2017-12

A comparison of inferential analysis methods for multilevel studies: Implications for drawing conclusions in animal welfare science

Stevens, Kara

10.1016/j.applanim.2017.08.002
Applied Animal Behaviour Science
Elsevier

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

A comparison of inferential analysis methods for multilevel studies: Implications for drawing conclusions in animal welfare science

Kara N. Stevens[1,2], Lucy Asher[3], Kym Griffin[4,5], Mary Friel[4,6], Niamh O'Connell[4] and Lisa M. Collins[1,6]

[1] School of Life Sciences, University of Lincoln, Brayford Pool, Lincoln, LN6 7DL

[2] Medical Statistics, Plymouth University Peninsula Schools of Medicine and Dentistry, PL6 8BX

[3] Centre of Behaviour and Evolution, Institute of Neuroscience, Newcastle University, NE2 4HH

[4] Institute for Global Food Security, Queen's University Belfast, Belfast, BT9 5BN

[5] School of Animal, Rural and Environmental Sciences, Nottingham Trent University, NG25 0QF

[6] Faculty of Biological Sciences, University of Leeds, Leeds, LS2 9JT

Corresponding author: Lisa M. Collins

Address: Faculty of Biological Sciences, University of Leeds, Leeds, LS2 9JT

Telephone: 00 44 11 33 43 59 40

Email: L.Collins@leeds.ac.uk

Running head: Model selection for multilevel repeated measures

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

ABSTRACT

Investigations comparing the behaviour and welfare of animals in different environments have led to mixed and often conflicting results. These could arise from genuine differences in welfare, poor validity of indicators, low statistical power, publication bias, or inappropriate statistical analysis. Our aim was to investigate the effects of using four approaches for inferential analysis of datasets of varying size on model outcomes and potential conclusions. We considered aggression in 864 growing pigs over six weeks as measured by ear and body injury score and relationships with: less and more enriched environments, pig's relative weight, and sex. Pigs were housed in groups of 18 in one of four pens, replicating the experiment 12 times. We applied four inferential models that either used a summary statistic approach, or else fully or partially accounted for complexities in study design. We tested models using both the full dataset ($n = 864$) and also using small sample sizes ($n = 72$).

The most appropriate inferential model was a mixed effects, repeated measures model to compare ear and body score. Statistical models that did not account for the correlation between repeated measures and/or the random effects from replications and pens led to spurious associations between environmental factors and indicators of aggression, which were not supported by the initial exploratory analysis. For analyses on smaller datasets ($n = 72$), due to the effect size and number of independent factors, there was insufficient power to determine statistically significant associations.

Based on the mixed effects, repeated measures models, higher body injury scores were associated with more enrichment (coef. est. = 0.09, $p = 0.02$); weight (coef. est. = 0.05, $p < 0.001$); pen location on the right side (coef. est. = 0.08, $p = 0.03$) and at the front of the experimental room (coef. est. = 0.11, $p = 0.003$). By comparison, lower ear injury scores were associated with more enrichment (coef. est. = -0.51, $p = 0.005$) and pen location at the front of the experimental room (coef. est. = -0.4, $p = 0.02$). These observed differences support the hypothesis that injuries to the body and ears arise from different risk factors. Although calculation of the minimum required sample size prior to conducting an experiment and selection of the inferential analysis method will contribute to the

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

validity of the study results, conflict between the outcomes will require further investigation via different methods such as sensitivity and specificity analysis.

Word count: 400

Key Words: Study design, sample size, mixed effect models, pig, animal health, animal welfare.

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

1. INTRODUCTION

The statistician George Box stated “all models are wrong, but some are useful” (Box and Draper, 1987); which raises the question, how do we determine which statistical model, or in other terminology, inferential analysis method, is most appropriate? In recent years, a spotlight has been directed at the transparency of animal research methodology, with low rates of methodological reporting being associated with less scientific rigour and lower reproducibility (Vogt et al 2016, Ionnides et al 2009, Kilkenny et al 2009). Articles pertaining to animal research have been criticised in the past for their design, statistical analysis and reporting (McCance, 1995; Kilkenny et al., 2009; Sargeant et al., 2010). The publication of a list of guidelines for animal research known as the ARRIVE guidelines (Kilkenny et al., 2010), has helped to improve the quality of animal research (Gulin et al., 2015). These guidelines highlight the importance of choosing the appropriate experimental assessments, sample sizes and statistical inferential analysis methods. It is important to ensure the sample size is sufficient to test the study hypothesis, but also bearing in mind the ethical and financial implications of using an unnecessarily large sample size within an experiment. There is a plethora of techniques to produce sample size estimates, and the appropriate technique will depend on the inferential analysis used for a study. Sample size can often be quite difficult to calculate for more complex designs, though the importance of conducting these calculations accurately has been well communicated, particularly in clinical trials literature (Freiman et al., 1978; Biau et al., 2008).

Discussion in this area naturally leads into consideration of the methodology of the statistical analysis conducted on the collected data. Many of the papers focussing on the quality of research using animals have primarily targeted experimental design, animal numbers, and reporting, but have not discussed the appropriate analysis of what can often be complex datasets. Precise replication of a published study is rarely performed, and typically different studies will use different experimental designs and statistical inferential techniques to address the question. Although this can make comparisons between published studies difficult, agreement in the overall conclusions under such circumstances can be considered strong evidence for the named association, though more subtle or complex relationships may potentially be missed. An identified significant treatment effect across studies through use of meta-analysis, is typically considered to be robust evidence for an association, and also allows the magnitude of the effect size to be more precisely estimated than in single studies considered in isolation (Borenstein et al., 2009). However meta-analysis also has limitations, for example when few studies have been published in an area, when they differ substantially, or when the inferential analysis used is inappropriate for the design.

Within the field of animal welfare, many published results on a particular issue are mixed or conflicting, leading to somewhat mixed messages about what the most appropriate solution for an identified welfare hazard might be. To some extent, it is possible that this is at least partly due to publication bias (e.g. Hopewell et al., 2009; Brown et al., 2017) and the drive for novelty rather than further support for a set of hypotheses in published research. However, the lack of agreement between studies may be due to other factors – the differences may reflect genuine differences between the studies, arising for reasons as yet unmeasured or unaccounted for. They may be due to the use of indicators that have not been thoroughly validated in all respects for the species in question (Cronbach & Meehl, 1955). Finally, the observed lack of agreement may be due to inappropriate statistical analysis, leading to masking of true effects, or the discovery of false positives.

Even when two studies ask a very similar research question with largely similar methodology, mixed results can emerge. A typical example of this can be found in studies that investigate causes, and consequently solutions, for aggression in pigs. For example, Beattie et al. (1996) investigated whether an enrichment object or floor space had more influence on pig behaviour. Their analysis showed that duration of harmful behaviour was significantly higher in less enriched pens, and measured pig aggressive behaviours had no significant association with space allowance. By comparison, Turner et al. (2000) found that smaller space allowances were associated with more skin lesions and longer-lasting aggressive events. These studies were similar in a number of respects, except that Turner et al. (2000) regularly adjusted pen sizes to maintain a consistent stocking density (weight per m^2) throughout the experiment, whereas Beattie et al. (1996) maintained pen dimensions (hence stocking density would increase throughout the study). Consequently, the two studies are incomparable with conventional meta-analytic approaches. Variation in the indicators used could also potentially explain differences in model outcomes For example, different indicators of injuries in pigs result in differences in the final conclusion, even if the studies use otherwise similar experimental designs and methods for inferential analysis. In relation to the provision of straw for pigs, different indicators of aggression have lead to different conclusions; for example, Lahrmann et al. (2015) found reduced shoulder injuries for straw-housed pigs, whereas Morgan et al. (1998) found that straw-housed pigs performed more aggressive interactions and Statham et al. (2011) and Arey and Franklin (1995) have both reported no significant effect of the provision of straw on outbreaks of aggression. Aggression can, and indeed, has been described and measured using a wide variety of indicators. Examples of indicators for aggression are: duration of fights and number of bites (Andersen et al. (2000); prevalence of giving/ receiving belly nosing, mounting, ear and tail biting, and biting the pen bars, chains or other pen details (Brunberg et al. (2011); the ratio of aggressive events to social interactions (Drickamer et al., 1999); skin lesions on different body areas (Desire et al., 2016). Frequently, there is little or no overlap between studies, or construct validation to demonstrate that all indicators recorded measure what they are proposed to measure (e.g. tail biting has been considered an indicator of aggression; however this has been reconsidered in more recent years, e.g. Taylor et al., 2010).

Here we used a study investigating aggression in pigs to compare differences between two areas for the assessment of skin injuries (believed to be indicative of aggression in pigs), an ear score and a composite body score (Conte et al. 2012), and the effects of analysing the data via four inferential methods: (i) generalised linear models; (ii) repeated measures analysis; (iii) linear mixed effect models; and (iv) linear mixed effect models for repeated measures. We compare the significant associations between the two injury assessments and the covariates detected via the exploratory and four methods of inferential analysis. These four approaches were chosen because, to varying degrees, these models could account for some of the features of the data and model parameters could be directly interpreted.

Methods (i)-(iii) were considered sub-optimal relative to (iv), as these models were unable to account for correlation in the repeated measures, and /or random effects from the hierarchical structure in the data (pens within replication). We hypothesised that not accounting for random effects from the pens

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

within replication and correlation between repeated measures will either result in additional spurious relationships and/or mask possible significant relationships between our injury assessments and the covariates. By ignoring random effects, we hypothesise there will be more statistically significant associations with environmental factors, and by ignoring the repeated measurements, we hypothesise the association between injury score and time covariate will be more complex.

We investigated the effects of sample size within multilevel designs by analysing the data from different replications (n=18 pigs * 4 pens per replicate) as separate studies, and comparing the coefficient estimates from each of these analyses. A reduced sample size leads to a decrease in power, which means it is more difficult to identify the environmental factors associated with the injury scores. We hypothesize, that with a reduced sample size, there will be fewer statistically significant associations between injury scores and environmental factors.

2. METHODS AND MATERIALS

2.1 Animals and Housing

The study was conducted at the Agri-Food and Biosciences Institute, Hillsborough, County Down, Northern Ireland. The study used commercial crossbreed PIC 337 (Large White x Landrace) pigs. Pigs received a commercial weaner diet ad libitum and water was always available, according to the standard practices on the farm.

Each pig was weighed when they were four weeks and again at ten weeks old. The pigs' sex and weights at 4 weeks of age were used by the stockman to balance the groups to achieve a similar average weight and 50:50 sex ratio in each group of 18 individuals. Groups were then allocated at random to one of four pens. The pigs remained in these pens for a period of approximately six weeks, and the study was replicated twelve times, which led to a sample size of 864.

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

Pigs were assigned to one of four pens for the study that were contained within an experimental room situated in a long shed, which was divided into a series of similar rooms, with floor to ceiling solid walls between each room. Two types of pen environment were used within this study. Pens 1 and 3 were classed as more enriched environments; these pens were 2.18 m × 5.16 m in dimension with deep straw bedding (replenished weekly). Pens 2 and 4 were classed as less enriched environments, these were 2.18 m × 3.42 m in dimension, and no straw was provided. Both pens had floors constructed from concrete and were partially slatted, however in the more enriched pens (1 and 3) the slats were covered with plywood to prevent straw falling into the slurry system. In all pens, suspended wooden blocks were provided as standard enrichment.

Pens 1 and 2 were located on the left side of the experimental room and pens 3 and 4 were located on the right. The adjacent room on the right (next to pens 3 and 4) almost always contained weaner pigs, whereas the adjacent room on the left (next to pens 1 and 2) was frequently empty, or was occasionally used to house sows that could not enter farrowing crates. The difference in directional noise from each adjacent room was balanced in the experimental design by having one pen of each treatment type on both sides of the room. Two of the four pens were located next to the front of the room (pen 2 and pen 3), and the other two pens were located at the back next to an internal corridor.

The pigs were kept commercially, hence decisions relating to culling and health were made by the farm manager, as part of the standard on-farm procedures. Outbreaks of aggression leading to injury were observed only on video footage, analysed typically several weeks after recording took place. Animals that were observed to have high body scores were reported to farm staff, and monitored closely by farm staff and researchers for a period of 7 days after. No animals were culled for the purposes of this study, though as noted in section 2.3, a small number of animals (n=9 out of 862 pigs) died during the study period due to poor health or failure to thrive.

2.2 Assessment of Injury

An assessment of each individual's injuries was completed at three time points after entering the pens: (1) On day 4; (2) Between days 8 – 17; (3) Between days 29 and 39. At each assessment each pig was

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

scored on the following body areas: left and right ear; snout; left and right shoulder; front and back legs; left and right flank; left and right hindquarter; and back; using a six point scaling system, as defined in figure 1 (Conte et al. 2012). As part of the standard practice on the farm, 50% of the tail was docked within the first 24 hours after birth for every pig, this meant that tail score had limited value as an indicator for aggression.

2.2.1 Indicators of Aggression

Ear and body score were considered as indicators of aggression. At each assessment time point, the ear score was recorded as the higher observed injury score on either the left or right ear (possible score 0-5), and the body score was recorded as the sum score of the back, left and right shoulder, flank and hindquarters scores (possible score 0 – 25).

Due to the method used to construct the body score, based on the Conte et al (2012) scale, the two elements of frequency of injury and severity are confounded, especially for lower values. In our dataset, body score ranged between zero and 25, suggesting body score could be analysed as a continuous variable. A histogram plot of the log transformed body score implied we could assume the data followed a Gaussian distribution.

Each ear was scored on a scale between zero and five, with a score of zero signifying no injuries or damage, and a score of five indicating the presence of many deep red lesions. As very few pigs were identified with a score of 3 or more, categories 3 to 5 were combined, so that the ear score categories represented: 0 = no injuries; 1 = one small superficial lesion; 2 = more than one small, superficial lesion; or one red (ie deeper than score 1) superficial lesion; 3 = one or more deep lesions, or more than one red superficial lesions. Initial exploratory analysis suggested that the relationship between the housing conditions, sex and weight were similar for pigs with an ear score of 0 or 1. Therefore, these two groups were combined to simplify subsequent inferential analyses.

2.3 Statistical Analysis

As injury assessments were made at three irregularly spaced points in time, the assessments for an individual pig could be correlated, but the strength of the correlation may differ because of the variable time differences. Replicating the study 12 times may cause significant random effects for each pen within replication. The differences could be caused by the combination of pigs within a pen, or even associated with unmeasured external influences (e.g. weather conditions, handler behaviour, noise). Using weight at 4 and 10 weeks of age, we produced estimates of each individual's intermediate weights by fitting a linear model between the two time points. Although growth is usually statistically modelled by a curve, plots of the expected growth curves in Carr (1998) indicated that a linear estimate of pig weight would be an appropriate approximation over the short time scale used in this study.

We calculated individual relative weights in each pen within replication, in line with previous research indicating that an individual's relative size compared with its group mates is more important than its actual size (Nettle et al., 2013). Andersen et al. (2000) found no significant difference in number of bites between groups of pigs with low and high weight variability, which suggested removing pen differences would have no adverse effects. This is similar to comparing a pig's weight rank, but also accounts for variable weight differences between pigs.

Missing data were due to human error in data entry, and death or culling of the individual pig during the course of the study, either due to poor health or failure to thrive.

The plots and statistical analyses were produced using the statistical program R (Team, 2015) using the multgee (Touloumis, 2016), ordinal (Christensen, 2015), and lme4 (Bates et al., 2015) packages to produce the statistical models.

2.3.2 Exploratory Analysis

Before applying any statistical test or fitting a statistical model to data, it is important to perform appropriate exploratory analysis. Choosing the right method to explore the data will depend on the

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

question being addressed. As these data consisted of observations measured over time, we aimed to explore how body and ear score changed over time.

We plotted each pig's body score over time and fitted a Gaussian kernel smooth estimator to pigs within each category (i.e. by treatment enrichment level). A kernel estimator is a non-parametric method of fitting a line between two continuous variables. If there is uncertainty about the form of this relationship (i.e. linear, quadratic, etc.), visual inspection of plots of the data can provide insight into this. An appropriate bandwidth is determined, with bigger bandwidths creating smoother lines. We selected a bandwidth of 15, as injury assessments took place every 14 days on average (more details of kernel estimators can be found in Wand and Jones (1994)). As we were treating ear score as an ordinal variable, we looked at the proportional change of pigs within each category, and used the same methods as outlined above for body score.

2.3.3 Inferential Analysis

The data from this experiment possessed a hierarchical structure, where we had repeated measurements for each pig, within a pen, within a replication. There are various methods that can be applied to this type of data, depending on the assumptions one makes. We compared the results of four methods of analysis on body and ear score, where each method considered different aspects of the study design: (i) ignored the study design; (ii) considered correlation in the repeated measurements; (iii) considered random effects from the hierarchical structure; (iv) considered the correlation structure and the random effects. Table 1 provides a comparison of the different inferential methods considered in this paper. Depending on the study design, it indicates which inferential method would be appropriate for different types of data.

(i) Ignoring study design (without accounting for repeated measures or hierarchical structure)

To demonstrate the effects of ignoring the study design completely, i.e. not accounting for repeated measures of individuals and random effects, we fitted a generalised linear model (GLM) to body and

DOI : https://doi.org/10.1016/j.applanim.2017.08.002

ear score. Specifically a log linear model (LLM) was fitted to body score and a cumulative logistic regression model (CLM) was fitted to ear score.

(ii) Repeated measures (without accounting for hierarchical structure)

As we assumed body score is continuous, we performed a multivariate analysis of covariance (MANCOVA) with a Gaussian distribution. This methodology compares the means of all the different possible groups and determines whether a significant difference is present when accounting for a possible time-dependent correlation between the assessments. We accounted for the replications within this inferential analysis using an error structure for individuals within replications. MANCOVA assumes that the assessments measured are taken at equally spaced points in time, and the difference in time is the same for each individual. Only individuals with complete data are included.

As ear score is an ordinal variable, we fitted a cumulative logistic regression model for repeated measures. To account for repeated measurements of the ear score, the parameters were estimated via generalized estimating equations (GEE), which allow for the presence of a possible time-dependent correlation between ear score assessments made at different times. However, a covariate for the replication was also included to account for the possible differences between replications.

(iii) Hierarchical structure (without accounting for repeated measures)

To remove the effect of the repeated measures we produced a summary variable for each pig. The summary variable for body score was simply the mean of the log transformed body score across each of the three repeated measures. The summary variable for ear score was slightly more complicated. Often categorical variables are summarised by their median or modal value. However, as the median and mode are not influenced by extreme values, it meant that severe injuries were missed. Therefore, we summed the ear score for each replication, then combined some of the categories according to the frequency and level of injury the category represented to bring the score in line with the original scoring system. The new ear score categories were 0 = less than 2 occurrences of superficial lesions,

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

or 1 occurrence of a deep lesion; 1 = 1 occurrence of a deep lesion and 1 occurrence of a superficial lesion or 3 occurrences of superficial lesion; 2 = more than 1 occurrence of a deep lesion.

To account for the random effects of pen within replication we fitted a mixed effects linear regression model (LME) to the mean log body score

$$y_{i,j} = \alpha + X_{i,j}\boldsymbol{\beta} + Z_{i,j}\boldsymbol{\delta}_i,$$

Equation 1

and a cumulative logistic mixed effects regression model (CLME) to the re-categorized sum of ear score

$$\text{logit}\big(\Pr[Y_{i,j} < k]\big) = \alpha_k + X_{i,j}\boldsymbol{\beta} + Z_{i,j}\boldsymbol{\delta}_i,$$

Equation 2

where: $y_{i,j}$ is the mean log body score; $Y_{i,j}$ is the ear score category for k=0,1,2; α is the intercept whereas α_k is the intercept for the k^{th} cumulative logit; $\boldsymbol{\beta}$ is a vector of fixed effects coefficient estimates; $X_{i,j}$ are the fixed covariates design vector for the j^{th} pig, in the i^{th} replication $\boldsymbol{\delta}_i$is a vector of the random effects for replication i; and $Z_{i,j}$ is a design vector of the random effects.

An important difference between the GLM and a mixed effects model comes from the estimation of the variance. In a GLM only the variance of the individual pigs is required, whereas now an estimate for the variance for the individual pigs and the replications is required.

(iv) Hierarchical data with repeated measures

To account for both the hierarchical design and repeated measurements within this study, we fitted the log linear and cumulative logistic, mixed effects model as defined in eEquation 3Equation 4:

$$\log\big(y_{i,j,t}\big) = \alpha + X_{i,j,t}\boldsymbol{\beta} + Z_{i,j,t}\boldsymbol{\delta}_i,$$

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

Equation 3

$$\mathrm{logit}\big(\Pr[Y_{i,j,t} < k]\big) = \alpha_k + X_{i,j,t}\boldsymbol{\beta} + Z_{i,j,t}\boldsymbol{\delta}_i.$$

Equation 4

These are very similar to Equation 1Equation 2, and in fact, the mathematical representation only requires the addition of a subscript *t* to denote the time element in the random effects model. See Twisk (2012) for more details on this type of analysis.

Computationally, as we are treating body score as a continuous Gaussian distributed variable, estimation of the coefficients and the variance for the replications and individuals in Equation 3 can be accomplished via GEE. However, there is no software available currently which can produce a mixed effects cumulative logistic regression model with repeated measures where the correlation between each observation depends on the time difference between repeated measures.). We concluded that as we only had three repeated observations, estimation of the random effects was more important than using GEE to account for a time dependent correlation structure for ear score. However, a random effect term for each pig was included instead, as it assumes the correlation between observations is constant over time.

Small Sample Sizes

To investigate the effects of small sample sizes, a repeated measures model was fitted to the data of each replication. This led to 12 statistical models, one for each replication, which each consisted of 72 pigs per model/replication (18 pigs assigned to 1 of 4 pens), with a maximum of three skin lesion assessments each, giving a total of number of observations of 216 per model. Each GLM consisted of the same covariates, which were equivalent to the covariates in the final hierarchical repeated measures model.

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

3. RESULTS

For 862 individual pigs we had a measurement for at least one of the injury assessments. For body score there were two pigs with missing data for the first observation, seven pigs with missing data for the second observation and nine pigs with missing data for the third observation. For ear score there were three pigs with missing data for the first observation, seven pigs with missing data for the second observation and 10 pigs with missing data for the third observation.

3.1 Body Score

3.1.1 Exploratory Analysis

The plots of the kernel smooth estimators in figure 2 a) – e) depict a cubic relationship with time. The kernel estimators of log body score are between 1 and 2 at the first examination (day 0), with a decline in log body score by the second examination (days 8-17), but by the third examination (days 29-39) there is an increase. All covariate groups mirror this pattern.

However, the slopes for each replication varied, as shown in figure 2 a), thus implying a random slope for replication over time was required. Figure 2 b) of the Gaussian kernel smooth estimators for each pen was used to determine whether different housing features were worth investigating. It is clear that pigs within pen 3 tended to have a higher body score than any of the other three pens, which all appeared to be quite similar. There was a difference between the intercept and a slight difference between the slopes for each pen.

The plots in figure 2 c) to e) further identify differences between the pens. Comparing the score of the different environments in figure 2 c), the difference between the less and more enriched environments is only evident after approximately 14 days. This implies an interaction between time and environment. The plot in figure 2 d) shows that pigs in the pens to the front of the experimental room had a consistently higher body score than pigs in the pens located at the back. We also observed that pigs in pens on the right side of the room had a higher body score than those in pens on the left side of the room, as shown in figure 2 e).

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

The plot in figure 2 f) is a scatter plot of body score by standardised relative weight. The blue line is the kernel smooth estimator using a bandwidth of 0.75. Less than 3% of the standardised weight values were either > 2 or < -2, which meant there were insufficient values to produce a reliable estimate of the relationship between body score and relative weight. However, the plot suggested that for a relative weight between -2 and 2, the relationship was linear and as weight increased so did log body score.

3.1.2 Inferential Analysis

Table 2 contains all the summary statistics for the fixed effects (coefficient estimate, standard error, Student's t-value and p-value) for the most appropriate model, (iv) LLME + GEE, and the p-values for all fixed effects for the three comparison methods, (i) LLM, (ii) MANCOVA and (iii) LLME. If a p-value was greater than 0.05 it was not included in the table. In all the statistical models the enrichment level, location of the pen (left/right side, front/back of the experimental room) was significantly associated with body score. Relative weight was a significant component in 3 out of the 4 statistical models.

The LLME + GEE model accounted for a random intercept and slopes over time for pens within replications, and a Gaussian correlation structure between observations for each pig. There was a significant cubic relationship with time, this can also be seen in figure 2 (a)-(e) of the kernel estimators. The significant relative weight coefficient implied that a unit increase in relative weight resulted in a 0.05 increase in log body score, which equates to a 5% increase in body score. On average, pigs on the right side of the room had a 0.094 higher log body score, i.e. their body score was 9.9% higher than those on the left side of the room. Also pigs with more enrichment and those in pens located at the front of the experimental room had higher log body scores by 0.124 (13.2% increase in body score) and 0.09 (9.4% increase in body score), respectively.

3.1.3 Small Sample Sizes

Figure 3 a) is a box plot of the coefficient estimate when using GEE to analyse each replication; when the random effect for replication was not included, with the fixed effect coefficient estimates under LLME + GEE model (table 2) included as a red cross. The box plot for relative weight was the only one where the whiskers of the plot did not include zero, implying this was the only covariate with a significant association with log body score for all but one replicate. This suggested that the coefficient estimate for relative weight should remain fairly consistent across replications. For pen location (left/ right, front/back of the experimental room), and more enriched pens, the coefficient estimates showed greater variance.

The median coefficient estimates were: weight = 0.04; right side of experimental room = 0.1; location to the front = 0.14; and more enriched environment = 0.11. Comparing these values with the coefficients estimates of the LLME + GEE model in table 2 we see that these values are quite similar, and encouraging as a form of sensitivity analysis. Within one replication, there are 216 observations. If we were to perform a t-test on these 216 observations to detect the largest effect size of 0.14 in log body score, assuming the standard deviation was 0.6 (estimated from the entire dataset), then we would have ≈40% power to detect this difference. This does not account for the repeated measures, which would reduce the power further.

3.2 Ear Score

3.2.1 Exploratory Analysis

From figure 4 there is evidence of a cubic relationship between ear score and time when comparing the proportion of pigs with an ear score of 0 with 1 and/or 2 (all plots on the left), where there is a decrease, plateau, then further decrease. However, the plots comparing the proportions observed in 0 and/or 1 with 2 (plots on the right) appear to be exponentially decaying.

The plots in figure 4 show the proportional change in the pigs observed within each ear score group with Gaussian kernel estimators to convey how the relationship between ear score changes over time

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

for different housing features. In figure 4 a) the variability in the shape of the relationship between ear score and time for the different replications indicate a different slope for each replication over time is required. However, in figure 4 b) the estimators for each pen have a similar shape, but different intercepts. There are clear differences in figures 4 c) and d) between environment and location next to the front or the back of the experimental room.

3.2.2 Inferential Analysis

Table 3 shows all the summary statistics for fixed effects (coefficient estimate, standard error, Student's t-value and p-value) for the cumulative logistic mixed effects regression model with random effect for pigs, (iv) CLME +1, and significant p-values for fixed effects from the three comparator methods (i) CLM, (ii) GEE and (iii) CLME. Within each statistical model, ear score was shown to have a significant association with the level of enrichment and the front/back pen location.

The CLME+1 model included random intercept and slope terms for pen within replication to account for the differences between replications over time, and a random intercept for each pig to account for the correlation between repeated measures. To discuss our findings, we use odds ratios (i.e. exponential transformation of the coefficients), so we can quantify the percentage increase or decrease in odds that will result in the increase or decrease in ear injury score. In the CLME +1 model, pigs in more enriched pens had 40% lower odds (Confidence Interval, CI: 14%, 58%) of having a higher ear score compared to pigs in less enriched pens. Similarly, pigs in a pen located at the front of the room had 33% lower odds (CI: 5%, 53%) of having a higher ear score.

3.2.3 Small sample sizes

We fitted a CLME model to each replication with a random intercept for each individual. Figure 3 b) contains the box plot of the coefficient estimates from the ordinal logistic regression of ear score for each replication. The fixed effect coefficient estimates under CLME+1 (table 3) are included as a red cross in figure 3 b). There was a wide range of values for the coefficients from each replication (median coefficient estimate for more enriched environment = -0.55; front of experimental room = -

0.21). Comparing the coefficient estimates for CLME and CLME+1, there was little difference between pen enrichment estimates (0.04), but a larger difference between pen location estimates (0.19).

3.3 Inference method comparisons

For both types of injury score, the key associations between the injury score and environmental factors were statistically significant across all four statistical models. Although, the magnitude of the relationship and the direction was not always the same between the most appropriate statistical model from approach (iv), and the other three statistical models, using methods (i) to (iii). The model via approach (iii) for both injury scores provided no insight into changes in injury over time, as this information was removed when summarising the injury scores.

Table 2 details the level of association between body score and the environmental factors for each inferential method. Approach (i), the LLM, did not account for the repeated measure correlation or random effects, and there was an additional significant association between body score and tail injury. Whereas for approach (ii), the MANCOVA, which only accounted for repeated measurements, there was a significant association between body score and sex. Neither of these associations were evident in the exploratory analysis or in the most appropriate approach (iv). However, the association between body score and weight was not statistically significant in approach (iii), the LLME model, but the evidence from exploratory analysis and most appropriate model indicated there was a relationship between these two variables.

In table 3 the statistical models from methods (i), CLM, and (ii), GEE, did not account for the random effects of pen within replication that led to high order degree polynomials with the day, 7 and 5 respectively. There was no evidence in the exploratory analysis or the final most appropriate model (CLME + 1), that this type of association between ear score and time was valid.

4. DISCUSSION

Comparing models where each incorporated different aspects of the study design demonstrated how important using the most appropriate inferential analysis is when producing valid results. By appropriately accounting for all sources of variation within the multilevel structure of the data (i.e. pens within replications) and considering the potential time-dependent correlation between observations, we increased the likelihood of identifying the true associations between the covariates and injury scores. We also found that there was a strong agreement between exploratory and inferential analysis, and associations seemed to be plausible.

In the most appropriate model for the data (repeated measures, mixed model), the strong significant association of ear and body injury score with the non-linear time component is suggestive of a complex relationship between behaviour and time. This observation was only possible because of the repeated observations within pigs, and further validated by the replications of the study. Although the variation in the inter-assessment interval time increased the statistical difficulty of the analysis, it did mean that there was more information available about changes in injury score over a wider range of interval differences. Ear and body injury score were both associated with the enrichment level and front location of pen within the experimental room, although the direction of this association changed for both covariates between injury scores. More enriched pens (coef. est. = -0.51, $p = 0.005$) and pens at the front of the experimental room (coef. est. = -0.4, $p = 0.02$) were both associated with a reduction in ear score, whereas those in more enriched pens (coef. est. = 0.09, $p = 0.02$), and pens at the front of the experimental room (coef. est. = 0.11, $p = 0.003$) had a higher body score. Body score was also associated with weight and pen location on the right side of the experimental room, such that as weight increased so did body score (coef. est. = 0.05, $p < 0.001$), and those pigs in pens on the right side of the experimental room also had a higher body score (coef. est. = 0.08, $p = 0.03$).

In this study, we investigated the impact of fitting statistical models that account for none, some and all of the known structural features of a multilevel dataset. We also analysed the effect of small sample size upon the most appropriate model. Similar investigations comparing inferential analyses have been conducted in human and non-human medical literature (Hu et al., 1998; Wang and Goonewardene, 2004), though this is the first example to the authors' knowledge in animal welfare.

In using an analytical approach that did not match the study design (approach (i): CLM), variance within the dataset that was associated with either the hierarchical structure or the correlational structure between repeated observations was not accounted for. This approach (CLM) led to predictions of a complicated relationship between ear injury score and time, with a 7-degree polynomial predicted to describe the relationship. For body score, the CLM predicted a cubic (i.e. 3-degree polynomial) relationship with time, just as was predicted by the most appropriate model (CLME+1). The high degree polynomial relationships predicted here result from poor estimation of variance, due to the models attempting to explain variation in the data using only the covariates, without the underlying hierarchical structure accounted for.

Including the correlation of the repeated measurements for approach (ii) via MANCOVA for body score and GEE for ear score did increase the p-values, but it did not account for the substantial variation caused by the random effects. Hence, there was an additional relationship between body score and sex, and the association between ear score and day was now a 5-degree polynomial. One substantial drawback back with MANCOVA is the strict format required of the data, i.e. equally spaced repeated measures with no missing values. Using GEE analysis is more flexible and the observations do not necessarily have to be equally spaced. However as the correlation coefficients between repeated measurements of ear score were all less than 0.3, and the differences between the estimators for replications and pens from the plots in figure 3 a) and b) appeared quite high, this suggested the random effects terms for replication and pen were more important than accounting for the correlation structure between repeated measurements. By replicating the study, we were able to gain insight into differences between pens, which we had not considered for inclusion in our experimental design prior to conducting the study; in particular, this would have been beneficial for the location of the pens within the experimental room. Although we accounted for differences in noise level with left/right side counter-balancing of the treatments, and accounted for potential differences between pens at the front (near the door) versus at the back of the room with front/back counter-balancing of treatments, we did not rotate the pens, which would have allowed us to account for the additional locational differences detected in the data. Although we were unable to fully explain the reason for differences between pen locations within the experimental room, we were able to identify that pen location was a source of variation and we could therefore statistically remove any undue influence this was having on other covariates within the model. Differences observed between replications could be related to weather conditions, handlers and many other features not measured as part of this study. Despite being unable to quantify all variation between replications, we believe that replication on other farm sites would help to build up a more general picture across contexts.

Summary measures of both body and ear score were used in approach (iii), which resulted in lost information about the nature of the relationships of body and ear score across time. Using this approach, we were unable to identify a significant association between body score and weight via the LLME model, but we detected a significant relationship between ear score and weight using the CLME, as compared to the final appropriate model.

In the final approach (iv) for body score and ear score, there was evidence of a cubic relationship with time for both injury scores. However, the direction of the coefficient estimates for day, day^2 and day^3 differed between body and ear injury scores. For body scores, the coefficients for time were positive for day and day^2 and negative for day^3, whereas for ear score they were negative for day and day^3 and positive for day^2. This result implies that the underlying behaviour indicated by proxy from these injury scores changed over time. For example, the initial decline in scores could be associated with pigs becoming acquainted with one another as a hierarchy within a pen was established within the first week (Barnett et al., 1994; Arey, 1999).

In both the final ear score and body score statistical models there was a significant association with pen location (front/back of the room) and enrichment level (see section 3.2.2). Pigs in pens located at the front of the room had lower odds of having a higher ear score (table 3), but higher odds of a higher body score (table 2). Pigs in more enriched pens had lower ear scores (as described in section 3.2.2,

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

table 3). This result supports previous findings that aggressive events are reduced in larger pen sizes (Fraser et al., 1991; Turner et al., 2000). Whereas the LME + GEE model for body score implies that more enriched pens resulted in higher body injury scores.

Finding clear differences in the predictors for ear and body scores lends support to the hypothesis that they have different underlying causes. Injuries to the ear are mainly received during aggressive interactions (McGlone, 1985). Injuries to the body on the other hand, whilst accrued through aggression, can also be the result of increased activity and play (Munsterhjelm et al. 2009; Camerlink et al., 2013). Unfortunately, as tails were docked at birth we were not able to use tail injury as another comparator, although research suggests that the majority of tail injuries reflect exploratory motivation rather than aggression (Taylor et al., 2010). Applying a similar study to undocked pigs may provide further detailed insight into aggression and the underlying motivating behaviours that lead to injuries. Statistical techniques used to determine the validity in medical screening tests, such as a receiver operator curve (ROC) analysis (Fawcett, 2006) or Bland-Altman test (Bland & Altman, 1986), may be used to compare indicators of aggression to determine if they are a measure of the same quantity.

Whilst the final model selected is appropriate for the experimental design, it is not perfect. There are currently no developed statistical methods available to analyse categorical outcome variables with a time dependent correlation structure between repeated measures within a hierarchical model (such as the random effects of replications within pens described within section 2.1). As such, we could not account for both the correlational structure and hierarchy of the study design within current statistical methodology. One possible solution could be to develop a statistical model with a probit link rather than a logit link, as the probit link is associated with the Gaussian distribution, and it may be easier to define a time dependent correlation structure with this compared to the logit link. However, the interpretation of the probit link can be difficult as there are no direct interpretations of the coefficients, instead it is necessary to refer to the marginal effects of the regressors (see Liao (1994) for more details), and the estimation of the coefficients would be computationally intensive.

Differences between the results of the four inferential methods highlight the importance of initial exploratory analysis in determining whether resulting significant associations are realistic, particularly as all four methods used are technically appropriate, albeit with varying degrees of fit to the experimental design. Strong evidence of a relationship in the exploratory analysis should translate to a significant association observed within the inferential analysis. Although measures were taken into account for layout of the experimental room, it was not possible to completely account for the extent of this effect, and it was through exploratory analysis that we were provided with greater insight into the magnitude and nature of the effect.

By analysing each replication separately, we were able to demonstrate how sample size affects the final coefficient estimates. The decrease in data resulted in insufficient power to detect significant associations, although the calculated medians of almost all the replications' coefficient estimates were consistent with our full final models. The results clearly demonstrate that analysis of small sample sizes may lead investigators to believe there was no association between the indicators for aggression and covariates, whereas it could be the study is under-powered to detect the effect size (i.e. the conclusion would be a type 2 error). As a simple demonstration, we performed a power calculation to detect a mean difference in body score of 0.18 and standard deviation of 0.6, based on summary statistics of enrichment level in the fifth week. The power calculation found that to detect such a difference with 80% power at the 5% level of significance, a sample size of 176 pigs (total 352) assigned to each enrichment level was required.

This study demonstrates through examples, how the type of indicator measured, the sample size and choice of statistical analysis can affect model outputs and conclusions drawn. We also highlight the importance of using an appropriate indicator to reflect the behaviour under investigation. The correct inferential analysis is important for meaningful results, which are not only plausible, but also supported by the exploratory analysis. To ensure the quality of animal science reports it is vital that a study consists of an appropriate sample size, with statistical analysis appropriate for the study design. These findings provide further support for the ARRIVE guidelines, but we feel that additional steps may improve the quality of research by ensuring studies are designed based upon the inferential analysis best equipped to answer the research question. It may be valuable to consider following similar procedures as in medical trials with the formulation of a protocol and detailed documentation of any unexpected and additionally planned deviations, which may subsequently affect the inferential analysis. This way, while best laid plans may still go awry in practice, there will be a clear plan to ensure that robust and appropriate analysis of the data can still be conducted.

ACKNOWLEDGEMENTS

The authors would like to thank AFBI for use of experimental room and care of the animals.

ETHICS STATEMENT

All procedures described were approved by the University of Lincoln's Ethics Committee on 8/9/2015, code COSREC62. This research was conducted at the Agri-Food and Biosciences Institute, Northern Ireland and conformed to the Association for the Study of Animal Behaviour's guidelines on the use of animals in research: http://asab.nottingham.ac.uk/ethics/guidelines.php.

FUNDING STATEMENT

This work was supported by the BBSRC (grants BB/K002554/1 and BB/K002554/2). MF was supported by a Department for Employment and Learning Northern Ireland studentship and Queen’s University Belfast.

REFERENCES

Andersen, I.L., Andenæs, H., Bøe, K.E., Jensen, P., Bakken, M., 2000. The effects of weight asymmetry and resource distribution on aggression in groups of unacquainted pigs. Appl. Anim. Behav. Sci. 68, 107-120.

Arey, D.S., 1999. Time course for the formation and disruption of social organisation in group-housed sows. Appl. Anim. Behav. Sci. 62, 199-207.

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

Arey, D.S., Franklin, M.F., 1995. Effects of straw and unfamiliarity on fighting between newly mixed growing pigs. Appl. Anim. Behav. Sci. 45, 23-30.

Barnett, J.L., Cronin, G.M., McCallum, T.H., Newman, E.A., 1994. Effects of food and time of day on aggression when grouping unfamiliar adult pigs. Appl. Anim. Behav. Sci. 39, 339-347.

Bates, D., Maechler, M., Bolker, B., Walker, S., 2015. lme4: Linear mixed-effects models using Eigen and S4, http://CRAN.R-project.org/package=lme4.

Beattie, V.E., Walker, N., Sneddon, I.A., 1996. An investigation of the effect of environmental enrichment and space allowance on the behaviour and production of growing pigs. Appl. Anim. Behav. Sci. 48, 151-158.

Biau, D.J., Kernéis, S., Porcher, R., 2008. Statistics in Brief: The Importance of Sample Size in the Planning and Interpretation of Medical Research. Clinical Orthopaedics and Related Research 466, 2282-2288.

Bland, J. M. & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. Lancet i, 301-310.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to meta-analysis. Chichester: Wiley.

Box, G.E., Draper, N.R., 1987. Empirical model-building and response surfaces. Wiley New York.

Brown, A.W., Mehta, T.S., Allison, D.B., 2017. Publication bias in science: What is it, why is it problematic, and how can it be addressed? In: *The Oxford Handbook of the Science of Science Communication*. Ed: K. Hall Jamieson, D. Kahan, D.A. Scheufele. Oxford University Press, New York, USA.

Brunberg, E., Wallenbeck, A., Keeling, L.J., 2011. Tail biting in fattening pigs: Associations between frequency of tail biting and other abnormal behaviours. Appl. Anim. Behav. Sci. 133, 18-25.

Camerlink, I., Turner, S. P., Bijma, P., Bolhuis, J. E. (2013). Indirect genetic effects and housing conditions in relation to aggressive behaviour in pigs. PloS one.;8:e65136.

Carr, J., 1998. Garth Pig Stockmanship Standards. 5m Publishing.

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

Christensen, R.H.B., 2015. ordinal: Regression Models for Ordinal Data, http://www.cran.r-project.org/package=ordinal.

Conte, S., P. G. Lawlor, N. O'Connell, and L. A. Boyle. 2012. Effect of split marketing on the welfare, performance, and carcass traits of finishing pigs1. J. Anim. Sci. 90:373-380.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. Pyschological Buletin. 52: 281-302.

Desire, S., Turner, S.P., D'Eath, R.B., Doeschl-Wilson, A.B., Lewis, C.R.G., Roehe, R., 2016. Prediction of reduction in aggressive behaviour of growing pigs using skin lesion traits as selection criteria. Animal 10, 1243-1253.

Drickamer, L.C., Arthur, R.D., Rosenthal, T.L., 1999. Predictors of social dominance and aggression in gilts. Appl. Anim. Behav. Sci. 63, 121-129.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters 27: 861-874.

Fraser, D., Phillips, P.A., Thompson, B.K., Tennessen, T., 1991. Effect of straw on the behaviour of growing pigs. Appl Anim Behav Sci 30, 307-318.

Freiman , J.A., Chalmers , T.C., Smith , H.J., Kuebler , R.R., 1978. The Importance of Beta, the Type II Error and Sample Size in the Design and Interpretation of the Randomized Control Trial. New England Journal of Medicine 299, 690-694.

Gulin, J., Rocco, D., Garca-Bournissen, F., 2015. Quality of Reporting and Adherence to ARRIVE Guidelines in Animal Studies for Chagas Disease Preclinical Drug Research: A Systematic Review. PLoS Negl. Trop. Dis. 9, 1-17.

Hopewell, S., Loudon, K., Clarke, M. J., Oxman, A. D., Dickersin, K. (2009). Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database of Systematic Reviews*, Issue 1. Art. No.: MR000006.

Hu, F.B., Goldberg, J., Hedeker, D., Flay, B.R., Pentz, M.A., 1998. Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. American Journal of Epidemiology 147, 694-703.

Ioannidis, J. P., Allison, D. B., Ball, C. A., Coulibaly I., Cui X., Culhane, A. C., Falchi, M., Furlanello, C., Game, L., Jurman, G., Mangion, J., Mehta, T., Nitzberg, M., Page, G. P., Petretto, E., van Noort, V. (2009). Repeatability of published microarray gene expression analyses. Nature Genetics. 41:149–155

Kilkenny, C., Browne, W.J., Cuthill, I.C., Emerson, M., Altman, D.G., 2010. Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research. PLoS Biol 8, 1-5.

Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M.F.W., Cuthill, I.C., Fry, D., Hutton, J., Altman, D.G., 2009. Survey of the Quality of Experimental Design, Statistical Analysis and Reporting of Research Using Animals. PLoS ONE 4, 1-11.

Lahrmann, H.P., Oxholm, L.C., Steinmetz, H., Nielsen, M.B.F., D'Eath, R.B., 2015. The effect of long or chopped straw on pig behaviour. Animal 9, 862-870.

Liao, T.F., 1994. Interpreting probability models: Logit, probit, and other generalized linear models. Sage.

McCance, I., 1995. Assessment of statistical procedures used in papers in the Australian Veterinary Journal. Aus. Vet. J. 72, 322-329.

McGlone, J.J., 1985. A Quantitative Ethogram of Aggressive and Submissive Behaviors in Recently Regrouped Pigs. J. Anim. Sci. 61, 556-566.

Morgan, C.A., Deans, L.A., Lawrence, A.B., Nielsen, B.L., 1998. The effects of straw bedding on the feeding and social behaviour of growing pigs fed by means of single-space feeders. Appl. Anim. Behav. Sci. 58, 23-33.

Munsterhjelm, C., Peltoniemi, O. A., Heinonen, M., Halli, O,. Karhapaa, M., et al. (2009). Experience of moderate ¨a bedding affects behaviour of growing pigs. Appl Anim Behav Sci. 118:42–53.

Nettle, D., Monaghan, P., Boner, W., Gillespie, R., Bateson, M., 2013. Bottom of the Heap: Having Heavier Competitors Accelerates Early-Life Telomere Loss in the European Starling, *Sturnus vulgaris*. PLoS ONE 8, e83617.

Sargeant, J.M., Thompson, A., Valcour, J., Elgie, R., Saint-Onge, J., Marcynuk, P., Snedeker, K., 2010. Quality of Reporting of Clinical Trials of Dogs and Cats and Associations with Treatment Effects. J. Vet. Intern. Med. 24, 44-50.

Statham, P., Green, L., Mendl, M., 2011. A longitudinal study of the effects of providing straw at different stages of life on tail-biting and other behaviour in commercially housed pigs. Appl. Anim. Behav. Sci. 134, 100-108.

Taylor, N.R., Main, D.C.J., Mendl, M., Edwards, S.A., 2010. Tail-biting A new perspective. Veterinary Journal 186, 137-147.

Team, R.C., 2015. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.

Touloumis, A., 2016. multgee: GEE Solver for Correlated Nominal or Ordinal Multinomial Responses, http://www.cran.r-project.org/package=multgee.

Turner, S.P., Ewen, M., Rooke, J.A., Edwards, S.A., 2000. The effect of space allowance on performance, aggression and immune competence of growing pigs housed on straw deep-litter at different group sizes. Livest. Prod. Sci 66, 47-55.

Twisk, J.W.R., 2012. Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide. 2 ed. Cambridge University Press, Cambridge.

Vogt L, Reichlin TS, Nathues C, Würbel H (2016) Authorization of Animal Experiments Is Based on Confidence Rather than Evidence of Scientific Rigor. PLoS Biol 14(12): e2000598.

Wand, M.P., Jones, M.C., 1994. Kernel Smoothing. Chapman & Hall/CRC.

Wang, Z., Goonewardene, L.A., 2004. The use of MIXED models in the analysis of animal experiments with repeated measures data. Canadian Journal of Animal Science 84, 1-11.

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

FIGURE LEGENDS

Figure 1: The six-point scaling system used to assess injuries to pig's body areas and outline of body areas for injury scoring; Ears, Snout, Shoulders, Legs, Back, Flanks, Hind quarters and Tail.

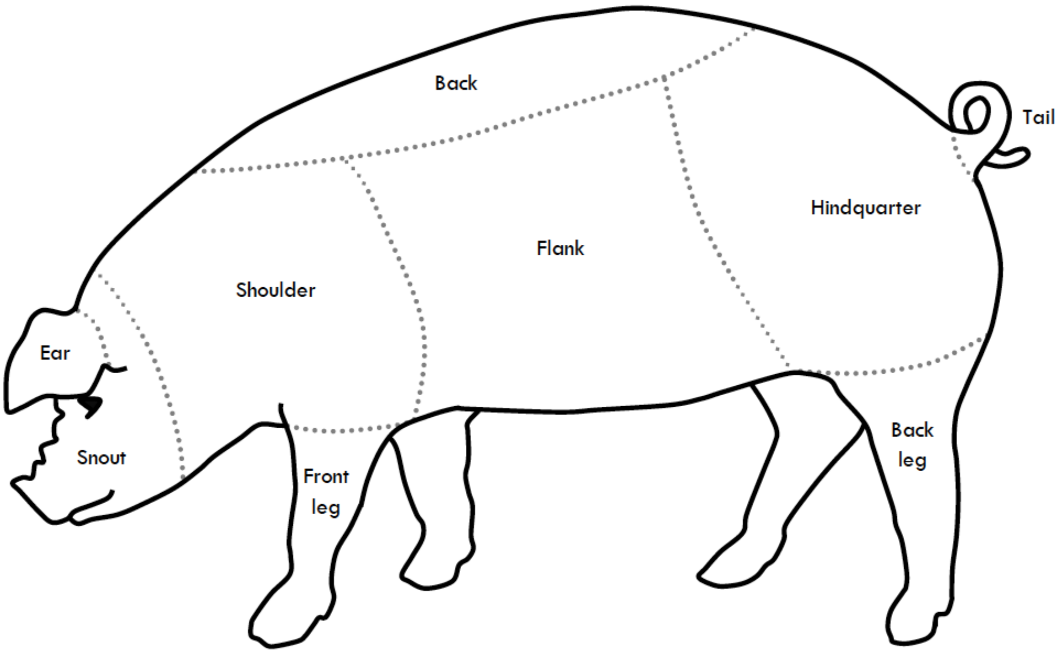
Figure 2: Plots of the log transformed body score by day with a Gaussian kernel smooth estimator with a bandwidth of 15 for a) replication; b) pen; c) enrichment; d) location to the front or back of the experimental room; e) location on either side of the experimental room. The light grey area depicts the time period the second injury assessments were gathered, all points gathered after this period are the third injury assessments and all points before are the first; f) Plot of the pig's relative weight for each pen within replication by log body score with a Gaussian kernel smooth estimator with bandwidth of 4. The grey area of the plot indicates the region where 95% of the data is located, and where the kernel estimator will be most reliable.

Figure 3: a) Box plot of the fixed effect coefficient estimates for the log linear regression model for body score for each replication. The red crosses represent the fixed effect coefficient estimates for the LLME + GEE from table 2. b) Box plot of the fixed coefficient estimates from the ordinal logistic regression of ear score for each replication. The red crosses represent the fixed effect coefficient estimates for the CLME +1 in table 3.cross.

Figure 4: Left plots: observed proportion with an ear score of 0 and 1/2. Right plots: observed proportion with an ear score of 0/1 and 2, with Gaussian kernel estimators with a bandwidth of 15 for a) replications; b) pens; c) enrichment; or d) location to the front or the back of the experimental room. The light grey area depicts the time period the second injury assessments were gathered, all injury assessments gathered after this period are the third injury assessments and all injury assessments before are the first.

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

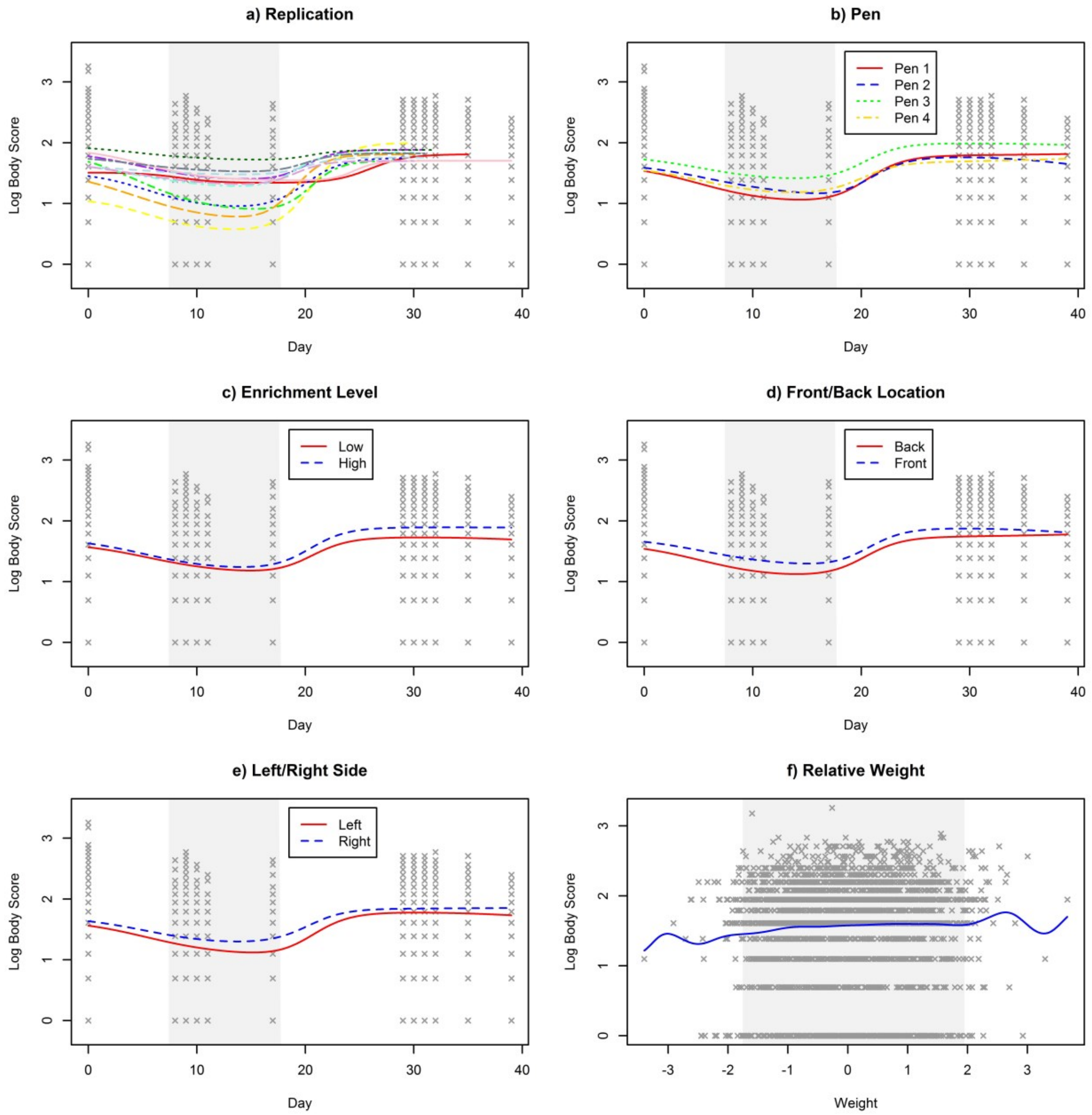
Figure 1



Score	Scaling System
0	No injuries.
1	One small superficial lesion.
2	More than one small, superficial lesion; or just one red (deeper than score 1) but still superficial lesion.
3	One or several big and deep lesions. If deep, only one single lesion. If not so deep, several red lesions.
4	One very big, deep and red lesion. Or many deep, red lesions.
5	Many, very big, deep and red lesions covering the skin area.

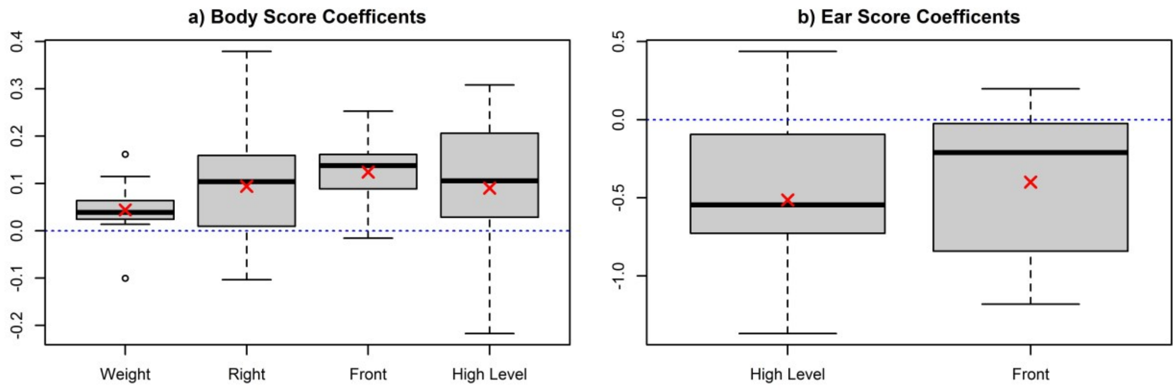
This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

Figure 2



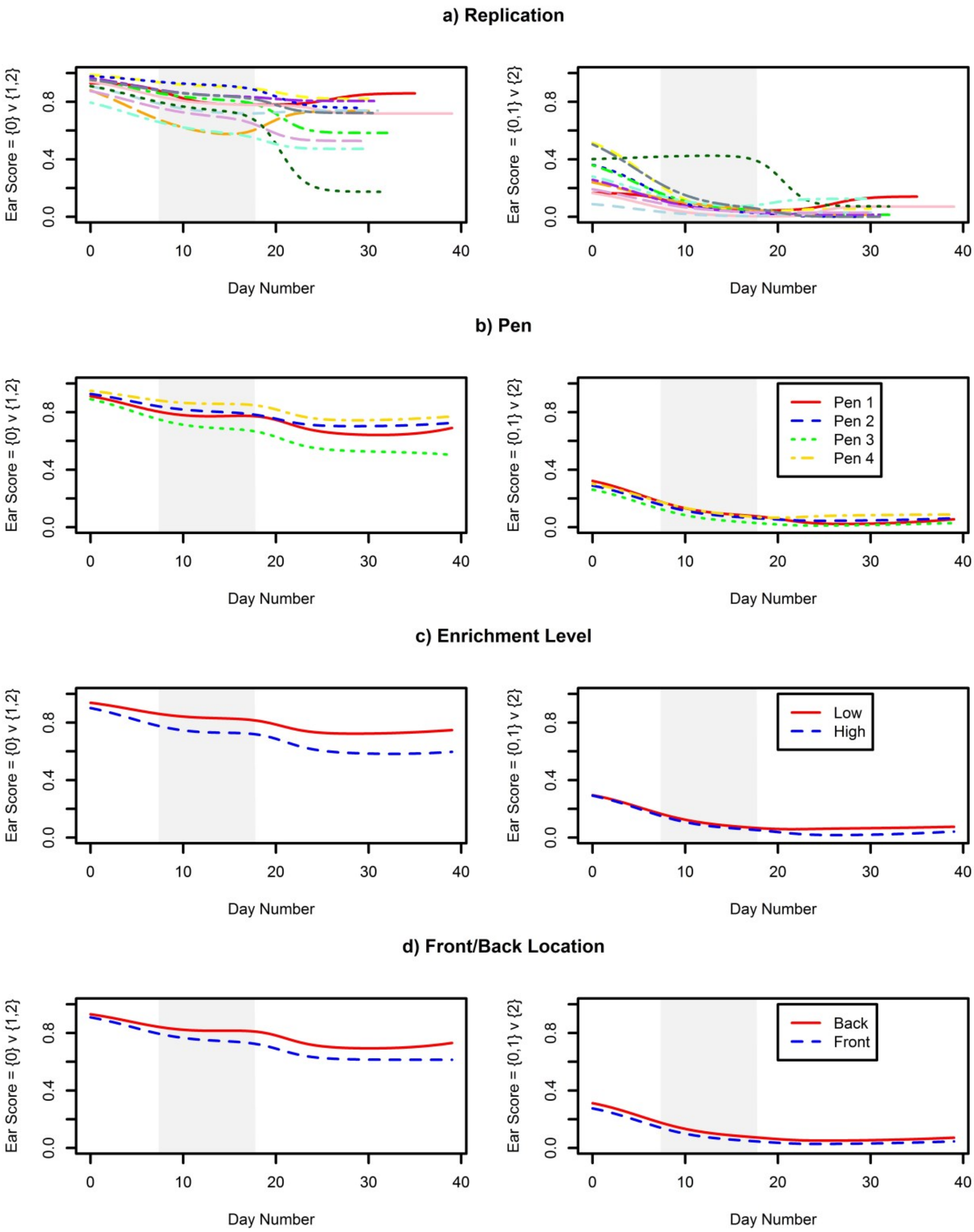
This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

Figure 3



This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

740 **Figure 4**



741

742

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

Table 1

Data	**Inferential Method**				
	MANCOVA	*GLM*	*LME*	*GEE*	*LME + GEE*
Univariate		C O			
Multivariate	C				
Repeated				C O	
Hierarchical			C O		
Repeated + Hierarchical					C

Table 1: Types of data that can be analysed using different inference methods, where C represents continuous data and O represents ordinal data. MANCOVA=Multivariate Analysis of Covariance; GLM=Generalised linear model; LME=Linear mixed effects model; GEE=General Estimating Equation model.

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

Table 2

	LLME + GEE				**LLME**	**MANCOVA**	**LLM**
				n	***n***		
Pigs				862	862	855	862
Body Score				2565	862	2550	2556
	β	***SE***	***t***	***p***	***p***		
Day	5.87	2.47	2.38	0.0173			< 0.0001
Day²	11.45	2.35	4.87	< 0.0001			< 0.0001
Day³	-6.39	1.30	-4.93	< 0.0001			< 0.0001
More Enriched	0.09	0.04	2.40	0.0224	0.0151	0.0003	0.0003
Location: Right	0.08	0.04	2.26	0.0307	0.0109	0.0018	< 0.0001
Sex						0.0041	
Weight	0.05	0.01	3.41	0.0007		0.0278	0.0013
Location: Front	0.11	0.04	3.16	0.0034	0.0011	0.0003	< 0.0001

Table 2: Summary statistics for inferential analysis of Body Score via the: log linear mixed effects model for repeated measures (LLME + GEE); linear mixed effects model of pig's mean log body score (LME); multivariate analysis of covariance (MANCOVA) of log body score, and a log linear regression model (LLM). Where: ***n*** is the number of pigs/body score assessment; ***β*** is the parameter estimate; ***SE*** is the standard error; ***t*** is the Student's t test statistic and ***p*** is the probability value associated with each covariate. **Day** is the day within the trial that observations were recorded; **More Enriched** refers to pens that had more enrichment (compared with Less Enriched); **Location: Right** refers to pens on the right side of the room (compared to pens on the left side of the room); **Location: Front** refers to pens at the front of the room (compared to pens at the back of the room).

This is an accepted authors draft of a paper published in Applied Animal Behaviour Science
Online publication complete: 6-SEP-2017
DOI : https://doi.org/10.1016/j.applanim.2017.08.002

Table 3

	CLME + 1				**CLME**	**GEE**	**CLM**
				n		***n***	
Pigs				862	862	862	862
Ear Score				2572	862	2572	2572
	β	*SE*	*t*	*p*		*p*	
Day	-51.68	5.75	-8.99	< 0.0001		< 0.0001	< 0.0001
Day^2	31.30	5.74	5.45	< 0.0001		< 0.0001	< 0.0001
Day^3	-13.56	6.51	-2.08	< 0.0369		0.0453	0.0003
Day^4						< 0.0001	< 0.0001
Day^5						0.0194	< 0.0001
Day^6							0.0255
Day^7							< 0.0001
More Enriched	-0.51	0.18	-2.79	0.0053	0.0131	< 0.0001	< 0.0001
Weight					0.0302		
Location: Front	-0.40	0.18	-2.25	0.0247	0.0328	< 0.0001	< 0.0001

Table 3: Summary statistics for inferential analysis of Ear Score via the: cumulative logistic mixed effects model with rep, pen and pig random effects (CLME + 1); cumulative logistic mixed effects model with rep and pen random effects for summary ear score (CLME); cumulative logistic regression model for repeated measures (GEE); the cumulative logistic regression model (CLM). Where: ***n*** is the number of pigs/ear score assessment; ***β*** is the parameter estimate; ***SE*** is the standard error; ***t*** is the Student's t test statistic and ***p*** is the probability value associated with each covariate. **Day** is the day within the trial that observations were recorded; **More Enriched** refers to pens that had more enrichment (compared with Less Enriched); **Location: Front** refers to pens at the front of the room (compared to pens at the back of the room).