apparent prevalence, spot on for primiparous cows and slightly overestimated for multiparous cows.

### 5.2.2 Downloadable application

The Bayesian model is available to download from (https://github.com/VeresKatalin/PTBC). The software accepts herd data as xlsx file in the following format:

| | HERD_ID | COW_ID | MULTIPAR | AGE | POS |
|---|---|---|---|---|---|
| 1 | 1000 | 1001 | 0 | 2.72 | 0 |
| 2 | 1000 | 1002 | 0 | 2.17 | 0 |
| 3 | 1000 | 1003 | 0 | 2.33 | 0 |
| 4 | 1000 | 1004 | 0 | 2.50 | 0 |
| 5 | 1000 | 1005 | 0 | 2.49 | 0 |

HERD_ID: integer, herd ID, COW_ID: cow ID, integer, unique, MULTIPAR: parity indicator, 0 for primiparous cows, 1 for multiparous cows, POS: milk ELISA positivity indicator, 1 for test-positive cows, 0 for test-negative cows.

### 5.2.3 Rule of thumb for estimating herd-level PTBC infection

The model was run for all herds tested in the national screening program and a linear regression was fitted to the results. The estimated true prevalence was considered as dependent variable and the apparent prevalence from raw data as explanatory variable. The relationship between apparent and true prevalence is illustrated in **Figure 19** and **Figure 20**.

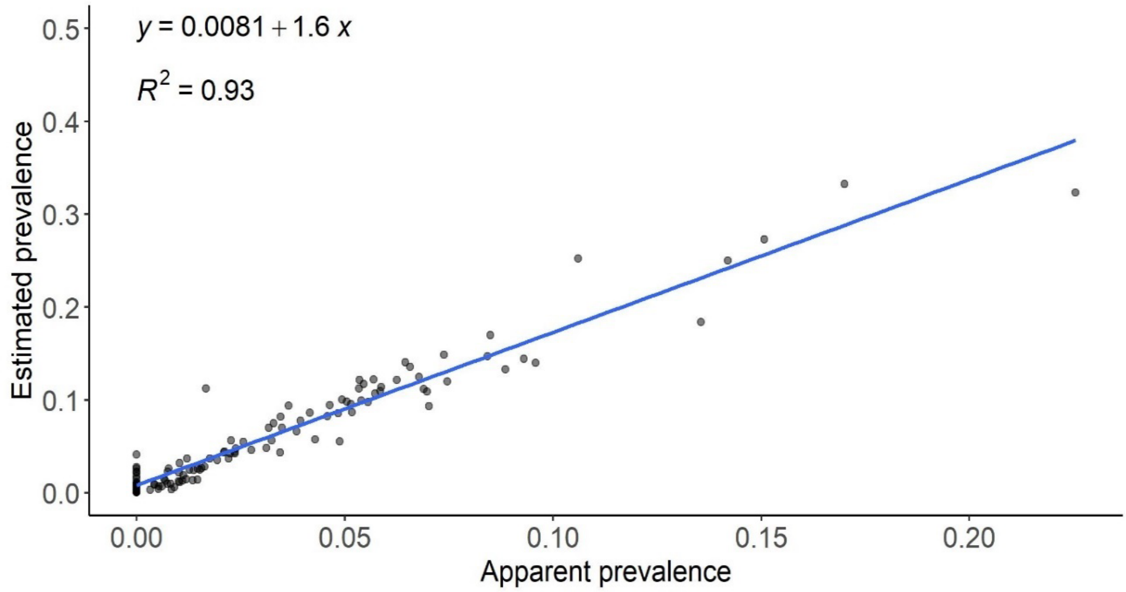



***Figure 19*** *Linear regression of apparent and estimated true paratuberculosis prevalence of primiparous cows based on the results of the Bayesian model. Each data point represents a herd.*

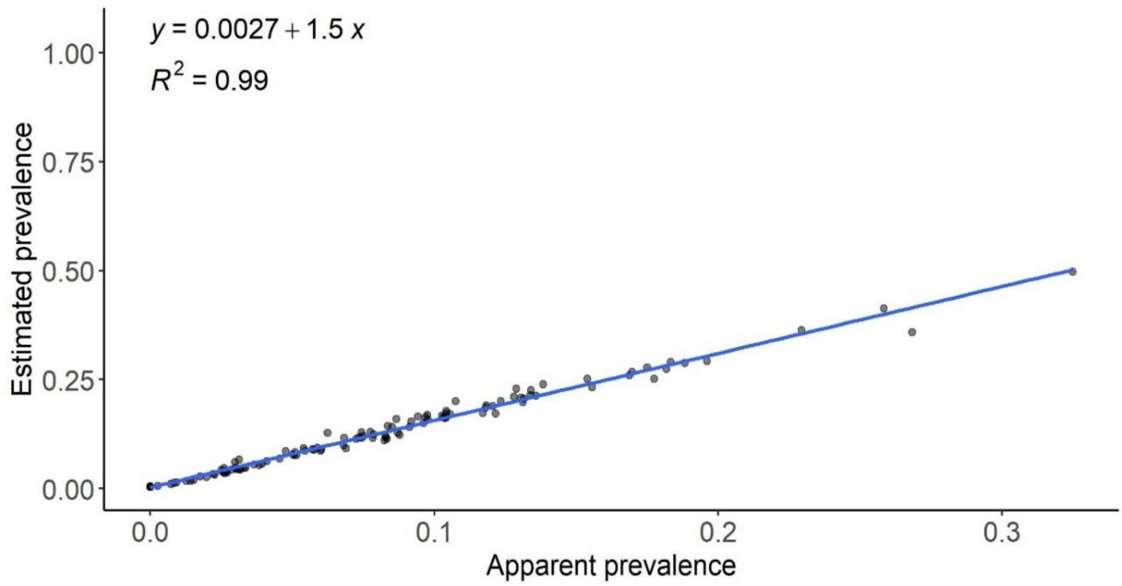
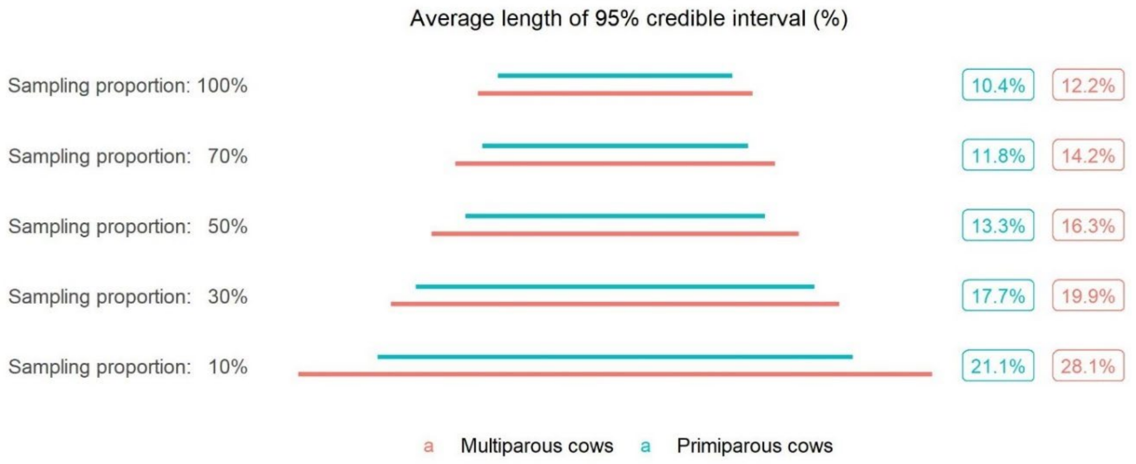



***Figure 20*** *Linear regression of apparent and estimated true paratuberculosis prevalence of multiparous cows based on the results of the Bayesian model. Each data point represents a herd.*

The intercept is close to zero in both cases, so the apparent prevalence multiplied by a simple factor can be used to obtain an estimate of the true prevalence. The true prevalence can be expressed as 1.6 times the apparent prevalence for primiparous cows and 1.5 times for multiparous cows.

### 5.2.4 Simulation study

Based on the model results from 500 simulated herds, the true prevalence fell within the 95% credible interval for 90% and 87% of herds for primiparous and multiparous cows, respectively. At a lower sampling percentage, there was no substantial change in the proportion of cases in which the credible interval contained the true parameter. The proportion remained around 90% for both primiparous and multiparous cows. However, the width of the credible interval increased considerably as shown in **Figure 21**, indicating the decreased accuracy of the estimate.




The percentages on the right show the length of the 95% credible interval.

***Figure 21*** *Reliability of Bayesian model estimates at different sampling frequencies.*

### 5.2.5 International model

The model infers the infection status of an individual herd and estimates the true prevalence for primiparous and multiparous cows from individual cow-level data and from prior information to incorporate regional characteristics.

In the analysis of the infection status of the simulated herds, based on the value of Bayes factor (BF), the infection status of two uninfected herds was classified as “strongly refuted” (BF<0.05). Both infected and uninfected herds appeared in the “refuted” (0.05≤BF<0.33) infection category. All herds rated as “supported” (3≤BF<20) and “strongly supported” (20≤BF) were indeed infected. For 42 herds, the results were “weakly refuted” (0.33≤BF<1) or “weakly supported” (1≤BF<3). Overall, the infection status was correctly identified in 66 of the 78 pseudo-herds that were not classified as weakly refuted or weakly supported.

When estimating the posterior probability of infection of the herds (PP), prior information on herd true prevalence was also taken into account. Out of 120 pseudo-herds, 53 truly infected herds were classified as “strongly supported” (0.95≤PP), with no herds falling into the “strongly refuted” (PP<0.05) category. Only one truly infected herd was misclassified into the “refuted” (0.05≤PP<0.25) group. The results were “weakly refuted” (0.25≤PP<0.50) or “weakly supported” ($0.50 \leq PP < 0.75$) for 48 herds. Overall, the infection status was correctly identified for 69 out of the 72 pseudo-herds with explicit results.

Summary measures obtained for within-herd true prevalence estimates demonstrated acceptable accuracy. For primiparous cows, the coverage, i.e., the proportion of herds where the simulated within-herd prevalence fell within the CrI, was at least 95%, demonstrating the accuracy of the Bayesian model and the prior distributions of primiparous CWHP in all regions. For multiparous cows the coverage was less accurate in Lombardy (85%) and Veneto (71%). The mean half length of the CrI measures the uncertainty of estimates. In all regions, it is of the same order of magnitude as the mean primiparous CWHP values, and it is slightly lower than the mean multiparous CWHP values, which seems acceptable from a practical point of view.

## 5.3 The relationship between PTBC infection and survival in Hungarian dairy herds

Based on the marginal model, the ELISA-positive group had an adjusted hazard ratio of 2.02 (95% CI: 1.5–2.7).This is a population-level estimate, meaning that ELISA positivity has a multiplicative effect on the hazard of culling of the whole cow population, with a factor of 2.02, taking into account the age and the herd of the cows.

According to the frailty model, the relative risk (adjusted hazard ratio) of the ELISA positive group was 2.03 (95% CI: 1.7-2.4). ELISA positivity has a multiplicative effect with a factor of 2.03 on the hazard of culling, taking into account the age and the herd of the cows. The between-herd heterogeneity is captured by the random effects. The variance of the logarithmically scaled random effects is 0.03, corresponding to a standard deviation of 0.17. The median absolute difference (MAD) of the hazard of two random clusters is 1.18. The effect of the exposure (2.03) compared to the MAD indicates a moderate effect of between herd heterogeneity on the hazard.

From the stratified Cox regression with individual level strata the unadjusted HR estimate of the ELISA positive group was 2.13 (95% CI: 1.75-2.60) meaning that ELISA positivity multiplies the hazard of culling with a factor of 2.13.

# 6 Discussion

## 6.1 Discussion of the results

### 6.1.1 Prevalence study

The evaluation of the PTBC situation in Hungary was updated using a robust methodology that allowed the inclusion of historical prior information. Although the estimates from the reassessment are in line with the results from the previous Hungarian study [4], the above changes turn the new results into a contribution to the field. In addition, the study results are expressed in terms of easily interpretable means. To my knowledge, the program and the rule of thumb based on the model to estimate the prevalence of interrelated subgroups in individual herds are without precedent in the veterinary literature.

Before placing the main results of the study in an international context, it is important to note that the estimates of HTP and CWHP vary widely across previous studies due to several factors that influence the results at different stages of the estimation process. Climate conditions, farming systems, herd size [67], genetic susceptibility of the breed [112] and the age distribution of the tested population affect the underlying infection level. Hot and dry climates and extensive or grazing farming systems are both associated with lower infection levels [113]. Older cows infected as a calf have a higher detectable infection level due to the progression of the disease toward stages with higher antibody production and more intensive shedding [70]. Apparent prevalences can also differ because there is no reference standard test to detect MAP infection. Factors such as the target condition (infected, infectious, clinical disease or seropositivity), sampling and testing strategies, the biological sample (tissue, serum, milk, or environmental faeces), the test method (ELISA, AGID, CFT, PCR), the test-kit and the positive-negative cut-off levels vary across studies. The age distribution of the population also affects the results as the sensitivity of the tests is age-dependent [18]. Finally, the quality of the data and the choice of analytical approach may further inflate the discrepancy between estimates. Therefore, it is important to exercise caution when comparing prevalence estimates from different studies.

The present HTP estimate is higher than the reported range of other studies from European countries using Bayesian methodology, namely 19.2-76.5% based on individual serum ELISA test in Italy [35] and 75–92% based on individual milk ELISA test in Denmark [106]. According to Nielsen and Toft’s literature review, HTP prevalence in farmed animals in Europe exceeds 50% and within-herd true prevalence is higher than 3–5% in several countries such as in Austria, the Netherlands, Germany, Italy [20]. The relatively high HTP in Hungary may be due to the absence of PTBC control programs and the existence of large herds, which are more likely to be infected by MAP and can experience substantial production losses related to PTBC, despite the stricter biosecurity provisions on large-scale farms [55, 67, 114].

According to Nielsen and Toft, typically, a high proportion of herds are infected with MAP, however, only a low proportion of animals screened positive within herds ranging from 2.7% to 15% [20]. Our marginal CWHP estimates are consistent with these results. These CWHP estimates were later used as prior information when estimating the prevalence of individual herds. The higher CWHP in multiparous cows can be attributed to the higher infectious pressure [70] and increasing proportion of ELISA-detectable MAP infections with the age of animal [18].

The positive difference between the mean and the median estimate for both $CWHP_1$ and $CWHP_2$ indicate right skewness of the posterior distributions. This means that there is a high probability that $CWHP_1$ or $CWHP_2$ is unusually large.

The high maximal correlation of $CWHP_1$ and $CWHP_2$ demonstrates a strong link between the conditional within-herd prevalence of primi- and multiparous cows from the same the herd. Apart from the correlation between the prevalence of primiparous and multiparous cows, we can quantify the differences using prevalence ratio. This ratio is a relative risk showing how much more common the infection is among multiparous cows than among primiparous cows. The high prevalence ratio of infection found between multiparous and primiparous is a measure of the effect of the parity and the age of the cows.

The random herd and parity effects correspond to the heterogeneity of the prevalence between herds. Furthermore, they highlight the source of correlation between parity groups in herds. The estimation of its SD cannot be directly interpreted. Instead, to assess the heterogeneity of the prevalence between herds, we can derive the standard deviation of CWHP based on the SD of the random effects. In the application, the SD of $CWHP_1$ and $CWHP_2$ are of the same magnitude as their means indicating strong heterogeneity of the disease prevalence between herds.

The lack of effect of region and herd size on the HTP and CWHP estimate suggests a homogeneity of the source population and justifies the selection of large intensive dairy cattle herd from Hungary as target population.

The sensitivity analyses showed that the HTP estimate from the model is sensitive to the change in HTP prior. Care should therefore be taken to choose the right prior distribution. It is reasonable to base the choice of the HTP prior on historical data or expert opinion. The mean $CWHP_1$ and $CWHP_2$ estimates proved to be sensitive to the change of the age-dependent sensitivity parameters. Varying specificity parameter did not affect the model estimates. By adopting an informative uniform prior distribution for specificity parameters within the range of published values (0.986-0.996) independently for each herd, rather than assuming a common fixed specificity value of 0.995, I found no substantial changes in the posterior estimates of HTP, $CWHP_1$, and $CWHP_2$ caused by the introduced between-herd heterogeneity. In addition, assuming the same specificity in all herds, with an informative uniform prior in the interval 0.986-0.996, also did not markedly affect the model estimates. When simulating data with different underlying HTP values from the range of the historical prior the main estimates remained on target.

The TP/AP ratios of 1.5 and 1.6 of individual herds found for primi- and multiparous cows are comparable to the ratios of 1.4, 1.5 and 1.5 published in the Danish study by Verdugo et al. [106] for three consecutive years, 2011, 2012 and 2013 in overall.

It is important to note that the international model used synthetic datasets to illustrate the approach, and therefore, regional prevalence estimates should not be interpreted as measures of PTBC prevalence in these areas. When generating synthetic data, I undertook a pragmatic approach: resampling a Hungarian herd and replacing test results to reflect regional herd sizes and prevalences. This technique may not fully capture the unique characteristics of herds in other regions, such as different age distributions, genetic backgrounds, or management practices, all of which can influence the age-sensitivity relationship and the true prevalence. The results for non-Hungarian regions demonstrate the flexibility of the model rather than its accuracy in these specific locations, where no local validation has taken place. Yet, the international model can be customized for any region using real local data and prior information.

When assessing the infection status of a herd, misclassifications and indeterminate classifications arise. From a Bayesian perspective, misclassification of infected and not infected herds occurs because probabilities should preferably be interpreted as indicators rather than used as classifiers. Errors arise from forcing clear-cut decisions that are fundamentally based on probabilistic information. Indeterminate classification may arise from a low prevalence in a high herd prevalence region or from a small sample size. If the herd with weak results was not fully tested, then herd managers may try to test a larger sample of cows. In case of a low prevalence in a high herd prevalence region, evaluating the effectiveness of herd management practices in preventing the spread of the disease between herds is recommended. If the practices introduced are appropriate, the herd can be considered uninfected; otherwise, it is reasonable to assume that the herd is infected.

To assess the potential level of infection, the estimated CWHP prevalences need to be evaluated along with the width of the credible intervals. This approach provides a comprehensive picture of the infection status of the herd. To confirm the results obtained or to observe changes in the infection status of the herd, it is recommended to retest the animals after a certain time interval and analyze the test results by fitting the proposed Bayesian model with updated, most recent prior distributions. Given the economic and health damage caused by PTBC [55], it is worthwhile to take measures to prevent and reduce infection even if the farm has been classified as “weakly refuted”.

The prevalence studies have several strengths. First, the large sample size of tested cows and the nationwide coverage of 24.3% of all dairy cows in 2019 in Hungary is unique. Second, all milking cows in the screened herds were sampled, which is not a common occurrence due to the high costs involved. The use of a historical HTP prior helps to reinforce the reliability of our model. The reliability of the estimates of our model was tested by sensitivity runs and simulation study, as suggested by McAloon et al. [105]. These analyses confirmed the stability of the model and the reliability of the estimates.

The voluntary nature of the testing scheme in our study may have introduced selection bias. However, a previous study by Ritter et al. [115] using environmental fecal samples from 174 Canadian dairy farms found no association between participation in a voluntary PTBC control program and herd-level MAP prevalence. In addition, the participation rate was high, 26.8 % of all performance tested herds. This suggests that self-selection is unlikely to have significantly altered our results.

The ELISA test has a low sensitivity in the animal-level detection of MAP. To detect MAP infection milk ELISA tests were used, which perform similarly to serum ELISA if calving related cow characteristics are taken into account for the latter [116, 117]. Further, milk tests on lactating cows have the advantage of being non-invasive and easy to perform from routinely collected samples. The cost effectiveness of these tests is counterbalanced by their inability to detect animals at the early stages of MAP infection due to their reduced sensitivities during the subclinical phase of MAP infection when antibody levels and fecal shedding are low [80]. Nevertheless, at the herd level, these tests may perform well since the high to absolute specificity allows inference of herd infection from a single positive test result.

### 6.1.2 Survival study

The results of our survival study are in line with previous research. Smith et al. [56] found that MAP infection detected by serum ELISA and faecal occultation tests was positively correlated with culling. In a study of nearly 8,000 cows in 38 US dairy herds, they found that cows with positive serum ELISA tests were 1.8 times more likely to be removed from the herd compared to their ELISA negative counterparts [54]. This value is very close to my estimated HR of around 2. In a previous study in a Hungarian dairy herd of 900 cows, MAP ELISA positive cows also had higher culling and mortality rates than their seronegative counterparts [52].

The present study focuses on the expected survival time from PTBC ELISA testing to culling. Since the infection precedes serological testing, the survival time from infection is not known. Thus, the research does not answer the theoretical question of how much earlier PTBC-positive individuals are culled relative to the time of infection. A further limitation is that the positive and negative cow groups do not coincide with the group of truly infected and truly uninfected cows due to the low sensitivity of the ELISA test. This means that the true difference is likely to be larger than the detected risk difference.

## 6.2 Discussion of the methodology

### 6.2.1 Prevalence study

#### 6.2.1.1 The applications

I presented a novel Bayesian modeling approach for disease prevalence that accounts for the hierarchy of data and the correlation at subgroup level. The proposed methodological innovation fits well into the existing literature as generalization of the well-established Hanson’s and Branscum’s model [24, 25]. The correlated prevalences of the subgroups are modelled by beta distributions within the same framework. The use of beta distribution to model the true prevalence of infectious diseases is widely accepted in Bayesian methodology [32, 34, 35, 55]. In these studies, one common assumption is the homogeneity of disease prevalence within commercial units (herds). In my approach, subgroups are allowed to have different marginal means, and the association of the prevalence of the subgroups within the same herd is accounted for through correlated random herd effects. The correlation coefficient is estimated from the model. Modelling subgroup level correlated prevalence by beta distributions allows estimating subgroup means and gives new insights into their interrelation.

This method is designed to estimate the mean prevalence, while also allowing for convenient calculation of the median. For methods directly modeling the median, the calculation of the mean estimate is in general a lengthy, cumbersome procedure. Median and mean can highlight different aspects of the posterior distribution and, when used together provide a natural measure of skewness. Subgroup means can be summarized into a meaningful weighted population average. The concept of mean is familiar to end-users; the interpretation of the results is seamless. The posterior distributions from the model offer the possibility to compute various additional measures from MCMC iterates and via simulation techniques.

Both the susceptibility of animals to PTBC infection and the detectability of infection are age-dependent [18]. The parity used as a subgroup in the model is a natural categorization of the cows and is strongly correlated with the age of the animals. The breakdown of the herd by parity results in more or less homogeneous groups with highly different infection prevalence. It is defined by important milestones in a cow's life and is compatible with herd management categories. The probability of a positive ELISA test result is two to three times lower for primiparous cows than for multiparous [78, 118]. With the novel methodology, it is possible to give more insight into PTBC status of the country at the level of parity groups.

In individual herds, for specificities close to 1, the ratio of apparent to true prevalence depends primarily on the sensitivity of the test according to relation (4.21) [1]. To examine the ratio of apparent to true prevalence, we also need to take into account the age dependence of sensitivity as described by the Meyer formula [17]. The vast majority of primiparous cows are 2-3 years old and the age variance is low, so the sensitivity in this category is relatively stable. Although age variance is higher for multiparous cows, sensitivity at higher ages is less dependent on age variation and heterogeneity according to the Meyer formula [17]. Therefore, the relationship between true and apparent prevalence is linear and stable and can be well estimated. Fitting a linear regression is theoretically appropriate.

For the international model, prior distributions were defined according to the following principles. Ideally, such priors are informed by posterior results from previous studies. Hungary and Denmark have published posterior distributions from PTBC prevalence studies using milk ELISA test (section 5.1), [106]. For these two countries, posterior HTP and CWHP distributions were directly used as prior along with the age-dependent sensitivity and constant specificity values from the Hungarian study. For the other four regions, where only serum ELISA test results were available, HTP and CWHP priors were derived from published expert elicitation [34, 35, 106, 107]. Consequently, the resulting posterior distributions may suggest bigger uncertainty than observed in the reality.

For the variance of the random herd effect and additive random parity effects, I initially derived priors from region-specific expert elicitation, but this led to unrealistically high variance estimates and consequent shrinkage toward the mean in some regions. To ensure greater accuracy in the estimates, I used the posterior distribution obtained from the Hungarian study as the prior distribution. While this is a limitation of our approach, it is important to note that assuming the same prior of the variance of the random effects does not imply the same posterior variance of the within-herd animal level prevalence across all regions. The posterior variance of the CWHP also depends on the regional mean of CWHP and the observed data of the target herd.

An age-dependent fixed value was applied for the sensitivity of the diagnostic test, together with a fixed value for its specificity. Using prior distributions instead of fixed values may result in unstable or biased estimates and may lead to an unidentifiable model. The use of fixed values is confirmed by detailed sensitivity analysis for the possible range of Se and Sp, based on available published values.

Combined with high-quality local data and prior information, the international model serves as a robust and adaptable framework to support evidence-based decision making at the herd level. The novelty of the method is the estimation of both the probability and the Bayes factor of infection of the target herd, combined with the estimation of infection prevalence within homogeneous subgroups of the herd, specifically parity groups. The approach can be readily generalized to other types of subgroups, such as different breeds, lactation stages, housing pens, or health status.

The methodology used in the prevalence study is the direct application of a novel class of models, the Distribution based marginalized multilevel models. The modeling process illustrates well the merits of the methodology.

#### 6.2.1.2 The added value of Distribution based marginalized multilevel models

The general dispersion modelling framework (4.7) provides great flexibility in the analysis of the variances and correlations of cluster-specific random effects. The role of the random effects in the model can be assessed and interpreted by examining the standard deviations of the (appropriately transformed) conditional, cluster specific means, evaluated at certain values of the marginal means and possibly other parameters of the DMMM. The decomposition of the total marginal variance of the outcome into the mean of the conditional variance plus the variance of the conditional mean (4.8) offers another way to assess the weight and importance of the cluster specific random effects relative to the observed outcome data.

The Pearson correlations of the random effects quantify the linear association between them. From these, the maximal correlations [97], or conceptually equivalently, the normal scoring dependence measures [93, 96] between the conditional means (4.2) can be established. They show the strength and direction of monotonic relationship between the conditional means.

I investigated the relationship between conditional and marginal means, examining how the conditional means change within a cluster as the marginal means change. This difference generally depends on the random effect, but under certain conditions, the conditional means can be transformed so that their difference locally approximates the change in the marginal mean. The transformation eliminates the heterogeneity of magnitudes attributable to the common random effect from the difference of the transformed conditional means. Consequently, the differences of the transformed conditional and the corresponding marginal means can be compared without respect to the actual magnitude of the common random effect. The transformation pulls the difference in the conditional means towards the difference in the marginal means. This is an approximate decomposition of the transformed conditional mean into the sum of the marginal mean and the transformed cluster-specific random effect. Based on this decomposition, a factor of proportionality reflecting the relative changes of transformed conditional and marginal means was introduced. If the marginal mean is a monotonically increasing and differentiable function of a linear predictor, then the cluster-specific means can be approximated by generalized linear model with random effect, and the marginal means by a generalized linear model, both having the same link function and covariates. In the two models, the ratio of the coefficients or intercepts of the fixed effects is the factor of proportionality. If the coefficient of the induced conditional model is greater/smaller than the coefficient of the induced marginal model means that the association between the covariate and the outcome is more/less pronounced within clusters than marginally at the population level.

The DMMM is a novel, innovative modelling technique appropriate to calculate the true, underlying population level and herd-level mean prevalences of relevant within-herd subgroups, taking into account the hierarchical structure of the data. This method sheds light on the interrelation of subgroup prevalences and enables us to determine whether, and to what extent, the effect of model covariates is more pronounced at the population or herd level.

### 6.2.2 Survival study

Latent variables are not observed or measured directly; their values are inferred from other observed variables. They are concepts or constructs having a potential effect on some observed variables. In complex hierarchical models, random cluster or subject effects are represented by latent variables. These variables are part of even the most basic statistical models, e.g. the ordinary linear regression - although they are rarely mentioned as latent variables in this context; the unexplained, random sampling variation incorporates the effect of unknown, unmeasured or just overlooked variables on the outcome. Individual values of the latent variables are of secondary interest, and usually we model only their variance. Cluster-level latent variables account for the association of the observations within the same cluster [119, 120].

The heterogeneity of the data (represented by individual or cluster level latent variables) can be handled using marginal models, conditional models and the mixture of them (MMM). Although these modelling strategies serve a common purpose, they have different targets of inference thus the interpretation of their results is different. The choice between these options should be made according to the research question, on subject-matter grounds.

The marginal estimate is adequate to support country or region level infection-control decisions but cannot be used to draw conclusions for herds or for individual animals. In mixed models for clustered data, within-cluster correlation is related to cluster-level random effects. The values of the individual random effects are of secondary interest and usually not inferred, thus we only model their variance. The heterogeneity between clusters is captured by the spread of the random effects. In the context of survival analysis random effects are called frailties, which act multiplicatively on the hazard. The estimated variance of the logarithmically scaled random effects is a measure of the unobserved heterogeneity of the clusters. By including the frailty terms, we incorporate the implicit correlation structure of the data into the survival model. The estimate from the frailty model is a conditional estimate, which can be interpreted both at a herd and at a cow level, meaning that a randomly selected ELISA positive cow has a 2.03 times higher hazard to be culled than a randomly selected ELISA negative cow from the same herd.

The stratified Cox model [111] is an extension of the Cox proportional hazards model that handle heterogeneity by stratification. Observations are grouped into strata according to the levels of all possible combinations of selected categorical variables. The equation of the general Cox model is specified separately for each stratum, with different baseline hazard functions but the same coefficients across strata. Thus, the resulting HR would be the same for each stratum. This model accounts for the heterogeneity between strata by allowing different baseline hazard functions for each stratum. Additionally, stratification covariates do not have to satisfy the proportional hazard assumption. The estimation is done by maximizing the product of stratum-specific partial-likelihood functions. Baseline hazards are nuisance parameters and are cancelled out. The resulting HR – although its value is the same in each stratum – will be conditional on the strata. The terminology for this is unadjusted HR, in contrast to a crude HR from a model without stratification and to an adjusted HR (aHR) from a model including additional covariates as adjusting factors.

Baseline hazard functions in a stratified Cox analysis show a strong analogy to latent variables in mixed models, though their effect is not condensed into a sole random number from a prespecified distribution, but is described by an arbitrary function. Baseline hazard functions are nuisance parameters which cancel out from the likelihood. These functions can be estimated, but the estimation is usually not effective. The resulting HR estimates can be interpreted at stratum level. That said, the stratified Cox model can be considered as a conditional model. The HR estimate is adjusted for matching factors, and it means that a randomly selected ELISA positive cow has a 2.13 times higher hazard to be culled than a randomly selected ELISA negative cow from the same herd and from the same age group.

While generally, the stratified Cox regression uses covariate-based strata to relax the proportional hazard assumption, here I used the individual layers to allow for different baseline hazard function for all subjects. The special, stratified Cox methodology with individual level strata was inspired from one of my human epidemiological studies [121].

In this epidemiological research, I used data from linked nationwide Danish registries to conduct a cohort study of the association between herpes zoster (HZ) and dementia during 1997–2017 [121]. The risk of dementia in patients older than 40 years who were first diagnosed with HZ (case group) was compared with that of a random control group matched 1:5 for sex, year of birth and date of diagnosis, selected from the general population. I identified 247,305 people with HZ and 1,235,890 matched comparators. Based on our directed acyclic graph, a large number of possible baseline confounders were identified. Follow-up started on the day of the patient's HZ diagnosis in the case group, for both patients and their matched 5 controls who had not yet undergone HZ at the start of follow-up. Patients were followed until the earliest of the following: first diagnosis of dementia, death, emigration, or end of the study (December 31, 2017). People in the comparison cohort who were diagnosed with HZ during follow-up were censored and included in the HZ cohort (with their own matched comparators) at the time of diagnosis. They contributed follow-up time to the comparison cohort until the date of HZ diagnosis and to the exposed cohort afterwards. This decision was based on an important lesson learned. Several prior studies [122–125] may have been affected by selection bias, with unknown magnitude or direction. In these studies, follow-up began at the start of the study period or on the diagnosis date for people with HZ, whereas comparators were selected from a population without HZ at any time during the study period. This sampling technique, which conditions on future information on the absence of HZ, may have selected people who were disproportionally healthy (and therefore less likely to develop HZ) or who died during the study period (and therefore had less chance to develop HZ), both of which would affect dementia risk. Cohorts were characterized by covariates and accumulated person-time. To account for the correlation within the same matching set (strata) I used stratified Cox regression to compute unadjusted HRs with 95% CIs for the HZ cohort compared with the matched comparison cohort. Because visual inspection of log-log plots showed that curves crossed or merged after 6-12 months of follow-up for most comparisons, I divided the follow-up period into 0-1 year and >1 year after the index date.

The human study served as a model for the survival study of the cows in terms of the methodology, of the solutions to possible complications and of the type of results the method can provide. The special methodology was validated on a large dataset, with a considerable set of confounders and on complex epidemiological questions. To my knowledge, this method is so far unknown in veterinary science. The experiences from this human study pave the way for using this innovative methodology on big databases and complex questions in veterinary science.

The stratified Cox methodology with matched exposed and control patients is an effective tool to fit conditional, subject specific survival models to time-to-event data. Although a few electronic databases of time-to-event veterinary data exist, to my knowledge, this approach has not yet been applied in veterinary science.

# 7 Conclusions

## 7.1 Major conclusions

The awareness of the difference between the apparent and the true prevalence is key in the control of infectious diseases. Bayesian methods can be used to estimate the true prevalence, helping decision-makers assess the production losses caused by the infection and develop appropriate preventive and control measures. The complex hierarchical structure of the data requires a new modeling approach that allows inference of correlated subgroup means within economic units. In response to this need, the DMMM framework is a useful novel tool for prevalence estimation providing simple and flexible control over the marginal and conditional means.

The synergy of DMMM and Bayesian methodology is innovative and forward-looking while at the same time fitting well into the proven framework led by Hanson, Branscum and McAloon. Incorporating correlated within-herd subgroup prevalences allows us to estimate various measures characterizing the interrelation of the prevalences in important subgroups. This is a powerful tool providing easily interpretable subgroup-level mean estimates. Allowing for a direct, distributional representation of subgroup level prevalences, this approach is also convenient to use as a basis for the calculation of further measures. For instance, the conditional estimates can be summarized into population means with simple weighting. Median and other characteristics of interest are easy to infer from posterior distributions.

The application of the method to estimate the prevalence of PTBC in Hungary illustrate the implementation and the merits of the novel methodology. The use of the new methodology along with a more recent and uniquely large dataset and historical priors confirm and refine the results of the latest PTBC screening in Hungary.

The Bayesian model provides an estimate of the true within-herd prevalence of infection in primiparous and multiparous cows for individual herds. To the best of my knowledge, this is the first Bayesian model for estimating the true PTBC infection at the level of correlated subgroups within an individual herd in Hungary using priors from a national PTBC survey. Based on the results of the model, I described a straightforward method for estimating the true prevalence for situations in which practitioners do not have the opportunity to run the model. The international model's potential utility for prevalence estimation and infection status inference was also demonstrated across diverse regional settings.

The culling data (survival dataset) are time-to-event data; thus, the use of survival analysis is appropriate when modelling the relationship of seroprevalence and culling. The stratified Cox model with individual observations as strata can be considered analogous to the DMMM, providing at the same time conditional and quasi-population level estimates.

The results of the applications showed that PTBC infection is present in almost every herd in Hungary. Therefore, setting up an obligatory, routine control-program is desirable. The strategy of infection status reassessment and the choice of control measures should be based on the prevalence of the disease. The true, underlying prevalence was proven to markedly differ from the apparent prevalence. The proposed methodological innovations could provide a framework for assessment of the infection status of the country.

According to our results, there is substantial heterogeneity of the PTBC prevalence at herd level, there can be herds with particularly high infection level. Thus, the PTBC prevalence of the individual herds is also important to estimate. The online application we provide, and the thumb rules are quick and easy tools to estimate the true prevalence based on test results.

The two times culling risk of the seropositive cows – even if the ELISA test results are not known to decision makers – show that these cows present a risk to the cost-effective operation of the farm. Careful consideration of pros and cons should be given upon deciding on keeping these animals in the herd.

## 7.2 Future directions, suggestions

Applied to a region or to a country, the proposed hierarchical modeling framework for estimating PTBC prevalence in dairy cattle - taking into account both herd-level and within-herd parity group structure - facilitates animal health management decision-making at the national level. It paints a clearer picture on the country's infection status and helps to adopt an adequate surveillance strategy. It lays down a methodological foundation for regularly assessing screening results and for evaluating the efficiency of control measures. Future prevalence studies could consider incorporating other locally relevant factors such as climate, production type, breed, age group, genetic markers, or cohabiting species into the model, to improve fit and to better reflect real-world observations.

The model for estimating the true infection prevalence of subgroups in individual herds along with the downloadable tool and the rule of thumb are a valuable tool for practitioners to evaluate the position of their herd's infection level within the country and to understand within-herd dynamics of the infection. Based on the results, they can reshape management practices likely to favor the spread of infection. Raising the awareness of the difference between apparent and true prevalence of the practitioners not only supports decision making but could largely contribute to the success of a national-level control program.

DMMM, the novel methodological framework is a flexible and broadly applicable tool. While it was originally inspired by the problem of prevalence estimation of PTBC in dairy cattle, it holds considerable potential for wider use. First, it could be applied to other livestock species such as sheep, goats, or pigs, and extended to different infectious diseases where hierarchical or nested sampling structures are common. These might include mastitis, bovine viral diarrhea, or respiratory conditions. Furthermore, the model lends itself to application in other countries, allowing for cross-country comparisons of prevalence under varying herd structures, management practices, or diagnostic protocols. The method is also relevant beyond veterinary diagnostics, in public health and epidemiology, where nested data structures – such as female and male patients within hospitals and across the country – are common. In addition to the applications where misclassification is due to the imperfect diagnostic tests, the framework could be used to evaluate the validity of disease records found in veterinary or human health registries, where misclassification can happen due to diagnostic or administrative errors.

With technological advances, tracking data is now also available in veterinary medicine. Survival analysis is the sine qua non of the adequate analysis of time-to event data. It is a versatile tool that can be used to analyze mortality, assess the risk of an event, predict the prognosis of a disease and identify factors that influence the above. Thus, it is important to promote the use of its various methods in veterinary medicine. Survival analysis methods, while well-established in human medicine, remain underutilized in veterinary science but offer significant potential benefits. These techniques can be particularly valuable for analyzing time-to-event data such as time to disease onset, treatment failure, or death in animals. Unlike traditional methods, survival analysis appropriately handles censored data – cases where the event has not occurred by the study's end – which is common in veterinary studies due to follow-up loss or study duration limits.

Advanced survival models, including marginal, frailty or stratified Cox proportional hazards models accounting for clustered data typical in veterinary epidemiology (e.g. animals nested within farms or geographic regions) can provide deeper insights into disease dynamics, risk factors, and treatment effects in animal populations. Integrating these underused methodologies into veterinary research and practice promises enhanced disease control strategies and better animal health outcomes.

Although stratified Cox models are widely used in human epidemiology, they use with individual-level strata is scarce and to my knowledge there are no previous veterinary applications of this kind at all. Stratified Cox proportional hazards models with individual-level strata offer a powerful tool for analyzing time-to-event data when there is heterogeneity across subjects that cannot be fully captured by measured covariates. In this approach, each individual is treated as a separate stratum, allowing the baseline hazard function to vary between strata while assuming that the effects of explanatory variables (e.g. infection status, seroprevalence or treatment) are constant across them. By stratifying on individual-level units, the model controls for these baseline differences, thus reducing bias.

Finally, by integrating survival analysis methods with Bayesian latent class prevalence modelling, researchers could assess the true impact of the infection - rather than the impact of the seroprevalence - on economic outcomes such as productivity, fertility, or treatment costs.

These potential extensions highlight the versatility of the methodology used in the present study and its promise as a valuable tool in both veterinary science and broader biomedical research.

# 8 New and novel scientific results

**1.** I introduced a new class of models, the *Distribution based marginalized multilevel models*. This framework provides easily interpretable marginal and conditional estimates and gives new insight into the interrelation of subgroups.

**2.** I illustrated the use of the new framework by modeling the prevalence of PTBC infection in Hungarian dairy cattle herds. I consolidated previous findings using a recent, extensive dataset and historical priors. While the apparent HP was 96.6% based on ELISA test result, mean HTP was estimated to a lower level of 92.5% (95% CrI: 88.0%; 96.1%). Mean CWHP was estimated to 8.4% (6.6%; 10.4%) for primiparous and 15.8% (13.5%; 18.4%) for multiparous cows, in contrast with 3.3% and 8% of AP, respectively.

**3.** I modelled PTBC prevalence in parity groups within individual herds and developed a rule of thumb for simple estimation of true prevalence. The true prevalence can be expressed as 1.6 times the apparent prevalence for primiparous cows and 1.5 times for multiparous cows. The model can be customized for any international region using local data and prior information.

**4.** I fitted a special stratified Cox regression model to human data to examine the effect of herpes zoster on dementia. Building on the lessons learned from that study, I adapted this technique to the veterinary field and performed a survival analysis to assess the impact of PTBC infection on the longevity of Hungarian dairy cows.

# 9 References


1. Rogan WJ, Gladen B (1978) Estimating prevalence from the results of a screening test. Am J Epidemiol 107:71–76. https://doi.org/10.1093/oxfordjournals.aje.a112510

2. Heagerty PJ, Zeger SL (2000) Marginalized Multilevel Models and Likelihood Inference. Statistical Science 15:1–26. https://doi.org/10.1214/ss/1009212671

3. Hanson TE, Johnson WO, Gardner IA (2003) Hierarchical models for estimating herd prevalence and test accuracy in the absence of a gold standard. JABES 8:223. https://doi.org/10.1198/1085711031526

4. Ózsvári L, Lang Zs, Monostori A, Kostoulas P, Fodor I (2020) Bayesian estimation of the true prevalence of paratuberculosis in Hungarian dairy cattle herds. Prev Vet Med 183:105124. https://doi.org/10.1016/j.prevetmed.2020.105124

5. Bokusheva R, Kimura S (2016) Cross-Country Comparison of Farm Size Distribution. OECD, Paris

6. Broughan JM, Judge J, Ely E, Delahay RJ, Wilson G, Clifton-Hadley RS, Goodchild AV, Bishop H, Parry JE, Downs SH (2016) A review of risk factors for bovine tuberculosis infection in cattle in the UK and Ireland. Epidemiology & Infection 144:2899–2926. https://doi.org/10.1017/S095026881600131X

7. Krieger M, Eisenberg S, Donat K, Campe A (2023) High-Impact Risk Factors for Mycobacterium avium ssp. paratuberculosis in Dairy Herds in Germany. Animals 13:1889. https://doi.org/10.3390/ani13121889

8. Puerto-Parada M, Arango-Sabogal JC, Paré J, Doré E, Côté G, Wellemans V, Buczinski S, Roy J-P, Labrecque O, Fecteau G (2018) Risk factors associated with Mycobacterium avium subsp. paratuberculosis herd status in Québec dairy herds. Prev Vet Med 152:74–80. https://doi.org/10.1016/j.prevetmed.2018.02.010

9. Rothman KJ, Greenland S, Lash TL (2008) Modern epidemiology, 3rd edition, thoroughly revised and updated. Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia

10. Greiner M, Gardner IA (2000) Epidemiologic issues in the validation of veterinary diagnostic tests. Prev Vet Med 45:3–22. https://doi.org/10.1016/S0167-5877(00)00114-8

11. Diggle PJ (2011) Estimating Prevalence Using an Imperfect Test. Epidemiology Research International 2011:e608719. https://doi.org/10.1155/2011/608719

12. Reiczigel J, Földi J, Ózsvári L (2010) Exact confidence limits for prevalence of a disease with an imperfect diagnostic test. Epidemiology & Infection 138:1674–1678. https://doi.org/10.1017/S0950268810000385

13. Lang Z, Földi J, Ózsvári L, Reiczigel J (2013) Relationship between seropositivity and prevalence on the basis of epidemiological analysis of data of BHV-1 eradication program in Hungary (Szeropozitivitás és prevalencia összefüggései hazai BHV-I- mentesítés adatainak járványtani elemzése alapján). Magy Allatorvosok Lapja 135:525–534

14. Lang Z (2015) Epidemiológiai és ökológiai mutatók statisztikai és valószínűségelméleti modellezése, térbeli szerkezetének vizsgálata. https://doi.org/10.14751/SZIE.2015.045

15. Lang Z, Reiczigel J (2014) Confidence limits for prevalence of disease adjusted for estimated sensitivity and specificity. Prev Vet Med 113:13–22. https://doi.org/10.1016/j.prevetmed.2013.09.015

16. Reiczigel J, Singer J, Lang Z (2017) Exact inference for the risk ratio with an imperfect diagnostic test. Epidemiology & Infection 145:187–193. https://doi.org/10.1017/S0950268816002028

17. Meyer A, Bond K, Van Winden S, Green M, Guitian J (2018) A probabilistic approach to the interpretation of milk antibody results for diagnosis of Johne's disease in dairy cattle. Prev Vet Med 150:30–37. https://doi.org/10.1016/j.prevetmed.2017.11.016

18. Nielsen SS, Toft N, Okura H (2013) Dynamics of Specific Anti-Mycobacterium avium Subsp. paratuberculosis Antibody Response through Age. PLOS ONE 8:e63009. https://doi.org/10.1371/journal.pone.0063009

19. Gardner I (2002) The utility of Bayes' theorem and Bayesian inference in veterinary clinical practice and research. Australian Veterinary Journal 80:758–761. https://doi.org/10.1111/j.1751-0813.2002.tb11347.x

20. Nielsen SS, Toft N (2009) A review of prevalences of paratuberculosis in farmed animals in Europe. Preventive Veterinary Medicine 88:1–14. https://doi.org/10.1016/j.prevetmed.2008.07.003

21. Branscum AJ, Gardner IA, Johnson WO (2005) Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. Preventive Veterinary Medicine 68:145–163. https://doi.org/10.1016/j.prevetmed.2004.12.005

22. Gelman A, Vehtari A, Dunson DB, Rubin DB, Stern HS, Carlin JB (2014) Bayesian data analysis, Third edition. CRC Press, Boca Raton

23. Rutjes AWS, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PMM (2007) Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess 11:iii, ix–51. https://doi.org/10.3310/hta11500

24. Branscum AJ, Gardner IA, Johnson WO (2004) Bayesian modeling of animal- and herd-level prevalences. Prev Vet Med 66:101–112. https://doi.org/10.1016/j.prevetmed.2004.09.009

25. Hanson TE, Johnson WO, Gardner IA, Georgiadis MP (2003) Determining the infection status of a herd. JABES 8:469–485. https://doi.org/10.1198/1085711032561

26. Meletis E, Conrady B, Hopp P, Lurier T, Frössling J, Rosendal T, Faverjon C, Carmo LP, Hodnik JJ, Ózsvári L, Kostoulas P, van Schaik G, Comin A, Nielen M, Knific T, Schulz J, Šerić-Haračić S, Fourichon C, Santman-Berends I, Madouasse A (2024) Review state-of-the-art of output-based methodological approaches for substantiating freedom from infection. Front Vet Sci 11:. https://doi.org/10.3389/fvets.2024.1337661

27. Johnson WO, Su C-L, Gardner IA, Christensen R (2004) Sample size calculations for surveys to substantiate freedom of populations from infectious agents. Biometrics 60:165–171. https://doi.org/10.1111/j.0006-341X.2004.00143.x

28. McAloon CG, Doherty ML, Whyte P, O'Grady L, More SJ, Messam LLMcV, Good M, Mullowney P, Strain S, Green MJ (2016) Bayesian estimation of prevalence of paratuberculosis in dairy herds enrolled in a voluntary Johne's Disease Control Programme in Ireland. Prev Vet Med 128:95–100. https://doi.org/10.1016/j.prevetmed.2016.04.014

29. Branscum AJ, Johnson WO, Thurmond MC (2007) Bayesian Beta Regression: Applications to Household Expenditure Data and Genetic Distance Between Foot-and-Mouth Disease Viruses. Australian & New Zealand Journal of Statistics 49:287–301. https://doi.org/10.1111/j.1467-842X.2007.00481.x

30. Muskens J, Barkema HW, Russchen E, van Maanen K, Schukken YH, Bakker D (2000) Prevalence and regional distribution of paratuberculosis in dairy herds in the Netherlands. Veterinary Microbiology 77:253–261. https://doi.org/10.1016/S0378-1135(00)00310-2

31. Bauman CA, Jones-Bitton A, Menzies P, Toft N, Jansen J, Kelton D (2016) Prevalence of paratuberculosis in the dairy goat and dairy sheep industries in Ontario, Canada. Can Vet J 57:169–175

32. Liapi M, Leontides L, Kostoulas P, Botsaris G, Iacovou Y, Rees C, Georgiou K, Smith GC, Naseby DC (2011) Bayesian estimation of the true prevalence of Mycobacterium avium subsp. paratuberculosis infection in Cypriot dairy sheep and goat flocks. Small Ruminant Research 95:174–178. https://doi.org/10.1016/j.smallrumres.2010.09.010

33. Lombard JE, Gardner IA, Jafarzadeh SR, Fossler CP, Harris B, Capsel RT, Wagner BA, Johnson WO (2013) Herd-level prevalence of Mycobacterium avium subsp. paratuberculosis infection in United States dairy herds in 2007. Prev Vet Med 108:234–238. https://doi.org/10.1016/j.prevetmed.2012.08.006

34. Pozzato N, Capello K, Comin A, Toft N, Nielsen SS, Vicenzoni G, Arrigoni N (2011) Prevalence of paratuberculosis infection in dairy cattle in Northern Italy. Prev Vet Med 102:83–86. https://doi.org/10.1016/j.prevetmed.2011.07.001

35. Sposato A, Fanelli A, Cordisco M, Trotta A, Galgano M, Corrente M, Buonavoglia D (2022) Bayesian estimation of prevalence of Johne's disease in dairy herds in Southern Italy. Prev Vet Med 199:105552. https://doi.org/10.1016/j.prevetmed.2021.105552

36. Yang DA, Heuer C, Laven R, Vink WD, Chesterton RN (2017) Estimating the true prevalence of bovine digital dermatitis in taranaki, New Zealand using a bayesian latent class model. Prev Vet Med 147:158–162. https://doi.org/10.1016/j.prevetmed.2017.09.008

37. Griswold ME, Swihart BJ, Caffo BS, Zeger SL (2013) Practical Marginalized Multilevel Models. Stat 2:10.1002/sta4.22. https://doi.org/10.1002/sta4.22

38. Collett D (2014) Modelling Survival Data in Medical Research, 3rd edition. Chapman and Hall/CRC, Boca Raton

39. Hosmer D, Lemeshow S (1999) Applied Survival Analysis: Regression Modelling of Time to Event Data. Wiley, New York

40. Klein JP, Houwelingen HC van, Ibrahim JG, Scheike TH (2013) Handbook of Survival Analysis. Chapman and Hall/CRC, New York

41. Kaplan EL, Meier P (1958) Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association 53:457–481. https://doi.org/10.1080/01621459.1958.10501452

42. Schober P, Vetter TR (2018) Survival Analysis and Interpretation of Time-to-Event Data: The Tortoise and the Hare. Anesth Analg 127:792–798. https://doi.org/10.1213/ANE.0000000000003653

43. Rai S, Mishra P, Ghoshal UC (2021) Survival analysis: A primer for the clinician scientists. Indian J Gastroenterol 40:541–549. https://doi.org/10.1007/s12664-021-01232-1

44. Selvin S (2004) Life Table Analysis: An Introduction. In: Selvin S (ed) Statistical Analysis of Epidemiologic Data. Oxford University Press

45. Gehan EA (1965) A Generalized Wilcoxon Test For Comparing Arbitrary Singly-Censored Samples. Biometrika 52:203–223

46. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG (1976) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br J Cancer 34:585–612

47. Ishak KJ, Kreif N, Benedict A, Muszbek N (2013) Overview of Parametric Survival Analysis for Health-Economic Applications. PharmacoEconomics 31:663–675. https://doi.org/10.1007/s40273-013-0064-3

48. Cox DR (1972) Regression Models and Life-Tables. Journal of the Royal Statistical Society: Series B (Methodological) 34:187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

49. Manning E, Collins M (2022) Epidemiology of paratuberculosis. pp 22–28

50. Vass-Bognár B, Bakony M, Baumgartner W, Khol JL, Jurkovich V (2022) Association between calf rearing technology and farm-level paratuberculosis infection in Hungarian dairy farms. Prev Vet Med 207:105719. https://doi.org/10.1016/j.prevetmed.2022.105719

51. Vass-Bognár B, Khol JL, Baumgartner W, Fornyos K, Papp M, Abonyi-Tóth Z, Bakony M, Jurkovich V (2024) Investigating the Prevalence of Paratuberculosis in Hungarian Large-Scale Dairy Herds and the Success of Control Measures over Four Years. Animals 14:151. https://doi.org/10.3390/ani14010151

52. Fodor I, Matyovszky B, Biczó A, Ózsvári L (2014) The losses due to paratuberculosis and its control in a Hungarian large-scale Holstein-Friesian dairy farm (A paratuberkulózis kártétele és az ellene való védekezés egy hazai nagyüzemi holstein-fríz tehenészetben). Magy Allatorvosok Lapja 136:213–222

53. Kudahl AB, Nielsen SS (2009) Effect of paratuberculosis on slaughter weight and slaughter value of dairy cows. J Dairy Sci 92:4340–4346. https://doi.org/10.3168/jds.2009-2039

54. Lombard JE, Garry FB, McCluskey BJ, Wagner BA (2005) Risk of removal and effects on milk production associated with paratuberculosis status in dairy cows. J Am Vet Med Assoc 227:1975–1981. https://doi.org/10.2460/javma.2005.227.1975

55. Ózsvári L, Harnos A, Lang Z, Monostori A, Strain S, Fodor I (2020) The Impact of Paratuberculosis on Milk Production, Fertility, and Culling in Large Commercial Hungarian Dairy Herds. Frontiers in Veterinary Science 7:778. https://doi.org/10.3389/fvets.2020.565324

56. Smith RL, Strawderman RL, Schukken YH, Wells SJ, Pradhan AK, Espejo LA, Whitlock RH, Van Kessel JS, Smith JM, Wolfgang DR, Gröhn YT (2010) Effect of Johne's disease status on reproduction and culling in dairy cattle. J Dairy Sci 93:3513–3524. https://doi.org/10.3168/jds.2009-2742

57. Jurkovich V, Bognár B, Balogh K, Kovács-Weber M, Fornyos K, Szabó RT, Kovács P, Könyves L, Mézes M (2016) Effects of subclinical Mycobacterium avium ssp. paratuberculosis infection on some physiological parameters, health status and production in dairy cows. Acta Veterinaria Hungarica 64:301–312. https://doi.org/10.1556/004.2016.029

58. Martins EG, Oliveira P, Oliveira BM, Mendonça D, Niza-Ribeiro J (2018) Association of paratuberculosis sero-status with milk production and somatic cell counts across 5 lactations, using multilevel mixed models, in dairy cows. J Dairy Sci 101:7638–7649. https://doi.org/10.3168/jds.2017-13746

59. McAloon CG, Whyte P, More SJ, Green MJ, O'Grady L, Garcia A, Doherty ML (2016) The effect of paratuberculosis on milk yield—A systematic review and meta-analysis. J Dairy Sci 99:1449–1460. https://doi.org/10.3168/jds.2015-10156

60. Pritchard TC, Coffey MP, Bond KS, Hutchings MR, Wall E (2017) Phenotypic effects of subclinical paratuberculosis (Johne's disease) in dairy cattle. J Dairy Sci 100:679–690. https://doi.org/10.3168/jds.2016-11323

61. Rasmussen P, Barkema HW, Mason S, Beaulieu E, Hall DC (2021) Economic losses due to Johne's disease (paratuberculosis) in dairy cattle. Journal of Dairy Science 104:3123–3143. https://doi.org/10.3168/jds.2020-19381

62. Abubakar I, Myhill D, Aliyu SH, Hunter PR (2008) Detection of Mycobacterium avium subspecies paratuberculosis from patients with Crohn's disease using nucleic acid-based techniques: A systematic review and meta-analysis. Inflammatory Bowel Diseases 14:401–410. https://doi.org/10.1002/ibd.20276

63. Feller M, Huwiler K, Stephan R, Altpeter E, Shang A, Furrer H, Pfyffer GE, Jemmi T, Baumgartner A, Egger M (2007) Mycobacterium avium subspecies paratuberculosis and Crohn's disease: a systematic review and meta-analysis. The Lancet Infectious Diseases 7:607–613. https://doi.org/10.1016/S1473-3099(07)70211-6

64. Garvey M (2018) Mycobacterium avium subspecies paratuberculosis: A possible causative agent in human morbidity and risk to public health safety. Open Vet J 8:172–181. https://doi.org/10.4314/ovj.v8i2.10

65. Lombard JE (2011) Epidemiology and Economics of Paratuberculosis. Veterinary Clinics of North America: Food Animal Practice 27:525–535. https://doi.org/10.1016/j.cvfa.2011.07.012

66. Nielsen SS, Bjerre H, Toft N (2008) Colostrum and milk as risk factors for infection with Mycobacterium avium subspecies paratuberculosis in dairy cattle. J Dairy Sci 91:4610–4615. https://doi.org/10.3168/jds.2008-1272

67. Whittington R, Donat K, Weber MF, Kelton D, Nielsen SS, Eisenberg S, Arrigoni N, Juste R, Sáez JL, Dhand N, Santi A, Michel A, Barkema H, Kralik P, Kostoulas P, Citer L, Griffin F, Barwell R, Moreira MAS, Slana I, Koehler H, Singh SV, Yoo HS, Chávez-Gris G, Goodridge A, Ocepek M, Garrido J, Stevenson K, Collins M, Alonso B, Cirone K, Paolicchi F, Gavey L, Rahman MT, de Marchin E, Van Praet W, Bauman C, Fecteau G, McKenna S, Salgado M, Fernández-Silva J, Dziedzinska R, Echeverría G, Seppänen J, Thibault V, Fridriksdottir V, Derakhshandeh A, Haghkhah M, Ruocco L, Kawaji S, Momotani E, Heuer C, Norton S, Cadmus S, Agdestein A, Kampen A, Szteyn J, Frössling J, Schwan E, Caldow G, Strain S, Carter M, Wells S, Munyeme M, Wolf R, Gurung R, Verdugo C, Fourichon C, Yamamoto T, Thapaliya S, Di Labio E, Ekgatat M, Gil A, Alesandre AN, Piaggio J, Suanes A, de Waard JH (2019) Control of paratuberculosis: who, why and how. A review of 48 countries. BMC Vet Res 15:198. https://doi.org/10.1186/s12917-019-1943-4

68. Windsor PA, Whittington RJ (2010) Evidence for age susceptibility of cattle to Johne's disease. The Veterinary Journal 184:37–44. https://doi.org/10.1016/j.tvjl.2009.01.007

69. Kirkpatrick BW, Shook GE (2011) Genetic Susceptibility to Paratuberculosis. Veterinary Clinics of North America: Food Animal Practice 27:559–571. https://doi.org/10.1016/j.cvfa.2011.07.003

70. Fecteau M-E (2018) Paratuberculosis in Cattle. Veterinary Clinics of North America: Food Animal Practice 34:209–222. https://doi.org/10.1016/j.cvfa.2017.10.011

71. Stabel JR (2000) Transitions in immune responses to *Mycobacterium paratuberculosis*. Veterinary Microbiology 77:465–473. https://doi.org/10.1016/S0378-1135(00)00331-X

72. World Organisation for Animal Health (2021) Paratuberculosis

73. Glover I, Bradley A, Green M, McAloon CG, Hyde R, O'Grady L (2025) Use of a hidden Markov model for interpretation of serial cow milk paratuberculosis antibody enzyme-linked immunosorbent assay results adjusted for milk yield and quality. Preventive Veterinary Medicine 235:106413. https://doi.org/10.1016/j.prevetmed.2024.106413

74. Sweeney RW (2011) Pathogenesis of Paratuberculosis. Veterinary Clinics of North America: Food Animal Practice 27:537–546. https://doi.org/10.1016/j.cvfa.2011.07.001

75. Gonda MG, Chang YM, Shook GE, Collins MT, Kirkpatrick BW (2007) Effect of *Mycobacterium paratuberculosis* infection on production, reproduction, and health traits in US Holsteins. Preventive Veterinary Medicine 80:103–119. https://doi.org/10.1016/j.prevetmed.2007.01.011

76. Jolly A, Fernández B, Stempler A, Ingratta G, Postma G, Boviez J, Lombardo D, Hajos S, Mundo SL (2022) Antibodies from healthy or paratuberculosis infected cows have different effects on Mycobacterium avium subspecies paratuberculosis invasion in a calf ileal loop model. Veterinary Immunology and Immunopathology 245:110381. https://doi.org/10.1016/j.vetimm.2022.110381

77. Vass-Bognár B, Jánosi Sz, Baumgartner W, Khol JL, Jurkovich V (2020) A paratuberculosis diagnosztizálásának lehetőségei szarvasmarhában. Magy Allatorvosok Lapja 142:10 pp. 579-592. , 14 p.

78. Nielsen SS, Enevoldsen C, Gröhn YT (2002) The Mycobacterium avium subsp. paratuberculosis ELISA response by parity and stage of lactation. Prev Vet Med 54:1–10. https://doi.org/10.1016/s0167-5877(02)00008-9

79. Sweeney RW, Whitlock RH, McAdams S, Fyock T (2006) Longitudinal Study of ELISA Seroreactivity to Mycobacterium Avium subsp. Paratuberculosis in Infected Cattle and Culture-Negative Herd Mates. J VET Diagn Invest 18:2–6. https://doi.org/10.1177/104063870601800102

80. Collins MT, Wells SJ, Petrini KR, Collins JE, Schultz RD, Whitlock RH (2005) Evaluation of Five Antibody Detection Tests for Diagnosis of Bovine Paratuberculosis. Clin Diagn Lab Immunol 12:685–692. https://doi.org/10.1128/CDLI.12.6.685-692.2005

81. Beaver A, Sweeney RW, Hovingh E, Wolfgang DR, Gröhn YT, Schukken YH (2017) Longitudinal relationship between fecal culture, fecal quantitative PCR, and milk ELISA in Mycobacterium avium ssp. paratuberculosis-infected cows from low-prevalence dairy herds. Journal of Dairy Science 100:7507–7521. https://doi.org/10.3168/jds.2017-12928

82. Cazer CL, Mitchell RM, Cicconi-Hogan KM, Gamroth M, Richert RM, Ruegg PL, Schukken YH (2013) Associations between Mycobacterium avium subsp. paratuberculosis antibodies in bulk tank milk, season of sampling and protocols for managing infected cows. BMC Vet Res 9:234. https://doi.org/10.1186/1746-6148-9-234

83. Sweeney RW (1996) Transmission of Paratuberculosis. Veterinary Clinics of North America: Food Animal Practice 12:305–312. https://doi.org/10.1016/S0749-0720(15)30408-4

84. Mortier RAR, Barkema HW, Negron ME, Orsel K, Wolf R, De Buck J (2014) Antibody response early after experimental infection with Mycobacterium avium subspecies paratuberculosis in dairy calves. Journal of Dairy Science 97:5558–5565. https://doi.org/10.3168/jds.2014-8139

85. Vass-Bognár B, Bakony M, Fornyos K, Baumgartner W, Khol JL, Jurkovich V (2025) When is the best time to test paratuberculosis positivity? Observations from a follow-up study in Hungarian dairy herds. Front Vet Sci 12:. https://doi.org/10.3389/fvets.2025.1570915

86. Weber MF, Kelton D, Eisenberg SWF, Donat K (2024) Progress in Paratuberculosis Control Programmes for Dairy Herds. Animals 14:1127. https://doi.org/10.3390/ani14071127

87. Tuberquia-López BC, Correa-Valencia NM del P, Hernández JM, Fernández-Silva JA, Ramírez-Vásquez NF (2022) Paratuberculosis control strategies in dairy cattle: A systematic review. Open Veterinary Journal 12:525–525. https://doi.org/10.5455/OVJ.2022.v12.i4.16

88. Rényi A (1989) Valószínűségszámítás, 5. kiadás. Tankönyvkiadó, Budapest

89. McElreath R (2020) Statistical rethinking: a Bayesian course with examples in R and Stan, Second edition. CRC Press, Boca Raton

90. Gill CJ, Sabin L, Schmid CH (2005) Why clinicians are natural bayesians. BMJ 330:1080–1083

91. Gory JJ, Craigmile PF, MacEachern SN (2021) A class of generalized linear mixed models adjusted for marginal interpretability. Statistics in Medicine 40:427–440. https://doi.org/10.1002/sim.8782

92. Guolo A, Varin C (2014) Beta regression for time series analysis of bounded data, with application to Canada Google® Flu Trends. The Annals of Applied Statistics 8:74–88. https://doi.org/10.1214/13-AOAS684

93. Xue-Kun Song P (2000) Multivariate Dispersion Models Generated From Gaussian Copula. Scandinavian Journal of Statistics 27:305–320. https://doi.org/10.1111/1467-9469.00191

94. Feng C, Hongyue W, Tu X, Kowalski J (2012) A Note on Generalized Inverses of Distribution Function and Quantile Transformation. Applied Mathematics 03:2098–2100. https://doi.org/10.4236/am.2012.312A289

95. Ferrari S, Cribari-Neto F (2004) Beta Regression for Modelling Rates and Proportions. Journal of Applied Statistics 31:799–815. https://doi.org/10.1080/0266476042000214501

96. Song P (2007) Correlated Data Analysis: Modeling, Analytics, and Applications

97. Yu Y (2008) On the maximal correlation coefficient. Statistics & Probability Letters 78:1072–1075. https://doi.org/10.1016/j.spl.2007.10.006

98. Veres K, Lang Z, Monostori A, Kostoulas P, Ózsvári L (2024) Bayesian latent class modelling of true prevalence in animal subgroups with application to bovine paratuberculosis infection. Prev Vet Med 224:106133. https://doi.org/10.1016/j.prevetmed.2024.106133

99. Christensen R, Johnson W, Branscum A, Hanson TE (2010) Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians. CRC Press, Boca Raton

100. R Core Team (2021) R: A Language and Environment for Statistical Computing. https://www.R-project.org/

101. Stevenson M, Nunes ES with contributions from T, Heuer C, Marshall J, Sanchez J, Thornton R, Reiczigel J, Robison-Cox J, Sebastiani P, Solymos P, Yoshida K, Jones G, Pirikahu S, Firestone S, Kyle R, Popp J, Jay M, Reynard C, Cheung A, Singanallur N, Szabo A, Rabiee A (2023) epiR: Tools for the Analysis of Epidemiological Data. https://CRAN.R-project.org/package=epiR

102. Stan Development Team (2023) RStan: the R interface to Stan. https://mc-stan.org/

103. Denwood MJ (2016) runjags: An R Package Providing Interface Utilities, Model Templates, Parallel Computing Methods and Additional Distributions for MCMC Models in JAGS. Journal of Statistical Software 71:1–25. https://doi.org/10.18637/jss.v071.i09

104. Gabry J, Mahr T. (2019) Plotting for Bayesian Models. https://mc-stan.org/bayesplot/. Accessed 6 Jan 2022

105. McAloon CG, Doherty ML, Whyte P, Verdugo C, Toft N, More SJ, O'Grady L, Green MJ (2019) Low accuracy of Bayesian latent class analysis for estimation of herd-level true prevalence under certain disease characteristics-An analysis using simulated data. Prev Vet Med 162:117–125. https://doi.org/10.1016/j.prevetmed.2018.11.014

106. Verdugo C, Toft N, Nielsen SS (2015) Within- and between-herd prevalence variation of Mycobacterium avium subsp. paratuberculosis infection among control programme herds in Denmark (2011–2013). Prev Vet Med 121:282–287. https://doi.org/10.1016/j.prevetmed.2015.07.012

107. Verdugo C, Valdes MF, Salgado M (2018) Within-herd prevalence and clinical incidence distributions of Mycobacterium avium subspecies paratuberculosis infection on dairy herds in Chile. Prev Vet Med 154:113–118. https://doi.org/10.1016/j.prevetmed.2018.03.022

108. Kass RE, Raftery AE (1995) Bayes Factors. Journal of the American Statistical Association 773–795. https://doi.org/10.1080/01621459.1995.10476572

109. Gronau QF, Singmann H, Wagenmakers E-J (2020) bridgesampling: An R Package for Estimating Normalizing Constants. Journal of Statistical Software 92:1–29. https://doi.org/10.18637/jss.v092.i10

110. Balan TA, Putter H (2020) A tutorial on frailty models. Stat Methods Med Res 29:3424–3454. https://doi.org/10.1177/0962280220921889

111. Therneau TM, Grambsch PM (2000) Modeling Survival Data: Extending the Cox Model. Springer, New York, NY

112. Mallikarjunappa S, Brito LF, Pant SD, Schenkel FS, Meade KG, Karrow NA (2021) Johne's Disease in Dairy Cattle: An Immunogenetic Perspective. Front Vet Sci 8:. https://doi.org/10.3389/fvets.2021.718987

113. Pavlik I, Bartl J, Dvorska L, Svastova P, du Maine R, Machackova M, Yayo Ayele W, Horvathova A (2000) Epidemiology of paratuberculosis in wild ruminants studied by restriction fragment length polymorphism in the Czech Republic during the period 1995–1998. Vet Microbiol 77:231–251. https://doi.org/10.1016/S0378-1135(00)00309-6

114. Garcia AB, Shalloo L (2015) Invited review: The economic impact and control of paratuberculosis in cattle. Journal of Dairy Science 98:5019–5039. https://doi.org/10.3168/jds.2014-9241

115. Ritter C, Wolf R, Adams CL, Kelton DF, Pickel C, Mason S, Orsel K, De Buck J, Barkema HW (2016) Short communication: Herd-level prevalence of Mycobacterium avium ssp. paratuberculosis is not associated with participation in

a voluntary Alberta Johne's disease control program. Journal of Dairy Science 99:2157–2160. https://doi.org/10.3168/jds.2015-10356

116. Lombard JE, Byrem TM, Wagner BA, McCluskey BJ (2006) Comparison of Milk and Serum Enzyme–Linked Immunosorbent Assays for Diagnosis of Mycobacterium Avium Subspecies Paratuberculosis Infection in Dairy Cattle. J VET Diagn Invest 18:448–458. https://doi.org/10.1177/104063870601800504

117. Nielsen SS, Agger JF (2003) Comparison of Milk and Serum Enzyme-Linked Immunosorbent Assays Adjusted for Time-Related Factors for the Detection of Paratuberculosis in Dairy Cattle. Acta Veterinaria Scandinavica 44:P64. https://doi.org/10.1186/1751-0147-44-S1-P64

118. McAloon CG, O'Grady L, Botaro B, More SJ, Doherty M, Whyte P, Saxmose Nielsen S, Citer L, Kenny K, Graham D, Green M (2020) Individual and herd-level milk ELISA test status for Johne's disease in Ireland after correcting for non-disease-associated variables. J Dairy Sci 103:9345–9354. https://doi.org/10.3168/jds.2019-18018

119. Agresti A (2013) Categorical Data Analysis, 3rd Edition. John Wiley & Sons

120. Fitzmaurice GM, Laird NM, Ware JH (2011) Applied Longitudinal Analysis, Second Edition. Wiley, Hoboken, New Jersey

121. Schmidt SAJ, Veres K, Sørensen HT, Obel N, Henderson VW (2022) Incident Herpes Zoster and Risk of Dementia. Neurology 99:e660–e668. https://doi.org/10.1212/WNL.0000000000200709

122. Bae S, Yun S-C, Kim M-C, Yoon W, Lim JS, Lee S-O, Choi S-H, Kim YS, Woo JH, Kim SY, Kim S-H (2021) Association of herpes zoster with dementia and effect of antiviral therapy on dementia: a population-based cohort study. Eur Arch Psychiatry Clin Neurosci 271:987–997. https://doi.org/10.1007/s00406-020-01157-4

123. Chen VC-H, Wu S-I, Huang K-Y, Yang Y-H, Kuo T-Y, Liang H-Y, Huang K-L, Gossop M (2018) Herpes Zoster and Dementia: A Nationwide Population-Based Cohort Study. J Clin Psychiatry 79:16m11312. https://doi.org/10.4088/JCP.16m11312

124. Choi HG, Park BJ, Lim JS, Sim SY, Jung YJ, Lee SW (2021) Herpes Zoster Does Not Increase the Risk of Neurodegenerative Dementia: A Case-Control Study. Am J Alzheimers Dis Other Demen 36:15333175211006504. https://doi.org/10.1177/15333175211006504

125. Tsai M-C, Cheng W-L, Sheu J-J, Huang C-C, Shia B-C, Kao L-T, Lin H-C (2017) Increased risk of dementia following herpes zoster ophthalmicus. PLoS One 12:e0188490. https://doi.org/10.1371/journal.pone.0188490

126. Sora VM, Panseri S, Nobile M, Di Cesare F, Meroni G, Chiesa LM, Zecconi A (2022) Milk Quality and Safety in a One Health Perspective: Results of a Prevalence Study on Dairy Herds in Lombardy (Italy). Life (Basel) 12:786. https://doi.org/10.3390/life12060786

127. Vaona F (2019) Comparison of potential effects on the profitability of the US MPP application on dairy farms in Veneto (Italy) and Wielkopolska (Poland). University Degli di Padova

128. Gonzalez S (2024) Dairy and Products Annual. USDA

129. Gebelein H (1941) Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung. ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik 21:364–379. https://doi.org/10.1002/zamm.19410210604

130. Lancaster HO (1957) Some properties of the bivariate normal distribution considered in the form of a contingency table. Biometrika 44:289–292. https://doi.org/10.1093/biomet/44.1-2.289

# 10 Publications of the author

## 10.1 Related scientific publications in peer-reviewed scientific journals with impact factor

**Veres K,** Lang Zs, Ózsvári L (2025) Bayesian Assessment of True Prevalence of Paratuberculosis Infection in Dairy Herds and Their Parity Subgroups. Pathogens 14:900.
https://doi.org/10.3390/pathogens14090900

**Veres K**, Lang Zs, Monostori A, Kostoulas P, Ózsvári L (2024) Bayesian latent class modelling of true prevalence in animal subgroups with application to bovine paratuberculosis infection. Prev Vet Med 224:106133. https://doi.org/10.1016/j.prevetmed.2024.106133

Schmidt SAJ, **Veres K,** Sørensen HT, Obel N, Henderson VW (2022) Incident Herpes Zoster and Risk of Dementia. Neurology 99:e660–e668. https://doi.org/10.1212/WNL.0000000000200709

**Veres K**, Lang, Zs, Ózsvári, L. (2025). Application of survival analysis in veterinary practice – The relationship between PTBC infection and culling in dairy cattle herds (Túléléselemzés alkalmazása az állatorvosi gyakorlatban – A paratuberkulózis fertőzöttség és a selejtezés kapcsolata tejelő szarvasmarha állományokban). Magy Allatorvosok Lapja 147:573-595.
https://doi.org/10.56385/magyallorv.2025.10.579-595

**Veres K,** Lang Zs, Monostori A, Ózsvári L (2024) Bayesian modelling in practice. Estimation of within-herd paratuberculosis prevalence in dairy cattle herds (Bayes-i modellezés a gyakorlatban – tejelő tehénállományok állományon belüli paratuberkulózis érintettségének becslése). Magy Allatorvosok Lapja 146:323–337.
https://doi.org/10.56385/magyallorv.2024.06.323-337

## 10.2 Related conference presentations

Bárdos K, Máté M, Lang Zs, **Veres K**, Czopowicz M, Kaba J, Ózsvári L (2023): Main keeping conditions and CAE infection of Hungarian goat herds. 31st International Congress of the Hungarian Association of Buiatrics, 11-14. October 2023, Keszthely, Hungary, 152-160. pp. ISBN: 978-615-81413-3-8

Lang Zs, Ózsvári L, **Veres K** (2023): Comparison of conditional and marginal means in distribution based marginalized multilevel models. 44th Annual Conference of the International Society for Clinical Biostatistics, 27-31 August 2023, Milan, Italy, Abstract book, 151.

## 10.3 Scientific publications not related to the topic of the doctoral research

Kurt G, Troelsen FS, **Veres K,** Sørensen HT, Erichsen R (2025) Risk of venous thromboembolism following colorectal cancer surgery in patients with and without inflammatory bowel disease. Am J Gastroenterol. https://doi.org/10.14309/ajg.0000000000003761

Sundbøll J, **Veres K,** Troelsen FS, Würtz M, Bøtker HE, Sørensen HT (2025) Myocardial infarction and short- or long-term risk of a subsequent cancer diagnosis: a Danish Nationwide Cohort Study. Journal of Thrombosis and Haemostasis 23:1023–1033. https://doi.org/10.1016/j.jtha.2024.12.008

Sun Y, **Veres K**, Hasselbalch HC, Frederiksen H, Østgård LSG, Horváth-Puhó E, Henderson VW, Sørensen HT (2024) Myeloproliferative Neoplasms and Dementia Risk: A Population-Based Cohort Study. Eur J Haematol. https://doi.org/10.1111/ejh.14297

Gadgaard NR, **Veres K,** Henderson VW, Pedersen AB (2024) Frozen Shoulder and the Risk of Parkinson's Disease: A Danish Registry-Based Cohort Study. Clin Epidemiol 16:447–459. https://doi.org/10.2147/CLEP.S463571

Korsgaard S, Troelsen FS, **Veres K**, Fuglsang CH, Sørensen HT (2024) Evaluation of Left Truncation and Censoring When Changing the Use of the International Classification of Diseases Eighth Revision Codes to Tenth Revision Codes in the Danish National Patient Registry. Clin Epidemiol 16:319–327. https://doi.org/10.2147/CLEP.S456171

Adelborg K, **Veres K,** Horváth-Puhó E, Clouser M, Saad H, Sørensen HT (2024) Risk and adverse clinical outcomes of thrombocytopenia among patients with solid tumors—a Danish population-based cohort study. Br J Cancer 1–8. https://doi.org/10.1038/s41416-024-02630-w

Laugesen K, **Veres K,** Hernandez-Diaz S, Chiu Y-H, Oberg AS, Hsu J, Rinaudo P, Spaan M, van Leeuwen F, Sørensen HT (2023) Overweight or obesity in children born after assisted reproductive technologies in Denmark: A population-based cohort study. PLoS Med 20:e1004324. https://doi.org/10.1371/journal.pmed.1004324

Collin LJ, **Veres K,** Gradus JL, Ahern TP, Lash TL, Sørensen HT (2022) Pre-existing stress-related diagnoses and mortality: a Danish cancer cohort study. Cancer 128:1312–1320. https://doi.org/10.1002/cncr.34036

Nørgaard M, **Veres K,** Sellebjerg FT, Svingel LS, Foch C, Boutmy E, Sabidó M, Magyari M (2021) Incidence of malignancy in multiple sclerosis: A cohort study in the Danish Multiple Sclerosis Registry. Mult Scler J Exp Transl Clin 7:20552173211053939. https://doi.org/10.1177/20552173211053939

Arnesen CAL, **Veres K,** Horváth-Puhó E, Hansen J-B, Sørensen HT, Brækkan SK (2021) Estimated lifetime risk of venous thromboembolism in men and women in a Danish nationwide cohort: impact of competing risk of death. Eur J Epidemiol. https://doi.org/10.1007/s10654-021-00813-w

Egenvall M, Martling A, **Veres K,** Horváth-Puhó E, Wille-Jørgensen P, Høirup Petersen S, Laurberg S, Sørensen HT, Syk I, Andersen PV, Christensen H, Nielsen JD, Jess P, Madsen MR, Pedersen AG, Østergaard E, Andersson PH, Bengtsson J, Bragmark M, Buchwald P, Egenvall M, Farahnak P, Folkesson J, Påhlman L, Goldinger M,

Heuman R, Lindberg K, Martling A, Näsvall P, Ottosson J, Sandzén B, Barberousse C, Horváth-Puhó E, Laurberg S, Påhlman L, Renehan A, Smedh K, Syk I, Sørensen HT, Wille-Jørgensen P (2021) No benefit of more intense follow-up after surgery for colorectal cancer in the risk group with elevated CEA levels – An analysis within the COLOFOL randomized clinical trial. European Journal of Surgical Oncology 47:2053–2059. https://doi.org/10.1016/j.ejso.2021.03.235

Mark-Christensen A, Erichsen R, **Veres K**, Laurberg S, Sørensen HT (2021) Rectal Cancer Risk and Survival After Total Colectomy for IBD: A Population-Based Study. Diseases of the Colon & Rectum 64:583–591. https://doi.org/10.1097/DCR.0000000000001810

Søgaard KK, **Veres K**, Vandenbroucke-Grauls CMJE, Vandenbroucke JP, Sørensen HT, Schønheyder HC (2020) Community-Acquired Escherichia coli Bacteremia after Age 50 and Subsequent Incidence of a Cancer Diagnosis: A Danish Population-Based Cohort Study. Cancer Epidemiol Biomarkers Prev 29:2626–2632. https://doi.org/10.1158/1055-9965.EPI-20-0705

Ording AG, Horváth-Puhó E, **Veres K,** Glymour MM, Rørth M, Sørensen HT, Henderson VW (2020) Cancer and risk of Alzheimer’s disease: Small association in a nationwide cohort study. Alzheimers Dement. https://doi.org/10.1002/alz.12090

Gribsholt SB, Cronin-Fenton D, **Veres K,** Thomsen RW, Ording AG, Richelsen B, Sørensen HT (2020) Hospital-diagnosed overweight and obesity related to cancer risk: a 40-year Danish cohort study. J Intern Med 287:435–447. https://doi.org/10.1111/joim.13013

Bengtsen MB, **Veres K,** Nørgaard M (2020) First-time postmenopausal bleeding as a clinical marker of long-term cancer risk: A Danish Nationwide Cohort Study. Br J Cancer 122:445–451. https://doi.org/10.1038/s41416-019-0668-2

Adelborg K, **Veres K,** Sundbøll J, Gregersen H, Sørensen HT (2019) Risk of cancer in patients with fecal incontinence. Cancer Med 8:6449–6457. https://doi.org/10.1002/cam4.2509

Nørgaard M, **Veres K,** Didden EM, Wormser D, Magyari M (2019) Multiple sclerosis and cancer incidence: A Danish nationwide cohort study. Mult Scler Relat Disord 28:81–85. https://doi.org/10.1016/j.msard.2018.12.014

Sundbøll J, Thygesen SK, **Veres K,** Liao D, Zhao J, Gregersen H, Sørensen HT (2019) Risk of cancer in patients with constipation. Clin Epidemiol 11:299–310. https://doi.org/10.2147/CLEP.S205957

Sundbøll J, Larsen AP, **Veres K,** Adelborg K, Sørensen HT (2019) Cardiovascular event rates and trajectories of LDL-cholesterol levels and lipid-lowering therapy in patients with atherosclerotic cardiovascular disease: A population-based cohort study. Thromb Res 183:124–130. https://doi.org/10.1016/j.thromres.2019.09.034

Søgaard KK, **Veres K,** Nørgaard M, Djurhuus JC, Sørensen HT (2019) Pyelonephritis in persons after age 50 as a clinical marker of urogenital cancer. Clin Microbiol Infect 25:87–91. https://doi.org/10.1016/j.cmi.2018.04.003

Ording AG, **Veres K,** Horváth-Puhó E, Glymour MM, Rørth M, Henderson VW, Sørensen HT (2019) Alzheimer’s and Parkinson’s Diseases and the Risk of Cancer: A Cohort Study. J Alzheimers Dis 72:1269–1277. https://doi.org/10.3233/JAD-190867

Mark-Christensen A, Erichsen R, **Veres K,** Laurberg S, Sørensen HT (2019) Extracolonic cancer risk after total colectomy for inflammatory bowel disease: A population-based cohort study. J Crohns Colitis. https://doi.org/10.1093/ecco-jcc/jjz199

Ahern TP, **Veres K,** Jiang T, Farkas DK, Lash TL, Sørensen HT, Gradus JL (2018) Adjustment disorder and type-specific cancer incidence: a Danish cohort study. Acta Oncol 57:1367–1372. https://doi.org/10.1080/0284186X.2018.1465586

Nørgaard M, **Veres K,** Ording AG, Djurhuus JC, Jensen JB, Sørensen HT (2018) Evaluation of Hospital-Based Hematuria Diagnosis and Subsequent Cancer Risk Among Adults in Denmark. JAMA Netw Open 1:e184909. https://doi.org/10.1001/jamanetworkopen.2018.4909

Hansen AT, **Veres K,** Prandoni P, Adelborg K, Sørensen HT (2018) Retinal vein thrombosis and risk of occult cancer: A nationwide cohort study. Cancer Med 7:5789–5795. https://doi.org/10.1002/cam4.1803

Sundbøll J, **Veres K,** Horváth-Puhó E, Adelborg K, Sørensen HT (2018) Risk and Prognosis of Cancer After Lower Limb Arterial Thrombosis. Circulation 138:669–677. https://doi.org/10.1161/CIRCULATIONAHA.117.032617

Ording AG, **Veres K,** Farkas DK, Adelborg K, Sørensen HT (2018) Risk of cancer in patients with epistaxis and haemoptysis. Br J Cancer 118:913–919. https://doi.org/10.1038/bjc.2017.494

Jiang T, **Veres K,** Körmendiné Farkas D, Lash TL, Toft Sørensen H, Gradus JL (2018) Post-traumatic stress disorder and incident fractures in the Danish population. Osteoporos Int 29:2487–2493. https://doi.org/10.1007/s00198-018-4644-0

Christensen DH, **Veres K,** Ording AG, Jørgensen JOL, Cannegieter SC, Thomsen RW, Sørensen HT (2018) Risk of cancer in patients with thyroid disease and venous thromboembolism. Clin Epidemiol 10:907–915. https://doi.org/10.2147/CLEP.S158869

Hansen AT, **Veres K,** Horváth-Puhó E, Ehrenstein V, Prandoni P, Sørensen HT (2017) Pregnancy-related venous thromboembolism and risk of occult cancer. Blood Adv 1:2059–2062. https://doi.org/10.1182/bloodadvances.2017010231

Schmidt SAJ, **Veres K,** Ording AG, Farkas DK, Fogh K, Sørensen HT (2016) Venous ulcers and risk of occult hematological or other cancers: a nationwide cohort study. Blood 128:874–877. https://doi.org/10.1182/blood-2016-03-705038

# 11 Appendix

## 11.1 International model

### 11.1.1 Tables

**Table A1.** Mean size of dairy cattle herds reported in the literature and the size of herds used in the region-specific synthetic data.

| Region | Mean Herd Size [1] | Herd Size (Synthetic Data) | Source |
|---|---|---|---|
| Denmark | 195, 205, 208 [2] | 200 | [106] |
| Southern Italy | 315 | 300 | [35] |
| Lombardy | 191 | 200 | [126] |
| Veneto | 100 | 100 | [127] |
| Chile | 83 | 100 | [128] |

[1] Large variability. [2] Data from three consecutive years.

**Table A2.** Priors used along with Hungarian data.

| Variable | Prior | Median (95th Percentile) |
|---|---|---|
| **HTP** | beta(150.589, 12.250) | 0.927 (0.955) |
| $\boldsymbol{\mu_1}$ | beta(65.700, 715.900) | 0.084 (0.101) |
| $\boldsymbol{\mu_2}$ | beta(134.195, 716.181) | 0.158 (0.179) |
| $\boldsymbol{\sigma^2}$ | inverse gamma(37.030, 4.620) | 0.126 (0.167) |
| $\boldsymbol{\sigma_1{}^2}$ | inverse gamma(5.330, 0.070) | 0.014 (0.032) |
| $\boldsymbol{\sigma_2{}^2}$ | inverse gamma(6.050, 0.090) | 0.016 (0.034) |

HTP—herd true prevalence; $\mu_1$—mean conditional within-herd prevalence for primiparous cows in the region; $\mu_2$—mean conditional within-herd prevalence for multiparous cows in the region; $\sigma^2$—variance of the herd random effect; $\sigma_1{}^2$—variance of the parity random effect in primiparous cows; $\sigma_2{}^2$—variance of the parity random effect in multiparous cows.

**Table A3.** Priors used in the study.

| Region | Parameter | Priors and Parameters Used in the Analysis | Median (95th Percentile) |
|---|---|---|---|
| Denmark | HTP | beta(425.110, 144.641) | 0.746 (0.756) |
| | $\mu_1$ | beta(1.770, 43.210) | 0.033 (0.095) |
| | $\mu_2$ | beta(3.850, 45.090) | 0.073 (0.150) |
| Southern Italy | HTP | beta(5.030, 7.040) | 0.412 (0.650) |
| | $\mu_1$ | beta(2.766, 46.335) | 0.050 (0.118) |
| | $\mu_2$ | beta(6.018, 47.403) | 0.108 (0.019) |
| Northern Italy (Lombardy, Veneto) | HTP | beta(13.320, 6.280) | 0.686 (0.838) |
| | $\mu_1$ | beta(1.079, 18.347) | 0.041 (0.157) |
| | $\mu_2$ | beta(2.347, 18.788) | 0.099 (0.239) |
| Chile | HTP | beta(14.200, 0.700) | 0.971 (0.999) |
| | $\mu_1$ | beta(12.896, 172.151) | 0.068 (0.103) |
| | $\mu_2$ | beta(28.061, 173.629) | 0.138 (0.181) |

[1] HTP—herd true prevalence; $\mu_1$—mean conditional within-herd prevalence for primiparous cows in the region; $\mu_2$—mean conditional within-herd prevalence for multiparous cows in the region.

**Table A4.** Information from the literature used to inform prior prevalence distributions of the model [8,10,12,23,].

| Region (Source of Prior Information) | Variable | Distribution from the Literature | Information |
|---|---|---|---|
| Denmark | HTP | beta (425.11, 144.64) | 0.747 (0.781) [a] |
| (posterior) | CWHP | beta (3.584, 45.398) | 0.055 (0.16) [b] |
| Southern Italy | HTP | beta (5.03, 7.04) | 0.4 (0.2) [b] |
| (prior) | CWHP | beta ($\mu\psi$,$(1-\mu)\psi$) | – |
| | CWHP($\mu$) | beta (3.14, 31.7) | 0.09 (0.18) [c] |
| | CWHP($\psi$) | gamma (11.16, 11.31) | – |
| Northern Italy (Lombardy, Veneto) | HTP | beta (13.32, 6.28) | 0.70 (>0.50) [b] |
| (prior) | CWHP | beta ($\mu\psi$,$(1-\mu)\psi$) | – |
| | CWHP($\mu$) | beta (1.53, 15.69) | 0.035 (<0.22) [d] |
| | CWHP($\psi$) | gamma (8.81, 1.42) | 0.2 (<0.30) [e] |
| Chile | HTP | beta (14.2, 0.7) | 0.97 (0.99) [e] |
| (prior) | CWHP | beta ($\mu\psi$,$(1-\mu)\psi$) | – |
| | CWHP($\mu$) | beta (22.2, 176.9) | 0.11 (0.15) [e] |
| | CWHP($\psi$) | gamma (9.1, 4.6) | 0.25 (0.30) [e] |

HTP - herd true prevalence; CWHP - conditional within-herd prevalence; Se - sensitivity of ELISA test; Sp - specificity of ELISA test; [a] - median and 95th percentile; [b] - mode and 5th percentile; [c] - estimates based on expert opinion: the mean prevalence is 0.09 with 95% certainty that it is not more than 0.18, where the expert is also confident that 90% of all herds have a prevalence less or equal to 0.35 with 95% certainty that it does not exceed 0.45; [d] - mode and 95th percentile; [e] - median and 95th percentile of the distribution of 90th percentile of the within-herd TP.

### 11.1.2 Derivation of priors for conditional within-herd prevalence of subgroups used in the international model

Historical information about the distribution of herd true prevalence (HTP) and within-herd true prevalence (CWHP) of PTBC infection were collected from the literature to obtain informative priors. The priors from the studies of different regions had the following form:

$$HTP \sim beta(\alpha_{HTP}, \beta_{HTP}),$$

$$mCWHP \sim beta(\mu\psi, (1-\mu)\psi),$$

$$\mu \sim beta(\alpha_{\mu}, \beta_{\mu}), E(\mu) = \frac{\alpha_{\mu}}{\alpha_{\mu} + \beta_{\mu}},$$

$$\psi \sim gamma(\alpha_{\psi}, \beta_{\psi}), E(\psi) = \alpha_{\psi} \times \beta_{\psi},$$

where mCWHP is the mean conditional within-herd prevalence, $\mu$ is the expected value, and $\psi$ is the precision parameter of the distribution of mCWHP.

The model requires the following priors:

- Herd true prevalence: $HTP \sim beta(\alpha_{HTP}, \beta_{HTP})$.
- Mean conditional within-herd prevalence for primiparous cows: $mCWHP_1 \sim beta(\mu_1\psi_1, (1-\mu_1)\psi_1)$.
- Mean conditional within-herd prevalence for multiparous cows: $mCWHP_2 \sim beta(\mu_2\psi_2, (1-\mu_2)\psi_2)$.
- Variance of the herd random effect: $\sigma^2 \sim inv.gamma(\alpha_{\sigma}, \beta_{\sigma})$.
- Variance of the additive parity effect for primiparous cows: ${\sigma_1}^2 \sim inv.gamma(\alpha_{\sigma_1}, \beta_{\sigma_1})$.
- Variance of the additive parity effect for multiparous cows: ${\sigma_2}^2 \sim inv.gamma(\alpha_{\sigma_2}, \beta_{\sigma_2})$.

The $\mu_1$ and $\mu_2$ parameters of $mCWHP_1$ and $mCWHP_2$ priors were derived from the $mCWHP$ prior using additional assumptions. Based on Hungarian data, the proportion of primiparous cows (*p*) was assumed to be 40% and the ratio of $\mu_1$ and $\mu_2$ (*R*) was assumed to be 2.

Assumptions:

- Population proportion of primiparous cows: *p*.
- Ratio of $\mu_1$ and $\mu_2$: *R*.

$$E(\mu) = p\mu_1 + (1-p)\mu_2, \quad \text{(A1)}$$

$$\frac{\mu_2}{\mu_1} = R. \quad \text{(A2)}$$

From (A1) and (A2), we have $\mu_1 = \frac{E(\mu)}{p(1-R)+R}$ and $\mu_2 = \frac{R \times E(\mu)}{p(1-R)+R}$.

### 11.1.3 Generation of synthetic data

The following priors were required for simulating the synthetic data:

- Conditional within-herd prevalence for primiparous cows: $CWHP_1 \sim beta(\mu_{CWHP1}\psi_{CWHP1}, (1-\mu_{CWHP1})\psi_{CWHP1})$
- Conditional within-herd prevalence for multiparous cows: $CWHP_2 \sim beta(\mu_{CWHP2}\psi_{CWHP2}, (1-\mu_{CWHP2})\psi_{CWHP2})$

$\mu_{CWHP1}$ and $\mu_{CWHP2}$ were chosen to be the mean $\mu_1$and $\mu_2$ of $mCWHP_1$ and $mCWHP_2$ priors, respectively. To be coherent with the model, $\psi_{CWHP1}$ and $\psi_{CWHP2}$ were assigned the $\psi$ parameters from the Hungarian study.

To model the joint infection probabilities of primiparous and multiparous cows within a herd, I used a Gaussian copula [93] with a correlation coefficient of $\rho = 0.89$ [98]. The construction is the following: Let $Z_1 \sim N(0,1), Z_2 \sim N(0,1)$ with $Corr(Z_1, Z_2) = 0.89$. Let $U_1 = \Phi(Z_1)$ and $U_2 = \Phi(Z_2)$, where $\Phi$ is the cumulative standard normal distribution function. The joint distribution of the infection probabilities $CWHP_1$ and $CWHP_2$ of the parity groups is derived by transforming the uniformly distributed $U_1$ and $U_2$ using the inverse cumulative distribution functions of the respective prior distributions. Let $X_1 = F^{-1}_{CWHP_1}(U_1)$ and $X_2 = F^{-1}_{CWHP_2}(U_2)$, where $F^{-1}_{CWHP}$ is the inverse of the beta cumulative distribution function of $CWHP$. Thus, $(X_1, X_2)$ are random variables with the desired marginal distributions (that of $CWHP_1$ and $CWHP_2$) and a joint dependence structure induced via the Gaussian copula with $\rho = 0.89$.

## 11.2 Maximal correlation

***Definition:***

The maximal coefficient of correlation $R_M$ is defined as:

$R_M(X,Y) := \sup Corr\big(g(X), h(Y)\big),$ (A3)

where

$0 < var\big(g(X)\big) < \infty$ and $0 < var\big(h(Y)\big) < \infty$ . (A4)

***Theorem 1*** [129, 130]:

If (X, Y) is bivariate normal then R(X, Y) = |corr(X, Y)|.

***Consequence:*** $(b_i + b_{i1}, b_i + b_{i2})$ is bivariate normal, which implies that

$R_M(b_i + b_{i1}, b_i + b_{i2}) = |corr(b_i + b_{i1}, b_i + b_{i2})| = \rho.$

***Claim:***

$R_M(CWHP_{i1}, CWHP_{i2}) = q.$

## 11.3 Distribution based marginalized multilevel models

### 11.3.1 Likelihood inference

The full likelihood function related to (4.1-4.4) is

$L = \prod_{i=1}^{N} L_i,$ (A5)

in which the factor of the $i$th subject (or cluster) equals

$L_i = p_Y\big(\boldsymbol{Y_i} | \mu_{\boldsymbol{i}}^{c}(\boldsymbol{\theta_i}, \boldsymbol{a_i}, \boldsymbol{D}), \boldsymbol{a_i}, \boldsymbol{v_i}\big) \times p(\boldsymbol{a_i}|\boldsymbol{D}),$ (A6)

where $p_Y$ and $p$ denote the probability mass function or the density function according to the data types of their arguments, $\boldsymbol{Y_i} = \big(Y_{i1}, \dots, Y_{in_i}\big)^T$, $\mu_{\boldsymbol{i}}^{\boldsymbol{c}} = \big(\mu_{i1}^{c}, \dots, \mu_{in_i}^{c}\big)^T$, which is calculated by (1.7), $\boldsymbol{\theta_i} = \big(\theta_{i1}, \dots, \theta_{in_i}\big)^T$, and $\boldsymbol{v_i} = \big(v_{i1}, \dots, v_{in_i}\big)^T$, $i = 1, \dots, N$. We emphasize that $\boldsymbol{\theta_i}$ contains the reparameterized vectors $\theta_{ij}$ in (A6), so both the marginal means $\mu_{ij}$ and the variances of the random effects $\boldsymbol{Z_{ij}^T D Z_{ij}}$ are among the components of $\theta_{ij}$, $i = 1, \dots, N, j = 1, \dots, n_i$.

One possible frequentist method to estimate the model parameters is to maximize the marginal likelihood

$M = \prod_{i=1}^{N} M_i,$ (A7)

where

$$M_i = \int L_i \, d\boldsymbol{a_i}, i = 1, \dots, N. \tag{A8}$$

A possible way to integrate out the random effects $a_i$ in (A8) is using adaptive Gauss-Hermite quadrature.

Bayesian methods are also available for fitting the previous model. In this approach generally the full likelihood function (A5) is used. Additionally, prior distributions are needed for the model parameters $\boldsymbol{\theta_i}$, $\boldsymbol{v_i}$ and $\boldsymbol{D}$, $i = 1, \dots, N$.

### 11.3.2 Comparison of conditional and marginal means

I investigated the magnitudes of conditional means relative to marginal means in DMMMs to explore the way they are connected. I provided methods and conditions for determining the increase or attenuation of conditional means relative to marginal means in DMMMs. I introduced a transformation that separates the marginal and conditional effects additively on the same scale, then I formed sufficient conditions of increase or attenuation of transformed conditional means relative to transformed marginal means. A factor of proportionality reflecting changes of transformed conditional and marginal means is also introduced.

#### 11.3.2.1 General case

Let

$$u = \Phi\left(\boldsymbol{Z^T a} \times \left(\boldsymbol{Z^T D Z}\right)^{-\frac{1}{2}}\right) \tag{A9}$$

be the transformed random effect, which is uniformly distributed in $[0, 1]$. Throughout this section let $\mu_0$ and $\mu$ be two marginal means close to each other. Let $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta}$ be the corresponding parameters of the distribution $H$ in (4.2), respectively, where their first components are $\theta_{01} = \mu_0$ and $\theta_1 = \mu$, and all the other components of $\theta$ coincides with that of $\theta_0$. Assuming the same random effect $u$, we intend to compare the conditional means $\mu^c = H^{-1}(\boldsymbol{\theta}, u)$ and $\mu_0^c = H^{-1}(\boldsymbol{\theta_0}, u)$.

Generally, the raw difference $\mu^c - \mu_0^c$ depends on $u$, therefore we need to construct a transformation $T(\boldsymbol{\theta}, x)$ in such a way that $T(\boldsymbol{\theta_0}, \mu^c) - T(\boldsymbol{\theta_0}, \mu_0^c)$ approximately does not depend on the common random effect $u$, if the marginal mean $\mu$ is sufficiently close to $\mu_0$.

Once we have a suitable transformation, we will be able to characterize the behavior of the transformed conditional means. According to the theoretical result (**Theorem 2**), this transformation $T$ exists and it does not depend on $u$ if and only if it has the form

$$T(\boldsymbol{\theta_0}, x) = l(\boldsymbol{\theta_0}) + m(\boldsymbol{\theta_0}) \times T_0(\boldsymbol{\theta_0}, x), \tag{A10}$$

where $l$ and $m$ are arbitrary real functions, and

$$T_0(\boldsymbol{\theta_0}, x) = -\int_{\mu_0}^{x} \frac{\partial_2 H(\boldsymbol{\theta_0}, t)}{\partial_1 H(\boldsymbol{\theta_0}, t)} dt. \tag{A11}$$

Using first order Taylor approximation, the following holds locally, in $\mu_0$:

$$T_0(\boldsymbol{\theta_0}, \mu^c) - T_0(\boldsymbol{\theta_0}, \mu_0^c) \approx \mu - \mu_0. \tag{A12}$$

This can be interpreted as the transformation $T_0$ pulls the deviation between conditional means towards the difference between marginal means.

Rearranging of the above difference, we get

$$T_0(\boldsymbol{\theta_0}, \mu^c) \approx \mu + \left(T_0\left(\boldsymbol{\theta_0}, H^{-1}(\boldsymbol{\theta_0}, u)\right) - \mu_0\right), \tag{A13}$$

which is an approximate decomposition of the transformed conditional mean $\mu^c$ into the marginal mean $\mu$ plus the transformed conditional mean depending only on $u$, and the reference parameter vector $\boldsymbol{\theta_0} = (\mu_0, \dots)^T$.

For the transformed marginal means analogous approximate relationship exists:

$$T_0(\boldsymbol{\theta_0}, \mu) - T_0(\boldsymbol{\theta_0}, \mu_0) \approx -\frac{\partial_2 T_0(\boldsymbol{\theta_0}, \mu_0)}{\partial_1 T_0(\boldsymbol{\theta_0}, \mu_0)} \times (\mu - \mu_0). \tag{A14}$$

(A14) expands to:

$$T_0(\boldsymbol{\theta_0}, \mu) \approx -\frac{\partial_2 H(\boldsymbol{\theta_0}, \mu_0)}{\partial_1 H(\boldsymbol{\theta_0}, \mu_0)} \times \mu + \left(T_0(\boldsymbol{\theta_0}, \mu_0) + \frac{\partial_2 H(\boldsymbol{\theta_0}, \mu_0)}{\partial_1 H(\boldsymbol{\theta_0}, \mu_0)} \times \mu_0\right). \tag{A15}$$

From (A12) and (A14), the limit of the ratio of differences of transformed conditional means to differences of transformed marginal means is

$$\rho(\boldsymbol{\theta_0}) = \lim_{\mu \to \mu_0} \frac{T_0(\boldsymbol{\theta_0}, \mu^c) - T_0(\boldsymbol{\theta_0}, \mu_0^c)}{T_0(\boldsymbol{\theta_0}, \mu) - T_0(\boldsymbol{\theta_0}, \mu_0)} = -\frac{\partial_1 H(\boldsymbol{\theta_0}, \mu_0)}{\partial_2 H(\boldsymbol{\theta_0}, \mu_0)}. \tag{A16}$$

$\rho(\boldsymbol{\theta})$ is a factor of proportionality reflecting changes of transformed conditional and marginal means.

### 11.3.2.2 Special case when the marginal mean depends on a linear predictor

Consider the special model in which the marginal mean (4.1) depends on a linear predictor,

$$\mu = h(\boldsymbol{X\beta}), \tag{A17}$$

where $h$ is a strictly monotone increasing differentiable inverse link function, $X$ is the vector of covariates, $\beta$ is the vector of coefficients, and $\boldsymbol{X\beta}$ is the linear predictor. In this case, it is possible a to define the transformation T in a such way that the linear approximation of the transformed conditional and marginal mean represent a conditional generalized linear model with random effects, and a marginal generalized linear model, respectively, with the same link function and covariates $T(\boldsymbol{\theta_0}, x)$ and $\boldsymbol{X}$, where the ratio of the coefficients of each covariate in the two models is equal to the factor of proportionality $\rho(\boldsymbol{\theta_0})$. More precisely,

$$T(\boldsymbol{\theta_0}, \mu^c) \approx -\rho(\boldsymbol{\theta_0})h^{-1}(\mu_0) + \rho(\boldsymbol{\theta_0})\boldsymbol{X\beta} + T\big(\boldsymbol{\theta_0}, H^{-1}(\boldsymbol{\theta_0}, u)\big), \tag{A18}$$

$$T(\boldsymbol{\theta_0}, \mu) \approx -h^{-1}(\mu_0) + \boldsymbol{X\beta}, \tag{A19}$$

where $T(\boldsymbol{\theta_0}, x) = \frac{T_0(\boldsymbol{\theta_0}, x)\rho(\boldsymbol{\theta_0})}{h'\left(h^{-1}(\mu_0)\right)}$.

As T is an affine transformation, the magnitudes of the coefficients in the induced conditional and marginal models locally reflect the relative magnitude of the differences of the conditional and marginal means.

We proved mathematically that if $-\frac{\partial_1 H(\boldsymbol{\theta_0}, x)}{\partial_2 H(\boldsymbol{\theta_0}, x)}$ as a function of $x$ is strictly concave (convex) everywhere in its domain, then $\rho(\boldsymbol{\theta_0}) > 1$ $(< 1)$, i.e., the magnitudes (i.e., absolute values) of the transformed differences of the conditional means are larger (smaller) than the magnitudes of the corresponding transformed marginal differences (**Theorem 3**). Consequently, the magnitudes of the coefficients in the induced conditional model are larger (smaller) than the magnitudes of the corresponding coefficients in the induced marginal model.

### 11.3.3 Proofs

#### 11.3.3.1 Proof of Theorem 1

$R_M(CWHP_{i1}, CWHP_{i2}) = \sup Corr\big(g(CWHP_{i1}), h(CWHP_{i2})\big)=$

$= su\,p\; Corr\big(g_1(b_i + b_{i1}), h_1(b_i + b_{i2})\big) = R_M(b_i + b_{i1}, b_i + b_{i2}) = q$, where

$g_1$(x)=g($B^{-1}_{\mu_1,\psi_1}\left(\Phi\left(\frac{x}{(\sigma^2+\sigma_1^2)^{\frac{1}{2}}}\right)\right)$) and $h_1$(x)=g($B^{-1}_{\mu_2,\psi_2}\left(\Phi\left(\frac{x}{(\sigma^2+\sigma_2^2)^{\frac{1}{2}}}\right)\right)$) respectively.

#### 11.3.3.2 Proof of Theorem 2

From the equation $H\big(\boldsymbol{\theta}, H^{-1}(\boldsymbol{\theta}, u)\big) \equiv u$ we have by differentiating with respect to μ

$$\partial_1 H\Big(\boldsymbol{\theta}, H^{-1}(\boldsymbol{\theta}, u)\Big) + \partial_2 H\Big(\boldsymbol{\theta}, H^{-1}(\boldsymbol{\theta}, u)\Big) \times \partial_\mu H^{-1}(\boldsymbol{\theta}, u) \equiv 0,$$

hence

$$\partial_\mu H^{-1}(\boldsymbol{\theta}, u)\Big|_{\boldsymbol{\theta} = \boldsymbol{\theta_0}} = -\frac{\partial_1 H\Big(\boldsymbol{\theta_0}, H^{-1}(\boldsymbol{\theta_0}, u)\Big)}{\partial_2 H\Big(\boldsymbol{\theta_0}, H^{-1}(\boldsymbol{\theta_0}, u)\Big)}. \qquad \text{(A20)}$$

We have

$$\partial_\mu T\Big(\boldsymbol{\theta_0}, H^{-1}(\boldsymbol{\theta}, u)\Big)\Big|_{\boldsymbol{\theta} = \boldsymbol{\theta_0}} = \partial_2 T\Big(\boldsymbol{\theta_0}, H^{-1}(\boldsymbol{\theta_0}, u)\Big) \times \partial_\mu H^{-1}(\boldsymbol{\theta}, u)\Big|_{\boldsymbol{\theta} = \boldsymbol{\theta_0}}. \qquad \text{(A21)}$$

Suppose $\partial_\mu T\Big(\boldsymbol{\theta_0}, H^{-1}(\boldsymbol{\theta}, u)\Big)\Big|_{\boldsymbol{\theta} = \boldsymbol{\theta_0}} = m(\boldsymbol{\theta_0})$ does not depend on $u$. From (A20) and (A21) we obtain

$$\partial_2 T(\boldsymbol{\theta_0}, t) = -m(\boldsymbol{\theta_0}) \times \frac{\partial_2 H(\boldsymbol{\theta_0}, t)}{\partial_1 H(\boldsymbol{\theta_0}, t)}, \qquad \text{(A22)}$$

with $t = H^{-1}(\boldsymbol{\theta_0}, u)$. We deduce (A10) and (A11) from (A22) by integration.

Suppose now that (A10) and (A11) hold. From (A11), (A20), and (A21) we have

$$\partial_\mu T_0\Big(\boldsymbol{\theta_0}, H^{-1}(\boldsymbol{\theta}, u)\Big)\Big|_{\boldsymbol{\theta} = \boldsymbol{\theta_0}} = -\frac{\partial_2 H\Big(\boldsymbol{\theta_0}, H^{-1}(\boldsymbol{\theta}, u)\Big)}{\partial_1 H\Big(\boldsymbol{\theta_0}, H^{-1}(\boldsymbol{\theta}, u)\Big)} \times -\frac{\partial_1 H\Big(\boldsymbol{\theta_0}, H^{-1}(\boldsymbol{\theta_0}, u)\Big)}{\partial_2 H\Big(\boldsymbol{\theta_0}, H^{-1}(\boldsymbol{\theta_0}, u)\Big)} = 1,$$

and from (A10) we obtain $\partial_{\mu} T\left(\boldsymbol{\theta}_{\mathbf{0}}, H^{-1}(\boldsymbol{\theta}, u)\right)\Big|_{\boldsymbol{\theta}\,=\,\boldsymbol{\theta}_{\mathbf{0}}} = m(\boldsymbol{\theta}_{\mathbf{0}})$, which does not depend on $u$.

#### 11.3.3.3 Proof of Theorem 3

Assume $u$ is distributed uniformly in the [0, 1] interval, $\boldsymbol{\theta} = (\mu, \ldots)^T, \boldsymbol{\theta} \in \Theta$ is a vector of parameters of $H$, and $\mu^c = H^{-1}(\boldsymbol{\theta}, u)$. We have

$$\mu = E\mu^c = E(H^{-1}(\boldsymbol{\theta}, u)). \tag{A23}$$

Applying (A20), we have

$$\left|\partial_{\mu} H^{-1}(\boldsymbol{\theta}, u)\right| = \left|\frac{\partial_1 H\left(\boldsymbol{\theta}, H^{-1}(\boldsymbol{\theta}, u)\right)}{\partial_2 H\left(\boldsymbol{\theta}, H^{-1}(\boldsymbol{\theta}, u)\right)}\right| = \left|\frac{\partial_1 H(\boldsymbol{\theta}, \mu^c)}{\partial_2 H(\boldsymbol{\theta}, \mu^c)}\right| \leq M(\mu^c), \tag{A24}$$

therefore, using Lebesgue's dominated convergence theorem, we can differentiate under the expectation sign in (A23), which yields

$$1 = E\partial_{\mu} H^{-1}(\boldsymbol{\theta}, u)\Big|_{\boldsymbol{\theta}\,=\,\boldsymbol{\theta}_{\mathbf{0}}} = E\left(-\frac{\partial_1 H\left(\boldsymbol{\theta}_{\mathbf{0}}, H^{-1}(\boldsymbol{\theta}_{\mathbf{0}}, u)\right)}{\partial_2 H\left(\boldsymbol{\theta}_{\mathbf{0}}, H^{-1}(\boldsymbol{\theta}_{\mathbf{0}}, u)\right)}\right) = E\left(-\frac{\partial_1 H(\boldsymbol{\theta}_{\mathbf{0}}, \mu_0^c)}{\partial_2 H(\boldsymbol{\theta}_{\mathbf{0}}, \mu_0^c)}\right). \tag{A25}$$

Suppose $-\frac{\partial_1 H(\boldsymbol{\theta}_{\mathbf{0}}, z)}{\partial_2 H(\boldsymbol{\theta}_{\mathbf{0}}, z)}$ is strictly concave everywhere in its domain. Then we have from Jensen’s inequality, and (A16)

$$1 < -\frac{\partial_1 H(\boldsymbol{\theta}_{\mathbf{0}}, E\mu_0^c)}{\partial_2 H(\boldsymbol{\theta}_{\mathbf{0}}, E\mu_0^c)} = -\frac{\partial_1 H(\boldsymbol{\theta}_{\mathbf{0}}, \mu_0)}{\partial_2 H(\boldsymbol{\theta}_{\mathbf{0}}, \mu_0)} = \rho(\boldsymbol{\theta}_{\mathbf{0}}). \tag{A26}$$

The other case is treated similarly.

## 12 Acknowledgement

First and foremost, I would like to express my deep gratitude to Dr. Zsolt Lang, my supervisor, for his patient guidance, enthusiastic encouragement and useful critiques. I am especially grateful to Dr. László Ózsvári, my co-supervisor, for gathering the data, validating and strengthening the thesis as veterinary expert, for providing guidance, helping me in administrative tasks and for keeping my progress on schedule. My thanks are also extended to Dr. Jenő Reiczigel for his support in certain administrative tasks related to the doctoral procedure.

I also owe a debt of gratitude to my anonymous reviewers who inspired me to go always deeper which helped me to improve my manuscripts.

Last, but not least, I wish to thank my family for their support, patience and encouragement throughout my study.