University of Nottingham
UK | CHINA | MALAYSIA

Development and evaluation of predictive models for real-time decision making on dairy farms

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy
January 2026

Ian David Glover BVSc DipECBHM MRCVS
School of Veterinary Medicine and Science
University of Nottingham

Abstract

Global demand for dairy products is forecast to rise over the coming decades in response to human population growth. Technological innovations are important for ensuring that this demand is met in a socially, environmentally and economically sustainable manner. Predictive models are a cornerstone of precision livestock farming, leveraging data collected on farms for making accurate predictions regarding the health or welfare of livestock. Model predictions can be incorporated into evidence-based, real-time decisions, informing management changes or disease treatments to optimise sustainability. The aim of this thesis was to develop and evaluate such predictive models for use in real-time decision making by farmers or their advisors.

Data pertaining to two prominent diseases of dairy cattle, paratuberculosis and mastitis, were utilised in this thesis. In Chapter 2, a model was developed which predicts the probability of cure of a case of clinical mastitis. Such predictions are useful for making treatment decisions. For example, cows with a sufficiently low cure probability can be selected for non-antimicrobial treatment. The model in Chapter 2 is noteworthy for the absence of a predictor variable indicating the species of bacterial pathogen; predictors in the model include only cow- and herd- level variables and do not require pathogen-identification.

In Chapters 3 and 4, modelling was used to improve interpretation of serial individual cow antibody ELISA results collected within paratuberculosis control-schemes. Interpretation of a series of multiple results, as opposed to a single test result, is non-trivial, and is likely to lead to misclassification of cows with regards to risk of paratuberculosis infection. In Chapter 3 repeated Bayesian updating was investigated, using a likelihood ratio derived from a

pair of Bayesian models to update the posterior probability of infection at each paratuberculosis test result. The poor calibration of the posterior probabilities in Chapter 3 prompted the application of unsupervised learning (hidden Markov models) to longitudinal ELISA result data in Chapter 4, yielding a model which assigns cows to one of a number of ordinal latent states. The latent states are likely to represent different stages of disease, and can provide an early warning of disease progression. Of particular interest in Chapter 4 was the disclosure of a group of cows (state 3) whose test results would usually be considered "negative" according to conventional interpretation of antibody ELISA results, and yet are at heightened risk of progressing to positive test results. In Chapter 5, the novel latent states were compared with the framework most-commonly employed in the United Kingdom (the "J-status") for interpreting serial paratuberculosis ELISA results. State 3 cows were found to be often concurrently low-risk according to this framework. Furthermore, predictions from the hidden Markov model developed in Chapter 4 were informative in a model created in Chapter 5 that predicts the probability of the onset in the subsequent twelve months of the highest-risk category ("J5") according to the conventional paratuberculosis risk framework. These predictions are expected to be useful in making pre-emptive management decisions (for example breeding decisions) according to cows' forecasted risk of progression to "J5" status. In Chapter 6, the implications of the latent states were studied further using inferential models which quantified the association of the latent states with 305-day milk yield, risk of service and risk of conception. Cows in the highest latent states were found to have reduced milk yield compared to those in the lowest states, corroborating the hypothesis that the latent states are aligned with disease progression.

In this thesis, some novel statistical techniques were explored, yielding practically-useful validated prediction models for real-time decision making, and advancing knowledge of the analysis of routinely-collected low-dimensional farm data.

Acknowledgements

I would like to express my gratitude to all those who facilitated this PhD for me. In particular I would like to thank Professor Martin Green, Professor Andrew Bradley, Professor Bobby Hyde and Dr. Luke O'Grady both for enabling me to embark on the PhD in the first place and for providing absolutely superb guidance along the way. I honestly could not have wished for better supervisors. I would also like to thank every one of my colleagues at Quality Milk Management Services, who have been consistently supportive and understanding whilst I attempted (sometimes not altogether successfully) to balance doing a PhD with being a useful member of the team. I am hugely grateful at this stage, and throughout my career, for all of the hard work done by farmers; on this particular occasion, it is ultimately the dedication of conscientious farmers which has resulted in the availability of large, accurate datasets, without which this PhD would have been a "non-starter". My thanks go also to my family for their enthusiastic support of me in pursuing this PhD. Finally, to my kids, Oscar and Bea, thank you for keeping me cognizant of the really important things in life.

Contents

Figures

Tables

Published works

Chapter 2: A derivative of this chapter has been submitted for publication in the Journal of Dairy Science.

Chapter 4: A derivative of this chapter has been published in Preventive Veterinary Medicine under the title **"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"** (https://doi.org/10.1016/j.prevetmed.2024.106413).

Chapter 1. Introduction

1.1. The growing demand for food

The global human population is expected to increase to approximately 10 billion by the year 2050 (FAO, 2017), accompanied by a rise in global food demand of between 35 and 56% between 2010 and 2050 (van Dijk et al., 2021). Dairy products are an important component of human diets, being the third largest source of protein for humans globally, and a crucial source of micronutrients, notably calcium and essential amino acids (Smith et al., 2022). Worldwide dairy production is increasing as a consequence (FAO, 2024).

1.2. Sustainability

Whilst dairy farming can contribute significantly to meeting the growing nutritional demands of the global human population, it must do so in a sustainable manner and in the face of considerable challenges. Sustainability in agriculture can be considered in terms of three criteria: environmental, economic and societal (Capper and Williams, 2023; de Wit et al., 1995). Environmental sustainability is concerned with limiting the impact of food production on the local and global environment. Societal sustainability dictates that food is produced in accordance with high welfare standards and that it augments human wellbeing through outcomes other than provision of nutrients, for example through employment or food safety. Economic sustainability refers to the long-term economic viability of food production, which requires profitability and preservation of local resources.

In a global context, factors which may limit environmental sustainability of dairy farming include climate change, depletion or deterioration of natural resources (e.g. land, water and energy) and deforestation (Capper and Bauman, 2013; de Wit et al., 1995; FAO, 2017; Kyriazakis et al., 2024; National Research Council of the National Academies, 2015; Peterson and Mitloehner, 2021). Currently much attention is paid to the impact of livestock production systems on the environment, particularly with regards to climate change. In 2020, agrifood systems were the source of approximately 30% of global greenhouse gas (GHG) emissions (Gerber et al., 2013), whilst livestock farming is estimated to contribute 14% of global GHG emissions. (FAO, 2022). Ruminant livestock farming has received particular scrutiny due its emissions of methane, a potent GHG; methane from ruminants is the second largest component of global agricultural greenhouse gas emissions (second only to deforestation) (FAO, 2022). Cattle milk production by itself is the cause of one fifth of the global GHG emissions attributed to all types of livestock farming (Gerber et al., 2013). In summary, modern dairy farming faces three key challenges (Michelena et al., 2025). These are 1) increasing output to meet the growing global demand for food; 2) doing so in an environmentally, societally and economically sustainable manner; and 3) minimising disease and maintaining animal welfare in the context of larger herd sizes.

1.2.1. The impact of livestock disease on sustainability

Emissions (including but not limited to GHGs) from farming are simply the proportion of inputs (for example feed or fertiliser) which is not captured, and therefore goes to waste (Kyriazakis et al., 2024).

Emissions of GHGs (e.g. carbon dioxide, nitrous oxide and methane) represent a loss of energy, organic matter and nitrogen from the process of food production (Gerber et al., 2013). In general terms, disease reduces efficiency (e.g. by reducing feed conversion efficiency, reducing yields or increasing culling) of livestock enterprises and therefore increases emissions intensity (emissions per unit of food produced) (Kyriazakis et al., 2024; von Soosten et al., 2020). As an example, a study investigating the environmental impact of mastitis, a prominent dairy production disease, found that an improvement in udder health of dairy herds could reduce GHG emissions from dairy farming by between 2.5 and 12% (Capper and Williams, 2023). Whilst the environmental impacts of many other animal diseases have not been well-quantified (Capper and Williams, 2023), it has been postulated that animal disease in general causes the loss of 20% of the protein produced globally by livestock (WOAH, 2008).

Reduction of livestock diseases can have significant benefits to society beyond maximising food production whilst limiting environmental impacts. For example, zoonotic infections in animals are a common cause of human foodborne illnesses, and can lead to pandemic disease in humans (Perry et al., 2018). Disease incidence is positively associated with antimicrobial usage (AMU) (Lardé et al., 2021), and in turn AMU for management of livestock diseases is implicated in the growth of antimicrobial resistance (AMR) in human infections (O'Neill, 2016). One route through which use of antimicrobials in livestock could influence resistance in human infections is through residues of antibiotics in food. The majority of antibiotic doses administered in dairy herds

are for the treatment or prevention of intramammary infections (IMIs) (Kuipers et al., 2016; Lardé et al., 2021; Stevens et al., 2016), and reduction in the prevalence of IMIs in dairy herds is associated with reduced antibiotic residues in milk (Nguyen et al., 2022). In addition, in Western societies high standards of animal welfare (and by extension low rates of disease) are of increasing importance to consumers (Buller et al., 2018).

The impact of livestock disease can also be considered from a purely economic perspective. The financial loss caused to the global dairy industry by twelve prominent dairy cattle diseases has been estimated at $US 64.7 billion ($341 / cow / year), with subclinical ketosis and clinical mastitis (CM) being the two most costly diseases in North America, Europe and worldwide (Rasmussen et al., 2024). Clearly, minimising animal diseases using cost-effective interventions improves the economic sustainability of livestock farming.

1.2.2. The role of technology in sustainability

Efficiency of milk production, and therefore emissions intensity (GHG per unit of milk produced), can be improved by the adoption of technological innovations (Tullo et al., 2019). As an example, intensification of dairy farming in the United States since the 1940s has led to approximately a 40% reduction in carbon emissions per kilogramme of milk produced (Capper et al., 2009). Reduction of emissions through improvements in efficiency are dependent on adoption of technology (Gerber et al., 2013). Global regional variation in the extent of implementation of technology in dairy farming means that milk production in many areas is less efficient,

and the environmental impacts per unit of milk production globally are greater than in regions where dairy farming is technologically most developed (FAO, 2022).

Whilst the key driver of adoption of technology may be increasing efficiency, for example through reduced disease prevalence (Lovarelli et al., 2020) or through optimizing feed inputs (Wathes et al., 2008), such innovations can, and should arguably be required to, also have a beneficial effect on animal welfare, food safety, the environment, individual farmer wellbeing and society as a whole. Adoption of digital technologies in livestock farming can facilitate prediction, and therefore prevention, of animal diseases by providing accurate and objective information for use in decision-making by producers (Neethirajan, 2023). Detection of disease in individual animals by farmers is especially challenging given increasing average herd size, but technology can be used to monitor animals on an intermittent or continuous basis (Neethirajan, 2023), permitting early detection of adverse health conditions. Therefore, considered adoption of technology should be viewed in a positive light in the context of animal welfare, alleviating societal concerns regarding the ethics of livestock farming. Wellbeing of producers can be improved by the increased profitability and freedom from time-consuming and repetitive monitoring (Neethirajan, 2023). Society as a whole can benefit not only from improved access to high quality food, but also for example through improved food safety, achievable through earlier detection and therefore elimination of zoonotic diseases, improved traceability (Banhazi et al., 2012), and a reduced requirement for antimicrobials in livestock

farming as part of a strategy to minimise AMR in human infections (Neethirajan, 2023).

1.2.3. Digitalisation of farming and precision livestock farming

Digitalisation refers to the use of any digital tools by farmers, and includes an array of techniques for the collection and processing of on-farm data, ultimately providing valuable insights to inform management decisions (Kleen and Guatteo, 2023). The term precision livestock farming (PLF) is often used to refer to a highly-automated type of digitalisation where management is informed or controlled automatically according to the outputs from complex computer algorithms processing continuously-collected sensor data (Kleen and Guatteo, 2023). The components required for PLF are 1) on-farm sensors (e.g. cameras, microphones, pressure sensors etc.) providing continuous or high-frequency data regarding animals or their environment; 2) a model used to interpret the data and predict an outcome (e.g. a health condition such as mastitis or lameness); 3) integration of predictions into on-farm decision-making or automated actuators (e.g. for regulating the animals' environment). One of the aims of PLF is to enable intensive monitoring of animals, herds or the environment in the context of modern dairy farming, where increasing herd sizes lead to farmers being responsible for the health and welfare of larger numbers of animals (Berckmans, 2017; Neethirajan, 2023; Wathes et al., 2008).

PLF is often defined purely with regards to continuous or high-frequency monitoring by automated sensors (Berckmans, 2014; Wathes et al., 2008). However, intermittently-collected data, for

example veterinary health records or results of diagnostic tests, can also be part of PLF. Despite not being continuously monitored, intermittent data can be incorporated into predictive models facilitating on-farm decision-making (Michelena et al., 2025). Therefore, use of such data can be framed as a branch of PLF, with the same goal of environmentally, socially and economically sustainable maximisation of food production.

PLF techniques have been developed for detection of a variety of dairy cow health conditions, including mastitis, lameness, heat stress and sub-acute ruminal acidosis (SARA), as well as for detecting the onset of physiological reproductive events (e.g. oestrus and calving) (Michelena et al., 2025).

1.3. Statistical learning

Central to the concept of PLF is the predictive model which processes the data. Predictive modelling is a computer technology which creates models to identify associations between variables in data. By precisely describing the associations between particular variables (e.g. clinical signs) and an outcome (e.g. presence or absence of a disease), these models can be used to predict the outcome on novel unlabelled (unknown outcome) data. The novel data could be collected in a different location, or at a future timepoint. The overall aim of this methodology is to make accurate predictions and inform decision making (Kuhn and Johnson, 2013). Predictive modelling can be viewed as analogous to human learning, where perceptual information (the data) is processed by a trained model (the brain) during decision-making (Griffiths and Tenenbaum, 2006). Hence predictive modelling is also referred to

as statistical learning, which in turn is an important component of machine learning and artificial intelligence (Gefeller et al., 2017; Haug and Drazen, 2023).

1.3.1. The multiple linear regression model

Models which predict a continuous outcome are commonly termed regression models. One such model is the multiple linear regression model, which takes the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i,$$

where y_i is the value of the dependent variable for the i^{th} subject drawn from, for example, a Gaussian distribution, β_0 is the intercept, $\beta_{1\ldots k}$ are coefficients for the k independent variables, $x_{1\ldots k}$ are the independent variables of interest and ε_i is an error term with a Gaussian distribution and mean zero. This equation defines that y_i is associated linearly and additively with the variables of interest. The aim of model training is to estimate the values of β_0 and $\beta_{1\ldots k}$ (the model parameters), so that $\hat{y}_i$ (the predicted value of the dependent variable for the i^{th} subject) can be calculated in novel data if the values of $x_{i1\ldots ik}$ are known. Model parameters may be estimated by choosing those which minimise the sum of squared errors (SSE) (Roustaei, 2024), defined as:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

where y_i is the observed value of the dependent variable and $\hat{y}_i$ is the predicted value of the dependent variable for the i^{th} subject. The term $y_i - \hat{y}_i$ is known as the residual, that is the difference between predicted and observed values of the dependent variable. The SSE

in this context is known as the loss function, to be minimised when fitting the model.

1.3.2. The generalised linear model

The linear model can be adapted for use where the dependent variable has a discrete, rather than continuous, distribution, in which case the task is referred to as classification rather than regression. As an example, the dependent variable can have a Bernoulli distribution (0 or 1, true or false). Such models are an extension of the linear regression model, and are referred to as generalised linear models (GLMs). Classification models are used to predict the probability of the dependent variable having a certain discrete value (Kuhn and Johnson, 2013). In GLMs, the predictor variable values are again mapped to the dependent variable with a linear additive equation. However, in regression models the distribution of y in the linear equation must be continuous from $-\infty$ to ∞, requiring some transformation of the discrete dependent variable through a link function (Bewick et al., 2005). In logistic regression the link function is the logit (the natural logarithm of the odds).

The logistic regression equation is:

$$logit(\pi_i) = ln\frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik},$$

where π_i is the probability of the dependent variable for the i^{th} subject being positive, β_0 is the intercept, $\beta_{1\ldots k}$ are coefficients for the k independent variables and $x_{1\ldots k}$ are the independent variables of interest. The error term (residual) is absent from this equation, but is implicit in the binomial distribution of the dependent variable

(Hosmer and Lemeshow, 2002). The i^{th} subject is classified according to the model-predicted probability, π_i, and a tuneable cutpoint (diagnostic threshold). For example, if the chosen cutpoint is 0.5, then all subjects with $\pi_i \geq 0.5$ are predicted to have a positive value of the dependent variable. The model parameters may be estimated by optimising the model likelihood (the loss function) (Bewick et al., 2005), which is defined as:

$$\sum_{i=1}^{n} f(y_i \mid \Theta),$$

where f is the probability mass function of the dependent variable y_i given the model parameters, Θ.

1.3.3. Mixed effects models

For linear and generalized linear regression models, subjects within the data must be independent and identically distributed (i.i.d.). That is, for every variable each row of data is drawn from the same probability distribution, and the values of variables for one subject are not correlated with those of other subjects. In real-world data, such an assumption rarely holds true (Hoffman and Walters, 2022); subjects often reside in clusters, with subjects being more like individuals within the same cluster than in other clusters. As an example in the current context, mixed effects models are useful to account for clustering between herds when cows are more similar to other cows within the same herd than they are to cows in other herds. Furthermore, in longitudinal data where each subject is observed multiple times, observations within the same individual are expected to be more similar than observations from different subjects (Bell et al., 2019). As a result, the residual will vary between subjects or groups of subjects, and cannot be specified in the

regression equation by a single error term ε. Therefore, random effect terms are present in the mixed effects model, capturing between-cluster residual variation in the intercept β_0 or the coefficients (regression slopes) $\beta_{1\ldots k}$ (Hoffman and Walters, 2022). These random effect terms are usually specified to have a Gaussian distribution (mean zero) (Hoffman and Walters, 2022).

The hierarchy of random effects in mixed models can take various forms. In the simplest form, a two-level random intercept model (e.g. cows nested within herds), variance between herds of the intercept β_0 is described by one additional error term. As an extension of this model, coefficients of the level 1 (cow-level) independent variables can also be permitted to vary between herds (a random intercept and random slope model). In a three-level nested model (for example cows nested within herds nested within geographical regions), further error terms are introduced permitting the intercept β_0 or coefficients $\beta_{1\ldots k}$ to vary at a third level (e.g. between regions). As well as such nested random effects structures, random effects may be crossed. This is applicable when level 1 subjects can be members of more than one higher-level cluster. For example, data may be collected during two separate clinical trials conducted on the same farms. In this case, cows are nested within both farm and clinical trial; two cows on any given farm may be enrolled in different trials.

The linear mixed model equation is:

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \cdots + \beta_k x_{ijk} + \mu_{0j} + \mu_{1j} x_{1j} + \mu_{2j} x_{2j} + \cdots + \mu_{kj} x_{kj} + \varepsilon_{ij},$$

where y_{ij} is the value of the dependent variable for the i^{th} subject in the j^{th} group, $x_{ij1\ldots k}$ and $\beta_{1\ldots k}$ are the values and coefficients of the k independent variables (fixed effects), μ_{0j} and $\mu_{1j\ldots kj}$ are terms permitting variance of the intercept and independent variable coefficients between level-2 groups, and ε_{ij} is the residual for the i^{th} subject in the j^{th} group. Inclusion of random effects in a model serves three purposes: 1) The variance of the random effects may be informative if there is interest in how much variation there is between groups in the data; 2) when using mixed effects models for making predictions on novel data, accounting for between-group heterogeneity in the model permits predictions to be made based on only the fixed-effects parameters, thus making predictions under the assumption that the novel data is drawn from an "average" population (Harrison et al., 2018), and 3) standard errors of the fixed-effects coefficients are often more conservative in mixed effects models, compared with models containing only fixed effects, helping to avoid false statistical inferences (Bell et al., 2019; Hoffman and Walters, 2022).

The linear mixed model can also be adapted to the case in which the dependent variable has a discrete (e.g. Bernoulli) distribution, through use of a link function such as the logit; such models are known as generalised linear mixed models (Dean and Nielsen, 2007). Various methods exist for estimating the parameters of a generalised linear mixed model, including maximising the model

likelihood, or approximations thereof, and Markov Chain Monte Carlo (MCMC) methods (Section 1.3.7) (Bolker et al., 2009; Tuerlinckx et al., 2006).

1.3.4. Inference, hypothesis-testing and prediction

Statistical models are commonly used for one of three purposes: data-exploration, inference or prediction (Tredennick et al., 2021). Data-exploration involves examining the correlation between variables in a dataset (for example to formulate new hypotheses), and inference (or hypothesis testing) means producing accurate estimates of the association of a small number of variables of interest with an observed response, providing evidence of the underlying mechanisms of the process being studied (Bzdok and Ioannidis, 2019). Simpler models, such as linear regression models, are ideally suited to inference as their fitted parameters are readily interpretable, which is desirable for decision support tools in healthcare settings (Tohka and van Gils, 2021).

Model-selection processes, and therefore the resulting most useful model, vary with the purpose of the model. Explorative models naturally include more variables than inferential models, the latter of which test an *a priori* hypothesis regarding a small number of predictors (Tredennick et al., 2021); inclusion of variables in the inferential model is guided by expert knowledge of the underlying real-world process (Bzdok and Ioannidis, 2019). Predictive models are created with one purpose, that is to make accurate predictions on novel data (Kuhn and Johnson, 2013). Since inference is of minor relevance to the goal of predictive modelling, that is the ability to make accurate predictions on novel data, variables with spurious

associations with the observed response may be included in predictive models, and the parameter estimates are of little interest (Tredennick et al., 2021). This has enabled the development of highly sophisticated machine learning algorithms (so called “black box” models, for example artificial neural networks) for creation of predictive models whose fitted parameters are not easily interpretable in the manner of simpler regression models (Bi et al., 2019). Whilst such sophisticated models may provide highly accurate predictions given appropriate model-training data, use of simpler models may still be more desirable than “black box” models due their portability, that is the ease with which the model can be implemented outside of the context in which it was created; the intercept and coefficients of a linear regression model are all that is required for making predictions on new data, whereas more complex models may require the use of software or data used during model-training for making predictions (Hawkins, 2004).

1.3.5. Generalisability and regularisation

Generalisability is the ability of a model to predict the outcome of interest in the general population with acceptable error, and can be assessed by measuring the error of predictions made on a suitable sample of data not used in model training, tuning or selection. Central to the concept of generalisability of predictive models is the bias-variance trade-off. The sum of squared errors (a measure of the difference between model-predicted and observed values of the dependent variable) comprises three components, bias, variance and irreducible error.

SSE is defined as follows (Hastie et al., 2009):

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = bias^2 + variance + \varepsilon,$$

where ε is the irreducible error. Bias is the inverse of the ability of a model to describe the relationship between predictor variables and an outcome in the data, and variance is the tendency of the model to be influenced by the spurious relationships ("noise") which are particular to the training data (Kuhn and Johnson, 2013). Bias and variance are negatively correlated (hence there is a "bias-variance trade-off" when selecting the best model), with simpler models tending to have less variance but more bias, and more complex models tending to have more variance but less bias. The hypothetical model which describes perfectly the data-generating process would have no bias or variance, so that any residuals between predicted and observed values of the outcome would be due solely to irreducible error. The process of creating predictive models therefore involves selecting a model with minimal bias and variance, which describes as closely as possible the true underlying relationships between variables in the process being studied. The accuracy of predictions made on training data of a model with excess variance will be good (Kuhn and Johnson, 2013), but the prediction accuracy will be disappointing when using such a model for predicting on novel "hold-out" data (Miller et al., 2024). Such models are said to be overfit, since they have captured spurious relationships (noise) in the training data. Overfitting is a result of training overly-complex models, for example models with many covariates or variable transformations, or models able to capture particularly complex interactions between covariates (Hawkins, 2004). Underfit models, those with excess bias, will make poor

predictions on both training and testing data. Overfit and underfit models are unlikely to be generalisable. However, overfitting is more likely than underfitting. Underfitting (e.g. exclusion of an important variable from a model) results in a larger deterioration of the loss function than overfitting (Yates et al., 2023), in accordance with the quadratic $bias^2$ term in the above equation.

Figure 1.1 illustrates a small simulated dataset with two variables, x and y, which are related by the function $y = x^2 + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$, along with predictions made by three models of varied complexity trained on the data. The simplest model, the linear regression $y = \beta_0 + \beta_1 x$, does not adequately describe the curvilinear association defined by the function, $y = x^2$, used to generate the data. This model has excess bias and is therefore underfit, as it fails to adequately describe the true nature of the association between x and y. The most complex model, a machine learning algorithm with excess variance in the context of these data, is able to capture the "noise" which arises due to the error term ε in the function used to generate the data. This model is therefore overfit. Neither the underfit or overfit models are expected to produce satisfactory predictions on novel samples of data created using the same data-generating function. The best model in this case is the model with intermediate complexity $y = \beta_0 + \beta_1 x + \beta_2 x^2$, which satisfactorily captures the true relationship between x and y.

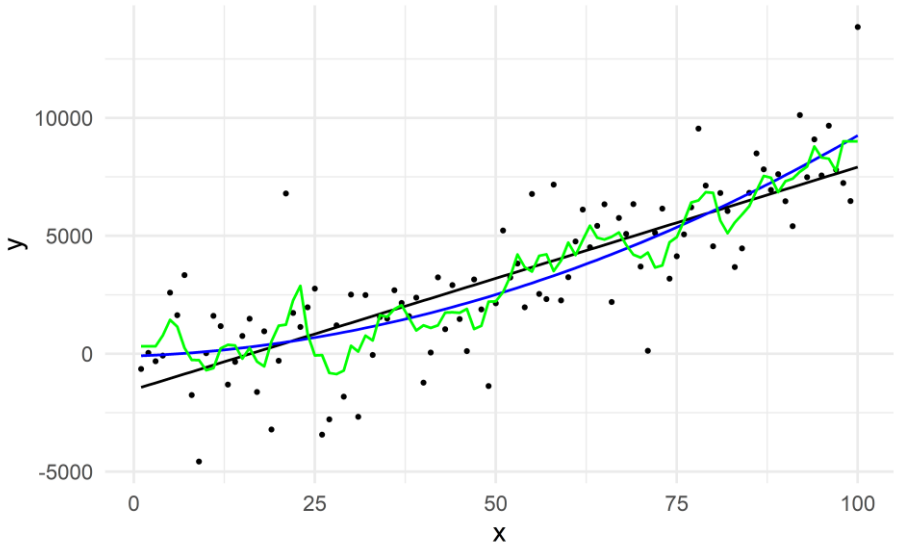

Figure 1.1: The bias-variance trade off.
The curvilinear function $y = x^2 + \varepsilon$, represented by black points, with the predictions of three models of varied complexity shown as lines. In order of increasing complexity, the models are $y = \beta_0 + \beta_1 x$ (black), $y = \beta_0 + \beta_1 x + \beta_2 x^2$ (blue) and a k-nearest neighbours algorithm (green).

Regularisation is the control of model complexity in order to adjust the bias-variance trade-off to avoid overfitting (Friedrich et al., 2023). One method of regularisation is penalisation, which imposes an additional penalty on the loss function used in model-fitting.

For example, in least absolute shrinkage and selection operator (LASSO) regression, the linear regression loss function (SSE) is modified as follows:

$$Penalised\ SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P}|\beta_j|$$

where $|\beta_j|$ is the absolute value of the coefficient for the j^{th} predictor, of P total predictor variables, and λ is a parameter controlling the magnitude of the penalty term.

The penalty term in LASSO regression, $\lambda \sum_{j=1}^{P} |\beta_j|$, known as the L1 penalty, tends to shrink coefficients towards zero, and for some variables can shrink the coefficient to zero, effectively eliminating that variable from the model. LASSO regression can therefore be used for variable selection. An alternative penalised regression, known as Ridge regression, uses a penalty term (the L2 penalty) based on the square of the values of the coefficients:

$$Penalised\ SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} \beta_j^2$$

The penalty term in Ridge regression also has the effect of shrinkage on the coefficients, but unlike LASSO will not shrink them to zero. However, shrunk coefficients are not necessarily representative of the underlying data-generating process and so penalised linear regression is unreliable for inference (Heinze et al., 2018). Both the L1 and L2 penalty terms can be used simultaneously in a machine-learning algorithm known as the Elastic Net, which is described later.

Other regularisation methods include the use of regularising priors in a Bayesian analysis, use of mixed effects models where the specified random effect distribution shrinks random effects (intercepts and slopes) towards zero, and tuning of model hyperparameters to shift the balance of bias and variance (Friedrich et al., 2023).

Some models have tuneable hyperparameters which control the complexity of the model (Kuhn and Johnson, 2013), for example through varying the degree of penalisation of coefficients (e.g.

Elastic Net) or pruning of trees in tree-based models (Friedrich et al., 2023).

In the Elastic Net model, the Ridge and LASSO regression loss functions,

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{P} \beta_j^2$$

and

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_2 \sum_{j=1}^{P} |\beta_j|$$

are combined in a penalised loss function with two tuneable hyperparameters, α and λ:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{P} {\beta_j}^2 + \lambda_2 \sum_{j=1}^{P} |\beta_j|,$$

where

$$\lambda_1 = \lambda * \frac{(1-\alpha)}{2}, \text{ and}$$

$$\lambda_2 = \lambda * \alpha$$

Therefore as λ increases, the degree of penalisation also increases. Adjustment of the other hyperparameter, α, controls the relative severity of the two penalties (L1 and L2), such that higher values of α will increase the L1 penalty and decrease the L2 penalty, and *vice versa* for lower values of α. Adjustment of the two hyperparameters α and λ can effectively control model complexity, avoiding under- and over-fitting.

1.3.6. Model fitting, selection and validation

Scoring rules are used to calculate the difference within each pair of model-predictions and observed outcomes (Gneiting and Raftery, 2007), yielding a single numerical value for each pair. Models are fit by minimising a function (e.g. average, sum or product) of these numerical scores over all data, thus optimising model parameters (in which case the function is known as a "loss function"). During subsequent model assessment the scoring rule is used to assess the accuracy of predictions of a fitted model (and can thus be referred to as a "performance metric").

The output of the loss function is therefore a measure of the difference between desired and actual model outputs (predictions) over all data (Terven et al., 2025). If the loss function is based on the mean of the scoring rule across all data points, it takes the general form (Wang et al., 2022):

$$\min_{f} \frac{1}{n} \sum_{i=1}^{n} L_{\Theta}\big(f(x_i)\big) + \lambda R(f) \,,$$

where n is the number of subjects, L is the loss function associated with model parameters Θ, $f(x_i)$ is the function of x (the model) applied to data for the i^{th} subject, $R(f)$ is an optional penalty applied if regularisation is desired, and λ represents one or more parameters controlling the degree of penalisation (model hyperparameters). Performance metrics (for assessing performance of fitted models) do not normally contain the penalty term, $\lambda R(f)$. In linear regression the loss function is usually the SSE (also known as the squared error loss), which reflects the difference (residual) between predicted and observed values of the

continuous dependent variable. A related loss function is root mean squared error (RMSE), defined as:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Choice of scoring rule is multifaceted, and must take into consideration the intended use-case of the model, and the nature of the data. Scoring rules for models whose output has a continuous distribution (e.g. linear regression, or a classification model used for making probabilistic predictions), are usually a product of the residual, $y - \hat{y}$, whereas development of models whose output is discrete (e.g. a classification model whose desired output is a strict category rather than a probability) may in addition employ scoring rules indicating whether each predicted class was correct. Veracity of predicted classes can be illustrated using a confusion matrix counting the number of correct and incorrect classifications (Table 1.1). Loss functions can then be derived which favour desirable model properties, for example maximising accuracy (the proportion of class predictions which are correct) or maximising positive or negative predictive values as required. As a practical example, a classification model designed to detect as many cases of a disease as possible could prioritise sensitivity (true positive rate) (Miller et al., 2024), in which case the loss function and performance metric should do so also.

	Predicted	
Observed	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Table 1.1: A confusion matrix for a model predicting a Boolean outcome, from which scoring rules for classifiers can be derived.
Subjects are cross-classified according to predicted and observed categorical values of the dependent variable. TP = number of true positives (where predicted and observed values are both positive), FP = number of false positives, FN = number of false negatives, TN = number of true negatives. Loss functions and performance metrics are thus derived, for example accuracy = (TP + TN) / (TP + FP + TN + FN), sensitivity = TP / (TP + FN), specificity = TN / (TN + FP), positive predictive value = TP / (TP + FP), negative predictive value = TN / (TN + FN).

For classification models, if data is unbalanced (i.e. if one class of the dependent variable considerably outnumbers another class), care must be taken to select a scoring rule which accounts for the imbalance (Kuhn and Johnson, 2013). Failure to do so can result in models which predict poorly on novel data with a different prevalence (Rainio et al., 2024; Tohka and van Gils, 2021). Additionally, since the confusion matrix is a function of prevalence, use of a scoring rule derived from the confusion matrix (Erickson and Kitamura, 2021) can result in models which are naïve in that they are adept at predicting one class but not the other. For example, in a dataset where 90% of observed class labels are positive, use of accuracy (proportion of class predictions which are correct) as the loss function could lead to a model which achieves optimal (90%) accuracy simply by always predicting the majority class. Scoring rules which are not functions of the prevalence (e.g. sensitivity, specificity, area under the receiver operating curve (ROC)) are useful in the face of class imbalance. Class weights can also be incorporated into some scoring rules (for example weighted logarithmic loss) so that inaccurate predictions made on subjects

with one class label can be penalised more heavily than poor predictions made on subjects with other class labels, thus prioritising accurate prediction of one class (e.g. the minority class) over another (Spies and Ng, 2025).

A related concept is the threshold chosen for deriving class predictions from numerical model outputs. For example, in logistic regression, the model outputs are the probabilities for each subject of each class of the dependent variable. Class predictions are then made according to a predetermined probability threshold (Spies and Ng, 2025) around which probabilistic predictions are categorised (e.g. as positive or negative). Choice of threshold will influence the confusion matrix, and therefore scoring rules which are functions of the confusion matrix will in turn be affected. Use of scoring rules which are insensitive to the threshold (e.g. area under the receiver operating curve) would be an appropriate choice if the intention is to optimise the model through tuning of this threshold.

Residuals may have positive or negative values, such that summation or averaging of the scoring rule may lead to favourable values of the loss function or performance metric despite large residuals (Naser and Alavi, 2023). To counter this issue, some scoring rules (for example SSE) are a product of the square of the residual, so that positive and negative residuals do not cancel each other out. However, the resulting non-linear relationship between residuals and score penalises relatively large residuals to a disproportionately great extent, biasing parameter estimates in the presence of outliers (Dumre et al., 2024). An alternative approach to avoiding cancellation of residuals of opposite signs is to use the

absolute value of each residual (Naser and Alavi, 2023), in which case the metric is less sensitive to outliers.

“Proper” loss functions optimise model parameters so that model predictions are at least as good as any other hypothetical model (the loss function may have multiple minima at various parameter values), whereas “strictly proper” loss functions are minimised by models whose predictions are better than any other hypothetical model (Gneiting and Raftery, 2007). In the context of this thesis, in which the focus is on models with a probabilistic output, a strictly proper loss function is therefore most appropriate when well-calibrated numerical predictions are sought.

Test data against which model performance is assessed is ideally taken from a different sample (Tohka and van Gils, 2021), providing reassurance that a selected model is generalisable rather than applicable to just the training data. However, collection of multiple data samples is often practically difficult, so the data from one sample may be partitioned into two or more subsets, yielding designated model-training and -testing datasets. Evaluation of models using data not used in any way during training avoids the problem of resubstitution, in which apparent model performance is assessed using the model-training data, yielding over-optimistic performance estimates (Tohka and van Gils, 2021). Test data must be used once only, to estimate the performance of the final selected model (Bradshaw et al., 2023). The scenario where information in the test dataset is unwittingly used for model development or selection is known as “data leakage”, and leads to overoptimistic estimates of model performance (Kapoor and Narayanan, 2023), as

the model has been inadvertently trained in part using information within the testing data.

Model performance, properly assessed, is crucial for model-selection and informs the modeler of the likely performance of the selected model when deployed on new data. The method of data-splitting is important for ensuring model generalisability (Tohka and van Gils, 2021) and for this reason there are a number of important considerations. Data leakage can occur when the method of splitting is naïve to dependence within non-i.i.d. data (Kapoor and Narayanan, 2023). For example, with repeated-measures data (multiple rows per subject or group), leakage occurs if data are split such that data for individual animals or groups are present in both the training and testing subsets; information regarding animals in the testing subset is also then present in the training subset. To avoid data-leakage, data should be split such that individuals in a particular group are not present in multiple subsets. For classification tasks with imbalanced data (where the prevalences of the different classes of the dependent variable differ), it is also important to employ stratified splitting; data are split such that the prevalence of the classes is approximately equal between subsets (Tohka and van Gils, 2021).

An elaboration of the data-splitting method is to create three subsets (training, validation and testing) (Bradshaw et al., 2023). Models trained using the training set are selected according to their performance against the validation set, and the performance of one final selected model is assessed against the testing set. Again, the testing set is used once only, for assessing a final model. The aim of

this data-splitting method is to reduce the risk of overfitting to the training set.

One drawback to the use of data-splitting for provision of a test dataset is the element of luck. Random splitting of a data sample, even when stratifying by class label and accounting for dependency within the data, could by chance yield a testing dataset for which a model gives particularly good or bad predictions, resulting in biased estimates of model generalisability (Miller et al., 2024). Therefore, there is a degree of instability of performance metrics produced by data-splitting, especially if the amount of data available is limited (e.g. less than 1000 rows) (Tohka and van Gils, 2021).

An approach for providing more stable estimates of model performance is resampling. Resampling methods take random samples of the model-training data, fitting a specific model of interest (e.g. with a certain set of independent variables or certain hyperparameters) to the sample, and then assessing the performance of the model on the data outside of the sample (the "holdout" or "out-of-bag" data). Repetition of this process, with averaging or pooling of a performance metric across all samples, provides realistic estimates of the expected predictive ability of a model created using many hypothetical similar datasets (Bates et al., 2024), and mitigates the risk of selecting an overfit model (Bradshaw et al., 2023). Many models (with varied predictors, parameters and hyperparameters) can be created and assessed using resampling, which can therefore be used both for selecting from a pool of candidate models as well as estimating the selected model's performance. Once the most favourable model specifications have been selected, a final model is trained on all of

the data. Common methods of resampling include cross validation (CV) and bootstrap resampling.

Cross validation is the process of estimating model predictive accuracy by repeatedly splitting the data and calculating the average performance across all subsets (Arlot and Celisse, 2010). In *k*-fold cross validation (*k*-fold CV), the data is split into *k* approximately equal size folds. One fold is held out as testing data and a model is trained on the remaining folds. Performance is then assessed using the held-out fold. This is repeated until all of the *k* folds have been used once as testing data, and performance metrics are pooled or averaged. The model specifications resulting in optimal performance are used to train a final model on the entirety of the data. Choice of the value of *k* is somewhat arbitrary, but typical values are 5 or 10 (Kuhn and Johnson, 2013). Cross-validated estimates of model performance can be considered in terms of a bias-variance trade off (Arlot and Celisse, 2010; Wilimitis and Walsh, 2023). Higher values of *k* cause the models within CV to be trained on a sample of data which more closely resembles the whole, thereby reducing the bias of performance estimates for the final model trained on the entirety of the data (Arlot and Celisse, 2010; Yates et al., 2023). If *k* is at least 10, bias of CV estimates of model performance will be negligible (Yates et al., 2023). However, increasing values of *k* will also increase variance (Arlot and Celisse, 2010), as the set of models created during CV will be tested using smaller (more variable) held-out folds. Excess variance of the CV estimates reduces the consistency of the procedure, meaning model-selection and -assessment may be unstable (Heinze et al., 2018). Choice of *k* can therefore be used to adjust the bias-variance

trade-off of the CV procedure, and is informed in part by the perceived signal-to-noise ratio in the data; if the ratio is large (strong signal), minimal bias is most appropriate, prompting higher values of *k*, whereas CV with noisy data will benefit from some bias and less variance, and therefore smaller values of *k* may be appropriate (Arlot and Celisse, 2010).

The cross-validated performance of the model depends on exactly how the data is folded (Krstajic et al., 2014), which is always to some extent random, and therefore *K*-fold CV is usually repeated multiple times, re-folding the data at the start of each repetition. This reduces the variance of estimates of model performance (Arlot and Celisse, 2010) as the CV estimates utilise more resamples of data.

There is a special case of *k*-fold CV, known as leave-one-out cross validation (LOOCV) where *k* is equal to the number of data points *n*. The data are split *n* times, each time partitioning the data such that a different single data point is held out as the test set (Arlot and Celisse, 2010). This approach maximises the data available for each model, and each training set created during LOOCV is almost identical to the whole dataset, minimising the CV bias. However, LOOCV has high variance and is computationally more expensive than *k*-fold CV (Arlot and Celisse, 2010).

In its simplest form, *k*-fold cross validation uses random folding of the data. However, for the same reasons which apply to simple data splitting, when data are imbalanced or non-i.i.d. respectively, stratification or splitting with respect to groups are more appropriate approaches to creating folds (Kuhn and Johnson, 2013).

Furthermore, for models with random effects, splitting such that individuals in each particular group are present in only one fold provides estimates of the accuracy of model predictions not conditioned on the random effects, therefore estimating the generalisability of the model to hypothetical new data from novel groups (Yates et al., 2023). Such approaches are often referred to as "leave one group out" or "leave *n* groups out" CV.

Cross validation is commonly used for the dual purposes of model selection and assessment of the performance of the selected model. Using the CV methods described above for both purposes would result in data leakage that would positively bias CV estimates of model performance, because model development (e.g. optimisation of predictor subset, parameters and hyperparameters) and model assessment both utilise the same folds of data (Tohka and van Gils, 2021). This problem can be avoided by the use of nested CV, which separates the tasks of model selection and evaluation (Yates et al., 2023). However, nested CV is more complex and computationally-expensive, so non-nested CV is often employed regardless (Wilimitis and Walsh, 2023), leading to overoptimistic estimates of model performance. Nested CV consists of an inner and an outer loop, as shown in Figure 1.2. Of note is the fact that different models can be selected by the CV procedure within each outer fold. The nested CV does not therefore indicate one optimal model, but instead provides an estimate of the performance of a hypothetical model created and selected by applying the pipeline contained within the inner loop to the entirety of the data (Bradshaw et al., 2023; Krstajic et al., 2014). To obtain a final optimal model, the pipeline is implemented using the entirety

of the data, yielding one model whose performance has been estimated by nested CV.

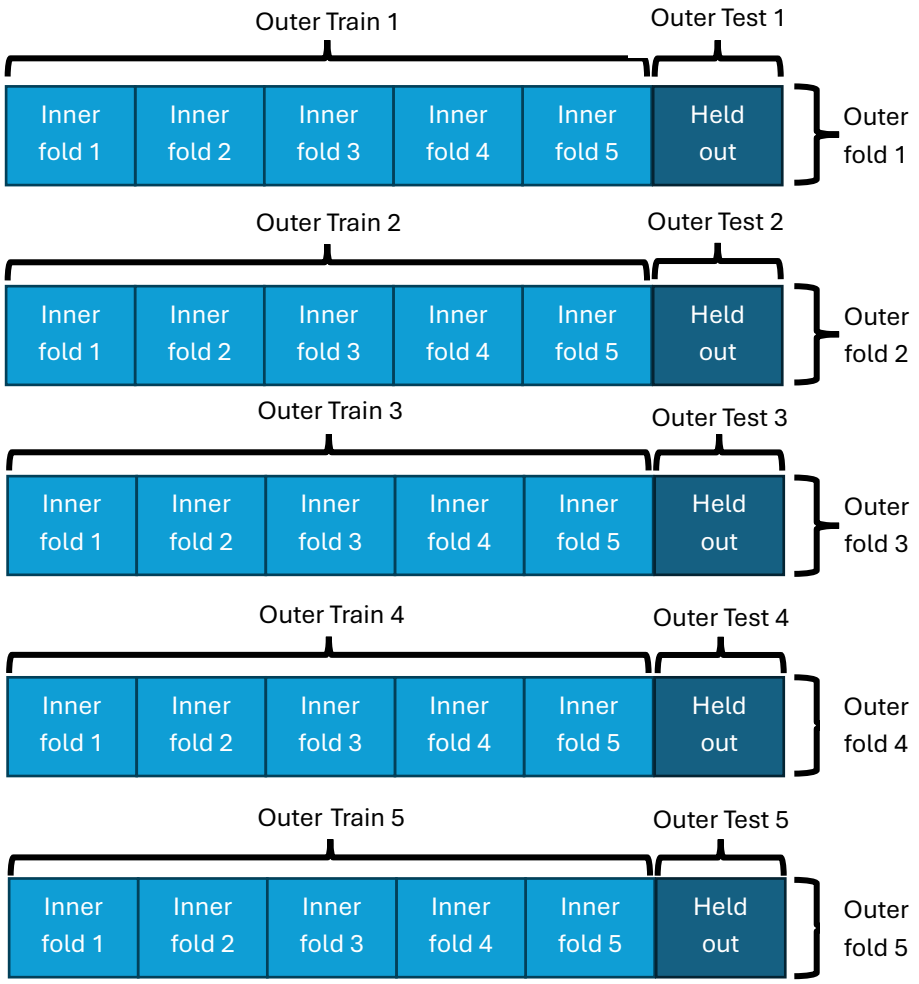

Figure 1.2: The procedure of nested cross validation, demonstrating the inner and outer loops

The data are folded into a number of outer folds (in this case 5). Each outer fold is itself folded, yielding a held-out outer test set and a number (in this case 5) of inner folds. In the inner loop, non-nested cross-validation is performed using the inner folds, yielding an optimal model for each outer fold. In the outer loop, the optimal model for each outer fold is then evaluated using the held out outer test set for that outer fold. Model performance metrics are averaged across the five outer folds.

Bootstrapping is the process whereby data is repeatedly sampled many times with replacement, yielding multiple resamples of the same size as the original data (Kuhn and Johnson, 2013). Models are created using the resamples and evaluated using the data not contained in the resamples (the "out of bag" data) (Tohka and van Gils, 2021). In general, model performance estimates obtained from bootstrapping have lower variance than *k*-fold CV due to the high number of resamples used (Kuhn and Johnson, 2013). However, bias is higher compared with 5- or 10-fold CV, since the data is sampled with replacement and on average each bootstrap sample (training set) contains only 63.2% of the total data (Bradshaw et al., 2023); each resample resembles the original data less than those produced during *k*-fold CV utilising a routine (e.g. 5 or greater) number of folds. In 10-fold CV, for example, each training set includes 90% of the total data.

The model likelihood, that is the probability of observing the data given the parameters of the model being scrutinised, is defined as:

$$L(\Theta) = \prod_{i=1}^{n} f(x_i|\Theta),$$

where Θ denotes the model together with its parameters, and $f(x_i|\Theta)$ is the probability of the i^{th} data point, x_i, given the model. The likelihood is therefore a measure of how well the model describes the data, and can be used for model selection. Given a large dataset, $L(\Theta)$, which is a product of n factors, may be impractically small and so, for convenience, the natural logarithm of the likelihood, $\ln(L(\Theta))$ is routinely used. For comparing nested models (e.g. with or without a particular predictor variable), a likelihood ratio test (LRT) can be used to assess if one model fits the

data significantly better than the other. The likelihood ratio test statistic when comparing two models, Θ_1 and $\Theta_{2,}$ is:

$$-2 * [\ln L(\Theta_1) - \ln L(\Theta_2)]$$

The difference between the logarithmically-transformed likelihoods of the two models is equal to the ratio of their likelihoods on the linear scale. The test statistic is compared to a chi-squared distribution, yielding a P-value indicating if one model is significantly better-fitting than the other. However, since the model likelihood in this case denotes the fit of the model to the training data, model-selection by optimisation of the model likelihood risks overfitting.

Analogous to penalised regression techniques, to mitigate overfitting a penalty term can be added to the model likelihood, yielding the general form of the information criterion (IC) (Dziak et al., 2020):

$$IC = -2 * \ln L(\Theta) + A_n . p,$$

where $-2 * \ln L(\Theta)$ is the deviance of the model Θ, p is the number of predictors in the model, and A_n is a variable which controls the degree of penalisation.

Since choice of A_n controls the severity of the penalty associated with the number of parameters p in the model, A_n is used to control the degree of complexity of the model selected by optimising the IC; higher values of A_n will reduce the degree of overfitting and favour relatively more parsimonious models (Dziak et al., 2020). In addition, by adding some bias to the model-selection process, the IC estimates the predictive performance of a model when applied

to novel data (Ding et al., 2018). In contrast to the LRT, ICs can be used to select between non-nested models (Heinze et al., 2018). There exist several ICs (Zhang et al., 2023), motivated by different use-cases and differing in their properties, as described below.

Akaike's information criterion (AIC) is an IC in which the variable A_n is equal to 2 (Akaike, 1974). Use of AIC is particularly appropriate under the assumption that there is no single "true" model which has generated the observed data, but rather there exist several satisfactory candidate models capable of adequate predictions on novel data (Heinze et al., 2018). AIC is therefore not consistent in that it will not reliably select the "true" data-generating model (Zhang et al., 2023), but this is of little relevance when the aim is to create a predictive model. The selected model under AIC is expected to be the one amongst the set of candidate models which minimises prediction-error on novel data; AIC is well-suited for model selection when the aim is to create a predictive model (Dziak et al., 2020). AIC can be shown to be asymptotically equivalent, as the size of the dataset n approaches infinity, to a cross-validated estimate of model predictive performance (Stone, 1977).

The Bayesian information criterion (BIC) employs the penalty term $A_n = \log(n) * p$, such that model likelihood is penalised more heavily with larger datasets. When n is large, ICs which are not a product of n (e.g. AIC) may select models which include weak or spurious associations between variables (Ding et al., 2018). In contrast, as n approaches infinity, BIC is expected to consistently select the "true" data-generating model (that with the largest posterior probability) if it is amongst the pool of candidate models (Heinze et al., 2018; Zhang et al., 2023). In this context, rather than

being the model which flawlessly generates the observed data, a "true" model can be a model which is consistent enough with the data to be practically useful for one's aims (Gelman and Shalizi, 2013). BIC is therefore particularly well-suited to scenarios in which inference, rather than prediction, is the goal (Ding et al., 2018). Compared with AIC, BIC tends to select more parsimonious models (Dziak et al., 2020). Asymptotically, model comparison using BIC is equivalent to comparing models using the Bayes' Factor (likelihood ratio) (Ding et al., 2018), which is discussed further in section 1.3.7.

The widely-applicable information criterion (WAIC) (Watanabe, 2010) is a further example of an IC, which can be used for estimating the generalisation error of Bayesian models. Similar to AIC and BIC, WAIC incorporates a penalty for model complexity, based on the effective number of model parameters (Vehtari et al., 2017). The WAIC penalty is the sum of the posterior variance of the log predictive density for each data point y_i:

$$\sum_{i=1}^{n} Var_{post}\left(\log P(y_i|\Theta)\right)$$

This penalty term evaluates the proportion of the variance of the log posterior predictive density which is attributable to changes in the model parameters; greater model complexity leads to a greater degree of penalisation. WAIC is appropriate for comparing singular models, whereas AIC and BIC are not (Watanabe, 2010). Singularity describes models where multiple values of one or more parameters yield the same model output. Examples of models that are known to be singular include Gaussian mixture models and Hidden Markov models, the latter of which are employed in Chapter 4.

Asymptotically, WAIC is equivalent to leave-one-out CV (Piironen and Vehtari, 2017).

Choice of model-selection method is based on theoretic and practical considerations. If enough diverse data is available, simple splitting of the data may be employed, yielding a "held-out" test dataset (Hastie et al., 2009) against which candidate models can be assessed. However, in many studies the data are limited in size, motivating the use of an IC or CV for model selection. Use of an IC is convenient and computationally inexpensive, and involves only choosing the most appropriate criterion. For example, if the modelling goal is inference or hypothesis-testing, use of BIC to find the "true" model would be an appropriate choice, whereas for predictive models, AIC is more likely than BIC to select a model with good predictive ability. However, these properties of AIC and BIC are based on asymptotic theory for the hypothetical scenario where the size of the dataset n approaches infinity. CV is a more flexible approach; because it directly estimates prediction error using multiple "held-out" testing subsets, it can be used to select and assess any type of model (parametric or non-parametric), whereas ICs are suited to models which are fit by maximising the likelihood. However, one must choose the data-splitting protocol employed within CV (e.g. the number of folds in *k*-fold CV), which is not a trivial task and yet is crucial in controlling bias and variance of both the CV estimates and the selected model. CV can also be computationally-expensive, especially when using large datasets or selecting from a pool of many candidate models.

A variety of information criteria were used in the studies in this thesis. Owing to its properties conducive to accurate predictions,

AIC was employed for selection of predictive models in Chapters 2 and 4, whereas the inferential nature of the models in Chapter 6 prompted the use of BIC for creation of a parsimonious "true" model. WAIC was employed to select the best-fitting Bayesian linear models in Chapter 3. Assorted CV techniques were also employed. Non-nested and nested leave-one-group-out CV were used for selection and evaluation of Elastic Net models in Chapter 3; in Chapter 4 CV was used to select between spline functions in linear models; repeated *k*-fold CV informed model-selection in Chapter 5. A testing dataset was "held-out" for model-evaluation in Chapters 2, 3 and 5.

1.3.7. Bayesian models

Bayesian statistical methods estimate the conditional probability of an event B, according to Bayes' theorem (Bayes, 1763):

$$P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$$

The term $P(B|A)$ is the posterior probability of event B given (conditioned on) the occurrence of event A. The theorem incorporates prior knowledge of the probability of event B as the prior probability, $P(B)$. The likelihood, $P(A|B)$, is derived from available data, and is the probability of event A occurring given that event B also occurs. The denominator $P(A)$ scales the posterior probability such that the posterior probabilities of all conceivable mutually exclusive events sum to 1. Evaluation of the posterior probability using Bayes' theorem can be viewed as a process of updating; prior knowledge of the probability of the event B is updated using new data according to the likelihood.

Bayes' theorem can be applied to estimate the posterior probability of a model Θ (and therefore the hypothesis which it represents) given the data x as follows:

$$P(\Theta|x) = \frac{P(x|\Theta) * P(\Theta)}{P(x)}$$

Here Θ represents all model parameters (e.g. type of model, coefficients, hyperparameters) which characterise the model and determine model outputs. Prior beliefs are updated according to the information contained in the data via the likelihood $P(x|\Theta)$ (Van de Schoot et al., 2014), yielding the posterior probability that the model (hypothesis) is correct (Gerß and Vonthein, 2025). In the presence of data x and two competing hypotheses, represented by models Θ_1 and Θ_2, Bayes' formula can be reparameterised using odds instead of probabilities as follows:

$$\frac{P(\Theta_1|\, x)}{P(\Theta_2|\, x)} = \frac{P(x|\, \Theta_1)}{P(x|\, \Theta_2)} * \frac{P(\Theta_1)}{P(\Theta_2)}$$

The posterior odds are therefore equal to the likelihood ratio, $\frac{P(x|\, \Theta_1)}{P(x|\, \Theta_2)}$, also referred to as the Bayes Factor, multiplied by the prior odds. The Bayes Factor represents the strength of evidence in the data for Θ_1 over Θ_2 (Younas, 2024). This form of Bayes theorem is convenient for repeated Bayesian updating when data is longitudinal, as the posterior odds calculated according to the data at each timepoint can be used as prior odds for updating the posterior odds according to the data at the next time point.

Bayesian analysis differs from classical frequentist methods in a number of practical and conceptual ways. Conceptually, Bayesian probabilities represent the degree of belief in some statement (De

Finetti, 1974), whereas frequentist probabilities (e.g. confidence intervals) represent the expected incidence of an event over a hypothetical infinite number of repetitions of a trial (Van de Schoot et al., 2014; Younas, 2024). Bayesian analysis can provide estimates of the probability that a given hypothesis is true, in contrast to frequentist methods which can only make qualitative statements about the validity of a null hypothesis (Gelman and Shalizi, 2013; Gerß and Vonthein, 2025; Goligher et al., 2024; Puga et al., 2015). Another distinctive feature of Bayesianism is the use of prior probabilities, representing prior beliefs formed from available relevant evidence, for example from results of previous trials (Gerß and Vonthein, 2025) or from expert opinion. The informativeness of a prior can vary with the strength of prior evidence. For example, highly-informative, precise (small variance) priors strongly influence the posterior probability, whereas uninformative (diffuse) priors, important for incorporating a lack of prior knowledge into the model, have less influence relative to the data informing the likelihood function (Van de Schoot et al., 2014). Incorporation into the model of priors derived from previous similar research allows a progressive increase in certainty about a hypothesis (Puga et al., 2015), which is pivotal in the accumulative nature of scientific discovery (Van de Schoot et al., 2014). Priors can also be used to penalise the likelihood (by rendering the data less influential on the posterior probability), thereby influencing the bias-variance trade-off and resulting in regularisation (Gelman and Shalizi, 2013). Whilst in the frequentist paradigm the data are considered random (one sample from a hypothetical infinite number gathered during repeated trials) and model parameters are

unknown but fixed quantities, Bayesian data are fixed (Younas, 2024) and parameters are unknown random quantities with probability distributions (Van de Schoot et al., 2014).

For all but the simplest Bayesian models, analytic calculation of the posterior density is impractical, and therefore methods have been developed for approximating the joint posterior (of any or all model parameters) by drawing multiple samples from it (van Ravenzwaaij et al., 2018). The posterior can be repeatedly sampled using MCMC methods. In MCMC, the joint posterior distribution is iteratively and semi-randomly sampled. At each iteration, the model likelihood corresponding to a certain joint posterior distribution is assessed; the parameter values for each sample are semi-random, guided in part by the likelihood corresponding to the previous sample (van Ravenzwaaij et al., 2018), resulting in convergence on parameter values which maximise the likelihood.

In Chapter 3, Bayes factors are used to assess the strength of evidence for each of a pair of Bayesian models at each timepoint within longitudinal data. The two models represent a pair of hypotheses, that the animal is infected with a disease, or that it is not infected. The probabilistic distributions of the likelihood functions of Bayesian models permit the estimation of the probability of observing any numerical test result given each of the twin hypotheses. The prior odds (the posterior odds from the previous timepoint) are updated at each timepoint using the likelihood ratio (Bayes Factor), yielding fresh posterior odds of infection. The Bayesian models are fit using a variation of MCMC known as the "no U-turn sampler", which converges more efficiently than standard MCMC (Hoffman and Gelman, 2014).

1.3.8. Unsupervised learning

Unsupervised learning is a method for understanding the latent structure of multivariable data where there is no target, y, for a model to predict. Instead, the aim is to estimate the joint distribution of a vector of variables, to ascertain if this joint distribution comprises a mixture of simpler distributions (Hastie et al., 2009). One use of unsupervised learning is "clustering", whereby knowledge of the multivariate mixture distribution allows the allocation of subjects to distinct latent subpopulations, where subjects within each subpopulation ("cluster") are more similar to each other than to subjects in other subpopulations. Such methods are highly useful for discovering hidden patterns in data. In Chapter 4, a hidden Markov model (HMM) (Rabiner, 1989) is employed for clustering of time-series data, allocating cows to latent states according to the results of repeated diagnostic tests. HMMs comprise a first-order Markov chain, each link in the chain representing a consecutive moment in time. At each timepoint, an animal has a probability of residing in each of a number of latent states; these state probabilities are in turn a function of transition and emission probabilities, respectively describing the risks of transitioning from any given state to another, and the distribution associated with each latent state of one or more measured variables. HMMs may be fitted using frequentist or Bayesian techniques; the HMM in Chapter 4 is fitted using maximum likelihood (frequentist) estimation.

In this thesis, statistical learning was applied to data in the context of two important disease of dairy cows, mastitis and paratuberculosis, which are discussed below.

1.4. Bovine mastitis

Mastitis, or inflammation of the mammary gland, which is usually due to bacterial IMI (Bradley, 2002), remains an important disease of dairy farming. Intramammary infections result in subclinical or clinical mastitis. Subclinical mastitis is characterised by the absence of clinical signs and an increased concentration of leukocytes in milk, a measure of which is referred to as the somatic cell count (SCC). Clinical mastitis results in observable signs of inflammation, such as changes in the consistency or colour of milk, and sometimes heat and swelling of the affected quarters, or systemic illness (Panchal et al., 2024).

Mastitis presents a number of significant challenges for the dairy industry due to its ubiquity and its implications for profitability, animal welfare and human health. The incidence of CM in some European countries has been reported in recent years to be between 10 and 32 cases per 100 cow-years (Hanks and Kossaibati, 2025; Kingshay, 2025; Lam et al., 2013; Leach et al., 2024; Rajala-Schultz et al., 2021). Each case of CM has been estimated by meta-analysis to cost on average €195 (Raboisson et al., 2020), suggesting a total expense of between €1950 and €6240 (£1690 and £5406) per 100 cows per year. This is in addition to other costs related to compromised udder health, for example those due to subclinical mastitis. Clinical mastitis is a painful condition (Fitzpatrick et al., 2013) and therefore is to the detriment of animal welfare. Presence of zoonotic pathogens in mastitic milk poses direct risks to human health (Maity and Ambatipudi, 2021).

Reduction of unnecessary AMU in agriculture is a key recommendation for minimising antimicrobial resistance globally (O'Neill, 2016). Treatment of CM is one of the predominant reasons for AMU on dairy farms (Kuipers et al., 2016; Lardé et al., 2021; Leite de Campos et al., 2021), and therefore management of this health condition represents an important opportunity for AMU reduction. To optimise treatment with regards to cure potential, AMU should be limited to only those CM cases which have a good chance of responding to antibiotics (de Jong et al., 2023b). Intramammary infections caused by certain Gram-negative pathogens may be assumed to have a high rate of spontaneous cure (Fuenzalida and Ruegg, 2019a; Lago et al., 2011). Therefore, one strategy to reduce AMU is to withhold antibiotics from culture-negative cases and cases caused by Gram-negative bacteria (de Jong et al., 2023a). Another strategy is to withhold antibiotics from cases with a low spontaneous cure rate but which also scarcely benefit from the use of antimicrobials (Krömker and Leimbach, 2017). If the probability of cure of a case is sufficiently small, it is not considered prudent to use antimicrobials. Such cases tend to be chronic IMIs, characterised by a history of previous CM events or elevated SCC (Schmenger et al., 2020). Alternatives to antimicrobial therapy include symptomatic treatment, for example with anti-inflammatory medication (Krömker et al., 2021), drying off, or culling.

In Chapter 2 a mixed-effects logistic regression model is created which predicts the probability of cure of a case of CM. Predicted cure probabilities are useful when making CM treatment decisions at the individual cow level.

1.5. Paratuberculosis

Paratuberculosis (pTB), caused by *Mycobacterium avium* subsp. *paratuberculosis* (MAP), is a chronic disease, predominantly of ruminants. Granulomatous enteritis leads to chronic diarrhoea, weight loss and death. Paratuberculosis is associated with other adverse outcomes, including reduced milk yield (McAloon et al., 2016), increased culling (Smith et al., 2010), impaired udder health (Pritchard et al., 2017; Rossi et al., 2017), and reduced fertility (Reynolds et al., 2023). In addition to the impact of pTB on animal health, there is concern over the zoonotic potential of MAP (Waddell et al., 2015).

Initial MAP infection usually occurs in young animals (less than one year old) (Windsor and Whittington, 2010), often via the faeco-oral route, although other, less likely routes are considered possible (for example *in utero*, via ingestion of MAP-containing milk or colostrum, and potentially via inhalation) (Barkema et al., 2018). Paratuberculosis is a chronic disease whose progression can be considered with regards to clinical signs, immunological response and potential for transmission. For example, clinically pTB can be divided into four consecutive stages (Tiwari et al., 2006); 1) a "silent" stage, during which no adverse effects are detectable; 2) a subclinical stage, associated with reduced productivity; 3) a clinical stage, with signs such as diarrhoea and weight-loss; and 4) an "advanced" stage, characterised by severe clinical signs and death. Progression of disease is often associated with a shift from predominantly cell-mediated immunity to predominantly humoral immunity, resulting in the presence of circulating MAP-specific immunoglobulins (Begg et al., 2018). As the disease progresses,

infected animals are also increasingly likely to become infectious via shedding of MAP (Espejo et al., 2012; Stabel et al., 2014) or via *in-utero* transmission (Whittington and Windsor, 2009).

A number of national or regional pTB control programmes exist, focussing on limiting transmission of infection via management changes to reduce exposure of susceptible animals to MAP, and removing the source of infection via culling of infected animals (Whittington et al., 2019). Vaccination has been employed in several pTB control programmes (Whittington et al., 2019). However, scarce evidence provides only a suggestion of vaccine efficacy in cattle (Corbiere et al., 2023), and moreover vaccination is not implementable in many regions due to its influence on bovine tuberculosis test results (Coad et al., 2013). Management changes to limit transmission include calving infected cows in a separate environment, cleaning of calving pens after each calving, removal of calves from dams promptly after birth, and feeding colostrum or milk only from cows not suspected of being infected (Nielsen and Toft, 2011). However, attempts at reduction in spread via management are hindered by a combination of practical difficulties in implementing management changes, increased labour requirements and a perceived lack of cost-benefit amongst farmers (Morrison and Rose, 2023). As a result, compliance with recommended management practices can be limited (Nielsen and Toft, 2011). In addition, management and culling decisions are highly reliant on timely detection of infected animals (Garcia and Shalloo, 2015), but poor diagnostic test performance is a fundamental challenge to early detection. Eradication of pTB from herds has therefore proven difficult in all impacted countries

around the world (Weber et al., 2024). The aspiration of the majority of control programmes is reduction of prevalence rather than eradication (Whittington et al., 2019), and only a slow reduction in herd- or animal-level prevalence is reported (for example in Germany, the Netherlands and Italy) (Weber et al., 2024). Optimising the rapid detection of potentially infectious animals could significantly reduce transmission of infection, ultimately improving the efficacy and cost-benefit of pTB control programmes.

Commonly used pTB diagnostic tests include those which detect MAP directly (e.g. culture or polymerase chain reaction testing of faecal samples), or indirectly by quantification of MAP antibodies in serum or milk, utilizing enzyme-linked immunosorbent assays (ELISA) (Whittington et al., 2019). Antibody ELISA testing of milk samples offers a practical and low-cost option for surveillance, since testing can be performed on routinely-collected milk-recording samples, and is therefore used commonly in pTB control programmes (Whittington et al., 2019). Numerical ELISA test results (sample-to-positive (S/P) ratios) are usually dichotomised into “positive” or “negative” categories according to a predefined cutpoint. Within-herd prevalence of pTB infection is often low, typically less than 15% (Whittington et al., 2019). Consequently, to avoid excessive false positive results, MAP ELISA tests are characterised by a diagnostic cut-point favouring specificity over sensitivity (Nielsen et al., 2002). Sensitivity for detection of infectious animals is therefore generally low, and has been estimated at between 21 and 61% (Nielsen and Toft, 2008). A further difficulty in identifying potentially infectious animals arises from the nature of the immune response to MAP infection. There is

a considerable delay between MAP infection and the initiation of a humoral immune response, which produces detectable circulating MAP-specific immunoglobulins (Begg et al., 2018). A field study found that the average age of animals at their first positive antibody ELISA result was approximately four years (Espejo et al., 2012).

Given the delayed onset of immunoglobulin production, and the modest sensitivity of the milk ELISA, cows can be tested repeatedly throughout their productive lifetimes, to identify potentially infectious animals as soon as possible. This strategy is used for example in the United Kingdom and Denmark, where all lactating cows in participating herds are tested four times per year using a milk antibody ELISA (Nielsen, 2009; Nielsen and Toft, 2011). However, difficulty lies in interpretation of patterns of serial test results when classifying animals with regards to risk of being infected or infectious. In the two most commonly-employed frameworks for assigning risk-categories to cows according to serial ELISA results (Table 3.1), arbitrary classification rules are derived empirically (Nielsen, 2009). As an example, within the J-Status classification, animals which receive two positive (S/P ≥ 30%) results within four consecutive ELISA tests are allocated to the highest risk category ("J5") (Taylor et al., 2024); the focus on the most recent four tests is an arbitrary choice, and moreover the period of time over which those four tests must occur is poorly defined. Furthermore, in order to improve the positive predictive value of high-risk categories, animals are usually required to have two or more positive test results to be assigned the highest risk level, reducing the likelihood of infectious animals being classified as high-risk. To the author's knowledge, the J-statuses have not

been validated with regards to the probability of an animal being infected or contagious.

Despite an intensive testing strategy, the prompt identification of infectious animals is not assured. Whilst there is some correlation between milk ELISA test results and faecal shedding of MAP (Beaver et al., 2017), the temporal relationship between the first positive antibody ELISA result and the onset of faecal MAP shedding is inconsistent. Some animals have a positive ELISA result prior to the onset of faecal shedding, whilst others do not (Nielsen and Toft, 2006), and ELISA results in animals shedding MAP in faeces fluctuate over time (Laurin et al., 2017). Failure to identify some cows which are infectious could be in part due to the dichotomisation of ELISA results into positive and negative categories. An S/P ratio less than a given threshold, especially a threshold which favours specificity over sensitivity, does not imply zero probability of disease (Collins, 2002; Fischer, 2021). In addition, MAP milk ELISA S/P ratios are influenced by non-disease factors, such as milk yield and somatic cell count (McAloon et al., 2020). Such factors can influence whether an S/P ratio falls above or below any given threshold. In summary, the dichotomous interpretation of test results using the diagnostic ELISA thresholds, and interpretation of serial dichotomous results, has the potential to inappropriately assign infected cows to low-risk categories, and is too crude to enable the timely identification of cows at high risk of being infected or infectious that is required for successful pTB control.

Two different modelling approaches to objective interpretation of serial pTB test results were taken in Chapters 3 and 4. In Chapter 3, the likelihood functions of two predictive models were compared with respect to MAP ELISA results, yielding Bayes factors which were used for repeated Bayesian updating of the posterior probability of MAP infection. In Chapter 4, a hidden Markov model was used to assign animals to latent classes based on repeated ELISA results.

1.6. Thesis aims

1.6.1. Overall aims

The aim of this thesis was to create predictive models that can be used by farmers or their advisors in real-time when making management decisions. The focus of predictive modelling was on two prevalent and impactful diseases of dairy cows: clinical mastitis and paratuberculosis. Both are diseases that are often the focus of intricate control strategies on farm, and which provide a rich source of data for use in statistical learning. As such, these two diseases are fertile ground for applying predictive models for improved on-farm decision making.

Various types of statistical model were explored, including supervised (mixed-effects models, penalised regression, Bayesian models) and unsupervised (hidden Markov models) methods. Model selection and evaluation methods were considered in depth and applied as appropriate according to the aims of each model. A key feature of the classification models in this thesis is their probabilistic output. Predicted probabilities are a useful quantity, not least because farms differ with regards to their management

priorities and attitude to risk. The provision of continuous, probabilistic model outputs, as opposed to discrete, categorical predictions (e.g. cure vs failure to cure) allows farmers or their advisors to make nuanced management decisions in the context of the individual farm's needs, and fully accounting for the predicted probability of an event of interest. Furthermore, all models were created with consideration of their intended practical use; owing to their portability, parametric models were exclusively used in order to simplify implementation of models in software applications which can be used by farmers or their advisors.

1.6.2. Chapter 2

The aim of Chapter 2 was to create and evaluate a model for prediction of the probability of cure of a case of clinical mastitis. A generalised linear mixed model was trained on a dataset containing information regarding cow-and herd-level factors which were thought to influence cure.

1.6.3. Chapter 3

The aim of Chapter 3 was to create a method for objective interpretation of serial pTB ELISA results. Sequential Bayesian updating was used to calculate the posterior probability of pTB-infection in individual animals. Three models were created; one penalised regression model for prediction of the prior probability of infection using information available in early life, before an animal has had its first MAP ELISA result; and a pair of Bayesian regression models predicting the probability density of an ELISA test result (S/P ratio), given that the cow is infected or non-infected with pTB. The

Bayesian models provided the likelihood ratio (Bayes Factor) for each iteration of the Bayesian updating.

1.6.4. Chapter 4

A second method of objective interpretation of serial pTB ELISA results was developed in Chapter 4. Two types of model were created. Firstly, linear models were constructed that predict a pTB ELISA result for an individual animal, given data describing non-disease factors (e.g. age and milk yield) which were thought to be associated with the S/P ratio. Following adjustment of results for non-disease factors, adjusted S/P ratios were used to train a hidden Markov model for prediction of the probability that an animal resides in each of a number of ordinal latent states which are hypothesised to represent disease progression.

1.6.5. Chapter 5

The aim was to create and evaluate a model for predicting the probability of onset of an established pTB high-risk state ("J5" status) in individual cows within the subsequent 12 months. Generalised linear mixed models were trained on data that included animals' pTB-testing history alongside other factors associated with disease progression.

1.6.6. Chapter 6

Inferential models (generalised linear mixed models, and discrete-time logistic survival models) were created to investigate the association of the latent HMM states with cow fertility and milk-production. The aim was to determine if progress through the ordinal states is associated with negative effects on production.

Chapter 2. Predicting the probability of cure in bovine clinical mastitis

In this chapter, a generalised linear mixed-effects model is created and evaluated. The model predicts the probability of cure of a case of clinical mastitis using information at the cow- and farm-level.

2.1. Introduction

The resolution of a case of CM is dependent on cow, herd, pathogen and treatment factors (Degen et al., 2015). Several factors at the cow- and herd-level have been found to be associated with the probability of cure of IMIs. At the cow-level, chronicity of infection plays an important role; previous cases of CM (in this lactation or previous lactations), or a history of elevated somatic cell counts, are associated with poorer cure rates (Samson et al., 2016; Svennesen et al., 2023; Wilm et al., 2023). In addition to case chronicity, factors associated with likelihood of cure include age (or parity) (Fuenzalida and Ruegg, 2019a; Krömker et al., 2021; Schmenger and Krömker, 2020; Wilm et al., 2023), stage of lactation (McDougall, 2003; McDougall et al., 2007), the number of quarters affected (McDougall et al., 2007), severity of clinical signs (Fuenzalida and Ruegg, 2019b; McDougall et al., 2007; Wenz et al., 2005), season (Bradley and Green, 2009) and the herd new IMI rate (Wilm et al., 2023).

An accurate estimate of the probability of cure of a CM case would facilitate rational decision making with regards to treatment. Knowledge of cure probability could be used for identifying cases which are unlikely to benefit from antimicrobial treatment. Such

knowledge might have additional uses. For example, it has been suggested that predicted cure probability could be used for selecting cases which would benefit from extended courses of antibiotics (McDougall et al., 2019). Analysis of predicted cure rates aggregated at the herd level might also lead to inference regarding within-herd mastitis epidemiology.

Different methods exist for determining if a CM case has cured. Bacteriological cure is considered the gold standard, is often defined as the absence in one or more post-treatment quarter milk samples of a causal pathogen detected in a pre-treatment sample (Ruegg, 2021), and is useful in a research setting. However, arguably a more practically useful definition is cytological cure, which is a return to low SCC during a defined period of time subsequent to CM. Cytological cure is dependent on bacteriological cure (Schmenger and Krömker, 2020) as persistence of infection should result in ongoing elevated SCC, and post-treatment SCCs are significantly negatively associated with bacteriological cure (Deluyker et al., 1999; Fuenzalida and Ruegg, 2019b; Pinzón-Sánchez and Ruegg, 2011; Vasquez et al., 2016). Moreover, achieving cytological cure is important for producers, who must minimise the SCC of saleable milk (Wilm et al., 2023).

Whilst factors associated with cure rates of CM have been investigated in previous research, to the author's knowledge no validated predictive models for estimating CM cure probability have so far been described. The aim of this study was therefore to create and validate a predictive model for the cure probability of clinical mastitis that would be of practical value in commercial dairy herds.

2.2. Materials and Methods

2.2.1. Data collation

A generalised linear mixed model (GLMM) was created to predict the probability of cure of a case of CM. Data for model-development were available from a previous study (Green et al., 2007), providing accurate records of CM cases and milk-recording results on 52 farms over a period of approximately 2 years between 2003 and 2005. Initially the mastitis case data included cow and farm identification, date of each mastitis event and information on which quarters were affected. Milk-recording data included calving dates, parity and the results of regular milk-testing (SCC, yield and milk quality). The data were merged and collated in commercial dairy data-analysis software (TotalVet version 3, Quality Milk Management Services LTD., Wells, U.K.), facilitating the calculation of several cow- and herd-level independent variables possibly associated with the probability of cure of a case of CM (Table 2.1). Univariable plots of the association of each predictor with the dependent variable were examined to assess the need for variable transformations. Some variables were categorised to permit testing of a non-linear association with the probability of cure and to accommodate missing values. For example, variables describing SCC measured prior to the mastitis event had a missing category in case a cow had not had any previous milk-recordings. The binary dependent variable indicated if cure occurred following a mastitis event in an individual cow.

Variable level	Variable	Description	Type	Stratum
Cow	Days-in-milk (DIM)	Day of lactation at mastitis event	Ordinal	0-99
				100-199
				200-299
				300+
	Parity	Parity at mastitis event	Ordinal	1
				2
				3
				4
				5+
	SCCM1	Somatic cell count (x 1000 cells/ml) at the most recent milk-recording, in the same parity, prior to the mastitis event	Ordinal	1-49
				50-99
				100-149
				150-199
				200-499
				500+
				Missing
	SCCM2	Somatic cell count (x 1000 cells/ml) at the penultimate milk-recording, in the same parity, prior to the mastitis event	Ordinal	1-49
				50-99
				100-149
				150-199
				200-499
				500+
				Missing
	SCCM3	Somatic cell count (x 1000 cells/ml) at the antepenultimate milk-recording, in the same parity, prior to the mastitis event	Ordinal	1-49
				50-99
				100-149
				150-199
				200-499
				500+
				Missing
	Repeat case	Variable indicating if the mastitis event is the first (index) case, or a repeat case, during the current lactation.	Boolean	No
				Yes
	Days since last case	The interval (days) since the previous mastitis event for this cow	Ordinal	0-99
				100-199
				200-299
				300+
				Never
	Number of cases previous parity	The number of mastitis events in the previous parity	Ordinal	0
				1
				2
				3+
	Number of quarters	Number of quarters affected at the mastitis event	Ordinal	1
				2
				3
				4
	Quarter position	Anatomical position of the affected quarter(s)	Categorical	Front
				Back
				Both
	Percentage SCC ≥ 100 previous parity	Proportion of milk recordings during the previous parity with SCC ≥ 100,000 cells/ml	Numerical	
	Percentage SCC ≥ 200 previous parity	Proportion of milk recordings during the previous parity with SCC ≥ 100,000 cells/ml	Numerical	
	Percentage SCC ≥ 500 previous parity	Proportion of milk recordings during the previous parity with SCC ≥ 100,000 cells/ml	Numerical	
	Season	Season in which mastitis event occurred	Categorical	Winter
				Spring
				Summer
				Autumn

Table continues...

Variable level	Variable	Description	Type	Stratum
Herd	Lactation new infection rate	Mean over the three most recent milk recordings of the proportion of cows in the herd with SCC <200,000 cells/ml at the previous recording which had SCC ≥ 200,000 cells/ml at the current recording	Numerical	
	Dry period cure rate	Mean over the three most recent milk recordings of the proportion of cows which had their first recording of lactation and whose SCC was ≥ 200,000 cells/ml at the last recording of the previous lactation, whose SCC was <200,000 cells/ml at the first recording of the current lactation	Numerical	
	Proportion chronically infected	Mean over the three most recent milk recordings of the proportion of cows with 2 SCCs ≥ 200,000 cells/ml out of their three most recent SCCs	Numerical	
	Proportion SCC >= 200	Mean over the three most recent milk recordings of the proportion of cows with SCC ≥ 200,000 cells/ml	Numerical	

Table 2.1: Predictor variables offered to a generalised linear mixed model for predicting the probability of cure of clinical mastitis.

A case was defined as having cured if one of the following was true:

1) At each of the three milk recordings after a case of CM within the same lactation, the composite SCC was below 200,000 cells/ml *and* the cow had no subsequent cases of CM during the period up to and including the date of the third subsequent milk recording.

2) At each of the two milk recordings after a case of CM within the same lactation, the composite SCC was below 100,000 cells/ml *and* the cow had no subsequent cases of CM during the period up to and including the date of the second subsequent milk recording.

Additionally, the first SCC after a case of CM was disregarded if it occurred within 14 days subsequent to the mastitis event, and if it was above the relevant threshold (100,000 or 200,000 cells/ml).

Cases failing to meet the above cure criteria were classified as non-cures, unless insufficient subsequent milk-recordings had occurred to establish if cure had occurred (e.g. in late lactation); such cases were removed from the data as described later.

2.2.2. Data splitting

Data were randomly split into two subsets, containing approximately 80% and 20% of the data, for model training and evaluation respectively. To prevent data leakage, data from individual farms appeared in either subset, but not both.

2.2.3. Data cleaning

Rows were removed from the data if insufficient milk-recording had occurred to define the value of some of the variables derived from somatic cell counts; the dependent variable, percentage of milk recordings in the previous parity with an elevated SCC (in multiparous animals), and the herd-level variables. For primiparous animals, the percentage of milk recordings in the previous parity with an elevated SCC was set to zero. In addition, some rows were missing information on which quarters were affected; these rows were also removed.

2.2.4. Statistical methods

Pairwise correlations between numerical predictor variables were calculated. One variable from each highly correlated (correlation coefficient ≥ 0.8) pair was selected for fitting multivariable models, by fitting univariable models and choosing the variable which resulted in the best model likelihood. Initially a model containing all candidate predictors was fit. Backwards stepwise variable

selection was conducted using AIC as a guide. Variables were retained if confounding was detected (≥ 10% change in the coefficient of a covariate significantly associated with the dependent variable). Quadratic transformations of numerical variables, and plausible two-way interactions, were offered to the model and retained if they improved AIC. For the final model, an optimal threshold (cutpoint) of predicted probability for making binary classification decisions (cure or not cure) was calculated as the average threshold which maximised Youden's index across 1000 bootstrap samples of the model-training data. Model calibration and classification accuracy of the final model were then assessed using the hold-out testing data set as follows. Probabilistic predictions (using fixed effects only) were made on the testing data, and were visualised with a histogram. A calibration plot (reliability diagram) (Fenlon et al., 2018) was created by dividing the predicted probabilities into ten bins, each bin containing an equal number of predictions, and plotting the proportion of cases which cured (the dependent variable) against the midpoint of each bin. The optimal threshold was used to classify predictions (cure or no cure), and classification was assessed by calculating sensitivity, specificity and balanced accuracy. To aid in making inferences from the final model, P-values were calculated using Satterthwaite's approximation of the degrees of freedom of a mixed model. Variables were considered statistically significant if the P-value was less than 0.05.

2.3. Results

Data-cleaning resulted in removal of 51.5% (6210/12068) of rows in the data. 30.7% (3699/12068) were removed due to there being

insufficient SCC information to determine whether cure occurred, 19.3% (2325/12068) due to missing information regarding which quarter was affected, 0.7% (81/12068) due to missing herd-level variables and 0.9% (105/12068) due to absence of previous lactation SCCs. Descriptions of the model-training and -testing datasets are given in Table 2.2 and Table 2.3. The herd-level variables representing the percentage of the herd chronically infected and the percentage of the herd with somatic cell count ≥ 200,000 cells/ml were highly correlated (Pearson's correlation coefficient 0.94). Since inclusion of the former variable in univariable models resulted in a favourable likelihood, the latter variable was not offered to any multivariable models. Parameters of the final model are shown in Table 2.4. The receiver operating characteristic curve for the model applied to the testing dataset is shown in Figure 2.1. The area under the ROC curve for the final model applied to the testing data was 0.71. The optimal predicted probability threshold for classification was 0.37 and when applied to the testing data resulted in sensitivity, specificity and balanced accuracy of 0.70, 0.62 and 0.66 respectively. A histogram of predicted probabilities made by the final model on the testing data is shown in Figure 2.2. The calibration plot is shown in Figure 2.3.

Variable	Stratum	Model Training data		Model testing data	
		n	%	n	%
Farm		36		13	
Cow		2598		868	
Mastitis events		4460		1398	
Parity	1	682	15.3%	181	12.9%
	2	864	19.4%	245	17.5%
	3	823	18.5%	271	19.4%
	4	660	14.8%	274	19.6%
	5+	1431	32.1%	427	30.5%
Days-in-milk	0 to 99	2496	56.0%	821	58.7%
	100 to 199	1321	29.6%	434	31.0%
	200 to 299	482	10.8%	112	8.0%
	300+	161	3.6%	31	2.2%
SCCM3 (x 1000 cells/ml)	1 to 49	723	16.2%	219	15.7%
	50 to 99	225	5.0%	71	5.1%
	100 to 149	156	3.5%	37	2.6%
	150 to 199	333	7.5%	86	6.2%
	200 to 499	437	9.8%	118	8.4%
	500+	358	8.0%	107	7.7%
	Missing	2228	50.0%	760	54.4%
SCCM2 (x 1000 cells/ml)	1 to 49	809	18.1%	253	18.1%
	50 to 99	277	6.2%	75	5.4%
	100 to 149	179	4.0%	61	4.4%
	150 to 199	491	11.0%	123	8.8%
	200 to 499	522	11.7%	162	11.6%
	500+	547	12.3%	159	11.4%
	Missing	1635	36.7%	565	40.4%
SCCM1 (x 1000 cells/ml)	1 to 49	818	18.3%	247	17.7%
	50 to 99	343	7.7%	95	6.8%
	100 to 149	245	5.5%	54	3.9%
	150 to 199	569	12.8%	167	11.9%
	200 to 499	523	11.7%	145	10.4%
	500+	1011	22.7%	342	24.5%
	Missing	951	21.3%	348	24.9%
Days since last case	0 to 99	1542	34.6%	446	31.9%
	100 to 199	415	9.3%	109	7.8%
	200 to 299	284	6.4%	85	6.1%
	300+	849	19.0%	266	19.0%
	Never	1370	30.7%	492	35.2%
Repeat case	No	2629	58.9%	875	62.6%
	Yes	1831	41.1%	523	37.4%
Number of cases previous parity	0	2602	58.3%	862	61.7%
	1	894	20.0%	307	22.0%
	2	461	10.3%	132	9.4%
	3+	503	11.3%	97	6.9%
Number of quarters	1	4174	93.6%	1275	91.2%
	2	211	4.7%	98	7.0%
	3	19	0.4%	7	0.5%
	4	56	1.3%	18	1.3%
Quarter position	Front	1879	42.1%	628	44.9%
	Back	2384	53.5%	694	49.6%
	Both	197	4.4%	76	5.4%
Season	Winter	1473	33.0%	490	35.1%
	Spring	953	21.4%	234	16.7%
	Summer	956	21.4%	236	16.9%
	Autumn	1078	24.2%	438	31.3%
Cured	No	2813	63.1%	877	62.7%
	Yes	1647	36.9%	521	37.3%

Table 2.2: Descriptive statistics of the categorical variables in the model-training and -testing datasets used for creation and evaluation of a generalised linear mixed model for predicting the probability of cure of a case of clinical mastitis.

Variable	Model training data						Model testing data					
	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Date of mastitis	04/03/2003	20/11/2003	17/03/2004	17/04/2004	26/10/2004	30/04/2005	08/03/2003	05/11/2003	16/03/2004	21/04/2004	07/11/2004	11/04/2005
Proportion SCC ≥ 100 previous parity[1]	0.00	0.09	0.38	0.41	0.70	1.00	0.00	0.10	0.38	0.42	0.73	1.00
Proportion SCC ≥ 200 previous parity[1]	0.00	0.00	0.11	0.24	0.40	1.00	0.00	0.00	0.14	0.24	0.40	1.00
Proportion SCC ≥ 500 previous parity[1]	0.00	0.00	0.00	0.09	0.10	1.00	0.00	0.00	0.00	0.09	0.11	1.00
Lactation new infection rate (%)	0.0	7.5	9.6	10.1	12.2	27.9	0.0	6.2	9.4	9.7	13.0	23.1
Dry period cure rate (%)	0.0	47.6	60.0	59.0	72.2	100.0	0.0	42.9	56.3	53.7	66.7	100.0
Percentage of herd chronically infected (%)	2.2	12.2	16.1	16.1	20.0	36.3	4.7	10.0	15.0	16.1	18.5	45.0
Percentage of herd SCC ≥ 200 (%)[1]	4.8	19.4	23.5	23.9	28.0	45.4	7.4	16.6	24.1	24.2	28.1	57.3
SCC1[2]	2	66	163.5	456.7	428	9999	4	67	173.5	624.2	548	9999
SCC2[2]	4	55	125	371.3	334	9999	6	53	126	453.2	381	8785
SCC3[2]	3	57	137	369.0	355	9999	6	58	133	424.1	342	9504
SCC4[2]	6	63	141	352.6	341	9340	6	63	146	435.4	372	8680

Table 2.3: Descriptive statistics of the numerical variables in the model-training and -testing datasets used for creation and evaluation of a generalised linear mixed model for predicting the probability of cure of a case of clinical mastitis.

[1]SCC (x 1000 cells/ml). e.g. 49 = 49,000 cells/ml.

[2]SCCs (x 1000 cells/ml) subsequent to the case of clinical mastitis. For example, SCC1 is the count at the first subsequent recording, SCC2 is at the second subsequent recording, and so on.

Variable	Estimate	Standard error	Odds ratio	Lower 95% confidence Interval	Upper 95% confidence interval	P-value
Intercept	0.111	0.265	1.117	0.664	1.879	0.677
Parity 1	Referent					
Parity 2	0.064	0.141	1.067	0.809	1.406	0.648
Parity 3	-0.007	0.150	0.993	0.740	1.332	0.961
Parity 4	0.122	0.160	1.130	0.825	1.547	0.446
Parity 5+	-0.195	0.150	0.823	0.613	1.104	0.194
SCCM1 1-49*	Referent					
SCCM1 50-99*	**-0.300**	**0.126**	**0.741**	**0.579**	**0.949**	**0.017**
SCCM1 100-149*	**-0.570**	**0.155**	**0.566**	**0.417**	**0.766**	**2.37E-04**
SCCM1 150-199*	**-0.704**	**0.178**	**0.495**	**0.349**	**0.701**	**7.77E-05**
SCCM1 200-499*	**-0.997**	**0.147**	**0.369**	**0.276**	**0.492**	**1.18E-11**
SCCM1 500+*	**-1.035**	**0.128**	**0.355**	**0.276**	**0.456**	**5.38E-16**
SCCM1 Missing	-0.016	0.134	0.984	0.757	1.280	0.907
SCCM2 1-49*	Referent					
SCCM2 50-99*	-0.083	0.132	0.921	0.710	1.193	0.533
SCCM2 100-149*	-0.337	0.176	0.714	0.506	1.008	0.056
SCCM2 150-199*	**-0.672**	**0.219**	**0.511**	**0.332**	**0.785**	**0.002**
SCCM2 200-499*	**-0.715**	**0.170**	**0.489**	**0.351**	**0.682**	**0.000**
SCCM2 500+*	**-0.448**	**0.170**	**0.639**	**0.458**	**0.891**	**0.008**
SCCM2 Missing	-0.093	0.153	0.911	0.676	1.229	0.543
SCCM3 1-49*	Referent					
SCCM3 50-99*	0.149	0.142	1.161	0.878	1.534	0.295
SCCM3 100-149*	-0.146	0.191	0.864	0.594	1.257	0.445
SCCM3 150-199*	-0.376	0.234	0.687	0.434	1.086	0.108
SCCM3 200-499*	-0.284	0.189	0.753	0.520	1.090	0.133
SCCM3 500+*	**-0.590**	**0.197**	**0.555**	**0.377**	**0.815**	**0.003**
SCCM3 Missing	-0.211	0.134	0.809	0.623	1.053	0.115
Days since previous mastitis 8-99	Referent					
Days since previous mastitis 100-199	-0.049	0.139	0.953	0.726	1.251	0.727
Days since previous mastitis 200-299	0.144	0.159	1.154	0.846	1.575	0.365
Days since previous mastitis 300+	0.001	0.112	1.001	0.805	1.247	0.989
Days since previous mastitis Never	0.186	0.108	1.204	0.974	1.488	0.086
Number cases previous parity 0	Referent					
Number cases previous parity 1	-0.079	0.109	0.924	0.746	1.144	0.469
Number cases previous parity 2	-0.203	0.140	0.816	0.620	1.074	0.148
Number cases previous parity 3+	**-0.603**	**0.149**	**0.547**	**0.408**	**0.733**	**5.32E-05**
Quarter position both	Referent					
Quarter position front	0.244	0.178	1.276	0.900	1.808	0.170
Quarter position back	0.045	0.176	1.046	0.741	1.477	0.797
Percentage SCC >= 100 previous parity*	**-0.012**	**0.002**	**0.988**	**0.985**	**0.991**	**6.28E-14**
Lactation new infection rate (%)	-0.021	0.016	0.979	0.948	1.010	0.180
Dry period cure rate (%)	0.002	0.002	1.002	0.998	1.005	0.413
Percentage of herd SCC >= 200 (%)*	0.004	0.009	1.004	0.986	1.021	0.698
Proportion SCC >= 100 previous parity*	0.242	0.156	1.274	0.938	1.729	0.121

Table 2.4: Parameters of the fixed effects of a generalised linear mixed model for predicting the probability of cure of a case of clinical mastitis.

Somatic cell counts measured at the most recent, penultimate and antepenultimate milk recordings are represented respectively by the variables SCCM1, SCCM2 and SCCM3.Statistically significant associations between covariates and the dependent variable are highlighted in bold.

**SCC (x 1000 cells/ml). e.g. 49 = 49,000 cells/ml.*

All levels of the SCCM1 variable were significantly associated with the probability of cure; increasing SCCM1 was negatively associated with the odds of cure. For example, compared to SCCM1 of less than 50,000 cells/ml, SCCM1 of at least 500,000 cells/ml had an odds ratio of 0.36 (95% C.I. 0.28 – 0.46). SCCM2 and SCCM3 were also significantly associated with reduced odds of cure, but only if SCCM2 or SCCM3 was greater than 150,000 or 500,000 cells/ml respectively. Variables describing udder health during the previous lactation were significantly associated with cure probability; the percentage of milk recordings in the previous lactation with a cell count of at least 100,000 cells/ml was negatively associated with odds of cure; the odds ratio associated with a one percent increase was 0.988 (95% C.I. 0.985 – 0.991), and mastitis cases were also less likely to cure if the cow had three or more cases of mastitis in the previous lactation (O.R. 0.547, 95% C.I. 0.408 – 0.733).

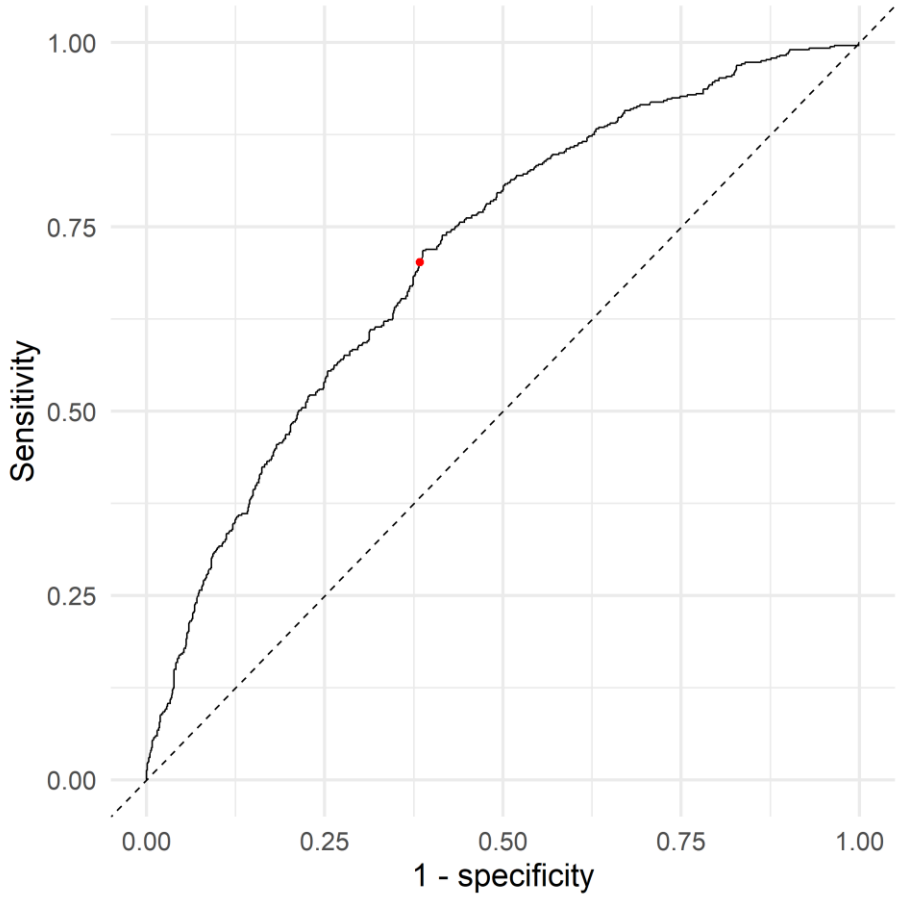

Figure 2.1: Receiver operating characteristic curve of a generalised linear mixed model for predicting mastitis cure probability.
The model was applied to a hold-out dataset (13 farms) for model validation. Area under the curve = 0.712. The red dot shows the model performance at the threshold which maximised Youden's index.

2.4. Discussion

The aim of this study was to create and validate a model for predicting the probability of cure of clinical mastitis cases. This would be of benefit to farmers in making rational treatment decisions, but to the author's knowledge such a validated model has not previously been reported. The probabilistic predictions from the final model were well-calibrated when assessed using held-out testing data Figure 2.3, indicating that the predicted

probabilities faithfully represent the chance of a CM case being cured. At the time of data collection, all CM cases on the subject farms were treated with antimicrobials. Therefore, the predictions can be interpreted as the probability of cure given antimicrobial treatment. The model achieves the primary aim of this study, which was to produce a model which can provide predicted probabilities of cure for use in making rational treatment decisions. These predictions would be most useful if provided to farmers in real time at the point when a treatment decision is required, for example during milking. This highlights the benefit of deploying GLMMs, which are highly portable, as predictive models; values of predictor variables can be easily combined with regression coefficients in a simple calculation on, for example, a mobile device. Therefore, farmers could be provided with instantaneous predictions based on simple information available at the time.

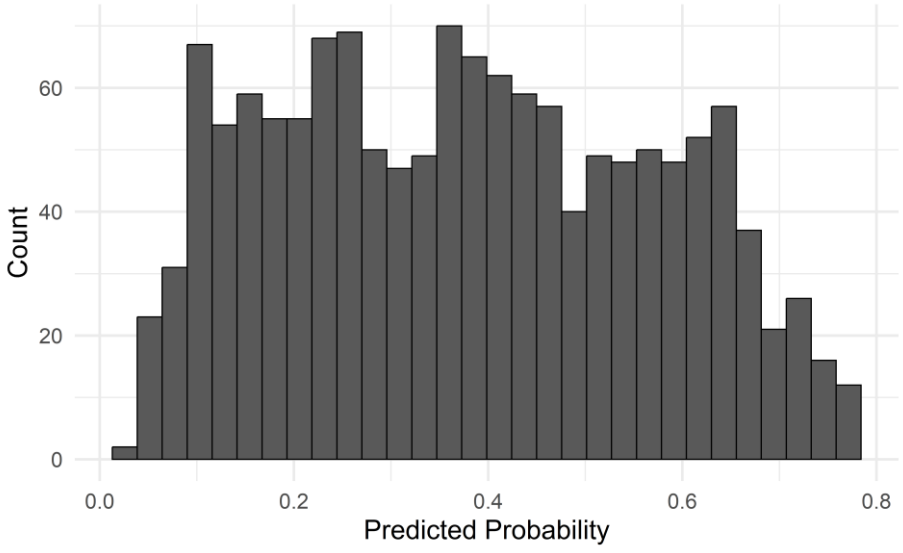

Figure 2.2: Histogram of probabilistic predictions made using a model forecasting the probability of cure of a case of clinical mastitis.

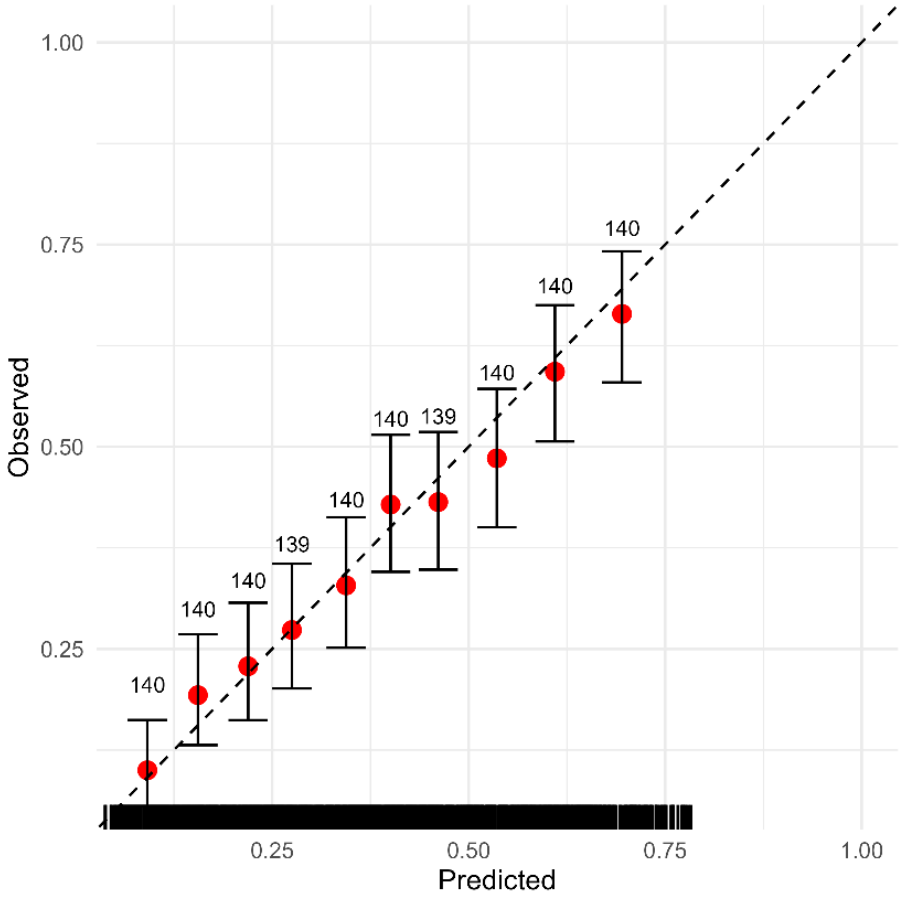

Figure 2.3: Calibration plot of a generalised linear mixed model for predicting the probability of cure of a case of clinical mastitis.
Predictions were divided into ten bins, each containing an equal number of predictions. For each bin, the proportion of cases that were observed to have cured was plotted against the midpoint of the bin (red points). 95% confidence intervals are shown as vertical error bars. The number of predictions within each bin is shown above the error bars. The rug plot above the x-axis illustrates the density of predictions across the range of probability.

Assessment of the ability of the model to correctly classify cases showed that binary cure predictions (cure or no-cure) were less reliable. The area under the ROC curve (0.71) was modest, as were the sensitivity (0.70) and specificity (0.62) even following tuning of the probability threshold used for classification. It can be seen from Figure 2.2 that many of the probabilistic model predictions, despite

being well-calibrated, were near the threshold (0.37) used for classification (cure or non-cure) and were therefore somewhat uncertain. These near-equivocal predicted probabilities were associated with modest classification accuracy for this model.

Practically for farmers, a model which produces well-calibrated probabilities, despite modest classification accuracy, is highly useful, because it facilitates nuanced treatment decisions which account for factors not accounted for in the model. For example, when making treatment decisions, farmers must balance the desire for AMU-reduction with the need for maintaining IMI cure-rates. In herds where observed cure-rates are low, farmers may wish to maximise cure rates by selecting a relatively low probability threshold for determining which cows to treat with antimicrobials, whereas in herds with high cure-rates a higher threshold could be employed resulting in fewer antimicrobial treatments. As another example, a farmer may wish to select a probability threshold according to the perceived or actual value of a cow; it could be desirable to treat high-yielding cows or cows of high genetic-merit despite a modest predicted cure probability; cows with comorbidities (for example lameness or infertility) might be treated only when the probability of cure is very high.

A similar recent study (Cruz et al., 2025) created a GLMM for predicting CM cure using a small dataset describing 22 cases on one farm. Classification accuracy was assessed by internal validation (resubstitution) and found to be good (area under the ROC curve was 0.89), but no external validation was attempted. Model calibration was not reported.

In the current study, predictor variables in the model significantly associated with the odds of cure were those describing recent SCC (SCCM1, SCCM2, SCCM3), the number of CM cases occurring in the previous parity, and the percentage of milk recordings during the previous parity with SCC of at least 100,000 cells/ml. This is consistent with previous research which has found associations between CM cure and chronicity of infection, as measured by previous elevated SCC or CM cases (Bradley and Green, 2009; Pinzón-Sánchez and Ruegg, 2011; Swinkels et al., 2013; Wilm et al., 2023).

It is noteworthy that this model provides well-calibrated probabilities of cure without the need to identify the causal pathogen. This is inconsistent with previous research which has found that cure rates vary with pathogen (Fuenzalida and Ruegg, 2019b; McDougall et al., 2007; Svennesen et al., 2023; Vasquez et al., 2016). The results of the current study indicate that knowledge of the causal pathogen is unnecessary for predicting cure, suggesting that pathogen has less influence on cure rates than is often assumed (de Jong et al., 2023b). However, it is also probable that some of the variables in the current model are themselves associated with pathogen type. For example, duration of elevated SCC both prior to and subsequent to CM varies with the species of bacterial pathogen; *Staphylococcus aureus* is associated with a period of elevated SCC which is relatively long and which often does not resolve during the same lactation, whereas *E. coli* CM is associated more with a rapid increase and subsequent rapid decrease in SCC (de Haas et al., 2004). Therefore, model features

such as previous SCCs which capture IMI chronicity are likely to reflect pathogen behaviour.

It is interesting to compare this model-based approach to the use of on-farm bacteriological culture (OFC) in CM treatment decisions. The OFC approach aims to withhold antimicrobial therapy from CM cases which are culture-negative or which are caused by Gram negative bacteria (de Jong et al., 2023b), based on the premise that Gram-negative IMIs, particularly *E. coli*, have a high rate of spontaneous cure. However, in one study 20% of all cases of *E. coli* CM were due to persistent infections (Bradley and Green, 2001), demonstrating that some *E. coli* IMIs do not spontaneously cure. In another study, the bacteriological cure rate of *E. coli* IMI in the absence of antimicrobial therapy was as low as 71.4% (Pyörälä and Pyörälä, 1998). Further, CM caused by some other common Gram-negative pathogens, for example *Klebsiella* spp., have been shown to have low spontaneous cure rates (Roberson et al., 2004) and to benefit from antimicrobials (Fuenzalida and Ruegg, 2019a). In summary it is uncertain that Gram negative IMIs do not benefit from antimicrobial treatment. Some of this uncertainty could be due to the influence of different strains of pathogen, exhibiting different virulence factors and tendencies to cause persistent infections (Bradley and Green, 2001). Use of the current model, whose features include measures of infection-persistence, rather than the name of the invading pathogen, would seem appropriate in light of this uncertainty. When making CM treatment decisions, the current model could therefore be used alongside, or as an alternative to, the OFC approach. It is also worth noting that, in contrast to the OFC approach which typically requires 24-hours to obtain results, and

necessitates the use of chargeable consumables, model-based cure predictions can be instantaneous and relatively cost-effective.

There are limitations to this study which are worth considering. Firstly, the data were a convenience sample, collected for a previous study (Green et al., 2007). Model performance was assessed using a held-out portion of the data, with the data partitioned with regard to farm, such that animals in any given herd could be present in either the model-training or -testing subsets, but not in both. This strategy avoids data leakage and is consistent with accurate estimates of the performance of a model when applied to data from novel groups (farms). Nevertheless, validation of the model using further data from novel farms would still be useful for establishing generalisability. Second, mastitis is not a static disease; aspects of its epidemiology and common treatment regimens can change over time (Leach et al., 2024; Ruegg, 2017). The data were collected during the period 2003 to 2005. To further establish the generalisability of this model, validation against data collected more recently would be informative.

Future research in this field could feature holistic predictive models which use a combination of cow, herd, pathogen and treatment factors. Such models would allow farmers to make informed CM treatment decisions based on, for example, the results of on-farm bacteriological culture as well as cow- and herd-factors, and accounting for various potential treatment options.

2.5. Conclusions

This study has provided a well-calibrated model for predicting the probability of cure of a case of clinical mastitis given cow- and herd-

variables which are readily available to those making real-time treatment decisions.

Chapter 3. Interpretation of serial paratuberculosis test results using repeated Bayesian updating

In this chapter and subsequent chapters, models are applied to data pertaining to paratuberculosis. In this chapter, penalised regression is used to make predictions regarding the probability that a calf was infected with MAP in early life. These predictions are used in turn as prior information in a process of repeated Bayesian updating applied at each consecutive pTB test result in an individual animal, yielding the posterior probability of disease.

3.1. Introduction

Paratuberculosis control schemes usually aim to reduce prevalence either through culling of cows thought to be infected and potentially contagious, through management changes intended to reduce disease transmission, or through a combination of both. Management changes can be targeted predominantly at only those cows thought to be contagious (Nielsen, 2009). In the United Kingdom, identification of infected and potentially contagious dairy cows is attempted by repeated testing of individual cow milk samples (MAP antibody ELISA). Cows are classified regarding risk of infection, according to one of two schemes for serial test result interpretation (Table 3.1). An aim of regular repeated testing is to identify contagious cows promptly (Nielsen, 2009). However, cows may shed MAP in faeces prior to receiving any positive test results (Laurin et al., 2017; Nielsen and Toft, 2006; Sweeney et al., 2006), potentially transmitting infection to others. This is consistent with

the modest sensitivity of the milk MAP antibody ELISA for detection of contagious cows (Nielsen and Toft, 2008) using the standard test interpretation; that is, classifying numerical ELISA results (S/P ratios) as positive or negative according to their value relative to a threshold recommended by the test manufacturer. Whilst prompt identification of all contagious animals would seem essential for minimising disease transmission, current methods are unlikely to do so.

In animals eventually classified as high-risk (following one or more positive ELISA results), it is possible that information in the form of S/P ratios of negative tests occurring earlier in life are predictive of later positivity. If so, such information could be used for predicting the probability that a cow will later be classified as infected (as defined by positive test results). Given repeated testing, the predicted probability could be updated whenever new information, in the form of a new test result, becomes available. Cows with a relatively high predicted probability could be managed accordingly to limit any potential contagion, thereby improving disease control.

3.1.1. The basis of repeated Bayesian updating

One method for updating a probability in light of new information is to apply Bayes' theorem, in which prior knowledge is combined with new evidence to provide a posterior probability. In the current context, the prior knowledge comprises that which is known about the probability of infection prior to a test result, and the new evidence is the test result.

J-Status	Definition
J0	The animal has never been classified as J5 status and all of the four most recent tests (within the last two years) gave a negative result. For animals with fewer than four tests in the last two years, all tests have returned a negative result.
J1	The animal has had only one test within the previous two years, which gave a negative result.
J2	The animal's antepenultimate test was positive, but there have been no other positive results amongst any of the most recent four tests occurring within the last two years. The animal has never been classified as J5 status.
J3	The animal's penultimate test was positive, but there have been no other positive results amongst any of the most recent four tests occurring within the last two years. The animal has never been classified as J5 status.
J4	The animal's latest result was positive, but there have been no positive results amongst any of the previous three tests occurring with the last two years. The animal has never been classified as J5 status.
J5	Two out of four consecutive tests gave a positive result at some point during the animal's life. Once an animal has been classified as J5 status, it can never revert to a different status and its status will be J5 for life.
QMMS Status	**Definition**
Currently Negative	The animal has never been classified as Positive status and the three most recent tests gave a negative result. For animals with fewer than three tests so far over their lifetime, all tests have returned a negative result. Note that the phrase currently negative is used, as the status of these animals may change according to future test results.
Uncertain	Either the most recent test was negative and the animal has had one positive result within the previous two tests, or the animal has had at least one uncertain[1] test result within the three most recent tests.
Provisionally Positive	The animal tested positive at the most recent test date, but the two previous tests gave an uncertain[1] or negative result. For animals with only one previous test result, that previous test result was uncertain or negative. For animals with no previous test results (e.g. freshly-calved heifers), the first test of their lifetime is positive.
Positive	Two out of three consecutive tests gave a positive result at some point during the animal's life. Once an animal has been classified as Positive status, it can never revert to a different status and its status will be Positive for life.

Table 3.1: Two similar disease-risk classification systems (J-status and QMMS status) used as part of routine disease monitoring, according to the milk Mycobacterium avium subsp. paratuberculosis ELISA test result history for individual animals.
[1]Uncertain (20% ≤ S/P < 30%) results are informative in the QMMS Status, but are not relevant to the J-status.

The posterior probability of infection, given a positive test result, is defined as:

$$P(D^+ \mid T^+) = \frac{P(T^+|D^+) \,.\, P(D^+)}{P(T^+)}$$

where $P(T^+|D^+)$ is the likelihood of a positive test result given that the animal has the disease (equal to test sensitivity), $P(D^+)$ is the prior probability of infection, and $P(T^+)$ is the probability of observing a positive test result in the population (derived from prevalence, test sensitivity and test specificity). Further equations can be derived which provide posterior probabilities in situations where the disease is either present or absent, or where the test result is positive or negative:

$$P(D^- \mid T^-) = \frac{P(T^-|D^-) \,.\, P(D^-)}{P(T^-)}$$

$$P(D^+ \mid T^-) = \frac{P(T^-|D^+) \,.\, P(D^+)}{P(T^-)}$$

$$P(D^- \mid T^+) = \frac{P(T^+|D^-) \,.\, P(D^-)}{P(T^+)}$$

Since $\frac{P(D^+)}{P(D^-)}$ is the odds of the animal being diseased, the posterior odds of disease given a positive diagnostic test result can be defined as:

$$\frac{P(D^+ \mid T^+)}{P(D^- \mid T^+)} = \frac{\left[\frac{P(T^+|D^+).\,P(D^+)}{P(T^+)}\right]}{\left[\frac{P(T^+|D^-).\,P(D^-)}{P(T^+)}\right]}$$

The equivalent for a negative test result is:

$$\frac{P(D^+ \mid T^-)}{P(D^- \mid T^-)} = \frac{\left[\frac{P(T^-|D^+).P(D^+)}{P(T^-)}\right]}{\left[\frac{P(T^-|D^-).P(D^-)}{P(T^-)}\right]}$$

The latter two equations can be simplified as:

$$\frac{P(D^+ \mid T^+)}{P(D^- \mid T^+)} = \frac{P(T^+|D^+)\,.P(D^+)}{P(T^+|D^-)\,.P(D^-)}$$

and

$$\frac{P(D^+ \mid T^-)}{P(D^- \mid T^-)} = \frac{P(T^-|D^+)\,.P(D^+)}{P(T^-|D^-)\,.P(D^-)}$$

where $\frac{P(T|D^+)}{P(T|D^-)}$ is the likelihood ratio (also referred to as the Bayes factor), and $\frac{P(D^+)}{P(D^-)}$ is the prior odds of disease.

Therefore, the posterior odds of disease are a product of the likelihood ratio (Bayes Factor) of the test result (positive or negative) and the prior odds, and deriving the post-test odds of disease in this way is referred to as Bayesian updating. With repeated tests, repeated Bayesian updating can be employed, at each test using the posterior odds calculated from the previous test result as the prior odds at the most recent test (Meyer et al., 2018). Prior information at the first test of an animal's life (for example the population disease prevalence) can be incorporated if available.

The likelihood ratio in the above equations is applicable when the diagnostic test returns a dichotomous (positive or negative) result. Repeated Bayesian updating has been applied in this way previously to interpret serial dichotomous MAP ELISA results (Meyer

et al., 2018). However, numerical test results (S/P ratios) can be used in place of dichotomous results (Fischer, 2021). It is feasible to train some predictive models (henceforth referred to as Bayesian Models) to forecast the probability of any numerical test result (S/P ratio) given disease status (*i.e.* $P(T|D^+)$ and $P(T|D^-)$), and to use predictions from these models to calculate the likelihood ratio, $\frac{P(T|D^+)}{P(T|D^-)}$, given any S/P ratio. Such models could also incorporate influential covariates and confounders, for example associations between non-disease factors (yield, SCC, age, stage of lactation, milk protein or butterfat concentration) and S/P ratio (McAloon et al., 2020). Prior information at the first test of an animal's life (at which point posterior odds from a previous test are not available) could be derived from models (henceforth referred to as Prior Models) utilising a variety of information (e.g. age, disease history of the dam and other close relatives) in addition to population disease prevalence.

3.1.2. Study aims

The aim of this study was to use repeated Bayesian updating (i.e. repeated applications of Bayes' theorem) to interpret serial milk MAP ELISA test results (S/P ratios), and to provide a predicted probability of a cow being classified as Johne's positive, according to a commonly-used Johne's disease risk classification system (Table 3.1). Predicted probabilities could serve as an early signal that a cow will become positive, facilitating reduction of disease transmission and improving control of pTB.

3.2. Materials and methods

A methodology was developed which provided continually updated probabilities of pTB infection in dairy cows repeatedly tested using a milk MAP antibody ELISA. The probability in an animal prior to its first ever test was provided by the Prior Model. This probability was then updated following each test result, utilizing predictions from the Bayesian Models. This process is illustrated in Figure 3.1. Finally, the practical utility of the methodology was assessed.

Data were available for herds milk-recording with Quality Milk Management Services (QMMS) LTD. For inclusion, herds were required to have conducted a minimum of 20 whole-herd pTB (individual cow milk MAP ELISA) tests over the ten-year period between 1st January 2011 and 31st December 2020, and to have carried out milk-recording on average at least four times per year during that same period. The MAP ELISA tests were carried out using a commercially available kit (IDEXX B.V., Hoofddorp, The Netherlands).

ELISA test results (optical densities) were converted to S/P ratios, and expressed in the data as percentages, as follows:

$$S/P\ Ratio\ = 100 * \frac{OD_{sample} - OD_{negative\ control}}{OD_{positive\ control} - OD_{negative\ control}}$$

where OD = optical density.

A whole-herd test was defined as an occasion on which at least 70% of animals in a herd had been tested.

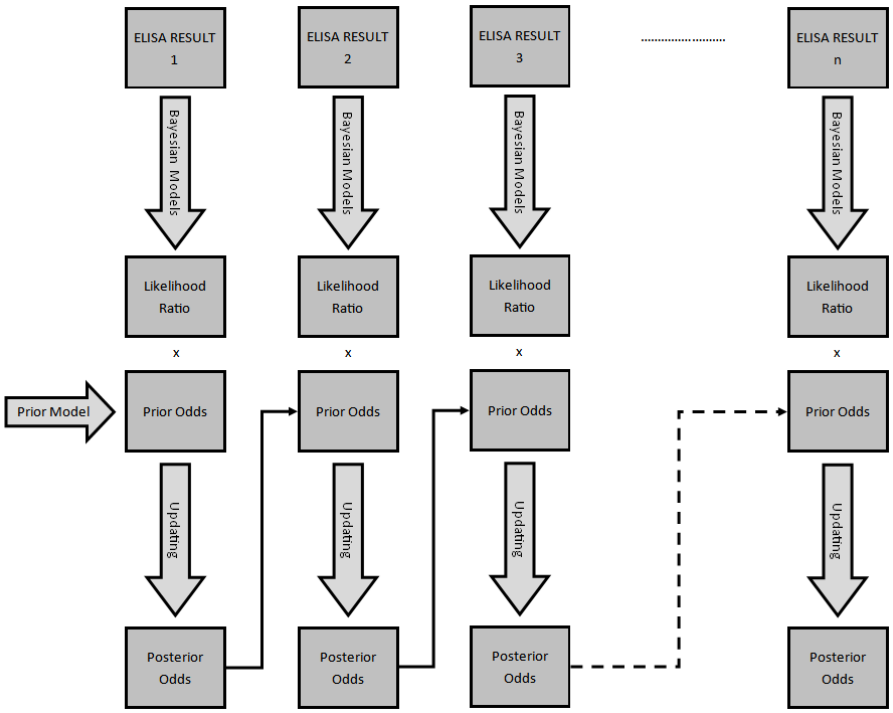

Figure 3.1: Repeated application of Bayesian updating for predicting the probability of Johne's disease infection in cows repeatedly tested (n times) using a milk antibody ELISA.

Posterior odds at each step are a product of the prior odds and the likelihood ratio. Likelihood ratios are provided by Bayesian Models predicting the probabilities of ELISA results given that an animal is infected or not infected. Prior odds at the first test are provided by a Prior Model incorporating herd prevalence and individual animal information. At subsequent tests, the prior odds are equal to the posterior odds derived from the previous test.

3.2.1. Prior Model

Paratuberculosis test results for all herds during the period between 1st December 2010 and 31st December 2021 were extracted from commercial dairy data analysis software (TotalVet Version 3, QMMS LTD., UK), and were merged with milk-recording data (yield, SCC, fat and protein concentration, days-in-milk (DIM), parity and age of cows on the date on which they were pTB-tested). To ensure that the herd-level variables (e.g. measures of pTB prevalence) in the data were relevant, animals were excluded from the data if they had not

been born in the herd (e.g. purchased cows), or if their birth date was unknown.

Independent variables fell into three categories:

1. Those variables reflecting direct transmission of MAP from dam to calf (one-to-one transmission) during the first few days of life;
2. Those measuring exposure of a newborn calf to MAP in its environment, for example through exposure to MAP-containing faeces from other dams in the same environment (one-to-many transmission);
3. Indicators of herd prevalence, representing the average risk of infection for animals within a herd.

The predictor variables are described in Table 3.2.

Johne's disease status was the binary dependent variable, indicating if the animal was classified as positive or negative for Johne's disease, similar to (Meyer et al., 2018), according to the following definitions:

- Johne's Positive (JP): An animal had at least three milk antibody ELISA tests during its lifetime, and two positive (S/P ≥ 30%) test results within three consecutive tests.
- Johne's' Negative (JN): An animal had at least nine tests over its lifetime, of which at least the last eight were negative (S/P < 30%), and had never met the definition of a JP animal.

The definition of a JP animal was the same as that used in the QMMS classification (Table 3.1). The requirement for numerous negative test results in the JN definition was intended to account for the

modest sensitivity (Nielsen and Toft, 2008) of the milk MAP antibody ELISA test; multiple repeated negative test results are required for confidence that an animal is not affected by pTB. The disease status of animals which did not meet the JP or JN definitions (for example animals with too few test results) was unknown.

An Elastic Net model was created for prediction of the probability of an animal being JP without utilising any ELISA test results. The data used to train the Elastic Net included only those animals defined as JP or JN (not unknown).

Model training was performed using the *caret* package in R version 4.4.2 (R Core Team, 2022). Leave-one-group-out cross validation (LOGOCV), was used to select the best subset of predictors to offer to the Elastic Net model, and to select optimal hyperparameter values. However, model development and evaluation should not be performed within the same CV procedure (section 1.3.6), and so expected model predictive performance was subsequently estimated using nested LOGOCV. Tuning of hyperparameters and selection of the best predictor subset was achieved using non-nested LOGOCV. All numerical variables were z-standardised by subtracting the mean of that variable in the data and dividing by the standard deviation of the variable in the data. The data were split, yielding a model-testing dataset containing one farm (the "leave-out" group), and a model-training dataset containing the remaining farms. Next, variable importance was calculated as follows. Multiple Elastic Net models were fitted to the model-training dataset, varying the hyperparameter values (α from 0 to 1, in increments of 0.025; λ from 0 to 2, in increments of 0.05).

For each model, probabilistic predictions were made on the model-testing dataset, and the weighted log loss of the predictions was calculated as follows:

$$Weighted\ log\ loss = \frac{1}{n}\sum_{1}^{n} w[o * \ln(p) + (1 - o) * \ln(1 - p)]$$

where o = 1 if the dependent variable is positive and o = 0 if the dependent variable is negative, p is the predicted probability, and w is the weight, given by the prevalence, or complement of the prevalence as appropriate, of the positive dependent variable.

For the model resulting in the best weighted log loss, variable importance (the absolute value of the coefficient for each predictor) was calculated. Variable importance was then used to select different-size subsets of the most important predictor variables. Next, multiple models were again fit to the model-training data, this time varying the hyperparameter values in the same fashion as previously, whilst also offering to the models either all predictors or a subset of predictors. Subsets used were the top 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20, 30 or 40 predictors according to variable importance. Weighted log-loss was calculated for predictions on the model- testing dataset (the “leave-out” farm). Thus, the best predictor subset and hyperparameter values were selected for the “leave-out” farm. The above procedure was conducted repeatedly, each time splitting the data so that a different farm was “held-out” as the model- testing dataset, until all farms had been “held-out” once. The final model (with its hyperparameter values and subset of predictors) was that which optimised the mean weighted log-loss across all data splits.

Category	Variable	Description	Type	Levels
Direct (one-to-one) transmission	Dam parity	Parity of the dam at the time of calving	Categorical	1
				2
				3
				4+
				Missing[1]
	Dam status at birth / at 12 months old / current	The QMMS Johne's status of the dam at the time of the animal's birth / when the animal was 12 months old / at her latest test	Categorical	Missing[1]
				Currently negative
				Uncertain
				Provisionally positive
				Positive
	Granddam current status	The QMMS Johne's status of the granddam at her latest test	Categorical	Currently negative
				Uncertain
				Provisionally positive
				Positive
	Proportion of siblings positive	The proportion of an animal's siblings whose QMMS Johne's status was positive at the latest test	Categorical	(0 - 0.333][2]
				(0.333 - 0.667]
				(0.667 - 1]
				Unknown[1]
	Proportion of siblings provisionally positive	The proportion of an animal's siblings whose QMMS Johne's status was provisionally positive at the latest test	Categorical	(0 - 0.333][2]
				(0.333 - 0.667]
				(0.667 - 1]
				Unknown[1]
Environmental (one-to-many) transmission	Proportion of vaguely proximal dams positive at birth / at 12 months old / current	The proportion of cows calving within 70 days either side of the animal's date of birth, whose QMMS Johne's status was positive at the time of birth / when the animal was 12 months old / at her latest test	Numeric	
	Proportion of vaguely proximal dams provisionally positive at birth / at 12 months old / current	The proportion of cows calving within 70 days either side of the animal's date of birth, whose QMMS Johne's status was provisionally positive at the time of birth / when the animal was 12 months old / at her latest test	Numeric	
	Proportion of proximal calves positive	The proportion of dairy calves born within 10 days either side of an animal's data of birth, whose QMMS Johne's status was positive at the latest test	Categorical	[0 – 0.333][2]
				(0.333 – 0.667]
				(0.667 – 1]
				Unknown[1]
	Proportion of proximal calves provisionally positive	The proportion of dairy calves born within 10 days either side of an animal's data of birth, whose QMMS Johne's status was provisionally positive at the latest test	Categorical	[0 – 0.333][2]
				(0.333 – 0.667]
				(0.667 – 1]
				Unknown[1]
	Proportion of vaguely proximal calves positive	The proportion of dairy calves born within 70 days either side of an animal's data of birth, whose QMMS Johne's status was positive at the latest test	Categorical	[0 – 0.25][2]
				(0.25 – 0.5]
				(0.5 – 1]
				Unknown[1]
	Proportion of vaguely proximal calves provisionally positive	The proportion of dairy calves born within 70 days either side of an animal's data of birth, whose QMMS Johne's status was provisionally positive at the latest test	Categorical	[0 – 0.167][2]
				(0.167 – 0.333]
				(0.333 – 1]
				Unknown[1]

Table continues...

Category	Variable	Description	Type	Levels
Herd prevalence	Proportion of herd currently positive	The mean of the proportions of the herd testing positive at the last herd test prior to the animal's birth and the next test after the animal's birth	Categorical	[0 - 0.0193][2] (0.0315 – 0.0527] (0.0527 – 0.183] Missing[1]
	Mean titre	The average of the mean titre at the last herd test prior to the animal's birth and the next test after the animal's birth	Categorical	[0.82 - 4.73][2] (4.73 – 6.05] (6.05 – 7.9] (7.9 – 27.6] Missing[1]
	Mean titre negative cows	The average of the mean titres of all titres below S/P = 10 at the last herd test prior to the animal's birth and the next test after the animal's birth	Categorical	[0.82 - 2.5][2] (2.5 - 2.89] (2.89 - 3.31] (3.31 - 5.4] Missing[1]

Table 3.2: Predictor variables offered to the model for prediction of the probability of an animal being classified as positive according to a commonly-used Johne's disease risk classification, at any point during its lifetime.
[1]Variables with missing values were categorised to allow a "missing" category. Variables describing proportions of relatives (siblings or horizontal relatives) or proximal were categorised to allow an "unknown" category in cases where no such relatives or calves existed.
[2]Square brackets indicate that the category is open at that end, i.e. that the value nearest the square bracket is included in this category.

The expected performance of a model fit using the above model selection procedure (selection of the number of predictors and the hyperparameter values, through optimisation of the weighted log-loss) was assessed using nested LOGOCV (Figure 1.2). The nested LOGOCV procedure comprised two loops (an inner loop nested within an outer loop). In the outer loop, the entire data were iteratively split, leaving out one farm at each iteration, each time yielding an outer training dataset (all but one farm) and an outer testing dataset (one farm). For each iteration of the outer loop (i.e. for each "held-out" farm), the inner loop procedure (the model selection procedure described above) was then performed using only the outer training dataset. The expected performance of the model selected by that procedure within each inner loop was

assessed using the outer testing dataset (the outer loop). This yielded an optimal model for each iteration ("held-out" farm) of the outer loop, together with an estimate of performance of that model. The tuned hyperparameters and predictor subset of the optimal model for each outer loop iteration were recorded.

Performance (calibration and discrimination) of each iteration's optimal model was assessed within the outer loop as follows. For calibration, probabilistic predictions were made on the outer testing dataset (the "held-out" farm), from which weighted log loss was calculated and calibration curves were plotted. For discrimination, an optimal predicted probability decision boundary was calculated first by making probabilistic predictions with the optimal model on its own training data (the outer trainset). Using the observed class labels (dependent variable) in the outer trainset, sensitivity and specificity were calculated at a variety of candidate decision boundaries. The optimal decision boundary was chosen as that which maximised the sum of sensitivity and specificity (Kuhn and Johnson, 2013), with the aim of optimising balanced accuracy in the context of an imbalanced dependent variable. Class predictions were then made on the outer testing dataset (the "held-out" farm) using the selected decision boundary; predictions were compared with observed class labels across all holdout folds combined, to estimate sensitivity, specificity and balanced accuracy. Distributions of the characteristics (hyperparameters, number of predictors and measures of discrimination) of all of the models selected during the outer loop were examined for insights into the expected stability of a model created using the model selection procedure employed.

3.2.2. Bayesian Models

Two models were created, which predicted the probability of a particular S/P ratio, given that an animal is either JP or JN respectively. Predicted probabilities were used for calculation of the likelihood ratio required for updating posterior odds (Figure 3.1).

Some data cleaning was required to remove implausible values of milk-recording data prior to further modelling. Rows of data were removed if DIM was greater than 730, or if yield, milk quality or SCC were missing. Data were limited to rows in which protein content was between 1% and 6%, and butterfat content was between 1.5% and 9%, similar to ranges recommended by the International Committee of Animal Recording (ICAR, 2014). Rows representing primiparous animals were removed if age was less than 20 months or greater than 60 months, and multiparous animals' rows were removed if age was less than 25 months or greater than 300 months. Days-in-milk was categorised due to the anticipated non-linear association between this feature and S/P ratio (the dependent variable in the Bayesian Models). Cows were excluded from modelling if their disease status was unknown (neither JP or JN).

The distribution of S/P ratios in the data, after accounting for repeated measures within cows, was visualised as follows, and was used to select the most appropriate distribution of the dependent variable (S/P ratio) in the models. For each cow in the data, one row of data was sampled at random. The distributions of S/P ratios in these combined samples, grouped by disease status (JP or JN), were visualised using density plots.

The data were then grouped by farm and split into training and testing sets, holding out approximately 20% of data in the testing set (to prevent data leakage, farms could appear in either the training or testing sets, but not in both). The training and testing datasets were further split according to disease status (JP or JN). This resulted in a total of four datasets in total (JP and JN training datasets and JP and JN testing datasets).

Bayesian generalised linear mixed models were created using a MCMC algorithm in the *brms* package in R version 4.2 (R Core Team, 2022). The models, whose dependent variable was the S/P ratio, were used to predict the probability of a test result occurring, given the disease status of the animal (the likelihood). One model was trained on JP training data, whilst the other was trained using JN training data. Default priors were selected by the *brms* package (weakly-informative student-t priors for the intercept and for the shape parameter of the lognormal distribution, non-informative flat uniform priors for covariates). Models were defined as:

$$\boldsymbol{y_i} \sim \boldsymbol{lognormal}(\boldsymbol{\mu_i}, \boldsymbol{\sigma})$$
$$\boldsymbol{\mu_i} = \boldsymbol{\alpha} + \boldsymbol{\beta x_i}$$
$$\boldsymbol{\alpha} \sim \boldsymbol{student\_t}\ (\boldsymbol{\nu}, \boldsymbol{\mu}, \boldsymbol{\sigma})$$
$$\boldsymbol{\beta} \sim \boldsymbol{Uniform}\ (-\infty, \infty)$$
$$\boldsymbol{\sigma} \sim \boldsymbol{student\_t}\ (\boldsymbol{\nu}, \boldsymbol{\mu}, \boldsymbol{\sigma})$$

Each model was fit using two parallel MCMC chains of 2000 iterations each. The first 500 iterations of each chain were discarded as burn-in. Convergence of models was confirmed if the Gelman-Rubin convergence diagnostic, *Rhat*, was no greater than 1.01. Traceplots were also examined to ensure that the MCMC algorithm had fully explored each parameter space.

A fixed value of 1 was added to the dependent variable (S/P ratio) to eliminate any zero values, permitting its representation as a lognormal distribution. The predictor variables offered to the models are described in Table 3.3. One predictor variable, the mean titre of negative cows (MTNC), was designed to measure variability between ELISA test kits, as well as inter- and intra-operator variability, which could bias ELISA results, by capturing variation in the average S/P ratio of uninfected animals.

Multiple models were created including every possible combination of the predictor variables, achieving an exhaustive search for the best predictor subset. Models were compared and selected using WAIC (Watanabe, 2010). The models included a random intercept for farm. Models including a random effect of cow failed to converge. Instead, to account for repeated measures within cows, one test result was sampled at random from each cow prior to model training.

Goodness-of-fit of the two models to the testing data was assessed by randomly drawing 100 samples from the posterior predictive distribution; these draws were compared visually (density plots) with the distribution of observed S/P ratios in the JP and JN testing datasets.

The likelihood of each test result was then derived using the Bayesian models as follows. For each row in the entire dataset, including those rows where disease status was unknown, for each of the two final models 1,000,000 draws were taken from the posterior predictive distribution. This resulted in a vector of predicted S/P ratios for each row of data given each of two

hypothetical scenarios (whether the animal was JP or JN). The likelihood ratio, $\frac{P(T|D^+)}{P(T|D^-)}$, for the observed S/P ratio was then calculated by dividing the number of elements within ± 3% of the observed S/P ratio in the JP vector, by that in the JN vector.

Posterior odds were then calculated as follows. For each pTB test in the data, including cows not meeting the JP or JN criteria, the posterior odds of disease were the product of prior odds and likelihood. For the first test of an animal's life, the prior odds $\frac{P(D^+)}{P(D^-)}$ were given by the predicted probability $P(D^+)$ from the Prior Model:

$$\frac{P(D^+)}{P(D^-)} = \frac{P(D^+)}{1 - P(D^+)}$$

Variable	Description	Type	Levels
Age	The animal's age in months[1]	Numeric	
Yield	Milk yield (litres)[1]	Numeric	
Log somatic cell count	Natural logarithm of composite somatic cell count[1]	Numeric	
Protein (%)	Composite % protein[1]	Numeric	
Butterfat (%)	Composite % butterfat[1]	Numeric	
Mean titre negative cows	Mean of all S/P ratios < 10% in the herd on the same day	Numeric	
Days in milk[3]	Stage of lactation	Categorical	[0 – 5][2] (5 – 10] (10 – 50] (50 – 100] (100 – 200] (200 – 300] > 300

Table 3.3: Independent variables offered to a model for prediction of S/P ratio of the animal, on the date on which the milk sample was taken.
[1]Measured on the day on which the milk sample was taken.
[2]Square brackets indicate that the category is open at that end, i.e. that the value nearest the square bracket is included in this category.
[3]Days in milk categories were included in the model as dummy variables.

For subsequent tests, the prior odds were equal to the posterior odds at the previous test.

To evaluate this methodology (application of Prior and Bayesian Models to observed data), likelihood ratios and posterior probabilities (converted from posterior odds) were visualised using histograms. To establish if the methodology could provide a useful early indication that an animal was likely to become positive, posterior probabilities of JP animals, prior to the onset of positivity (as indicated by repeated positive ELISA results) were compared with those of matched JN animals. Onset of positivity for a JP animal was the point at which it received the first of a pair of positive (S/P ≥ 30%) test results. Tests in JP animals were matched with tests in JN animals by farm and age (within a 3-month margin) at onset of positivity. JP animals were excluded from this analysis if pairing with a JN animal of similar age was not possible. In JP animals, the interval between each test result and the onset of positivity was calculated. In JN animals, the equivalent interval was the time difference between each test and the test conducted at an age most similar to the age of onset of positivity in the animal's JP counterpart. In this way, posterior probabilities of JN animals were temporally aligned with the onset of positivity of their positive counterparts. Data were then limited to those test results occurring prior to age at onset of positivity (or, in JN animals, prior to the test conducted at an age most similar to the age at onset of positivity of the matched JP animal). A boxplot was created, grouped by status (JP or JN) and categorised interval until onset of positivity, to compare posterior probabilities in JP animals, prior to the onset of

positivity, with temporally-aligned probabilities of matched JN animals.

Receiver operating characteristic analysis was used to further assess the predictive ability of the Bayesian updating methodology, and to identify thresholds of posterior probability which could facilitate discrimination between JN animals JP animals prior to onset of positivity. For each animal in the data (excluding cows which were neither JP or JN), one test (with its associated posterior probability), was sampled at random. For JP animals the sampled test occurred prior to the onset of positivity. Initially, analysis was performed using all sampled test results, to establish a threshold which would be predictive of onset of positivity for an animal at any point in the future. The selected threshold was that which minimised the difference between sensitivity and specificity.

Given the low prevalence of JP cows in the data, it was anticipated that positive predictive value (PPV) would be modest. Therefore, further analysis was performed to investigate the impact of the chosen threshold on predictive values, and to select an appropriate threshold which maximised PPV without overly compromising sensitivity. Therefore, an alternative threshold was chosen which maximised the F1-score, defined as follows:

$$\mathrm{F1} = 2 * \frac{PPV * sensitivity}{PPV + sensitivity}$$

where PPV is the positive predictive value. Maximisation of the F1 score selects a threshold which balances sensitivity and PPV. To understand the nature of posterior probabilities at different intervals prior to onset of positivity, optimal thresholds and

measures of discrimination were also calculated using different definitions of a positive outcome; each definition was onset of positivity within a certain number of months in the future (either 24, 18, 12 or 6 months).

Calibration of the posterior probabilities for prediction of onset of positivity in an animal within various timeframes was assessed by inspecting calibration plots, using onset of positivity ever in the future, or within 24, 18, 12 or 6 months, as the outcome. The calibration plots showed the proportion of animals, grouped by categorised posterior probability, which were JP and whose onset of positivity occurred within the chosen time frames.

Fifty animals were randomly sampled from each of three groups (JP, JN and unknown disease status). For each individual animal in the samples, test results and posterior probabilities were plotted alongside influential covariates (yield, milk quality, age, DIM etc.). Examination of these plots provided insight into the results of applying the methodology to serial test results in individual cows.

3.3. Results

3.3.1. Prior model

After inclusion of only animals born on the same farm, removal of 147 animals whose date of birth was unknown, and after excluding 22,314 animals which did not meet the JP or JN criteria, the data consisted of 7,548 animals on 47 farms. Further descriptive statistics are shown in Table 3.4 and Table 3.5.

Variable	Category	Number	Percent
Dam parity	1	1825	24.2
	2	1432	19.0
	3	1045	13.8
	4+	1697	22.5
	Missing	1549	20.5
Dam status at birth	Currently negative	2605	34.5
	Uncertain	123	1.6
	Provisionally Positive	94	1.2
	Positive	28	0.4
	Missing	4698	62.2
Dam status at 12 months old	Currently negative	4548	60.3
	Uncertain	238	3.2
	Provisionally Positive	155	2.1
	Positive	136	1.8
	Missing	2471	32.7
Dam status current	Currently negative	5147	68.2
	Uncertain	371	4.9
	Provisionally Positive	255	3.4
	Positive	455	6.0
	Missing	1320	17.5
Granddam current status	Currently negative	2498	33.1
	Uncertain	150	2.0
	Provisionally Positive	127	1.7
	Positive	152	2.0
	Missing	4621	61.2
Proportion of proximal calves positive	[0 – 0.333]	6831	90.5
	(0.333 – 0.667]	186	2.5
	(0.667 – 1]	68	0.9
	Unknown	463	6.1
Proportion of proximal calves provisionally positive	[0 – 0.333]	6982	92.5
	(0.333 – 0.667]	82	1.1
	(0.667 – 1]	21	0.3
	Unknown	463	6.1
Proportion of vaguely proximal calves positive	[0 – 0.25]	7302	96.7
	(0.25 – 0.5]	228	3.0
	(0.5 – 1]	7	0.1
	Unknown	11	0.1
Proportion of vaguely proximal calves provisionally positive	[0 – 0.167]	7383	97.8
	(0.167 – 0.333]	148	2.0
	(0.333 – 1]	6	0.1
	Unknown	11	0.1
Proportion of herd currently positive	(0 - 0.0193]	944	12.5
	(0.0193 - 0.0315]	687	9.1
	(0.0315- 0.0527]	768	10.2
	(0.0527 - 0.183]	836	11.1
	Missing	4313	57.1
Mean titre	(0.82 - 4.73]	1547	20.5
	(4.73 - 6.05]	814	10.8
	(6.05 - 7.9]	993	13.2
	(7.9 - 27.6]	986	13.1
	Missing	3208	42.5
Mean titre negative cows[1]	(0.82 - 2.5]	1456	19.3
	(2.5 - 2.89]	1084	14.4
	(2.89 - 3.31]	898	11.9
	(3.31 - 5.4]	884	11.7
	Missing	3226	42.7
Dependent Variable	Negative	6142	81.4
	Positive	1406	18.6

Table 3.4: Descriptive statistics of the categorical variables in the data used to train an Elastic Net model for prediction of Johne's status in individual animals.
[1]Mean titre negative cows is the mean of all ELISA results with S/P ratio < 30% on the test day in question.

Variable	Min	Lower Quartile	Median	Mean	Upper Quartile	Maximum
Proportion of proximal dams provisionally positive at 12 months old	0.00	0.00	0.00	0.02	0.02	1.00
Proportion of proximal dams provisionally positive current	0.00	0.00	0.00	0.03	0.05	1.00
Number of vaguely proximal dams	0.00	51.00	87.00	116.37	151.00	576.00
Proportion of vaguely proximal dams positive at birth	0.00	0.00	0.00	0.01	0.01	0.14
Proportion of vaguely proximal dams positive at 12 months old	0.00	0.00	0.01	0.02	0.03	0.24
Proportion of vaguely proximal dams positive current	0.00	0.02	0.05	0.06	0.09	1.00
Proportion of vaguely proximal dams provisionally positive at birth	0.00	0.00	0.00	0.01	0.02	0.21
Proportion of vaguely proximal dams provisionally positive at 12 months old	0.00	0.00	0.01	0.02	0.03	0.22
Proportion of vaguely proximal dams provisionally positive current	0.00	0.01	0.03	0.03	0.05	0.50

Table 3.5: Descriptive statistics of the continuous variables in the data used to train an Elastic Net model for prediction of Johne's status in individual animals.

The hyperparameter values of the final Prior Model were 0 and 0.1 for α and λ respectively. Owing to the α value of 0, the model was in effect a Ridge regression. The predictor variables included in the final Prior Model are shown in Table 3.6. Variable importance of features in the final model is shown in Figure 3.2. The number of vaguely proximal dams (calving within 70 days) was considerably more influential in the model than any other variables.

Characteristics of the optimal models for each outer loop iteration in the nested LOGOCV are shown in Table 3.7. The number of predictors (minimum = 8, maximum = 16) selected was considered

to be moderately stable; 12 predictors were selected in 59.6% (28/47) of folds, and 14 predictors were selected in a further 25.5% (14/47) of folds. Hyperparameter values were also stable. The alpha value (minimum = 0, maximum = 0.025) was 0 in 95.7% (45/47) of folds, and the lambda value (minimum = 0.05, maximum = 0.15) was 0.1 in 87.2% (41/47) of folds.

Feature	Description
Number of vaguely proximal dams	The number of cows calving within the period 70 days either side of the animal's date of birth
Granddam current status	The QMMS Johne's status of the granddam at her latest test
Proportion of herd currently positive	The mean of the proportions of the herd testing positive at the last herd test prior to the animal's birth and the next test after the animal's birth
Proportion of vaguely proximal calves positive	The proportion of dairy calves born within 70 days either side of an animal's data of birth, whose QMMS Johne's status was positive at the latest test
Dam current status	The QMMS Johne's status of the dam at her latest test
Proportion of horizontal relatives positive	The proportion of an animal's siblings, aunts and great-aunts whose QMMS Johne's status was positive at the latest test
Proportion of horizontal relatives provisionally positive	The proportion of an animal's siblings, aunts and great-aunts whose QMMS Johne's status was provisionally positive at the latest test
Proportion of vaguely proximal dams provisionally positive current	The proportion of cows calving within 70 days either side of the animal's date of birth, whose QMMS Johne's status was provisionally positive at her latest test
Mean titre	The average of the mean titre at the last herd test prior to the animal's birth and the next test after the animal's birth

Table 3.6: Predictor variables in the final elastic net Prior model selected by leave-one-group-out cross validation.

The measures of discrimination (sensitivity, specificity and balanced accuracy; Table 3.7) were considered to be relatively less stable across the folds of the nested cross validation. For all of the holdout datasets combined, sensitivity, specificity and balanced accuracy were 0.669, 0.449 and 0.559 respectively. Calibration of the final model was considered to be acceptable as a result of the calibration plot, shown in Figure 3.3.

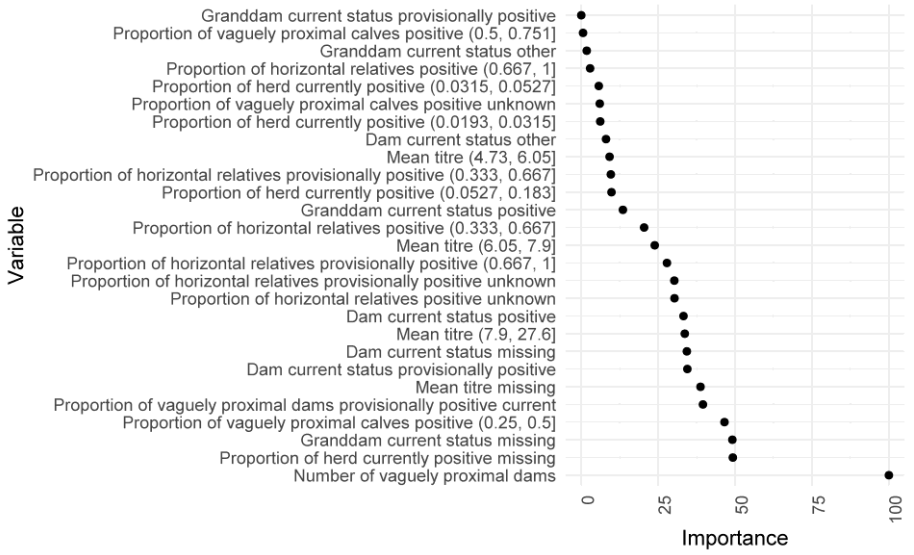

Figure 3.2: Variable importance of predictor variables in the final Elastic Net model for prediction of positive Johne's status at any time point during a cow's life.
Numerical predictors were z-standardised prior to model fitting. Variable importance was calculated as the absolute value of the coefficients. Note that for categorical predictors, variable strata are shown individually.

3.3.2. Model selection

Parameters of the two final Bayesian Models are shown in Table 3.9.

3.3.3. Model goodness of fit

For each of the two Bayesian Models, the samples drawn from the posterior predictive distribution of the testing dataset are shown with the distribution of observed S/P ratios in Figure 3.5. According to these density plots, the fit of both models was adequate.

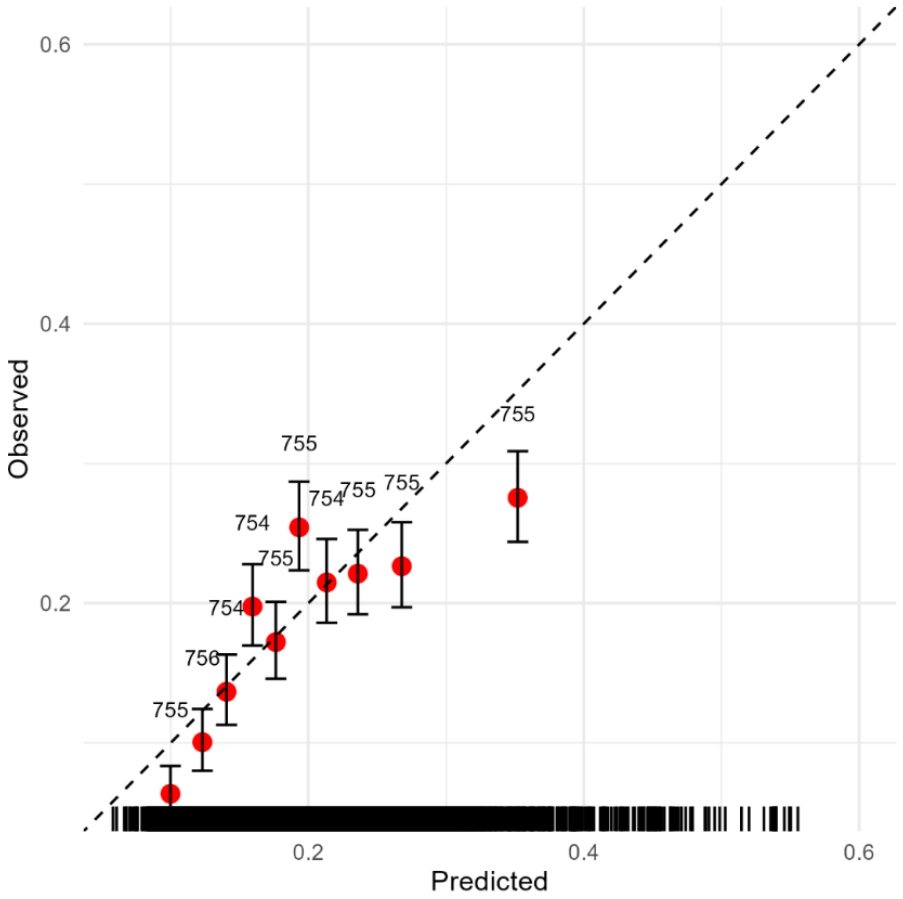

Figure 3.3: Calibration plot of predictions made on hold-out data within a nested cross validation procedure.
Data points were divided according to predicted probabilities into ten bins of equal number of data points. The proportion of data points within each bin which had a positive observed class label is shown on the y-axis. Number of data points within each bin is shown. 95% confidence intervals are shown as error bars. The rug plot on the x-axis shows the density of data points across the range of probability.

Iteration	Number of predictors	Alpha	Lambda	Weighted Log Loss	Optimal Probability Cutpoint	Sensitivity	Specificity	Balanced Accuracy
1	12	0	0.1	-0.046	0.109	0.444	0.782	0.613
2	10	0	0.1	-0.214	0.184	0.585	0.771	0.678
3	14	0	0.1	-0.364	0.178	0.561	0.664	0.612
4	14	0.025	0.05	-0.433	0.160	0.784	0.098	0.441
5	14	0	0.1	-0.334	0.209	0.951	0.025	0.488
6	12	0	0.1	-0.086	0.140	0.400	0.701	0.550
7	14	0	0.1	-0.338	0.180	0.912	0.140	0.526
8	16	0	0.1	-0.442	0.200	0.857	0.112	0.485
9	14	0	0.1	-0.231	0.168	0.000	0.409	0.205
10	12	0	0.15	-0.394	0.229	0.900	0.133	0.517
11	12	0	0.05	-0.151	0.246	0.842	0.000	0.421
12	12	0	0.1	-0.050	0.177	0.333	0.858	0.596
13	12	0	0.1	-0.251	0.180	0.833	0.656	0.745
14	14	0	0.1	-0.124	0.139	0.000	1.000	0.500
15	14	0	0.1	-0.076	0.183	1.000	0.478	0.739
16	12	0	0.1	-0.457	0.225	0.738	0.212	0.475
17	10	0	0.1	-0.321	0.155	0.405	0.524	0.464
18	12	0	0.1	-0.106	0.160	0.000	0.667	0.333
19	12	0	0.1	-0.419	0.153	0.680	0.245	0.462
20	10	0	0.1	-0.218	0.180	0.964	0.075	0.519
21	12	0	0.1	-0.119	0.156	0.636	0.846	0.741
22	12	0	0.1	-0.370	0.214	0.556	0.308	0.432
23	12	0	0.1	-0.158	0.212	0.889	0.504	0.696
24	14	0	0.1	-0.180	0.209	0.250	0.803	0.526
25	12	0	0.1	-0.182	0.213	0.750	0.516	0.633
26	12	0	0.1	-0.191	0.152	0.583	0.419	0.501
27	14	0.025	0.1	-0.431	0.186	0.825	0.417	0.621
28	12	0	0.1	-0.452	0.148	0.600	0.478	0.539
29	10	0	0.1	-0.166	0.204	0.859	0.096	0.478
30	12	0	0.1	-0.457	0.136	0.525	0.508	0.517
31	12	0	0.1	-0.285	0.132	0.486	0.335	0.411
32	12	0	0.1	-0.087	0.164	0.222	0.946	0.584
33	12	0	0.1	-0.376	0.211	0.917	0.290	0.603
34	12	0	0.1	-0.060	0.136	0.286	0.811	0.549
35	9	0	0.05	-0.135	0.148	0.192	0.695	0.443
36	8	0	0.1	-0.250	0.453	0.957	0.042	0.499
37	12	0	0.1	-0.319	0.187	0.863	0.150	0.507
38	12	0	0.1	-0.238	0.155	0.690	0.473	0.581
39	12	0	0.05	-0.427	0.132	0.308	0.618	0.463
40	12	0	0.05	-0.302	0.107	0.214	0.822	0.518
41	14	0	0.1	-0.097	0.143	0.111	0.889	0.500
42	12	0	0.1	-0.154	0.105	0.417	0.772	0.594
43	12	0	0.1	-0.355	0.244	1.000	0.375	0.688
44	14	0	0.1	-0.363	0.233	0.981	0.009	0.495
45	12	0	0.1	-0.314	0.203	0.889	0.165	0.527
46	12	0	0.1	-0.141	0.160	0.462	0.433	0.447
47	14	0	0.1	-0.211	0.189	0.500	0.524	0.512

Table 3.7: Number of features selected, hyperparameter values, weighted log loss and discrimination metrics of the optimal model in each iteration of the outer loop of a nested cross validation.

Variable	Category	Number (rows)	Percentage	Min	Lower Quartile	Median	Mean	Upper Quartile	Maximum
Age (months)				20.2	41.2	55.6	59.3	73.5	185.3
Yield (litres)				0.1	21.2	28.4	29.0	36.0	97.5
Somatic cell count (x 1000 cells/ml)				1	24	54	185.8	142	18,494
Protein (%)				1.3	3.2	3.4	3.4	3.7	6.0
Butterfat (%)				1.5	3.6	4.3	4.3	4.9	9.0
Mean titre negative cows				0.5	2.6	3.1	3.1	3.7	6.3
Days in milk				0	81.0	169	177.3	255	730
S/P ratio (%)				0	1.7	2.8	7.6	4.9	282.9
Disease status	Negative	69,711	88.7						
	Positive	8,883	11.3						

Table 3.8: Description of features in the data used for training Bayesian models predicting the probability of any given S/P ratio resulting from application of an ELISA test.
Each row in the data represented one ELISA test result.

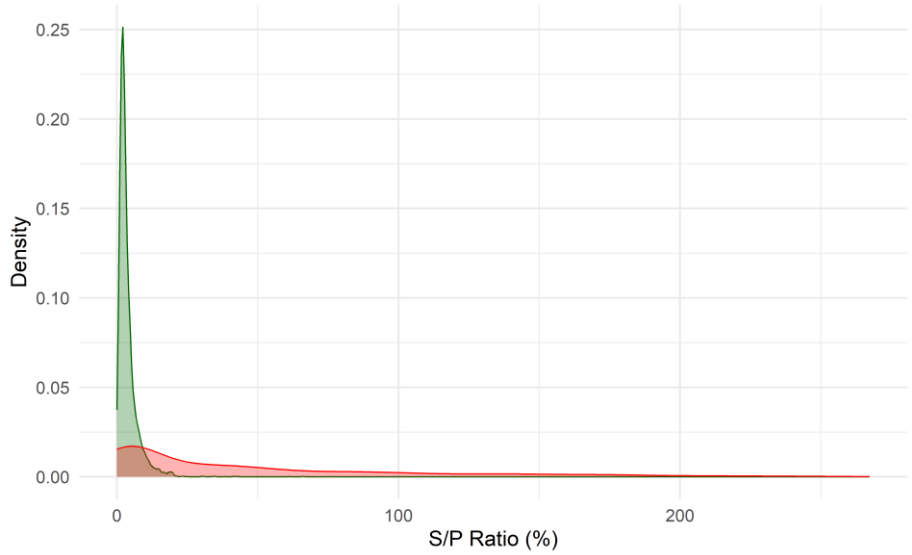

Figure 3.4: Density plots of S/P ratios in the sampled data (one test per cow sampled at random).
Red = Johne's Positive (JP) cows, green = Johne's negative (JN) cows. After splitting, 64.0 % (5866 / 7478 rows) of the data were in the JN training dataset, 14.5 % (1081 / 7478) were in the JP training dataset, 17.7 % (1325 / 7478) were in the JN testing dataset, and 3.8 % (287/7478) were in the JP testing dataset.

3.3.4. Model assessment

The histogram of the posterior probabilities resulting from applying the entire pipeline (Bayesian updating using the Prior and Bayesian Models) on the entire dataset (including cows not meeting the JP or JN criteria) is shown in Figure 3.6. Likelihood ratios ranged from 0.035 to infinity (median 0.327). Illustrations of the results of applying the pipeline to individual JP and JN animals are shown in Figure 3.7.

Matching JP and JN animals resulted in 777 pairs of cows. The boxplot of posterior probabilities of these pairs prior to onset of positivity is shown in Figure 3.8. The median posterior probability of JP cows increased as they approached the onset of positivity, and the distribution of their posterior probabilities was shifted right. This was in contrast to JN animals, whose posterior probability remained steady. Results of the ROC analysis are shown in Table 3.10. Maximisation of the F1 score resulted in modest improvement in PPV when compared to balancing sensitivity and specificity, but did lead to reduced sensitivity.

Calibration plots are shown in Figure 3.9. Consistent with the modest positive predictive values found, calibration of posterior probabilities was visibly poor for all of the selected time periods.

Negative Model

Group-level effects		Mean	Estimated error	Lower 95% Credible Interval	Upper 95% Credible Interval
Intercept standard deviation		0.08	0.02	0.05	0.12
Population-level effects					
Intercept		0.01	0.13	-0.25	0.27
Age (months)		0.00	0.00	0.00	0.00
Yield (litres)		0.00	0.00	0.00	0.00
Log somatic cell count		0.05	0.01	0.04	0.06
Protein (%)		0.16	0.02	0.12	0.20
Mean titre negative cows (%)		0.29	0.01	0.27	0.31
Days in milk	[0 – 5]	*Referent*			
	(5 – 10]	-0.34	0.09	-0.52	-0.17
	(10 – 50]	-0.48	0.09	-0.65	-0.32
	(50 – 100]	-0.46	0.09	-0.63	-0.30
	(100 – 200]	-0.32	0.09	-0.49	-0.15
	(200 – 300]	-0.13	0.11	-0.35	0.07
	> 300	-0.43	0.09	-0.60	-0.26
Lognormal distribution scale parameter					
Sigma		0.47	0.01	0.46	0.48

Positive Model

Group-level effects	Mean	Estimated error	Lower 95% Credible Interval	Upper 95% Credible Interval
Intercept standard deviation	0.41	0.08	0.27	0.60
Population-level effects				
Intercept	1.58	0.31	0.99	2.17
Age (months)	0.01	0.00	0.01	0.02
Yield (litres)	-	0.00	-0.04	-0.02
Log somatic cell count	0.10	0.04	0.03	0.18
Mean titre negative cows (%)	0.34	0.06	0.22	0.46
Lognormal distribution scale parameter				
Sigma	1.39	0.03	1.33	1.45

Table 3.9: Parameters of the two final Bayesian linear regression models predicting S/P ratio.

Standard deviation of the random intercept for farm is shown. The negative and positive models were trained on data from cows which were Johne's negative and positive respectively. Model predictions are therefore the expected S/P ratio of animals given their Johne's status and covariate values.

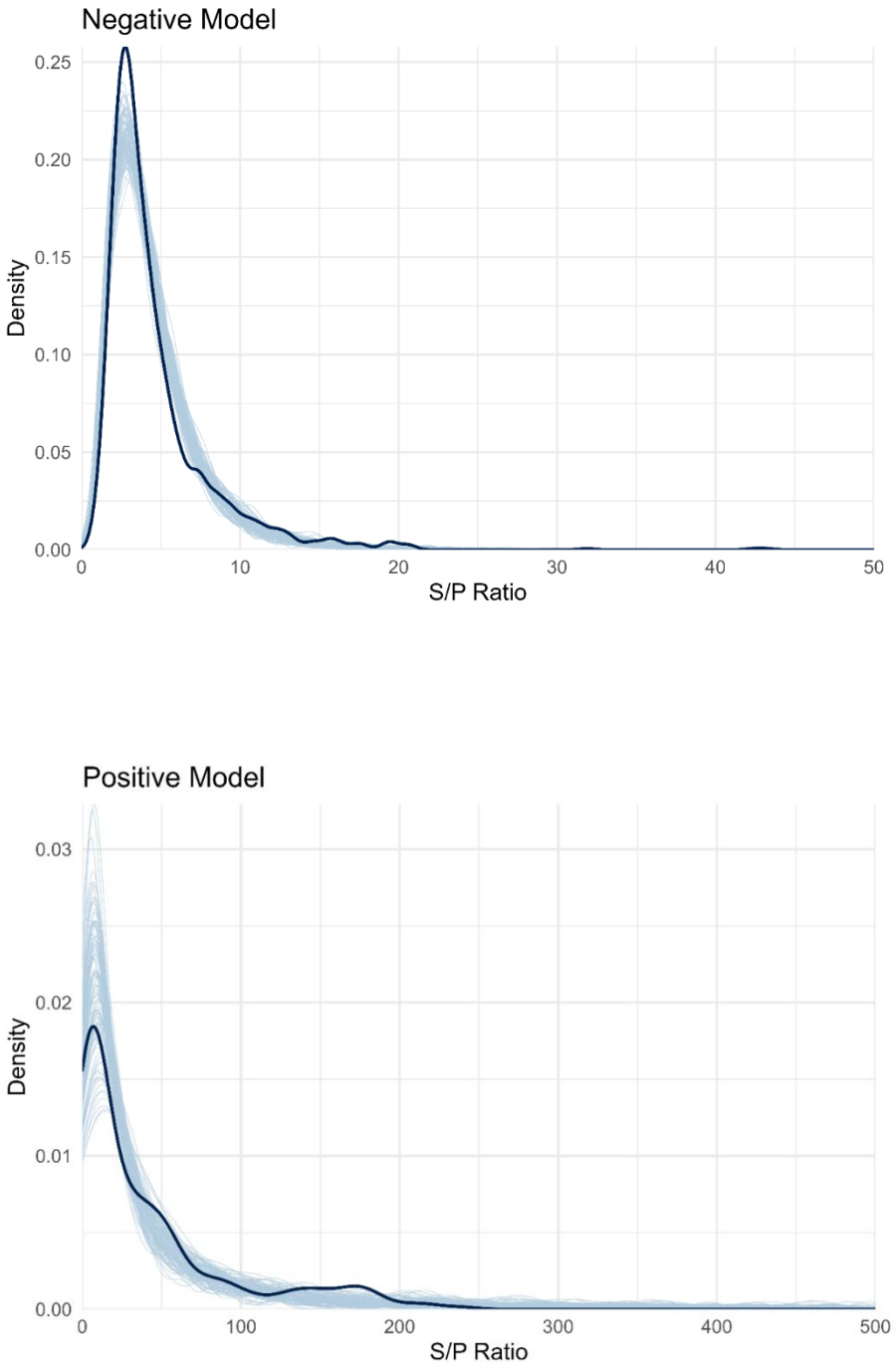

Figure 3.5: Comparison of 100 draws from the posterior predictive distribution (light blue) and the distribution of observed S/P ratios (dark blue) in the testing datasets, for the final Johne's negative (JN) and Johne's positive (JP) models.

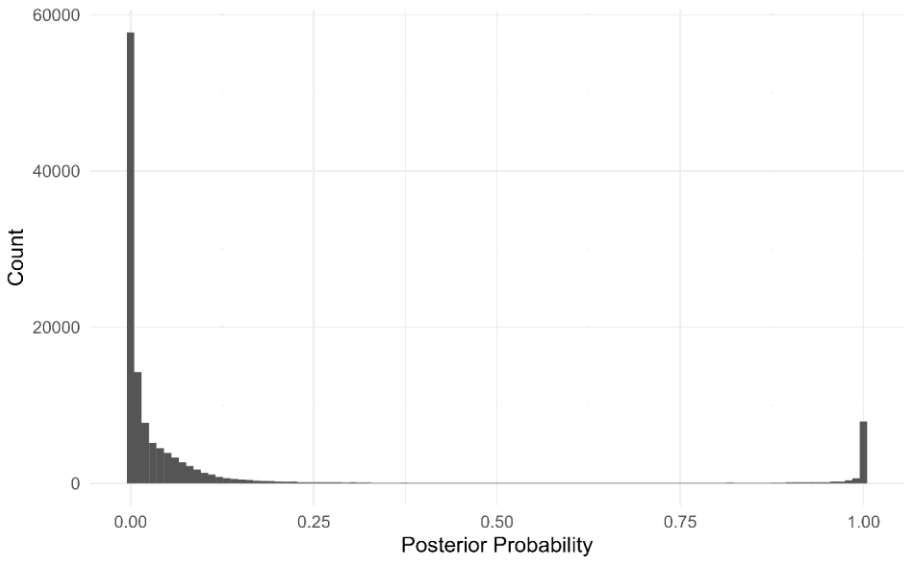

Figure 3.6: Histogram of the bimodal distribution of posterior probabilities resulting from applying the models to the entire dataset.

The individual cow plots disclosed two unanticipated aspects of the Bayesian updating methodology. Firstly, likelihood ratios were infinite on some occasions. Investigation showed that this event was rare; the likelihood ratio was infinite at 1.9 % (2421 / 126,062) of the tests in the entire dataset, in 7.0 % (1426 / 20,391) of cows. The S/P ratio distribution which resulted in infinite likelihood ratios had a median value of 93.3% (min 24.1%, mean 104.0%, max 282.9%). On these occasions the distribution of predicted S/P ratios predicted by the negative cow Bayesian Model did not include the observed S/P ratio. This led to the denominator of the likelihood ratio, $P(T|D^-)$, being zero, giving an infinite value for the likelihood ratio. In turn, the posterior odds at this and all subsequent tests took on a value of infinity, equating to a posterior probability of exactly 1. An example of this scenario is shown in Figure 3.10.

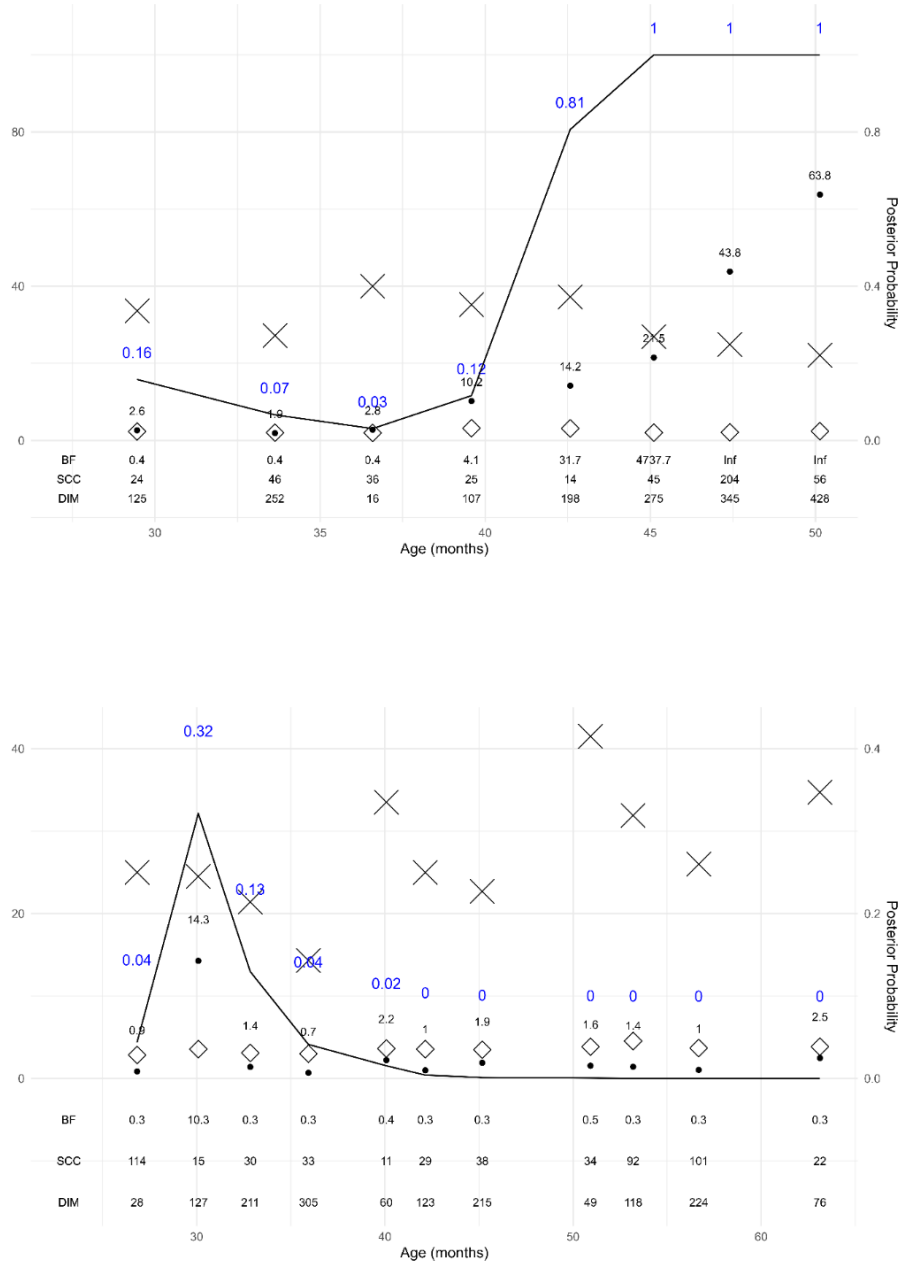

Figure 3.7: Posterior probabilities resulting from Bayesian updating of two individual cows.

Top: Johne's positive (JP) cow; bottom: Johne's negative (JN) cow. Posterior probability is shown as a black continuous line (right y-axis), annotated in blue with posterior probabilities. Individual ELISA tests are shown as points, with S/P ratios (left y-axis) shown in black above each point. Yield, and mean titre of negative cows are shown as crosses and diamonds respectively (left y-axis). Somatic cell count (SCC), days-in-milk (DIM) and likelihood ratio (Bayes' factor; BF) are shown just above the x-axis.

Figure 3.8 : Posterior probabilities of animals at paratuberculosis tests conducted prior to onset of positivity.
Green = Johne's negative (JN) cows, red = Johne's positive (JP) cows. For JP cows, onset of positivity was at the first of the pair of positive (S/P >= 30%) results which resulted in a positive status according to a commonly used interpretation of serial test results. For matched JN cows, onset of positivity was at the test conducted at an age most similar to, and within three months of, the age at onset of positivity of the cow's positive counterpart. Months prior to onset of positivity is categorised. Square brackets within age categories indicate that the upper value of each category is included in the category.

Method	Months until positive status	Threshold	Prevalence[1]	Balanced Accuracy	Sensitivity	Specificity	PPV[2]	NPV[3]	AUROC[4]
Balance sensitivity and specificity	Any	0.01	0.15	0.76	0.76	0.76	0.37	0.94	0.85
	24	0.01	0.12	0.76	0.76	0.76	0.32	0.95	0.84
	18	0.02	0.11	0.77	0.77	0.77	0.30	0.96	0.85
	12	0.02	0.08	0.78	0.78	0.78	0.25	0.97	0.85
	6	0.03	0.04	0.79	0.79	0.79	0.15	0.98	0.87
Maximise F1 score	Any	0.04	0.15	0.75	0.63	0.86	0.46	0.93	0.85
	24	0.07	0.12	0.73	0.56	0.90	0.44	0.93	0.84
	18	0.07	0.11	0.74	0.58	0.89	0.40	0.94	0.85
	12	0.11	0.08	0.71	0.50	0.93	0.40	0.95	0.85
	6	0.12	0.04	0.74	0.57	0.91	0.24	0.97	0.87

Table 3.10: Results of receiver operating characteristic analysis, showing the optimal threshold of posterior probability for tests occurring within different length periods prior to onset of positivity.
[1]Prevalence = proportion of animals for which onset of positivity occurred within the stated period.
[2]Positive predictive value; [3]Negative predictive value
[4]Area under the receiver operating characteristic curve.

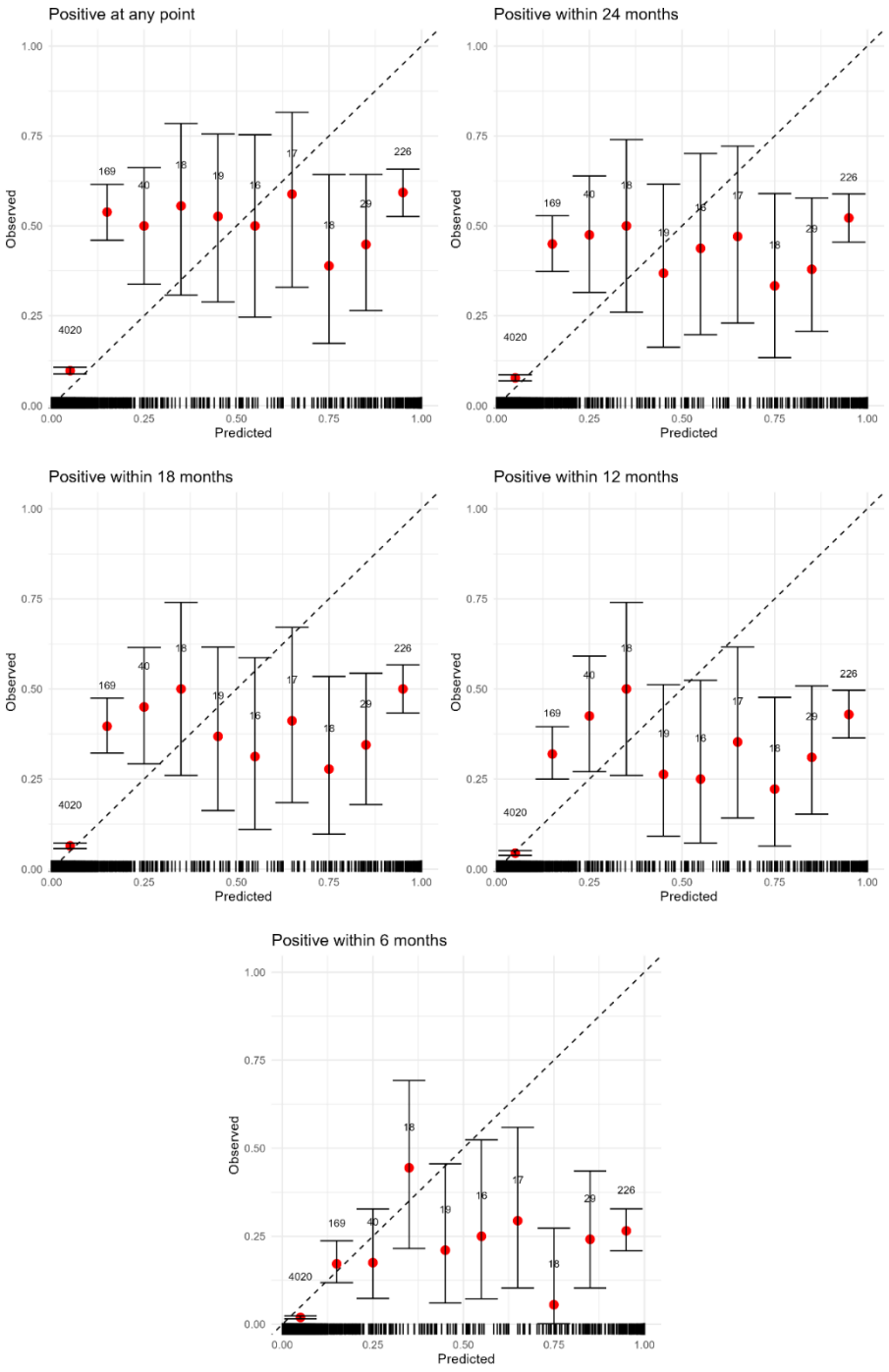

Figure 3.9: Calibration plots of posterior probabilities within various intervals prior to onset of positivity.
The number of subjects in each probability bin is shown above each point. 95% confidence intervals are shown as error bars. Rug plots on the x-axes show the distribution of posterior probabilities.

Secondly, some cows had a posterior probability of approximately zero following test results considerably above the test-positive threshold (S/P ≥ 30%), an example of which is illustrated in Figure 3.10. At 0.12% (146/126,062) of tests, in 0.7% (140/20,391) of cows, the posterior probability was less than 0.01 despite a concurrent S/P ratio of at least 30%. This scenario occurred in animals which were on average older (median 79.8 months vs. 45.0 months) than the data average, and which had had more tests (median 12 tests vs. 4 tests) than the average cow in the data. The posterior odds at these tests, despite S/P ratios being at least 30%, were lower than the average across all of the data (median 0.0003 vs. 0.008) and had a maximum value of 0.009. Repeated Bayesian updating, over many tests, with likelihood ratios less than zero, had led to exceptionally small values of posterior odds. When used as the prior for Bayesian updating, even a large likelihood ratio, such as occurred at a positive test result (S/P ≥ 30%), would not result in posterior odds sufficiently large to give a posterior probability noticeably above zero.

3.4. Discussion

This study used Bayesian updating to interpret the results of serial milk MAP antibody ELISA tests in dairy cows, incorporating numerical ELISA results (S/P ratios) and influential covariates, and providing real-time probabilistic predictions of Johne's disease status in individual cows. Aspects of this methodology for MAP ELISA interpretation have been previously investigated, for example use of likelihood ratios for numerical test results (Collins, 2002), consideration of non-disease factors (McAloon et al., 2020) and

Bayesian updating (Meyer et al., 2018). However, combining all these aspects is a novel approach.

The posterior predictions resulting from Bayesian updating could be useful for cow management by differentiating between animals of relatively low or high disease-risk. Whilst calibration of posterior probabilities was found to be poor, nevertheless it is noteworthy that the reported methodology has the potential to alert that certain cows are likely to enter a Johne's positive status in the future, with balanced accuracy approximately between 75 and 80% and AUROC between approximately 0.85 and 0.9.

One interesting finding is that likelihood ratios of negative (S/P < 30%) ELISA results commonly have a value above 1, suggesting that the threshold commonly employed for classifying a result as "positive" is inappropriate for sensitive detection of diseased animals. This is consistent with plentiful research which has found that MAP antibody ELISA result interpretation tends to favour specificity and positive predictive value over sensitivity (Beaver et al., 2017; Laurin et al., 2017; Nielsen and Toft, 2008, 2006; Sweeney et al., 2006). Another interesting result is that some ELISA results were associated with an infinite likelihood ratio, suggesting that on occasion a single high test-result gives confidence that an animal is pTB-infected. This is in contrast to pTB-risk classification schemes used commonly in the UK (Table 3.1), in which cows must receive two positive (S/P ≥ 30%) results before being confirmed positive.

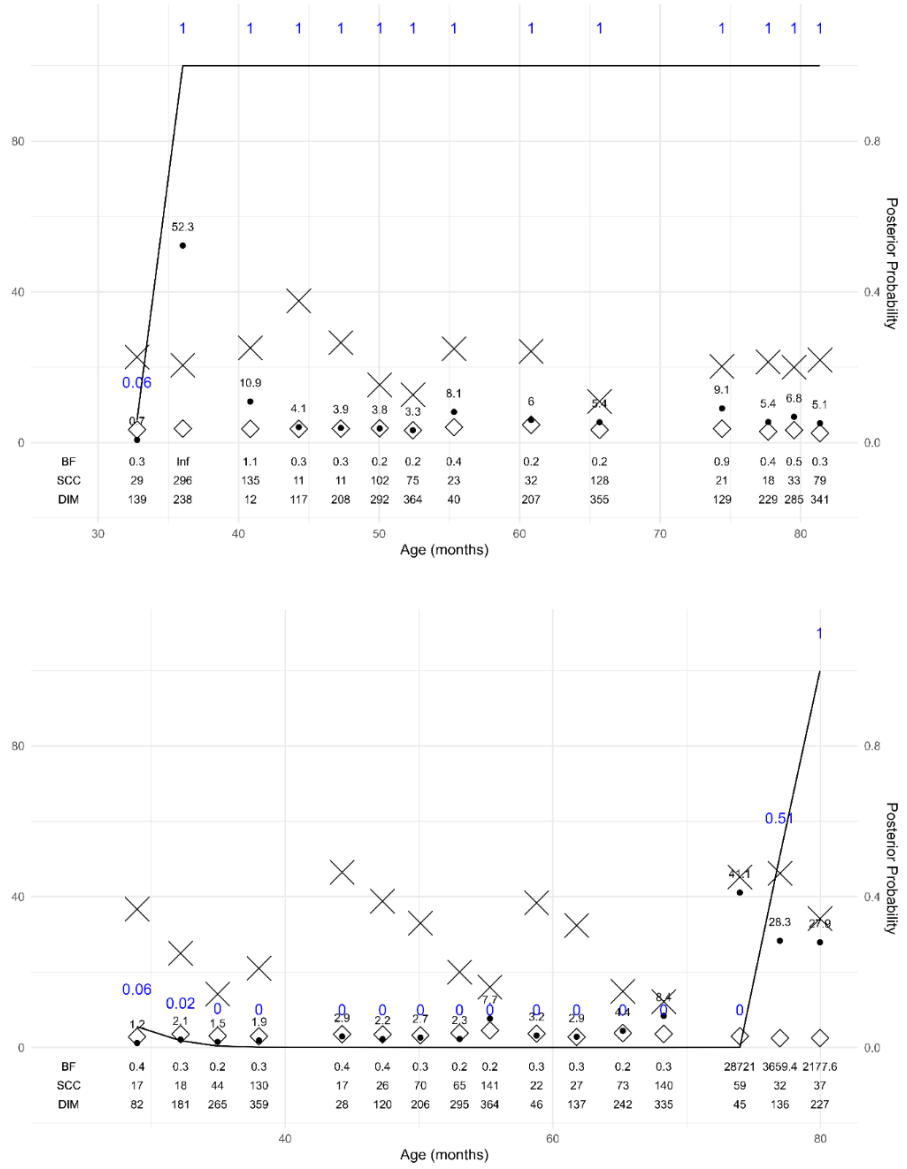

Figure 3.10: Application of the Bayesian updating methodology to the ELISA results of two cows.

Top: a cow for which the likelihood ratio was infinite at a test in early life, resulting in a posterior probability of 1 at this and all subsequent tests. Bottom: a cow for which the posterior probability remained low despite an S/P ratio considerably above the test-positive threshold (30 %). At the antepenultimate test (S/P = 41.1 %), the posterior probability remains approximately zero despite a likelihood ratio of 28,721. The prior and posterior odds at this test were 0.00000001 and 0.0003 respectively.

Test results (S/P ratios) are shown as points, with the value of the S/P ratio shown in black above each point. The likelihood ratio (Bayes' Factor; "BF") at each test is shown above the x-axis, along with the somatic cell count (SCC) and days-in-milk (DIM). Yield is represented by crosses, mean titre of negative cows by diamonds. The posterior probability is shown as a black line, and the values of posterior probability at each test are printed in blue above the line.

The Prior Model provides probabilistic predictions of disease status using only information regarding the chances that an animal was infected with MAP in early life (for example variables capturing the likely disease state of other animals present in the same environment at the time of birth, or capturing apparent herd prevalence). Overall calibration (Figure 3.3) of the Prior Model predictions is likely to be acceptable (as determined by nested cross validation), raising the possibility that the predictions could be used in farm management (e.g. breeding decisions in nulliparous animals) in the absence of test results. The usefulness of the Prior Model for classifying animals is less certain, as nested cross validation demonstrated that accuracy metrics (sensitivity and specificity) were likely to be unstable on novel unseen data (Table 3.7). In the current context, the Prior Model was used only for provision of prior information (itself a novel approach) in the Bayesian updating.

Use of likelihood ratios for interpretation of MAP antibody ELISA results has been previously described (Collins, 2002). In that study, likelihood ratios were estimated for a range of S/P ratios in serum MAP antibody ELISA results. Consistent with the current study, likelihood ratios were often greater than 1 for S/P ratios below the recommended test-positive cutpoint, suggesting that dichotomisation of S/P ratios using the recommended cutpoint leads to a lack of sensitivity for identifying infected animals. Bayesian updating for interpretation of cow serial MAP antibody ELISA test results has also been previously reported (Meyer et al., 2018). Using methods similar to the current study, posterior probabilities were determined for individual cows following each

ELISA test, by combining prior information and likelihood ratios. Likelihood ratios accounted for age of the animal, but used only a dichotomous (positive or negative) interpretation of results. A change-point model has also been reported which provides posterior probabilities that a cow has reached a state of positivity (Wang et al., 2011). In that study, and similar to the current work, numerical S/P ratios were interpreted by the model in the context of some non-disease factors (yield and age), and the output (probability that a change point had been reached) was probabilistic. However, the implications of applying this model were scarcely explored.

The data used for this study were from a convenience sample of herds utilising QMMS LTD. for milk-recording and pTB-testing. MAP ELISA test results are known to vary by laboratory and geographical region (Barden et al., 2020), so generalisability of the current methodology to data from other sources is not assured. The current results are also likely to be dependent on the inherently subjective criteria used to define the disease status of cows (JP or JN). Meyer *et al.* (2018), in their study of Bayesian updating with dichotomised S/P ratios, found in a sensitivity analysis that altering the criteria defining positive and negative cows resulted in considerable changes to calculated ELISA test sensitivity and specificity.

Bayesian updating over all previous test results in each animal is most appropriate if the animal's status is static throughout its lifetime; in such a scenario, test results occurring at any stage in an animal's life would be expected to be typical of the animal's disease state (i.e. infected or uninfected), and the posterior probabilities would be expected to represent the likelihood that an animal was

infected. However, the immune status of an infected cow is not static, but instead has distinct phases, typically characterised by a period of seronegativity followed by seroconversion (Begg et al., 2018; Tiwari et al., 2006). Test results occurring early in an infected cow's life are therefore unlikely to be distinguishable from those of uninfected animals. In animals tested frequently prior to seroconversion, the posterior odds could become exceptionally low (as in Figure 3.10), giving unacceptably low posterior probabilities in some cows even following positive test results. Repetition of the current analysis, using Bayesian updating over only the most recent few test results, would perhaps lead to better calibrated posterior probabilities.

3.5. Conclusions

This study has demonstrated the potential of using numerical test results (S/P ratios) for interpretation of serial MAP antibody ELISA results, providing probabilistic predictions of an animal's disease status whilst accounting for influential covariates. Furthermore, S/P ratios below the diagnostic threshold (S/P < 30%) can potentially provide an early warning that an animal is approaching a positive disease status as defined by a commonly used pTB risk-classification. Such information would be useful for farmers and advisors when making management decisions. However, practical application of this method would require improved calibration of its predicted probabilities as well as demonstration of its generalisability.

Chapter 4. Interpretation of serial paratuberculosis test results using a hidden Markov model

An alternative method to interpretation of serial MAP antibody ELISA results was motivated by the findings in Chapter 3. Multiple linear regression models were used to assess and correct for the associations between non-disease factors and the numerical MAP ELISA result (S/P ratio). Corrected ELISA results were then used to train a hidden Markov model for clustering of longitudinal pTB testing data. Cows can be allocated to one of several different latent classes according to model predictions.

4.1. Introduction

Hidden Markov models (Rabiner, 1989) are useful to cluster subjects in time-series data. In HMMs, subjects move through a sequence of several latent states, such as stages of a disease. The latent states are not observed (hidden), but are associated with an observed variable, for example a milk test result, whose probability is a mixture distribution in which each component corresponds to a latent state. Test results are used to infer in which latent state an animal resides.

The aim of this study was to create a HMM which objectively assigns animals to latent states according to numerical (not dichotomised) test values of serial milk MAP antibody ELISA results, whilst accounting for non-disease factors which influence test results. In this model, cows could move between latent states over time,

which is appropriate for representing the dynamic nature of cows' immune responses to MAP-infection (Begg et al., 2018; Tiwari et al., 2006).

Such a model may aid pTB control by revealing early patterns in individual cow serial test results which are currently overlooked and potentially indicative of significant risk of being infected or infectious.

4.2. Materials and Methods

Data analysis proceeded in two stages: 1) Creation and evaluation of models for adjusting pTB test results (S/P ratios) based on non-disease variables; and 2) building and evaluating HMMs for time-series clustering of animal states utilising adjusted pTB test results. At each of these two stages, data were first cleaned appropriately for model creation. Model predictions made on raw (unclean) data were then examined to help understand the implications of applying the models to the sort of data routinely collected by milk-recording organisations.

4.2.1. Creation and evaluation of models for adjusting paratuberculosis test results

The target population were UK dairy herds engaged in regular pTB milk ELISA testing and regular milk recording. The source population were all herds milk recording with Quality Milk Management Services (QMMS) LTD., UK. All farms were included if, during the ten-year period between 1st January 2011 and 31st December 2020, they had conducted at least 20 whole-herd pTB tests and had conducted milk-recording on average at least four times per year. A whole herd pTB test was defined as an occasion on

which at least 70% of lactating animals had been tested. Testing consisted of quantification of MAP antibodies in individual cow milk samples, using a commercially available ELISA (IDEXX B.V., Hoofddorp, The Netherlands). In accordance with the policy of QMMS LTD., which specifies that pTB testing should not occur within 60 days following an intradermal tuberculosis test, it was assumed that pTB tests were not carried out during this period. Test results and milk recording data for all animals that had calved at least once were retrieved via dairy analysis software (TotalVet Version 3, QMMS LTD., UK). The resulting dataset consisted of one row per individual animal pTB test result, with any associated milk recording data for that animal on the same day, including parity, age, DIM, yield, SCC and milk quality (% protein and % butterfat). ELISA test results (optical densities) were converted to S/P ratios, and expressed in the data as percentages, as follows:

$$S/P\ Ratio\ = 100 * \frac{OD_{sample} - OD_{negative\ control}}{OD_{positive\ control} - OD_{negative\ control}}$$

where OD = optical density.

Somatic cell counts and S/P ratios were transformed due to their right-skewed distribution, by adding one (to eliminate zero-values) and then calculating the natural logarithm. To account for possible variability in ELISA test kits (Köhler et al., 2022), day-to-day laboratory procedures and herd average immune response, a new variable was created, henceforth referred to as the mean titre of negative cows (MTNC). This variable was the mean of all S/P ratios below 10% at each test day in each herd. It was anticipated that S/P ratios below 10% would largely represent animals which were not mounting a humoral immune response, and so variation in MTNC

between test days in a herd would be a measure of any variation in ELISA results which was not due to MAP infection.

Due to differences between primi- and multiparous animal groups in the association of some non-disease factors (e.g. yield, DIM) with S/P ratios, data were split into primiparous and multiparous sets. Cleaning of both datasets was performed with the aim of removing erroneous measurements. Rows were removed if their values for any predictor variables were implausible, as follows. Given a realistic minimum age at first calving (AFC) of 20 months, rows of the primiparous data with age less than this value were removed, as were rows of primiparous data with age greater than 60 months. Within the data, calvings or abortions only triggered a new lactation if they occurred following a gestation of at least 4.5 months. Therefore, for multiparous data, rows with age less than 25 months (minimum AFC, plus a short interval between calving and conception, plus 4.5 months) were excluded. A minimum DIM of zero days (indicating the day of parturition) was permitted, and rows with DIM greater than 730 days were removed. Rows with protein less than 1% or greater than 6% were excluded, as were rows with butterfat less than 1.5% or greater than 9%. The range of values for S/P ratio, yield, SCC and MTNC were all considered plausible. Prior to training of the linear models only, rows with missing milk-recording data (yield, SCC, protein or butterfat) were also removed. Missing data were deemed missing at random (Rubin, 1976).

Similar to the methods of McAloon *et al.* (2020), linear models were used to quantify the association of the natural logarithm of the pTB test result (LogSP; dependent variable), with age, lactation stage, SCC, milk yield and quality, as well as MTNC, for each of the primi-

and multiparous datasets. Models were fitted using the *lme4* package (Bates et al., 2015) in R version 4.2 (R Core Team, 2022). Univariable plots were used to assess the linearity of the association between each predictor and LogSP. Bivariable plots (showing the association of LogSP with one predictor, at varying values of a second predictor) were used to assess the presence of interactions between predictors. Non-linear associations between predictors and LogSP were modelled using cubic b-splines (McAloon et al., 2020). K-fold repeated cross validation (10 folds, 10 repeats) of univariable models for each feature was employed to find the optimum number of b-spline degrees of freedom. The optimum number was taken as that which minimised RMSE whilst, to prevent overfitting, minimising the difference between apparent and cross-validated RMSE. Predictions from each resulting univariable spline model were plotted for a range of the predictor, to check for overfitting, and if necessary, the number of degrees of freedom was reduced until the resulting plot showed a smooth curvilinear association.

Multivariable mixed effects models (LogSP; dependent variable) with random intercepts for farm, and cow nested in farm, were fitted including b-spline and interaction terms as appropriate. For each of the primiparous and multiparous datasets, two models were fit; one model was offered all the predictors, whilst a reduced model was offered only age, DIM and MTNC as predictors. The reduced models were later used for prediction in cases where milk recording data were unavailable. Features were selected in a backwards-stepwise manner using AIC. At each step, the variable whose exclusion resulted in the lowest AIC was removed. Variables were

nevertheless retained in the model if confounding was evident, as indicated by a change of at least 10% in the coefficient of a covariate significantly associated with the dependent variable. Biologically plausible two-way interactions suggested by the bivariable plots were included if they improved AIC. Normality of random effects was evaluated through histograms and Q-Q plots. Marginal effects plots were examined to visualise the association of different features with LogSP.

Adjustment of S/P ratios proceeded in a similar manner to previously published methods (McAloon et al., 2020). One of the four models was used to predict LogSP for each animal in the raw, unclean data, at each timepoint; i) a model for primiparous cows when milk recording data were present, ii) a model for multiparous cows when milk recording data were present, iii) a model for primiparous cows when milk recording data were missing and iv) a model for multiparous cows when milk recording data were missing. Predicted LogSP values were converted back to S/P ratio, so that predictions were the expected S/P ratios, given covariate values, irrespective of true disease status. Residuals (the difference between the actual and predicted S/P ratios) were calculated. Next, the full models (including milk-recording features) were used to predict an S/P ratio for two hypothetical animals, one primiparous and one multiparous, with median values of all predictors. These predictions represented the expected S/P ratio of 'average' primiparous and multiparous animals. Finally, S/P ratios adjusted for covariates were calculated by adding the model residual for each data point to the model-predicted S/P ratio for the hypothetical average animal.

Adjusted S/P ratios were compared with unadjusted S/P ratios by inspection of scatter plots. Adjusted S/P ratios which were at least ten units more or less than the unadjusted S/P ratio were considered unusual. Covariate values of these unusual data points were summarised to explore the number of times and the circumstances in which S/P values would be adjusted to such an extent.

4.2.2. Building and evaluating hidden Markov models for time-series clustering

The raw data were again cleaned to remove erroneous measurements. Rows of data with implausible values for age, DIM, protein or butterfat were identified using the same criteria as were applied previously. To fit models to time-series data, it was important that every test result was present for any animal in the training data. For this reason, if an animal had any implausible values in the data, then all the data for that animal was excluded. Animals having been tested for pTB only once during the period were also removed, as their data would not be informative in time-series clustering. S/P ratios were adjusted for influential covariates as described previously. Negative values of adjusted S/P ratios were set to zero in line with the test kit manufacturer's recommendations for reporting results. A fixed value of one was added to each adjusted S/P ratio to permit its representation as a lognormal mixture distribution in the subsequent models.

Hidden Markov models, comprising continuous-time first-order Markov processes, were utilised for clustering of time-series data. When fitting such models, there are three types of parameter to be

estimated; initial state probabilities (the probabilities of an animal being in each state prior to any pTB test results), transition probabilities (the probabilities of an animal moving from a certain state to another state over a defined period of time) and emission probabilities (the probability density functions of the numerical ELISA test results corresponding to each state). Transitions between states were permitted only between adjacent states (e.g. from 1 to 2, but not from 1 to 3). However, in a continuous time HMM, transitions are expected to occur during the interval between observations (pTB test results). A subject can therefore appear to move from one state to a non-adjacent state from one observation to the next because of unobserved transitions occurring in between ELISA test results. Emission probabilities (S/P ratios) were specified as lognormal distributions. Models were trained using the *msm* package (Jackson, 2011) in R version 4.2 (R Core Team, 2022).

The aim of modelling was to produce a predictive model to support decision making by farmers regarding management of individual cows. It was anticipated that predictions from a model with many states would be difficult to interpret and would therefore be of limited use to producers; therefore, models with up to four latent states were considered from the outset. Parity, categorised as "1", "2" or "3+", was included as a covariate for all transition probabilities. To aid convergence, the number of parameters to be estimated was reduced by constraining LogSP variance to be equal across states, and by specifying the initial state probabilities prior to fitting. Appropriate initial state probabilities were determined by comparing the likelihood of models fit with a variety of different values. The optimum number of states was selected by comparing

AIC of the different models. Maximum likelihood calculations in the *msm* package require initial values to be supplied for the parameters to be estimated (transition probabilities, and location and scale of the emission probability functions). These initial values are given in Table A.1 (Appendix A). Sensitivity of models to the initial values supplied was assessed by varying these values and comparing fitted model parameters.

Transition probability matrices for each parity category, corresponding to the median interval between pTB tests in the data, were extracted from the final model. The matrices were used to determine the dynamics of state membership, and to compare relative risks of transitioning to another state from different starting states and in different parities. Bootstrap 95% confidence intervals were calculated for the transition probabilities and relative risks, using 100 bootstrap samples of the data. To visualise the distribution of adjusted S/P ratios associated with each state, probability density functions were plotted using the scale and location parameters of the emission distributions from the fitted model.

The adequacy of the model fit was assessed by comparison of simulated and observed S/P ratios (Gelman et al., 1996; McElreath, 2018) as follows. A dataset of 100,000 animals was simulated, and animals were allocated to a state according to the model initial state probabilities. The animals were then simulated to transition at four-month intervals (approximately equal to the mean interval between tests in the data) to other states, until a total of six transitions had occurred, mimicking the passage of two years. Animals moved between states according to the transition

probabilities for the mean parity in the data. At each timepoint, for each animal in this simulated dataset, one random draw was taken from the lognormal distribution defined by the HMM emission parameters for the state in which that animal resided at that timepoint. A fixed value of one was subtracted from this random draw to convert it to a simulated adjusted S/P ratio. Two thousand random samples of adjusted S/P ratios were taken from each of the simulated and observed datasets and were compared visually using histograms. The process was repeated multiple times, comparing samples of S/P ratios from the simulated dataset with samples of S/P ratios from different folds of the observed dataset. Folds were created by dividing the observed data according to one particular feature of interest. For example, for continuous features (DIM, yield, SCC, protein, butterfat and MTNC), folds were created by splitting the data into quartiles of the feature of interest, yielding four folds. Three folds were created for parity (one for each of the parity categories 1,2 and 3+). Five folds were created containing data from different farms. The validity of posterior predictions was therefore assessed across a range of scenarios.

To appraise model predictions, two types of probability were calculated for each animal at each timepoint, using the forward and Viterbi algorithms respectively (Rabiner, 1989); 1) filter probabilities (the probability of an animal being in each state given all of its pTB test results up to that timepoint) and 2) the hidden path (given the filter probabilities at the end of the time series for each cow, the most likely path taken through the hidden states). Plots of individual cows, showing their adjusted titres and associated filter probabilities and hidden paths, were studied to further understand

model predictions at the individual animal level. For each herd, plots were created showing the proportion of cows in each state over time according to hidden paths.

4.3. Results

4.3.1. Creation and evaluation of models for adjusting paratuberculosis test results

Prior to cleaning, the data comprised 252,789 pTB test results, from 39,419 animals on 47 farms. After cleaning, the data comprised 220,179 results, from 37,792 animals on 47 farms. Testing occurred between 7th December 2010 and 22nd December 2021. Details of the number of rows of the uncleaned data with implausible values are given in Table A.2 (Appendix A). Descriptive statistics of the data before and after cleaning are shown in Table A.3 and Table A.4 (Appendix A).

Potential non-linear associations, modelled using cubic b-splines, were observed between DIM and S/P ratio in both the primiparous and multiparous data. The numbers of b-spline degrees of freedom for DIM in the primiparous and multiparous models were 8 and 11 respectively.

The parameters of the four final models are shown in Table A.5 (Appendix A). Conditional effects of the features specified as b-spline terms are shown in Figure A.1 (Appendix A). The relationship between DIM and S/P ratio in all models was curvilinear, such that S/P ratios were lowest during mid lactation, higher towards late lactation, and higher still during early lactation. In the multiparous full model, there were interactions between log SCC and yield, between DIM and yield, and between log SCC and DIM. In the

primiparous full model, there were interactions between log SCC and yield, log SCC and protein, and between protein and DIM. DIM and age interacted in the primiparous reduced model. Interaction effects are shown in Figure A.2 (Appendix A).

The adjusted S/P ratio was at least 10 units less than the unadjusted S/P ratio in 0.8% (2,005 / 252,789) of rows of the data. These unusual cases had a median adjusted S/P ratio of -1, and at least one covariate whose value was lower than the 5th or greater than the 95th percentile (i.e. extreme).

4.3.2. Building and evaluating hidden Markov models for time-series clustering

After cleaning, the data comprised 224,852 test results, from 33,230 animals on 47 farms. Descriptive statistics of the dataset after cleaning are shown in Table 4.1. Original and adjusted S/P ratios in the cleaned data are shown Figure A.3 (Appendix A).

AIC indicated that the HMM with four states was optimal. The initial state probabilities were 0.5, 0.4, 0.05 and 0.05 for states 1 to 4 respectively. Sensitivity analysis demonstrated that the model fit was insensitive to the initial values supplied during model fitting. Halving or doubling the initial transition intensity values or the initial values for emission probability distribution location parameters resulted in negligible changes to HMM parameters.

Feature	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum
Date	07/12/10	25/02/15	24/05/17	12/03/17	31/07/19	22/12/21
S/P Ratio (%)	0.0	1.7	2.8	6.7	5.2	282.9
Adjusted S/P Ratio (%)	0.0	1.4	2.4	5.9	4.2	282.3
Parity	1.0	1.0	2.0	2.6	3.0	16.0
Age (months)	20.3	35.7	47.2	53.3	65.0	228.6
Days in milk (days)	0.0	78.0	163.0	173.9	250.0	730.0
Milk yield (litres)	0.1	21.1	28.0	28.7	35.6	97.5
Somatic Cell Count (x 1000 cells/ml)	1.0	23.0	52.0	185.9	136.0	18494.0
Protein (%)	1.2	3.2	3.4	3.4	3.7	6.0
Butterfat (%)	1.5	3.6	4.2	4.3	4.9	9.0
Mean titre negative cows (% S/P)	0.5	2.6	3.1	3.1	3.7	6.3

Table 4.1: Descriptive statistics of the features in the cleaned dataset used for building hidden Markov models.
Data were taken from cows (n = 33,230) in 47 dairy herds in the United Kingdom which were engaged in regular milk recording and paratuberculosis testing. Herd size during the study ranged from 14 to 1169 (median = 179.0, mean = 247.6).

Extracted transition matrices for the HMM (Table 4.2) showed that, in almost all instances, over a period equal to the median intertest interval (3.91 months), animals were more likely to remain in their current state than to transition to another state. The exception to this was for primiparous animals in state 3, which were most likely to transition to state 2. The probabilities of transitioning away from states 1 or 2 were particularly low, and the probability of transitioning away from state 1 became lower with increasing parity. The probability of transitioning away from states 3 and 4 were somewhat higher but again decreased with increasing parity. Relative risks of animals in different parities moving, over a period equal to the median intertest interval, to state 4 from each of the other states are shown in Table 4.3. These relative risks

demonstrate that progression to state 4 was substantially more likely from higher states when compared to lower states. For example, primiparous animals were approximately 10 times more likely to transition from state 2 to state 4 than from state 1 to state 4, and approximately 15 times more likely to transition from state 3 to state 4 than from state 2 to state 4. In multiparous animals the differences in relative risks were larger. Animals in parity 3 or greater were approximately sixty times more likely to transition to state 4 from state 2 compared to state 1, and approximately seventeen times as likely to transition to state 4 from state 3 compared to state 2. The distributions of adjusted S/P ratios according to the emission parameters of the fitted model are shown in Figure 4.1.

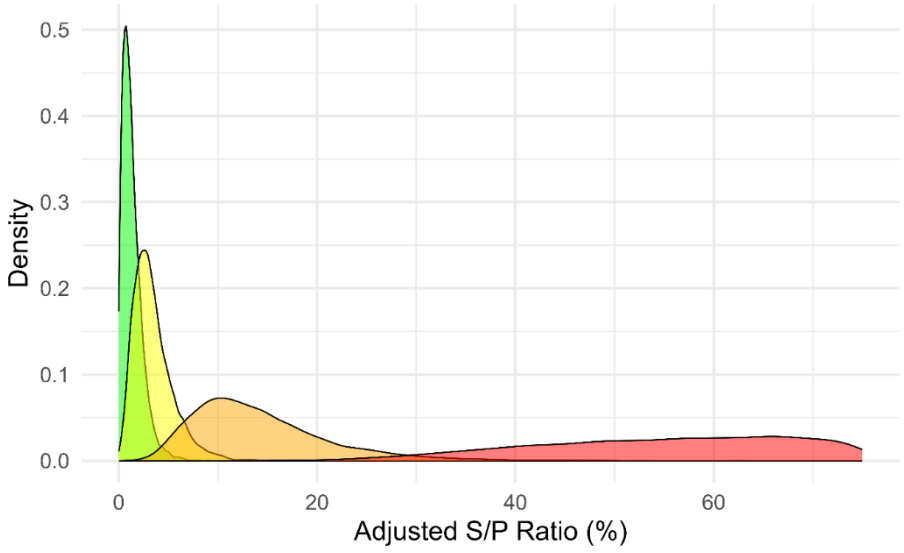

Figure 4.1: Emission (S/P ratio) probability distributions corresponding to the latent states in a hidden Markov model.
Green = state 1, yellow = state 2, orange = state 3, red = state 4. The x-axis is truncated at a value of 75 % to improve visualisation of the lower states.

Posterior predictive checks (comparing the distributions of simulated and observed adjusted S/P ratios) did not demonstrate poor predictions in any ranges of features in the data. The prevalence over time of cows in different hidden states in an example herd, according to hidden paths, is shown in Figure A.4 (Appendix A), and model predictions on two example cows are shown in Figure A.5 (Appendix A).

		To state			
	From state	**1**	**2**	**3**	**4**
Parity 1	1	0.762 (0.75 - 0.774)	0.228 (0.217 - 0.24)	0.009 (0.008 - 0.01)	0.001 (0.001 - 0.001)
	2	0.014 (0.011 - 0.017)	0.917 (0.912 - 0.922)	0.062 (0.059 - 0.064)	0.008 (0.007 - 0.009)
	3	0.004 (0.003 - 0.005)	0.443 (0.419 - 0.476)	0.434 (0.405 - 0.464)	0.12 (0.112 - 0.128)
	4	0.001 (0.001 - 0.001)	0.175 (0.157 - 0.196)	0.372 (0.341 - 0.412)	0.452 (0.396 - 0.497)
Parity 2	1	0.9 (0.89 - 0.91)	0.096 (0.086 - 0.106)	0.003 (0.003 - 0.004)	0 (0 - 0)
	2	0.034 (0.031 - 0.038)	0.904 (0.899 - 0.909)	0.057 (0.055 - 0.06)	0.005 (0.004 - 0.005)
	3	0.005 (0.004 - 0.006)	0.251 (0.232 - 0.268)	0.635 (0.616 - 0.654)	0.11 (0.101 - 0.119)
	4	0.001 (0.001 - 0.001)	0.048 (0.04 - 0.057)	0.258 (0.226 - 0.288)	0.694 (0.659 - 0.733)
Parity 3+	1	0.958 (0.954 - 0.962)	0.04 (0.036 - 0.044)	0.002 (0.002 - 0.002)	0 (0 - 0)
	2	0.086 (0.082 - 0.092)	0.827 (0.82 - 0.833)	0.081 (0.077 - 0.085)	0.006 (0.005 - 0.006)
	3	0.012 (0.011 - 0.014)	0.228 (0.209 - 0.25)	0.666 (0.645 - 0.686)	0.094 (0.089 - 0.1)
	4	0.001 (0.001 - 0.002)	0.038 (0.033 - 0.043)	0.23 (0.206 - 0.252)	0.731 (0.702 - 0.757)

Table 4.2: Transition probability matrices for the hidden Markov model with four hidden states.
Probabilities represent the probability of animals in different parities moving from one state to another over a period equal to the median interval between paratuberculosis tests (3.91 months). Bootstrap 95% confidence intervals are shown in brackets.

	From state	Relative Risk
Parity 1	1	1 (1 - 1)
	2	10 (9 - 11)
	3	153 (139 - 167)
	4	577 (471 - 673)
Parity 2	1	1 (1 - 1)
	2	27 (24 - 30)
	3	624 (548 - 722)
	4	3,948 (3,230 – 4,771)
Parity 3+	1	1 (1 - 1)
	2	62 (56 - 69)
	3	1,066 (943 – 1,217)
	4	8,264 (7,070 – 9,556)

Table 4.3: Relative risks in a hidden Markov model of an animal finishing in state 4 (the highest state) after a period equal to the median interval between paratuberculosis tests (3.91 months).
Relative risks are according to an animal's state at the beginning of that period, compared to state 1.

4.4. Discussion

Paratuberculosis is a significant disease of dairy farming, associated with reduced productivity (McAloon et al., 2016), poor udder health (Pritchard et al., 2017; Rossi et al., 2017) and increased culling (Smith et al., 2010). Repeated testing of individual cow milk samples using an antibody ELISA is a common strategy in pTB control programmes. Difficulties in interpretation of serial test results could lead to misclassification of cows with regards to disease status, hindering pTB control. The novel approach presented here provides for a more nuanced, objective interpretation of serial milk ELISA test results, potentially improving detection of pTB and aiding control programmes.

The application of hidden Markov models to serial adjusted milk MAP antibody ELISA results has produced a model with four latent states. Given a series of milk ELISA test results, the model provides

the probabilities of an animal being in each of the four states at any timepoint. Each state is associated with a different distribution of S/P ratio (Figure 4.1). However, the state probabilities themselves are dependent on all previous test results up to the most recent test and are mediated through the emission and transition probabilities of the fitted HMM. Therefore, an animal may have a high probability of being in a state even if their most recent test result is not typical of that state. States 1 and 2 are the most prevalent starting states according to the initial state probabilities (probabilities of 0.5 and 0.4 respectively), and states 1 and 2 are similar to each other with regards to S/P ratios. States 3 and 4 are rather distinct however, both from states 1 and 2 and from each other. The transition probabilities of the HMM (Table 4.2), and the calculated relative risks of transitioning to state 4 from each of the other states (Table 4.3), show that cows are more likely to remain in states 1 or 2 than to transition to states 3 and 4, and a minority of animals reside in states 3 or 4 at any point. Moreover, current state is an indicator of the probability of progressing towards the highest states, with cows in state 3 for example having a considerably higher probability of progressing to state 4 than cows currently in lower states.

The states within the HMM are dependent primarily on the adjusted ELISA test results. In infected animals, the magnitude of antibody ELISA results is correlated somewhat with disease-progression (Beaver et al., 2017; Koets et al., 2015) and extent of faecal shedding of MAP (Beaver et al., 2017; Laurin et al., 2017; Nielsen and Toft, 2006; Sweeney et al., 2006). Therefore, it is probable that the HMM states are also correlated to some degree with disease progression and faecal shedding. Under this assumption, the states might

plausibly be interpreted as ordinal categories of disease progression, or of risk of transmission via faecal shedding of MAP. With increasing parity, the probability of transitioning away from states 1, 3 and 4 decreases. This would be expected if the ordinal HMM states are aligned with disease progression, as an animal's state should converge on either uninfected or late-stage disease with increasing age and number of tests. It is however acknowledged that the association between parity and transition probabilities may be confounded by culling. Correlations between HMM states and extent of faecal shedding, or probability of being infectious via other routes (*in utero*, or through production of MAP-contaminated milk) cannot be established with the data available in this study but could be a key area for future research.

The distribution of S/P ratios corresponding to state 4 (the state likely to represent the highest risk category) is largely above the conventional test-positive cut-point (S/P ≥ 30%), which favours specificity (Collins, 2002; Nielsen et al., 2002) and positive predictive value. Animals within state 4 therefore have the highest probability of being infected. States 1, 2 and 3 are predominantly associated with S/P ratios below the test-positive cut-point (S/P < 30%). S/P ratios below 30% would be commonly disregarded when dichotomising milk antibody ELISA results, however our results suggest that cows with S/P ratios below this threshold should not be treated as one uniform low-risk group. A distinct group of animals in state 3 which have not returned any positive (S/P ≥ 30%) milk ELISA results are substantially more likely to progress to state 4 in future than cows in states 1 and 2, and could be more infectious at present or in the future when compared with cows in states 1 or

2. Identification and management of state 3 cows, which have hitherto generally been disregarded, has the potential to further reduce within-herd MAP transmission rates on farms engaged in pTB control schemes. Notwithstanding their likely progression towards state 4, pending further evaluation of the likelihood that state 3 cows are currently infectious or infected, management of these cows should be limited to measures aimed at reducing the probability of MAP transmission to other animals, and should not include culling. Animals in state 1 or 2 appear to be the most likely to be free of disease and, especially with increasing age, are unlikely to ever transition to higher risk states or to rise above the traditional S/P threshold of ≥ 30%. Cows in state 1 or state 2 may represent lower risks with regards to breeding replacement animals or sources of colostrum.

Hidden Markov models have previously been applied to animal health data, for example to model herd-level freedom from infectious disease (van Schaik et al., 2023), and on pTB data in limited settings, for example for inferring common patterns of progression of faecal MAP shedding using comparatively small datasets (Ceres et al., 2020; Louzoun et al., 2015). To the author's knowledge the use of a HMM for clustering of cows according to serial milk ELISA results, used commonly in pTB control schemes, has not previously been attempted. Wang et al. (2011) fit a change-point model, which estimated the probability, given a series of ELISA and faecal MAP culture results, that a transition from a negative to a positive state had occurred in an individual animal. The change-point model included only two latent states (in contrast to the four states in our model) but, similar to our study, adjusted

S/P ratios for some covariates (yield and age). However, the wider implications of this model (e.g. the significance of the change-point in the context of disease progression, and the application of this model to large empirical datasets) were not explored.

Several studies (Eisenberg et al., 2015; Machado et al., 2018; McAloon et al., 2020; Nielsen et al., 2002) have investigated the associations between S/P ratios and milk yield, quality or SCC, and our findings are broadly in agreement with all such studies. The non-linear association of DIM with S/P ratio agrees with other research (McAloon et al., 2020; Nielsen et al., 2002), which also revealed higher S/P ratios in early and late lactation. This association has been postulated to be due to an increased concentration of non-pTB-specific immunoglobulin G (IgG) in colostrum, as well as concentration of IgG by low yields in early and late-lactation milk (Eisenberg et al., 2015). Of the covariates in the linear models, the association between MTNC and S/P ratio had the largest effect size. A change in MTNC from 2 to 6% S/P was associated with a difference of 5.84% and 6.88% S/P in individual cow test results in primi- and multiparous animals respectively. Inclusion of MTNC as a model covariate is a novel approach to account for variation in test results due to non-disease factors (e.g. variability between test kits or lab protocol, or variation due to a humoral immune response of uninfected animals exposed to environmental MAP), and is based on the assumption that uninfected animals would have relatively low S/P ratios. If MTNC variation was due mainly to test kit variability or variation of lab protocol, it would be expected that distribution of MTNC would vary little between farms (assuming random allocation of test kits to farms). Indeed, on over half of farms in the data, the

mean and median MTNC over all test dates lay between 3.0 and 3.5%, suggesting modest variation in MTNC between farms. However, further investigation of the usefulness of MTNC as an indicator of test-kit or lab protocol variability, and the exact range of S/P ratios which should be averaged to calculate the MTNC figure, will require further research.

This study used a sample of data from one milk-recording organization and from forty-seven farms. Whilst the farms included in this study are not atypical for dairy farms in the United Kingdom, and represent a variety of farming systems, the dataset is nevertheless a convenience sample. It is unknown to what extent it is representative of the wider population of dairy cows, in the United Kingdom or globally, and to what extent the pTB test results would differ if performed by a different laboratory. Geographical region, breed and laboratory have all been found to be associated with odds of testing positive by milk antibody ELISA (Barden et al., 2020). Therefore, the generalisability of this model has not been established, and comparative work using other, larger datasets would be important.

4.5. Conclusions

This work has provided novel insights into the interpretation of the results of serial milk pTB antibody ELISA tests. Application of hidden Markov models has identified four latent states, the highest and lowest of which appear to be broadly representative of cows which would also be considered high- and low-risk within existing pTB control schemes. An intermediate cluster (state 3) has been identified, associated predominantly with test results below the

test-positive threshold routinely used, and containing animals which are distinct from most test-negative cows and are at increased risk of progressing to the highest state. Presence in state 3 appears to represent an early warning of disease progression in an individual animal, and therefore identification of such animals could be used to inform pTB control programmes.

Chapter 5. Predicting paratuberculosis high-risk (J5) status in dairy cows using a novel method of test interpretation

In this chapter, predictive models are created which forecast the onset of "J5" status, a strategically important indication of pTB positivity widely adopted by the United Kingdom dairy industry.

5.1. Introduction

A novel method for interpretation of serial pTB test results was developed in Chapter 4. In this method, ELISA results are first adjusted for non-disease factors, then a hidden Markov model is used to allocate an animal to one of four latent states according to the series of adjusted results. The latent states are positively correlated with ELISA S/P ratios, which themselves are correlated with disease progression and faecal shedding of MAP (Beaver et al., 2017; Koets et al., 2015; Laurin et al., 2017; Nielsen and Toft, 2006; Sweeney et al., 2006). It is therefore hypothesised that the HMM states are indicative of disease progression or contagious potential.

Of practical use would be the ability to predict the probability that a cow will become a definite positive (for example, reach J5 status; Table 5.1) in the next 12 months. Such a prediction would allow farmers to manage high-risk cows at an earlier stage (for example to calve in an isolated environment or to alter breeding policy). Therefore, the aim of this study was to evaluate whether latent HMM ELISA states can be used to predict the probability of an individual cow being classified as J5 in the next 12 months.

Status	Definition
J0	The animal has had at least two negative test results and has not had a positive result within the last three tests.
J1	The animal has only had one test in its lifetime, and the result was negative.
J2	The animal had a positive test result 2 tests previous, but the most recent 2 test results have been negative.
J3	The previous test result was positive, but the most recent was negative.
J4	The most recent test result was positive, but the animal has not had another positive within the previous 3 tests.
J5	2 or more positive results within 4 tests at any point in the animal's history. Once an animal is deemed J5, it is not possible to revert to any other status.

Table 5.1: Paratuberculosis disease risk status (J-status) according to a classification system commonly used by United Kingdom Milk Recording Organisations.
Green = low risk, yellow = medium risk, red = high risk.

5.2. Materials and Methods

5.2.1. Data collation

Data were collated from a sample of herds that undertake milk-recording with Quality Milk Management Services (QMMS) LTD., UK. To be eligible for inclusion, herds were required to have undergone at least twenty whole-herd pTB tests during the ten-year period between 1st January 2011 and 31st December 2020 and to have conducted milk-recording on average at least four times per year during that period. A whole herd pTB test was defined as an event in which at least 70% of lactating animals were tested for pTB using

composite milk samples; concentrations of MAP antibodies in milk were measured using a commercially available ELISA (IDEXX B.V., Hoofddorp, The Netherlands). The data included all pTB tests conducted from 1st December 2010 up to and including 31st December 2021.

5.2.2. Data cleaning

Cleaning of data was undertaken to permit modelling and to remove erroneous measurements. Animals whose date of birth was unknown (and hence whose ages when tested for pTB were unknown) were removed from the data. A minimum realistic age at first calving was defined as twenty months, and rows of data with age less than this value were removed. Animals which were present in the data of more than one herd (due to movement between study herds during the period encapsulated in the data) were removed to ensure that herd-level variables were reliable. To ensure that the model was based on recent test results (see below), animals were excluded if the interval between any of their pTB tests exceeded 183 days. Animals with only one pTB test result were also excluded, as subsequent pTB tests were required to determine an animal's outcome (whether J5 status was reached within the subsequent twelve months). A flow diagram of the process of data-preparation is shown in Figure 5.1.

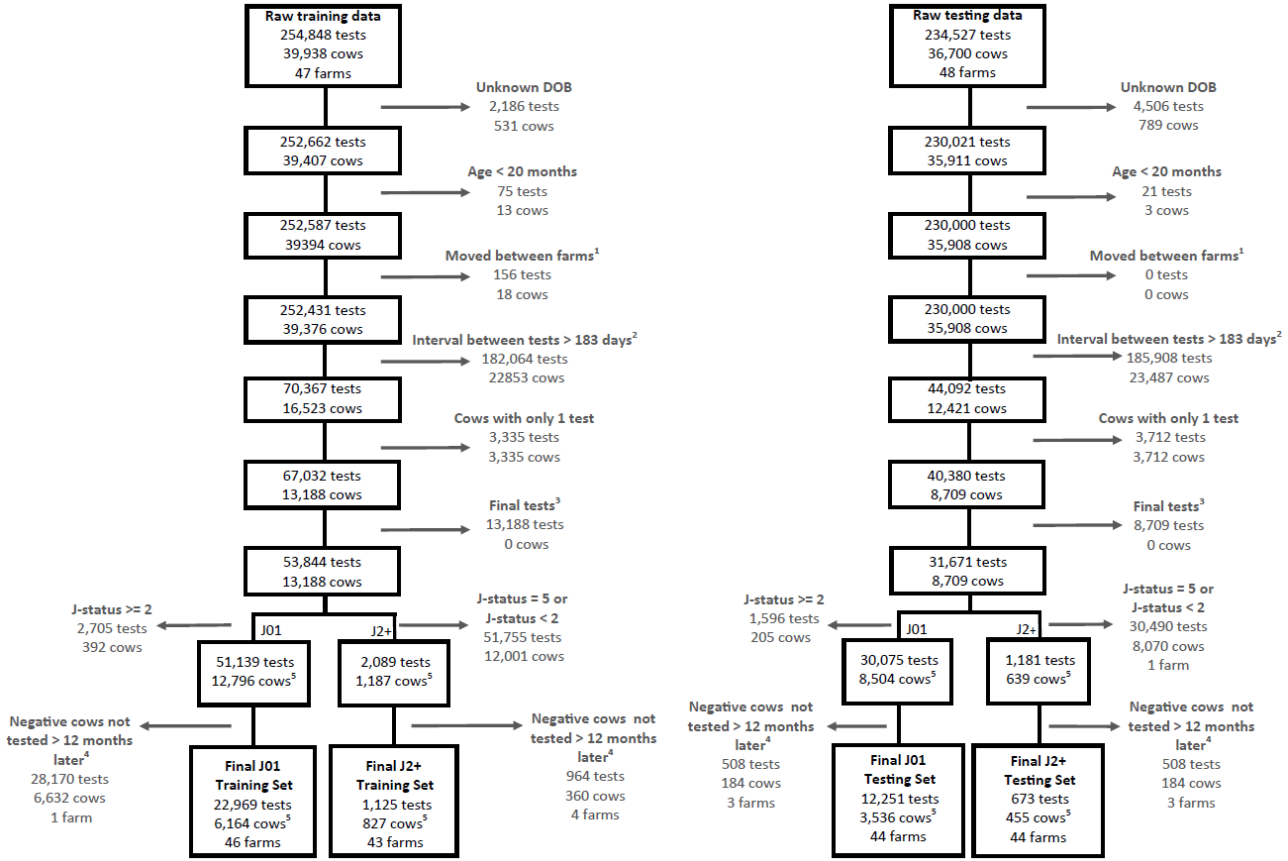

Figure 5.1: Flow diagram illustrating the process of preparing data for training and validating generalised linear mixed models for prediction of the onset of a positive J-status in individual dairy cows in the subsequent 12 months.

Arrows represent removal of the data described. Retained data is shown in boxes.

[1]All data for a cow were removed if she moved between study farms during the data-collection period.

[2]All data for a cow were removed if the interval between any two of her tests was greater than 183 days.

[3]To allow calculation of the dependent variable, rows of data were retained for a cow only if a subsequent test occurred.

[4]To train models predicting the outcome given survival and subsequent testing occurring over a period of greater than one year, rows were removed if a cow was not tested again during the subsequent 12-month period.

[5]Owing to fluctuation in individual animal J-statuses over time, data from individual cows could be present in both J01 and J2+ datasets.

5.2.3. Allocation of latent states

Through application of the method developed in Chapter 4, animals were assigned a pTB state at each testing occasion based on all the milk MAP antibody ELISA test results up to the current timepoint. Briefly, ELISA results, reported as S/P ratios, were adjusted to account for the non-disease factors, reported in Chapter 4, which influence the results. These non-disease factors were age, lactation-stage, milk yield, somatic cell count, milk sample concentrations of protein, and the mean of all results with an S/P ratio below 10% on the same herd test day. At each test date for each animal, given all adjusted S/P ratios for that animal up to and including the timepoint in question, the HMM was used to predict the probability of an animal being in each of four ordinal states, and to allocate the animal into the state with the highest probability.

5.2.4. Relationship between contemporary HMM states and J-statuses

Data were in long form, each row representing a timepoint at which timepoint according to a pTB-risk classification system commonly employed in the UK (Table 5.1) was calculated. Prior to creation of predictive models, the relationship between HMM states and contemporary J-statuses was assessed as follows. The herd with the fewest number of unique animals in the data was identified. To avoid bias due to clustering within animals and herds, the same number of unique animals were sampled at random from every herd, and from each of these animals one row of data (one test result) was then sampled at random. For each HMM state, the proportion of rows in the sampled data in which an animal was

concurrently in each of the six J-statuses was calculated and visualised in a bar chart.

5.2.5. Creation of predictive models

Generalised linear mixed models with a logit link were built to predict the probability of a cow becoming positive J5-status within 12 months, given her survival to the end of that period, and given that she was tested at least once every 183 days (twice yearly). Models were fit using the *lme4* package (Bates et al., 2015) in R version 4.2 (R Core Team, 2022). Farm was included as a random effect.

Positivity according to J-status was defined as follows. For cows that had not had a positive ELISA result (S/P ≥ 30%) within the most recent three tests (J0 or J1 cows), a positive outcome was the occurrence of the first of a pair of positive results (within four consecutive tests) which resulted in J5 status. For cows which had already had a positive result within the most recent three tests (J2, J3 or J4 cows), the positive outcome was the occurrence of the second of the pair of positive (S/P ≥ 30%) results. Thus, relevant outcomes were created for cows in all J-statuses.

The data were split into two subsets, containing either cows which had not received a positive result (J0 or J1 cows), or which had received a positive result (J2, J3 or J4 cows) within the most recent three tests (henceforth referred to as "J01" or "J2+" data). The binary dependent variable for the J01 data described the occurrence (or lack thereof) in the subsequent twelve months of the first of a pair of positive test results which resulted in an animal being classified as J5. The equivalent dependent variable for the J2+ data indicated

whether the second of a pair of positive results occurred in the next twelve months and therefore if the animal was classified as J5.

Data were limited to those rows in which the dependent variable could be defined. The value of the dependent variable could only be determined when there existed a subsequent pTB test for a cow. Therefore, the last row of data for each animal was not included in the training data, except insofar as was needed for defining the dependent variable at each timepoint. The dependent variable could only be defined as negative (positive status not occurring within 12 months) if cows were not lost to follow-up. Therefore, rows of data for cows not becoming positive within 12 months were retained only if the cow had at least one pTB test subsequent to the 12-month period. Both subsets included cows which had fewer than three previous test results.

Independent variables (Table 5.2) were of both time-varying and time-invariable types. Time-varying features were predominantly those describing HMM states, which change over time as cows are repeatedly tested using the ELISA. Time-invariant predictors were those describing the state of the dam near the time of calving.

To create a model with good predictive ability, models with different combinations of fixed effects were assessed using cross validation. Variables describing the three most recent HMM states for an animal were included in each of four different ways: 1) Three separate categorical predictors, describing the most-likely HMM state at each of the most recent three tests, and including a "missing" category when animals had had fewer than three previous tests, 2) the probabilities of the animal being in each of

HMM states 1 to 3 at the most recent test (the probability for state 4 was not included, as it is the complement of the sum of probabilities for states 1 to 3 and therefore its inclusion would cause collinearity), plus categorical predictors describing the most-likely HMM state at each of the two previous tests, 3) an interaction between the current most-likely HMM state and the probability of the animal being in that state, plus categorical predictors describing the most likely HMM state at the previous two tests, and 4) a single categorical predictor, each category representing a different permutation (including "missing") of possible HMM states at the most recent three tests.

Variable	Type	
Age	Time-varying	Categorical (20-50, 50-80, 80+)
Probability current state 1	Time-varying	Numeric
Probability current state 2	Time-varying	Numeric
Probability current state 3	Time-varying	Numeric
Probability current state 4	Time-varying	Numeric
Most likely Current state	Time-varying	Categorical (1,2,3 or 4)
Most likely state one test ago	Time-varying	Categorical (1,2,3,4 or Missing)
Most likely state two tests ago	Time-varying	Categorical (1,2,3,4 or Missing)
Ever state 3 three or more tests ago	Time-varying	Boolean (Yes/No)
Ever state 4 three or more tests ago	Time-varying	Boolean (Yes/No)
Proportion of herd currently state 4 (rolling 3 test average)	Time-varying	Numeric
Maximum periparturient dam state[1]	Time-invariant	Categorical (1,2,3,4 or Unknown)
Latest dam State prior to calving[2]	Time-invariant	Categorical (1,2,3,4 or Unknown)

Table 5.2: Predictor variables in a generalised linear mixed model predicting the probability of becoming positive J5 status within the next twelve months.
States and State probabilities are calculated by a hidden Markov model developed for interpretation of serial milk paratuberculosis antibody ELISA results in dairy cows (Chapter 4).
[1]The maximum (highest) State of the dam during the period between 12 months prior and 12 months subsequent to calving. "Unknown" if the dam was not tested during this period.
[2]The latest State of the dam within six months prior to calving. "Unknown" if the dam was not tested within the six months prior to calving.

Models using one of these four formulations, plus one of all possible combinations of the other fixed effects (the variables indicating whether the animal had ever previously been state 3 or 4, age category, the rolling proportion of the herd which were HMM state 4, and dam states around the time of calving) were assessed.

Cross validation was 10-fold, repeated 10 times. To prevent data-leakage, folds were created such that a farm could be present in either the training or testing data, but not in both. Weighted log loss was the loss function used to compare models. Weighted log loss assesses the calibration of the predicted probabilities of the logistic model whilst accounting for imbalance in the dependent variable. The selected combination of fixed effects was the simplest (fewest terms) which resulted in a cross-validated weighted log loss within one standard error of the minimum loss (Kuhn and Johnson, 2013). The weighted log loss was calculated as follows:

$$w * [o * ln(p) + (1 - o) * ln(1 - p)],$$

where o is the value (1 or 0) of the binary dependent variable, p is the predicted probability, and w is the weight, given by the prevalence, or complement of the prevalence as appropriate, of the positive dependent variable. Normality of random effects was checked through visualisation of histograms.

5.2.6. Evaluation of predictive models

A second 'test' dataset was collated for evaluation of model predictive performance. These data included any herds milk-recording with QMMS LTD. which had undertaken at least 10 whole-herd pTB tests over a five-year period between January 2016 and December 2020, and which were not included in the data used for

model training. Data were collated from these farms up to and including 13th May 2024. These testing data were processed in the same way as the model-training data, yielding the same variables, and were similarly split into two subsets (J01 and J2+) according to presence or absence of recent positive test results.

Predictive performance (calibration) of the two final models was assessed using the testing data. Probabilistic model predictions were made on the relevant testing data subsets (J01 or J2+), setting the random effects terms in the models to zero. Model calibration was assessed visually through calibration curves. The distribution of predicted probabilities for both models was visualised with histograms.

5.3. Results

5.3.1. Data cleaning

Prior to cleaning and splitting, the training data comprised 254,848 test results, for 39,938 cows from 47 farms. Cleaning caused removal of 187,816 rows, resulting in 67,032 rows in 13,188 cows on 47 farms. The testing data originally comprised 234,527 test results in 36,700 cows from 48 farms, and cleaning removed 194,147 rows, leaving 40,380 test results for 8,709 cows from 48 farms. The process of data preparation (including cleaning and further steps) is illustrated in Figure 5.1. Descriptive statistics of the final training and testing datasets are shown in Table 5.3.

5.3.2. Relationship between contemporary HMM states and J-statuses

The relationship between HMM states and J-statuses is shown in Figure 5.2. All the animals in HMM state 4 were concurrently either J4 or J5. Of animals in HMM states 1 and 2, 99.1% (2,511/2,533) and 97.7% (3,211/3,286) were concurrently either J0 or J1. The animals in HMM state 3 were more varied with regards to concurrent J-status; 64.4% (326/506), 15.4% (78/506), 2.6% (13/506), 3.6% (18/506), 8.1% (41/506) and 5.9% (30/506) were J0, J1, J2, J3, J4 and J5 respectively. Almost four-fifths (79.8%) of HMM state 3 cows were in the two lowest-risk categories (J0 or J1) according to J-status.

5.3.3. Model training

Use of the single categorical predictor describing each possible permutation of the most recent three HMM states was problematic when modelling the J2+ data; presence of positive results within the most recent three tests led to many diverse combinations of recent HMM states, resulting in a predictor variable with many categories. Therefore, models with this predictor were not fit to the J2+ data.

5.3.4. Model results

For both the J01 and J2+ models, the model with the best weighted log loss, and all models whose weighted log loss was within one standard error of the best, included as predictors the probabilities of the animal being in each of the HMM states at the most recent test, plus categorical predictors describing the most-likely HMM state at each of the two previous tests. Parameters of the two final models are shown in Table 5.4.

Variable		J01 Training	J01 Testing	J2+ Training	J2+ Testing
pTB tests *n*		22969	12251	1125	673
Cows *n*		6164	3536	827	455
Farms *n*		46	44	43	44
Test date	Min	21/12/10	28/06/12	21/12/10	18/07/11
	Max	22/02/21	12/07/23	09/07/21	08/11/23
Age (months)	Min	20.2	20.8	20.9	21.9
	Median	36.0	36.1	42.1	45.5
	Mean	40.4	43.2	45.7	50.8
	Max	155.8	164.6	146.8	160.9
Parity	Min	1.0	1.0	1.0	1.0
	Median	1.0	1.0	2.0	2.0
	Mean	1.6	1.9	1.9	2.4
	Max	10.0	10.0	10.0	9.0
Inter-test interval (days)	Min	20.0	22.0	28.0	22.0
	Median	96.0	98.0	99.0	106.0
	Mean	104.7	105.9	111.0	111.4
	Max	183.0	183.0	183.0	183.0
Most-likely pTB state *n* (%)	1	8871 (38.6)	4398 (35.9)	17 (1.5)	27 (4.0)
	2	12607 (54.9)	7162 (58.5)	197 (17.5)	132 (19.6)
	3	1491 (6.5)	691 (5.6)	340 (30.2)	201 (29.9)
	4	0	0	571 (50.8)	313 (46.5)
J status *n* (%)	1	16990 (74.0)	8749 (71.4)	0	0
	2	5979 (26.0)	3502 (28.6)	0	0
	3	0	0	152 (13.5)	103 (15.3)
	4	0	0	213 (18.9)	135 (20.1)
	5	0	0	760 (67.6)	435 (64.6)
Positive within 12 months *n* (%)	No	22115 (96.3)	11790 (96.2)	503 (44.7)	339 (50.4)
	Yes	854 (3.7)	461 (3.8)	622 (55.3)	334 (49.6)

Table 5.3: Description of the training and testing datasets used for creation of models predicting the probability of cows becoming J5 status.
The number of rows of data in each category is shown, with the percentage of rows in the data in brackets.

Variable		J01 Model			J2+ Model		
		Estimate	Standard Error	Odds Ratio (95% CI)	Estimate	Standard Error	Odds Ratio (95% CI)
Intercept		5.404	1.920		1.574	0.377	
Current probability state 1		-10.400	1.903	0 (0 - 0.001)	-4.127	1.034	0.016 (0.002 - 0.122)
Current probability state 2		-9.336	1.903	0 (0 - 0.004)	-4.524	0.366	0.011 (0.005 - 0.022)
Current probability state 3		-7.422	1.933	0.001 (0 - 0.026)	-2.180	0.259	0.113 (0.068 - 0.188)
Highest probability state last test	1	Referent					
	2	0.007	0.134	1.007 (0.774 - 1.311)	0.765	0.347	2.15 (1.089 - 4.244)
	3	0.017	0.179	1.017 (0.717 - 1.444)	1.338	0.354	3.812 (1.903 - 7.635)
	4				0.945	0.411	2.572 (1.149 - 5.758)
	Missing[1]	-0.193	0.137	0.824 (0.631 - 1.078)	0.016	0.365	1.016 (0.497 - 2.077)
Highest probability state 2 tests ago	1	Referent					
	2	0.080	0.135	1.083 (0.832 - 1.41)	-0.699	0.320	0.497 (0.265 - 0.931)
	3	-0.064	0.198	0.938 (0.636 - 1.384)	-0.370	0.366	0.691 (0.337 - 1.415)
	4				-0.024	0.428	0.976 (0.422 - 2.259)
	Missing[1]	-0.080	0.131	0.923 (0.714 - 1.195)	-0.516	0.335	0.597 (0.309 - 1.151)
Dam latest state[2]	1	Referent					
	2	-0.315	0.174	0.73 (0.519 - 1.026)			
	3	-0.448	0.300	0.639 (0.355 - 1.15)			
	4	0.049	0.388	1.05 (0.491 - 2.246)			
	Unknown	0.118	0.146	1.126 (0.846 - 1.497)			
Proportion of herd state 4		14.395	2.408	1.785×10^6 (1.591×10^4 – 2.003×10^8)			
Ever state 3 previously[3]					-0.057	0.279	0.945 (0.547 - 1.631)

Table 5.4: Parameters of two final models for prediction of the probability of positivity (J5) according to J-status.

[1]Previous highest probability states were designated "missing" if a cow had insufficient previous tests.

[2]The state of the dam at her most recent test preceding calving if this occurred within 6 months prior to the calving date. Otherwise "Unknown".

[3]Ever in state 3, up to and including 3 tests ago.

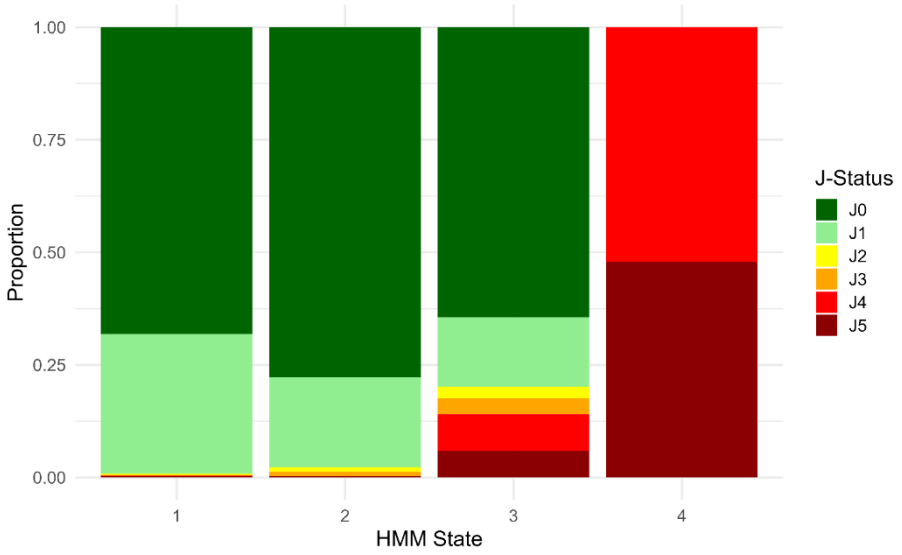

Figure 5.2: The proportion of sampled rows of data which were in each latent state of a hidden Markov model (HMM state) and concurrently in each of the six J-statuses.

In the J01 model, the probabilities of currently being in states 1, 2 and 3 were all negatively associated with the dependent variable. The negative coefficient for the state 1 probability was of greater magnitude than that for state 2, which in turn was of greater magnitude than state 3. After accounting for covariates in the model, a cow with a high probability of being in state 1 would therefore have a lower predicted probability of J5 status than a cow with a high probability of being in state 2 (and likewise with state 2 compared to state 3). In the J2+ model, the probabilities of being currently in states 1, 2 and 3 were again negatively associated with the probability of becoming J5. Of the three, the probability of state 2 had the largest effect, followed by state 1 and then state 3. Cows with a high probability of being in state 2 would have the lowest predicted probability of becoming J5, followed by cows with a high probability of being currently state 1, and then those with a high

probability of state 3. J2+ animals included cows which were concurrently HMM state 4. Since the probability of HMM state 4 is the complement of the sum of probabilities for the other three states, it can be said for J2+ cows that the probability of state 4 was positively associated with the probability of becoming J5. In the J2+ model, the state with the highest probability at the last (penultimate) test was associated with the probability of becoming J5, with cows previously in state 1 having a lower probability. Also, in the J2+ model, the probability of becoming J5 was reduced if two tests ago the cow was most likely to be in state 2 compared to state 1. The proportion of the herd currently in state 4 was highly positively associated with future J5 status for currently J0 or J1 cows.

5.3.5. Model validation

Histograms of predicted probabilities in the testing data are shown in Figure 5.3. The predicted probabilities of the J01 model were right-skewed (min 0.004, max 0.71, median 0.02, mean 0.03), with most predictions being low probability. In contrast, the predictions of the J2+ model (min 0.04, max 0.95, median 0.61, mean 0.54) were distributed more evenly throughout the probability range 0 to 1. Calibration plots are shown in Figure 5.4 and demonstrate satisfactory calibration of predicted probabilities for both models.

5.4. Discussion

The use of an unsupervised learning model (HMM) in Chapter 4 for objective interpretation of serial milk MAP antibody ELISA results, identified the potential for using HMM states in pTB control schemes. The current study compared the novel method with an

established method (the J-classification), by investigating the association between HMM state and current or future J Status.

The study in Chapter 4 identified potential advantages of the novel HMM classification with regards to pTB control. Cows can be identified which, despite never having received a positive (S/P ≥ 30%) ELISA result, are at higher risk of progression to the highest HMM state, which itself is characterised by positive (S/P ≥ 30%) test results. The current study has demonstrated that such cows (HMM state 3 cows) are often classified concurrently as low-risk (J0 or J1) according to J-status (Figure 5.2). The models presented in this study now provide well-calibrated probabilities, derived in part from the current HMM states, of cows progressing to J5 status during the subsequent 12 months, and show that HMM state 3 cows are at greater risk of progression to J5 than cows in states 1 or 2.

The existence in our data of cows which are considered relatively high-risk according to the HMM (state 3 cows) and yet are classified as low-risk according to J-status (J0 or J1) is unsurprising given the modest sensitivity (Nielsen and Toft, 2008) of the milk MAP antibody ELISA when dichotomising results using a standard test-positive threshold (e.g. S/P ≥ 30%). Cows can shed MAP in faeces despite giving negative ELISA test results (Laurin et al., 2017), raising the possibility that HMM state 3 cows are at greater risk of being contagious. Cows which repeatedly test positive (e.g. J5 cows) are most likely to shed MAP in faeces (Nielsen, 2008). However, for cows which do progress to positive ELISA results, faecal MAP shedding can occur prior to the first positive result (Nielsen and Toft, 2006; Sweeney et al., 2006) and therefore prior to J5 or even J4 status. The current models facilitate the identification of cows

which are likely to progress to J5 (and therefore are likely to shed MAP) at an earlier stage. The predicted probabilities of progression to J5 would be useful as a decision-support tool for a farmer when making management decisions in a pTB control programme. Lack of on-farm resources, for example space for segregating cows, has been identified as an obstacle in implementing pTB control (Morrison et al., 2024; Morrison and Rose, 2023). The ability to rank cows according to their probability of progressing to J5 status could be used to prioritise the allocation of resources to those cows with the highest probabilities. Furthermore, model predictions are for a period of twelve months, facilitating breeding decisions. For example, cows with a low probability of progression to J5 in the subsequent 12-month period could be considered safest for producing replacement dairy heifers.

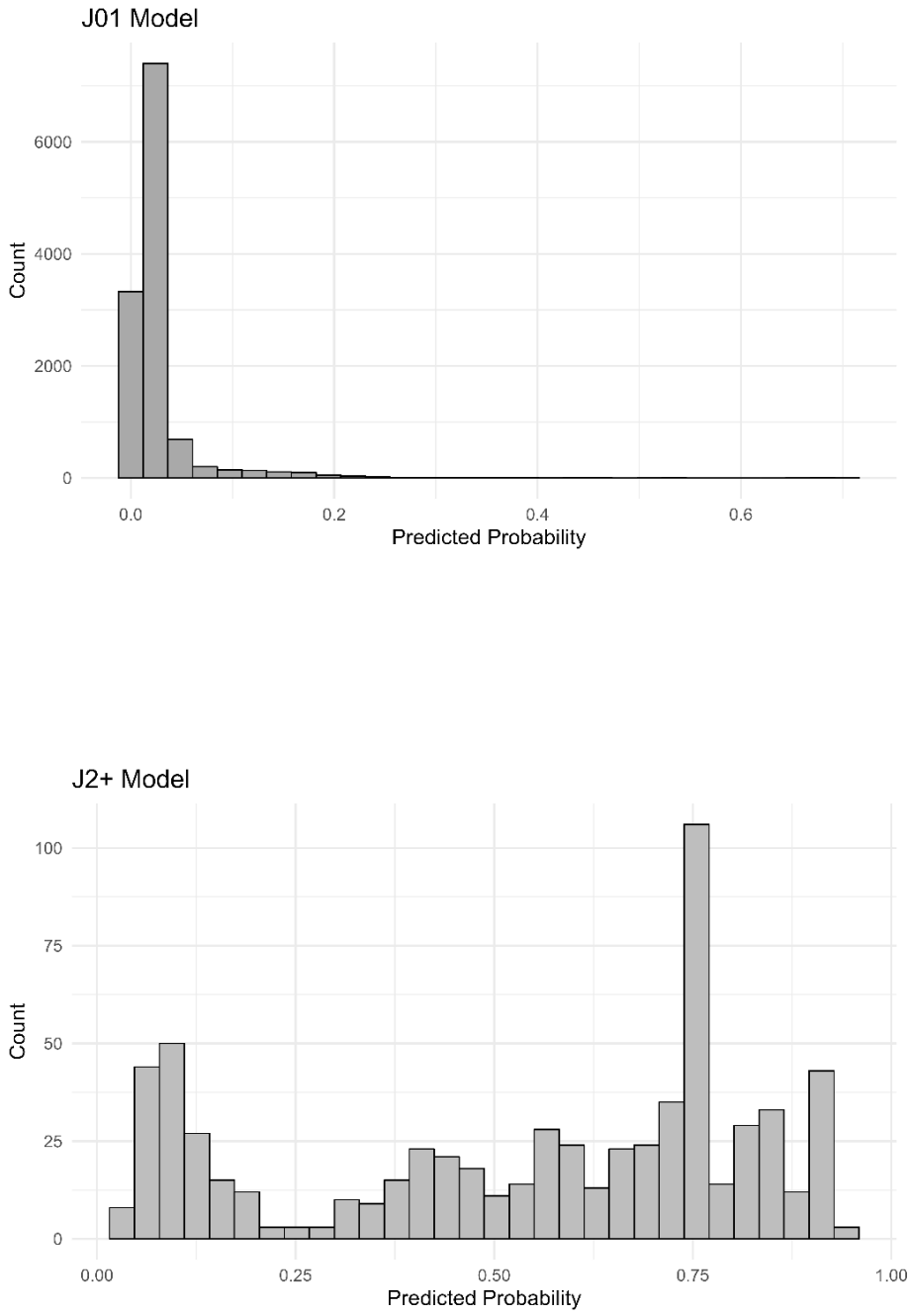

Figure 5.3: Histograms showing the distribution of predicted probabilities from the final models predicting positivity ("J5" status) for two groups of cows.
Top: cows with no positive (S/P ≥ 30%) MAP antibody ELISA results within the three most recent tests (J0 or J1 cows). Bottom: Cows with at least one positive MAP antibody ELISA result within three recent tests (J2, J3 or J4 cows).

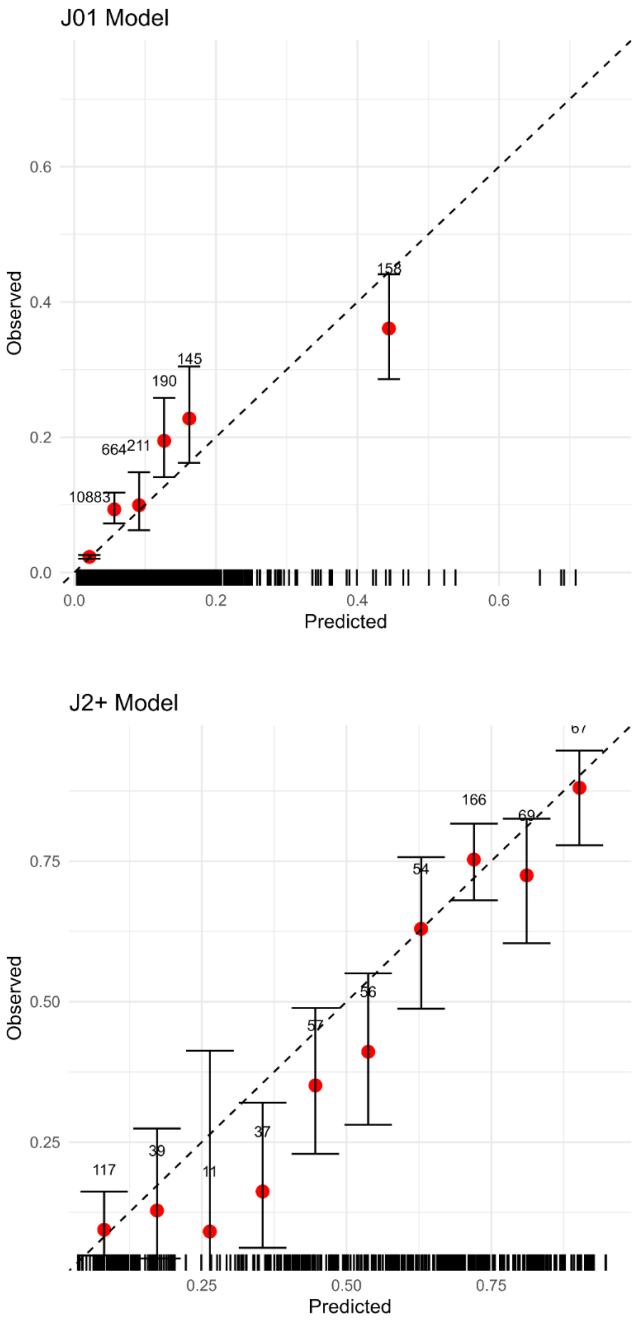

Figure 5.4: Calibration plots of the final models predicting positivity ("J5" status) for two groups of cows.

Predictions were divided into ten bins of equal probability interval. 95% confidence intervals are shown as error bars. The number of data points within each bin are shown. Top: Cows with no positive (S/P >= 30%) MAP antibody ELISA results within the three most recent tests (J0 or J1 cows). To aid visualisation, bins containing fewer than 100 data points were merged with the next (higher probability) bin. Bottom: Cows with at least one positive MAP antibody ELISA result within three recent tests (J2, J3 or J4 cows). The rug plots on the x-axes illustrate the densities of data points across the ranges of predicted probabilities.

Both final models utilise the current HMM state probabilities, and the best candidate models (loss function within one standard error of the best value) all included these probabilities, demonstrating the value of the probabilistic HMM output. Simply classifying cows into risk categories (e.g. J-status or QMMS Status), omits highly useful information contained in the numerical S/P ratio. For example, cows within one category may have considerably different risks as revealed by the HMM state probabilities. For the J01 model, cows with high probability of being in HMM state 1 had the lowest probability of progression to J5, followed in turn by HMM states 2 and 3. For the J2+ model, cows with a high probability of being in HMM state 4 have the highest probability of progression to J5, followed by cows with a high probability of state 3, then cows with a high probability of being in states 1 or 2. Interestingly, the J2+ model does not evidence a difference in risk between cows in states 1 or 2 (Table 5.4). This could be due to the sparsity of HMM state 1 cows in the J2+ data, reflected in the wide odds ratio confidence interval for the current probability of state 1.

The distribution of predicted probabilities of the J01 model were right-skewed, as expected (Figure 5.3). According to the low prevalence in the J01 data of cows which become J5 (Table 5.3), cows which have not had a recent positive test result become J5 only infrequently. In contrast, the predicted probabilities of the J2+ model were distributed more evenly.

Owing to imperfect test specificity (Nielsen and Toft, 2008), cows repeatedly tested with a MAP antibody ELISA can give occasional positive test results in the absence of faecal MAP shedding and subsequent positive test results (Sweeney et al., 2006). The J2+

model appears highly useful for differentiating such animals, following a positive test result, from those which are likely to go on to receive subsequent positive test results.

The ability to predict future MAP ELISA results from current or past results has been investigated previously. Cows giving a positive serum MAP ELISA result were more likely to receive a positive result again at their next test when compared with cows giving a negative result (Hirst et al., 2002). In the same study, for cows giving a positive ELISA result, there was a positive correlation between the S/P ratio of that result, and the likelihood of receiving a positive result at the subsequent test, corroborating the view that simple dichotomisation of test results into positive or negative categories ignores important information contained in the S/P ratio. Similarly, in another study (Imada et al., 2024), machine learning made use of previous test results to predict the next result in series of individual cow milk MAP antibody ELISA results. Again, numerical S/P ratios, in the form of the mean of recent test results, were found to be highly useful in predicting the next result. In contrast to the models created in the current study, however, prediction of one subsequent test result limits predictions to a short time-period and cannot be used to predict the onset of J5 status unless an animal has already received one positive test result. Additionally, the models presented by Imada *et al*. (2024) were used to predict the class (positive or negative) of the next result, but not for making probabilistic predictions. To the author's knowledge, probabilistic prediction of a cow's J-status, over a period as long as one year, has not previously been reported.

Several caveats should be considered. Farms included were a convenience sample of UK dairy herds which had conducted milk-recording and pTB testing through one laboratory. Choice of laboratory has been found to be associated with the probability of a positive (S/P ≥ 30%) MAP antibody ELISA result (Barden et al., 2020). Moreover, the data used for model validation was taken from the same laboratory (but different farms), and the testing set used for assessment of the J2+ was particularly small. Quantity of data was restricted by including only cows which were tested at intervals of a maximum of six months.

The random effects structure of the models is also noteworthy. A random effect of cow within herd would seem appropriate, but fitting models with such a structure proved impossible. Owing to imbalance in the dependent variable (Table 5.3) due to low prevalence of J5 cows, and the fact that some cows had few pTB tests, the dependent variable had only one value (1 or 0) for many animals in the data. Therefore, animal identification was strongly associated with the dependent variable, and models including a random effect of cow converged poorly, or not at all. Modelling of larger datasets might overcome this problem. Similar work including data from more farms, tested by different laboratories, and testing at different intervals, would be useful in further refining and establishing the generalisability of these models.

This study suggests that the HMM states provide a more sensitive or prompt method of detection of high-risk cows, potentially facilitating pTB control. Further research is now required to establish if state 3 cows are at significant risk of currently being

contagious (through faecal MAP shedding or other routes), in addition to being at elevated risk of progression to being contagious.

5.5. Conclusions

The probabilities of cows being in each of four states, according to a hidden Markov model used to interpret serial milk MAP antibody ELISA results, can be used to predict the probability of individual cows progressing to the highest category of risk according to the J-classification system commonly used in the United Kingdom pTB control programme. Such probabilities could be used by farmers when making rational decisions regarding management of individual cows to limit disease transmission.

Chapter 6. Production and fertility implications of the latent states revealed by a hidden Markov model for interpretation of serial paratuberculosis test results

In this chapter, inferential models are created to examine the associations between HMM state and milk production, and between state and the risk of service conception. Linear and logistic mixed-effects models are employed, as well as discrete-time survival models.

6.1. Introduction

Positive pTB test results have been reported to be associated with reduced milk production (McAloon et al., 2016), increased intramammary infections (Pritchard et al., 2017; Rossi et al., 2017), and impaired fertility (Ozsvari et al., 2020; Reynolds et al., 2023) amongst other effects, though these associations are not consistent across studies. For example, test-positivity has been reported to be not associated with fertility (Kennedy et al., 2016; Martins et al., 2024; VanLeeuwen et al., 2010), or even associated with improved fertility (Gonda et al., 2007; Marcé et al., 2009; Smith et al., 2010). The association between positive results and milk yield is the best-established (Griss et al., 2024; McAloon et al., 2016).

Speculative reasons for discrepant research results include variation of disease effects over the duration of this chronic disease (Aly et al., 2010; Beaudeau et al., 2007; Pritchard et al., 2017; Tiwari et al., 2007; Van Leeuwen et al., 2002), choice of diagnostic test

(Gonda et al., 2007; Hendrick et al., 2005; Johnson-Ifearulundu et al., 2000; Kennedy et al., 2016), survivorship bias (Smith et al., 2010) and misclassification bias, whereby infected animals are classified as non-infected, or *vice versa*.

Misclassification may occur for several reasons. One is that poor sensitivity of commonly used tests (Nielsen and Toft, 2008) is expected to result in low negative predictive value, and therefore presence of infected animals in negative control groups (Tiwari et al., 2007). Also, within-herd pTB prevalence is often low, generally below 15% (Whittington et al., 2019); combined with imperfect test specificity (Kennedy et al., 2016; Nielsen and Toft, 2008) this leads to mediocre positive predictive values and the presence of non-infected animals in experimental exposure groups. Other factors (described more fully in Chapter 1) potentially leading to misclassification include dichotomisation of the numerical ELISA result, and the confounding influence of non-disease factors (e.g. milk yield) on the ELISA result (McAloon et al., 2020).

A novel method of serial milk ELISA interpretation was developed in Chapter 4. Predictions from a hidden Markov model are used to allocate a cow to one of four states, according to all her milk ELISA results (S/P ratios adjusted for non-disease factors). The objective is to alleviate concerns regarding thresholds, non-disease factors and objective interpretation of serial tests. There is particular interest in one state (state 3), which represents cows whose S/P ratios tend to be less than the customary test-positive threshold but greater than typical negative results. Cows in state 3 have a relatively high probability of future positive results (Chapter 4), but work is required to fully establish if these cows are currently in an

early subclinical stage of pTB, which would suggest they should be managed to limit contagion. The aim of this study is therefore to assess if state 3 cows are less productive or less fertile, in a manner similar to test-positive cows.

6.2. Materials and Methods

6.2.1. Overview

Models were created to investigate the associations of the states predicted by the HMM developed in Chapter 4, with 305-day energy-corrected milk production, time to first service, time to conception, and odds of conception given a serve event. For the 305-day energy-corrected yield models, the variable of interest was the highest HMM state reached by a cow during the lactation in question. For the time to first service, time to conception, and odds of conception models, the variable of interest was the cow's HMM state at the latest pTB test. Due to their similarities with regards to S/P ratios (Figure 4.1), states 1 and 2 were merged for all analyses.

6.2.2. Eligibility

Data pertaining to milk production, pTB testing and fertility for individual cows were retrieved from commercial farm data analysis software (TotalVet Version 3, Quality Milk Management Services LTD., Wells, Somerset, UK). Herds were eligible for inclusion if they had conducted regular whole-herd paratuberculosis screening and milk recording through QMMS LTD. Datasets with particular inclusion criteria had been collated for a previous study (Chapter 5); the first dataset, used for training predictive models, included herds which had conducted at least 20 whole-herd pTB tests during the ten-year period between January 2011 and December 2020. A

second dataset had been compiled for testing the predictive models and included additional herds which had conducted at least 10 whole-herd pTB tests between January 2016 and December 2020. In addition, for inclusion herds were required to have conducted milk-recording on average at least four-times per year during the relevant period. Availability of a hold-out model-testing dataset was not a requirement for the current study, and so these two datasets were combined to maximise the amount of data available for training inferential models. A whole herd pTB test was defined as an occasion on which at least 70% of lactating animals were tested. pTB tests on individual cows were conducted using an enzyme-linked immunosorbent assay (ELISA, IDEXX B.V., Hoofddorp, The Netherlands).

Cow lactations were selected for inclusion in the data based on several criteria. For the 305-day yield models, if a cow had fewer than two pTB tests by 305 days of any lactation, then this lactation and all subsequent lactations for that cow were excluded. This ensured that included lactations were of at least 305 days duration and that there had been sufficient pTB testing to properly inform the HMM states over the cow's lifetime up to and including day 305 of the last included lactation for each cow. For fertility models, although sufficient historic pTB testing over the cow's lifetime was still required for determining HMM states with reasonable confidence, service or conception events were expected to routinely occur during early lactation, so modelling the association between HMM states and fertility was not dependent on the number of pTB tests within the same lactation. Therefore, for all fertility models, consecutive lactations containing two pTB tests prior to

305 DIM were included in the same way as for the 305-day yield models, but in addition one extra subsequent lactation was included regardless of the number of pTB tests occurring during that lactation. To ensure that HMM states were still reasonably current for the fertility models, a restriction was set on the maximum permitted interval from the latest pTB test to a service or conception, as described later.

6.2.3. Data preparation

For the 305-day energy-corrected yield models, the experimental unit was a lactation in an individual cow. All eligible multiparous lactations were included initially in the data. Modelling of data from solely multiparous cows allowed inclusion as a covariate of the previous lactation 305-day energy-corrected milk, which it was anticipated would be correlated with the yield of the current lactation (the dependent variable). Production metrics (yield, somatic cell count, % protein and % butterfat) were aggregated over 305 days using the test interval method (Sargent et al., 1968). Rows of data were removed if production metrics were missing or were implausible, using as a guide ranges recommended by the International Council for Animal Recording (ICAR, 2014) as follows. Given a minimum and maximum plausible daily yield of 3 and less than 100 litres of milk respectively, lactations were excluded if 305-day yield was less than 915 (305 * 3) or greater than or equal to 30,500 (305 * 100) litres. Lactations were removed if the average protein concentration was less than 1% or greater than 6%, or if butterfat concentration was less than 1.5% or greater than 9%. Lactations were also excluded if the cow's age at the start of lactation was less than 25 months, as age at second calving of less

than 25 months was deemed unrealistic. Finally, lactations in which cows conceived before 14 days-in-milk were considered implausible and were removed from the data.

For investigation of the association between the time-dependent HMM state and time to first service and conception, data from all eligible farms were prepared in long form. Each row of the data represented an individual day within a cow's lactation. All variables were treated as time-dependent, so their values could change on any given day. The binary dependent variable for the time to first service model indicated if the cow had been served on that day, and for the time to conception model, if the cow had conceived on that day. Conception was deemed to have occurred if a cow was subsequently diagnosed pregnant or, in the absence of a pregnancy diagnosis, if the cow failed to return to service within 50 days or if the service occurred within a plausible window (283 ± 24 days) prior to calving; if multiple serves occurred within this plausible window then the successful serve was deemed to be that closest to 283 days prior to calving.

Data from all eligible farms (including all cows regardless of eligibility) were initially cleaned by detecting and removing lactations within periods characterised by poor on-farm recording of fertility events (services, pregnancy diagnoses or calvings) (Reynolds and Hudson, 2023; Reynolds et al., 2023). The aim of this process was to minimise bias introduced by poor data quality. For each year in each farm, certain indicators of fertility data quality were calculated; proportion of lactations beginning in that year in which there was at least one serve event recorded; mean conception rate for all serves occurring during that year;

proportions of lactations in which the 365th day of lactation (the expected day of re-calving in a cow with a 365-day calving interval) fell within that year and which had a prospective calving interval (from the beginning of lactation to the subsequent calving) of less than 300 days or of greater than 600 days; the proportion of serves occurring within that year which had an unknown outcome (uncertain if conception occurred); and the proportion of lactations with a plausible service (283 ± 24 days prior to calving) occurring during that year. Lactations were removed from the data if they commenced during a year in which the proportion of lactations with at least one serve event was zero, if they included a serve event in a year in which the mean conception rate was less than 10% or greater than 65%, if the 365th day of lactation occurred in a year in which either the percentage of lactations with calving intervals less than 300 days or greater than 600 days was greater than 3% and 6% respectively, if a serve event occurred within a year in which the percentage of serves with an unknown outcome was greater than 5%, or if a serve event occurred in a year in which the percentage of lactations with a plausible service date was less than 90%. Thus, lactations were removed if they included data occurring in a year in which fertility event-recording was poor.

The above process excluded data which originated from periods on farms when fertility event recording was poor but did not address individual lactations with poor data quality during periods of overall satisfactory fertility event recording on farms. Such lactations were now detected and removed. Data were first limited to eligible cow-lactations (based on the lactation-level inclusion criteria described above). For the time to conception, and odds of conception models,

lactations were removed if it could not be determined if the cow had conceived at a service or not (if no pregnancy diagnosis was conducted and the cow was removed from the herd before 50 days had elapsed). Lactations were also removed if the age at calving was unknown or was implausible (less than 20 or 25 months, for primi- and multiparous cows respectively). Examination of histograms revealed that service or conception events were infrequent prior to 21 DIM and after 300 DIM. Therefore, it was decided to model the risk of service or conception within the period between these two lactation stages only, and to this end lactations were removed from the data for training the time to first service and time to conception models, if first service or conception respectively occurred prior to 21 DIM, and for lactations that remained in the data, rows with DIM less than 21 or greater than 300 were removed. Rows prior to the first pTB test or milk-recording of an animal's life, or rows in which the interval since the most recent pTB test or milk recording was greater than 93 or 31 days respectively were removed. This helped ensure that information regarding HMM states and milk production of cows was not overly historic. Rows were also removed if the most recent milk-recording occurred during the previous lactation, as milk production at the end of the previous lactation was not expected to be informative of a cow's current production level. For modelling the time to first service and conception respectively, rows subsequent to the first service or first conception in each lactation were removed, since cows were no longer eligible for first service or first conception. Rows with values of milk-recording variables (yield, protein or

butterfat) that were missing or outside the ICAR-based ranges previously described were also removed.

For the model assessing odds of conception given a service event, the time to conception dataset was then limited to rows (days) on which a serve event occurred. These data contained all the time-dependent variables used in the survival models; values of these variables therefore represented parameters (e.g. HMM state, yield) for a cow at her most recent pTB test or milk recording, respectively within 93 and 31 days prior to each service event.

6.2.4. Energy-corrected yield

For all models, energy-corrected milk (ECM) yield was calculated according to ICAR recommendations (ICAR, 2023), using the following formula;

$$ECM = \frac{fat\ yield\ (Kg) * 38.3 + protein\ yield\ (Kg) * 24.2 + milk\ yield\ (Kg) * 0.7832}{3.14}$$

where fat yield, protein yield and milk yield refer to production on the relevant test-day (for all fertility models) or aggregated over 305 days (for the 305 energy-corrected yield model) as appropriate.

6.3. Statistical analysis

Generalised linear mixed models were created using the *lme4* package in R version 4.2 (R Core Team, 2022). For the 305-day energy-corrected yield, the outcome variable was modelled as a Gaussian distribution in a linear regression, and for the fertility survival outcomes (first service or first conception) a discrete time survival analysis was conducted using binary logistic regression (logit link). For the odds of conception model, a binary logistic regression (logit link) was also used. Effect sizes were considered to

be significant with a P-value of less than 0.05. Confidence intervals and P-values were calculated in the *lmertest* R package using Satterthwaite's method of approximating the degrees of freedom in mixed models. All numerical variables were centred around their mean. Model assumptions, model residuals and normality of random-effect distributions were confirmed as is routine.

6.3.1. 305-day energy-corrected yield model

The dependent variable was 305-day energy-corrected milk yield. Best subset variable selection was conducted by fitting a model for each possible combination of predictors, and selecting the model which minimised BIC. Two predictors were forced into models regardless: the highest state reached by the cow during the current lactation, and the highest state reached by the cow at any point prior to the start of the current lactation. Covariates offered to the models were parity (1, 2, 3, 4 or 5+), age at the start of lactation (25-45, 45-70, 70-95, 95-120 and 120-145 months), the natural log of the mean somatic cell count over all milk recordings during this lactation, days-in-milk at conception (14-50, 51-100, 101-150,151-200, 201-250, 251-305, 306-500, 500+ days and No Conception), and the previous lactation energy-corrected yield (1000 – 6000, 6000 – 8500, 8500 – 11000, 11000 – 13500, 13500 – 23500 litres and missing). Random effects of cow, farm, and year nested within farm were included. To check that confounding variables had not been left out of the model, after the best subset of variables was selected, any excluded variables were re-introduced into the model and were retained if their inclusion resulted in at least a 10% change in the coefficients of one or more of the variables of interest (those describing the HMM state of the cow). Biologically plausible two-

way interactions were tested, and retained if their effect improved BIC.

6.3.2. Time to first service and conception survival models

The dependent variable was the occurrence, or lack thereof, on a particular day of a fertility event of interest (first service or first conception). Initially random effects of farm, cow and year within farm were included. However, the random effect of cow had zero variance and so was dropped from the models, leaving random effects of farm, and year nested in farm. Fitting a model for every combination of predictor variables was unfeasible given the size of the datasets, and therefore fixed effect variables were chosen by backwards selection, beginning with a model containing all predictors. Variables were removed if their exclusion resulted in an improved BIC value, and if confounding (≥ 10% change in a coefficient of the variable of interest) was not detected. Time-dependent fixed effects offered to the models were DIM, the HMM state of the cow at the most recent pTB test (within 93 days), the energy corrected yield and the natural logarithm of the somatic cell count of the cow at the most recent milk-recording (within 31 days), age at the start of the lactation, parity (1,2,3,4 and 5+) and a variable indicating if the cow had ever had a positive pTB test result (S/P ≥ 30%) prior to the most recent pTB test. The last variable was included to account for a likely reluctance of farmers to serve cows which had previously received a positive test result. The variables DIM and HMM state were forced into models regardless, and to permit a non-linear association between DIM and the odds of service or conception, quadratic and cubic transformations of DIM were also forced into all models. Biologically plausible interactions

were tested and retained if their inclusion resulted in improved BIC. The hazard functions for the first service and first conception models were examined as follows. For each model, the marginal effect of days-in-milk (the probability of first service or conception over time) was calculated by making predictions at varied values of days-in-milk and HMM state, and fixed values of all other predictor variables; fixed values were the mean and reference level for numerical and categorical predictors respectively. These predictions were converted to survival probabilities as follows:

$$S_t = \prod_{0}^{t} 1 - P_{it}$$

where S_t is the probability that the fertility event (first service or first conception) has occurred by t days-in-milk, and P is the probability of first service or conception at t days-in-milk, in a cow in HMM state i. Survival functions of both models were visualised by plotting the resulting survival probabilities by days-in-milk. The median interval from calving to first service or conception, for each HMM state, was then extracted.

6.3.3. Odds of conception model

The binary dependent variable indicated if a service resulted in a conception. Models were fit with random effects of farm, year nested within farm, and cow. HMM state at the most recent pTB test was forced into all models, and the set of candidate covariates was the same as those offered to the survival models. The best subset of covariates was selected from all possible combinations as that which optimised BIC whilst retaining confounders whose inclusion resulted in at least a 10% change in the coefficients of HMM state.

Plausible two-way interactions were included if they resulted in improvement of BIC.

6.4. Results

6.4.1. Descriptive statistics

The data used for training each of the four models are described in Tables 6.1 to 6.3.

6.4.2. 305-day energy-corrected yield model

Parameters of the final 305-day energy corrected yield model are shown in Table 6.4. Cows whose highest HMM state during a lactation was 3 or 4 respectively gave on average approximately 53 and 99 litres less energy corrected milk over 305 days when compared to cows which remained in states 1 or 2. In addition, cows which had reached state 4 in previous lactations gave approximately 108 litres less energy-corrected milk. Parity and the 305-day energy-corrected yield in the previous lactation were significantly associated with yield, and there was a statistically significant interaction between these two (Figure 6.1). This interaction revealed that the positive association between previous lactation yield and current lactation yield was most pronounced in cows that are currently in their second parity. Age at the start of lactation, days in milk at conception, and 305-day mean somatic cell count, were also associated significantly with yield.

Feature	Stratum	n	%	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum
Herds		91							
Cows		18432							
Parities		27681							
pTB Tests		82583							
305-day yield current lactation (litres)				921.6	7344.5	9245.5	9436	11340.6	28984
Energy-corrected 305-day yield current lactation (litres)				962.6	7707.6	9435.7	9589	11309.2	29213.7
305-day protein (%)				1.325	3.232	3.428	3.468	3.666	5.983
305-day butterfat (%)				1.552	3.644	4.171	4.207	4.705	8.96
Mean SCC (x10^3 cells/ml)				2.8	39	76	159.6	166.5	3987.2
Year				2011	2016	2018	2018	2020	2023
DIM at conception (days)	[14 - 50]	228	0.82%						
	(50-100]	8917	32.21%						
	(100 - 150]	9114	32.93%						
	(150 - 200]	3345	12.08%						
	(200 - 250]	1691	6.11%						
	(250 - 305]	847	3.06%						
	> 305	452	1.63%						
	No conception	3087	11.15%						
Age at lactation start (months)	[20 - 45]	13249	47.86%						
	(45 - 70]	11474	41.45%						
	(70 - 95]	2587	9.35%						
	(95 - 145]	371	1.34%						
Parity	2	14438	52.16%						
	3	7336	26.50%						
	4	3395	12.26%						
	5+	2512	9.07%						
Energy-corrected 305-day yield previous lactation (litres)	[1000 - 6000]	1967	7.11%						
	(6000 - 8500]	6357	22.97%						
	(8500 - 11000]	5542	20.02%						
	(11000 - 13500]	2235	8.07%						
	(13500 - 23500]	535	1.93%						
	Missing	11045	39.90%						
Highest HMM State current lactation	<= 2	22477	81.20%						
	3	4083	14.75%						
	4	1121	4.05%						
Highest HMM State all previous lactations	<= 2	22476	81.20%						
	3	4236	15.30%						
	4	969	3.50%						

Table 6.1: Descriptive statistics of the dataset used for training regression models investigating the association of HMM States with 305-day energy-corrected milk yield.

Feature	Stratum	First service model								First Conception model							
		n	%	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum	n	%	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum
Herds		81								81							
Cows		27,761								31,326							
Parities		49,153								59,833							
Days		1,584,214								3,053,586							
pTB Tests		55,307								79,572							
Milk Recordings		77,272								132,961							
DIM				21	43	63	77.9	89	300			21	59	88	105.8	137	300
Yield (litres)				3.0	24.2	30.4	31.7	38.1	99.0			3.0	24.8	31.5	32.5	39.3	99.0
Energy-corrected Yield (litres)				2.4	24.3	30.5	31.7	37.9	132.0			2.4	25.0	31.4	32.4	38.6	132.0
Protein (%)				1.2	3.1	3.3	3.4	3.6	6.0			1.2	3.1	3.3	3.4	3.6	6.0
Butterfat (%)				1.5	3.4	4.1	4.1	4.8	9.0			1.5	3.4	4.0	4.1	4.7	9.0
SCC ($x10^3$ cells/ml)				1	16	32	156.5	84	18,494			1	16	32	149.9	84	18,494
Age at start of lactation (months)				20.1	27.5	38.6	44.1	54.3	144.2			20.1	27.0	38.0	43.0	52.2	144.2
Days since last pTB test				0	9	20	26.0	35	93			0	12	25	32.0	50	93
Days since last milk-recording				0	6	13	13.6	21	31			0	6	13	13.9	21	31
Year				2011	2017	2019	2018	2020	2024			2011	2017	2019	2018	2020	2024
Parity (rows)	1	530,669	33.5%							1,070,741	35.1%						
	2	438,572	27.7%							883,373	28.9%						
	3	284,737	18.0%							543,215	17.8%						
	4	168,104	10.6%							298,460	9.8%						
	5+	162,132	10.2%							257,797	8.4%						
HMM State (rows)	<= 2	1,407,300	88.8%							2,752,181	90.1%						
	3	135,111	8.5%							239,123	7.8%						
	4	41,803	2.6%							62,282	2.0%						
First Service (rows)	0	1,552,560	98.0%							-	-						
	1	31,654	2.0%							-	-						
First Conception (rows)	0	-	-							3,021,241	98.9%						
	1	-	-							32,345	1.1%						

Table 6.2: Descriptive statistics of the datasets used for training regression models in discrete time survival analyses investigating the association of HMM States with the risk of first service and first conception.

Feature	Stratum	n	%	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum
Herds		78							
Cows		25,842							
Parities		44,009							
Serve events		74,472							
DIM				21	72	94	107.0	129	300
Yield (litres)				3.0	25.9	32.6	33.6	40.3	97.5
Energy-corrected Yield (litres)				2.4	25.9	32.1	33.1	39.3	99.2
Protein (%)				1.2	3.1	3.3	3.3	3.5	5.9
Butterfat (%)				1.5	3.3	4.0	4.0	4.7	9.0
SCC (x10^3 cells/ml)				1	15	29	125.8	72	1812
Age at start of lactation (months)				20.1	25.9	36.7	40.4	49.0	141.4
Days since last pTB test				0	15	28	34.5	53	93
Days since last milk-recording				0	7	14	14.4	22	31
Year				2011	2016	2019	2018	2020	2024
Parity (rows)	1	29,294	39.3						
	2	22,165	29.8						
	3	12,436	16.7						
	4	6,024	8.1						
	5+	4,553	6.1						
HMM State (rows)	≤2	68,272	91.7						
	3	5,120	6.9						
	4	1,080	1.5						
Conceived	0	42,596	57.2						
	1	31,876	42.8						

Table 6.3: Descriptive statistics of the dataset used for training a model to investigate the association between HMM state and odds of conception, given that a service event has occurred.

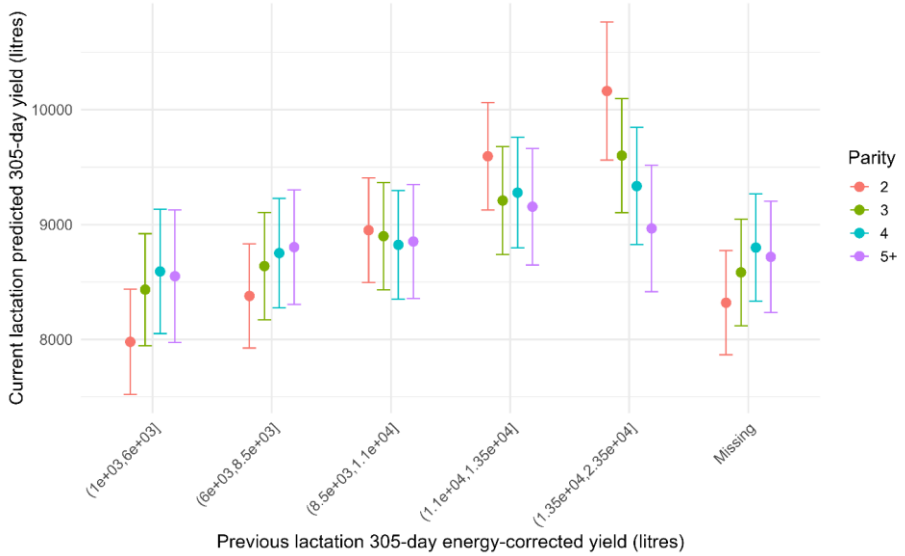

Figure 6.1: Marginal effect of the interaction between the fixed effects of parity and previous lactation 305-day energy-corrected milk yield from a linear mixed effects model.

Variable	Stratum	Estimate	Lower 95% C.I.	Upper 95% C.I.	P-value
Intercept		7979.32	7520.84	8437.80	<2.00E-16
Highest HMM State current lactation	<= 2	Referent			
	3	-52.85	-102.70	-3.00	**0.037733**
	4	-99.29	-187.88	-10.70	**0.028048**
Highest HMM State previous lactations	<= 2	Referent			
	3	-21.31	-73.76	31.14	0.425873
	4	-107.80	-209.81	-5.79	**0.038341**
Parity	2	Referent			
	3	454.68	265.25	644.10	**2.56E-06**
	4	612.89	310.59	915.18	**7.10E-05**
	5+	571.07	210.62	931.53	**0.001904**
Age at start of lactation (months)	<= 45	Referent			
	(45 - 70]	309.58	223.23	395.93	**2.16E-12**
	(70 - 95]	338.51	185.00	492.01	**1.55E-05**
	(95 - 145]	101.86	-113.21	316.93	0.353282
Natural log 305-day mean SCC		-174.66	-192.73	-156.60	**<2.00E-16**
305-day EC[1] yield previous lactation	<= 6e+03	Referent			
	(6e+03 - 8.5e+03]	399.31	307.66	490.95	**<2.00E-16**
	(8.5e+03 - 1.1e+04]	971.61	869.45	1073.76	**<2.00E-16**
	(1.1e+04 - 1.35e+04]	1615.28	1467.28	1763.27	**<2.00E-16**
	(1.35e+04 - 2.35e+04]	2182.93	1777.49	2588.36	**<2.00E-16**
	Missing	340.45	250.62	430.28	**1.13E-13**
DIM at conception	<= 50	Referent			
	(50 - 100]	99.23	-86.90	285.37	0.296071
	(100 - 150]	118.82	-67.27	304.91	0.210757
	(150 - 200]	186.60	-3.36	376.56	0.0542
	(200 - 250]	291.08	94.55	487.61	**0.0037**
	(250 - 305]	368.94	161.18	576.71	**0.000501**
	> 305	523.94	295.96	751.91	**6.68E-06**
	No conception	6.38	-184.74	197.50	0.947831
Parity 3 * 305-day EC[1] yield previous lactation	(6e+03 - 8.5e+03]	-194.59	-387.44	-1.73	**0.047991**
Parity 4 * 305-day EC yield previous lactation	(6e+03 - 8.5e+03]	-239.91	-556.32	76.51	0.137273
Parity 5 * 305-day EC yield previous lactation	(6e+03 - 8.5e+03]	-146.13	-498.05	205.79	0.415741
Parity 3 * 305-day EC yield previous lactation	(8.5e+03 - 1.1e+04]	-506.78	-700.68	-312.88	**3.04E-07**
Parity 4 * 305-day EC yield previous lactation	(8.5e+03 - 1.1e+04]	-740.36	-1054.19	-426.53	**3.79E-06**
Parity 5 * 305-day EC yield previous lactation	(8.5e+03 - 1.1e+04]	-669.12	-1024.87	-313.36	**0.000228**
Parity 3 * 305-day EC yield previous lactation	(1.1e+04 - 1.35e+04]	-840.05	-1066.24	-613.85	**3.47E-13**
Parity 4 * 305-day EC yield previous lactation	(1.1e+04 - 1.35e+04]	-929.56	-1269.19	-589.93	**8.21E-08**
Parity 5 * 305-day EC yield previous lactation	(1.1e+04 - 1.35e+04]	-1009.38	-1394.41	-624.35	**2.80E-07**
Parity 3 * 305-day EC yield previous lactation	(1.35e+04 - 2.35e+04]	-1016.41	-1480.01	-552.81	**1.74E-05**
Parity 4 * 305-day EC yield previous lactation	(1.35e+04 - 2.35e+04]	-1439.92	-1972.85	-907.00	**1.19E-07**
Parity 5 * 305-day EC yield previous lactation	(1.35e+04 - 2.35e+04]	-1766.97	-2343.88	-1190.05	**1.96E-09**
Parity 3 * 305-day EC yield previous lactation	Missing	-190.60	-376.13	-5.06	**0.044077**
Parity 4 * 305-day EC yield previous lactation	Missing	-133.23	-435.24	168.78	0.387254
Parity 5 * 305-day EC yield previous lactation	Missing	-172.13	-508.87	164.61	0.316416

Table 6.4: Parameters of the fixed effects of the final model of the association of HMM States with 305-day energy corrected yield.
EC yield = Energy-corrected yield.

6.4.3. Time to first service model

The parameters of the fixed effects of the final model are shown in Table 6.5. When compared with HMM states 1 and 2, state 4 was negatively associated with odds of first service (O.R = 0.775, 95% C.I. = 0.756 – 0.794). The time to first service for cows in state 3 was not significantly different to state 1 and 2 (O.R = 0.973, 95% C.I. = 0.885 – 1.062). The association between HMM state and time to first service was confounded by the variable indicating if the cow had ever previously received a positive test result. Exclusion of this variable led to a stronger negative association between HMM states 3 and 4 and odds of first service, and statistical significance ($P < 0.05$) of the association between state 3 and odds of first service. Energy-corrected yield, somatic cell count, and age at the start of lactation were all significantly associated with time to first service. The hazard and survival functions are shown in Figure 6.2 and Figure 6.3. The median intervals from calving to first service (and 95% confidence intervals) predicted by the model were 74 (71 - 77), 74 (71 - 78) and 80 (76 - 85) days for cows in HMM states 1 or 2, 3, and 4, respectively.

Variable	Stratum	Estimate	Standard error	Odds ratio	O.R. Lower 95% C.I.	O.R. Upper 95% C.I.	P-value
Intercept		-6.778	0.082				**<2.00E-16**
Days-in-milk		-0.031	0.001	0.969	0.968	0.971	**<2.00E-16**
Days-in-milk^2		0.031	0.001	1.032	1.030	1.034	**<2.00E-16**
Days-in-milk^3		0.044	0.001	1.045	1.044	1.046	**<2.00E-16**
HMM state	≤2	Referent			0.000		
	3	-0.027	0.045	0.973	0.885	1.062	0.232
	4	-0.255	0.010	0.775	0.756	0.794	**1.70E-08**
Energy-corrected milk yield (litres)		0.013	0.001	1.014	1.012	1.015	**<2.00E-16**
Somatic cell count (log)		-0.039	0.005	0.961	0.952	0.971	**<2.00E-16**
Ever previously positive (S/P >= 30)		-0.232	0.032	0.793	0.730	0.856	**6.75E-13**
Age at start of lactation (months)		-0.009	0.000	0.991	0.990	0.992	**<2.00E-16**

Table 6.5: Parameters of the fixed effects of the final model of the association of HMM states with time to first service.

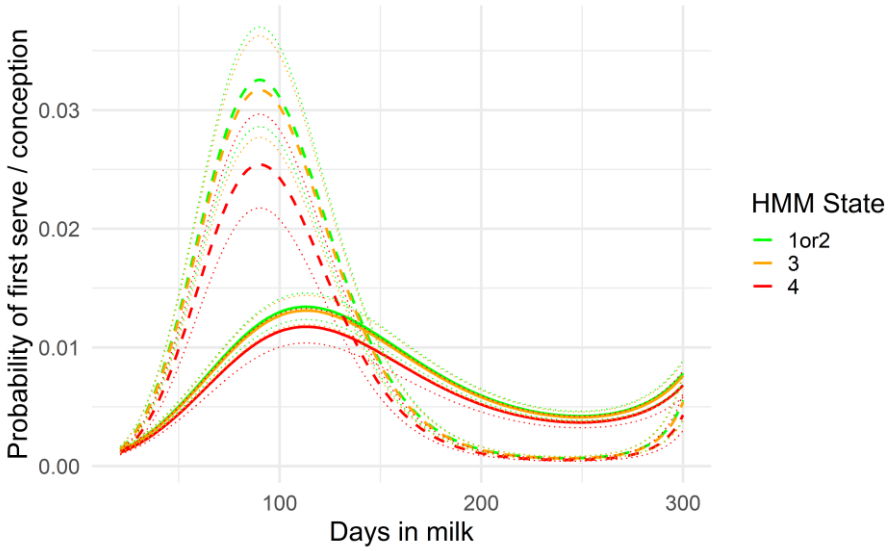

Figure 6.2: Hazard functions for cows in different HMM states in generalised linear mixed effects models employed in discrete-time survival analyses of time to first service and conception.
Dashed lines = first service. Solid lines = conception. 95% confidence intervals are displayed as dotted lines.

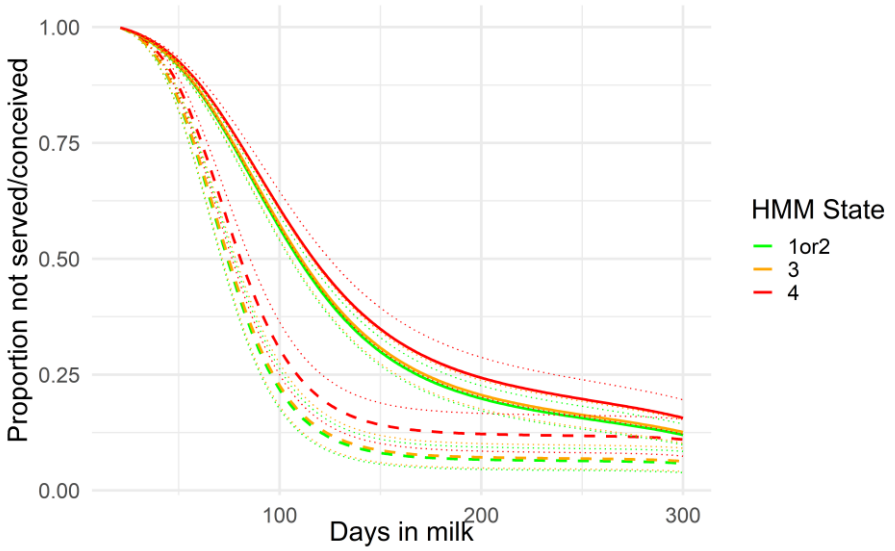

Figure 6.3: Survival functions of the final models for investigating the association of HMM state with the time to first service and conception, for cows in different HMM states.
Dashed lines = first service. Solid lines = conception. 95% confidence intervals are displayed as dotted lines.

6.4.4. Time to conception model

The parameters of the fixed effects of the final model are shown in Table 6.6. Cows in state 4 had lower odds of conception compared to states 1 or 2 (O.R = 0.873, 95% C.I. = 0.796 – 0.958). The odds of conception for state 3 were not significantly different to states 1 or 2. The variable indicating whether a cow had ever previously received a positive test result was identified as a confounder; its exclusion caused an increase in the magnitude of the negative association between HMM states 3 and 4 and odds of conception, and caused the HMM state 3 term to become significant ($P < 0.05$). Yield, somatic cell count, parity and age, were all associated with odds of conception. The median interval to conception for cows in HMM states 1 or 2, 3, and 4 were 109 (105-113), 110 (105 – 115) and 116 (109 – 124). Figures 6.2 and 6.3 illustrate the hazard and survival functions.

6.4.5. Odds of conception model

Model parameters are shown in Table 6.7. HMM state was not significantly associated with odds of conception, although HMM state 4 tended ($P = 0.076$) to be associated with increased odds of conception when compared with HMM state 1 or 2. Age at lactation start was negatively associated with odds of conception. The predicted probability of conception as a function of days in milk for cows in different states is illustrated in Figure 6.4.

Variable	Stratum	Estimate	Standard error	Odds ratio	O.R. Lower 95% C.I.	O.R. Upper 95% C.I.	P-value
Intercept		-5.444	0.054				**<2.00E-16**
Days-in-milk		-0.012	2.21E-04	0.988	0.988	0.989	**<2.00E-16**
Days-in-milk^2		0.007	4.91E-04	1.007	1.006	1.008	**<2.00E-16**
Days-in-milk^3		0.018	4.19E-04	1.019	1.018	1.019	**<2.00E-16**
HMM state	≤2	Referent					
	3	-0.024	0.023	0.976	0.933	1.021	0.2943
	4	-0.135	0.047	0.873	0.796	0.958	**0.0041**
Energy-corrected milk yield (litres)		0.008	7.54E-04	1.008	1.007	1.010	**<2.00E-16**
Log Somatic cell count (x 1000 cells/ml)		-0.031	0.005	0.970	0.961	0.979	**4.10E-11**
Ever previously positive (S/P ≥ 30)	0	Referent					
	1	-0.172	0.032	0.842	0.792	0.896	**5.20E-08**
Parity	1	Referent					
	2	0.056	0.022045	1.058	1.013	1.104	**0.0112**
	3	0.190	0.037356	1.209	1.124	1.301	**3.73E-07**
	4	0.283	0.054223	1.327	1.193	1.476	**1.83E-07**
	5+	0.542	0.077331	1.720	1.478	2.001	**2.37E-12**
Age at start of lactation (months)		-0.020	0.001307794	0.980	0.978	0.983	**<2.00E-16**

Table 6.6: Parameters of the fixed effects of the final model of the association of HMM states with time to first conception.

Variable	Stratum	Estimate	Standard error	Odds ratio	O.R. Lower 95% C.I.	O.R. Upper 95% C.I.	P-value
Intercept		-0.355	0.048				**1.05E-13**
HMM state	≤ 2	Referent					
	3	0.003	0.032	1.003	0.943	1.068	0.914
	4	0.117	0.066	1.124	0.988	1.279	0.076
Days-in-milk		-0.002	0.000	0.998	0.998	0.999	**9.41E-08**
Days-in-milk^2		0.003	0.001	1.003	1.002	1.005	**1.08E-05**
Days-in-milk^3		0.006	0.001	1.006	1.004	1.007	**<2.00E-16**
Age at start of lactation (months)		-0.008	0.001	0.992	0.991	0.993	**<2.00E-16**

Table 6.7: Parameters of the fixed effects of the final model of the association between HMM states and odds of conception at a service event

6.5. Discussion

This work has investigated the associations of the states predicted by an HMM for interpreting serial milk MAP antibody ELISA results, with milk production and fertility. The association between pTB and dairy cow fertility has been the subject of only a few studies (Griss et al., 2024). Notwithstanding the novelty of the HMM states for classifying cows with regards to pTB risk, any associations between these states and fertility would be of interest in the broader context of the implications of pTB. The association between pTB and milk yield is comparatively well-established by research, including a recent meta-analysis (McAloon et al., 2016). Nevertheless, the implications of the HMM states for milk production are of interest due to the novel nature of this method of interpretation of serial pTB ELISA test results.

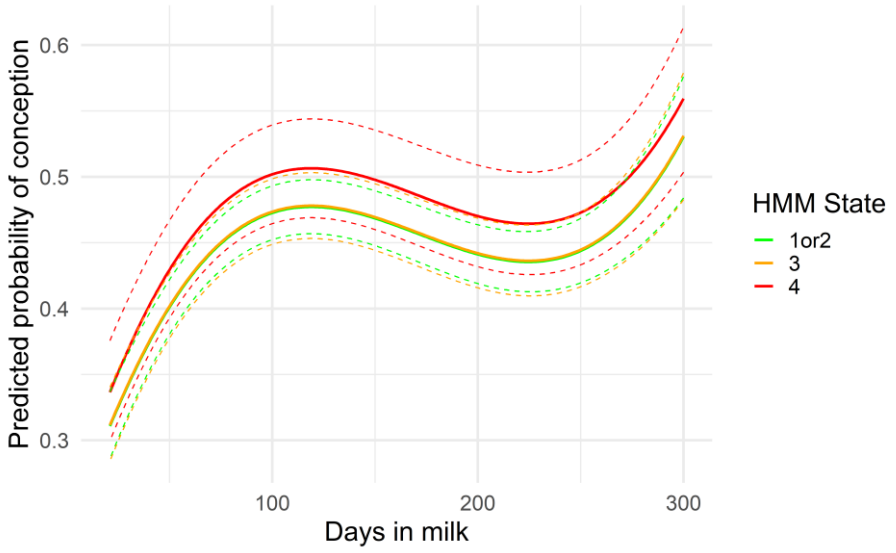

Figure 6.4: Predicted probability of conception as a function of days-in-milk, from a model estimating the odds of conception given the occurrence of a serve event.
95% confidence intervals are shown as shaded areas whose boundaries are dotted lines.

Several previous studies have classified cows with regards to pTB infection using a single ELISA result in cross-sectional data, or using the latest or highest result out of repeated tests in longitudinal data. Few studies (Nielsen, 2008; Nielsen et al., 2009; Pritchard et al., 2017) have investigated production or fertility in cows classified according to the results of multiple consecutive ELISA tests. The creation of an objective method of serial test interpretation, and subsequent interest in the implications of this method, motivated the current study.

Our results show that cows which reached HMM state 3 or 4 produced less energy-corrected milk in the concurrent lactation, compared with cows remaining in states 1 or 2. Additionally, less milk was produced during the current lactation if a cow had reached state 4 during any previous lactation. This second finding has an intuitive explanation. Cows which are in relatively late stages of pTB, and hence which would be expected to have received higher numerical test results (S/P ratios) over a relatively longer duration, would also be expected to be affected more severely.

The magnitude of yield losses found in this study (Table 6.4) appear modest in comparison to the losses found in some previous studies. Reaching HMM state 3 or 4 was associated with 305-day losses of approximately 53 or 99 litres of energy-corrected milk during the concurrent lactation (equivalent to 0.2 to 0.3 litres per day). For the average cow (with a median value of all categorical variables and a mean log somatic cell count) giving a model-predicted energy-corrected yield of 8438.6 litres, this equates to a

loss of 0.6% or 1.2% for states 3 or 4 respectively. Several studies have found greater production losses associated with milk pTB antibody ELISA positivity. For example, losses of between 254 and 1000 litres have been reported (Hendrick et al., 2005; Ozsvari et al., 2020; Sorge et al., 2011), or daily losses of between 0.34 and 4 litres (Nielsen et al., 2009; Pritchard et al., 2017). Those reported losses are broadly consistent with studies which have classified cows according to serum ELISA results; average 305-day losses of between 212 and 3248 litres or Kg of milk have been reported (Gonda et al., 2007; Lombard et al., 2005; Nordlund et al., 1996; Shook et al., 2012; Tiwari et al., 2007; Van Leeuwen et al., 2002), and daily losses of between 1 and 4 Kg (Beaudeau et al., 2007). A meta-analysis (McAloon et al., 2020) found an average daily loss of 1Kg regardless of test medium (milk or serum).

HMM state 4 is typically associated with S/P ratios which would be considered positive results (S/P ≥ 30%) (Figure 4.1), and therefore cows in state 4 could be expected to be associated with losses similar to those described above. However, herds in the current study have been subject to regular whole-herd pTB tests for several years, and so farm managers are likely to be taking steps to reduce within-herd pTB prevalence, including prompt culling of cows thought to be infected; infected cows in these data are therefore likely to be in relatively early stages of subclinical pTB. This is in contrast to the majority of previous studies, which classified each cow according to one ELISA result; such cows could be in relatively later stages of disease. It is interesting to note that in the study by Ozsvari *et al.* (2020), which found a relatively severe loss (1000 Kg over 305d), it was acknowledged that the ELISA-positive group

included some clinical (i.e. late-stage) cases. The results of the current study indicate an additional loss of approximately 108 litres over 305 days for cows which have also reached state 4 during a previous lactation. These cows are likely to be at a later disease stage than cows which have reached state 4 for the first time during the current lactation, and their total losses are predicted to be more consistent with the majority of published research. State 3 cows are typically associated with S/P ratios below the test-positive threshold (S/P = 30%; Figure 4.1), and so are likely to include cows even earlier in the disease course when yield losses would be expected to be less than that experienced by state 4 cows. Some studies have reported an increased yield in cows prior to their first positive ELISA result. If a similar effect is present in the sample used in this study, it could further explain the minor yield losses found. The results of the current study are consistent with the hypothesis that state 3 cows, which are overlooked when interpreting the ELISA test in a conventional manner (dichotomised at S/P 30%) are in the early subclinical stage of pTB.

Cows in HMM state 4 were found to have a calving to conception interval which is on average 7 days longer than cows in states 1 or 2. The results of the time to first service model suggest that this delay is to a large degree due to a reduced risk of service, and an average extension of the calving to first service interval of 6 days. It is unknown precisely to what extent the delay in service is the result of infertility (delayed return to cyclicity, or failure to express oestrus) or reluctance of farmers to breed cows which have had a positive ELISA result, but the confounding in both survival models between both states 3 and 4, and the variable indicating if the cow had ever

previously had a positive test result, is suggestive that farmers' breeding decisions play a substantive part in these results.

The diminished risk of conception is consistent with some previous studies which have found reduced fertility in ELISA test-positive cows (Johnson-Ifearulundu et al., 2000; Ozsvari et al., 2020; Reynolds et al., 2023) . However, other studies have found that at least some cows with positive ELISA results have relatively better fertility (Lombard et al., 2005; Reynolds et al., 2023). Reynolds *et al* (2023) found that cows experienced a period of improved conception rates prior to their first positive test. Similarly, the tendency ($P < 0.1$) in the odds of conception model for state 4 cows to be more likely to conceive given a serve event appears to mitigate somewhat the reduced risk of service in these cows. Overall, the current results appear to indicate that state 4 cows are less likely to be served, but that the risk of conception in these cows is not diminished and may be increased.

It is worth considering some limitations of this study. First, the data were from a convenience sample of farms conducting milk-recording and pTB testing through one laboratory. The selected farms operated commercial dairy operations and were assumed to be typical of UK dairy farms. Choice of laboratory has been found to influence pTB milk ELISA results (Barden et al., 2020). There cannot be complete confidence in the generalisability of the results presented here. The analysis also was not designed to investigate the association of HMM states with past or future performance and so does not help investigate the hypothesis that cows can experience a period of improved fertility (Reynolds et al., 2023) or production (Nielsen et al., 2009) prior to receiving positive ELISA

results. As with many similar studies, survivorship bias is also likely to have been present in the analysis of 305-day milk production; low-yielding cows and cows which have positive ELISA results are likely to be at increased risk of culling, and their absence from the data would bias the association between HMM state and yield towards the null, especially for state 4 cows whose test results are typically positive (S/P ≥ 30%). This could further explain the modest loss of yield in states 3 and 4.

Future similar research would benefit from a larger sample taken from other laboratories, and could include analysis of the association of HMM states with future and past fertility and milk yield. Investigation of the HMM states with other health conditions (e.g. mastitis, lameness or metabolic conditions) would also be of interest. Whilst this study suggests that state 3 cows are affected by pTB, it is unknown to what extent state 3 cows are contagious to herd-mates; future research should aim to address this uncertainty.

Chapter 7. Discussion

7.1. Overview

The aim of this thesis, that is to create predictive models for use by farmers or their advisors in real-time when making management decisions, has been achieved; in the process providing novel insights into some of the nuances of statistical learning applied to data collected on dairy farms, and enhancing knowledge of two impactful diseases, clinical mastitis and pTB. Considering this work in a wider context, it provides some practical tools for improving the productivity and welfare of cows on commercial farms, which is consistent with the need to meet the increasing global demand for safe, sustainable sources of nutritious dairy products. Insights gained from this work could facilitate ongoing research in the fields of predictive modelling, paratuberculosis and mastitis. Several of the models presented here have been deployed in commercial software used for analysis of dairy farm data, and are already used by farmers. Regarding predictive modelling, the focus of this thesis, insights have been provided into the use of modelling methods for longitudinal data which have been little-utilised previously within the dairy science literature (for example repeated Bayesian updating in Chapter 3, and the hidden Markov model in Chapter 4).

In Chapter 2, a model was created for predicting the probability of cure of a case of clinical mastitis, which is used in making real-time treatment decisions, and that also yields useful inferences regarding the factors most important in determining whether a cure occurs. Despite a substantial body of published research investigating such factors, to the author's knowledge models with

the ability to accurately predict cure probability have not previously been reported. The model in Chapter 2 should be useful for reducing AMU on farms, potentially mitigating the potential for development of AMR. Reduced AMU would in turn have economic benefits to farms, owing for example to reduced medicines expenditure and milk-discard.

Chapters 3 to 6 focus on paratuberculosis. Chapter 3 explored prediction of the probability of MAP-infection from data available during the early life of an animal, and demonstrated some difficulties in applying repeated Bayesian updating to serial pTB test results. Chapter 4 showed that the use of HMMs for unsupervised learning on longitudinal data is a novel and practical approach for objectively quantifying disease risk using serial ELISA results. In particular, this approach revealed the existence of a hitherto-unidentified group of cows (HMM state 3) that are at relatively high risk of progression despite giving ELISA results which are conventionally considered negative. This has substantial implications for pTB control programmes and therefore for management of the potential, albeit unproven (Waddell et al., 2015), risk of zoonotic MAP infections. Chapter 4 also affirmed existing knowledge of non-disease factors which can influence the results of the MAP antibody ELISA, and described a methodology which can account for them. In Chapter 5, models were developed for predicting the onset of “J5” status in individual cows within the subsequent 12 months, demonstrating the potential of the novel HMM states for predicting future pTB-status. Chapter 6 addressed the question of whether the HMM states are associated with fertility

and milk production, thereby helping to establish the relevance of the HMM states to real-world cow health outcomes.

7.2. Predictive modelling

7.2.1. The models and the data

Much recent research has been conducted into the application of machine learning to data collected on dairy farms, with a large proportion of studies reporting the use of highly-sophisticated algorithms such as neural networks and tree-based models (Slob et al., 2021). Such algorithms are well-suited to large or high-dimensional datasets ("big data"), with potentially many complex interactions between variables, enabling the identification of subtle but important relationships within the data. The combination of big data and sophisticated ML algorithms could be employed, for example, when modelling high-frequency sensor data. However, many data recorded on dairy farms are smaller in size and of low dimensions (e.g. fertility data, diagnostic test result data, mobility scoring, medicines records), as is the case with the datasets employed in this thesis. With small, low-dimensional data which does not include complex interactions, logistic regression performs as well at classification as complex ML models (van der Ploeg et al., 2014), and a systematic review of clinical prediction models developed for human medicine found that logistic regression has similar performance to more complex ML algorithms (Christodoulou et al., 2019). When the quantity of data available is modest, as in this thesis, simpler models (e.g. linear models) are an appropriate choice, whereas use of more complex model types would potentially lead to overfitting. All of the models employed in

this thesis are of a relatively simple parametric type, which is appropriate for the modest size of the data, and furthermore include readily-interpretable model parameters, providing useful insights into the disease processes studied and ensuring the model portability.

7.2.2. Random effects

Inclusion of random effects in models trained on non-i.i.d. data permits predictions to be made using only the intra-cluster (fixed) effects; predictions are then representative of an average group (cow or herd in this context) (Harrison et al., 2018). Figure 7.1 illustrates the variance partition coefficients for the GLMMs created in this thesis. In all but one of the models (the inferential model employed to investigate the association between HMM state and 305-day milk yield in Chapter 6), the herd random effect contributed only a modest proportion (less than 0.25) of total model variance. For these models, inclusion of a herd random effect is likely to have made little difference to the other model parameters and model performance. This provides reassurance that predictions made by these models would be generalisable to novel data from other herds, under the assumption that the herds in the model-training data were broadly representative of commercial dairy herds in general. The greater contribution by the herd random effect in the 305-day yield model is expected given the normal variation in yields of dairy herds resulting from, for example, differences in genetics or husbandry. The considerable contribution of the random effect of cow in the models used to adjust S/P ratios in Chapter 4 has two implications. Firstly, when using these models to adjust S/P ratios in novel data, errors in predictions for some individual cows may be

relatively large. This is intuitive because the S/P ratio (the dependent variable in these models) varies with pTB infection status, which is not included as a fixed effect in these models. Infected cows tend to have higher S/P ratios and so contribute to the variance of the cow random effect. Secondly, however, given the small proportion of pTB-infected cows in the data, it is also possible that other cow-level variables (for example breed or genetics), not included in the models, are an important factor in between-cow variation in S/P ratios. It is noteworthy that the variances of both random effects in the clinical mastitis cure prediction model (Chapter 2) were small; this reassures that this model is generalisable to novel data.

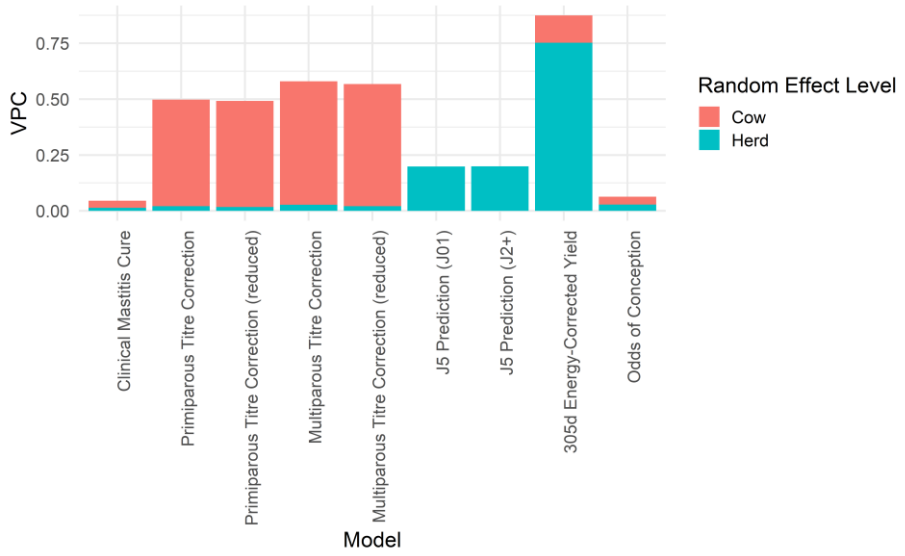

Figure 7.1: Variance partition coefficients of the herd- and cow-level random effects in linear mixed models created in this thesis.
VPC = Variance partition coefficient. For the generalised linear mixed models (GLMMs) with a logit link, the residual model (within-cluster) variance was defined as $\frac{\pi^2}{3}$ *according to the latent response formulation (Austin and Merlo, 2017).*

7.2.3. The Elastic Net

Use of the Elastic Net model was an appropriate choice in Chapter 3. The small quantity of data (n = 7,548) and the perceived low signal-to-noise ratio of the data prompted the use of regularisation in an attempt to avoid overfitting. In these data, the signal-to-noise ratio was expected to be low, as none of the independent variables directly described features of the particular animal whose disease status was the target of interest. This expectation is consistent with the hyperparameter values selected by each inner loop of the nested cross validation (Table 3.7). For the hyperparameter controlling the overall extent of regularization in the Elastic Net, λ, a low value (≤ 0.1) was consistently selected across the outer folds of the nested CV, indicating that little regularisation was applied. A plot of the loss function across all assessed hyperparameter values is shown in Figure 7.2. The loss function varies by only a small amount (≈ 0.007) across all values of the hyperparameters, indicating that regularisation failed to substantively improve model performance.

The application of minimal regularisation by the Elastic Net could occur with data which has a particularly high or low signal-to-noise ratio, either of which would result in little improvement in model performance through regularisation. The shape of the loss function in Figure 7.2 is not convex for any values of α; a convex function would indicate that there is an optimal level of regularisation for this model and this data, whereas here there appears to be no such optimum. This suggests a low signal-to-noise ratio. Model performance improved as the value of α approached zero, indicating that the L2 penalty was strongly preferred over the L1

penalty. In contrast to L1, the L2 penalty does not shrink coefficients all the way to zero, so does not select a subset of variables. The dominance of the L2 penalty suggests that there was not a small number of useful predictor variables in this data, but that many variables are correlated in a small way with the dependent variable (i.e. a low signal-to-noise ratio was likely). However, the pipeline that was assessed within the nested cross-validation included an initial step of selecting the optimum number of variables to offer to the Elastic Net algorithm. Table 3.7 shows that out of a total of thirty predictor variables, a maximum of 16 variables was selected by the nested CV. Preselecting the optimum number of variables reduced model complexity, potentially reducing the extent of regularisation required, and could also have led to the low alpha values (dominance of the L2 penalty) selected by the nested CV. Overall, the nested CV procedure suggests that the signal-to-noise ratio was low, although preselecting the number of variables may have affected the optimal hyperparameter values.

The low signal-to-noise ratio appears to have resulted in a model which is limited in its classification ability. The distribution of predicted probabilities (Figure 3.3) is narrow and confined to the lower range of probability, indicating that the model struggles to differentiate between subjects which are likely or unlikely to be infected with pTB. This led to a modest balanced accuracy (Table 3.7), although calibration of predicted probabilities was acceptable (Figure 3.3).

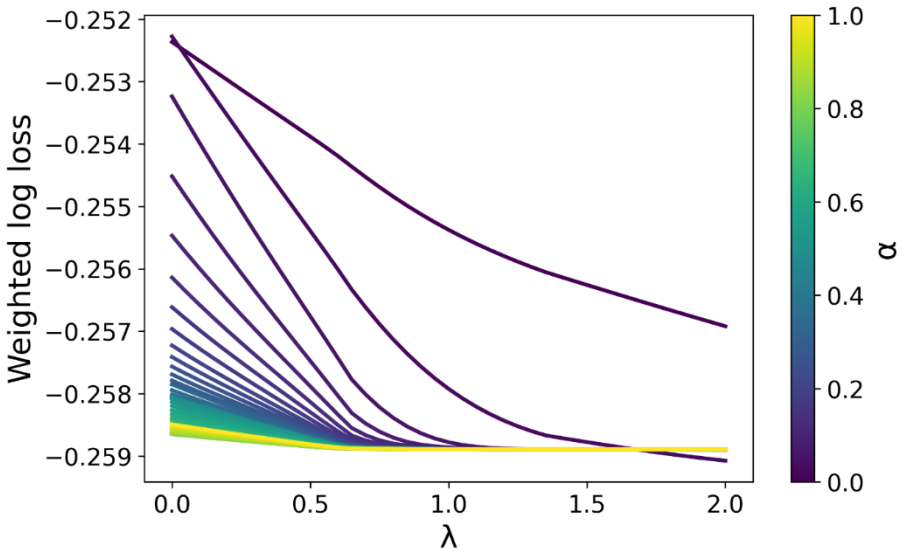

Figure 7.2: Hyperparameter tuning within the nested cross-validation procedure in Chapter 3.
Results of the nested cross-validation are pooled across all outer folds (farms). Lines are estimated by locally weighted scatterplot smoothing (LOWESS).

7.2.4. Repeated Bayesian Updating

Repeated Bayesian updating (Chapter 3) was used for interpretation of serial pTB ELISA results, yielding a posterior probability of disease in an animal following each repetition of the test. This approach to interpretation of sequential test results has been employed previously in the veterinary and human medical literature, for example in the context of pTB in cattle (Meyer et al., 2018) and COVID-19 in humans (Cao et al., 2021). With diseases of low prevalence, such as pTB, the positive predictive value of one positive dichotomised result of a test of imperfect specificity is expected to be low; the combination of the results of several tests using Bayesian updating should be useful in this scenario for improving the positive predictive value (Balayla, 2022). In addition, given the modest sensitivity of the pTB ELISA, the negative

predictive value of a single dichotomised test is also expected to be low, but should be improved with repeated testing.

Balayla (2022) proposes a method for calculating the number of dichotomous test results required in repeated Bayesian updating for obtaining a given positive or negative predictive value of the posterior probability, accounting for the sensitivity and specificity of the test and the disease prevalence. The calculation for the number of positive tests required for a given positive predictive value is:

$$n_{ppv} = \frac{\ln\left[\frac{\rho(\Phi - 1)}{\Phi(\rho - 1)}\right]}{\ln\left[\frac{a}{1 - b}\right]}$$

where n_i is the number of tests required, ρ is the desired positive predictive value, Φ is the disease prevalence, a is test sensitivity and b is test specificity. The equivalent equation for negative predictive value is:

$$n_{npv} = \frac{\ln\left[\frac{\tau\Phi}{(1 - \Phi)(1 - \tau)}\right]}{\ln\left[\frac{b}{1 - a}\right]}$$

where τ is the desired negative predictive value.

Figure 7.3 shows the number of sequential positive or negative dichotomised test results required for a positive or negative predictive value of 95% according to this method, for a conceivable range of pTB ELISA sensitivity. Owing to high specificity, a positive predictive value of at least 95% is assured after approximately two positive tests, whereas the number of tests required for a 95% negative predictive value can be almost four when test sensitivity is at its lowest, for example during the early stages of the disease.

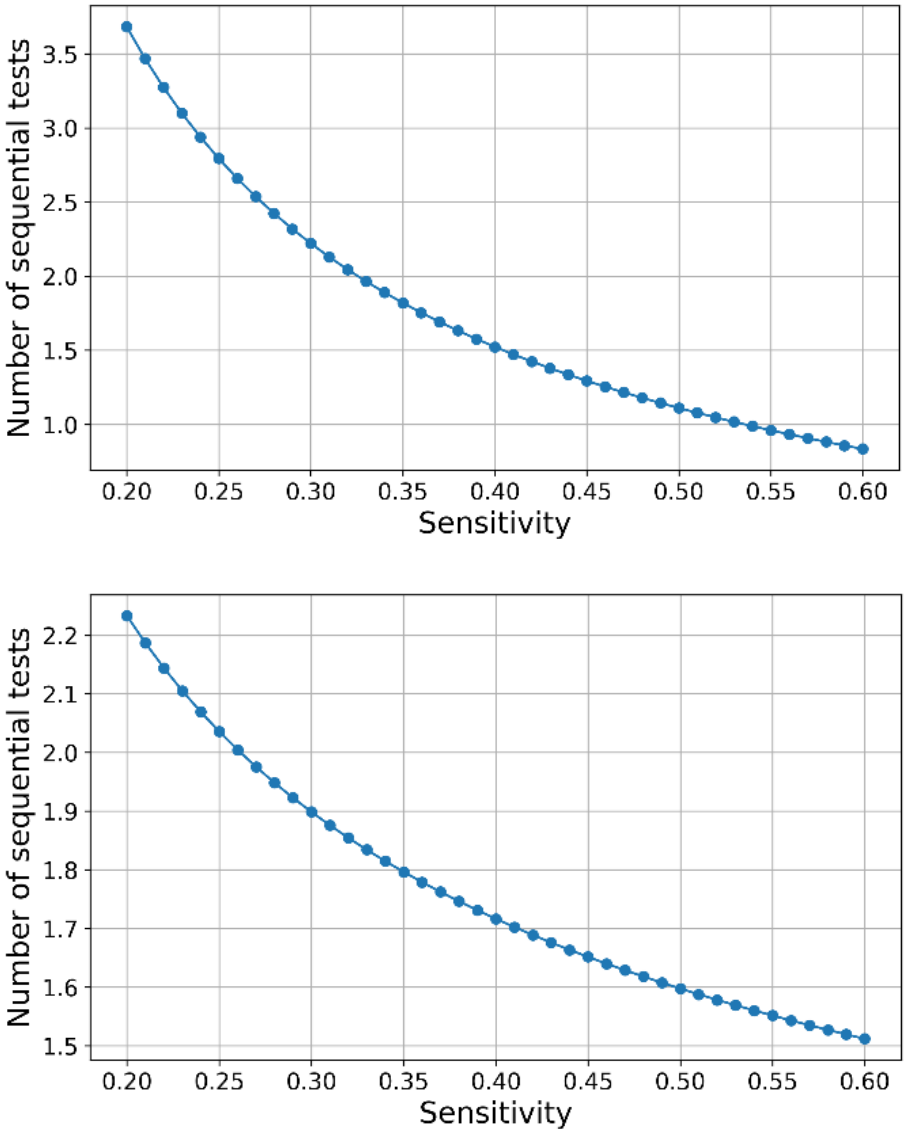

Figure 7.3: The number of sequential tests required for a 95% predictive value given known sensitivity of the diagnostic test
Top panel: negative predictive value, bottom panel: positive predictive value. Specificity was fixed at 0.98, prevalence at 0.1 (Balayla, 2022).

In the context of the continuous S/P ratios modelled in Chapter 3, as opposed to dichotomised test results, the posterior probabilities (post-test probabilities of disease) are equivalent to predictive values. According to the above rationale, the posterior probability

should decrease, or increase, with each sequential low, or high, S/P result, yielding a low or high posterior probability of disease after a small number of sequential results. If there is dependence between sequential test results, as is likely when repeatedly testing an individual animal using the same assay, the change in the posterior probability with each test will be amplified further (van Walraven et al., 2009). This is consistent with the problem identified in Chapter 3, and illustrated in Figure 3.10, whereby the posterior probability becomes so low in the face of repeated low S/P ratios that it remains low even when updated with a high S/P ratio. Figure 7.3 also shows that, in the scenario of no dependence between test results, only a small number (approximately 2) of sequential positive ELISA results would be required for a positive predictive value of 0.95. With dependence between sequential test results, a very high posterior probability of infection could result after only one test; this is less problematic given the need to identify potentially infected animals with high sensitivity. In summary, failure to account for dependence between sequential test results is likely to have exaggerated posterior probabilities, leading to their poor calibration (Figure 3.9).

Methods have been proposed for adjusting likelihood ratios in Bayesian updating to account for known dependence between tests (Chan et al., 2008; Fanshawe et al., 2024). However, application of these methods would be difficult in this context. Conditional correlation between sequential ELISA results is difficult to accurately assess as it is expected to vary over the course of the disease due to changes in the biological system (the cow and the disease agent) over time. Compounding this difficulty is the varied interval between sequential tests in individual animals.

Consideration should also be given to the condition predicted by the Bayesian updating. Prior to application of repeated Bayesian updating, it was necessary to define patterns of ELISA results which represented “positive” or “negative” animals. These definitions were used to create a dependent variable for the models used to derive the Bayes factors in repeated updating; since these definitions describe test outcomes, the posterior probabilities provided by the repeated Bayesian updating could be viewed as the probability not of disease or infection, but simply of seroconversion. Since seroconversion is not correlated perfectly with infection status or disease progress, the very low posterior probabilities resulting from repeated low S/P ratios could be viewed as an indicator that the animal has not yet seroconverted, but not so much as an indicator of disease status.

The subjectivity of defining “positive” and “negative” labels in the data in Chapter 3, and the poor calibration of the posterior probabilities, prompted the use of unsupervised learning in Chapter 4; in the hidden Markov model, disease status is not subjectively defined, but modelled as a number of latent states.

7.2.5. The hidden Markov model

Hidden Markov models are ideally suited to longitudinal data where the true state of a subject is unknown (latent). Paratuberculosis ELISA results are only a proxy measure of true disease status. Much data collected on dairy farms indirectly measure the condition of interest. For example, somatic cell counts, commonly used for the detection of intramammary infections, measure the inflammatory response to a pathogen, rather than detecting the pathogen itself;

intradermal tuberculosis testing similarly assesses the cell-mediated immune response to antigenic stimulation, rather than truly detecting infection. Data such as these would be well-suited to the application of HMMs. However, use of HMMs in the veterinary literature is rarely described. Use of an HMM for detection of intramammary infections using somatic cell count data has previously been reported (Detilleux, 2011). HMMs have also been applied in herd-level disease surveillance (Robertson et al., 2011; van Roon et al., 2022) and prediction (Saro et al., 2024), for animal behaviour recognition (Leos-Barajas et al., 2017; Pastell and Frondelius, 2018; Sumi et al., 2021), for lameness-detection in image data (Magee and Boyle, 2002), and for describing patterns of MAP-shedding in pTB cases (Ceres et al., 2020; Louzoun et al., 2015). HMMs have seen limited use in the human medical literature, for example in the diagnosis of cancer and heart arrythmias (Mor et al., 2021). This thesis expands the reported repertoire of use-cases of HMMs to include diagnosis of a chronic livestock disease based on serial test results, where the test results are indirect measures (proxies) of true disease status.

7.2.6. Model selection and generalisability

Throughout the thesis careful consideration was given to the methods used for model selection and assessment, aiming to create final models whose predictive ability is generalisable to novel data or whose parameters are reliable for inference. Cross-validation was considered the ideal method of model selection, and predictive performance on a hold-out testing dataset was considered a robust method of assessing final models; both cross-validation and assessment against a hold-out testset are appealing

methods as they directly measure the accuracy of model predictions against real data. Selection of the best subset of predictor variables, by conducting an exhaustive search of all possible combinations of features, ensures that the final model is selected from a maximal pool of possible models. Therefore, whenever possible, models were selected according to the results of an exhaustive search over all combinations of candidate features, using cross-validation, and model evaluation included assessment against a hold-out testing dataset. When this approach was computationally infeasible, owing to large datasets or a very large number of candidate models, information criteria and backwards feature selection were used as appropriate in place of cross-validation and an exhaustive predictor search. Table 7.1 provides an overview of model selection and assessment methods used in this thesis, as well as a justification for their use. Given that the aim is to deploy many of the models in this thesis for use by farmers, generalisability to novel herds is crucial. It is therefore important to consider the extent to which generalisability has been established by the model assessment methods used in this thesis. Whenever possible, the final selected models were assessed against a "held-out" portion of the data, consisting of data from multiple farms, which was not used in any way during model development or selection.

Techniques were employed to avoid data-leakage, which could favourably bias estimates of model performance; data were always split respecting groupings, so data from one farm could not appear both in the training and testing data, and any pre-processing of data (e.g. variable transformation, feature engineering, variable

standardisation) was performed in the same way on both subsets after the data had been partitioned (Yang et al., 2022). This method of validation against a "held-out" testset of several herds provides reassurance that a model does indeed generalise to other farms. However, it must also be noted that the "held-out" sets were often taken from the same data sample as that used to train the models, and therefore are assumed to be subject to similar biases as those in the model-training data (Wan et al., 2022) (e.g. similarities in disease prevalence, epidemiology, laboratory techniques, data-collection or processing), which would be expected to bias estimates of model performance. Ideally, testing data would be sampled separately from model-training data, further establishing generalisability to farms which differ in ways not captured by the variables in the model. Additionally, it cannot be ruled out that random splitting of data yielded model-testing subsets on which the selected models performed well by chance (Miller et al., 2024), although only one (the J2+ test set in Chapter 5) of the model-testing datasets were so small (e.g. less than 1000 rows) that this is likely (Tohka and van Gils, 2021). In Chapter 3, nested leave-one-group-out cross-validation was used as an alternative to simple data-splitting for assessment of the performance of the Elastic Net model. This decision was motivated by the routine use of resampling during selection of machine learning models; since cross-validation was employed for hyperparameter tuning, it was natural to extend this procedure to also provide estimates of model performance.

Table 7.1: Methods used in this thesis for model-selection and -assessment. *P = number of candidate predictor variables. n = dataset size (number of rows).*

Chapter	Model purpose	P	n	Methods		Justification
				Selection	Evaluation	
2	Prediction of clinical mastitis cure probability	18	5858	Backwards elimination of features, guided by AIC	Hold-out testing dataset	Exhaustive search for best predictor subset and cross-validation were both precluded by the large number of candidate predictors, which would lead to cross-validating a very large pool of candidate models. Therefore, backwards model selection was used, guided by AIC. AIC is appropriate for a predictive model.
3	Prediction of probability of pTB infection from early-life information	20	7548	Cross validation (non-nested leave-one-group out)	Cross validation (nested leave-one-group out)	Elastic net model-fitting is relatively computationally-inexpensive, so even with a large number of variables (20), cross-validation is possible, especially given the small size of the data. Leave-one-group out cross validation is useful for assessing the generalisability of models to novel groups.
3	Derivation of Bayes factors from models predicting S/P ratio	7	7478	Exhaustive search for best subset of predictors using information criterion (WAIC), plus threshold tuning using ROC analysis	Posterior predictive checking on hold-out testing dataset	Small number of candidate predictors and small dataset, so an exhaustive search for the best feature subset was possible. Bayesian cross-validation would be prohibitively computationally expensive, but WAIC is asymptotically equivalent to Bayesian leave-one-out cross-validation. Model goodness-of-fit was assessed using posterior predictive checks on a hold-out testing dataset, which is appropriate for Bayesian models.
4	Adjustment of S/P ratios for non-disease factors	6	220179	Cross-validation (selection of number of spline degrees of freedom) and backwards elimination of features guided by AIC.	Entire pipeline assessed after subsequent application of HMM.	A large dataset precluded cross-validation of full models or an exhaustive search for best feature subset. An information criterion was used to guide backwards feature elimination instead. AIC is appropriate for predictive models.
4	Unsupervised learning of adjusted S/P ratio data		224852	AIC	Comparison of simulated S/P ratios with observed data	Cross-validation not appropriate for unsupervised learning. AIC is appropriate for predictive models.
5	Prediction of the onset of J5 status	12	22969 and 1125	Exhaustive search for best subset of predictors using cross-validation	Hold-out testing dataset	The modest size of the data permitted cross-validation coupled with an exhaustive predictor subset search. The particularly small size of the J2+ data risked instability (high variance) of the cross-validation results, and therefore some bias was introduced by using *k*-fold cross validation as opposed to leave-one-group-out cross-validation.
6	Inferring association of HMM State with 305-day milk yield	<10	28583	Exhaustive predictor subset search, guided by BIC	None (inferential models)	Cross-validation coupled with exhaustive feature subset search were difficult with the size of the data, so an exhaustive search was conducted, guided by an information criterion. BIC is appropriate for inferential models.
6	Inferring association of HMM State with time to service and conception	<10	74474	Backwards feature selection, guided by BIC	None (inferential models)	Either cross-validation or an exhaustive search for the best subset of predictors were too computationally expensive, owing to the size of the data. Therefore, backwards feature elimination was carried out, guided by an information criterion. BIC is appropriate for inferential models.
6	Inferring association of HMM state with odds of conception, given that a service occurs	<10	55307	Exhaustive predictor subset search, guided by BIC	None (inferential models)	Cross-validation coupled with exhaustive feature subset search were difficult with the size of the data, so an exhaustive search was conducted, guided by an information criterion. BIC is appropriate for inferential models.

Use of CV minimises the concern that a model may fit the testing data by chance, since the model is assessed against many different partitions of the data. Furthermore, use of leave-one-group-out CV was appropriate for the non-i.i.d. data in this thesis, as it estimates model performance when applied to new groups. However, use of cross-validation in Chapter 3 may have inadvertently led to data-leakage. The pipeline in this chapter also included a pair of Bayesian models created from the same data as the Elastic net, although this pair of models was fit and assessed by partitioning data into training and testing subsets. Therefore, the Elastic Net model, which provided the prior odds to be updated at the first pTB test result of an animal's life, had been trained partially on data which was also used to assess the calibration of the entire pipeline. Estimates of the calibration of the posterior probabilities provided by the pipeline may have been somewhat favourably-biased by the consequent data-leakage. Nevertheless, given that posterior probability calibration was so poor (Figure 3.9), the pipeline in Chapter 3 is of little use for prediction of pTB risk anyway.

The unsupervised nature of the hidden Markov model in Chapter 4 meant that it could not be assessed using labels in a "held-out" testset or within a cross-validation procedure. Instead, the entire pipeline (Elastic Net and HMM) was assessed by comparing model-simulated and observed data (S/P ratios) (Gelman et al., 1996). Further validation of this pipeline could include similar comparisons using an entirely new dataset, for example from another laboratory.

Overall, whilst the methods used in this thesis suggest the models are generalisable, their validation against further samples of data is still desirable.

7.3. Clinical mastitis

Reduction of unnecessary AMU in agriculture is recommended for mitigating the proliferation of AMR (O'Neill, 2016). Clinical mastitis is the most common therapeutic indication for use of antimicrobials on dairy farms (Kuipers et al., 2016; Lardé et al., 2021; Leite de Campos et al., 2021). The model presented in Chapter 2 allows farmers to select mastitic cows with a low-probability of cure, and to withhold antimicrobial treatment. Given a low enough probability of cure, withholding antimicrobial treatment in these cows is expected to reduce AMU whilst impacting cure rates only minimally.

To achieve a reduction in AMU when treating CM, some farms have adopted an approach whereby milk samples from mastitic cows are cultured on-farm, and culture results available after approximately 24 hours are used to identify cases of mastitis caused by Gram-negative bacteria and cases resulting in no growth. Antimicrobials are not administered to such cases. Owing to the direct cost of implementing OFC, this approach is likely to result in a net cost to the farmer unless the proportion of IMIs in the herd which are caused by Gram-positive pathogens is small, for example less than 20% (Down et al., 2017). The aetiology of mastitis has changed over recent decades, such that Gram-negative infections are increasingly common (Bradley, 2002), but Gram-positive pathogens are still unlikely to be responsible for less than 20% of IMIs; a

systematic review of studies globally reporting mastitis aetiology demonstrated that the prevalence of Gram-positive infections tends to be considerably greater (Morales-Ubaldo et al., 2023). Even in 50 modern dairy herds in Wisconsin, postulated to contain a relatively large proportion of Gram-negative IMIs, over one quarter of IMIs were caused by Gram-positive bacteria (Oliveira et al., 2013). The model presented in Chapter 2 is highly portable and, in contrast to OFC, does not require the use of expensive consumable items. Its implementation can therefore be a relatively inexpensive route to effective antimicrobial stewardship (Dyar et al., 2017), reducing AMU on farms.

It is interesting to examine the potential reduction in AMU achievable through withholding antimicrobials from cases with low cure probability, and to compare it with AMU reductions reported in studies implementing selective treatment based on results of OFC. When adopting the OFC approach, reductions of between approximately 50 and 70% have been reported (Kock et al., 2018; Lago et al., 2011; Mansion-de Vries et al., 2016; Vasquez et al., 2017). Figure 7.4 shows the proportion of CM cases that would be treated with antimicrobials if all cases above a given threshold of predicted probability were treated; the data is for 12 herds whose mean predicted cure probability is in the bottom quartile of herds (i.e. herds with the largest proportion of cows “unworthy” of antimicrobial treatment). For these herds, to achieve a reduction in AMU of approximately 50%, antimicrobials should be witheld from all cows with a predicted probability of less than approximately 25 to 30%. With a cure probability as high as 25 or 30%, farmers may be reluctant to withold antimicrobial treatment, not wishing to

squander an opportunity to achieve a cure. However, Figure 7.4 also shows that a considerable reduction in AMU could be achieved in herds such as these with use of a lower probability threshold, which may be more acceptable to farmers. Overall, it is unlikely, given the extent of the AMU reduction of 50 to 70% reported when adopting OFC-based selective treatment protocols, that cure-probability-based treatment decisions would reduce AMU to a similar extent, unless farmers are willing to use a relatively high probability threshold (25 – 30%). Nevertheless, use of this model alongside, or as an alternative to, the OFC-approach could be a valid route to significantly reduce AMU on farms.

7.4. Paratuberculosis

Implementation of regional or national pTB control programs has in some cases been associated with reduced infection prevalence, but pTB-eradication, even from limited geographical regions, remains rare (Weber et al., 2024).

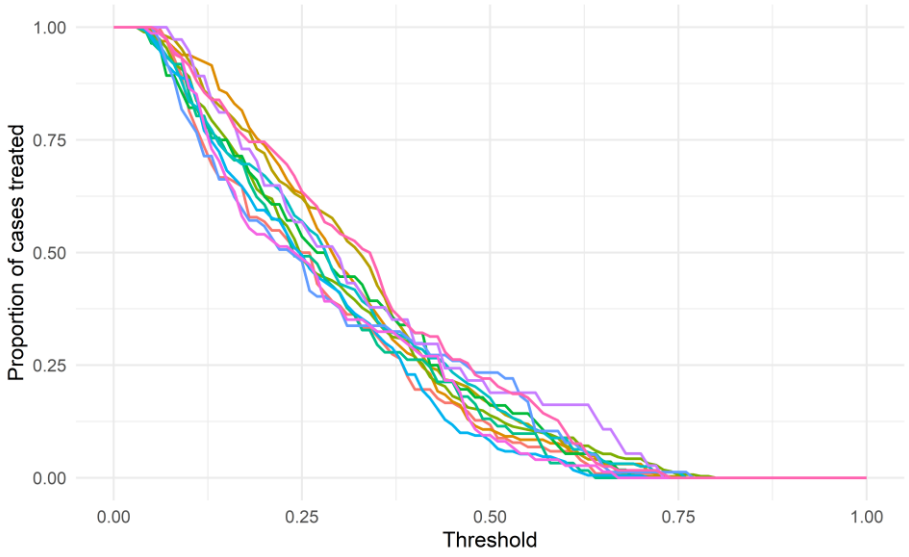

Figure 7.4: The proportion of clinical mastitis cases that would be treated with antimicrobials according to the chosen cure probability threshold.
The data shown is for the twelve farms whose mean cure probability was in the bottom quartile of herds. Each farm is represented by a different coloured line.

In the United Kingdom, for example, a recent survey found that the prevalence of antibody-ELISA test-positive cows, and the mean ELISA S/P ratio, had both fallen over the previous decade (Taylor et al., 2024), suggesting a concurrent reduction in infection prevalence. However, a fall in prevalence of test-positive cows could be brought about by reduction of the rate of new infections, or by an increased rate of culling of test-positive cows, the latter of which was indeed found to have increased (Taylor et al., 2024). Therefore, it is uncertain to what extent the successes of the UK pTB control programme are attributable to control of MAP-transmission relative to increased culling.

In the UK, farmers are incentivised to manage or cull cows which receive one or more positive (S/P ≥ 30%) ELISA test results, with the aim of limiting contagion. The median within-herd prevalence of

test-positive cows in the UK is approximately 0.02 (Taylor et al., 2024). Figure 7.5 illustrates the distribution of within-herd prevalence of state 3 cows, for the herds in the data used to train the HMM. The within-herd prevalence of state 3 cows in these data is approximately four times higher (median 0.086) than the national average prevalence of test-positive cows.

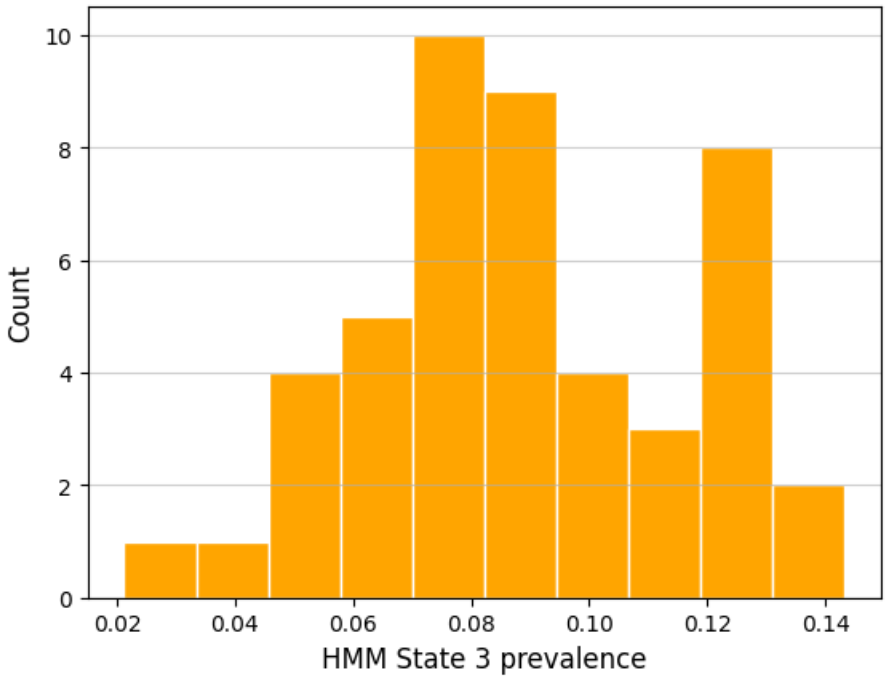

Figure 7.5: Within-herd prevalence of HMM state 3 cows for the herds included in the data used to train a hidden Markov model for interpretation of serial MAP antibody ELISA results.

It is unknown currently to what extent state 3 cows are contagious, but given that cows can shed MAP in faeces prior to ever receiving a positive (S/P ≥ 30%) ELISA result (Nielsen and Toft, 2006), it is probable that some state 3 cows, at a heightened risk of progression to a state in which they test positive, are also

contagious. The considerably higher prevalence of state 3 cows relative to test-positive cows suggests that, even in the scenario in which state 3 cows shed MAP to a lesser extent, they could collectively represent a risk of contagion comparable to that posed by test-positive cows. Furthermore, the majority of state 3 cows are not identified as high-risk ("J4" or "J5") by the prevailing UK pTB control framework (Figure 5.2). Use of the Chapter 4 pipeline for interpreting serial ELISA results has potential to identify a greater proportion of contagious cows, allowing farmers to reduce transmission through appropriate management.

One of the prominent outcomes of this thesis is the ability to predict the future pTB-status of cows. The Elastic Net model of Chapter 3 provides reasonably well-calibrated predictions, based on information available in early life, of the probability of an animal ultimately entering a state in which it repeatedly tests positive (S/P ≥ 30%). Despite being poorly calibrated (Figure 3.9), the posterior probabilities of the entire pipeline in Chapter 3 did tend to rise prior to the onset of positive status (Figure 3.8). The HMM of Chapter 4 capitalises on this signal (non-negligible but technically negative (S/P < 30%) ELISA results), providing predictions that an animal will enter state 4 (associated with positive ELISA results) in the future. The model in Chapter 5 then utilises HMM state probabilities to provide well-calibrated predictions of future "J5" status. Prediction of future pTB status has scarcely been attempted previously (Imada et al., 2024), but could be highly useful for alerting farmers at an earlier stage of potentially contagious cows.

7.5. Limitations of this thesis

The datasets analysed in this thesis were generally of modest size and, excepting Chapter 2, were convenience samples taken from farms milk-recording with one laboratory (QMMS LTD., UK). The herds included were on commercial dairy farms distributed across the UK, and so were assumed to be representative of the wider dairy cow population. Data comprising the same 47 farms was used for training of the models in Chapters 3 to 6 (whose focus was paratuberculosis), although a second dataset of a further 48 QMMS farms was used to validate the models in Chapter 5. MAP antibody ELISA test results are known to vary between laboratories (Barden et al., 2020). Therefore, the extent to which results generalise to the UK or wider populations of dairy cows remains unclear.

Regarding Chapter 2, the data were taken from farms milk-recording with a different organisation, but still comprised only a small number of farms. In addition, the data used in Chapter 2 were sampled approximately twenty years prior to this thesis. They were used in this thesis owing to their perceived accuracy, which was a product of the original study design (Green et al., 2007); farmers faithfully recorded every case of clinical mastitis occurring during the study period. This is in contrast to routinely-collected farm records of CM cases, in which it is suspected that farmers sometimes under-record the number of CM cases which have occurred (Leach et al., 2024). It is plausible that aspects of clinical mastitis have changed over the past two decades. For example, the prevalence of different pathogens and their exact nature (e.g. virulence, pathogenicity, antimicrobial sensitivity), and commonly-used treatment protocols, could have altered in this time, which in

turn could affect cure rates and therefore the accuracy of the model's predictions on contemporary data. Evaluation of the models in this thesis against some larger contemporary datasets originating from different laboratories would be useful.

The definition of targets for the predictive models should also be scrutinised. In all predictive models, the target was a proxy measure of the disease status of cows. In Chapter 2, the definition of clinical mastitis cure was based largely on somatic cell counts (cytological cure), rather than bacteriological cure *per se*. In Chapters 3, 4 and 5, the target was derived from MAP ELISA results, which are a proxy for pTB infection. Use of proxy measures of disease status, whilst commonly used for practical and economic reasons, may have introduced some bias into the use of model predictions as indicators of actual disease status.

7.6. Future research

7.6.1. Paratuberculosis

The implications of the HMM states have been partially explored. This thesis establishes that state 3 cows suffer production losses (Table 6.4) and are at relatively high risk of progression to state 4 (Table 4.3) and "J5" status (Table 5.4). This is suggestive that state 3 cows can also shed MAP, but further research is required to confirm this hypothesis. Suggested future studies could include measuring the concentration of MAP in faeces of cows in different states. Investigating if there is a correlation between the states of cows and their own calves, or between cows and any calves sharing an environment during early life, would also be useful for establishing if state 3 cows represent an infection risk to young calves.

It would be of interest to further investigate factors associated with transitions between states, particularly from state 3 to state 4. In the HMM in Chapter 4, transition probabilities were associated with parity. Inclusion of further covariates (e.g. recent health events, DIM, yield, breed) that modify the transition probabilities could lead to further understanding of the pTB disease process, as well as providing more accurate predictions as to which cows will progress to state 4. In this thesis, inclusion of additional covariates on the transition probabilities led to non-convergence of the HMM when fitting using maximum likelihood methods. Use of Bayesian (e.g. Markov Chain Monte Carlo) techniques instead may allow the fitting of more complex HMMs. Further, Bayesian techniques could be highly useful for obtaining models tailored to subpopulations of cows (e.g. of a particular breed, in a specific geographic region or even in an individual herd); the posterior probabilities of a global model (trained on a large heterogenous dataset) could be used as priors in a HMM fit to data from the subpopulation of interest, thereby providing a model for a specific use-case (e.g. in one large herd, or for cows of a particular breed) (Su et al., 2018).

The impact of deploying the HMM on farms should also be assessed. The ability of the model to disclose potentially pTB-infected animals at an earlier stage would be expected to lead to a reduction in within-herd pTB prevalence, providing that farmers manage such cows appropriately to limit contagion. This hypothesis should now be tested, for example by monitoring herd-wide MAP ELISA results in longitudinal data collected on participating farms.

7.6.2. Mastitis

The model in Chapter 2 predicts the probability of cure of a case of CM if treated with antimicrobials. Estimates of cure probability given other treatment protocols, for example non-antibiotic treatments such as anti-inflammatory drugs, would be highly useful, providing information to farmers regarding the loss of cure potential resulting from withholding of antimicrobial treatment. Under the assumption that Gram-negative IMIs benefit little from antimicrobial treatment, several studies have been conducted investigating cure rates in Gram-negative infections when not using antimicrobials (de Jong et al., 2023b). Collection of the required data describing cure rates of Gram-positive IMIs in the absence of antimicrobials is difficult due to ethical concerns, as Gram-positive infections are generally considered to benefit from antimicrobials (de Jong et al., 2023b). However, recent analysis of such data has indeed been reported (Krömker et al., 2021). The impact of deploying the model in Chapter 2 on farms is expected to be a reduction of AMU in cows with a low cure probability (e.g. chronic IMIs in older cows). The potential drop in cure rates when withholding antimicrobials from such cows cannot be large given the low probability of cure with antimicrobial treatment. However, for farmers who wish to withhold antimicrobials from Gram negative infections (for example through use of on-farm culture), quantification of the loss of cure potential caused by withholding antimicrobials from these CM cases would allow more-informed decision making. In addition, creation of models that predict cure probability given varied treatments would permit a full financial cost-benefit analysis of different treatment protocols in individual

cases or at the herd-level, accounting for the consequences of either achieving or failing to achieve a cure in each CM case.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19, 716–723. https://doi.org/10.1109/TAC.1974.1100705

Aly, S.S., Anderson, R.J., Adaska, J.M., Jiang, J., Gardner, I.A., 2010. Association between Mycobacterium avium subspecies paratuberculosis infection and milk production in two California dairies. J. Dairy Sci. 93, 1030–1040. https://doi.org/10.3168/jds.2009-2611

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Stat. Surv. 4, 40–79. https://doi.org/10.1214/09-SS054

Austin, P.C., Merlo, J., 2017. Intermediate and advanced topics in multilevel logistic regression analysis. Stat. Med. 36, 3257–3277. https://doi.org/10.1002/sim.7336

Balayla, J., 2022. Bayesian updating and sequential testing: overcoming inferential limitations of screening tests. BMC Med. Inform. Decis. Mak. 22, 6. https://doi.org/10.1186/s12911-021-01738-w

Banhazi, T.M., Lehr, H., Black, J.L., Crabtree, H., Schofield, P., Tscharke, M., Berckmans, D., 2012. Precision Livestock Farming: An international review of scientific and commercial aspects. International Journal of Agricultural and Biological Engineering 5, 1–10. https://doi.org/10.3965/j.ijabe.20120503.00?

Barden, M., Smith, R.F., Higgins, H.M., 2020. The interpretation of serial Johne's disease milk antibody results is affected by test characteristics, pattern of test results and parallel bovine tuberculosis testing. Prev. Vet. Med. 183, 105134. https://doi.org/10.1016/j.prevetmed.2020.105134

Barkema, H.W., Orsel, K., Nielsen, S.S., Koets, A.P., Rutten, V.P.M.G., Bannantine, J.P., Keefe, G.P., Kelton, D.F., Wells, S.J., Whittington, R.J., Mackintosh, C.G., Manning, E.J., Weber, M.F., Heuer, C., Forde, T.L., Ritter, C., Roche, S., Corbett, C.S., Wolf, R., Griebel, P.J., Kastelic, J.P., De Buck, J., 2018. Knowledge gaps that hamper prevention and control of Mycobacterium avium subspecies paratuberculosis infection. Transbound. Emerg. Dis. 65, 125–148. https://doi.org/10.1111/tbed.12723

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using lme4. J. Stat. Softw. 67, 1–48. https://doi.org/10.18637/jss.v067.i01

Bates, S., Hastie, T., Tibshirani, R., 2024. Cross-Validation: What Does It Estimate and How Well Does It Do It? J. Am. Stat. Assoc. 119, 1434–1445. https://doi.org/10.1080/01621459.2023.2197686

Bayes, T., 1763. Bayes 1763 An essay towards solving a problem in the doctrine of chances. https://academic.oup.com/biomet/article-abstract/45/3-4/296/234123

Beaudeau, F., Belliard, M., Joly, A., Seegers, H., 2007. Reduction in milk yield associated with Mycobacterium avium subspecies paratuberculosis (Map) infection in dairy cows. Vet. Res. 38, 625–634. https://doi.org/10.1051/vetres:2007021

Beaver, A., Sweeney, R.W., Hovingh, E., Wolfgang, D.R., Gröhn, Y.T., Schukken, Y.H., 2017. Longitudinal relationship between fecal culture, fecal quantitative PCR, and milk ELISA in Mycobacterium avium ssp. paratuberculosis-infected cows from low-prevalence dairy herds. J. Dairy Sci. 100, 7507–7521. https://doi.org/10.3168/jds.2017-12928

Begg, D.J., Plain, K.M., de Silva, K., Gurung, R., Gunn, A., Purdie, A.C., Whittington, R.J., 2018. Immunopathological changes and apparent recovery from infection revealed in cattle in an experimental model of Johne's disease using a lyophilised culture of Mycobacterium avium subspecies paratuberculosis. Vet. Microbiol. 219, 53–62. https://doi.org/10.1016/j.vetmic.2018.03.029

Bell, A., Fairbrother, M., Jones, K., 2019. Fixed and random effects models: making an informed choice. Qual. Quant. 53, 1051–1074. https://doi.org/10.1007/s11135-018-0802-x

Berckmans, D., 2017. General introduction to precision livestock farming. Animal Frontiers 7, 6–11. https://doi.org/10.2527/af.2017.0102

Berckmans, D., 2014. Precision livestock farming technologies for welfare management in intensive livestock systems. OIE

Revue Scientifique et Technique 33, 189–196. https://doi.org/10.20506/rst.33.1.2273

Bewick, V., Cheek, L., Ball, J., 2005. Statistics review 14: Logistic regression. Crit. Care 9, 112-118. https://doi.org/10.1186/cc3045

Bi, Q., Goodman, K.E., Kaminsky, J., Lessler, J., 2019. What is machine learning? A primer for the epidemiologist. Am. J. Epidemiol. 188, 2222–2239. https://doi.org/10.1093/aje/kwz189

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., White, J.S.S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. 24, 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Bradley, A.J., 2002. Bovine Mastitis: An Evolving Disease. The Veterinary Journal 164, 116–128. https://doi.org/10.1053/tvjl.2002.0724

Bradley, A.J., Green, M.J., 2009. Factors affecting cure when treating bovine clinical mastitis with cephalosporin-based intramammary preparations. J. Dairy Sci. 92, 1941–1953. https://doi.org/10.3168/jds.2008-1497

Bradley, A.J., Green, M.J., 2001. Adaptation of Escherichia coli to the bovine mammary gland. J. Clin. Microbiol. 39, 1845–1849. https://doi.org/10.1128/JCM.39.5.1845-1849.2001

Bradshaw, T.J., Huemann, Z., Hu, J., Rahmim, A., 2023. A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging.

Radiol. Artif. Intell. 5, e220232. https://doi.org/10.1148/ryai.220232

Buller, H., Blokhuis, H., Jensen, P., Keeling, L., 2018. Towards farm animal welfare and sustainability. Animals 8, 1–13. https://doi.org/10.3390/ani8060081

Bzdok, D., Ioannidis, J.P.A., 2019. Exploration, Inference, and Prediction in Neuroscience and Biomedicine. Trends Neurosci. 42, 251-262. https://doi.org/10.1016/j.tins.2019.02.001

Cao, L., Zhao, S., Li, Q., Ling, L., Wu, W.K.K., Zhang, L., Lou, J., Chong, M.K.C., Chen, Z., Wong, E.L.Y., Zee, B.C.Y., Chan, M.T.V., Chan, P.K.S., Wang, M.H., 2021. A Bayesian method for synthesizing multiple diagnostic outcomes of COVID-19 tests. R. Soc. Open Sci. 8, 201867. https://doi.org/10.1098/rsos.201867

Capper, J.L., Bauman, D.E., 2013. The role of productivity in improving the environmental sustainability of ruminant production systems. Annu. Rev. Anim. Biosci. 1, 469–489. https://doi.org/10.1146/annurev-animal-031412-103727

Capper, J.L., Cady, R.A., Bauman, D.E., 2009. The environmental impact of dairy production: 1944 compared with 2007. J. Anim. Sci. 87, 2160–2167. https://doi.org/10.2527/jas.2009-1781

Capper, J.L., Williams, P., 2023. Investing in health to improve the sustainability of cattle production in the United Kingdom: A

narrative review. Veterinary Journal 296–297, 105988. https://doi.org/10.1016/j.tvjl.2023.105988

Ceres, K.M., Schukken, Y.H., Gröhn, Y.T., 2020. Characterizing Christodoulou infectious disease progression through discrete states using hidden Markov models. PLoS One 15, e0242683. https://doi.org/10.1371/journal.pone.0242683

Chan, S.F., Deeks, J.J., Macaskill, P., Irwig, L., 2008. Three methods to construct predictive models using logistic regression and likelihood ratios to facilitate adjustment for pretest probability give similar results. J. Clin. Epidemiol. 61, 52–63. https://doi.org/10.1016/j.jclinepi.2007.02.012

Christodoulou, E., Ma, J., Collins, G.S., Steyerberg, E.W., Verbakel, J.Y., Van Calster, B., 2019. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J. Clin. Epidemiol. 110, 12-22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Coad, M., Clifford, D.J., Vordermeier, H.M., Whelan, A.O., 2013. The consequences of vaccination with the Johne's disease vaccine, Gudair, on diagnosis of bovine tuberculosis. Veterinary Record 172, 266. https://doi.org/10.1136/vr.101201

Collins, M.T., 2002. Interpretation of a commercial bovine paratuberculosis enzyme-linked immunosorbent assay by using likelihood ratios. Clin. Diagn. Lab. Immunol. 9, 1367–1371. https://doi.org/10.1128/CDLI.9.6.1367-1371.2002

Corbiere, F., Guellouz, D., Tasca, C., Foures, L., Dubaux, E., Foucras, G., 2023. Effects of Silirum®-Based Vaccination

Programs on Map Fecal Shedding and Serological Response in Seven French Dairy Herds. Animals 13, 1569. https://doi.org/10.3390/ani13091569

Cruz, L.V.F., Daros, R.R., Ostrensky, A., 2025. Factors Associated with Cure and Prediction of Cure of Clinical Mastitis of Dairy Cows. Dairy 6, 37. https://doi.org/10.3390/dairy6040037

De Finetti, B., 1974. Bayesianism: Its Unifying Role for Both the Foundations and Applications of Statistics', Int. Stat. Rev. 42, 117.

de Haas, Y., Veerkamp, R.F., Barkema, H.W., Gröhn, Y.T., Schukken, Y.H., 2004. Associations Between Pathogen-Specific Cases of Clinical Mastitis and Somatic Cell Count Patterns. J. Dairy Sci. 87, 95–105. https://doi.org/10.3168/jds.S0022-0302(04)73146-X

de Jong, E., Creytens, L., De Vliegher, S., McCubbin, K.D., Baptiste, M., Leung, A.A., Speksnijder, D., Dufour, S., Middleton, J.R., Ruegg, P.L., Lam, T.J.G.M., Kelton, D.F., McDougall, S., Godden, S.M., Lago, A., Rajala-Schultz, P.J., Orsel, K., Krömker, V., Kastelic, J.P., Barkema, H.W., 2023a. Selective treatment of nonsevere clinical mastitis does not adversely affect cure, somatic cell count, milk yield, recurrence, or culling: A systematic review and meta-analysis. J. Dairy Sci. 106, 1267–1286. https://doi.org/10.3168/jds.2022-22271

de Jong, E., McCubbin, K.D., Speksnijder, D., Dufour, S., Middleton, J.R., Ruegg, P.L., Lam, T.J.G.M., Kelton, D.F., McDougall, S., Godden, S.M., Lago, A., Rajala-Schultz, P.J., Orsel, K., De

Vliegher, S., Krömker, V., Nobrega, D.B., Kastelic, J.P., Barkema, H.W., 2023b. Invited review: Selective treatment of clinical mastitis in dairy cattle. J. Dairy Sci. 106, 3761-3778. https://doi.org/10.3168/jds.2022-22826

de Wit, J., Oldenbroek, J.K., van Keulen, H., Zwart, D., 1995. Criteria for sustainable livestock production: a proposal for implementation. Agric. Ecosyst. Environ. 53, 219–229. https://doi.org/10.1016/0167-8809(94)00579-4

Dean, C.B., Nielsen, J.D., 2007. Generalized linear mixed models: A review and some extensions. Lifetime Data Anal. 13, 497–512. https://doi.org/10.1007/s10985-007-9065-x

Degen, S., Paduch, J.H., Hoedemaker, M., Krömker, V., 2015. Einflussfaktoren auf die bakteriologische heilung boviner mastitiden. Tierarztliche Praxis Ausgabe G: Grosstiere - Nutztiere 43, 222–227. https://doi.org/10.15653/TPG-141082

Deluyker, H.A., Chester, S.T., Van Oye, S.N., 1999. A multilocation clinical trial in lactating dairy cows affected with clinical mastitis to compare the efficacy of treatment with intramammary infusions of a lincomycin/neomycin combination with an ampicillin/cloxacillin combination. J. Vet. Pharmacol. Ther. 22, 274–282. https://doi.org/10.1046/j.1365-2885.1999.00205.x

Detilleux, J.C., 2011. A hidden Markov model to predict early mastitis from test-day somatic cell scores. Animal 5, 175–181. https://doi.org/10.1017/S1751731110001746

Ding, J., Tarokh, V., Yang, Y., 2018. Model Selection Techniques: An Overview. IEEE Signal Process. Mag. 35, 16–34. https://doi.org/10.1109/MSP.2018.2867638

Down, P.M., Bradley, A.J., Breen, J.E., Green, M.J., 2017. Factors affecting the cost-effectiveness of on-farm culture prior to the treatment of clinical mastitis in dairy cows. Prev. Vet. Med. 145, 91–99. https://doi.org/10.1016/j.prevetmed.2017.07.006

Dumre, P., Bhattarai, S., Shashikala, H.K., 2024. Optimizing Linear Regression Models: A Comparative Study of Error Metrics, in: Proceedings - 4th International Conference on Technological Advancements in Computational Sciences, ICTACS 2024. Institute of Electrical and Electronics Engineers Inc., pp. 1856–1861. https://doi.org/10.1109/ICTACS62700.2024.10840719

Dyar, O.J., Huttner, B., Schouten, J., Pulcini, C., 2017. What is antimicrobial stewardship? Clinical Microbiology and Infection 23, 793-798. https://doi.org/10.1016/j.cmi.2017.08.026

Dziak, J.J., Coffman, D.L., Lanza, S.T., Li, R., Jermiin, L.S., 2020. Sensitivity and specificity of information criteria. Brief. Bioinform. 21, 553-565. https://doi.org/10.1093/bib/bbz016

Eisenberg, S.W.F., Veldman, E., Rutten, V.P.M.G., Koets, A.P., 2015. A longitudinal study of factors influencing the result of a Mycobacterium avium ssp. paratuberculosis antibody ELISA in milk of dairy cows. J. Dairy Sci. 98, 2345–2355. https://doi.org/10.3168/jds.2014-8380

Erickson, B.J., Kitamura, F., 2021. Magician's corner: 9. performance metrics for machine learning models. Radiol. Artif. Intell. 3, e200126. https://doi.org/10.1148/ryai.2021200126

Espejo, L.A., Godden, S., Hartmann, W.L., Wells, S.J., 2012. Reduction in incidence of Johne's disease associated with implementation of a disease control program in Minnesota demonstration herds. J. Dairy Sci. 95, 4141–4152. https://doi.org/10.3168/jds.2011-4550

Fanshawe, T.R., Nicholson, B.D., Perera, R., Oke, J.L., 2024. A review of methods for the analysis of diagnostic tests performed in sequence. Diagn. Progn. Res. 8. https://doi.org/10.1186/s41512-024-00175-3

FAO, 2022. Greenhouse gas emissions from agrifood systems Global, regional and country trends, 2000-2020. https://openknowledge.fao.org/items/74bfebdb-3272-4e6a-98f4-ee36c7146d44

FAO, 2017. The future of food and agriculture: trends and challenges. https://openknowledge.fao.org/server/api/core/bitstreams/2e90c833-8e84-46f2-a675-ea2d7afa4e24/content

FAO, 2024. Food Outlook Biannual Report on Global Food Markets. https://doi.org/10.4060/cd1158en

Fenlon, C., O'Grady, L., Doherty, M.L., Dunnion, J., 2018. A discussion of calibration techniques for evaluating binary and

categorical predictive models. Prev. Vet. Med. 149, 107–114. https://doi.org/10.1016/j.prevetmed.2017.11.018

Fischer, F., 2021. Using Bayes theorem to estimate positive and negative predictive values for continuously and ordinally scaled diagnostic tests. Int. J. Methods Psychiatr. Res. 30, e1868. https://doi.org/10.1002/mpr.1868

Fitzpatrick, C.E., Chapinal, N., Petersson-Wolfe, C.S., DeVries, T.J., Kelton, D.F., Duffield, T.F., Leslie, K.E., 2013. The effect of meloxicam on pain sensitivity, rumination time, and clinical signs in dairy cows with endotoxin-induced clinical mastitis. J. Dairy Sci. 96, 2847–2856. https://doi.org/10.3168/jds.2012-5855

Friedrich, S., Groll, A., Ickstadt, K., Kneib, T., Pauly, M., Rahnenführer, J., Friede, T., 2023. Regularization approaches in clinical biostatistics: A review of methods and their applications. Stat. Methods Med. Res. 32, 425-440. https://doi.org/10.1177/09622802221133557

Fuenzalida, M.J., Ruegg, P.L., 2019a. Negatively controlled, randomized clinical trial to evaluate intramammary treatment of nonsevere, gram-negative clinical mastitis. J. Dairy Sci. 102, 5438–5457. https://doi.org/10.3168/jds.2018-16156

Fuenzalida, M.J., Ruegg, P.L., 2019b. Short communication: Longitudinal study of quarter-level somatic cell responses after naturally occurring, nonsevere clinical mastitis diagnosed as culture negative, or caused by Escherichia coli or Klebsiella pneumoniae, and randomly assigned to a no-

treatment group or to receive intramammary ceftiofur. J. Dairy Sci. 102, 11476–11482. https://doi.org/10.3168/jds.2018-16190

Garcia, A.B., Shalloo, L., 2015. Invited review: The economic impact and control of paratuberculosis in cattle. J. Dairy Sci. 98, 5019–5039. https://doi.org/10.3168/jds.2014-9241

Gefeller, O., Hofner, B., Mayr, A., Waldmann, E., 2017. Predictive Modelling Based on Statistical Learning in Biomedicine. Comput. Math. Methods Med. https://doi.org/10.1155/2017/4041736

Gelman, A., Meng, X.-L., Stern, H., 1996. Posterior predictive assessment of model fitness via realized discrepancies. Stat. Sin. 6, 733-807.

Gelman, A., Shalizi, C.R., 2013. Philosophy and the practice of Bayesian statistics. British Journal of Mathematical and Statistical Psychology 66, 8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x

Gerber, P.J., Steinfeld, H., Henderson, B., Mottet, A., Opio, C., Dijkman, J., Falcucci, A., Tempio, G., 2013. Tackling climate change through livestock – A global assessment of emissions and mitigation opportunities. Food and Agriculture Organization of the United Nations (FAO), Rome. https://www.fao.org/4/i3437e/i3437e.pdf

Gerß, J.W.O., Vonthein, R., 2025. Introduction to Bayesian statistics: Part 36 of a series on the evaluation of scientific

publications. Dtsch. Arztebl. Int. 122, 271–276. https://doi.org/10.3238/arztebl.m2025.0035

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102, 359–378. https://doi.org/10.1198/016214506000001437

Goligher, E.C., Heath, A., Harhay, M.O., 2024. Bayesian statistics for clinical research. The Lancet 404, 1067-1076. https://doi.org/10.1016/S0140-6736(24)01295-9

Gonda, M.G., Chang, Y.M., Shook, G.E., Collins, M.T., Kirkpatrick, B.W., 2007. Effect of Mycobacterium paratuberculosis infection on production, reproduction, and health traits in US Holsteins. Prev. Vet. Med. 80, 103–119. https://doi.org/10.1016/j.prevetmed.2007.01.011

Green, M.J., Leach, K.A., Breen, J.E., Green, L.E., Bradley, A.J., 2007. National intervention study of mastitis control in dairy herds in England and Wales. Vet. Rec. 160, 287–293. https://doi.org/10.1136/vr.160.9.287

Griffiths, T.L., Tenenbaum, J.B., 2006. Optimal Predictions in Everyday Cognition. Psychol. Sci. 17, 767–773. https://doi.org/10.1111/j.1467-9280.2006.01780.x

Griss, S., Knific, T., Buzzell, A., Carmo, L.P., Schüpbach-Regula, G., Meylan, M., Ocepek, M., Thomann, B., 2024. A scoping review on associations between paratuberculosis and productivity in cattle. Front. Vet. Sci. 11, 1352623. https://doi.org/10.3389/fvets.2024.1352623

Hanks, J., Kossaibati, M., 2025. Key performance indicators (KPIs) for the UK national dairy herd. A study of herd performance in 500 Holstein/Friesian herds for the year ending 31st August 2024. Veterinary Epidemiology and Economics Research Unit, University of Reading, UK. https://panlivestock.com/wp-content/uploads/2025/01/NMR500Herds-2024_Final_Edited_10Jan25.pdf

Harrison, X.A., Donaldson, L., Correa-Cano, M.E., Evans, J., Fisher, D.N., Goodwin, C.E.D., Robinson, B.S., Hodgson, D.J., Inger, R., 2018. A brief introduction to mixed effects modelling and multi-model inference in ecology. PeerJ 2018. https://doi.org/10.7717/peerj.4794

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, Springer Series in Statistics. Springer New York, New York, NY. https://doi.org/10.1007/b94608

Haug, C.J., Drazen, J.M., 2023. Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. New England Journal of Medicine 388, 1201–1208. https://doi.org/10.1056/NEJMra2302038

Hawkins, D.M., 2004. The Problem of Overfitting. J. Chem. Inf. Comput. Sci. 44, 1-12. https://doi.org/10.1021/ci0342472

Heinze, G., Wallisch, C., Dunkler, D., 2018. Variable selection – A review and recommendations for the practicing statistician. Biometrical Journal. 60, 431-449. https://doi.org/10.1002/bimj.201700067

Hendrick, S.H., Kelton, D.F., Leslie, K.E., Lissemore, K.D., Archambault, M., Duffield, T.F., 2005. Effect of paratuberculosis on culling, milk production, and milk quality in dairy herds. J. Am. Vet. Med. Assoc. 227, 1302–1308. https://doi.org/10.2460/javma.2005.227.1302

Hirst, H.L., Garry, F.B., Salman, M.D., 2002. Assessment of test results when using a commercial enzyme-linked immunosorbent assay for diagnosis of paratuberculosis in repeated samples collected from adult dairy cattle. J. Am. Vet. Med. Assoc. 220, 1685–1689. https://doi.org/10.2460/javma.2002.220.1685

Hoffman, L., Walters, R.W., 2022. Catching Up on Multilevel Modeling Cross-level interaction: a model-estimated constant slope for the multiplicative combination of two predictor variables measured for different dimensions of sampling. Annu. Rev. Psychol. 73, 659-689. https://doi.org/10.1146/annurev-psych-020821-103525

Hoffman, M.D., Gelman, A., 2014. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo, J. Mach. Learn. Res. 15, 1593-1623.

Hosmer, D.W.., Lemeshow, S., 2002. Applied logistic regression. Wiley.

ICAR, 2023. International Council for Animal Recording Guidelines. Section 2-Guidelines for Dairy Cattle Milk Recording Section 2-Dairy Cattle Milk Recording.

https://www.icar.org/Guidelines/02-Overview-Cattle-Milk-Recording.pdf

ICAR, 2014. International Agreement of Recording Practices ICAR Recording Guidelines. https://pecuaria.pt/docs/Guidelines_2014.pdf

Imada, J., Arango-Sabogal, J.C., Bauman, C., Roche, S., Kelton, D., 2024. Comparison of Machine Learning Tree-Based Algorithms to Predict Future Paratuberculosis ELISA Results Using Repeat Milk Tests. Animals 14, 1113. https://doi.org/10.3390/ani14071113

Johnson-Ifearulundu, Y.J., Kaneene, J.B., Sprecher, D.J., Gardiner, J.C., Lloyd, J.W., 2000. The effect of subclinical Mycobacterium paratuberculosis infection on days open in Michigan, USA, dairy cows. Prev. Vet. Med. 46, 171–181. https://doi.org/10.1016/S0167-5877(00)00145-8

Kapoor, S., Narayanan, A., 2023. Leakage and the reproducibility crisis in machine-learning-based science. Patterns 4, 100804. https://doi.org/10.1016/j.patter.2023.100804

Kennedy, A.E., Byrne, N., Garcia, A.B., O'Mahony, J., Sayers, R.G., 2016. Analysis of Johne's disease ELISA status and associated performance parameters in Irish dairy cows. BMC Vet. Res. 12, 43. https://doi.org/10.1186/s12917-016-0667-y

Kingshay, 2025. Dairy costings focus annual report. https://www.kingshay.com/wp-content/uploads/Dairy-Costings-Focus-Report-2025.pdf

Kleen, J.L., Guatteo, R., 2023. Precision Livestock Farming: What Does It Contain and What Are the Perspectives? Animals 13, 779. https://doi.org/10.3390/ani13050779

Kock, J., Wente, N., Zhang, J., Paduch, S., Leimbach, D., Klocke, D., Gelfert, C.C., Kromker, V., 2018. Udder health effects of an evidence-based mastitis therapy concept in Northwestern Germany. Milk Science International - Milchwissenschaft 71, 14–20.

Koets, A.P., Eda, S., Sreevatsan, S., 2015. The within host dynamics of Mycobacterium avium ssp. paratuberculosis infection in cattle: Where time and place matter. Vet. Res. 46, 61. https://doi.org/10.1186/s13567-015-0185-0

Köhler H., Wichert A., Donat K., 2022. Variation in the Performance of Different Batches of Two Mycobacterium avium Subspecies paratuberculosis Antibody ELISAs Used for Pooled Milk Samples. Animals 12, 442. https://doi.org/10.3390/ani12040442

Krömker, V., Leimbach, S., 2017. Mastitis treatment—Reduction in antibiotic usage in dairy cows. Reproduction in Domestic Animals 52, 21–29. https://doi.org/10.1111/rda.13032

Krömker, V., Schmenger, A., Klocke, D., Mansion-de Vries, E.M., Wente, N., Zhang, Y., Leimbach, S., 2021. Non-inferiority Trial Investigating the Efficacy of Non-steroidal Anti-inflammatory Drugs and Antimicrobial Treatment of Mild to Moderate Clinical Mastitis in Dairy Cows With Long-lasting Udder

Diseases. Front. Vet. Sci. 8, 660804. https://doi.org/10.3389/fvets.2021.660804

Krstajic, D., Buturovic, L.J., Leahy, D.E., Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. J. Cheminform. 6, 10. https://doi.org/10.1186/1758-2946-6-10

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4614-6849-3

Kuipers, A., Koops, W.J., Wemmenhove, H., 2016. Antibiotic use in dairy herds in the Netherlands from 2005 to 2012. J. Dairy Sci. 99, 1632–1648. https://doi.org/10.3168/jds.2014-8428

Kyriazakis, I., Arndt, C., Aubry, A., Charlier, J., Ezenwa, V.O., Godber, O.F., Krogh, M., Mostert, P.F., Orsel, K., Robinson, M.W., Ryan, F.S., Skuce, P.J., Takahashi, T., Van Middelaar, C.E., Vigors, S., Morgan, E.R., 2024. Improve animal health to reduce livestock emissions: quantifying an open goal. Proceedings of the Royal Society B: Biological Sciences 291, 20240675. https://doi.org/10.1098/rspb.2024.0675

Lago, A., Godden, S.M., Bey, R., Ruegg, P.L., Leslie, K., 2011. The selective treatment of clinical mastitis based on on-farm culture results: I. Effects on antibiotic use, milk withholding time, and short-term clinical and bacteriological outcomes. J. Dairy Sci. 94, 4441–4456. https://doi.org/10.3168/jds.2010-4046

Lam, T.J.G.M., Van Den Borne, B.H.P., Jansen, J., Huijps, K., Van Veersen, J.C.L., Van Schaik, G., Hogeveen, H., 2013. Improving bovine udder health: A national mastitis control program in the Netherlands. J. Dairy Sci. 96, 1301–1311. https://doi.org/10.3168/jds.2012-5958

Lardé, H., Dufour, S., Archambault, M., Massé, J., Roy, J.P., Francoz, D., 2021. An observational cohort study on antimicrobial usage on dairy farms in Quebec, Canada. J. Dairy Sci. 104, 1864–1880. https://doi.org/10.3168/jds.2020-18848

Laurin, E.L., Sanchez, J., Chaffer, M., McKenna, S.L.B., Keefe, G.P., 2017. Assessment of the relative sensitivity of milk ELISA for detection of Mycobacterium avium ssp. paratuberculosis infectious dairy cows. J. Dairy Sci. 100, 598–607. https://doi.org/10.3168/jds.2016-11194

Leach, K.A., Holsey, H.J., Bradley, A.J., Green, M.J., 2024. Improvement of mammary gland health in 81 'sentinel herds' in England and Scotland between 2012 and 2021. Veterinary Record 194, e3605. https://doi.org/10.1002/vetr.3605

Leite de Campos, J., Kates, A., Steinberger, A., Sethi, A., Suen, G., Shutske, J., Safdar, N., Goldberg, T., Ruegg, P.L., 2021. Quantification of antimicrobial usage in adult cows and preweaned calves on 40 large Wisconsin dairy farms using dose-based and mass-based metrics. J. Dairy Sci. 104, 4727–4745. https://doi.org/10.3168/jds.2020-19315

Leos-Barajas, V., Photopoulou, T., Langrock, R., Patterson, T.A., Watanabe, Y.Y., Murgatroyd, M., Papastamatiou, Y.P., 2017.

Analysis of animal accelerometer data using hidden Markov models. Methods Ecol. Evol. 8, 161–173. https://doi.org/10.1111/2041-210X.12657

Lombard, J.E., Garry, F.B., Mccluskey, B.J., Wagner, B.A., 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

Louzoun, Y., Mitchell, R., Behar, H., Schukken, Y., 2015. Two state model for a constant disease hazard in paratuberculosis (and other bovine diseases). Vet. Res. 46, 67. https://doi.org/10.1186/s13567-015-0189-9

Lovarelli, D., Bacenetti, J., Guarino, M., 2020. A review on dairy cattle farming: Is precision livestock farming the compromise for an environmental, economic and social sustainable production? J. Clean. Prod. 262, 121409. https://doi.org/10.1016/j.jclepro.2020.121409

Machado, G., Kanankege, K., Schumann, V., Wells, S., Perez, A., Alvarez, J., 2018. Identifying individual animal factors associated with Mycobacterium avium subsp. paratuberculosis (MAP) milk ELISA positivity in dairy cattle in the Midwest region of the United States. BMC Vet. Res. 14, 28. https://doi.org/10.1186/s12917-018-1354-y

Magee, D.R., Boyle, R.D., 2002. Detecting lameness using 'Re-sampling Condensation' and 'multi-stream cyclic hidden

Markov models.' Image Vis. Comput. 20, 581–594. https://doi.org/10.1016/S0262-8856(02)00047-1

Maity, S., Ambatipudi, K., 2021. Mammary microbial dysbiosis leads to the zoonosis of bovine mastitis: A One-Health perspective. FEMS Microbiol. Ecol. 97, fiaa241. https://doi.org/10.1093/femsec/fiaa241

Mansion-de Vries, E.M., Lücking, J., Wente, N., Zinke, C., Hoedemaker, M., Krömker, V., 2016. Comparison of an evidence-based and a conventional mastitis therapy concept with regard to cure rates and antibiotic usage. Milk Science International - Milchwissenschaft 69, 27–32.

Marcé, C., Beaudeau, F., Bareille, N., Seegers, H., Fourichon, C., 2009. Higher non-return rate associated with Mycobacterium avium subspecies paratuberculosis infection at early stage in Holstein dairy cows. Theriogenology 71, 807–816. https://doi.org/10.1016/j.theriogenology.2008.10.017

Martins, E., Oliveira, P., Correia-Gomes, C., Mendonça, D., Ribeiro, J.N., 2024. Association of Mycobacterium avium paratuberculosis serostatus with age at first calving, calving interval, and milk production in dairy cows. J. Dairy Sci. 107, 3916–3926. https://doi.org/10.3168/jds.2023-23516

McAloon, C.G., O'Grady, L., Botaro, B., More, S.J., 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

McAloon, C.G., Whyte, P., More, S.J., Green, M.J., O'Grady, L., Garcia, A.B., Doherty, M.L., 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

McDougall, S., 2003. Intramammary treatment of clinical mastitis of dairy cows with a combination of lincomycin and neomycin, or penicillin and dihydrostreptomycin. N. Z. Vet. J. 51, 111–116. https://doi.org/10.1080/00480169.2003.36349

McDougall, S., Agnew, K.E., Cursons, R., Hou, X.X., Compton, C.R.W., 2007. Parenteral treatment of clinical mastitis with tylosin base or penethamate hydriodide in dairy cattle. J. Dairy Sci. 90, 779–789. https://doi.org/10.3168/jds.S0022-0302(07)71562-X

McDougall, S., Clausen, L., Hintukainen, J., Hunnam, J., 2019. Randomized, controlled, superiority study of extended duration of therapy with an intramammary antibiotic for treatment of clinical mastitis. J. Dairy Sci. 102, 4376–4386. https://doi.org/10.3168/jds.2018-15141

McElreath, R., 2018. A Bayesian course with examples in R and Stan. Chapman and Hall/CRC, New York. https://doi.org/10.1201/9781315372495

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

Michelena, Á., Fontenla-Romero, Ó., Luis Calvo-Rolle, J., 2025. A review and future trends of precision livestock over dairy and beef cow cattle with artificial intelligence. Log. J. IGPL. 33, jzae111 . https://doi.org/10.1093/jigpal/jzae111

Miller, C., Portlock, T., Nyaga, D.M., O'Sullivan, J.M., 2024. A review of model evaluation metrics for machine learning in genetics and genomics. Frontiers in Bioinformatics 4, 1457619. https://doi.org/10.3389/fbinf.2024.1457619

Mor, B., Garhwal, S., Kumar, A., 2021. A Systematic Review of Hidden Markov Models and Their Applications. Archives of Computational Methods in Engineering 28, 1429–1448. https://doi.org/10.1007/s11831-020-09422-4

Morales-Ubaldo, A.L., Rivero-Perez, N., Valladares-Carranza, B., Velázquez-Ordoñez, V., Delgadillo-Ruiz, L., Zaragoza-Bastida, A., 2023. Bovine mastitis, a worldwide impact disease: Prevalence, antimicrobial resistance, and viable alternative approaches. Vet. Anim. Sci. 21, 100306. https://doi.org/10.1016/j.vas.2023.100306

Morrison, R., Hanks, J., Orpin, P.G., Strain, S., Taylor, E.N., Rose, D.C., 2024. Challenges for the management of Johne's disease in the UK: Expectation management, space, 'free riding', and vet-farmer communication. Prev. Vet. Med. 231, 106295. https://doi.org/10.1016/j.prevetmed.2024.106295

Morrison, R., Rose, D.C., 2023. Factors that influence dairy farmers' decision to implement Johne's Disease control practices: A systematic review. Prev. Vet. Med. 220, 106053. https://doi.org/10.1016/j.prevetmed.2023.106053

Naser, M.Z., Alavi, A.H., 2023. Error Metrics and Performance Fitness Indicators for Artificial Intelligence and Machine Learning in Engineering and Sciences. Architecture, Structures and Construction 3, 499–517. https://doi.org/10.1007/s44150-021-00015-8

National Research Council of the National Academies, 2015. Critical Role of Animal Science Research in Food Security and Sustainability, Critical Role of Animal Science Research in Food Security and Sustainability. National Academies Press, Washington, D.C. https://doi.org/10.17226/19000

Neethirajan, S., 2023. The Significance and Ethics of Digital Livestock Farming. AgriEngineering 5, 488–505. https://doi.org/10.3390/agriengineering5010032

Nguyen, B.T., Briggs, K.R., Nydam, D. V., 2022. Dairy production sustainability through a one-health lens. J. Am. Vet. Med. Assoc. 261, 12–16. https://doi.org/10.2460/javma.22.09.0429

Nielsen, S.S., 2009. Use of diagnostics for risk-based control of paratuberculosis in dairy herds. In Pract. 31, 150–154. https://doi.org/10.1136/inpract.31.4.150

Nielsen, S.S., 2008. Transitions in diagnostic tests used for detection of Mycobacterium avium subsp. paratuberculosis

infections in cattle. Vet. Microbiol. 132, 274–282. https://doi.org/10.1016/j.vetmic.2008.05.018

Nielsen, S.S., Enevoldsen, C., Gröhn, Y.T., 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

Nielsen, S.S., Krogh, M.A., Enevoldsen, C., 2009. Time to the occurrence of a decline in milk production in cows with various paratuberculosis antibody profiles. J. Dairy Sci. 92, 149–155. https://doi.org/10.3168/jds.2008-1488

Nielsen, S.S., Toft, N., 2011. Effect of management practices on paratuberculosis prevalence in Danish dairy herds. J. Dairy Sci. 94, 1849–1857. https://doi.org/10.3168/jds.2010-3817

Nielsen, S.S., Toft, N., 2008. Ante mortem diagnosis of paratuberculosis: A review of accuracies of ELISA, interferon-γ assay and faecal culture techniques. Vet. Microbiol. 129, 217–235. https://doi.org/10.1016/j.vetmic.2007.12.011

Nielsen, S.S., Toft, N., 2006. Age-specific characteristics of ELISA and fecal culture for purpose-specific testing for paratuberculosis. J. Dairy Sci. 89, 569–579. https://doi.org/10.3168/jds.S0022-0302(06)72120-8

Nordlund, K. V, Goodger, W.J., Pelletier, J., Collins, M.T., 1996. Associations between subclinical paratuberculosis and milk production, milk components, and somatic cell counts in dairy herds. J. Am. Vet. Med. Assoc. 208, 1872–6.

Oliveira, L., Hulland, C., Ruegg, P.L., 2013. Characterization of clinical mastitis occurring in cows on 50 large dairy herds in Wisconsin. J. Dairy Sci. 96, 7538–7549. https://doi.org/10.3168/jds.2012-6078

O'Neill, J., 2016. Tackling drug-resistant infections globally: final report and recommendations. https://amr-review.org/sites/default/files/160525_Final paper_with cover.pdf

Ozsvari, 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. Front. Vet. Sci. 7. https://doi.org/10.3389/fvets.2020.565324

Panchal, J., Patel, A., Patel, S., Goswami, D., 2024. Understanding mastitis: Microbiome, control strategies, and prevalence – A comprehensive review. Microb. Pathog. 187, 106533. https://doi.org/10.1016/j.micpath.2023.106533

Pastell, M., Frondelius, L., 2018. A hidden Markov model to estimate the time dairy cows spend in feeder based on indoor positioning data. Comput. Electron. Agric. 152, 182–185. https://doi.org/10.1016/j.compag.2018.07.005

Perry, B.D., Robinson, T.P., Grace, D.C., 2018. Review: Animal health and sustainable global livestock systems. Animal 12, 1699–1708. https://doi.org/10.1017/S1751731118000630

Peterson, C.B., Mitloehner, F.M., 2021. Sustainability of the Dairy Industry: Emissions and Mitigation Opportunities. Frontiers in

Animal Science 2, 1–11. https://doi.org/10.3389/fanim.2021.760310

Piironen, J., Vehtari, A., 2017. Comparison of Bayesian predictive methods for model selection. Stat. Comput. 27, 711–735. https://doi.org/10.1007/s11222-016-9649-y

Pinzón-Sánchez, C., Ruegg, P.L., 2011. Risk factors associated with short-term post-treatment outcomes of clinical mastitis. J. Dairy Sci. 94, 3397–3410. https://doi.org/10.3168/jds.2010-3925

Pritchard, T.C., Coffey, M.P., Bond, K.S., Hutchings, M.R., 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

Puga, J.L., Krzywinski, M., Altman, N., 2015. Points of significance: Bayesian statistics. Nat. Methods 12, 377–378. https://doi.org/10.1038/nmeth.3368

Pyörälä, S.H.K., Pyörälä, E.O., 1998. Efficacy of parenteral administration of three antimicrobial agents in treatment of clinical mastitis in lactating cows: 487 cases (1989-1995). J. Am. Vet. Med. Assoc. 212, 407–412.

R Core Team, 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77, 257–286. https://doi.org/10.1109/5.18626

Raboisson, D., Ferchiou, A., Pinior, B., Gautier, T., Sans, P., Lhermie, G., 2020. The use of meta-analysis for the measurement of animal disease burden: Losses due to clinical mastitis as an example. Front. Vet. Sci. 7. https://doi.org/10.3389/fvets.2020.00149

Rainio, O., Teuho, J., Klén, R., 2024. Evaluation metrics and statistical tests for machine learning. Sci. Rep. 14, 6086. https://doi.org/10.1038/s41598-024-56706-x

Rajala-Schultz, P., Nødtvedt, A., Halasa, T., Persson Waller, K., 2021. Prudent Use of Antibiotics in Dairy Cows: The Nordic Approach to Udder Health. Front. Vet. Sci. 8. https://doi.org/10.3389/fvets.2021.623998

Rasmussen, P., Barkema, H.W., Osei, P.P., Taylor, J., Shaw, A.P., Conrady, B., Chaters, G., Muñoz, V., Hall, D.C., Apenteng, O.O., Rushton, J., Torgerson, P.R., 2024. Global losses due to dairy cattle diseases: A comorbidity-adjusted economic analysis. J. Dairy Sci. 107, 6945–6970. https://doi.org/10.3168/jds.2023-24626

Reynolds, J., Hudson, C., 2023. Herd health: approach to assessing fertility data quality. In Pract. 45, 222–230. https://doi.org/10.1002/inpr.311

Reynolds, J.A., Bradley, A.J., Sherwin, V.E., Remnant, J.G., Hudson, C.D., 2023. Associations between Johne's disease and fertility in UK dairy herds. Veterinary Journal 298–299. https://doi.org/10.1016/j.tvjl.2023.106015

Roberson, J.R., Warnick, L.D., Moore, G., 2004. Mild to moderate clinical mastitis: Efficacy of intramammary amoxicillin, frequent milk-out, a combined intramammary amoxicillin, and frequent milk-out treatment versus no treatment. J. Dairy Sci. 87, 583–592. https://doi.org/10.3168/jds.S0022-0302(04)73200-2

Robertson, C., Sawford, K., Gunawardana, W.S.N., Nelson, T.A., Nathoo, F., Stephen, C., 2011. A hidden Markov model for analysis of frontline veterinary data for emerging zoonotic disease surveillance. PLoS One 6, e24833. https://doi.org/10.1371/journal.pone.0024833

Rossi, G., Grohn, Y.T., Schukken, Y.H., Smith, R.L., 2017. The effect of Mycobacterium avium ssp. paratuberculosis infection on clinical mastitis occurrence in dairy cows. J. Dairy Sci. 100, 7446–7454. https://doi.org/10.3168/jds.2017-12721

Roustaei, N., 2024. Application and interpretation of linear-regression analysis. Medical Hypothesis, Discovery, and Innovation in Ophthalmology 13, 151–159. https://doi.org/10.51329/mehdiophthal1506

Rubin, D. 1976. Inference and missing data. Biometrika 63, 581-592.

Ruegg, P.L., 2021. What Is Success? A Narrative Review of Research Evaluating Outcomes of Antibiotics Used for Treatment of Clinical Mastitis. Front. Vet. Sci. 8, 1–14. https://doi.org/10.3389/fvets.2021.639641

Ruegg, P.L., 2017. A 100-Year Review: Mastitis detection, management, and prevention. J. Dairy Sci. 100, 10381–10397. https://doi.org/10.3168/jds.2017-13023

Samson, O., Gaudout, N., Schmitt, E., Schukken, Y.H., Zadoks, R., 2016. Use of on-farm data to guide treatment and control mastitis caused by Streptococcus uberis. J. Dairy Sci. 99, 7690–7699. https://doi.org/10.3168/jds.2016-10964

Sargent, F.D., Lytton, V.H., Wall, O.G., 1968. Test Interval Method of Calculating Dairy Herd Improvement Association Records. J. Dairy Sci. 51, 170–179. https://doi.org/10.3168/jds.S0022-0302(68)86943-7

Saro, J., Ducháček, J., Brožová, H., Stádník, L., Bláhová, P., Horáková, T., Hlavatý, R., 2024. Discrete Homogeneous and Non-Homogeneous Markov Chains Enhance Predictive Modelling for Dairy Cow Diseases. Animals 14, 2542. https://doi.org/10.3390/ani14172542

Schmenger, A., Krömker, V., 2020. Characterization, cure rates and associated risks of clinical mastitis in Northern Germany. Vet. Sci. 7, 1–17. https://doi.org/10.3390/vetsci7040170

Schmenger, A., Leimbach, S., Wente, N., Zhang, Y., Biggs, A.M., Kroemker, V., 2020. Implementation of a targeted mastitis therapy concept using an on-farm rapid test: Antimicrobial consumption, cure rates and compliance. Veterinary Record 187, 401. https://doi.org/10.1136/vr.105674

Shook, G.E., Chaffer, M., Wu, X.L., Ezra, E., 2012. Genetic parameters for paratuberculosis infection and effect of

infection on production traits in Israeli Holsteins. Anim. Genet. 43, 56–64. https://doi.org/10.1111/j.1365-2052.2012.02349.x

Slob, N., Catal, C., Kassahun, A., 2021. Application of machine learning to improve dairy farm management: A systematic literature review. Prev. Vet. Med. 187, 105237. https://doi.org/10.1016/j.prevetmed.2020.105237

Smith, N.W., Fletcher, A.J., Hill, J.P., McNabb, W.C., 2022. Modeling the Contribution of Milk to Global Nutrition. Front. Nutr. 8, 1–7. https://doi.org/10.3389/fnut.2021.716100

Smith, R.L., Strawderman, R.L., Schukken, Y.H., Wells, S.J., Pradhan, A.K., Espejo, L.A., Whitlock, R.H., Van Kessel, J.S., Smith, J.M., Wolfgang, D.R., Gröhn, Y.T., 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

Sorge, U.S., Lissemore, K., Godkin, A., Hendrick, S., Wells, S., Kelton, D., 2011. Associations between paratuberculosis milk ELISA result, milk production, and breed in Canadian dairy cows. J. Dairy Sci. 94, 754–761. https://doi.org/10.3168/jds.2010-3404

Spies, N.C., Ng, D.P., 2025. Performance metrics for machine learning solutions in laboratory medicine. Lab. Med. 56, 597-607. https://doi.org/10.1093/labmed/lmaf013

Stabel, J.R., Bradner, L., Robbe-Austerman, S., Beitz, D.C., 2014. Clinical disease and stage of lactation influence shedding of

Mycobacterium avium subspecies paratuberculosis into milk and colostrum of naturally infected dairy cows. J. Dairy Sci. 97, 6296–6304. https://doi.org/10.3168/jds.2014-8204

Stevens, M., Piepers, S., Supré, K., Dewulf, J., de Vliegher, S., 2016. Quantification of antimicrobial consumption in adult cattle on dairy herds in Flanders, Belgium, and associations with udder health, milk quality, and production performance. J. Dairy Sci. 99, 2118–2130. https://doi.org/10.3168/jds.2015-10199

Stone, M., 1977. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion, Source: Journal of the Royal Statistical Society. Series B (Methodological) 39, 44-47. https://www.jstor.org/stable/2984877

Su, T.-L., Jaki, T., Hickey, G.L., Buchan, I., Sperrin, M., 2018. A review of statistical updating methods for clinical prediction models. Stat. Methods Med. Res. 27, 185–197. https://doi.org/10.1177/0962280215626466

Sumi, K., Maw, S.Z., Zin, T.T., Tin, P., Kobayashi, I., Horii, Y., 2021. Activity-integrated hidden markov model to predict calving time. Animals 11, 1–12. https://doi.org/10.3390/ani11020385

Svennesen, L., Skarbye, A.P., Farre, M., Astrup, L.B., Halasa, T., Krömker, V., Denwood, M., Kirkeby, C., 2023. Treatment of mild to moderate clinical bovine mastitis caused by gram-positive bacteria: A noninferiority randomized trial of local penicillin treatment alone or combined with systemic treatment. J. Dairy Sci. 106, 5696–5714. https://doi.org/10.3168/jds.2022-22993

Sweeney, R.W., Whitlock, R.H., McAdams, S., Fyock, T., 2006. Longitudinal study of ELISA seroreactivity to Mycobacterium avium subsp. paratuberculosis in infected cattle and culture-negative herd mates. Journal of Veterinary Diagnostic Investigation 18, 2–6. https://doi.org/10.1177/104063870601800102

Swinkels, J.M., Cox, P., Schukken, Y.H., Lam, T.J.G.M., 2013. Efficacy of extended cefquinome treatment of clinical Staphylococcus aureus mastitis. J. Dairy Sci. 96, 4983–4992. https://doi.org/10.3168/jds.2012-6197

Taylor, E.N., Channa, K., Orpin, P.G., Hanks, J., Taylor, N.M., 2024. Changes in the Johne's disease situation in GB dairy herds over 10 years, as revealed by regular milk ELISA data. Prev. Vet. Med. 231, 106295. https://doi.org/10.1016/j.prevetmed.2024.106317

Terven, J., Cordova-Esparza, D.M., Romero-González, J.A., Ramírez-Pedraza, A., Chávez-Urbiola, E.A., 2025. A comprehensive survey of loss functions and metrics in deep learning. Artif. Intell. Rev. 58, 195. https://doi.org/10.1007/s10462-025-11198-7

Tiwari, A., VanLeeuwen, J.A., Dohoo, I.R., Keefe, G.P., Haddad, J.P., Tremblay, R., Scott, H.M., Whiting, T., 2007. Production effects of pathogens causing bovine leukosis, bovine viral diarrhea, paratuberculosis, and neosporosis. J. Dairy Sci. 90, 659–669. https://doi.org/10.3168/jds.S0022-0302(07)71548-5

Tiwari, A., VanLeeuwen, J.A., McKenna, S.L.B., Keefe, G.P., Barkema, H.W., 2006. Johne's disease in Canada Part I: Clinical symptoms, pathophysiology, diagnosis, and prevalence in dairy herds. Canadian Veterinary Journal 47, 874–882.

Tohka, J., van Gils, M., 2021. Evaluation of machine learning algorithms for health and wellness applications: A tutorial. Comput. Biol. Med. 132, 104324. https://doi.org/10.1016/j.compbiomed.2021.104324

Tredennick, A.T., Hooker, G., Ellner, S.P., Adler, P.B., 2021. A practical guide to selecting models for exploration, inference, and prediction in ecology. Ecology 102. https://doi.org/10.1002/ecy.3336

Tuerlinckx, F., Rijmen, F., Verbeke, G., De Boeck, P., 2006. Statistical inference in generalized linear mixed models: A review. British Journal of Mathematical and Statistical Psychology 59, 225-255. https://doi.org/10.1348/000711005X79857

Tullo, E., Finzi, A., Guarino, M., 2019. Review: Environmental impact of livestock farming and Precision Livestock Farming as a mitigation strategy. Science of the Total Environment 650, 2751-7760. https://doi.org/10.1016/j.scitotenv.2018.10.018

Van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J.B., Neyer, F.J., van Aken, M.A.G., 2014. A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research. Child Dev. 85, 842–860. https://doi.org/10.1111/cdev.12169

van der Ploeg, T., Austin, P.C., Steyerberg, E.W., 2014. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med. Res. Methodol. 14, 137. https://doi.org/10.1186/1471-2288-14-137

van Dijk, M., Morley, T., Rau, M.L., Saghai, Y., 2021. A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050. Nat. Food 2, 494–501. https://doi.org/10.1038/s43016-021-00322-9

Van Leeuwen, J.A., Keefe, G.P., Tiwari, A., 2002. Seroprevalence and productivity effects of infection with bovine leukemia virus, Mycobacterium avium subspecies paratuberculosis, and Neospora caninum in Maritime Canadian dairy cattle. Bov. Pract. 86–91. https://doi.org/10.21423/bovine-vol36no2p86-91

van Ravenzwaaij, D., Cassey, P., Brown, S.D., 2018. A simple introduction to Markov Chain Monte–Carlo sampling. Psychon. Bull. Rev. 25, 143–154. https://doi.org/10.3758/s13423-016-1015-8

van Roon, A.M., Madouasse, A., Toft, N., Santman-Berends, I.M.G.A., Gethmann, J., Eze, J., Humphry, R.W., Graham, D., Guelbenzu-Gonzalo, M., Nielen, M., More, S.J., Mercat, M., Fourichon, C., Sauter-Louis, C., Frössling, J., Ågren, E., Gunn, G.J., Henry, M.K., van Schaik, G., 2022. Output-based assessment of herd-level freedom from infection in endemic situations: Application of a Bayesian Hidden Markov model. Prev. Vet. Med. 204, 105662. https://doi.org/10.1016/j.prevetmed.2022.105662

van Schaik, G., Madouasse, A., van Roon, A.M., More, S. J., Graham, D.A., Frossling, J., Gethmann, J., Fourichon, C., Mercat, M., Agren, E., Sauter_louis, C., Gunn, G., Eze, J., Humphry, R., Henry, M.K., Guelbenzu, M., Nielen, M., Santman-Berends, I.M.G.A., 2023. Rev. Sci. Tech. 42, 210-217. https://doi.org/10.20506/RST.42.3364.

van Walraven, C., Austin, P.C., Jennings, A., Forster, A.J., 2009. Correlation between serial tests made disease probability estimates erroneous. J. Clin. Epidemiol. 62, 1301–1305. https://doi.org/10.1016/j.jclinepi.2009.04.010

VanLeeuwen, J.A., Haddad, J.P., Dohoo, I.R., Keefe, G.P., Tiwari, A., Tremblay, R., 2010. Associations between reproductive performance and seropositivity for bovine leukemia virus, bovine viral-diarrhea virus, Mycobacterium avium subspecies paratuberculosis, and Neospora caninum in Canadian dairy cows. Prev. Vet. Med. 94, 54–64. https://doi.org/10.1016/j.prevetmed.2009.11.012

Vasquez, A.K., Nydam, D. V., Capel, M.B., Ceglowski, B., Rauch, B.J., Thomas, M.J., Tikofsky, L., Watters, R.D., Zuidhof, S., Zurakowski, M.J., 2016. Randomized noninferiority trial comparing 2 commercial intramammary antibiotics for the treatment of nonsevere clinical mastitis in dairy cows. J. Dairy Sci. 99, 8267–8281. https://doi.org/10.3168/jds.2016-11258

Vasquez, A.K., Nydam, D. V., Capel, M.B., Eicker, S., Virkler, P.D., 2017. Clinical outcome comparison of immediate blanket treatment versus a delayed pathogen-based treatment

protocol for clinical mastitis in a New York dairy herd. J. Dairy Sci. 100, 2992–3003. https://doi.org/10.3168/jds.2016-11614

Vehtari, A., Gelman, A., Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat. Comput. 27, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

von Soosten, D., Meyer, U., Flachowsky, G., Dänicke, S., 2020. Dairy Cow Health and Greenhouse Gas Emission Intensity. Dairy 1, 20–29. https://doi.org/10.3390/dairy1010003

Waddell, L.A., Rajic, A., Stärk, K.D.C., McEwen, S.A., 2015. The zoonotic potential of Mycobacterium avium ssp. Paratuberculosis: A systematic review and meta-analyses of the evidence. Epidemiol. Infect. 143, 3135-3157. https://doi.org/10.1017/S095026881500076X

Wan, B., Caffo, B., Vedula, S.S., 2022. A Unified Framework on Generalizability of Clinical Prediction Models. Front. Artif. Intell. 5. https://doi.org/10.3389/frai.2022.872720

Wang, C., Turnbull, B.W., Nielsen, S.S., Gröhn, Y.T., 2011. Bayesian analysis of longitudinal Johne's disease diagnostic data without a gold standard test. J. Dairy Sci. 94, 2320–2328. https://doi.org/10.3168/jds.2010-3675

Wang, Q., Ma, Y., Zhao, K., Tian, Y., 2022. A Comprehensive Survey of Loss Functions in Machine Learning. Annals of Data Science 9, 187–212. https://doi.org/10.1007/s40745-020-00253-5

Watanabe, S., 2010. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. Journal of Machine Learning Research 11, 3571–3594. https://doi.org/https://doi.org/10.5555/1756006.1953045

Wathes, C.M., Kristensen, H.H., Aerts, J.M., Berckmans, D., 2008. Is precision livestock farming an engineer's daydream or nightmare, an animal's friend or foe, and a farmer's panacea or pitfall? Comput. Electron. Agric. 64, 2–10. https://doi.org/10.1016/j.compag.2008.05.005

Weber, M.F., Kelton, D., Eisenberg, S.W.F., Donat, K., 2024. Progress in Paratuberculosis Control Programmes for Dairy Herds. Animals 14, 1127. https://doi.org/10.3390/ani14071127

Wenz, J.R., Garry, F.B., Lombard, J.E., Elia, R., Prentice, D., Dinsmore, R.P., 2005. Short communication: Efficacy of parenteral ceftiofur for treatment of systemically mild clinical mastitis in dairy cattle. J. Dairy Sci. 88, 3496–3499. https://doi.org/10.3168/jds.S0022-0302(05)73034-4

Whittington, R., Donat, K., Weber, M.F., Kelton, D., Nielsen, S.S., Eisenberg, S., Arrigoni, N., Juste, R., Sáez, J.L., Dhand, N., Santi, A., Michel, A., Barkema, H., Kralik, P., Kostoulas, P., Citer, L., Griffin, F., Barwell, R., Moreira, M.A.S., Slana, I., Koehler, H., Singh, S.V., Yoo, H.S., Chávez-Gris, G., Goodridge, A., Ocepek, M., Garrido, J., Stevenson, K., Collins, M., Alonso, B., Cirone, K., Paolicchi, F., Gavey, L., Rahman, M.T., 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, A.N., Piaggio, J., Suanes, A., De Waard, J.H., 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

Whittington, R.J., Windsor, P.A., 2009. In utero infection of cattle with Mycobacterium avium subsp. paratuberculosis: A critical review and meta-analysis. Veterinary Journal 179, 60–69. https://doi.org/10.1016/j.tvjl.2007.08.023

Wilimitis, D., Walsh, C.G., 2023. Practical Considerations and Applied Examples of Cross-Validation for Model Development and Evaluation in Health Care: Tutorial. JMIR AI 2. https://doi.org/10.2196/49023

Wilm, J., Krömker, V., Kirkeby, C., Gussmann, M., 2023. Lactational treatment of bovine mastitis—Development over time and factors influencing cytological cure. J. Dairy Sci. 106, 5740–5752. https://doi.org/10.3168/jds.2022-22841

Windsor, P.A., Whittington, R.J., 2010. Evidence for age susceptibility of cattle to Johne's disease. Veterinary Journal 184, 37–44. https://doi.org/10.1016/j.tvjl.2009.01.007

WOAH, 2008. 76th General Session of the OIE, Bulletin No. 4. https://www.woah.org/fileadmin/Home/eng/Publications_&_Documentation/docs/pdf/bulletin/Bull_2008-4-ENG.pdf

Yang, J., Soltan, A.A.S., Clifton, D.A., 2022. Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. NPJ Digit. Med. 5, 69. https://doi.org/10.1038/s41746-022-00614-9

Yates, L.A., Aandahl, Z., Richards, S.A., Brook, B.W., 2023. Cross validation for model selection: A review with examples from ecology. Ecol. Monogr. 93. https://doi.org/10.1002/ecm.1557

Younas, A., 2024. Beyond 'statistical significance': A nontechnical primer of Bayesian statistics and Bayes factors for health researchers. J. Eval. Clin. Pract. 30, 1218–1226. https://doi.org/10.1111/jep.14032

Zhang, J., Yang, Y., Ding, J., 2023. Information criteria for model selection. Wiley Interdiscip. Rev. Comput. Stat. 15. https://doi.org/10.1002/wics.1607

Appendices

Appendix A: Chapter 4 supplementary materials

Model		2-State Model		3-State Model			4-State Model			
Transition	**From state**	**To state**								
		1	**2**	**1**	**2**	**3**	**1**	**2**	**3**	**4**
	1	0.0	0.1	0.0	0.1	0.0	0.0	0.1	0.0	0.0
	2	0.1	0.0	0.1	0.0	0.1	0.1	0.0	0.1	0.0
	3	-	-	0.0	0.1	0.0	0.0	0.1	0.0	0.1
	4	-	-	-	-	-	0.0	0.0	0.1	0.0
Emission	**Parameter**	**State**								
		1	**2**	**1**	**2**	**3**	**1**	**2**	**3**	**4**
	Natural log mean (95% C.I.)	1.3	3	1.3	2.1	3.9	0.9	1.3	2.1	3.9
	Natural log standard deviation (95% C.I.)	0.5	0.7	0.5	0.6	0.8	0.5	0.5	0.6	0.8

Table A.1: Initial estimates when fitting models for the transition intensity matrix and for the emission probability function parameters for each hidden state.

Variable	Minimum[1]	Maximum[1]	No. rows below minimum	No. rows above maximum
Age of primiparous animals (months)[2]	20	60	63 (0.02%)	929 (0.37%)
Age of multiparous animals (months)[2]	25	-	83 (0.03%)	-
DIM (days)	0	730	0 (0%)	824 (0.33%)
Protein (%)	1	6	20270 (8.02%)	14344 (5.67%)
Butterfat (%)	1.5	9	21764 (8.61%)	14921 (5.90%)

Table A.2: The number (%) of rows in the raw dataset whose value for a variable was outside of a range considered to be plausible.
The percentage is calculated as a percentage of the entire dataset (252,789 rows). Note that some rows had values which were implausible for more than one variable.
[1]The minimum and maximum plausible values for each variable.
[2]The minimum and maximum plausible ages were different for primi- and multiparous cows. There was no maximum plausible age for multiparous cows.

Feature	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum
Date	07/12/2010	25/02/2015	24/05/2017	12/03/2017	31/07/2019	22/12/2021
S/P Ratio (%)	0.0	1.7	2.8	6.8	5.2	282.9
Parity	1.0	1.0	2.0	2.6	3.0	16.0
Age (months)	0.0	35.6	47.5	53.6	65.5	895.2
Days in milk (days)	0.0	78.0	163.0	177.2	252.0	1996.0
Milk yield (litres)	0.0	20.0	27.5	27.5	35.2	99.1
Somatic Cell Count (x 1000 cells/ml)	0.0	23.0	51.0	188.3	136.0	18494.0
Protein (%)	0.0	3.1	3.4	3.4	3.7	18.4
Butterfat (%)	0.0	3.5	4.2	4.2	4.9	21.8
Mean titre negative cows (% S/P)	0.5	2.6	3.1	3.1	3.7	6.3

Table A.3: Descriptive statistics of the features in the raw data (prior to cleaning to remove implausible values).
Data were taken from cows (n = 39,419) in 47 dairy herds in the United Kingdom which were engaged in regular milk recording and paratuberculosis testing.

Feature	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum
Date	07/12/2010	05/03/2015	03/07/2017	08/04/2017	27/08/2019	22/12/2021
S/P Ratio (%)	0.0	1.8	2.9	6.7	5.2	282.9
Parity	1.0	1.0	2.0	2.6	3.0	16.0
Age (months)	20.2	35.8	47.5	53.7	65.6	228.6
Days in milk (days)	0.0	78.0	163.0	175.3	252.0	730.0
Milk yield (litres)	0.1	21.2	28.0	28.8	35.6	99.1
Somatic Cell Count (x 1000 cells/ml)	1.0	23.0	52.0	185.7	137.0	18494.0
Protein (%)	1.2	3.2	3.4	3.4	3.7	6.0
Butterfat (%)	1.5	3.6	4.2	4.3	4.9	9.0
Mean titre negative cows (% S/P)	0.5	2.6	3.1	3.2	3.7	6.3

Table A.4: Descriptive statistics of the features in the data used for training linear models (after removal of rows with implausible feature values).
The data included cows (n = 37,792) in 47 dairy herds in the United Kingdom which were engaged in regular milk recording and paratuberculosis testing.

	Primiparous Model (Full)			Primiparous Model (Reduced)			Multiparous Model (Full)			Multiparous Model (Reduced)		
Variable	Estimate	95% Confidence Interval	Standard Error	Estimate	95% Confidence Interval	Standard Error	Estimate	95% Confidence Interval	Standard Error	Estimate	95% Confidence Interval	Standard Error
Intercept	-1.317	-1.68 - -0.954	0.185	1.939	1.716 - 2.162	0.114	0.195	-0.036 - 0.425	0.118	1.297	1.241 - 1.353	0.029
B-Spline (DIM) 1	2.349	1.829 - 2.869	0.265	-2.492	-2.884 - -2.1	0.200	-0.792	-1.15 - -0.434	0.183	-1.111	-1.179 - -1.043	0.035
B-Spline (DIM) 2	0.635	0.267 - 1.002	0.187	-1.549	-1.801 - -1.297	0.128	-0.530	-0.74 - -0.319	0.107	-0.997	-1.036 - -0.958	0.020
B-Spline (DIM) 3	1.296	0.891 - 1.701	0.207	-2.120	-2.416 - -1.824	0.151	-0.781	-1.043 - -0.518	0.134	-1.165	-1.215 - -1.115	0.025
B-Spline (DIM) 4	1.085	0.748 - 1.422	0.172	-1.830	-2.085 - -1.575	0.130	-0.753	-0.982 - -0.523	0.117	-1.118	-1.161 - -1.075	0.022
B-Spline (DIM) 5	1.380	1.065 - 1.695	0.161	-1.983	-2.234 - -1.732	0.128	-0.792	-1.039 - -0.545	0.126	-1.131	-1.177 - -1.085	0.024
B-Spline (DIM) 6	0.933	0.432 - 1.433	0.256	-2.151	-2.546 - -1.756	0.201	-0.873	-1.111 - -0.634	0.122	-1.095	-1.139 - -1.051	0.023
B-Spline (DIM) 7	0.933	0.03 - 1.836	0.461	-1.493	-2.248 - -0.738	0.385	-0.911	-1.15 - -0.672	0.122	-1.032	-1.076 - -0.988	0.022
B-Spline (DIM) 8	1.112	0.228 - 1.996	0.451	-2.034	-3.087 - -0.981	0.537	-0.881	-1.114 - -0.649	0.119	-0.886	-0.929 - -0.844	0.022
B-Spline (DIM) 9							-0.404	-0.759 - -0.049	0.181	-0.451	-0.512 - -0.39	0.031
B-Spline (DIM) 10							-0.869	-1.5 - -0.237	0.322	-0.622	-0.726 - -0.519	0.053
B-Spline (DIM) 11							-0.706	-1.36 - -0.053	0.333	-0.664	-0.775 - -0.552	0.057
Protein (%)	0.519	0.424 - 0.614	0.049				0.192	0.18 - 0.203	0.006			
Yield (litres)	-0.001	-0.003 - 0.001	0.001				-0.015	-0.019 - -0.01	0.002			
Log (SCC (x 1000 cells/ml))	-0.035	-0.079 - 0.01	0.023				0.186	0.155 - 0.217	0.016			
Age (months)	0.010	0.009 - 0.012	0.001	-0.024	-0.032 - -0.017	0.004	0.005	0.005 - 0.005	0.000	0.006	0.005 - 0.006	0.000
Mean titre negative cows (% S/P)	0.301	0.294 - 0.307	0.003	0.308	0.301 - 0.315	0.003	0.289	0.284 - 0.293	0.002	0.298	0.294 - 0.303	0.002
B-Spline (DIM) 1 x Protein	-0.810	-0.962 - -0.657	0.078									
B-Spline (DIM) 2 x Protein	-0.335	-0.442 - -0.228	0.055									
B-Spline (DIM) 3 x Protein	-0.564	-0.679 - -0.449	0.059									
B-Spline (DIM) 4 x Protein	-0.511	-0.605 - -0.417	0.048									
B-Spline (DIM) 5 x Protein	-0.596	-0.682 - -0.509	0.044									
B-Spline (DIM) 6 x Protein	-0.427	-0.563 - -0.29	0.070									
B-Spline (DIM) 7 x Protein	-0.426	-0.667 - -0.186	0.123									
B-Spline (DIM) 8 x Protein	-0.472	-0.704 - -0.241	0.118									
Yield (litres) x Log (SCC (x 1000 cells/ml))	-0.002	-0.002 - -0.001	0.000				-0.001	-0.002 - -0.001	0.000			

Table A.5: Regression coefficients of the fixed effects of the four final multivariable models describing the association between covariates and log S/P ratios (MAP antibody ELISA results) in dairy cows in 47 UK dairy herds.

DIM = days in milk. SCC = somatic cell count.

Table continues...

Variable	Primiparous Model (Full)			Primiparous Model (Reduced)			Multiparous Model (Full)			Multiparous Model (Reduced)		
	Estimate	95% Confidence Interval	Standard Error	Estimate	95% Confidence Interval	Standard Error	Estimate	95% Confidence Interval	Standard Error	Estimate	95% Confidence Interval	Standard Error
B-Spline (DIM) 1 x Age (months)				0.056	0.042 - 0.07	0.007						
B-Spline (DIM) 2 x Age (months)				0.024	0.015 - 0.033	0.005						
B-Spline (DIM) 3 x Age (months)				0.039	0.028 - 0.049	0.005						
B-Spline (DIM) 4 x Age (months)				0.030	0.021 - 0.039	0.004						
B-Spline (DIM) 5 x Age (months)				0.036	0.027 - 0.044	0.004						
B-Spline (DIM) 6 x Age (months)				0.046	0.035 - 0.058	0.006						
B-Spline (DIM) 7 x Age (months)				0.031	0.013 - 0.049	0.009						
B-Spline (DIM) 8 x Age (months)				0.044	0.022 - 0.066	0.011						
B-Spline (DIM) 1 x Yield (litres)							0.022	0.015 - 0.029	0.003			
B-Spline (DIM) 2 x Yield (litres)							0.010	0.006 - 0.014	0.002			
B-Spline (DIM) 3 x Yield (litres)							0.014	0.009 - 0.019	0.003			
B-Spline (DIM) 4 x Yield (litres)							0.012	0.008 - 0.017	0.002			
B-Spline (DIM) 5 x Yield (litres)							0.012	0.008 - 0.017	0.002			
B-Spline (DIM) 6 x Yield (litres)							0.014	0.009 - 0.019	0.002			
B-Spline (DIM) 7 x Yield (litres)							0.015	0.01 - 0.019	0.002			
B-Spline (DIM) 8 x Yield (litres)							0.012	0.008 - 0.017	0.002			
B-Spline (DIM) 9 x Yield (litres)							0.002	-0.005 - 0.009	0.004			
B-Spline (DIM) 10 x Yield (litres)							-0.003	-0.016 - 0.009	0.006			
B-Spline (DIM) 11 x Yield (litres)							0.013	0.001 - 0.025	0.006			
B-Spline (DIM) 1 x Log (SCC (x 1000 cells/ml))							-0.132	-0.181 - -0.082	0.025			
B-Spline (DIM) 2 x Log (SCC (x 1000 cells/ml))							-0.077	-0.106 - -0.047	0.015			
B-Spline (DIM) 3 x Log (SCC (x 1000 cells/ml))							-0.082	-0.119 - -0.045	0.019			
B-Spline (DIM) 4 x Log (SCC (x 1000 cells/ml))							-0.084	-0.116 - -0.052	0.016			
B-Spline (DIM) 5 x Log (SCC (x 1000 cells/ml))							-0.083	-0.118 - -0.048	0.018			
B-Spline (DIM) 6 x Log (SCC (x 1000 cells/ml))							-0.080	-0.114 - -0.047	0.017			
B-Spline (DIM) 7 x Log (SCC (x 1000 cells/ml))							-0.070	-0.104 - -0.036	0.017			
B-Spline (DIM) 8 x Log (SCC (x 1000 cells/ml))							-0.047	-0.08 - -0.014	0.017			
B-Spline (DIM) 9 x Log (SCC (x 1000 cells/ml))							-0.060	-0.113 - -0.007	0.027			
B-Spline (DIM) 10 x Log (SCC (x 1000 cells/ml))							0.018	-0.081 - 0.116	0.050			
B-Spline (DIM) 11 x Log (SCC (x 1000 cells/ml))							-0.065	-0.171 - 0.041	0.054			

Table A.5 (continued): Regression coefficients of the fixed effects of the four final multivariable models describing the association between covariates and log S/P ratios (MAP antibody ELISA results) in dairy cows in 47 UK dairy herds.
DIM = days in milk. SCC = somatic cell count.

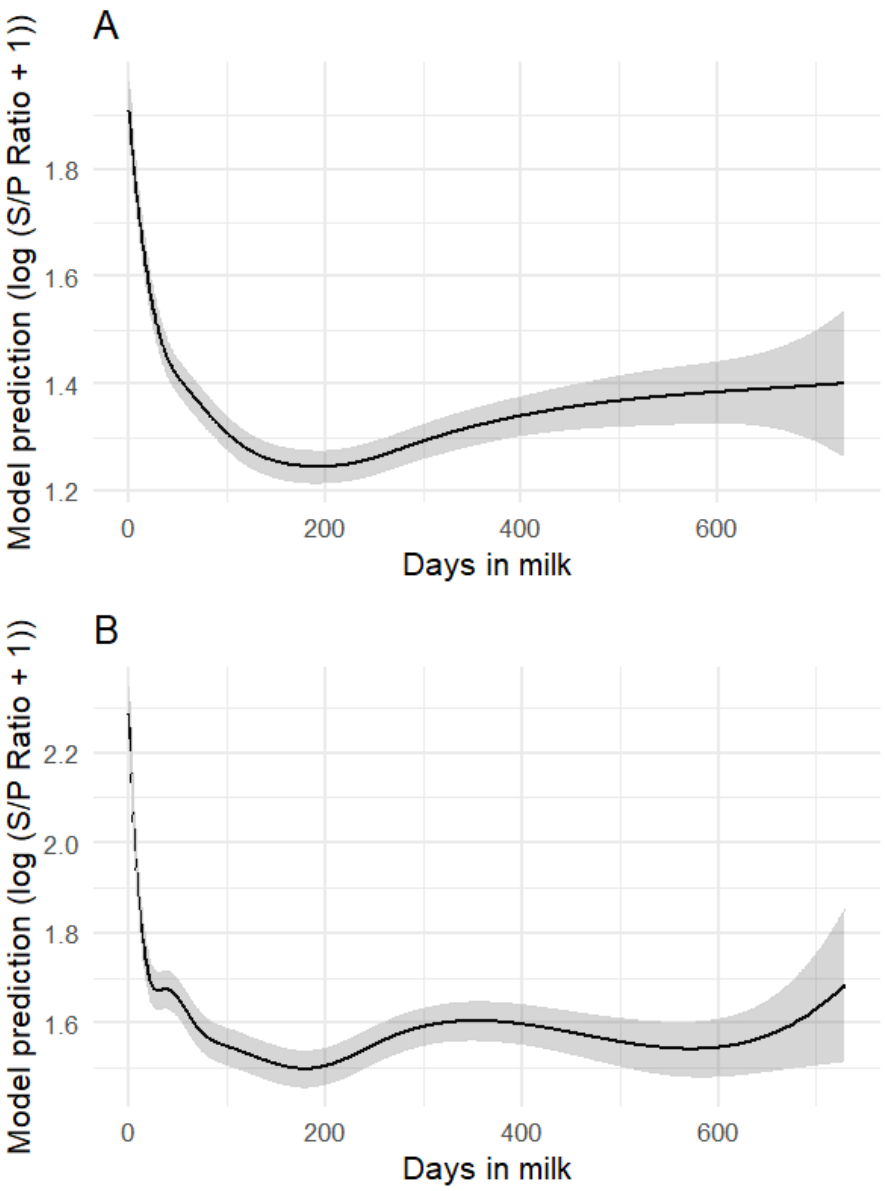

Figure A.1: Days in milk (primiparous model); B) Days in milk (multiparous model). *95% confidence intervals are shown as shaded areas.*

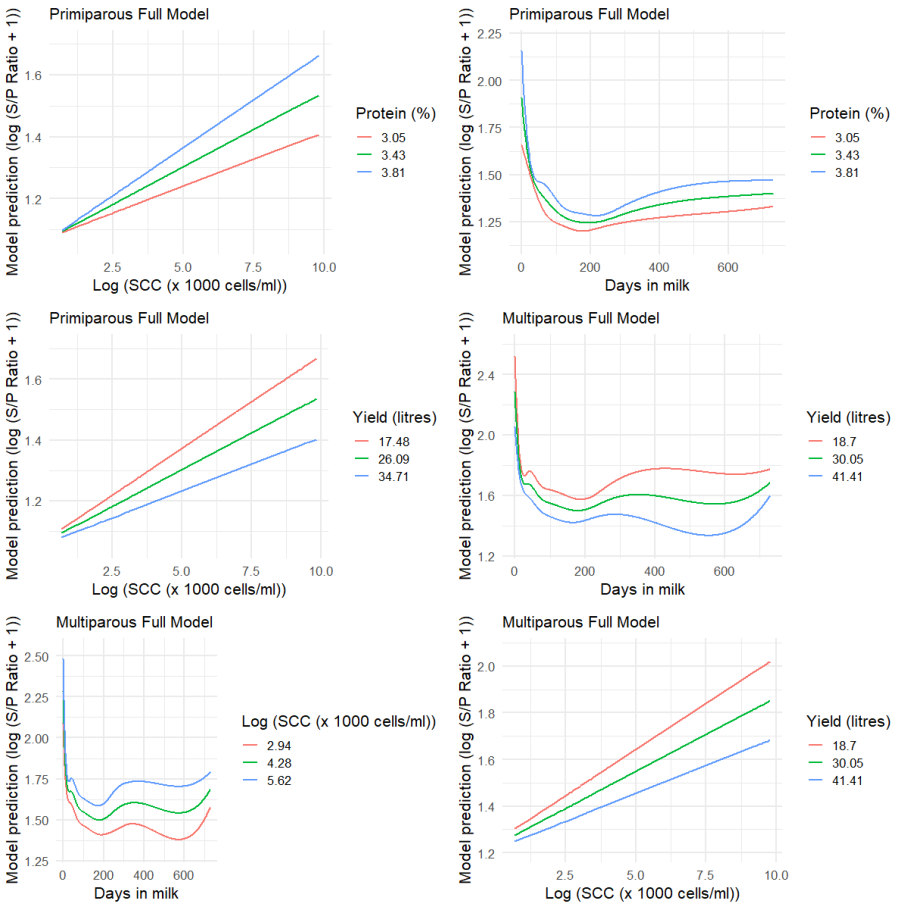

Figure A.2:Interaction terms in the full linear models used for adjusting S/P ratios according to influential non-disease factors.

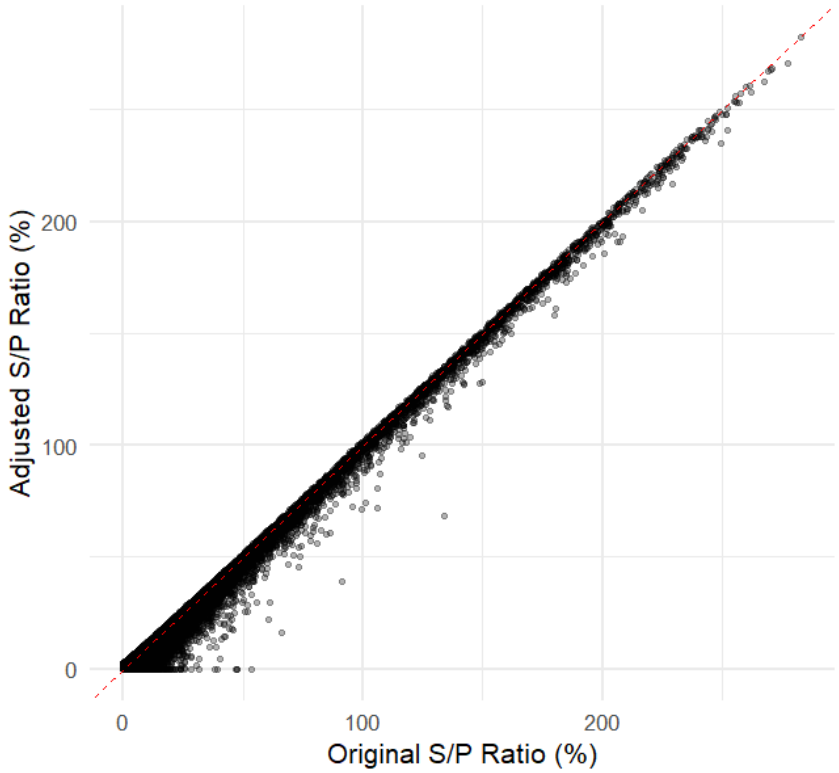

Figure A.3: Scatterplot of S/P ratio (%) values in the data used to train hidden Markov models, before and after adjusting for influential covariates.

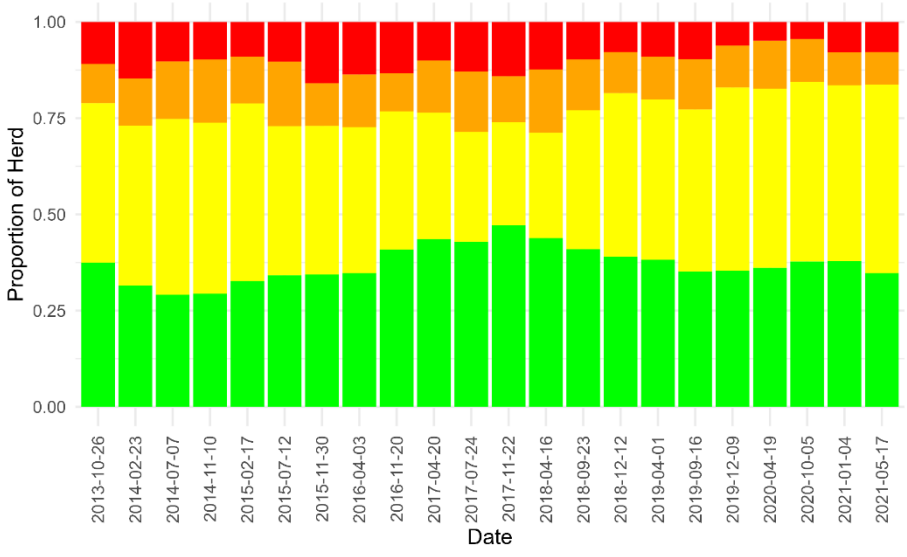

Figure A.4: The prevalence over time of cows in different states, according to the hidden state path calculated using the Viterbi algorithm, in one example herd.

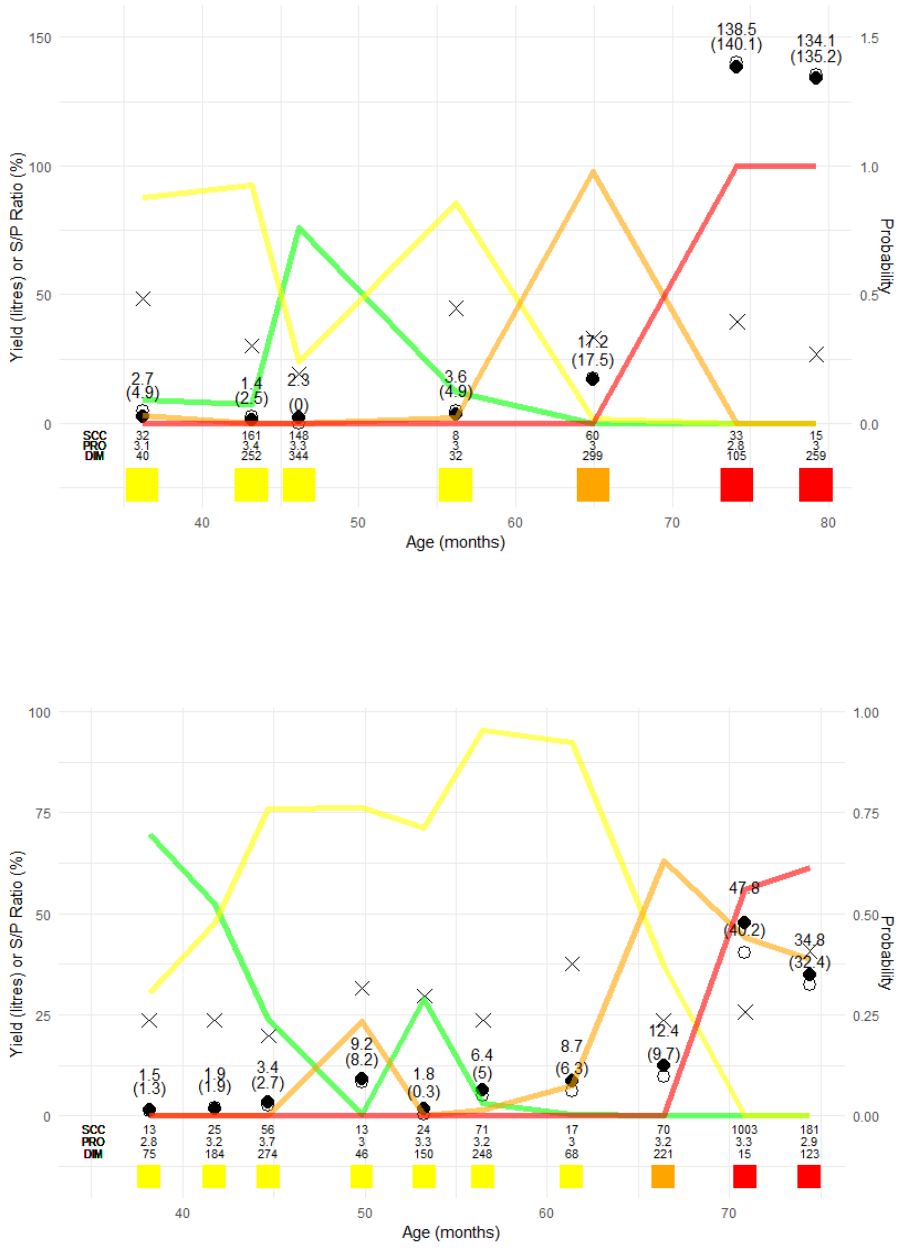

Figure A.5: Predictions from a 4-state hidden Markov model for interpretation of serial milk paratuberculosis antibody ELISA test results on two example cows.
The original S/P ratios (%) are represented by solid black circles. The adjusted S/P ratios (%) are represented by hollow circles. Numeric values of original and, in brackets, adjusted S/P ratios are shown above these circles. Milk yield is represented by crosses. Coloured lines represent the filter probability (calculated using the forward algorithm) over time for each hidden state. The coloured squares indicate the most likely hidden state path (calculated using the Viterbi algorithm). Somatic cell count (SCC), % protein (PRO) and days-in-milk (DIM) are shown as text. Green = state 1, yellow = state 2, orange = state 3, Red = state 4.