Comparison of Three Alternative Methods for Analysis of Equine Faecal Egg Count Reduction Test Data

Denwood, M.J.[*,a], Reid, S.W.J.[a], Love, S.[b], Nielsen, M.K.[c], Matthews, L.[a], McKendrick, I.J.[d], Innocent, G.T.[d]

[a] *Boyd Orr Centre for Population and Ecosystem Health, Institute of Comparative Medicine, Faculty of Veterinary Medicine, University of Glasgow, Bearsden Road, Glasgow, G61 1QH, UK.*
[b] *Veterinary Companion Animal Sciences, Institute of Comparative Medicine, Faculty of Veterinary Medicine, University of Glasgow, Bearsden Road, Glasgow, G61 1QH, UK.*
[c] *Department of Large Animal Sciences, Faculty of Life Sciences, University of Copenhagen, Hoejbakkegaard Alle 5, DK-2630 Taastrup, Denmark.*
[d] *Biomathematics & Statistics Scotland (BioSS), The King's Buildings, Edinburgh, EH9 3JZ, UK.*

Abstract

The Faecal Egg Count Reduction Test (FECRT) is the most widely used method of assessing the efficacy of anthelmintics, and is the only *in vivo* technique currently approved for use with horses. Equine Faecal Egg Count (FEC) data are frequently characterised by a low mean, high variability, small sample sizes and frequent zero observations. Accurate analysis of the data therefore depends on the use of an appropriate statistical technique. Analyses of simulated FECRT data by methods based on calculation of the empirical mean and variance, non-parametric bootstrapping, and Markov chain Monte Carlo (MCMC) were compared. The MCMC method consistently outperformed the other methods, independently of the sample size and distribution from which the data were generated. Bootstrapping produced notional 95% confidence intervals containing the true parameter as little as 40% of the time with sample sizes of less than 50. Analysis of equine FECRT data yielded inconclusive results in 53 of 63 (84%) datasets, suggesting that the routine use of prior sample size calculations should be adopted to ensure sufficient data are collected. The authors conclude that computationally intensive parametric methods such as MCMC should be used for analysis of FECRT data with sample sizes of less than 50, in order to avoid making erroneous inference about the true efficacy of anthelmintics in the field. Software to perform all three types of analyses documented here is freely available in the form of an add-on package to the R statistical programming language from http://cran.r-project.org/web/packages/bayescount/index.html.

Key words: MCMC, bootstrap, WAAVP, FECRT, anthelmintic efficacy, equine

[*]Corresponding author
Email address: m.denwood@vet.gla.ac.uk (Denwood, M.J.)

1. Introduction

The Faecal Egg Count Reduction Test (FECRT) is the most widely used method of assessing the *in vivo* efficacy of anthelmintics against parasitic nematodes of horses, sheep and cattle (Coles et al., 2006; Kaplan, 2002), and is an essential tool in the process of monitoring the increasing prevalence of anthelmintic resistance. The test is known to have several limitations, including the variability of Faecal Egg Count (FEC) data (Uhlinger, 1993), leading to a relatively variable FECRT result (Miller et al., 2006). This is especially true in equine FEC data, where effects such as differing age related immunity (Klei and Chapman, 1999) and differences in grazing management (Dopfer et al., 2004) impact on the observed FEC. Combined with the small group sizes and frequent zero FEC observations (Kaplan, 2002; Nielsen et al., 2006) often encountered with horses, this high variability between animals and low mean FEC introduce difficulties in analysis of equine FECRT data which do not arise to the same extent in analysis of FECRT data obtained from cattle or sheep.

The method currently advocated by the World Association for the Advancement of Veterinary Parasitology (WAAVP) involves calculating the empirical mean and variance before and after treatment, and calculating the empirical mean reduction and estimates of the 95% confidence interval for the true reduction using these figures (Coles et al., 1992). This method takes no account of the difference between uncertainty regarding the true mean of a sample, introduced by the Poisson variability of the counting process, and variability in the true mean of different samples. Calculation of 95% confidence intervals in this manner also assumes that the distribution of error for the mean is symmetrical on the log scale, although parameter likelihoods (and therefore errors) have been reported to be skewed for FEC data (Denwood et al., 2008), potentially justifying this assumption.

A non-parametric bootstrapping approach has recently been suggested as an appropriate method to generate confidence limits from equine FECRT data (Vidyashankar et al., 2007). The technique involves re-sampling and summarising the observed data, and makes no assumptions about the underlying distribution or processes generating the data (Mooney and Duval, 1993), or the parameter error structure. Non-parametric bootstrapping approaches are therefore widely used and extremely useful when the underlying distribution of data is unknown. A fundamental assumption underlying this approach is that the data obtained are fully representative of the population, an assumption which risks being violated when dealing with small sample sizes, giving misleading results. A non-parametric bootstrapping approach is more complex and time consuming than the currently advocated WAAVP method, however the use of facilities like Excel spreadsheet macros and basic computer programs potentially allow different data to be analysed relatively quickly.

Alternative options for analysis of FECRT data include computationally intensive parametric methods. These include parametric bootstrapping, or the likelihood profiling method proposed by Torgerson et al.

(2005), but here Markov chain Monte Carlo (MCMC) (Gilks et al., 1998) is used as an example. Each of these methods requires the use of a parametric distribution in order to describe the FEC data. The negative binomial is the most frequently used parametric distribution for FEC data, and is equivalent to the gamma-Poisson compound distribution implemented here (for the derivation see Vose (2004)). Conceptually, this represents a population of Poisson distributions with gamma distributed means, where the Poisson distributions account for counting variability in observed FEC within a sample, and the gamma distribution describes the variability between samples. The latter could arise as a combination of several factors, including the aggregated distribution of eggs in faeces, variations in worm fecundity over time, variations in faecal consistency, and variations in the numbers of worms present, which are impossible to separate using only a single faecal sample per individual. For the MCMC model, pre-treatment data are assumed to follow a single gamma-Poisson (negative binomial) distribution, while post treatment data are distributed as a different gamma-Poisson distribution, with a mean value which has been scaled relative to the pre-treatment mean, and a value for variability which has separately been scaled relative to the pre-treatment variability. This allows inference on the true change in mean egg shedding, with an additional parameter reflecting the true change in variability between egg counts. From this model, estimates of the mean anthelmintic efficacy and the variability in anthelmintic efficacy between animals can be obtained. The advantage of an MCMC based approach is that the different sources of variability can be taken into consideration, leading to more accurate estimates of the uncertainty of true parameter estimates. Disadvantages of this approach include the comparatively high computational effort required to implement the method, and the need to make distributional assumptions about the processes generating the data. FEC between animals is well described by a gamma-Poisson (negative binomial) distribution, however alternatives include zero-inflated distributions (Denwood et al., 2008; Nødtvedt et al., 2002), and the use of a lognormal distribution to describe the variability in means (Morrison, 2004).

The bootstrapping and MCMC procedures also have the advantage of attempting to define the full distribution of likely values for the true FEC reduction. This allows the results to be presented in a more intuitive way, such as a probability that the true egg count reduction is less than a given percentage, typically the published efficacy of the drug used. This probability, $\hat{p}$, is relatively easy to estimate using numerical integration for both the bootstrapping and MCMC methods, and can allow the group to be classified as 'Susceptible' if $0 < \hat{p} < 2.5\%$, 'Possible resistant' if $2.5 < \hat{p} < 50\%$, 'Probable resistant' if $50 < \hat{p} < 97.5\%$, or 'Confirmed resistant' if $97.5 < \hat{p} < 100\%$. These definitions allow a distinction to be made between confirmed resistance and the lack of evidence of susceptibility, which is lacking in the current interpretation of lower 95% confidence interval and empirical mean reduction statistics described by Coles et al. (1992).

Given the worldwide importance of anthelmintic resistance, there is an urgent need to improve and standardise the statistical method used to analyse such data (Coles et al., 2006; Kaplan, 2002). Compared to the case with ruminants, the relatively small sample sizes, high variability between counts, and relatively

low pre-treatment mean FEC frequently encountered with equine FECRT data may provide a challenge to the use of a non-parametric bootstrapping procedure, since there are relatively few data points from which to sample. There is also often insufficient data to be able to analyse the underlying distribution, which prevents validation of the choice of distribution used by the MCMC analysis. The aim of this study was to assess the usefulness of 95% confidence intervals generated using these three methods using simulated data, and then to assess the impact of the assumptions being made for each method.

2. Materials and Methods

2.1. Statistical Analysis

The analysis currently recommended by the World Association for the Advancement of Veterinary Parasitology was performed as described by Coles et al. (1992). Bootstrapping was conducted using a function written by the author in the R statistical programming language (R Development Core Team, 2008). New pre-and post-treatment pseudo-datasets were sampled from each dataset, and the mean reduction calculated 10,000 times. The mean estimate and 95% confidence intervals for each dataset were then calculated and recorded from these 10,000 iterations.

Bayesian MCMC analysis was performed using a bespoke model, implemented using JAGS (Plummer, 2008) for the MCMC simulation. The model fits a gamma-Poisson distribution to the pre and post-treatment data, with parameters for pre- and post-treatment means and shape parameters. The pre-treatment mean and shape parameters are given minimally informative prior distributions spanning all values that are seen in real FECRT data for each parameter. Post-treatment mean and shape parameters are calculated by multiplying the pre-treatment mean and shape parameters by a "change in mean" and "change in shape" parameter, respectively. The "change in mean" is given an uninformative $Beta(1,1)$ prior, and the "change in shape" a diffuse lognormal prior with a mean of one. The true % FEC reduction is derived from $(1 - change\ in\ mean) * 100$. Calling JAGS to run each simulation and summarising of MCMC chains was automated using the runjags package (Denwood, 2008) for R, with two chains. Convergence was assessed using the Gelman-Rubin statistic (Gelman and Rubin, 1992), and necessary sample size using Raftery and Lewis's diagnostic (Raftery and Lewis, 1995). The median estimate and 95% credible intervals for the true egg count reduction were calculated in R using the MCMC output.

For all three methods, credible intervals for the proportion of datasets with the true reduction parameter contained within the nominal 95% confidence intervals were calculated using a Bayesian approach with an uninformative $Beta(1,1)$ prior. The mean relative size of these confidence intervals was calculated using equation (1).

$$confidence\ interval\ size = \frac{\sum \frac{U-L}{T}}{N} \tag{1}$$

Where L denotes the lower confidence interval, U the upper confidence interval, T the true parameter value, and N the number of datasets

To assess the accuracy of the median estimates, the relative root-mean-square-error (RMSE) was calculated using the simulated (true) value for each parameter. The RMSE can also be thought of as the standard deviation of the ratio between each median estimate and the simulated values; however it should be noted that this is not equivalent to the accepted meaning of the term "standard deviation". The term relative RMSE will be used to avoid confusion.

2.2. Comparisons of methods for analysis of FECRT data

A total of 1000 parameters for a simulated FECRT were generated in the R statistical programming language. The true proportional FEC reduction was simulated from a $Uniform(0.75, 1)$ distribution, so that true egg count reductions varied from reduced efficacy to efficacious reductions. The pre-treatment mean number of eggs counted (equal to FEC if the egg counting technique had an egg detection threshold of 1 EPG), and sample size (number of animals) were chosen to reflect the values seen in real equine FECRT data obtained from 63 typical Danish equine datasets. The 2.5% and 97.5% quantiles for observed pre-treatment mean and sample size were used as the lower and upper bounds of the distributions used to generate the parameters. Pre-treatment mean was taken from a $Uniform(1.45, 53.1)$ distribution, and sample size per group was sampled randomly from integers between 4 and 16 inclusive with each integer having an equal probability of selection. The coefficient of variation (cv) between samples before treatment was sampled from a $Uniform(1, 1.41)$ distribution (corresponding to a pre-treatment shape parameter of the gamma distribution, k, of between 1 and 0.5), and the proportional increase in cv after treatment was sampled from the same distribution (corresponding to a post-treatment shape parameter of between $1 * 1 = 1$ and $0.5 * 0.5 = 0.25$). These values were also chosen to reflect the values most likely to be encountered in real FECRT data; published values of k are usually less than one (Shaw et al., 1998), and differing efficacy of anthelmintic between animals would be expected to result in an increase in variability post-treatment.

In order to test the implications of the distributional assumptions made by the MCMC and WAAVP methods, simulated datasets were generated using the following three different distributions of underlying sample means; gamma-Poisson (negative binomial), multi-modal lognormal-Poisson, and uniform-Poisson. For each dataset, the meta-population mean and variance was the same for all distributions. The number of modes for each multi-modal lognormal-Poisson distribution was sampled as between two and ten for each dataset, and a separate lognormal distribution used to describe the distribution of modes within the group. These modes conceptually represent sub-groups within the population, with the population variance split equally between the two compound lognormal distributions for each animal. If the simulated parameter mean and variance required negative parameter value for the lower limit of the uniform-Poisson distribution, then a log-uniform distribution was used instead (that is, a distribution which is uniform on the log scale).

Pre- and post-treatment egg count data were generated using each of these three distributions with the 1000 parameter values, to simulate a FECRT for a total of 3000 datasets. These datasets were then analysed using each of the three methods. More details regarding the generation of these data are available from the corresponding author.

2.3. Equine FECRT data

The MCMC and bootstrap methods were applied to equine FECRT data obtained from 63 typical Danish equine establishments, with a median (range) of 9 (6-22) animals per dataset. For these data, a modified MCMC method using zero-inflated gamma-Poisson distributions in place of the gamma-Poisson distributions was used, in addition to the MCMC method described previously. For each dataset, the probability, $\hat{p}$, that the observed FEC reduction was less than the "desired" FEC reduction was calculated by numerical integration of the posterior estimates for true FEC reduction. From this, the dataset was classified as 'Susceptible' if $0 < \hat{p} < 2.5\%$, 'Possible resistant' if $2.5 < \hat{p} < 50\%$, 'Probable resistant' if $50 < \hat{p} < 97.5\%$, or 'Confirmed resistant' if $97.5 < \hat{p} < 100\%$. In this study, the "desired" FEC reduction was set 95%, corresponding to the best estimate of the efficacy of the drug used in a naïve population. This figure represents the minimum population mean FEC reduction we would expect from a fully susceptible group of animals if we were able to observe the true mean FEC before and after treatment, and could be adjusted for both methods with other datasets depending on the drug used and desired tolerance in true efficacy.

2.4. Bootstrap analysis

A more complex analysis of the performance of the bootstrapping method was performed using gamma-Poisson data. Sample size was drawn from the set {5, 10, 20, 30, 40, 50, 60, 70, 80, 90 & 100}, and pre-treatment mean number of eggs counted from the set {1, 5, 10, 20, 30, 40, 50, 75 & 100}. Each of these 99 combinations was used to generate 1000 datasets using two gamma-Poisson distributions and a true FEC reduction randomly generated from a $Uniform(0.75, 1)$ distribution. For each dataset, the parameter value used for pre-treatment *cv* was either 1 or 1.41, and post-treatment change in *cv* either 1 or 1.41. Each dataset was analysed using the bootstrap method to provide a median estimate and 95% confidence intervals as before.

3. Results

3.1. Comparisons of methods for analysis of FECRT data

Of the 3000 datasets, 35 of the gamma-Poisson datasets, 32 of the multi-modal lognormal-Poisson datasets, and 33 of the (log) Uniform-Poisson datasets gave an empirical reduction of 100%. The median (95% confidence interval) simulated true reduction for these empirical 100% reduction datasets was

99.13% (82.23% - 99.97%). As the post-treatment variance for these datasets was 0, the WAAVP method of calculating 95% confidence intervals could not be applied. In practice, these datasets would be assumed to represent a 100% reduction, so 95% confidence limits of 100% to 100% were assigned to these datasets. The non-parametric bootstrapping approach generated the same confidence limits for these datasets, since all possible combinations of datapoints give a 100% reduction.

In Figure 1, the proportion of true reductions that were contained within the notional 95% confidence intervals for each method with all datasets are shown (95% credible intervals calculated using a Bayesian method with an uninformative prior). There is no evidence that the MCMC method did not estimate true 95% confidence intervals for both the gamma-Poisson and (log) Uniform data, but the confidence was lower for the multi-modal data. Non-parametric bootstrapping and the WAAVP method both returned notional 95% confidence intervals that contained the true value between 85% and 90% of the time for all data types. Discounting the datasets with an empirical reduction of 100% improved the apparent performance of the bootstrapping and WAAVP methods, although both methods still generated lower estimates of confidence than the MCMC method for all data types (data not shown).

In Table 1, the mean relative size of the notional 95% confidence intervals for each method and dataset are shown. The relative RMSE for each combination is shown in Table 2. The MCMC method returned on average slightly larger 95% confidence limits than the other methods for each dataset, although when datasets with 100% apparent reductions were excluded, the three methods produce similarly sized 95% confidence intervals (data not shown). The 95% confidence intervals were largest for the (log) Uniform-Poisson data, and most narrow for the multi-modal data. The MCMC median estimates produced a lower relative RMSE than the bootstrapping median and WAAVP mean estimates in every case. The bootstrapping median and WAAVP mean estimates generally had a similar relative RMSE, although those produced by the bootstrapping method were lower. As for the relative size of 95% confidence intervals, the relative RMSE was smallest for each method for the multi-modal data and largest for the (log) Uniform-Poisson data.

3.2. Analysis of equine FECRT data

The probabilities of resistance returned by the bootstrapping and modified MCMC method relative to the first MCMC method are shown in Figure 2. The probabilities were greater for MCMC than bootstrapping in all but one case, indicating that bootstrapping consistently estimated the true efficacy to be higher than the estimates produced by MCMC. Estimates produced by the modified MCMC method using the zero-inflated gamma-Poisson distribution were very similar to those produced by the first MCMC method using the uni-modal gamma-Poisson distribution.

In Table 3, the classifications made for each dataset using each method are shown. None of the datasets were classified as confirmed susceptible using the MCMC method, and of 14 (22%) classified as confirmed susceptible with the bootstrap method, four (6%) were classified as 'probable resistant' using the MCMC

method. In addition, seven (11%) of the datasets were classified as 'confirmed resistant' using MCMC and only 'probable resistant' using the bootstrap method. There was insufficient information in the data to determine either confirmed resistance or susceptibility for 53 datasets (84%) using MCMC and 46 datasets (73%) using the bootstrap method.

3.3. Bootstrap analysis

The effect of increasing pre-treatment mean FEC and sample size on the ability of the bootstrapping method to accurately predict the true FEC reduction is shown in Figure 3 and Figure 4. As pre-treatment mean FEC increased, the 95% confidence intervals were more reliable, although this affect appeared to be less pronounced with an increase in mean above ten counted eggs at sample sizes 20 and greater. With sample sizes of five and ten, the notional 95% confidence intervals contained the true parameter no more than 90% of the time, and as little as 40% of the time with a very low mean FEC. At sample sizes 20 to 40, the 95% confidence intervals contained the true parameter between 90% and 95% of the time for pre-treatment mean FEC of over ten counted eggs. This improved to between around 93% and 95% for sample sizes of 50 and above with pre-treatment mean FEC of ten counted eggs and above. Even with a sample size of 100, the notional 95% confidence intervals contained the true parameter between only 89% and 93% of the time with a pre-treatment mean FEC of one egg counted, and between 92% and 95% with a pre-treatment mean FEC of five eggs counted. Conversely, the confidence of the estimates produced by the MCMC method were not decreased by a reduced mean and sample size, with notional 95% confidence intervals containing the true value 99% of the time with a mean of 1 and sample size of 5, 97% of the time with a mean of 100 and sample size of 5, 97% of the time with a mean of 1 and sample size of 100, and 96% of the time with a mean of 100 and sample size of 100 (95% credible intervals not shown).

4. Discussion

For all datasets, simulated from each of the distributions tested, the MCMC method provided confidence intervals with the best defined properties, as well as the most precise median estimates for the true FEC reduction. The size of the 95% confidence intervals produced was slightly greater for the MCMC method, but not when datasets with empirical reductions of 100% were removed. This indicates that the MCMC methods were producing more appropriate 95% confidence intervals, rather than merely larger 95% confidence intervals. This was the case not only for data simulated from a gamma-Poisson distribution, where the MCMC method using the same distribution would be expected to perform well, but also using data simulated from different distributions. The performance of the MCMC method was less optimal using the multi-modal data, but even here it out-performed the other two methods. In addition, the modified MCMC method (based on a zero-inflated gamma-Poisson distribution) produced much more similar results to the

first MCMC method (based on a uni-modal gamma-Poisson distribution) than the bootstrap procedure for the analysis of equine FECRT data. This implies that the distributional assumptions made by the MCMC method has less practical impact on the analysis of these types of FECRT data than the assumption that bootstrapping a limited number of data points can capture all the variability of an inherently very variable system. Vidyashankar et al. (2007) propose dealing with this effect by taking into account the inter-farm variability. The intention of this paper was to assess the performance of each method when analysing individual datasets in the absence of any other comparable datasets, so that taking into account inter-farm variability would not have been possible. The MCMC method is also capable of analysing data from multiple sites, for example by defining a distribution of efficacy that describes the mean FEC reduction at each site and using this extra information to reduce uncertainty in the estimate for the true mean efficacy. However, by directly describing the variability structure in FEC data, parametric techniques eliminate the necessity for data from additional sites (where none is available), and allow efficacy to be analysed at an individual farm level.

Several of the datasets generated with parameters similar to observed equine FECRT data gave an empirical reduction of 100%, even where the true mean reductions were close to 75%. These datasets present difficulties when using both the WAAVP and bootstrap methods, which were unable to generate appropriate 95% confidence limits. Nineteen (19%) of these datasets were simulated using empirical reductions of less than 95%, and so represent a consistent source of false negatives for these methods. The MCMC method was the only method examined in this paper which is capable of analysing datasets with 100% empirical reductions in an appropriate fashion.

It is also apparent from the analysis presented here that analysis of a single equine FECRT dataset will often prove inconclusive. Using the MCMC method, only 10 of the datasets from equine field studies were classified as 'confirmed resistant' and 0 as 'confirmed susceptible', with the remaining 53 (84%) datasets containing insufficient information to be sure if the true drug efficacy was reduced or not. This is consistent with the conclusions made by Miller et al. (2006), that the results of a FECRT based on an arithmetic mean reduction can be inconsistent. The utility of the method could be increased by performing a suitable sample size calculation prior to performing the FECRT, and increasing the number of samples taken and/or reducing the egg detection threshold accordingly. This is probably not practical for routine clinical tests, however, due to the added cost and time associated with taking more samples and counting more eggs. A more useful solution might be to use a process control approach for routine surveillance, combined with the use of a more detailed FECRT with prior sample size calculations to calculate the required number of samples to take when the process control indicated a possible problem. This may represent both a more efficient use of resources, and a greater overall diagnostic test sensitivity and specificity, than the current use of repeated reduction tests viewed in isolation and without the necessary sample size calculations.

In this paper, the efficacy of reductions were classified according to the probability that the true reduction

was below a given threshold, which is not consistent with the method currently advocated by the WAAVP. This departure was made to allow a distinction to be drawn between cases where there is clear evidence for resistance and cases where there is insufficient evidence to demonstrate acceptable efficacy. Using the classification scheme currently used by the WAAVP, which involves consideration of the mean estimate and lower 95% confidence interval only (Coles et al., 1992), it is not possible to make this distinction in the absence of suitable power calculations. This limitation may lead to confusion over the clinical interpretation of FECRT analysis results.

The more flexible and intuitive output produced by the MCMC and bootstrap methods, including the ability to produce a single probability that the true reduction is less than a given value, make them both more attractive methods than the current WAAVP recommendation. It is evident that the MCMC method outperformed the bootstrap method in this study, however this may not be true when the data has a larger sample size or mean. Since the true distribution of data is unknown, the most conservative estimate would be to use the data at which the MCMC method performed worst. This produced notional 95% confidence intervals with a true estimated confidence of 93%. The bootstrapping procedure returned notional 95% confidence intervals with a true confidence greater than or equal to 93% only when the sample size was at least 40 with a pre-treatment mean FEC of 40 counted eggs or more, or with a sample size of at least 50 with a pre-treatment mean FEC of ten counted eggs or more. This suggests that the MCMC method should be used in preference to the bootstrap method with a sample size of less than 40 with a pre-treatment mean FEC of 40 counted eggs (equal to, for example, 1000EPG with an egg detection threshold of 25 EPG), or with a sample size of less than 50 with smaller pre-treatment mean FEC. The authors expect that similar results could be obtained using any computationally intensive parametric method such as parametric bootstrapping, likelihood profiling, or MCMC sampling from the likelihood without the use of prior information. With larger datasets, the data distribution independence and reduced computational effort associated with the non-parametric bootstrap procedure make this method more attractive.

5. Conclusions

Using data simulated with similar values of mean and sample size to those observed in equine FECRT data, both the method currently advocated by the WAAVP and a non-parametric bootstrap method failed to provide true 95% confidence intervals for the FEC reduction. In order to avoid making erroneous inference regarding the true efficacy of anthelmintics in the field, computationally intensive parametric methods such as MCMC should therefore be used with sample sizes of less than 50. The large proportion of inconclusive results returned from analysis of equine FECRT data suggests that the routine use of prior sample size calculations should be adopted to ensure sufficient data is collected. Software to perform all three types of analyses documented here is freely available in the form of an add-on package to the R statistical programming

language from http://cran.r-project.org/web/packages/bayescount/index.html.

6. Acknowledgements

This research was produced as part of the DEFRA-funded VTRI project 0101. BioSS is partly funded by the Scottish Government. The authors are grateful to Stig Petersen (Equilab Laboratory, Joerlunde Overdrev 7, DK-3500 Slangerup, Denmark) for providing the equine FECRT data discussed.

References

Coles, G. C., Bauer, C., Borgsteede, F. H., Geerts, S., Klei, T. R., Taylor, M. A., Waller, P. J., 1992. World Association for the Advancement of Veterinary Parasitology (W.A.A.V.P.) methods for the detection of anthelmintic resistance in nematodes of veterinary importance. Vet. Parasitol. 44 (1-2), 35–44.

Coles, G. C., Jackson, F., Pomroy, W. E., Prichard, R. K., von Samson-Himmelstjerna, G., Silvestre, A., Taylor, M. A., Vercruysse, J., 2006. The detection of anthelmintic resistance in nematodes of veterinary importance. Vet. Parasitol. 136 (3-4), 167–185.

Denwood, M., 2008. runjags: Run Bayesian MCMC Models in the BUGS syntax from Within R. R package version 0.9.2.
URL http://cran.r-project.org/web/packages/runjags/

Denwood, M. J., Stear, M. J., Matthews, L., Reid, S. W. J., Toft, N., Innocent, G. T., 2008. The distribution of the pathogenic nematode *Nematodirus battus* in lambs is zero-inflated. Parasitology 135 (10), 1225–1235.

Dopfer, D., Kerssens, C. M., Meijer, Y. G. M., Boersema, J. H., Eysker, M., 2004. Shedding consistency of strongyle-type eggs in Dutch boarding horses. Vet. Parasitol. 124 (3-4), 249–258.

Gelman, A., Rubin, D., 1992. Inference from Iterative Simulation using Multiple Sequences. Statistical Science 7, 457–511.

Gilks, W. R., Richardson, S., Spiegelhalter, D. J., 1998. Markov chain Monte Carlo in practice. Chapman and Hall, Boca Raton, Fla.
URL http://www.loc.gov/catdir/enhancements/fy0646/98033429-d.html

Kaplan, R. M., 2002. Anthelmintic resistance in nematodes of horses. Vet. Res. 33 (5), 491–507.

Klei, T. R., Chapman, M. R., 1999. Immunity in equine cyathostome infections. Vet. Parasitol. 85 (2-3), 123–133.

Miller, C. M., Waghorn, T. S., Leathwick, D. M., Gilmour, M. L., 2006. How repeatable is a faecal egg count reduction test? N. Z. Vet. J. 54 (6), 323–328.

Mooney, C. Z., Duval, R. D., 1993. Bootstrapping: a nonparametric approach to statistical inference. Sage Publications, 2455 Teller Road, Thousand Oaks, CA 91320.
URL http://www.sagepub.com/booksProdDesc.nav?prodId=Book3980&

Morrison, D. A., 2004. Technical variability and required sample size of helminth egg isolation procedures: revisited. Parasitol. Res. 94 (5), 361–366.

Nielsen, M. K., Haaning, N., Olsen, S. N., 2006. Strongyle egg shedding consistency in horses on farms using selective therapy in Denmark. Vet. Parasitol. 135 (3-4), 333–335.

Nødtvedt, A., Dohoo, I., Sanchez, J., Conboy, G., DesCôteaux, L., Keefe, G., Leslie, K., Campbell, J., 2002. The use of negative binomial modelling in a longitudinal study of gastrointestinal parasite burdens in Canadian dairy cows. Can. J. Vet. Res. 66 (4), 249–257.

Plummer, M., 2008. Just Another Gibbs Sampler (JAGS).
URL http://calvin.iarc.fr/~martyn/software/jags/

R Development Core Team, 2008. 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

Raftery, A. E., Lewis, S. M., 1995. The number of iterations, convergence diagnostics and generic Metropolis algorithms. In Practical Markov Chain Monte Carlo (W.R. Gilks, D.J. Spiegelhalter and S. Richardson, eds.). London, U.K.: Chapman and Hall.

Shaw, D. J., Grenfell, B. T., Dobson, A. P., 1998. Patterns of macroparasite aggregation in wildlife host populations. Parasitology 117 (Pt 6), 597–610.

Torgerson, P. R., Schnyder, M., Hertzberg, H., 2005. Detection of anthelmintic resistance: a comparison of mathematical techniques. Vet. Parasitol. 128 (3-4), 291–298.

Uhlinger, C. A., 1993. Uses of fecal egg count data in equine practice. Compendium On Continuing Education For the Practicing Veterinarian 15 (5), 742–749.

Vidyashankar, A. N., Kaplan, R. M., Chan, S., 2007. Statistical approach to measure the efficacy of anthelmintic treatment on horse farms. Parasitology 134 (Pt.14), 2027–2039.

Vose, D., 2004. ModelAssist for @Risk. Risk Thinking, Ltd.
URL http://www.vosesoftware.com/modelassist.htm

	Gamma-Poisson	**Multi-modal**	**Uniform-Poisson**
Bootstrapping	0.702	0.459	0.803
WAAVP	0.673	0.532	0.786
MCMC	0.746	0.555	0.803

Table 1: **Mean relative size of 95% confidence intervals for the true mean FEC reduction produced by each method from the analysis of 1000 datasets simulated using each distribution**

	Gamma-Poisson	**Multi-modal**	**Uniform-Poisson**
Bootstrapping	1.69	1.49	2.6
WAAVP	1.7	1.53	2.66
MCMC	1.61	1.46	1.77

Table 2: **Relative root-mean-square-error for median or mean estimate for the true mean FEC reduction produced by each method from the analysis of 1000 datasets simulated using each distribution**

		Bootstrap			
		Sus.	Poss. res.	Prob. res.	Res.
MCMC	Susceptible	0	0	0	0
	Possible resistant	10	8	0	0
	Probable resistant	4	9	22	0
	Resistant	0	0	7	3

Table 3: **The number of datasets assigned to each category of estimated efficacy status by MCMC and bootstrap analysis for 63 individual Danish equine datasets (median (range) of 9 (6-22) animals per dataset)**

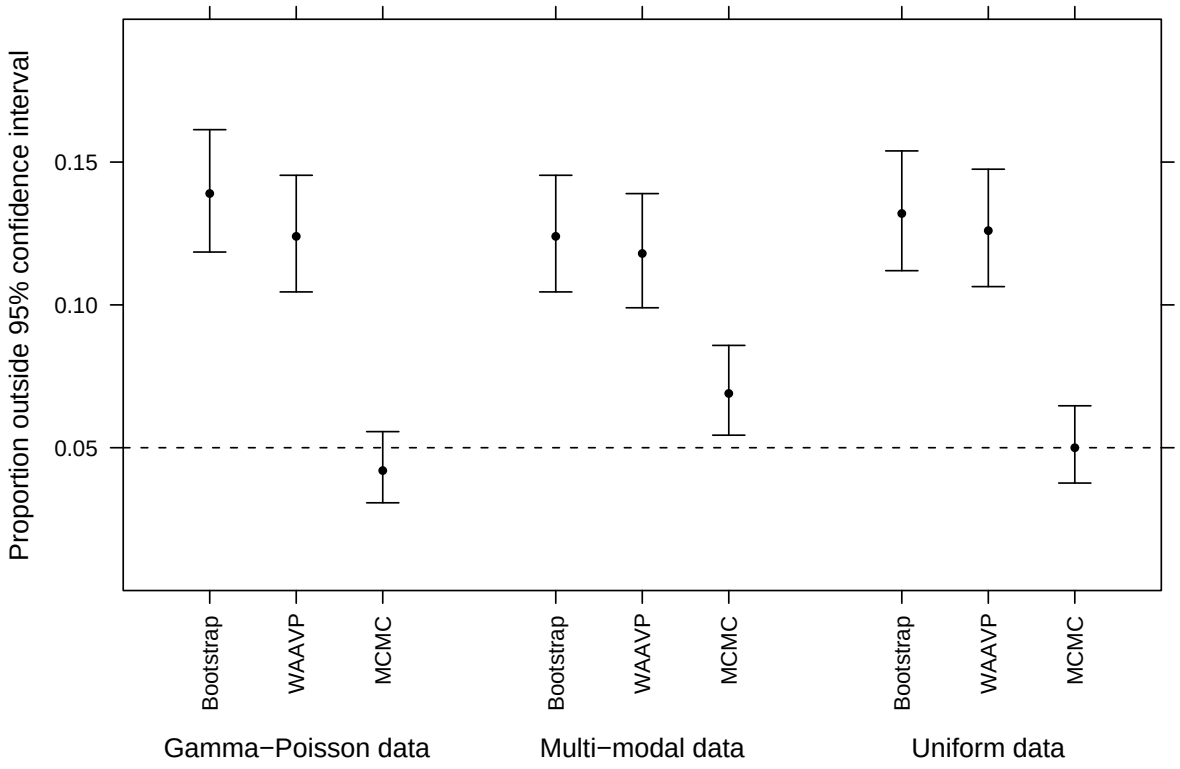

Figure 1: **The proportion of 95% confidence intervals not containing the simulated true mean FEC reduction parameter for each method from the analysis of 1000 datasets simulated using each distribution**

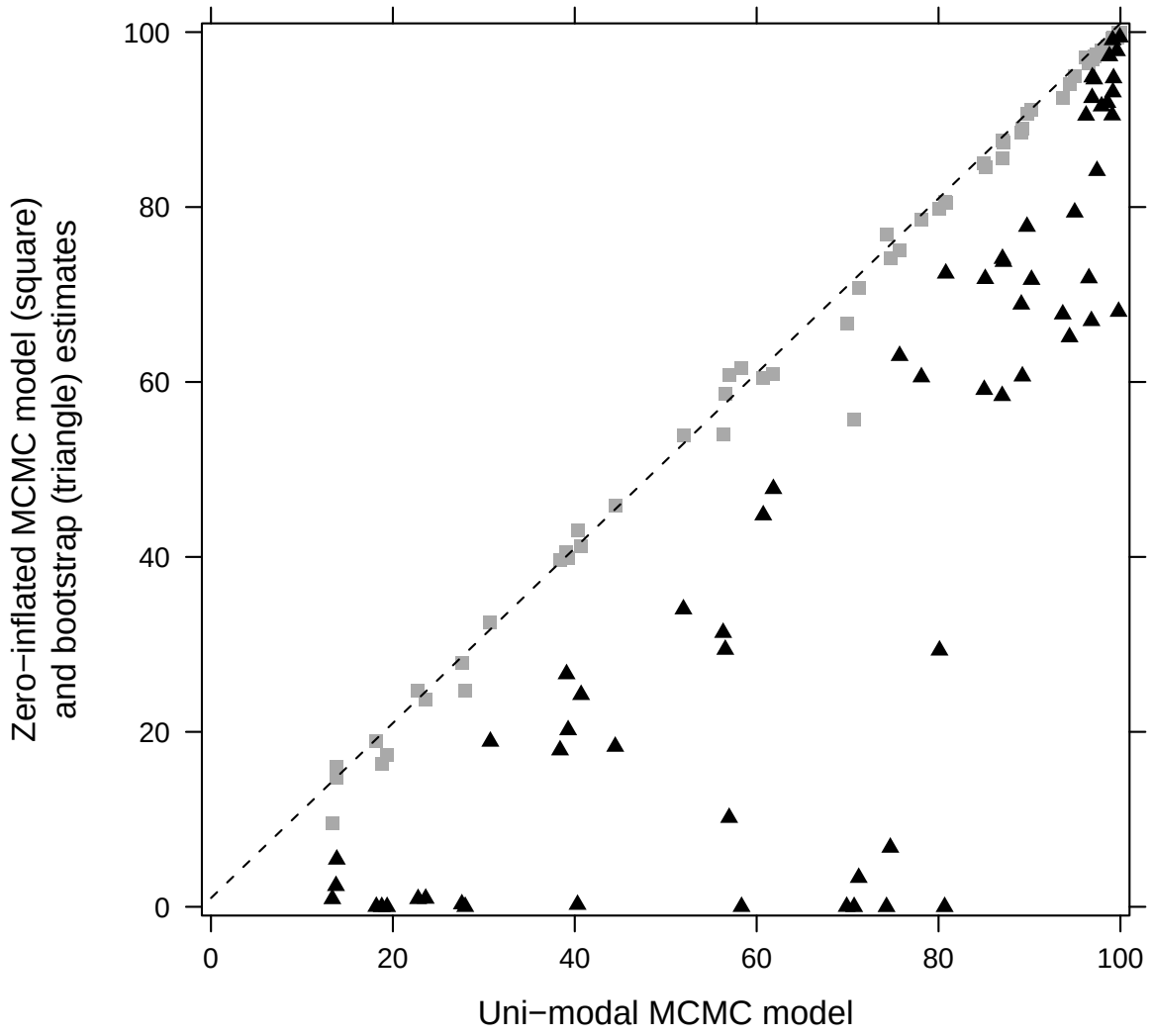

Figure 2: **Comparison of the estimated probability of efficacy $< 95\%$ returned for 63 individual Danish equine datasets by bootstrapping and an MCMC method based on a zero-inflated gamma-Poisson distribution, relative to an MCMC method based on a uni-modal gamma-Poisson distribution (median (range) of 9 (6-22) animals per dataset)**

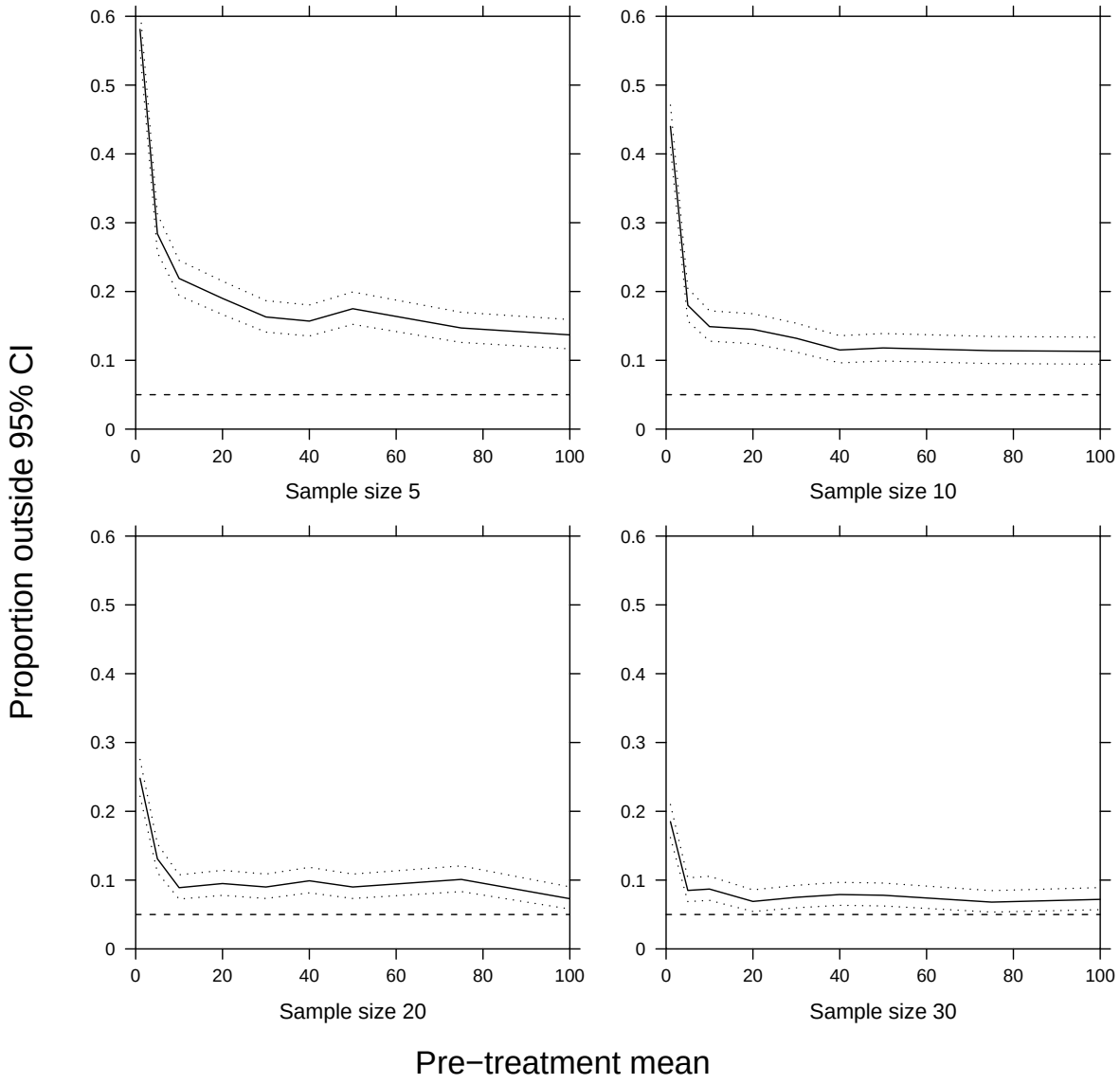

Figure 3: **Proportion of 95% confidence intervals produced using the bootstrap method that did not contain the true parameter from 1000 simulated datasets at each pe-treatment mean number of eggs counted (95% credible intervals in dotted lines). Sample sizes 5, 10, 20 and 30 shown.**

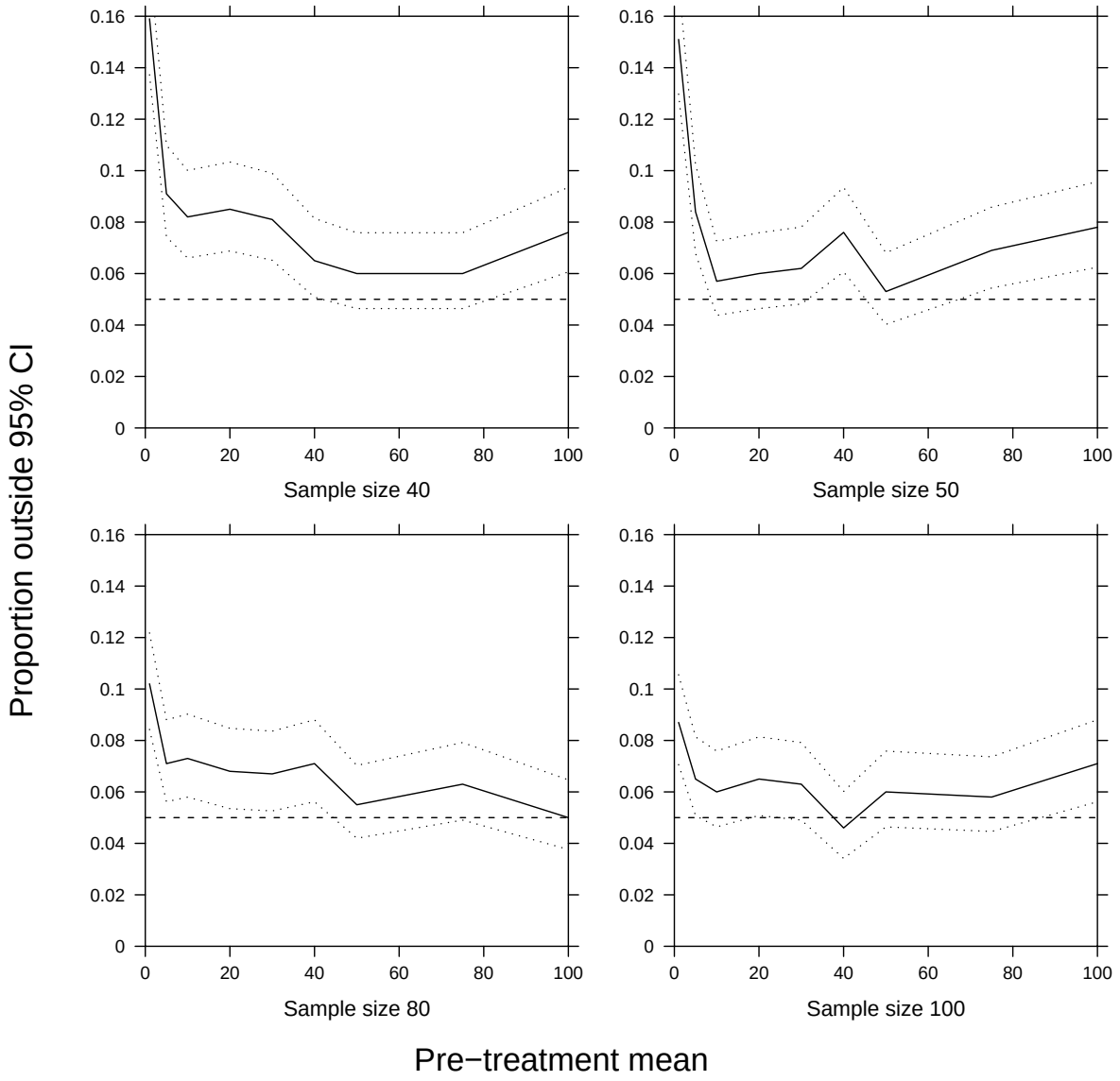

Figure 4: **Proportion of 95% confidence intervals produced using the bootstrap method that did not contain the true parameter from 1000 simulated datasets at each pe-treatment mean number of eggs counted (95% credible intervals in dotted lines). Sample sizes 40, 50, 80 and 100 shown.**