Referee A initial report

1. I agree that a Poisson model without an offset is implausible... although including the offset doesn't stop you from testing whether herd density (number of herds per unit area) has an effect. Put another way, if you don't include an offset the natural null hypothesis for the parameter in question is theta=1, not theta=0.

2. But a binomial model makes more sense...with or without herd density as a risk factor.

3. I don't know how reliable the standard asymptotics are, wrt parameter std errors or residual deviance, for this data-configuration. It might be worth running some simulations of the preferred model to check.

4. Torgesen'd analysis using glmmTMB is not strictly the same as a quasi-Poisson analysis using glm. The former fits a model with a Normal random effect in the linear predictor for which the unconditional variance and mean are not proportional, as assumed in the latter. Also, glmTMB appears to use a Laplace approximation to the likelihood, which may or may not be good enough. For the binomial model you still have the choice between quasi-binomial or Normal random effect in the linear predictor should you wish to avail yourself of it. I would be surprised if it made a big difference to the inference on the treatment effect (modulo comment 3).

5. You don't define ´triplet' in your note. It would be helpful to the reader to do so. Incidentally, why did Christl et al ignore the data from the reactive culling arm? Even if they weren't interested in its effect, it would have given more residual degrees of freedom, which might have helped with the asymptotics for the residual deviance.

Author response to Referee A

1 and 2: Glad you concur with using a binomial model.

The regions are all of similar area, so baseline herd numbers are correspond to actual herd density. When fitting a model to assess the treatment effect, we need to include the Triplet effect to get the best fitting model---once that's done herd density has no significant effect. I think that any more detailed study of the dependence of incidence on herd density is a bit beyond what I wanted to cover in this note, which is the significance or otherwise of the treatment effect.

3. I ran some simulations and both the parameter standard errors and the distribution of the residual deviance are right in line with the asymptotics (not surprisingly because of the binomial assumptions).

4. The natural way of allowing for over-dispersion in the binomial case is to use a betabinomial in glmmTMB. Interestingly, the maximum likelihood estimation of the dispersion parameter is essentially infinite...so this suggests we should use the binomial rather than the quasibinomial and not make any correction for overdispersion. I think I'd rather err on the conservative side and use the quasi...but will also mention the analysis using the betabinomial/binomial.

5. The main text talks about triplets which is why I didn't discuss it in the note itself. The reactive culling arm of the experiment was terminated at an earlier stage because of evidence that it made things worse. I'd prefer to stick to the data that everyone else has analysed....

Referee B report (standard type) with subsequent author responses (bold type)

From an infectious disease epidemiologist's viewpoint (i.e. my area of expertise), your analysis seems a sensible and valid approach. You analyse the number of breakdowns during the trial as binomially distributed with the number of herds as 'trials' and then consider the effect of overdispersion (extra-binomial variation) to adjust standard errors and confidence intervals and reduce the likelihood of Type 1 error. Assuming my interpretation is correct, this is equivalent to modelling the incidence risk i.e. the proportion of herds that breakdown over a given time-period, where the numerator is the number of incident breakdowns and the denominator is an estimate of the number of herds at risk over the three-year period of the trial (or more specifically the log odds of a herd breaking down over the time period). You also model the log odds ratio of new herd breakdowns compared to historical breakdowns by including historical breakdowns as an offset, which again seems a novel and valid approach.

The response variable in the Donnelly et al paper is also the number of breakdowns (logged), and this is assumed to be Poisson distributed and a function of four covariates: the log of the number of herds, the triplet, the log of the number of previous breakdowns and the intervention.

Torgerson et al question the validity of the Donnelly et al approach, arguing that the model is mis-specified, leading to Type 1 errors. If I understand correctly, their criticism focuses on the lack of consideration of overdispersion (extra Poisson variation) and the way the model incorporates the number of herds. They argue that the (log) number of herds at risk should be included as an offset, rather than an explanatory covariate. In their models, the response variable is in essence the scaled incidence of breakdown i.e. the number of herds that breakdown, scaled by the number of herds at risk, and the exponent of each coefficient is equivalent to the relative rate of breakdown, relative to the intercept. This seems broadly similar to your initial approach but with an overdispersed Poisson rather than an overdispersed binomial error distribution. As an aside, a strict herd-time-at-risk approach would assume that once a herd has broken down during the trial it is no longer at risk of being an incident breakdown and should be removed from the denominator. Proportional hazards and failure time models may be considered more appropriate here but given the co-authors on the Donnelly et al paper I'm sure this was considered and rejected, most likely due to uncertainty in the actual number of herds at risk in each year.

I have a few extra observations:

- Torgerson justifies the use of fixing the coefficient for the log of the number of herds as 1, and including this variable as an offset, by arguing this is more consistent with 'standard' epidemic theory whereby transmission is "dependent on density (i.e. contact rate)". This is an oversimplification in general terms, as transmission can be density or frequency-dependent, but most likely somewhere between the two. Here density dependence assumes the rate of contact increases with the density of the population, and the per capita force of infection increases with the density of infecteds (animals or herds). In contrast, frequency-dependent transmission assumes the contact rate is independent of the density and the force of infection increases with the prevalence of

infection in the population (see attached reference). If the areas of the triplet trial regions are the same, these are broadly equivalent of course and this may not be an issue for the RBCT study. However, it is also an oversimplification in the specific scenario of the RBCT study as bovine Tb is likely to be transmitted both directly from herd to herd, and indirectly through wildlife (i.e. the primary hypothesis under investigation in the triplet study). If bovine Tb is maintained in the badger population (i.e. with an R0>1) and cattle infections are just spillover events, it may not be reasonable to assume that the number of incident breakdowns is likely to be "linearly associated with the number of herds in the baseline". It will be more likely to scale with badger populations and badger infection.

I think that what this means is that the dependence of incidence on herd numbers should, be such that the coefficient fitted for log(herd numbers) in the Poisson model should be, if anything, greater than 1. This lends even more credence to Torgerson's approach as opposed to the original approach. We don't have any info on the badger population or infection, but because the triplets are reasonably geographically co-located, one would assume that this would to some extent be subsumed in the triplet effect. I did experiment with including herd numbers (density) in the quasi-binomial analysis. The relevant coefficient is not significantly different from zero, and the model including the variable has a higher AICc than one without. So I hope it's OK for the current purpose just to present the "Torgerson" analysis of assuming that infection is proportional to herd numbers.

- In my view, it's also worth considering the RBCT experiment in terms of causal inference and examining how this relates to model specification. The variable (log) historical number of breakdowns can be considered a measure of 'infection pressure' in the region prior to the onset of the trial and it seems reasonable to consider this as an explanatory covariate. However, this variable may also be correlated with the number of herds at risk in each area, as well as unmeasured covariates related to badger density and infection. This could potentially lead to confounding. This includes unmeasured confounding resulting from differences in badger population densities, badger infection prevalence and cattle herd sizes. This could affect the estimate of the coefficient associated with the number of herds in the Donnelly model. I note you fix the historical breakdown variable in the most parsimonious binomial model and model the log odds ratio between the observed and historical breakdown rates. I would be interested to see the comparison between the output of this model and the model with the number of historical breakdowns as an explanatory covariate?

This is model 1 in the original analysis. It makes very little difference. The AICc score is marginally higher. The coefficient of historical log odds is not significantly different from 1. The treatment effect is still significant at the 5% level but not the 2.5% level, and the confidence interval is very similar. I have put a sentence or two to this effect in the note.

- All models show a significant 'triplet effect' with marked variation between triplets even after the inclusion of covariates related to both the triplet and the response variable (i.e. historical breakdowns and number of herds). Presumably, the study was designed in this way as an RCT in order to control for confounding by unmeasured variables, including badger density and infection. Given the small sample size, randomisation of the intervention was unlikely to control for all residual confounding, hence the need to include some potential confounders, whilst being mindful of over-fitting. As herd

numbers vary markedly between triplet, the inclusion of triplet as a covariate may have subsumed some of variation attributable to herd numbers, again affecting the parameter estimate for the baseline number of herds in the Donnelly model.

See remarks made following the first bullet point.

- Torgerson et al conducted some post hoc analyses of the residuals. It may be worth doing something similar for the binomial model? This could help determine which model is likely to be least mis-specified, and the impact of both Type 1 and Type 2 errors.

Examining the residuals shows nothing untoward but with such a small sample I am a bit hesitant to take too much notice of any analysis of residuals.

I hope you find my comments helpful.

Extremely helpful. Given that this is intended only as a short note, I've decided that I'd rather leave things more or less as they are, though I have slightly simplified the presentation. I am really grateful for your comments and general support for the approach.

The author sent a revised version to the referees together with his responses and both referees wrote to say they were now content.