Bayesian zero-inflated predictive modelling of herd-level *Salmonella* prevalence for risk-based surveillance

Benschop, J.*(1), Spencer, S. (2), Alban, L (3), Stevenson, M. (1) and French, N. (1)

(1) EpiCentre, Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand.
(2) School of Mathematical Sciences, University of Nottingham, Nottingham, UK.
(3) Danish Agricultural & Food Council. Vinkelvej 11, DK-8620 Kjellerup, Denmark

**corresponding author* : j.benschop@massey.ac.nz

Abstract

The national control programme for *Salmonella* in Danish swine herds introduced in 1993 has led to a large decrease in pork-associated human cases of salmonellosis. The pork industry is increasingly focussed on the cost-effectiveness of surveillance while maintaining consumer confidence in the pork food supply. Using national control programme data from 2003 and 2004, we developed a zero-inflated binomial model to predict which farms were most at risk of *Salmonella*. We preferentially sampled these high-risk farms using two sampling schemes based on model predictions resulting from a farm's covariate pattern and its random effect. Zero-inflated binomial modelling allows assessment of similarities and differences between factors that affect herd infection status (introduction), and those that affect the seroprevalence in infected herds (persistence and spread). Both large (producing greater than 5,000 pigs per annum), and small herds (producing less than 2,000 pigs per annum) were at significantly higher risk of infection and subsequent seroprevalence, when compared with medium-sized herds (producing between 2,000 and 5,000 pigs per annum). When compared with herds being located elsewhere, being located in the south of Jutland significantly decreased the risk of herd infection, but increased the risk of a pig from an infected herd being seropositive. The model suggested that many of the herds where *Salmonella* was not detected were infected, but at a low prevalence. Using cost and sensitivity, we compared the results of our model-based sampling schemes to those under the standard sampling scheme, based on herd size, and the recently introduced risk-based approach. Model-based results were less sensitive but showed significant cost savings. Further model refinements, sampling schemes, and the methods to evaluate their performance are important areas for future work, and these should continue to occur in direct consultation with Danish authorities.

Introduction

New challenges for animal health surveillance for zoonotic disease include those associated with developing reduction strategies for surveillance systems for diseases that in the past represented an important risk, when today the risk to consumers is substantially reduced (Willeberg, 2006). *Salmonella* in Danish pork is an example of a disease that meets these criteria.

The Danish swine *Salmonella* surveillance-and-control programme (DSSCP) was instigated in 1993 by the Danish Ministry of Food, Agriculture and Fisheries in response to a human epidemic of salmonellosis (Baggesen et al., 1996). The programme's objective is to lower the prevalence of *Salmonella* so that domestically produced pork is no longer an important source of salmonellosis in humans (Mousing et al., 1997). The estimated number of cases of salmonellosis in humans in Denmark attributable to pork consumption decreased from 1,444 in 1993, to 215 in 2005 (Nielsen et al., 2001) (Ministry of Family and Consumer Affairs, 2006).

During 2008 there was a large and sustained outbreak of human salmonellosis due to *Salmonella enterica* serotype Typhimurium phage type U292 in Denmark. This became the largest outbreak recorded in Denmark since the present surveillance system became active in 1980 (Ethelberg et al., 2008). Locally produced pork and pork products have been suspected. However, the source has not been identified. In the face of this recent epidemic the current climate in Denmark is probably not conducive for proposing a

surveillance reduction strategy. Such strategies require a delicate balance between satisfying producer and industry concerns about cost-effective testing and maintaining consumer confidence in food supply. It makes sense that any strategy involving a reduction in testing should demonstrate an equal or greater sensitivity as the existing one, regardless of the potential efficiency gains.
It is possible to meet the differing needs of both consumer confidence in food supply and industry requirements for a surveillance reduction strategy if a targeted approach is used. Hereby populations with higher risk of infection are preferentially sampled. Our objective is firstly to develop a model that predicts which farms are most at risk of *Salmonella*. Secondly, we preferentially sample the high-risk farms and compare our results to those under: (1) the standard sampling scheme, based on herd size, and (2), the recently introduced risk-based approach (Ministry of Family and Consumer Affairs, 2006). In this way we are able to evaluate the impact of alternative sample collection strategies on overall system performance.

Materials and Methods

Data were obtained from three sources. Firstly, the Danish Central Husbandry Register provided a unique herd identifier, details of farm location, herd size and the number of sows in the herd. Secondly, the central database of the DSSCP provided the date of sampling, and the result of the Danish-mix ELISA for 2003 and 2004. The third source of data was the Danish Specific Pathogen Free (SPF) Company which provided health status details associated with each farm.
Four sampling schemes were used or developed:

(1) Original herd size-based sampling (OHS). The number of samples taken depended solely on herd size: the aim was to take 60, 75, or 100 samples annually from herds with an estimated annual kill of 200 – 2,000, 2,001 – 5,000, and greater than 5,000 slaughter pigs respectively (Alban et al., 2002). This scheme represents the bench-mark to which we compare the alternative sampling strategies.

(2) DMA risk-based sampling (DRB). We applied a modified version of the sampling criteria that was introduced by the DSSCP to herds in July 2005 (Enoe et al., 2003; Ministry of Family and Consumer Affairs, 2006). Our modification is that we have extended the time period over which herds are assessed to determine their prevalence to be the whole year, rather than the previous five months.

(3) Model derived risk-based sampling A (MRBA). We developed a targeted surveillance strategy based on our previous risk-factor, spatial and temporal analyses of the DSSCP data (Benschop et al., 2008a; Benschop et al., 2008b; Benschop et al., 2008c). All herds with a predicted median within-herd seroprevalence at or below a model determined cut-off in 2003 were identified as low risk and were placed on the DRB scheme. This prediction was based on the farm's covariate pattern and random farm effect. All other herds (above the predicted within-herd seroprevalence threshold) were left on the current sampling scheme for 2004 based on herd size (OHS).

(4) Model derived risk-based sampling B (MRBB). As in MRBA above, all herds with a predicted median within-herd seroprevalence at or below a model determined cut-off in 2003 were identified as low risk and were placed on the DRB scheme. The remaining herds were then assigned to two different sampling schemes depending on their predicted seroprevalence in 2003: (1) those with a predicted seroprevalence that was <0.25 or >0.55 were left on the current sampling scheme based on herd size; (2) those with a predicted seroprevalence of between 0.25 and 0.55 were more intensively sampled to provide 95% confidence that we were within 0.05 of the true value of the predicted seroprevalence. This range was chosen as these herds were near the cut-off for level 2 *Salmonella* status (0.40) (Alban et al., 2002).

The frequency histogram of the herd-level prevalence based on the actual test results from the OHS sampling strategy for 2003 and 2004 showed a large amount of variation with a predominance of test-negative herds. These test-negative herds can come from two types of disease-negative herds: (1) those that are truly uninfected and therefore every sample is negative, and (2); those that are, in fact, infected but provide insufficient samples to detect the presence of infection. This led us to propose a zero-inflated binomial (ZIB) approach to model herd level *Salmonella* prevalence as it reflected our understanding of what is happening on the farm. The ZIB model has two herd level outcomes: the probability of infection and, conditional on infection being present, an estimate of herd-level seroprevalence.

Variables that might explain both the presence of infection and herd level prevalence included herd size, farm location, number of sows present, and herd health status. We developed a logistic model within a Bayesian framework using WinBUGS version 1.4.1 (Gilks et al., 1994). This was extended to a zero-inflated binomial model and specified as follows:

$$cases[I] \sim Bin(pop[i], p[i])$$

Here the number of cases from the *i*th herd is binomially distributed as a function of the number of trials (tests for *Salmonella* antibodies in meat-juice) *pop[i],* and the probability of a test being positive (adjusted OD% >10), *p[i].*

We further defined:

$$p[i] = rho[i]*J[i]$$

Where *J[i]* is an indictor variable representing infection status of the *i*th herd, *rho[i]* is the sero-prevalence conditional on the presence of infection. The term *rho* therefore represents the probability of finding infection in a randomly chosen pig from an infected herd. The latent variable *J[i]* is distributed as:

$$J[i] \sim Bern(q[i])$$

Where *q[i]* is the probability of a herd being infected. This latent variable was modelled as:

$\log\left(\frac{q_i}{1-q_i}\right) = \alpha_0 + \alpha_1 x_{1i} + \ldots + \alpha_m x_{mi} + A_i$	Equation 1

The latent variable *rho[i]* was modelled as:

$\log\left(\frac{rho_i}{1-rho_i}\right) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_m x_{mi} + B_i$	Equation 2

We planned to use this model, based on 2003 data, to predict the probability of infection and seropositivity in 2004. These predictions inform our sampling strategies.

To check for consistency between years (2003 and 2004) we ran the model on both years of data separately and compared the magnitude and direction of the regression coefficients and the correlation between the random farm effects (A_i and B_i). This was thought to be important, because substantial changes in pig- and herd-level risks for infection (arising from, for example, changes in herd size or changes in the price of feed) from one year to the next could reduce the ability of the 2003 model to predict herd-level behaviour in 2004.

A scatter-plot of the median conditional sero-prevalence *q[i]* versus the median probability of infection *rho[i]* was used to identify the cut-off for the two model-derived risk-based sampling schemes. The results from all four sampling schemes were compared by considering cost, the number of false negative farms and the number of farms detected with a within herd sero-prevalence of ≥ 0.40.

Results

When we ran the model on both years of data separately there was no change in the sign of the estimated regression coefficients, and only minor changes in magnitude. There was moderate positive correlation between the 2003 and 2004 values of each of the random farm effects: A_i (0.18) and B_i (0.52). The scatter-plot of the median conditional sero-prevalence *q[i]* versus the median probability of infection *rho[i]* showed a partial distinction in predicted seroprevalence between herds that were detected as positive and those that were not. This provided us with our cut-off threshold of 0.09 for the sampling schemes.

Factors associated with the probability of a herd being infected included herd size and location. A herd producing less than 2,000, or greater than 5,000 pigs for slaughter per year had a 1.58 (95% CI: 1.18–2.11) and 2.08 (95% CI: 1.42–3.14) greater odds of infection with *Salmonella*, respectively, compared with herds producing between 2,000 and 5,000 pigs per year for slaughter. Herds in Sønderjylland had a lower risk of being infected than herds outside Sønderjylland (OR= 0.25, 95% CI: 0.19-0.33)

Factors associated with the level of seropositivity in a herd, given that the herd is infected, included herd size, location, the presence of sows and herd health status. The odds of a pig being seropositive in an infected small or large herd was increased by a factor of 1.16 (95% CI 1.06-1.28) compared with a pig being seropositive in an infected medium herd. Compared with farms located outside of Sønderjylland, the odds of pigs being *Salmonella* positive on farms within Sønderjylland was increased by a factor of 1.68 (95% CI: 1.51–1.86).
Table 1 shows the performance of each of the sampling schemes. The scheme with the lowest cost was MRBA; the one with the highest cost was OHS. The one with the lowest number of false negatives and highest sensitivity was OHS and the one with the highest number of false negatives and lowest sensitivity was MRBA and MRBB.

Table 1: Performance of four sampling schemes for surveillance for *Salmonella* in Danish finisher herds in 2004, n = 8151 herds.

Sampling scheme	OHS	DRB	MRBA	MRBB
Number of false negative farms[a]	731	1186	3257	3251
Sensitivity	0.91	0.85	0.60	0.60
Number of high positive farms[b]	304	849	1148	1199
Cost of sampling scheme (€1000)	1.118	959	372	479

[a] Farms infected in 2004 with *Salmonella* but not detected by the sampling scheme
[b] Farms the sampling scheme has detected at a *Salmonella* seroprevalence of ≥ 0.40

Conclusion

The use of the ZIB model has the potential to allow assessment of the extent of the similarities and differences between factors that affect herd infection status (introduction) and those that affect the seroprevalence in infected herds (persistence and spread). Our work would suggest that many of the farms where the disease has not been observed are actually infected but with low prevalence. Using cost and sensitivity, we compared the results of our model based sampling schemes to those under the standard sampling scheme, based on herd size, and the recently introduced risk-based approach. Model-based results were less sensitive but show significant cost savings. We believe that our framework for zero-inflated modelling has provided a useful starting point for further exploration of this technique in the design of surveillance systems.

(References can be obtained upon request)