Understanding how new evidence influences practitioners' beliefs regarding dry cow therapy: a Bayesian approach using probabilistic elicitation

H.M. Higgins [a*], J. Mouncey [b], I. Nanjiani [b] and A.J.C Cook [c]

[a] Institute of Veterinary Science, University of Liverpool, Leahurst Campus, Chester High Road, Neston, Wirral. UK. CH64 7TE. h.higgins@liverpool.ac.uk

[b] Westpoint Veterinary Group, Dawes Farm, Bognor Road, Warnham, West Sussex, RH12 3SH. jon.mouncey@westpointfarmvets.co.uk and ian.nanjiani@westpointfarmvets.co.uk

[c] Faculty of Health and Medical Sciences, School of Veterinary Medicine, Daphne Jackson Road, Guildford, Surrey, GU2 7AL. alasdair.j.cook@surrey.ac.uk

*Corresponding author telephone.: +44 (0) 151 795 6368, or +44 (0) 7833597029

E-mail: h.higgins@liverpool.ac.uk

Abstract

This study used probabilistic elicitation and a Bayesian framework to quantitatively explore how logically practitioners' update their clinical beliefs after exposure to new data. The clinical context was the efficacy of antibiotics versus teat sealants for preventing mammary infections during the dry period. While most practitioners updated their clinical expectations logically, the majority failed to draw sufficient strength from the new data so that their clinical confidence afterwards was lower than merited. This study provides quantitative insight into how practitioners' update their beliefs. We discuss some of the psychological issues that may be faced by practitioners when interpreting new data. The results have important implications for evidence-based practice and clinical research in terms of the impact that new data may bring to the clinical community.

Keywords:

Antimicrobial resistance

Bayesian analysis

Belief updating

Clinical priors

Evidence-based veterinary medicine

Probabilistic elicitation

1. Introduction

A Bayesian statistical framework is ideally suited to and increasingly being used in evidence-based medicine (Ashby, 2006). This approach is conceptually straightforward. There are always two sources of information, the new data arising from a recent experiment and the prior information (Spiegelhalter et al., 2000). Prior information is any pre-existing information of relevance to the parameter of interest that has not arisen from the new experiment. The prior information must be expressed in a quantitative format as a probability distribution, called 'the prior' (Garthwaite et al., 2005). The information originating from the new data is summarised by a likelihood function. To conduct the analysis, Bayes theorem is used to combine the prior with the likelihood function and produce a posterior probability distribution (Bayes, 1763). Bayes theorem expresses how the prior information should, logically, be updated in light of the new evidence. Hence the posterior distribution encapsulates everything that is now known about the parameter, having updated the prior with the new data (Spiegelhalter et al., 2004). If the prior information is weak (i.e. contains considerable uncertainty), and the new data comparatively strong, the posterior will be dominated by the new data, and vice versa.

By always including prior information formally in the statistical analysis itself, Bayesian statistics quantitatively places the new data *in the context of* pre-existing knowledge and addresses the question: how should the data change what we currently believe? (Spiegelhalter et al., 2004). It is, therefore, a formalisation of 'learning from experience' and hence evidence-based practice (Ashby, 2006). In contrast, the traditional (frequentist) statistical framework does not include prior information in the analysis itself and hence the reader is left to quantify for themselves how the new data should be combined with prior knowledge to arrive at a final answer. There are several possible choices for the prior information, including data from previously conducted experiments and 'off-the-shelf' theoretical

distributions aimed at representing different prior perspectives, for example, a ‘reasonable cautious sceptic’ (Spiegelhalter et al., 1994). Another possibility is to base the prior information on pre-existing clinical knowledge, which is then referred to as a ‘clinical prior’ (Chaloner and Rhame, 2001). In this case, the practitioner’s current belief needs to be captured in a numerical format as a probability distribution (Johnson et al., 2010a). The technique used to do this is called probabilistic elicitation (O'Hagan et al., 2006).

Currently, UK National Health Service data monitoring committees may use a Bayesian analysis to aid their decisions over when to terminate a trial (Spiegelhalter et al., 2004). Clinical priors obtained by eliciting doctors’ beliefs are combined with the accruing trial data. When the posterior distribution associated with the most sceptical clinical prior supports the new treatment, the trial may be trial stopped on the grounds that the new data will be sufficiently strong to convince the medical community (Fayers et al., 1997). The assumption underpinning this decision is that doctors will actually update their beliefs in keeping with Bayes theorem. There is literature on the intuitiveness of Bayesian logic in several non-clinical contexts (O'Hagan et al., 2006; Kynn, 2008). Experimental psychologists in the 1970’s questioned clinicians’ abilities to reason logically due to heuristics. These are quick mental strategies that people may employ instinctively to make judgements when faced with uncertainty. They can be effective but may lead to severe bias and error (Tversky and Kahneman, 1974; Cooke, 1991). For example, people often make estimates by starting from an initial value and amend this to arrive at a final answer. Even if the initial value is known to be arbitrary, people will typically give answers that are biased towards the initial value, the so-called anchoring phenomenon (Tversky and Kahneman, 1974). There is also psychological and behavioural literature indicating that people may react in a negative way when their beliefs are challenged, be that emotionally, cognitively or behaviourally (Brehm, 1966; Politi et al., 2007). Such negative reactions may contravene Bayesian logic.

In contrast, there is also recent work suggesting that people's judgements are very close to Bayesian estimates for certain tasks (Baker et al., 2006; Griffiths and Tenenbaum, 2006; Westover et al., 2011) and especially when information is represented in a way to facilitate Bayesian reasoning (Hoffrage et al., 2000; Gigerenzer, 2011). However, the existing literature has predominately involved undergraduate students and lay tasks (Phillips and Edwards, 1966; Griffiths and Tenenbaum, 2006). Given the increased emphasis on evidence based medicine and the increased use of Bayesian methods in clinical care, it is of real practical interest to understand how practitioners' update their beliefs compared to Bayes Theorem for clinical parameters such as incidence rates using the type of information that is published in medical journals. There is, however, a paucity of specific literature. To start to address this research gap, we used the clinical context of dry cow therapy to illustrate a simple practical method that may be used to quantitatively investigate how logically practitioners update their clinical beliefs compared to Bayes theorem for a continuous clinical parameter.

In the UK, blanket antibiotic dry cow therapy (BDCT) is a commonly used strategy to aid mastitis control. It involves the infusion of an intra-mammary antibiotic in all quarters of all cows at dry-off, irrespective of infection status. The aim is to cure any pre-existing intra-mammary infections (IMI) and prevent new IMI over the dry period. An alternative strategy is selective dry cow therapy (SDCT) whereby cows with a low probability of an IMI receive an internal teat sealant (ITS) instead of antibiotics to prevent new IMI. Using SDCT instead of BDCT can considerably reduce antibiotic use. A key clinical question underpinning the use of SDCT is whether practitioners believe that ITS is as effective as an antibiotic, or better, at preventing new IMI in uninfected quarters. Therefore, this study aimed to quantify practitioners' beliefs for the efficacy of an antibiotics versus ITS, before and after exposure to

new data. The study illustrates a practical way to quantitatively investigate how practitioners updated their beliefs compared to Bayes theorem.

2. Materials and methods

2.1. Recruitment of practitioners

In total, 20 practitioners were recruited from 6 practices in the South of England. Sample size was a pragmatic choice in keeping with existing research involving the elicitation of beliefs (Johnson et al., 2010a). Since we set out to demonstrate a method to quantitatively assess changes in belief, rather than to draw inferences to a wider population, we selected practices based on the convenience of being within a practical driving distance. Inclusion criteria were practitioners providing healthcare to cattle during their normal working hours. Voluntary signed consent was obtained. Individual face-to-face interviews lasting 30 minutes were conducted by HMH between 1 December 2014 and 31 January 2015, at the participants' workplaces. A standard script was used for consistency. The script was piloted on 3 practitioners to ensure the method was tenable. This pilot data is not included.

2.2. Clinical context and definition of elicited parameter, ϕ

The population of interest was cows with all four quarters uninfected at dry-off. To ensure everyone considered the same population and to avoid any potential confusion over uncertainty associated with diagnosing cows as uninfected, it was emphasised to participants that when giving their answers the population they must consider were cows that were *genuinely* uninfected (free from both major and minor pathogens) i.e. they must assume that this had been 100% reliably established.

The outcome of interest was the dry period new infection rate, defined as the percentage of uninfected quarters that acquire a new infection during the dry period. The new infection could be either a major or minor pathogen, and result in either a sub-clinical or clinical

infection. To avoid any potential ambiguity associated with diagnosing the new infections, it was emphasised to participants that when giving their answers they simply needed to consider how many quarters would *actually* acquire a new infection, and diagnosing this infection was not something they needed to think about.

The two treatments considered at dry-off were (i) a long acting intra-mammary suspension containing 0.250g cefalonium per syringe, administered correctly and (ii) an ITS, specifically a 4g intra-mammary suspension containing 65% bismuth subnitrate, administered correctly. To ensure everyone considered the same baseline, participants were told the infection rate with cefalonium was 30%. The elicited parameter, ϕ, was the *difference* in the infection rate if an ITS was used instead of cefalonium, i.e. how much higher (+) or lower (-) than 30% the infection rate would be with ITS. It was assumed that influencing factors other than which treatment was given remained constant and the dry period was 60 days. We chose to compare ITS to antibiotics rather than giving no treatment because giving no treatment in uninfected cows is generally not considered a tenable option in high yielding dairy cows in the UK, mainly due to delay in closure of the teat canal keratin plug (Dingwell et al., 2004). Furthermore, we elicited the difference in the infection rates, as opposed to other measures of relative efficacy, because we considered adding and subtracting to be the simplest way for participants to give their answer.

The probability that an ITS treated, uninfected quarter would be infected at calving was denoted by θ. After the interviews were finished, participants beliefs for θ were calculated using

$$\theta = \frac{30 + \phi}{100}. \qquad (1)$$

We could have elicited beliefs directly for θ (the infection rate with ITS conditional on a 30% infection rate with cephalonium), rather than the difference. However, by asking for the difference it was envisaged that it would facilitate participants to think more carefully about which treatment was better (the infection rate with cephalonium was marked in red on the chart, see Fig 1).

2.3. Probabilistic elicitation method to capture beliefs

A variety of different methods have been reported to elicit beliefs probabilistically (Johnson et al., 2010a). This study used a version of the roulette method (also called chip and bins) because it is a method that has been shown to be feasible, valid and reliable for capturing beliefs in a clinical setting (Johnson et al., 2010b). Current best practice for elicitation was followed (Garthwaite et al., 2005; O'Hagan et al., 2006). This included: (i) a face-to-face interview; (ii) providing a training exercise; (iii) use of a standardized script; (iv) a design that avoided heuristics; (v) provision of feedback; (vi) opportunity to revise responses and (vii) use of simple graphical methods.

Following the general methodology of Johnson et al. (2010b), participants were asked to express their beliefs probabilistically by indicating the weight of their belief for ϕ using 10 chips each worth 0.1 probability, and placing them in discrete 5 per cent intervals (the bins) over the range of ϕ. Coins, specifically 5 pence pieces, were used for the chips. Adhesive putty (Blu-Tack®, Bostik) was used to make the coins adhere to, but be easily detached from, a laminated sheet. This allowed participants to easily revise their answers.

The training exercise took approximately 5 minutes. Participants were shown 3 generic examples involving 2 treatments, A and B. Each example demonstrated a different belief and their meanings were explained (see Fig 1). No context was provided in order to reduce any bias. Example 1 (Fig 1) represents a practitioner who "believes confidently" that treatment B is definitely inferior to A, because they have assigned 0.8 probability (80% chance) that the infection rate will be higher with B by 25-30%. Example 2 represents a practitioner who believes that treatment B is definitely superior to A, but they hold a less confident belief compared to example 1. Example 3 represents a belief that favours treatment B, but allows some probability (0.3 in total) that A is superior. Afterwards, the examples were placed out of sight to mitigate bias.

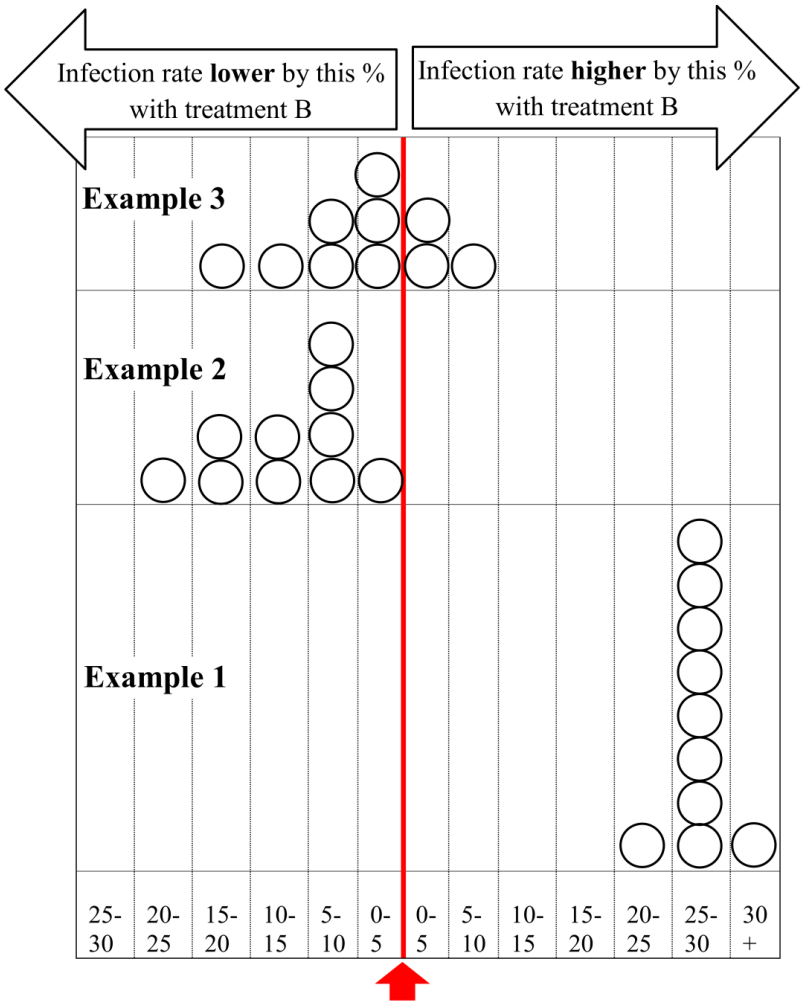

Fig.1. The 3 examples used for training. Each circle denotes a 0.10 probability. Example 1 represents a "confident" belief that treatment B is inferior to A. Example 2 represents the belief that treatment B is superior to A; it is a less confident belief relative to example 1. Example 3 represents a belief that favours treatment B, but allows some probability that A is superior.

2.4. Task structure and new data presented

The task itself started by capturing participants' current beliefs for ϕ as probability distributions using the roulette method (see Section 2.3). This distribution is denoted by ϕ_1. They were then shown the results (point estimate and 95% classical confidence interval) from

a clinical trial. The trial data was fictitious and followed a binomial distribution, $Y \sim \text{Binomial}(n, \theta)$, where n is the number of uninfected quarters at dry-off , θ is (as defined previously) the probability that an ITS treated, uninfected quarter is infected at calving, and the number of infected quarters were realisations, y, on the random variable Y. n was set to 1000, which resulted in 95% confidence intervals that were ~5% wide i.e. the width of one bin. The 95% binomial confidence intervals (Wilson, 1927) were calculated using the 'binom' package in the software program R (R Core Team, 2015).

The data, y, each practitioner was shown was dependent on their prior distribution, as follows. The point estimate for the data was centred in the 5% interval adjacent to the mode of their prior and favouring cefalonium. For example, the mode of the prior in Example 2, Figure 1, is in the lower -(5 to 10)% interval for ϕ (i.e. θ = 0.2-0.25), and the adjacent interval to this is -(0 to 5%), (i.e. θ = 0.25-0.3). Hence the point estimate of the data would be in the middle of this at ϕ = -2.5%. Using Eq. 1, this is $\theta = 0.275$, and with n=1000, this means $y = 275$ infected quarters. By varying the point estimate of the data according to each participant's prior belief, all participants had their prior beliefs challenged to the same extent with respect to central location.

Due to the properties of the binomial distribution, the width of the confidence interval each participant was shown differed slightly (range 4.4 - 5.4%) as it was related to their prior mode. However, in the context of the *precision of the task* as it was set (5% interval bins), differences in the strength of data they were shown was minor.

The final part of the task involved re-eliciting the practitioner's belief for ϕ as a probability distribution after seeing the new data using a new set of coins and laminated sheet (see Section 2.3). This probability distribution is denoted by ϕ_2. During the interviews, no

analysis was done and practitioners were *not* told how their updated beliefs compared to a Bayesian analysis.

2.5. Initial data manipulation

The raw data was entered into a Microsoft Excel spread sheet (v2010, Microsoft Corp). For each participant, their probability distributions ϕ_1 and ϕ_2 were transformed to probability distributions for θ using Eq. 1, to yield θ_1 and θ_2. Hence θ_1 was the practitioner's *prior distribution* and this will also be referred to as their *prior belief* for θ . Similarly, θ_2 was their *elicited posterior distribution* and this will also be referred to as their *elicited posterior belief* for θ. Since the raw data comprised discrete quantities of probability placed in fixed intervals, the mean and variance of θ_1 and θ_2 were simply calculated using

$$E(\Theta) = \sum_{all\,\theta} \theta_i \; p(\theta_i) \qquad \text{and} \qquad V(\Theta) = \left[\sum_{all\,\theta} \theta_i^2 \, p(\theta_i)\right] - [E(\Theta)]^2$$

where $E(\Theta)$ is the mean, $V(\Theta)$ the variance, θ took the possible values of $\{0.025, 0.075, 0.125, \ldots, 1\}$ corresponding to the mid-point of each interval and $p(\theta_i)$ is the probability placed in each interval. A parametric distribution from the Beta family was fitted to θ_1 and θ_2 for each practitioner by equating $E(\Theta)$ and $V(\Theta)$ to the first and second moments of the Beta family expressed in terms of its two hyper-parameters (α, β), and solving the two simultaneous equations (Gupta and Nadarajah, 2004):

$$E(\Theta) = \frac{\alpha}{\alpha + \beta} \qquad \text{and} \qquad V(\Theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{2}$$

From the fitted distributions, summary statistics including 95% equal tailed Bayesian credible intervals, means and standard deviations were calculated.

2.6. Conjugate Bayesian analysis

For each practitioner, Bayes theorem was used to combine their prior distribution with the new data they were shown, to produce a posterior distribution. This posterior distribution is subsequently referred to as their *Bayesian posterior distribution* (or *Bayesian belief*) and has probability density function, $\pi(\theta|\ y,n)$. It is a theoretical distribution that expresses how each practitioner *should* have updated their prior belief, according to Bayes theorem. Since Beta distributions were fitted to the prior beliefs and the new data followed a binomial distribution, this Bayesian analysis was conjugate with the Bayesian posterior distribution also taking the form of Beta distribution:

$$\theta_1 \sim Beta(\alpha,\beta)\ ;\ \pi(y|n,\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}\ ;$$

$$\theta|\ y,n \sim Beta\ (\alpha+y, n-y+\beta) \qquad (3)$$

where α and β were derived for each practitioner by solving Eqs. 2, $\pi(y|n,\theta)$ is the binomial likelihood function, n=1000 and y varied with the mode of θ_1 (see section 2.4)

By comparing their elicited posterior belief, θ_2, with the Bayesian posterior belief, calculated from Eq. 3 it was possible to quantify how close practitioners were to Bayesian logic. In this respect, for continuous parameters such as θ, there are two key elements to consider. First, whether the elicited posterior belief is centred as in keeping with Bayes theorem, which in this context could be termed their 'clinical expectation'. Second, whether the elicited posterior belief carries the appropriate uncertainty, which in this context could be termed their 'clinical confidence'. To make an assessment of each of these, we used the mean and the standard deviation of the elicited posterior distribution and the Bayesian posterior distribution. We produced scatterplots of the elicited posterior means versus Bayesian posterior means, and the elicited posterior standard deviations versus the Bayesian posterior

standard deviations. Other metrics for quantifying the overall distance between two probability distributions could have been used. However, they are less useful in this context because they do not allow a separate assessment of clinical expectations and clinical confidence, i.e. the first and second moments of the distributions.

2.7. Discrete Bayesian analysis

Since the raw data comprised discrete quantities of probability placed in fixed intervals it was also possible to use these directly as prior discrete probability distributions because it was possible to solve the summation in the denominator of Bayes theorem and calculate a Bayesian discrete posterior distribution for each practitioner as follows:

$$\pi(\theta|y,n) = \frac{\pi(\theta)\,\pi(y|\theta)}{\sum \pi(\theta_i)\,\pi(y|\theta_i)}$$

where $\pi(\theta|y,n)$ is the probability mass function of the Bayesian posterior distribution, $\pi(\theta)$ is the prior probability mass function, $\pi(y|\theta)$ is the binomial likelihood function, and the summation in the denominator is over possible values for θ of $\{0.025, 0.075, 0.125, \ldots, 1\}$. We compared the results obtained to those derived from the conjugate Bayesian analysis (section 2.6), in order to make an assessment of the sensitivity of the results to using the discrete versus the continuous form of Bayes theorem.

3. Results

3.1 Prior beliefs

Figure 2 presents summary statistics (95% credible interval and mean) for the practitioners' prior beliefs for θ derived from the fitted Beta distributions. In Fig. 2, practitioners are ordered vertically by their prior mean. The majority of practitioners had the mean of their distribution ≤ 0.3 suggesting that the clinical opinion of practitioners in this

sample was an expectation that ITS was either equivalent or superior to cefalonium. However, only 8 practitioners (numbers 1-8 in Fig. 2) were entirely convinced of the superiority of ITS in the sense that they had their entire 95% credible intervals ≤ 0.3 and hence gave minimal probability for ITS being the inferior treatment. Overall, there was heterogeneity in clinical beliefs with respect to the efficacy of ITS compared to cephalonium, both in terms of centre of location and variance of the prior distributions. As illustrated in Fig. 2, several pairs of practitioners had non-over-lapping 95% credible intervals indicating very different clinical opinions.

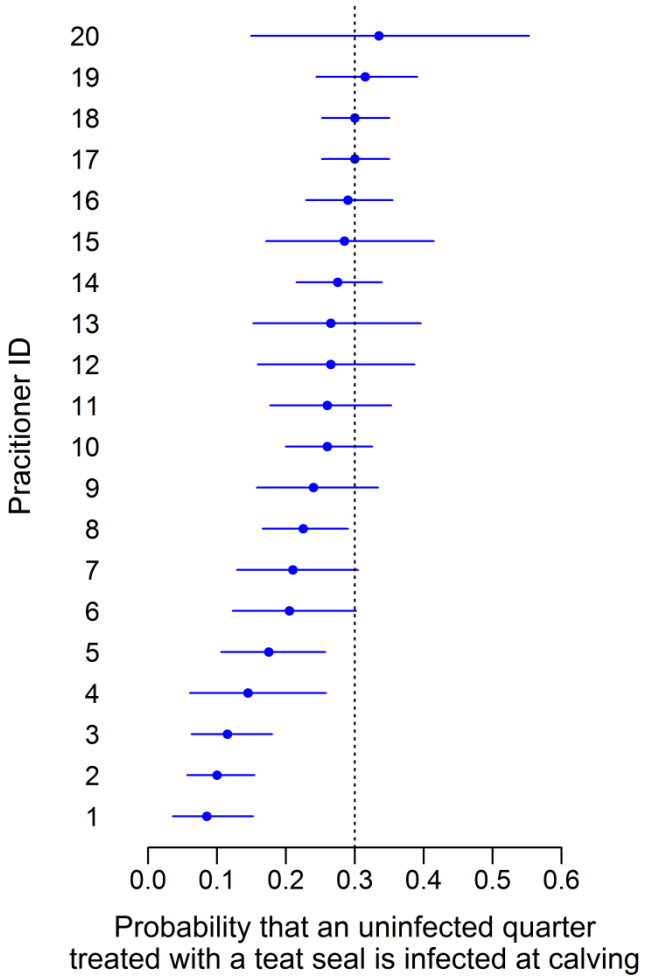

Fig. 2 Prior beliefs (95% credible interval and mean) for 20 practitioners for θ: the probability that an uninfected quarter treated with a teat seal is infected at calving, given a 0.3 probability with cephalonium.

3.2 Elicited posterior compared to Bayesian posterior distributions

For the 20 practitioners, Fig. 3 presents summary statistics for their prior, elicited posterior and Bayesian posterior distributions for θ derived from the conjugate Bayesian analysis; practitioners are ordered vertically by their prior mean.

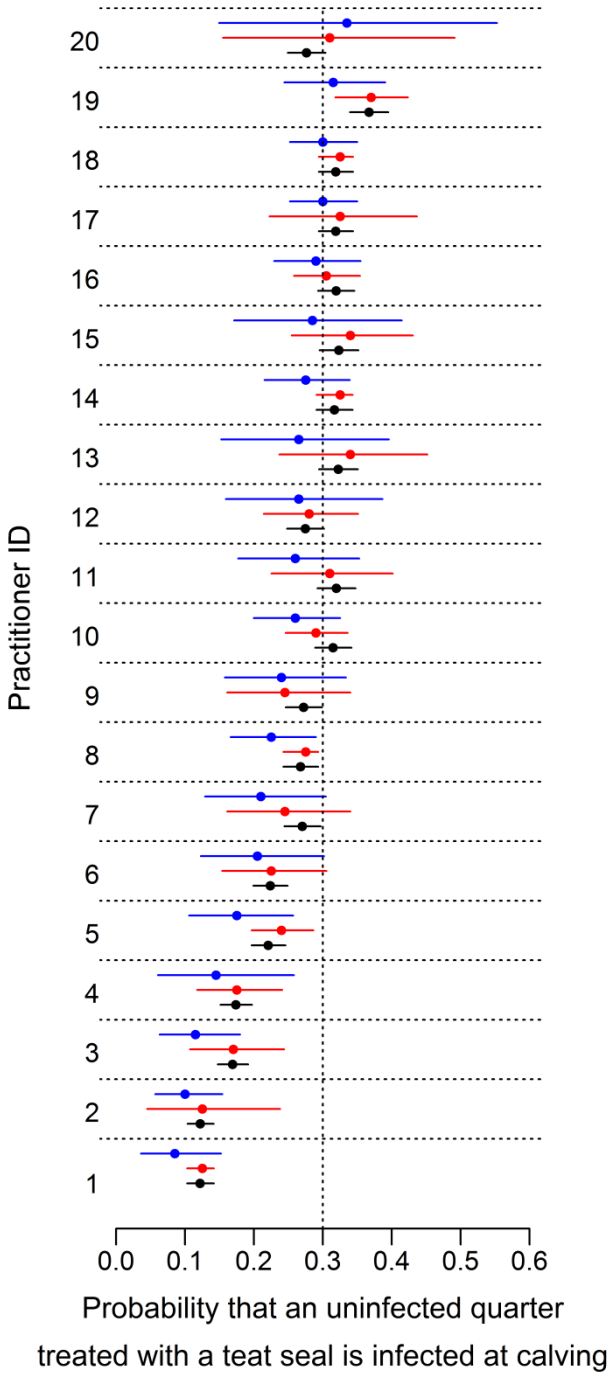

Fig. 3. Summary statistics (95% credible interval and mean) for the prior (blue), elicited posterior (red) and Bayesian posterior (black) distributions for 20 practitioners for θ: the

probability that an uninfected quarter treated with a teat seal is infected at calving, given a 0.3 probability with cephalonium.

In total, 4 out of the 20 practitioners, (numbered 1, 8, 14 and 18 in Fig. 3) updated their beliefs perfectly logically in light of the new data, both central location and variance (i.e. clinical expectation and confidence). For all the participants, their prior beliefs were weak relative to the strength of new data they were shown, and hence all the Bayesian posterior beliefs predominately reflected the new data.

Figure 4 presents a scatterplot of the elicited posterior versus the Bayesian posterior means. The dashed diagonal line in Fig. 4 is the line along which the elicited posterior mean equals the Bayesian posterior mean. Practitioners falling on this line were exactly in keeping with Bayes theorem. Figure 4 shows that most practitioners updated their clinical expectations either exactly, or close to, Bayesian logic.

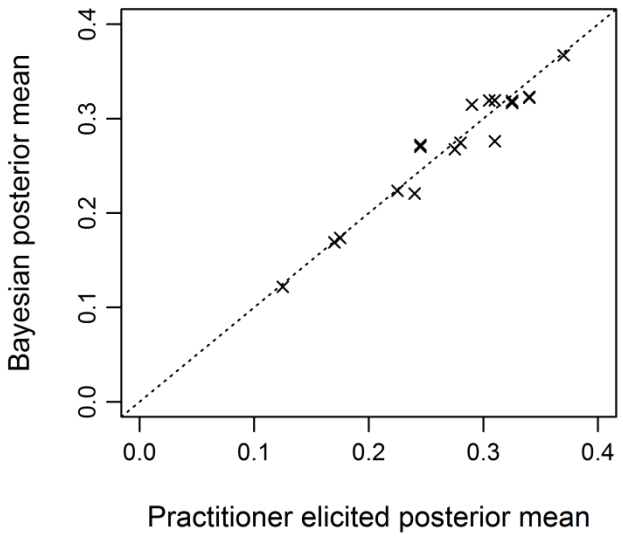

Fig. 4. Scatterplot of elicited posterior versus Bayesian posterior means for 20 practitioners regarding θ: the probability that an uninfected quarter treated with a teat seal is infected at calving, given a 0.3 probability with cephalonium.

In marked contrast, however, the majority did not update their uncertainty (i.e. their clinical confidence) in keeping with Bayesian logic. Figure 5 presents a scatterplot of the elicited posterior versus the Bayesian posterior standard deviations. The diagonal line denotes equality and hence practitioners falling on this line updated their uncertainty about θ exactly in keeping with Bayes theorem.

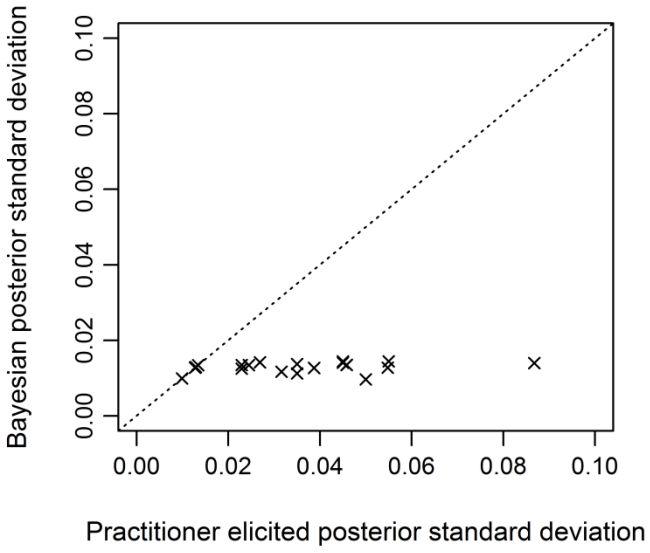

Fig. 5. Scatterplot of elicited posterior versus Bayesian posterior standard deviations for 20 practitioners regarding θ: the probability that an uninfected quarter treated with a teat seal is infected at calving, given a 0.3 probability with cephalonium.

In addition, while not apparent from Figure 3, eight practitioners (numbers 4, 5, 6, 10, 12, 15, 16, 20, Fig 3) had elicited posterior distributions with variance less than their prior but greater than their Bayesian posterior distribution, and this is reflected in the relative widths of their 95% credible intervals in Figure 3. In total, six had elicited posterior distributions with variances equal to their prior (numbers 3, 7, 9, 11, 13, 19 Figure 3), and two had elicited posterior distributions with variances greater than their prior (numbers 2 and 17, Figure 3); the latter is a considerable departure from Bayes theorem. Thus, most practitioners did not draw as much confidence from the new data as was logical.

3.3 Discrete versus conjugate Bayesian analysis

Updating the discrete prior distributions directly gave the same results as fitting parametric Beta distributions to the raw data and performing a conjugate Bayesian analysis.

4. Discussion

4.1 Clinical implications for the prescription of antibiotics versus teat sealants

The existing evidence suggests that in uninfected quarters at dry-off, using an ITS instead of an antibiotic is as effective at preventing new IMI (Rabiee and Lean, 2013). The priors we elicited suggest that some practitioners' are currently not entirely convinced of the efficacy of teat sealants when administered correctly. We chose not to give participants a review of the current evidence at the start of their interviews and therefore part of the variation in their prior beliefs may have been due to differences in their awareness or interpretation of the current literature.

The fact that some practitioners' are currently not entirely convinced of the efficacy of teat sealants may play a role in the implementation of SDCT in practice and hence it has potential implications for responsible antimicrobial prescribing because using SDCT instead of BDCT can considerably reduce antibiotic use. However, there are, of course, many important practical and psychological barriers to implementing SDCT in reality and the clinical beliefs of practitioners are just one of many factors in the broader perspective.

4.2 The use of Bayes theorem and a Bayesian approach in a clinical setting

Bayes theorem is derived directly from the fundamental axioms of probability theory. There is no controversy among statisticians that Bayes theorem *per se* is correct, in the sense that it is the logical way to update prior information based on new data (Spiegelhalter et al., 2004). Bayes theorem cannot think (!) and as such is impervious to psychological issues, heuristics or bias. In contrast, humans on occasion may fall foul of such factors when updating their beliefs. On the other hand, since humans can think they can also consider other factors when updating their beliefs which Bayes theorem cannot do, such as how trustworthy they consider the new data to be. Therefore, prima facie, it may be tempting to believe that it is not appropriate to use Bayes theorem as a gold standard for clinical belief updating. There are three key points here, however.

Firstly, when using Bayes theorem to update data there is an important, albeit somewhat implicit, underlying assumption which is that the new data is valid, i.e. that it was produced by a robust, appropriately designed and conducted scientific experiment, and that it yields information directly about the clinical parameter of interest.

Secondly, using Bayes theorem and a Bayesian approach does not mean that clinical judgement is not important. On the contrary, by taking a Bayesian approach, the existence and validity of clinical experience and judgement is formally incorporated into the analysis,

in the form of the prior information. This means, for example, that if a practitioner is confidently sceptical that a new treatment will not work, based on their current clinical experience and knowledge, this will be reflected in their prior probability distribution and hence also their posterior distribution and the inferences they will draw from the new data; with a strongly sceptical prior the new data would have to be very strong to yield a posterior that gives any support for the new treatment. Furthermore, once a practitioner has logically updated their beliefs with new data, clinical judgement and human thinking are crucial in order to take into consideration a multitude of contextual and social factors that will determine how the evidence should be used and applied in clinical practice.

Lastly, the use of a Bayesian approach in a clinical setting has been steadily growing over the last 25 years and it has now permeated all major areas of medical statistics (Ashby, 2006). It has been strongly argued by leading statisticians and psychologists that clinicians' prior beliefs should be elicited before a new clinical trial commences and a Bayesian approach used to facilitate the design, monitoring and interpretation of new data (Edwards et al., 1963; Parmar et al., 1994; Spiegelhalter et al., 1994; Berry, 1996; Parmar et al., 2001; O'Hagan and Luce, 2003).

4.3 Clinical expectations versus clinical confidence in belief updating for continuous parameters

It was an interesting contrast to observe that the majority of practitioners updated the central location (clinical expectation) of their distributions logically, but the variance (clinical confidence) of their beliefs illogically. For this task, the data was centred close to the participant's prior mode and hence in terms of central location per se, the data was not radically challenging any of the practitioners' prior beliefs. In contrast, the strength of the evidence they were shown was, for all participants, much stronger than the strength of their

prior beliefs (i.e. the variance of their prior distributions). Hence with respect to variance per se, the data was challenging the participants' prior beliefs to a reasonable degree. We hypothesize that this may to some extent explain the observed contrast; more negative reactions may have been evoked by the greater challenge to their prior beliefs with respect to the variance than central location, and negative reactions may impede updating in accordance with Bayes theorem.

4.4 Implications for evidence based veterinary medicine

Evidence based practice relies heavily on changing practitioners' beliefs by presenting them with new data. The results, however, provide quantitative support for the notion that new data which differs from practitioners' current beliefs can generate uncertainty (Ellsberg, 1961; Politi et al., 2007). Thus, new data may result in practitioners failing to draw enough confidence from the evidence or even in them having weaker beliefs upon which to base their clinical decisions than they did before. This 'psychological handicap', effectively a hurdle of doubt, is important for clinical researchers and data monitoring committees to bear in mind when assessing the strength of conviction that new data may bring to the clinical community.

A difficult decision when conducting a new clinical trial is when to stop it. One consideration is whether the accruing evidence is strong enough to be convincing to clinicians, even those who currently hold relatively sceptical beliefs about the new treatment that is being assessed by the trial. One way to make an assessment of the impact of the accruing results of a new trial on clinicians is to use a Bayesian approach (Fayers et al., 1997). Thus, as the results of a clinical trial accumulate, they could be shown to clinicians and their beliefs elicited probabilistically. By doing this at regular intervals of time during the clinical trial as an interim analysis, the decision over when to stop the trial would be facilitated; when the data is strong enough to result in previously sceptical clinicians having a posterior distribution that favours the new treatment, the trial could be stopped. Using the

roulette method employed here, it would not be arduous task to elicit the beliefs of clinicians on a regular basis during a clinical trial.

The results also support the view that some practitioners may benefit from assistance to appropriately adjust their current beliefs in the light of new evidence, and in particular it may be worthwhile training practitioners to be more comfortable with uncertainty. An interpretation section in clinical papers that presents a variety of prior beliefs and demonstrates how the trial result should influence them may be helpful. This would enable practitioners to self-evaluate how they should adjust their clinical beliefs, and help them to make the best use of data. In turn, this would facilitate the efficient uptake of new evidence into clinical practice. It is important for clinical researchers to make their results transparent and easily interpretable to all practitioners in the context of their current clinical beliefs, and a Bayesian framework is ideally suited to this. Furthermore, the type of task described here could be used to help teach the logical updating of clinical beliefs and the concepts of Bayesian statistics to clinicians as part of undergraduate or postgraduate training.

4.5 Assessing the updating of clinical beliefs

When comparing practitioners' belief updating to Bayes theorem, it is worth noting that it is difficult to differentiate a practitioner who did not express their prior belief accurately as a probability distribution from a practitioner who appeared to update their prior belief illogically. By following best practice for elicitation, we mitigated this potential bias. Nonetheless, it is possible that some practitioners may have specified priors that did not reflect their true beliefs. Interestingly other authors have elicited clinical priors from doctors and remarked that some doctors appeared to give over-confident answers given the available evidence (Chaloner and Rhame, 2001). We hypothesise that, psychologically, at least some

practitioners' may have difficulties separating their bedside manner from their own uncertainty with respect to the evidence.

Other psychological issues in the updating of clinical beliefs that could usefully be explored in further studies include the consequences of believing the new information for a practitioner's previous and future actions. Practitioners may be resistant to change and react against it simply because they have perceptions that treatments they have used have worked in the past and their prior beliefs are overly and inaccurately strong (Brehm, 1966).

Furthermore in practice, practitioners' will usually update their clinical beliefs without being required to specifically think and probabilistically quantify what they currently believe first; thus of primary importance is how they update their clinical beliefs without having to first think and express their beliefs probabilistically. This is unfortunate, given it is currently impossible to know if a person has updated their belief logically, without in some way ascertaining what it was they believed to begin with. Of relevance here, is the existing debate in the literature over whether practitioners' have beliefs that already exist, pre-formed and coherent, and hence are 'ready for the taking' by elicitation (Lindley et al., 1979), or alternatively, whether their beliefs exist in a more diffuse state, and hence are 'conjured up on the fly' in response to the elicitation task itself (Winkler, 1967); if the latter is true, then we speculate that any potential differences between measurable and actual belief updating in practice, may be greater.

4.6 Use of discrete versus continuous probability distributions

We chose the roulette method to probabilistically elicit beliefs which directly produced discrete prior probability distributions, and in our case it was possible to explicitly (by hand) calculate the summation in the denominator of Bayes theorem to yield a discrete posterior distribution. However many methods to elicit beliefs do not produce discrete

probability distributions, instead a small number of summary statistics are elicited and it is common practice and mathematically convenient to fit continuous parametric distributions to the raw data to represent the prior beliefs. With the prior beliefs expressed in this format, a conjugate analysis (or in more complex cases, sophisticated simulation techniques) are needed to solve Bayes theorem and derive the posterior distribution.

The extent to which a fitted parametric density function actually represents what a person believes is a non-trivial statistical problem within the field of probabilistic elicitation that has not yet been resolved (Garthwaite et al., 2005; O'Hagan et al., 2006; Oakley and O'Hagan, 2007). This is because to specify the uncertainty in a continuous random variable X uniquely as a probability distribution requires eliciting an infinite collection of probability statements from the person, $P(X \leq x) \, \forall \, x$, which is impossible. The person can only provide a finite summary of their beliefs.

In our case, given we could both fit parametric distributions to our raw data and use the discretely elicited prior distributions directly, it was of interest to do both and compare the results. We chose to present the results using 95% credible intervals and means derived from the conjugate analysis using the fitted parametric distributions for convenience and reader familiarity.

4.7 Limitations

The results are conditional on the task as it was set and in particular our choice of the roulette method and the way we chose to challenge their beliefs (strength and central location of the new data). It is possible that different results would be obtained using a different methodological approach, thus repetition of this type of study is warranted. Furthermore, our design does not take account of any clustering of the data which may be relevant and add complexity to the design of the task. In addition, any method must be acceptable to

practitioners themselves and as simple as possible. Our perception was that the practitioners in our sample found the methodology acceptable; however the training exercise is important.

The fact that this study used fictitious data may potentially have given rise to some psychological implications, particularly that the results may not have been believed. Steps were taken to overcome this, however, including specifically emphasising at the outset that the task required participants to 'use their imagination and really believe' the trial results. Indeed, in our experience it is crucial to remind practitioners at the end of the interview that the data are synthetic in order to avoid them transferring incorrect information to clients.

Conflict of Interest. None

Acknowledgements

Our thanks go to all the practitioners who participated in this study. This research was funded by the University of Surrey.

References

Ashby, D., 2006. Bayesian statistics in medicine: a 25 year review. Statistics in medicine 25, 3589-3631.

Baker, C.L., Tenenbaum, J.B., Saxe, R.R., 2006. Bayesian models of human action understanding. Advances in Neural Information Processing Systems 18, 99-106.

Bayes, T., 1763. An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society 53, 418.

Berry, D.a.S., D., 1996. Bayesian Biostatistics Marcel Dekker, New York.

Brehm, J.W., 1966. A theory of psychological reactance. Academic Press New York.

Chaloner, K., Rhame, F.S., 2001. Quantifying and documenting prior beliefs in clinical trials. Statistics in medicine 20, 581-600.

Cooke, R.M., 1991. Opinion and subjective probability. Oxford University Press New York.

Dingwell, R.T., Leslie, K.E., Schukken, Y.H., Sargeant, J.M., Timms, L.L., Duffield, T.F., Keefe, G.P., Kelton, D.F., Lissemore, K.D., Conklin, J., 2004. Association of cow and quarter-level factors at drying-off with new intramammary infections during the dry period. Preventive veterinary medicine 63, 75-89.

Edwards, W., Lindman, H., Savage, L.J., 1963. Bayesian Statistical Inference for Psychological Research. Psychological Review 70, 193-241.

Ellsberg, J.W., 1961. Risk, ambiguity, and the savage axioms. Quarterly Journal of Economics 75, 643-669.

Fayers, P.M., Ashby, D., Parmar, M.K.B., 1997. Tutorial in biostatistics: Bayesian data monitoring in clinical trials. Statistics in medicine 16, 1413-1430.

Garthwaite, P.H., Kadane, J.B., O'Hagan, A., 2005. Statistical methods for elicitating prior distributions. Journal of the American Statistical Association 100, 680-701.

Gigerenzer, G., 2011. What are natural frequencies? British Medical Journal 343, d6383.

Griffiths, T.L., Tenenbaum, J.B., 2006. Optimal predictions in everyday cognition. Psychological Science 17, 767-773.

Gupta, A.K., Nadarajah, S., 2004. Handbook of Beta Distribution and Its Applications. Marcel Dekker.

Hoffrage, U., Lindsey, S., Hertwig, R., Gigerenzer, G., 2000. Communicating statistical information. Science 290, 2261-2262.

Johnson, S.R., Tomlinson, G.A., Hawker, G.A., Granton, J.T., Feldman, B.M., 2010a. Methods to elicit beliefs for Bayesian priors: A systematic review. Journal of clinical epidemiology 63, 355-369.

Johnson, S.R., Tomlinson, G.A., Hawker, G.A., Granton, J.T., Grosbein, H.A., Feldman, B.M., 2010b. A valid and reliable belief elicitation method for Bayesian priors. Journal of clinical epidemiology 63, 370-383.

Kynn, M., 2008. The 'heuristics and biases' bias in expert elicitation. Journal of the Royal Statistical Society: Series A 171, 239-264.

Lindley, D.V., Tversky, A., Brown, R.V., 1979. On the reconciliation of probability assessments. Journal of the Royal Statistical Society: Series A 142, 146-180.

O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., Rakow, T., 2006. Uncertain judgements: Eliciting experts' probabilites. Wiley Chichester.

O'Hagan, A., Luce, B.R., 2003. A primer on Bayesian statistics in health economics and outcomes research. Centre for Bayesian statistics in health economics, Sheffield.

Oakley, J.E., O'Hagan, A., 2007. Uncertainty in prior elicitations: A non-parametric approach. Biometrika 94, 427-441.

Parmar, M.K., Spiegelhalter, D.J., Freedman, L.S., 1994. The CHART trials: Bayesian design and monitoring in practice. Statistics in medicine 13, 1297-1312.

Parmar, M.K.B., Griffiths, G.O., Spiegelhalter, D.J., Souhami, R.L., Altman, D.G., van der Scheuren, E., 2001. Monitoring of large randomised clinical trials: A new approach with Bayesian methods. The Lancet 358, 375-381.

Phillips, L.D., Edwards, W., 1966. Conservatism in a simple probability inference task. Journal of Experimental Psychology 72, 346-354.

Politi, M.C., Han, P.K.J., Col, N.F., 2007. Communicating the uncertainty of harms and benefits of medical interventions. Medical Decision Making 27, 681-695.

R Core Team, 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rabiee, A.R., Lean, I.J., 2013. The effect of internal teat sealant products (Teatseal and Orbeseal) on intramammary infection, clinical mastitis, and somatic cell counts in lactating dairy cows: A meta-analysis. Journal of Dairy Science 96, 6915-6931.

Spiegelhalter, D.J., Abrams, K.R., Myles, J.P., 2004. Bayesian approaches to clinical trials and health-care evaluation. Wiley Chichester.

Spiegelhalter, D.J., Freedman, L.S., Parmar, M.K., 1994. Bayesian approaches to randomized trials. Journal of the Royal Statistical Society Series A 157, 357-416.

Spiegelhalter, D.J., Myles, J.P., Jones, D.R., Abrams, K.R., 2000. Bayesian methods in health technology assessment: A review. Health Technology Assessment 4, pp.41.

Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: Heuristics and biases. Science 185, 1124-1131.

Westover, M.B., Westover, K.D., Bianchi, M.T., 2011. Significance testing as perverse probabilistic reasoning. BMC Medicine 9:20.

Wilson, E.B., 1927. Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association 22, 209-212.

Winkler, R.L., 1967. The assessment of prior distributions in Bayesian analysis. Journal of the American Statistical Association 62, 776-800.

Figures and captions (black and white versions)

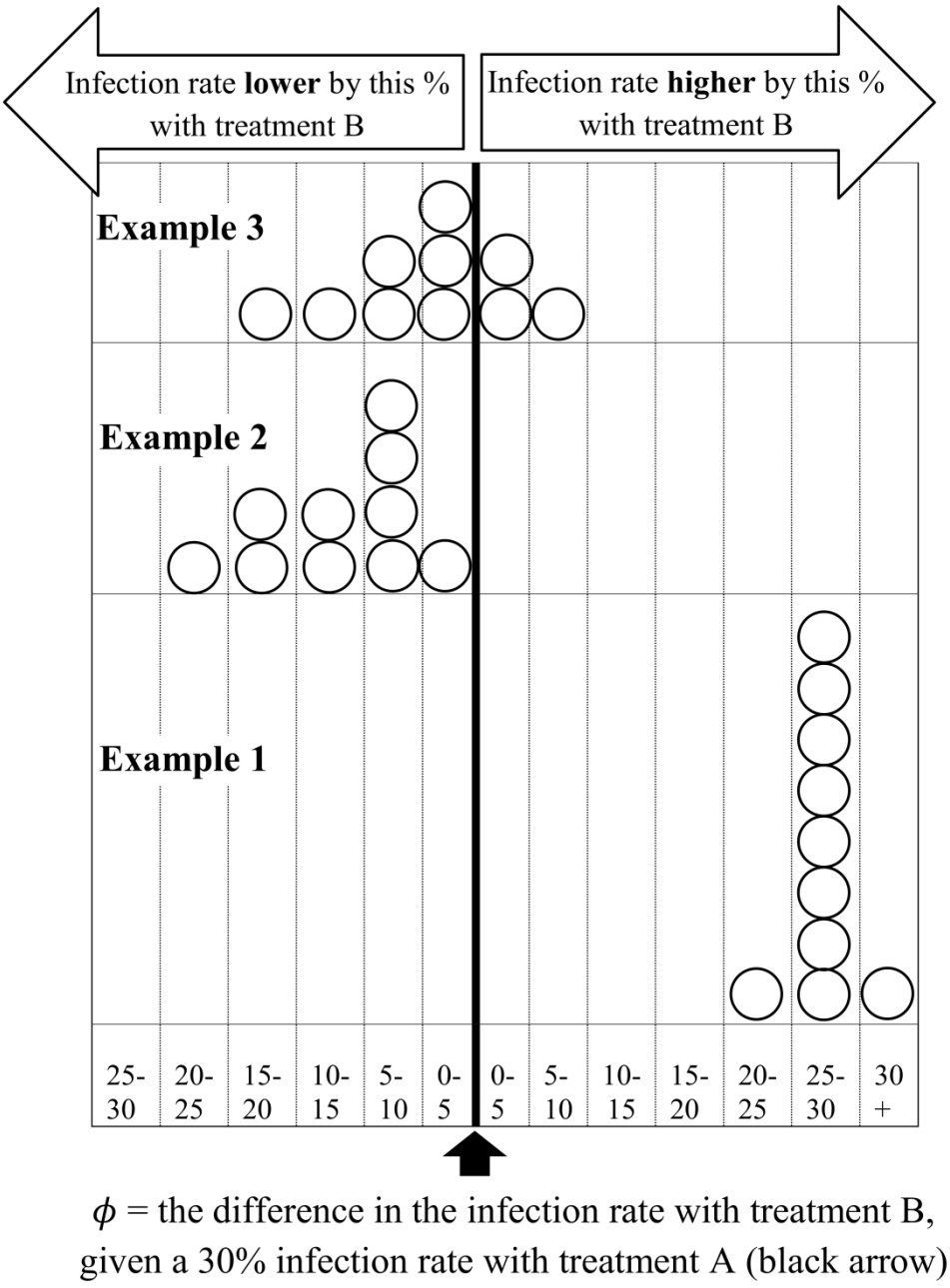

Fig.1. The 3 examples used for training. Each circle denotes a 0.10 probability. Example 1 represents a "confident" belief that treatment B is inferior to A. Example 2 represents the belief that treatment B is superior to A; it is a less confident belief relative to example 1. Example 3 represents a belief that favours treatment B, but allows some probability that A is superior.

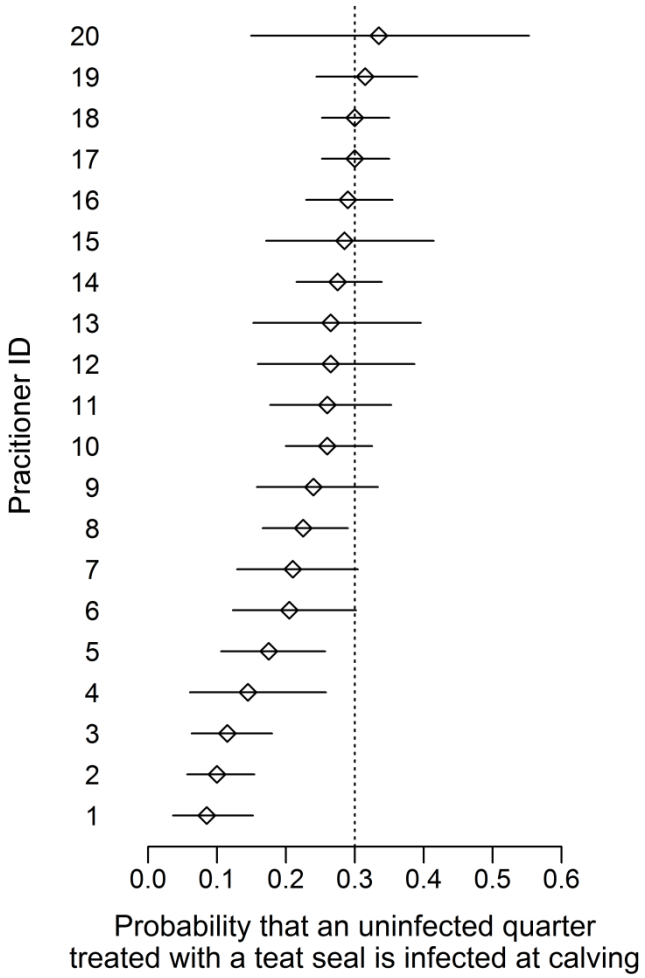

Fig. 2 Prior beliefs (95% credible interval and mean) for 20 practitioners for θ: the probability that an uninfected quarter treated with a teat seal is infected at calving, given a 0.3 probability with cephalonium.

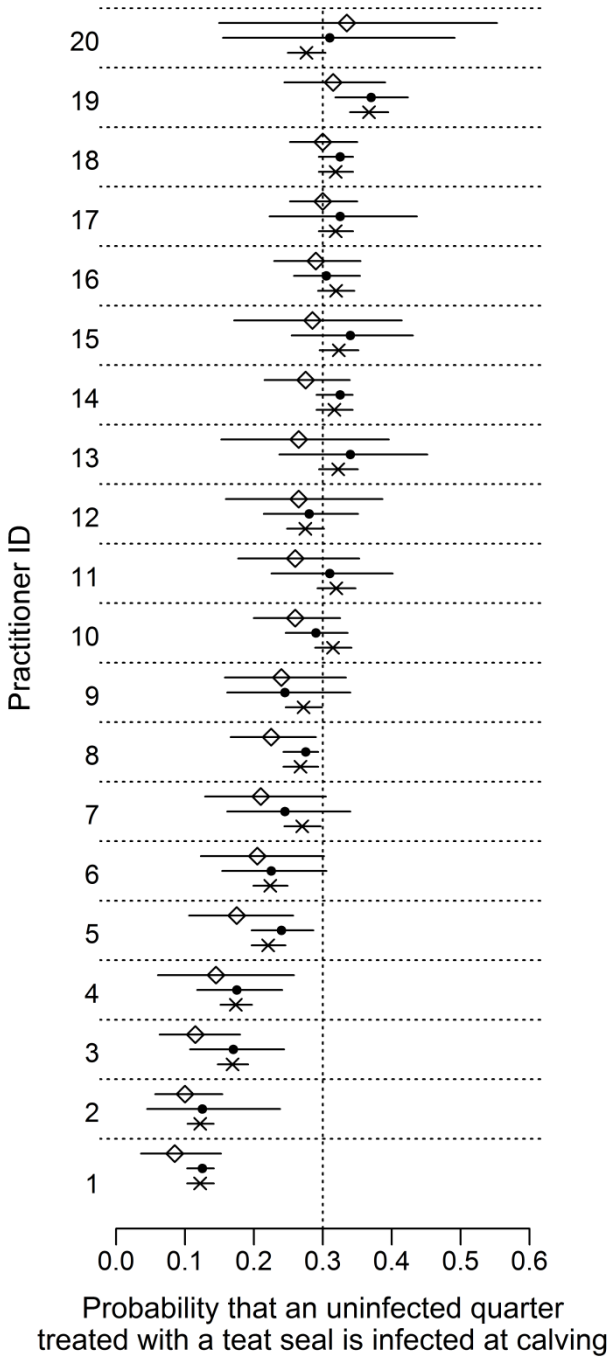

Fig. 3. Summary statistics (95% credible interval and mean) for the prior (diamond), elicited posterior (dot) and Bayesian posterior (cross) distributions for 20 practitioners for θ: the probability that an uninfected quarter treated with a teat seal is infected at calving, given a 0.3 probability with cephalonium.