



**Random effect selection in generalised linear models
a practical application to slaughterhouse surveillance data in Denmark**

Denwood, Matt; Houe, Hans; Forkman, Björn; Nielsen, Søren Saxmose

Publication date:
2015

Document version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Denwood, M., Houe, H., Forkman, B., & Nielsen, S. S. (2015). *Random effect selection in generalised linear models: a practical application to slaughterhouse surveillance data in Denmark*. Abstract from Annual Meeting of the Society of Veterinary Epidemiology and Preventive medicine, Ghent, Belgium.

RANDOM EFFECT SELECTION IN GENERALISED LINEAR MODELS: A
PRACTICAL APPLICATION TO SLAUGHTERHOUSE SURVEILLANCE DATA IN
DENMARK

M.J. DENWOOD^{*}, H. HOUE, B. FORKMAN AND S.S. NIELSEN

SUMMARY

We analysed abattoir recordings of meat inspection codes with possible relevance to on-farm animal welfare in cattle. Random effects logistic regression models were used to describe individual-level data obtained from 461,406 cattle slaughtered in Denmark. Our results demonstrate that the largest variance partition was at farm level for most codes, but there was substantial variation in reporting for some meat inspection codes between abattoirs. There was also substantial agreement for the relative under or over-reporting of different slaughter codes within individual abattoirs. This indicates that the sensitivity of routine surveillance in Denmark is affected by differences in the working practices between abattoirs, resulting in biased prevalence estimates. Therefore, it is essential to correct for the variation in reporting between abattoirs before meaningful inference can be made from prevalence estimates based on data derived from meat inspection.

INTRODUCTION

All carcasses from food producing animals in the European Union (EU) are subject to meat inspection according to EU legislation (Anon, 2004), for the primary purpose of ensuring the safety of the food. However, the idea of using this substantial resource of meat inspection data for purposes other than food safety has gained substantial traction over recent years (Dupuy et al., 2013; Harley et al., 2012; Klauke et al., 2013; Knage-Rasmussen et al., 2014; Vial and Reist, 2014). In Denmark, there has been a considerable emphasis on the possibility of using such data for monitoring on-farm animal health and welfare using related meat inspection codes as a proxy for compromised health or welfare on farm (Knage-Rasmussen et al., 2014; Nielsen et al., 2014). The high availability and low cost of meat inspection data are potentially very attractive, but concerns remain about the validity of using such data outside the original remit of food safety. For example, Nielsen et al. (unpublished) demonstrated the relatively poor correlation between the prevalence of some pig health conditions in meat inspection codes and findings from more detailed pathological examination of the same groups of animals. There is also the possibility of variation in reporting level between abattoirs due to differences in abattoir practices and the design/setup of meat inspection lines, which may complicate the comparison of disease prevalence reported from different abattoirs. Before comparing the prevalence of any meat inspection codes between farms, it is therefore essential to account for any difference in the apparent prevalence between farms that may be caused by differences in recording between abattoirs.

^{*} Matthew James Denwood, Department of Large Animal Sciences, University of Copenhagen, Grønnegårdsvej 8, 1870 Frederiksberg C, Denmark. Email: md@sund.ku.dk

The most common way of accounting for this type of clustering within epidemiological data is to use one or more random effect terms to describe the structural patterns in the data. These have been used for linear models as well as generalised linear models, including logistic regression models (Guo and Zhao, 2000; Li et al., 2011). These random effects are often used to account for known structural relationships that would otherwise break the independence assumptions of the response variable, and have also been shown to be useful to account for over-dispersion at the level of observation (Harrison, 2014). In these cases, the 'significance' of these terms is not of interest, and it is neither necessary nor desirable to consider removing the random effects in an attempt to improve the model fit. However, there are some situations such as that presented here, where the presence or absence of the random effects themselves is of interest, in which case some formal test of the significance of these random effects terms is desirable. In addition, it is very common to undertake some type of model selection process to eliminate candidate fixed effect variables from mixed effect models. In both of these cases, the validity of several commonly used methods of model selection is challenged by the presence of the random effects terms.

The most common method of comparing nested models is using some penalised measure of model fit such as the likelihood ratio test (LRT), variants of Akaike's information criterion (AIC), or the Bayesian information criterion (BIC). Each of these methods compares the maximum likelihood estimate of two or more nested models, with a correction for the relative complexity of the models. This penalty factor is derived from the relative number of degrees of freedom between the models, and is necessary to account for the natural ability of more complex models to produce a superior fit to any given dataset simply by chance. The number of degrees of freedom is relatively simple to calculate for purely fixed effects models, but for random effects and mixed models there is no straightforward method of determining the effective degrees of freedom contributed by the random effect. Standard methods such as the deviance information criterion (DIC) may be used within a Bayesian context (Spiegelhalter et al., 2002), but not in the standard frequentist maximum likelihood framework. Alternatively, a more computationally intensive approach can be taken to model comparison, using data simulated under the null model and fitted to both the null and alternative models. This allows the direct calculation of a distribution of expected likelihood ratio statistics, to which the observed likelihood ratio can be compared. The two advantages of this method are that no assumptions are made about the distribution of the likelihood ratio statistic, and there is no requirement to calculate the unknown quantity describing the number of degrees of freedom. This method has not to our knowledge been widely used within epidemiology, but a similar method has been recommended in the context of mixture models (McLachlan, 1987).

In this paper, we demonstrate the use of simulated data to provide a robust method of model selection for models including random effects. This method is applied to a random-effects logistic regression model describing the observed prevalence of various meat inspection codes, selected on the basis of animal welfare relevance and non-negligible prevalence, in Danish cattle. The primary interest from these data is inference regarding the relative contributions of variance between farms (interpreted as true difference in prevalence of these diseases), compared to variance between abattoirs (interpreted as differences in the sensitivity and specificity of disease recording between abattoirs).

MATERIALS AND METHODS

Meat inspection codes

Cattle meat inspection data were obtained for the entirety of 2012, from all eight cattle abattoirs that slaughtered >10,000 head of cattle in Denmark during 2012. All recorded meat inspection codes in accordance with Danish legislation (Anon, 2011) were available, with an individual animal level recording of presence or absence of each of these codes. The data were then divided into separate datasets describing animals under 18 months of age (n=212,826; denoted 'calves') vs. older cattle (n=248,580; denoted 'adults'), before being aggregated by abattoir (n=8), and farm of origin (n=7,020 for calves; n=10,721 for adults). Slaughter date information was disregarded, but the same time period (all of 2012) was collected for each abattoir. The number of animals slaughtered during the year and recorded frequency of each meat inspection code was obtained for each combination of abattoir of slaughter and farm of origin (this observation level grouping is referred to here as the 'group'). Some meat inspection codes (and combinations) were excluded from modelling analysis on the basis of the following criteria:

- Slaughter plant codes (mostly used for decision making at the abattoir), for example contaminated hide
- Codes possibly related to transport of animals to the slaughterhouse (on the basis that these may not have occurred on farm)
- Acute conditions, which could have occurred during transport (based on assessment by a professor of veterinary pathology in the Department of Veterinary Disease Biology, University of Copenhagen)
- Central nervous system conditions (on the basis that they are relatively non-specific and difficult to assess at the abattoir)
- Codes not related to animal welfare
- Codes relating to non-specific conditions

Related codes were grouped into code combinations where biologically sensible, based on consensus assessment including three of the authors (HH; SSN; BF) and two other experts at the University of Copenhagen. For example, healed fractures recorded separately for various different sites in the carcass were grouped into a single category. A final exclusion criterion was then imposed to remove any meat inspection codes (or combinations of related codes) with a sufficiently low observed prevalence in the calf or adult dataset so that logistic regression models could not reasonably be implemented. This prevalence threshold was set at an observed prevalence of greater than zero in a minimum of 50 separate groups of animals, with the criterion implemented independently for each code and age group. The final dataset describing the number of recorded codes and total number of animals in each group (abattoir of slaughter combined with farm of origin) was created separately for each of the selected codes from the overall calf and adult datasets.

Model fitting

For each of the separate datasets included in the analysis, a random effects logistic regression model was used to describe the recorded prevalence of each meat inspection code (or combination of codes) for each group of animals. A random effect representing the group was fitted to every model in order to account for the residual extra-Binomial variance associated with clustering of observations within each individual combination of abattoir and farm (the rationale for this is discussed in detail by Harrison, 2014). No fixed effects (other than an intercept term) were considered for inclusion in the model. All models were fitted using the `glmer` function of the `lme4` package (Bates et al., 2014) in R (R Development Core Team, 2014). Variance estimates for random effects were taken from the model summary statistics, and estimates for the individual factor level effects within random effects terms were extracted from the random effects terms using the `ranef()` function. Confidence intervals for all parameters were obtained using parametric bootstrapping from the final fitted model with 250 iterations, which was deemed sufficient to approximate the true confidence interval. Model fit was assessed by comparing the likelihood obtained from the full model to a 95% confidence interval for the likelihood of data simulated under the same model, using the same parametric bootstrap procedure.

Model selection method

The improvement in model fit yielded by each of the two random effects terms (farm and abattoir) was assessed sequentially by comparing the fit of a model without the random effect to be tested (the null model) to the fit of a model including the random effect to be tested (the alternative model). The model fit comparison was done by comparing the difference in log likelihood from the null and alternative models to a distribution of the same statistic generated from simulated data. The probability that the data are consistent with the null model is given by the proportion of simulated likelihood ratio statistics that are greater (or equal) in magnitude to the observed likelihood ratio statistic. Given a desired alpha error rate (in this case a p-value of 0.05), a cut-off can then be imposed for rejecting the null hypothesis in the same way as the usual likelihood ratio test, but without the assumption that the test statistic follows a chi-square distribution, or the requirement to specify a number of degrees of freedom. This exact numerical approximation to the p-value requires only a sufficient number of simulated likelihood ratio test statistics to ensure that the decision to accept or reject the null hypothesis is justified. This procedure was verified to produce the desired alpha error rate of 5% using a simulation study (data not shown).

The simulated test statistic was obtained as follows. First, the parameter estimates obtained from the null model were used to obtain a simulated dataset, with the individual random effect estimates for group resampled from the variance estimate for this random effect. Both null and alternative models were then re-fitted to the simulated data, before recording the difference in log likelihood observed between the two fitted models. This simulated likelihood ratio was then compared to the likelihood ratio observed from the true dataset to determine if the observed likelihood ratio was greater than that obtained using data simulated under the null model. This process was repeated until the target threshold of $p=0.05$ lay outside (either above or below) the 99% confidence interval for the true p-value. If the 99% confidence interval still contained 0.05 after 1000 bootstrap iterations (representing a high degree of certainty about a p-value very close to the cut-off value), the estimate at that point was used to determine the outcome of the likelihood ratio test.

Model selection procedure

A forward step-wise selection algorithm was used to test the candidate random effects. Starting from the simplest model we considered, with intercept and group random effect terms fitted, the additional random effects representing the farm of origin ('Farm') and abattoir of slaughter ('Abattoir') were sequentially considered for inclusion in the model using the method specified above. The 'Abattoir' random effect was first considered for inclusion in the model, followed by the 'Farm' random effect. In the cases where the 'Abattoir' effect was initially rejected for inclusion but the 'Farm' effect was included, the 'Abattoir' effect was subsequently re-tested for inclusion.

RESULTS

Data summary

The number of unique farms of origin for each abattoir, and the number of farms sending animals to different abattoirs, is shown in Figure 1. The majority of farms sent animals to only one abattoir, but around 12% of farms sent animals to up to four abattoirs. A small number of farms (38 in total) sent animals to five or six abattoirs, and a single farm sent one or more adult cow to all eight abattoirs during 2012 (data point omitted from Fig. 1). There were a total of 13,681 and 8,109 observed combinations of farm and abattoir (group) for adults and calves, respectively.

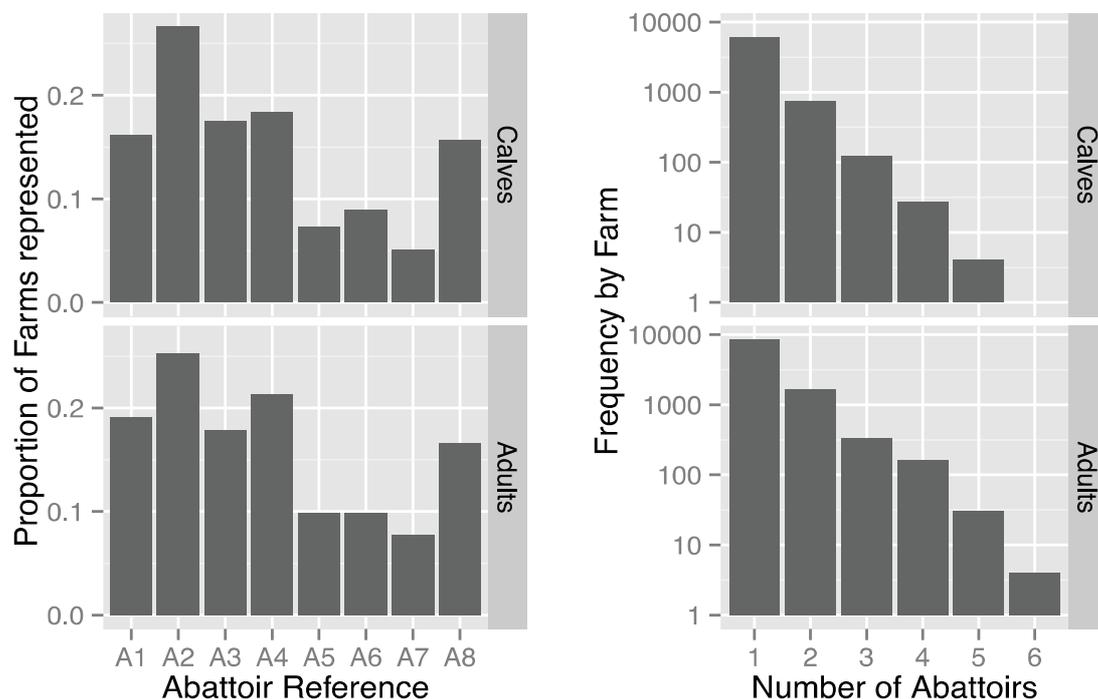


Fig. 1 The proportion of farms sending one or more animal to each abattoir (left) and number of farms (on the log scale) sending animals to multiple abattoirs (right) for calves (top) and adult cattle (bottom) in Denmark during 2012

Of the 92 slaughterhouse codes recorded, the following were removed: 6 slaughter plant codes; 8 transport-related codes; 16 acute conditions; 2 CNS conditions; 12 conditions not

related to animal welfare; and 4 non-specific codes. Of the remaining 44 slaughter codes, 19 were used individually and 25 were grouped into 9 separate combinations of related codes, giving a total of 28 code combinations each for calves and adults. Of these 28 codes, 5 were removed from the adult data, and 14 were removed from the calf data, due to low prevalence.

Modelling results

After exclusion of unsuitable meat inspection codes, a total of 37 datasets (23 adult; 14 calves) with a total of 24 disease codes were taken forward for modelling. Of these, there was no evidence for a random effect of Abattoir or Farm in 12 datasets. Descriptions of the remaining 18 unique codes together with random effect estimates and crude estimated prevalence are given in Table 1. Figure 2 shows the corresponding individual abattoir effect estimates (extracted from the random effect component of the fitted model) where the preferred model contained both farm and abattoir random effects.

Table 1. Model inference for the meat inspection codes used for cattle slaughtered in Denmark during 2012. Crude prevalence estimates are shown alongside random effects variance estimates for Farm and Abattoir where these were fitted in the final model. Estimates in italics indicate models with poor fit, and underlined estimates indicate non-overlapping 95% confidence intervals between abattoir and farm variance estimates.

| Code | Description | Calves | | | Adults | | |
|------|--|-----------|--------------------|--------------------|-----------|--------------------|--------------------|
| | | Prev % | Abb. σ^2 | Farm σ^2 | Prev % | Abb. σ^2 | Farm σ^2 |
| PYR | Pyrexia | 0.10 | 0.33 | 0.44 | 0.27 | -- | <i>0.3</i> |
| ENDO | Endocarditis | 0.03 | -- | -- | 0.19 | -- | <i>0.29</i> |
| LW | Lung worm | 0.03 | -- | -- | 0.19 | 3.18 | -- |
| CGE | Chronic gastroenteritis | 0.05 | -- | -- | 0.24 | 0.9 | 0.57 |
| RD | Renal disease | 0.19 | 1.54 | -- | 0.56 | 1.11 | -- |
| UD | Chronic uterine disease | 0.00 | -- | -- | 0.05 | -- | <i>2.07</i> |
| HD | Hoof disease | 0.03 | -- | -- | 0.23 | 8.1 | -- |
| OST | Osteomyelitis | 0.04 | -- | <i>2.02</i> | 0.16 | -- | -- |
| CA | Chronic arthritis and arthrosis | 0.43 | 0.19 | -- | 0.42 | -- | -- |
| MA | Muscle atrophy | 0.02 | -- | -- | 0.35 | 25.52 | -- |
| SE | Sores and eczema | 0.01 | -- | -- | 0.04 | -- | <i>0.04</i> |
| PNEU | Chronic pneumonia and pleuritis | 8.81 | 0.38 | 0.77 | 3.96 | 0.6 | 0.19 |
| CARD | Chronic pericarditis and traumatic reticulo-pericarditis | 4.00 | 0.2 | 0.49 | 5.33 | 0.25 | 0.19 |
| LD | Liver disease | 12.14 | <u>0.08</u> | <u>1.63</u> | 12.00 | <u>0.15</u> | <u>1.33</u> |
| ABS | Abscesses | 0.84 | 0.13 | 0.27 | 2.50 | 0.11 | 0.28 |
| HF | Healed fracture | 0.30 | -- | -- | 1.51 | <u>0.08</u> | <u>0.46</u> |
| SA | Skin abrasions | 1.14 | 0.2 | 0.48 | 6.70 | 0.27 | 0.31 |
| SR | Scabies and ringworm | 0.09 | <i>0.04</i> | -- | 0.01 | -- | -- |

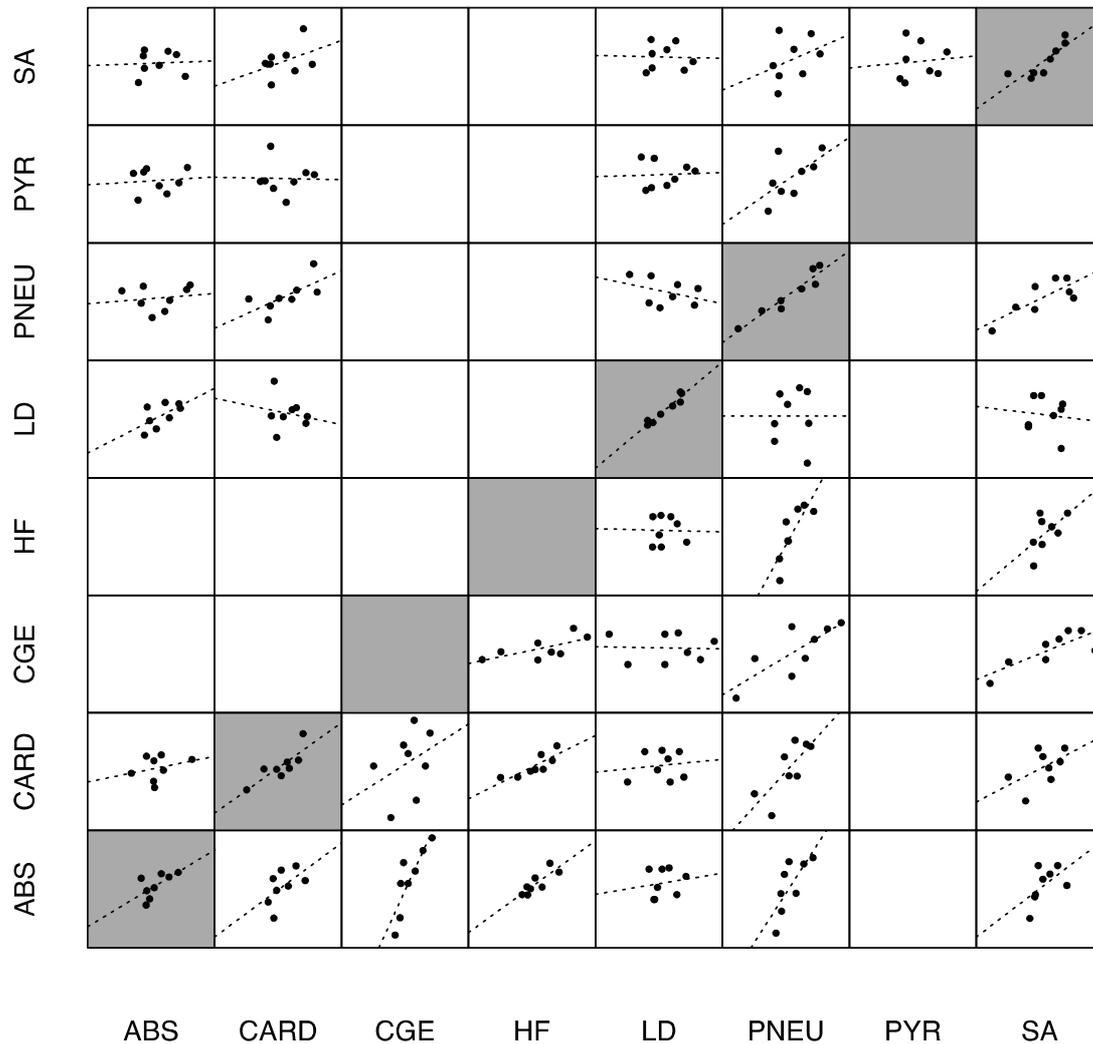


Fig. 2 Comparison between the eight individual abattoir estimates for preferred models with both farm and abattoir random effects, with regression lines shown dotted. The diagonal (grey shading) compares estimates from adults (x axis) and calves (y axis) for the same code. Off diagonal compares estimates between codes for calves (above diagonal) or adults (below diagonal). Axes are fixed to a scale of between -1.8 and 1.8 (on the logit probability scale).

There was evidence of random effects for both abattoir and farm for pyrexia (PYR) in calves; chronic gastroenteritis (CGE) and healed fracture (HF) in adults; and chronic pneumonia and pleuritis (PNEU), chronic pericarditis and traumatic reticulo-pericarditis (CARD), liver disease (LD), abscesses (ABS), and skin abrasions (SA) in both age groups. Where the full model was preferred, the mean estimate for the variance attributable to abattoir was smaller than that due to farm in 10 out of 13 cases, although the largest estimates for variance due to abattoir were obtained from lung worm (LW) and muscle atrophy (MA) in adults, and renal disease (RD) in both age groups, where no farm random effect was fitted. There was a significantly higher (non-overlapping 95% confidence intervals) variance estimate for abattoir compared to farm for healed fractures (HF) in adults, and liver disease (LD) in both adults and calves. There was no evidence for lack of fit for any model with both

farm and abattoir random effects, but a bootstrapped distribution of likelihoods obtained using simulated data did not contain the observed likelihood for five of the models including just a farm random effect, and one of the models including just an abattoir random effect (shown in italics). Some care should therefore be exercised when interpreting the results from these six models.

In addition to the overall random effect of abattoir, it may be useful to compare the estimates for individual abattoir effects contained within the fitted random effect structure. There was a clear positive linear relationship between the estimated abattoir effects for adults and calves for the same code (Figure 2; diagonal plots). For the six codes with highest observed prevalence, there was generally positive agreement within age groups between abattoir estimates for PNEU and other slaughter codes, as well as between ABS, CARD, HF, SA and other slaughter codes for adults. However, there was very little agreement between the abattoir effects for LD and the other conditions.

DISCUSSION

Of the 92 slaughter codes available, 44 were deemed by animal welfare experts to be potentially indicative of on-farm welfare. Even after grouping related conditions, only 4 of the 28 potentially relevant codes had a prevalence of over 1% in both adults and calves (an additional two codes had a prevalence of over 1% in adults only). Therefore, even with a dataset of over 100,000 animals grouped into approximately 10,000 groups of animals, considerable difficulties remained in modelling these rare outcomes. However, there was significant evidence for a random effect of abattoir (with or without a random effect of farm) in either adults or calves for 14 of the code combinations, which tended to be those with the highest prevalence. Where both farm and abattoir random effects were indicated, the estimate for the variance due to farm tended to be larger than that of abattoir. The relative importance of these standard deviation estimates to the prevalence on the logit scale can be compared using the square root of the estimates given in Table 1, although it should be noted that the abattoir random effect is likely to be under-estimated to a greater degree than that for farm due to the smaller number of grouping levels. However, the farm effect variance was estimated to be the larger of the two for three of the codes for adults (PNEU, CARD & CGE), and there are also a number of conditions for which there is evidence for an effect of abattoir but not for farm; LW, HD, MA in adults, CA in calves, and RD in both age groups. This observation is consistent with a greater degree of variation between abattoir practices in relation to these codes, for example examination of feet and kidneys, which is a highly plausible conclusion. It is also worth noting that the prevalence of some of these conditions varies considerably between age groups, which supports the decision to model data from calves and adults separately.

There was strong evidence that the individual abattoir effects were consistent between some codes, for example PNEU vs. ABS, CARD, CGE, HF and SA, indicating that the abattoir-specific factors which tended towards higher or lower levels of reporting are consistent between some abattoir codes. This is likely to be due to variation in inspection procedures, such as the amount of time available to examine carcasses and training procedures for staff within the same abattoir, as well as differences in recording practices between abattoirs, such as paper-based vs. electronic recording. One notable exception is LD, the abattoir code with the highest prevalence, which does not show any correlation with the abattoir effects for other codes, but does show near perfect correlation for the estimates in calves vs. adults. This may represent the different slaughterhouse procedure for examining

the liver as a discrete organ compared to many of the other conditions, which tend to be assessed directly from the carcass. However, it is also worth noting that the abattoir level variance for LD was small both in absolute terms and when compared to the relatively large estimates for the farm-level variance for the same condition.

The most likely interpretation of the observed consistency between abattoir effects for different codes is that abattoirs with high sensitivity of recording for some conditions tend also to have a high sensitivity of recording for other conditions. A more thorough inspection procedure at some abattoirs would be expected to result in greater sensitivity to detect a variety of different codes at these abattoirs, for example by having inspection procedures that better facilitate inspection of the carcass, leading to a higher detection rate for a number of different codes relative to an abattoir with different procedures. It may also be possible that recording one code on a carcass may lead to a more detailed inspection and therefore higher probability of detecting other codes in the same carcass. However, the correlations presented in Fig. 2 are for the overall abattoir effect of detecting these codes, and do not represent any possible correlation in detecting different codes on the same carcass, so we do not expect this possible effect to be a major contributor to the overall pattern observed. Another possible interpretation for the correlation between abattoir effects for different codes is that the recording practices do not differ between abattoirs, but a systematic bias exists in the decisions for farmers to send animals to different abattoirs based on the age, health status, breed or condition of their animals. This would lead to a discrepancy in the true prevalence of the relevant abattoir codes at different abattoirs, even after correcting for the farm of origin, because the abattoirs are receiving animals from different populations. However, given the small number of major abattoirs in Denmark and high proportion of farms sending animals to only one abattoir, we find the first interpretation to be most likely. Conversely, negative correlations between abattoir effects for different codes may imply a certain level of cross-classification between codes, although no such effect was observed in these data due to the grouping of related codes.

The use of random effects models to describe parameters of interest presented two main methodological challenges. Firstly, the standard likelihood ratio test could not be applied to the model selection, because the assumptions regarding the chi-square distribution of likelihood ratio are invalidated by the nature of random effects. We solved this problem using a computationally intensive method to directly simulate a distribution of likelihood ratios under our null hypothesis, to which the observed likelihood ratio could be applied (McLachlan, 1987). This method was verified to give the desired 5% type-I error rate based on simulated data, and is generally applicable to robust model selection in the context of mixed models. The second methodological aspect of our work concerns the 95% confidence intervals of the parameters of interest; in this case the variance estimates for farm and abattoir random effects, as well as the individual estimates for the effect of each abattoir contained within the overall random effects term. There similarly exists no viable distributional approximation on which to base 95% confidence intervals for these parameters, but parametric bootstrapping approaches can be used to generate these intervals directly. A further difference between fixed effect models and random effect models is the effect of 'shrinkage' in random effects models. This tends to pull the estimates for individual levels within random effects terms towards the centre of the distribution, and is known to lead to underestimates of the variance of random effects terms in some situations (Burnham & White, 2002). The assumption of an approximately normal distribution of effect estimates between farms and abattoirs is also important, although in practice inference made from random effects models is relatively robust to this assumption as long as the true distribution is

continuous and uni-modal. In our case, we believe that the use of random effects terms is justified because of the large number of theoretically additive factors contributing to the overall distribution of effects between farms and abattoirs, and any shrinkage of individual effect estimates towards zero is conceptually conservative and therefore also justifiable in this situation.

In conclusion, our results support the hypothesis that in most cases the majority of the observed variance is due to differences between farms, but we also show that abattoir-specific effects contribute a significant portion of the observed variation in reported prevalence of some slaughter codes. This indicates that the sensitivity of routine surveillance in Denmark is affected by differences in the working practices between abattoirs, and it is therefore crucial to account for these differences before making use of information derived from slaughter codes to make comparisons regarding animal welfare between farms.

ACKNOWLEDGEMENTS

This project was funded by The Danish Agrifish Agency under The Ministry of Food, Agriculture and Fisheries of Denmark. The authors would like to thank Sine Norlander Andreassen, Anne Marie Michelsen and Henrik Elvang Jensen for their valuable assistance interpreting and grouping the meat inspection codes.

REFERENCES

- Anon (2004). Regulation (EC) No 854/2004 of the European Parliament and of the Council of 29 April 2004 laying down specific rules for the organisation of official controls on products of animal origin intended for human consumption
- Anon (2011). Cirkulære om udøvelse af kødkontrol. CIR1H nr 9611 af 19/12/2011. (Government Circular on conduct of meat inspection in Denmark, in Danish). Danish Veterinary and Food Administration, Glostrup, Denmark
- Bates, D., Maechler, M., Bolker, B. and Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7. <http://cran.r-project.org/package=lme4>
- Burnham, K.P. and White, G.C. (2002). Evaluation of some random effects methodology applicable to bird ringing data. *J. Appl. Stat.* 29, 245–264. doi:10.1080/02664760120108755
- Dupuy, C., Morignat, E., Maugey, X., Vinard, J., Hendrikx, P., Ducrot, C., Calavas, D. and Gay, E. (2013). Defining syndromes using cattle meat inspection data for syndromic surveillance purposes: a statistical approach with the 2005-2010 data from ten French slaughterhouses. *BMC Vet. Res.* 9, 88. doi:10.1186/1746-6148-9-88
- Guo, G. and Zhao, H. (2000). Multilevel Modeling for Binary Data. *Annu. Rev. Sociol.* 26, 441–462. doi:10.1146/annurev.soc.26.1.441
- Harley, S., More, S., Boyle, L., Connell, N.O. and Hanlon, A. (2012). Good animal welfare makes economic sense : potential of pig abattoir meat inspection as a welfare surveillance tool. *Ir. Vet. J.* 65, 1. doi:10.1186/2046-0481-65-11

- Harrison, X. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ* 2, e616. doi:10.7717/peerj.616
- Klauke, T.N., Piñeiro, M., Schulze-Geisthövel, S., Plattes, S., Selhorst, T. and Petersen, B. (2013). Coherence of animal health, welfare and carcass quality in pork production chains. *Meat Sci.* 95, 704–11. doi:10.1016/j.meatsci.2013.03.022
- Knage-Rasmussen, K.M., Rousing, T., Sørensen, J.T. and Houe, H. (2014). Assessing animal welfare in sow herds using data on meat inspection, medication and mortality. *Animal* 1–7. doi:10.1017/S1751731114002705
- Li, B., Lingsma, H.F., Steyerberg, E.W. and Lesaffre, E. (2011). Logistic random effects regression models: a comparison of statistical packages for binary and ordinal outcomes. *BMC Med. Res. Methodol.* 11, 77. doi:10.1186/1471-2288-11-77
- McLachlan, G.J. (1987). On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture. *Appl. Stat.* 36, 318. doi:10.2307/2347790
- Nielsen, S.S., Michelsen, A.M., Jensen, H.E., Barington, K., Opstrup, K.V. and Agger, J.F. (2014). The apparent prevalence of skin lesions suspected to be human-inflicted in Danish finishing pigs at slaughter. *Prev. Vet. Med.* 117, 200–6. doi:10.1016/j.prevetmed.2014.08.003
- R Development Core Team (2014). R: A Language and Environment for Statistical Computing. <http://www.r-project.org>.
- Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit. *J. R. Stat. Soc. Ser. B Statistical Methodol.* 64, 583–639
- Vial, F. and Reist, M. (2014). Evaluation of Swiss slaughterhouse data for integration in a syndromic surveillance system. *BMC Vet. Res.* 10, 33. doi:10.1186/1746-6148-10-33