Contents lists available at ScienceDirect

Microbial Risk Analysis

journal homepage: www.elsevier.com/locate/mran

Evaluation of risk-based surveillance strategies for *Salmonella* Dublin in Danish dairy herds by modelling temporal test performance and herd status classification errors

Alessandro Foddai[a,*], Jørgen Nielsen[b], Liza Rosenbaum Nielsen[c], Erik Rattenborg[b], Hans Ebbensgaard Murillo[d], Johanne Ellis-Iversen[a,d]

[a] National Food Institute, Technical University of Denmark, KGS, Lyngby 2800, Denmark
[b] Danish Agriculture and Food Council, SEGES, Aarhus 8200, Denmark
[c] Department of Veterinary and Animal Sciences, University of Copenhagen, Frederiksberg C 1870, Denmark
[d] Danish Veterinary and Food Administration, Glostrup 2600, Denmark

ARTICLE INFO

Keywords:
Salmonella Dublin
Risk-based surveillance
Evaluation
Simulation
Temporal test performance
Status misclassification

ABSTRACT

The potential risk-based improvement of the *Salmonella* Dublin surveillance programme in Danish dairy herds was investigated, considering herd status misclassifications due to testing errors. The programme started in October 2002. Currently (early 2021) all dairy herds are classified based on quarterly bulk tank milk (BTM) testing with an indirect antibody ELISA (iELISA). Over the last two decades, the prevalence of herds classified as "likely infected" (levels 2,3) reduced remarkably. However, since 2015, the apparent prevalence has increased again, calling for improved surveillance and control to protect animal and human health. A deterministic simulation model based on data (2018–2019) from 2283 dairy herds in level 1 ("most likely free from infection"), was developed to estimate status misclassifications as false negative (FN) and false positive (FP) herds, under two testing strategies. These were: (A) the current system based on quarterly BTM testing only, and (B) an alternative strategy based on additional blood testing of up to eight calves, within herds at high risk of infection (HR). Both strategies were evaluated using three risk classification methods (I to III) and four sensitivity analysis scenarios (SA1-4), where different temporal performances were simulated for the iELISA in BTM. To apply strategy *B*, the best high-risk classification method (II), which combined managerial applicability and minimized errors, would require testing approximately 1000 calves across 127 HR herds. In that case, strategy A would cause 3 FNs and 67 FPs, by assuming annual BTM sensitivity (BTMSe) 95% conditional on a 1-year disease history and specificity (BTMSp) 97%. Whereas strategy *B* could cause a similar number of FNs, but 7 FPs more, assuming a sensitivity (Se) of 77% and specificity (Sp) of 99% in individual blood-samples (SA1). Assuming also quarterly BTMSe 53% and BTMSp 99.9% (SA4), strategy A derived 28 FNs and 2 FPs, while strategy B resulted in 6 FNs less and 8 FPs more. Therefore, strategy *B* could improve early detection of infected HR herds, while strategy A would avoid more unnecessary restrictions in false-positive herds. This improves knowledge on the potential use of additional blood testing in HR herds and illustrates how deterministic modelling can be used to improve disease surveillance and control.

1. Introduction

The bacterium *Salmonella enterica* subsp. *enterica* serovar Dublin (*S.* Dublin[1]) is a zoonotic pathogen infecting mostly cattle and leading to mortality and production losses (Richardson and Watson, 1971; Nielsen et al., 2012). Transmission to humans occurs through consumption of non-pasteurised dairy products, insufficiently cooked meat or occupational exposure to infected animals (Fierer, 1983; Helms et al., 2003; Harvey et al., 2017).

In Denmark, a surveillance programme of *S.* Dublin was initiated in 2002 covering all cattle herds (Anonymous, 2004; Nielsen et al., 2004; Nielsen, 2013a). Currently (early 2021), all dairy herds are tested

* Corresponding author.
E-mail address: alefo@food.dtu.dk (A. Foddai).

[1] A list of abbreviations is provided in the Appendix

https://doi.org/10.1016/j.mran.2021.100184
Received 12 May 2021; Received in revised form 28 July 2021; Accepted 14 September 2021
Available online 30 September 2021

quarterly in bulk tank milk (BTM), with an indirect antibody-detecting Enzyme-Linked Immunosorbent Assay (iELISA). This test provides results as an ODC%-value, which is a background corrected proportion of the test sample optical density (OD) to a known positive reference sample (Hoorfar et al., 1993; Hoorfar et al., 1995; Nielsen et al., 2004; Warnick et al, 2006).

During the last decades, the eradication programme has led to a reduced prevalence of dairy herds classified as "likely infected" (level 2 or 3), and thus, classifying more than 90% of the herds as "most likely free" from infection (level 1) (Warnick et al., 2006; SEGES, 2021). However, since 2015, an increase in prevalence has been observed (SEGES, 2021). At the time of writing, the highest apparent prevalence was reported in the region of Jylland – Syd (17.4%) and the average inter-regional prevalence (across 10 regions) was around 6.4% (SEGES, 2021).

During eradication programmes, re-increases of prevalence of infected dairy herds could be related to lowered temporal sensitivity (Thurmond, 2003) in BTM testing, for example when herd size increases (Foddai et al., 2014; Foddai et al., 2016). The size of Danish dairy herds is known to increase at high speed (Foddai et al., 2015; Danish Agriculture & Food Council, 2020). Larger sizes can cause a delay in the detection of antibodies, due to their high dilution in large milk tanks. Therefore, *S.* Dublin infected herds could be wrongly classified as "likely free from disease" (level 1), before a sufficiently high within-herd seroprevalence is reached in the lactating cows and the BTM turns positive at the iELISA. Until that point, these would be "false negative" (FN) herds. The time elapsing between disease introduction into a herd and its detection can be defined as an high-risk period (HRP) (Horst et al., 1997), because during that time window, the pathogen may be spread to other farms, while the farmer is unaware that animals are infected. To shorten the HRP and to allow for early detection of newly infected herds, temporal herd-level sensitivity (HSe) needs to be increased, e.g. by using more sensitive test(s) and/or by supplementary testing. A high HSe may generate a high negative predictive value (NPV) in testing negative herds, if the probability of infection remains low.

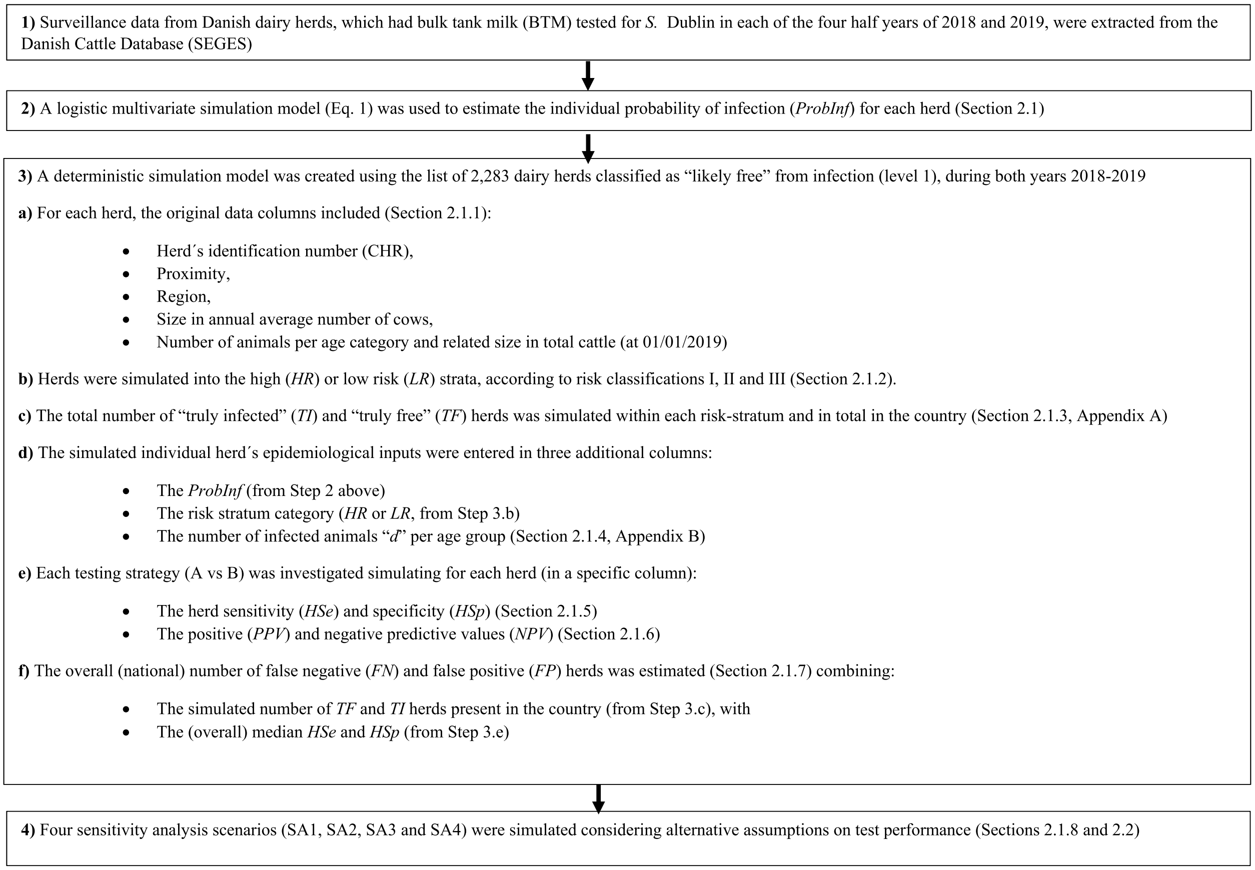
Improvement of the Danish eradication programme may also be achieved by applying risk-based surveillance (Stärk et al., 2006). In that case, cost-efficiency could be optimized by prioritising resources towards population strata at higher risk of: exposure, infection, detection, and/or transmission, while minimising consequences (Stärk et al., 2006; Cameron, 2012; Cameron et al., 2014; Hansen et al., 2018; Alban et al., 2020), such as disease spreading from FN herds to others, during the HRP.

The purpose of this study was to investigate alternative test-strategies for the Danish *S.* Dublin surveillance programme in dairy cattle herds, to underpin efforts aimed to reduce the prevalence of this infection in the Danish cattle population. This was accomplished by comparing alternative testing approaches to the current surveillance approach and assessing how the HSe and NPV could be improved, without reducing herd specificity (HSp) and positive predictive values (PPV) much. Thus, the potential numbers of FN and false positive (FP) herds were compared under alternative combinations of testing strategies and herds risk classification methods (I to III). This implied additional testing efforts towards specific population strata to make the programme risk-based. Annual and quarterly HRPs were considered to reflect changes of temporal sensitivity (BTMSe) and specificity (BTMSp) of the iELISA when used on BTM samples.

2. Materials and methods

A diagram resuming the main steps followed in this study, is shown in Fig. 1. Two alternative testing strategies were investigated:

A) Current strategy based on quarterly BTM antibody testing of all level 1 dairy herds



1) Surveillance data from Danish dairy herds, which had bulk tank milk (BTM) tested for *S.* Dublin in each of the four half years of 2018 and 2019, were extracted from the Danish Cattle Database (SEGES)

2) A logistic multivariate simulation model (Eq. 1) was used to estimate the individual probability of infection (*ProbInf*) for each herd (Section 2.1)

3) A deterministic simulation model was created using the list of 2,283 dairy herds classified as "likely free" from infection (level 1), during both years 2018-2019

a) For each herd, the original data columns included (Section 2.1.1):

- Herd´s identification number (CHR),
- Proximity,
- Region,
- Size in annual average number of cows,
- Number of animals per age category and related size in total cattle (at 01/01/2019)

b) Herds were simulated into the high (*HR*) or low risk (*LR*) strata, according to risk classifications I, II and III (Section 2.1.2).

c) The total number of "truly infected" (*TI*) and "truly free" (*TF*) herds was simulated within each risk-stratum and in total in the country (Section 2.1.3, Appendix A)

d) The simulated individual herd´s epidemiological inputs were entered in three additional columns:

- The *ProbInf* (from Step 2 above)
- The risk stratum category (*HR* or *LR*, from Step 3.b)
- The number of infected animals "*d*" per age group (Section 2.1.4, Appendix B)

e) Each testing strategy (A vs B) was investigated simulating for each herd (in a specific column):

- The herd sensitivity (*HSe*) and specificity (*HSp*) (Section 2.1.5)
- The positive (*PPV*) and negative predictive values (*NPV*) (Section 2.1.6)

f) The overall (national) number of false negative (*FN*) and false positive (*FP*) herds was estimated (Section 2.1.7) combining:

- The simulated number of *TF* and *TI* herds present in the country (from Step 3.c), with
- The (overall) median *HSe* and *HSp* (from Step 3.e)

4) Four sensitivity analysis scenarios (SA1, SA2, SA3 and SA4) were simulated considering alternative assumptions on test performance (Sections 2.1.8 and 2.2)

Fig. 1. Diagram representing the study layout.

B) Additional blood testing of up to eight calves within herds at high risk of infection (HR); while herds at low risk (LR) would continue to test only BTM.

In both strategies, the annual temporal BTMSe and BTMSp were simulated with values that reflected the performance of the iELISA after four consecutive quarters of a year (*i.e.* four BTM tests) as the HRP (Warnick et al., 2006). When strategy *B* was investigated for the same HRP, it was assumed that blood testing of calves was added in the 4th testing round. When both strategies were evaluated for single quarterly HRPs, the temporal BTMSe and BTMSp were simulated with values that reflected the performance of the iELISA considering a single BTM test and recent disease introduction.

2.1. *Logistic regression modelling to estimate the expected probability of herd infection*

National surveillance data from all dairy herds, which had BTM tested for *S.* Dublin antibodies in each of the four half years of 2018 and 2019, were extracted from the Danish Cattle Database (SEGES). The probability of infection (ProbInf) was calculated for each herd, based on estimates of the risk of becoming infected from a logistic regression model.

The logistic regression model was used to analyse a total of 2422 herds, after removing herds with annual average number of animals equal to zero and herds, which had been in level 2 in 2017. The latter were disregarded, because they were expected having a higher risk of re-becoming positive in 2018-2019 (Nielsen and Dohoo, 2012). Furthermore, the herds used for the logistic regression were in level 1 on 01/01/2018 and were divided into two groups: herds that stayed in level 1 and herds that switched to level 2 anytime during 2018-2019 (Table 1).

The status of each herd was assigned according to the rules of the surveillance and classification programme (Danish order No. 1326 of 29/11/2017; Anonymous, 2019). A herd was classified in level 1, if the average of four consecutive BTM tests was ≤ 25 ODC% and if in the 4th sample the increase from the average of the three previous BTM values was ≤ 20 ODC%. Three risk factors were considered for logistic regression:

1 Herd size: average number of cows present in the herd during the period 01/12/2018 to 30/11/2019.
2 Proximity: the number of neighbouring properties within a 5 km radius that were in official level 2 at any point in time during 2018-2019, and
3 Trade: the number of herds from which cattle were purchased during the two years.

Firstly, for each risk factor, a univariate analysis was carried out and the significance in predicting the *S.* Dublin level's change was evaluated for each herd. Secondly, five multivariable models were built to explore eventual associations and interactions between the three variables. Those were: model *A* = Size + Proximity; *B* = Size + Proximity + Trade; *C* = Size * Proximity + Trade (*i.e.* considering interaction between size and proximity); *D* = Size * Proximity * Trade (*i.e.* the full model considering interaction between all three factors), and *E* = Size * Proximity (*i.e.* without trade, but allowing an interaction between the two remaining factors).

Table 1
Number of Danish dairy herds in *S.* Dublin level 1 or 2 on 01/01/2018, which were considered for the logistic regression analysis carried out to estimate the herd's probability of infection (ProbInf).

Data from all of 2018 - 2019	Level 1 on 01/01/2018	
	Number of herds	Percentage
Constantly in level 1	2,286	94.4%
In level 2 at some point during the study period	136	5.6%
Total	2,422	100%

All interactions and Trade were non-significant, probably because purchase from level 2 and level 3 herds was prohibited during the study period (Danish order No. 1326 of 29/11/2017; Anonymous, 2019). Therefore, model A was finally used as:

$$\ln\left(\frac{p\ (x)}{1-\ p\ (x)}\right)=\beta_0+\beta_1 x_1+\beta_2 x_2 \tag{1}$$

Where *p*(*x*) was the probability that "*y*" (the binary variable: herd infected = 1 or not = 0) was equal to 1, *i.e. p*(*x*) is the infection probability, also called ProbInf in sections below. The ß$_0$ was the intercept, while x$_1$ and x$_2$ were the variables Size and Proximity with their respective regression coefficients ß$_1$ and ß$_2$. Thereafter, for each herd, the translation of ORs into ProbInf values was made by ProbInf = Odds / (1 + Odds), where Odds = exp (ß$_0$ + ß$_1$×$_1$ + ß$_2$×$_2$), since Odds = *p* (*x*) / (1 - *p* (*x*)) (Hosmer and Lemeshow, 1989).

2.1.1. *Overview of the deterministic simulation model used to evaluate the testing strategies*

To develop the deterministic simulation model, 2283 level 1 dairy herds were used (Tables 2 and 3), because they satisfied the following conditions during 2018-2019: (i) were BTM tested in all the four half years, (ii) had an annual average number of cows above zero, (iii) were always in level 1 and (iv) had at least one animal at 01/01/2019. Thus, the aim of this model was to compare the ability to detect infection using two testing strategies *A* and *B* in herds classified as likely *S.* Dublin free. The data variables used for each herd were:

- Herd's identification number (CHR).
- Proximity.

Table 2
Information on level 1 Danish dairy herds within the *S.* Dublin programme, according to different classifications of high risk and low risk herds.

	Classification I-II		Classification III	
Parameter	**HR herds**	**LR herds**	**HR herds**	**LR herds**
Number of herds	127	2,156	346	1,937
Proportion of herds (PrP)	5.6%	94.4%	15.2%	84.8%
Number of calves to test (n)	966	n.a.	2648	n.a.
Effective probability of infection (EPI)	21.6%	1.5%	9.6%	1.3%
Relative risk of infection (RR)	14.8	1	7.2	1
Truly infected (TI) herds	27	32	33	26
Truly free (TF) from infection herds	100	2,124	313	1,911

Herds classification method I = High risk (HR) herds and low risk (LR) herds divided using as cut-off the 95th percentile probability of infection (ProbInf). Classification II = HR herds had ≥ 8 neighbours in level 2 and ≥ 200 average cows per year. Classification III = HR herds had ≥ 200 annual average cows and were located in high prevalence regions (Himmerland, Jylland – Syd and Jylland – Sydvest). Classifications I and II led the same results, and thus, are presented together. n.a = not applicable.

Table 3

Number of *S*. Dublin level 1 dairy herds in each Danish region according to risk classification and stratum, with respective median number of neighbours in level 2 (proximity) and size in median number of cows or in total cattle. Within brackets are the 5th and 95th percentiles of each distribution.

	Classification I and II		Classification III	
Parameter	**HR herds**	**LR herds**	**HR herds**	**LR herds**
Neighbours in level 2	11 (8; 20)	2 (0; 11)	5 (1; 16)	2 (0; 12)
Size in number of cows	299 (207; 621)	148 (40; 435)	299 (206; 613)	139 (37; 403)
Size in total cattle	530 (333; 1169)	289 (84; 800)	538 (314; 1104)	273 (79; 729)
Region	**HR**	**LR**	**HR**	**LR**
Bornholm	0	26	0	26
Fyn	0	145	0	145
Himmerland	36	181	96	121
Jylland – Midt	7	243	0	250
Jylland - Midtvest	5	351	0	356
Jylland – Nord	3	312	0	315
Jylland – Øst	0	194	0	194
Jylland – Syd	49	438	182	305
Jylland – Sydvest	27	144	68	103
Sjælland	0	122	0	122

Herds classification method I = High risk (HR) herds and low risk (LR) herds HR herds divided using as cut-off the 95th percentile probability of infection (ProbInf). Classification II = HR herds had ≥ 8 neighbours in level 2 and ≥ 200 average cows per year. Classification III = HR herds had ≥ 200 annual average cows and were located in high prevalence regions (Himmerland, Jylland – Syd and Jylland – Sydvest). Classifications I and II led the same results, and thus, are presented together.

- ProbInf.
- Region of herd location.
- Herd size in annual average number of cows.
- Overall herd size from number of animals per age group at 01/01/2019 (young calves = 0–3 months old, old calves = 3–6 months, heifers-steers = 6–24 months, and adult cows > 2 years).

Throughout the paper, the terms “simulation” and “estimation” are used interchangeably, because although the inputs described above were obtained from data, others were simulated such as: the proportion of actually infected (undetected) level 1 herds, the within-herd disease epidemiology, and the temporal test performance. The simulation model was developed in R (R Core Team, 2013).

2.1.2. Classification of herds into high-risk and low-risk population groups

The deterministic simulation model reflected the differential risk of herd infection across 2283 level 1 dairy herds. Three alternative risk-based classification methods were investigated (Tables 2 and 3).

In classification I, the 95th percentile ProbInf (21.6%) was used as cut-off to split herds between the HR and LR strata. This was a statistical classification method.

In classifications II and III, cut-offs were defined using practical parameters, which would facilitate implementation in the surveillance programme.

In classification II, the HR herds had at least eight neighbours in level 2 and at least 200 cows in (annual) average. Those cut-offs corresponded to the minimum values observed in classification I. Nevertheless, the two classifications led to the same classification of herds across the two risk strata (Table 3).

In classification III, the HR herds had at least 200 cows in (annual) average and were located in the high prevalence regions (‘Himmerland’, ‘Jylland-Syd’ and ‘Jylland-Sydvest’) (Table 3).

2.1.3. Simulating between-herds infection: “truly infected” and “truly free” herds

The overall expected number of “truly” infected (TI), but classified as level 1 dairy herds, was simulated by multiplying the national median ProbInf (2.6%) times the 2283 herds. Accordingly, 59 TI and 2,224 (97.4%) “truly free” (TF) herds, were assumed in the country.

Thereafter, in each classification, the 59 TI herds were allocated to the HR and LR strata (Table 2) according to the proportion of herds (PrPs) and the herd́s individual relative risk (RRs) of infection within each stratum. Both parameters were used to calculate the related (within stratum) effective probability of infection (Martin et al., 2007a; 2007b), namely EPI_{HR} and EPI_{LR} (see Appendix A). These were multiplied for the number of HR and LR herds, to simulate the “TI” herds per stratum (Table 2).

Thus, Table 2 shows for each classification (I to III) and stratum (HR or LR): the number and PrPs of herds, the number of calves (n) to test under strategy *B* (in HR herds), the median EPI, the RR, and the number of simulated TI and TF herds.

Table 3 shows the number of herds per Danish region, the number of level 2 neighbours per herd (proximity), and the herd size in number of cows or in total cattle.

2.1.4. Simulating within-herd infection

Data on test results from infected herds is registered in the national Danish Cattle Database and was used to explore variability in within-herd infection epidemiology. It showed that approximately 16% of the level 2 herds had antibodies in cows only. In another 6% only calves 3–6 months old were antibody-positive, while the remaining 78% of herds had seropositive cattle in multiple age groups (unpublished data).

To account for this, the list with all 2283 level 1 dairy herds was randomized in Excel. In the top 366 (16%) herds, the infection was simulated only in adult cows and the total number of infected animals within that group was simulated as $d = \text{WGP} * \text{group size}$ (rounded up to the closest integer, *i.e.* $d \geq 1$). The WGP represented the within-group design seroprevalence at the day of testing (Martin et al., 2007a; 2007b). Similarly, in the 137 (6%) herds appearing from line 367 and downwards of the randomized list, the infection was simulated in the group of 3–6 months old calves only, and d was calculated as above. In the remaining 1783 (78%) herds, the overall within-herd design prevalence (WHP) was split between the different age groups (see Appendix B), following the same principles (Martin et al., 2007a; 2007b) applied in the previous section to simulate between-herds infection.

It must be noted that WGP and WHP, represented cut-offs (design prevalence) at which detection of the seropositive animal/s was expected to occur with the simulated (group or herd level) sensitivity at the day of testing (Martin et al., 2007a; 2007b). Both values were set at 10%. Nevertheless, only the number of infected 3–6 months old calves (d) was used to calculate the sensitivity gained from individual blood testing in strategy *B* (see next Section), because the aim was to evaluate how additional (risk-based) blood testing could increase overall temporal HSe, in HR herds where *S*. Dublin could be missed from testing the milking group.

2.1.5. Herd sensitivity and specificity

The overall HSe, HSp, NPV and PPV were estimated for each herd under each testing strategy. In strategy *A*, the HSe and HSp were similar to the BTMSe and BTMSp of the BTM iELISA, whereas for strategy *B*, both values were calculated. In both strategies, the BTMSe was assumed equal to 0%, for the 6% herds where infection was simulated only in calves older than three months. For the 16% herds simulated with only antibody positive cows, the group sensitivity (GSeOlderCalves) from testing of individual calves was set at 0%. Whereas, in herds with simulated infected calves, the GSeOlderCalves was estimated using a hypergeometric approximation (MacDiarmid, 1988).

$$\text{GSeOlderCalves} = 1 - (1 - \text{n}/\text{N} * \text{Se})^{\wedge\text{d}} \tag{2}$$

Where n = number of randomly tested calves (up to eight per herd) and N = number of calves present in the group. The Se was the sensitivity of the iELISA when used in individual blood (see Section 2.1.8), while d was the number of infected animals present within the group, as explained above. Thus, GSeOlderCalves represented the probability of detecting at least one antibody positive calf, if at least one was "truly" seropositive within the group. Moreover, from a managerial point of view, Eq. (2) assumed specificity = 100%. Hence, it was assumed that in the alternative surveillance programme (*B*), even one positive blood sample would classify the herd into level 2.

The group specificity obtained from individual blood testing, was estimated as GSpOlderCalves = Sp ^{n}. Where Sp was the individual diagnostic specificity (Section 2.1.8). The overall (parallel) HSe and HSp of testing strategy *B* (in HR herds), were then calculated assuming independence between groups and using Eqs. (3) and (4), respectively:

$$\text{HSeParallel} = 1 - (1 - \text{GSeOlderCalves}) * (1 - \text{BTMSe}) \tag{3}$$

$$\text{HSpParallel} = \text{GSpOlderCalves} * \text{BTMSp} \tag{4}$$

2.1.6. Negative and positive predictive values

The NPV and the PPV were estimated for both testing strategies within each risk strata, by using Eqs. (5) and (6) (Noordhuizen et al., 2001):

$$\text{NPV} = \frac{[\text{HSp} \ * \ (1 - \text{EPI}_{\text{H}})]}{[\text{HSp} \ * \ (1 - \ \text{EPI}_{\text{H}}) + (1 - \ \text{HSe}) \ * \ \text{EPI}_{\text{H}}]} \tag{5}$$

$$\text{PPV} \ = \ \frac{(\text{HSe} \ * \ \text{EPI}_{\text{H}})}{[(\text{HSe} \ * \ \text{EPI}_{\text{H}}) \ + \ (1 - \ \text{EPI}_{\text{H}}) \ * \ (1 - \ \text{HSp})]} \tag{6}$$

Where, the EPI_{H} represented the median effective probability of herd infection within the risk stratum (EPI_{HR} or EPI_{LR}), as explained in Section 2.1.3 and in the Appendix (A). It must be noted that, similar predictive values would have been obtained, if the within-stratum median ProbInf was instead used in Eqs. (5) and (6). The EPI_{HR} or EPI_{LR} were preferred as inputs, because they were more consistent with the number of TI and TF herds allocated by the model within each risk stratum (Table 2) from a national median ProbInf = 2.6%.

For testing strategy *B*, when Eqs. (5) and (6) were applied to HR herds, the HSe was set equal to the HSeParallel obtained from Eq. (3), while the HSp was set equal to the HSpParallel estimated in Eq. (4).

2.1.7. Number of false negative and false positive herds

The number of FN and FP herds were simulated in total and for each risk stratum (HR and LR). The overall median HSe simulated in Section 2.1.5 represented the probability that a TI herd is correctly classified as positive by the testing strategy, while the median HSp represented the probability that a TF herd is correctly classified as negative. The number of FN and FP herds were estimated as:

$$\text{FN} = \text{TI} * (1 - \text{HSe}) \tag{7}$$

$$\text{FP} = \text{TF} * (1 - \text{HSp}) \tag{8}$$

Where (1- HSe) was the probability that a TI was "wrongly" classified as negative and (1- HSp) was the probability that a TF was wrongly classified as positive, by the testing strategy used.

2.1.8. Simulating test performance as originally validated

In classifications I to III, the annual BTMSe and BTMSp were set at 95% and 97%, according to Warnick et al. (2006); who estimated those mean values when herds sizes were smaller than in the current situation and when the national herd prevalence was around 8%. The estimates represented the HSe and HSp after four consecutive quarterly BTM results (*i.e.* annual HRP), similarly to today's testing strategy (A).

For the blood testing of calves, the individual diagnostic Se was set at 85% or 77%, for cut-offs 25 ODC% or 50 ODC%, respectively. Whereas the related Sp was 88 or 95% (Nielsen et al., 2004).

2.2. Sensitivity analysis on iELISA's performance

A sensitivity analysis was carried out with four additional scenarios (SA1, SA2, SA3 and SA4), to investigate the impact of uncertainty on the current performance of the iELISA.

The risk-based classification II was used for all SA scenarios because: it combined the practicality of an eventual implementation and led the same HR and LR herds identified in the statistical classification I (Table 2). Classification III was disregarded for the sensitivity analysis, due to the very high number of HR herds and calves to test in blood (Table 2).

Moreover, for all SA scenarios, the individual blood Se and Sp were set at 77% and 99% respectively, for cut-off 50 ODC% (Nielsen and Ersbøl, 2004). The cut-off 25 ODC% was disregarded, because we knew it would have caused too many false positive herds (see results).

Scenario SA1 differed from the original scenario simulated under classification II, only because the Sp was increased from 95% (Nielsen et al., 2004) to 99% (Nielsen and Ersbøl, 2004).

Scenario SA2 differed from SA1, because the annual median BTMSe was reduced from 95 to 92%, while the BTMSp was increased from 97 to 98%. The BTMSe and BTMSp inputs used in SA2 were the minimum and maximum values estimated by Warnick et al. (2006). Thus, SA2 assumed that the current annual BTMSe could be lower and BTMSp higher than when the test was validated, due to current bigger herd sizes and higher antibodies dilutions in BTM.

In SA3, the BTMSe was set at 88% while the related BTMSp was increased to 99%, assuming that if the BTMSe reduced, the BTMSp improved in some way (*i.e.* if a TI herd was less likely to result positive, then also a TF herd was less likely to result false positive). Warnick et al. (2006) stated that "the probability of testing positive on the initial test was 88% or higher for all patterns, where the herd was infected in the current quarter". Accordingly, the iELISA performance assumed in SA3 represented the quarterly BTMSe as originally validated.

In SA4, the BTMSe was further reduced to 53% while the BTMSp was further increased to 99.9% (based on our opinion), to reflect the current impact of antibodies dilutions after quarterly HRPs.

3. Results

3.1. Testing strategy A: annual herd sensitivity, specificity and predictive values

For testing strategy A, the annual median HSe and HSp were 95 and 97% (Table 4a) in all herds classifications (I to III) and in both strata (HR or LR), similar to the values used as inputs for BTMSe and BTMSp (Warnick et al., 2006). Nevertheless, some of the related PPV differed across scenarios and strata, because the EPI_{HR} and EPI_{LR} differed too (Table 2).

In the HR herds, the annual median NPV was always around 99% (approximated), while the PPV ranged from 77% in classification III to 90% in classifications I-II.

In LR herds, with all three classifications, the NPV was 99% while the PPV was 30–32% (Table 4a).

3.2. Testing strategy B: annual herd sensitivity, specificity and predictive values

For testing strategy B, using the cut-off 25 ODC% for blood testing in HR herds, the annual median HSeParallel was around 98% in all three classifications (I to III), while the HSpParallel was 35%. The related median NPV was 98% (I-II) or 99% (III), while the PPV ranged from 14% (III) to 29% (I-II) (Table 4a).

In the same HR stratum, but using cut-off 50 ODC% for blood testing, the annual median HSeParallel was always around 97%, while the

Table 4a
Simulated annual median herd sensitivity, specificity and predictive values, for Danish dairy herds classified in level 1 within the *S.* Dublin eradication programme, for each combination of: testing strategy, risk classification procedure and risk stratum.

Risk classification I and II	HR-B-25	HR-B-50	HR-A	LR-A
HSe	98 %	97 %	95 %	95 %
HSp	35 %	64 %	97 %	97 %
NPV	98 %	99 %	99 %	99 %
PPV	29 %	43 %	90 %	32 %
	HR-B-25	**HR-B-50**	**HR-A**	**LR-A**
Risk classification III				
HSe	98 %	97 %	95 %	95 %
HSp	35 %	64 %	97 %	97 %
NPV	99 %	99 %	99 %	99 %
PPV	14 %	22 %	77 %	30 %

HSe = Herd temporal sensitivity; HSp = Herd temporal specificity; NPV = Herd negative predictive value, PPV = Herd positive predictive value. Testing strategy *A* = All herds (HR = high risk; LR = low risk) tested on BTM only. Strategy B = BTM testing in all herds plus blood testing in HR herds. Results presented using cut-off 25 or 50 ODC% (HR-B-25 or HR-B-50) for blood testing. Classification I = HR and LR herds divided using as cut-off the 95th percentile probability of infection (ProbInf). Classification II = HR herds had ≥ 8 neighbours in level 2 and ≥ 200 average cows per year. Classification III = HR herds had ≥ 200 annual average cows and were located in high prevalence regions (Himmerland, Jylland – Syd and Jylland – Sydvest). Classifications I and II led the same results, and thus, are presented together. In all three classifications (I to III) the original annual BTMSe 95% and BTMSp 97% (Warnick et al., 2006), were used. Whereas the blood diagnostic sensitivity (Se) was set at 85% or 77% for cut-offs 25 ODC% or 50 ODC%, and the specificity (Sp) was 88% or 95% (Nielsen et al., 2004).

Table 4b
Simulated median herd sensitivity, specificity and predictive values, for Danish dairy herds classified in Level 1 within the *S.* Dublin eradication programme, for each sensitivity analysis scenario (SA1, SA2, SA3 and SA4).

Scenario SA1	HR-B-50	HR-A	LR-A
HSe	97 %	95 %	95 %
HSp	90 %	97 %	97 %
NPV	99 %	99 %	99 %
PPV	72 %	90 %	32 %
Scenario SA2	**HR-B-50**	**HR-A**	**LR-A**
HSe	96 %	92 %	92 %
HSp	90 %	98 %	98 %
NPV	99 %	98 %	99 %
PPV	73 %	93 %	41 %
Scenario SA3	**HR-B-50**	**HR-A**	**LR-A**
HSe	94 %	88 %	88 %
HSp	91 %	99 %	99 %
NPV	98 %	97 %	99 %
PPV	75 %	96 %	57 %
Scenario SA4	**HR-B-50**	**HR-A**	**LR-A**
HSe	76 %	53 %	53 %
HSp	92 %	99.9 %	99.9 %
NPV	93 %	89 %	99 %
PPV	73 %	99 %	89 %

HSe = Herd temporal sensitivity; HSp = Herd temporal specificity; NPV = Herd negative predictive value, PPV = Herd positive predictive value. Testing strategy A = All herds (HR = high risk; LR = low risk) tested on BTM only. Testing strategy B = BTM testing in all herds plus blood testing in HR herds. All SA scenarios used the risk-based classification II and individual blood Se = 77% and Sp = 99% for cut-off 50 ODC% (Nielsen and Ersbøll, 2004) (*i.e.* only column HR-B-50 is reported compared to Table 4a). SA1 = BTMSe 95% and BTMSp 97% (Warnick et al., 2006) representing original annual BTM temporal performance. SA2 = BTMSe 92% and BTMSp 98% (Warnick et al., 2006) representing current annual BTM temporal performance. SA3 = BTMSe 88% (Warnick et al., 2006) and BTMSp 99% representing original quarterly BTM temporal performance. SA4 = BTMSe 53% and BTMSp 99.9% representing current quarterly BTM temporal performance (expert opinion).

HSpParallel was 64%. The median NPV was 99%, while the median PPV ranged from 22% (III) to 43% (I-II) (Table 4a).

In LR herds, blood testing was not simulated. Thus, the annual sensitivity, specificity and predictive values did not change compared to the estimates obtained for testing strategy A (Table 4a).

3.3. *False negative and false positive herds considering original annual test performance*

For strategy A, by assuming annual BTMSe 95% and BTMSp 97% (Warnick et al., 2006), a total of 3 FN (in all risk classifications), and 66 (III) or 67 (I-II) FP herds were simulated (Table 5).

For strategy B, 2 (III) or 3 (I-II) FN herds were obtained. In classifications I-II, 100 or 129 FP herds were simulated if cut-off 50 or 25 ODC% were used for blood testing. Whereas in classification III, the total number of FP herds was 169 or 261, respectively. For other differences between testing strategies and risk classifications, see Table 5.

3.4. *Output of sensitivity analysis under varying test performance*

HSe, HSp and predictive values estimated in the sensitivity analysis (SA scenarios) are shown in Table 4b. In this Section, focus is on the estimated number of FN and FP herds under varying test́s performances (Table 5).

In scenario SA1, a total of 3 FN and 67 FP herds were estimated for strategy A, while strategy B led to 3 FN and 74 FP herds (Table 5). The main difference between SA1 and the original scenario under classification II was that, with strategy B, 10 instead of 36 FP high risk herds were estimated in the former (Table 5), because the higher specificity used for individual blood testing (Sp = 99% instead of 95%) increased the HSpParallel from 64% (Table 4a, II) to 90% (Table 4b, SA1).

Scenario SA2 resulted in 1 (B) or 2 (A) FNs more and 22 (B) or 23 (A) FPs less than SA1 (Table 5). Thus, by assuming current lower annual temporal BTMSe (92 vs 95%) and higher BTMSp (98 vs. 97%) a slightly higher number of FN herds, but a remarkably lower number of FP herds were obtained in SA2, for both strategies.

In scenario SA3 (quarterly highest BTMSe = 88% and lowest BTMSp = 99%), a total of 7 FN and 22 FP herds were simulated for strategy A. Whereas testing strategy B resulted in 1 FNs less and 8 FPs more (Table 5).

In scenario SA4 (quarterly lowest BTMSe = 53% and highest BTMSp = 99.9%), 28 FN and 2 FP herds were simulated for strategy A. Whereas testing strategy B resulted in 6 FN less and 8 FP herds more (Table 5).

4. Discussion

In this study, the number of surveillance units (animals and herds) to test, as well as the potential herd status classification errors (FN and FP), were assessed for different combinations of testing strategies, risk-based population classifications and test performances. Simulation outputs were reported in two different forms: (a) as individual (median) HSe, HSp and predictive values (Table 4.a and b) and (b) as national number of FN and FP herds (Table 5). From a managerial perspective, estimating the national number of FNs and FPs, can give clearer information than just focusing on percentage estimates of HSe, HSp and predictive values. For example, showing that the individual HSe increases e.g. from 95 to 98% (Table 4.a first line) when changing from strategy A to strategy B is interesting, because it would mean that the percentage of detected TI high risk herds would increase of 3%. However, such a percentage alone is not enough to show improvement of the system at national level, in terms of absolute number of involved herds and related costs. For this purpose, the HSe and HSp evaluated at individual herd level had to be related to the number of TI and TF herds present in the country, to estimate FN and FP herds, because those two kinds of classification errors determine the actual improvement and sustainability of the of the system in the long run.

Our results could be used to inform eventual improvements of the *S.* Dublin eradication programme, especially to improve early detection. In

Table 5
Number of false negative and false positive level 1 Danish dairy herds within the *S*. Dublin eradication programme according to combinations of: testing strategy, risk-based classification, risk stratum and simulation scenario.

Scenario I and II	HR-B-25	HR-B-50	HR-A	LR-A
FN	1	1	1	2
FP	65	36	3	64
Scenario III	**HR-B-25**	**HR-B-50**	**HR-A**	**LR-A**
FN	1	1	2	1
FP	204	112	9	57
Scenario SA1	**HR-B-25**	**HR-B-50**	**HR-A**	**LR-A**
FN	n.a	1	1	2
FP	n.a	10	3	64
Scenario SA2	**HR-B-25**	**HR-B-50**	**HR-A**	**LR-A**
FN	n.a	1	2	3
FP	n.a	10	2	42
Scenario SA3	**HR-B-25**	**HR-B-50**	**HR-A**	**LR-A**
FN	n.a	2	3	4
FP	n.a	9	1	21
Scenario SA4	**HR-B-25**	**HR-B-50**	**HR-A**	**LR-A**
FN	n.a	7	13	15
FP	n.a	8	0	2

FN = Number of false negative herds; FP = Number of false positive herds. Testing strategy A = all herds (HR = high risk; LR = low risk) tested on BTM only. Testing strategy B = BTM testing in all herds plus blood testing in HR herds. Scenarios I to III are related to estimates presented in Table 4a. All SA scenarios are related to estimates reported in Table 4b, and thus, used the risk-based classification II and individual blood Se = 77% and Sp = 99% for cut-off 50 ODC% (Nielsen and Ersbøll, 2004) (*i.e.* only column HR-B-50 is reported for SA scenarios). SA1 = BTMSe 95% and BTMSp 97% (Warnick et al., 2006) representing original annual BTM temporal performance. SA2 = BTMSe 92% and BTMSp 98% (Warnick et al., 2006) representing current annual BTM temporal performance. SA3 = BTMSe 88% (Warnick et al., 2006) and BTMSp 99% representing original quarterly BTM temporal performance. SA4 = BTMSe 53% and BTMSp 99.9% representing current quarterly BTM temporal performance (expert opinion).

the final phases of an eradication programme, when the prevalence is low, HSe, early detection of newly infected herds and low number of FNs are related to each other and are extremely important, to maintain or optimize the epidemiological status and to reach the final goal of the programme. However, early detection is also challenging, because rare infections become more difficult to find as prevalence reduces. Then, herds misclassified as negative (due to HSe < 100%) while in fact infected with *S*. Dublin (FN), could spread disease during the HRP, and can have important consequences for the eradication progress, for the income of farmers and for human health. This study showed how risk-based surveillance could increase the temporal HSe and the related NPV in large HR herds, and thus, could reduce the number of FNs. At the same time, it was evaluated how FPs could be affected, because cost-efficiency of the system can be hampered by under-testing as well as by over-testing and imposing unnecessary control actions on FP farms.

4.1. Impact of risk classification on managerial applicability of risk-based surveillance

By definition, the application of risk-based surveillance relies on an adequate characterisation of the different populations ́ risk strata. The estimates of the individual herds' probabilities of infection and of the relative risk were based on actual national data and on risk factors quantified from the same population and time frame. In line with Martin et al. (2007a), (2007b), the RR inputs were combined with the respective PrPs of farms to which they applied (Table 2), to estimate the within-stratum effective probability of infection: namely EPI_{LR} and EPI_{HR} (Appendix A). Thereafter, the total 59 TI herds simulated from the national median ProbInf were split across the two strata, in a standardized and objective manner, according to each EPI value (Table 2; Appendix A). Through the application of this method of surveillance evaluation, the importance of the applied risk-based classification can be clearly distinguished from the effects of the investigated sampling strategy and test (Martin et al., 2007a; 2007b; Foddai et al., 2020).

Classification I divided the 2,283 level 1 herds across the two risk strata, by using as cut-off the 95^{th} percentile ProbInf, which was estimated for each herd by logistic regression. Whereas classification II was based on herd size and on number of infected neighbours, and classification III was based on size and region of location. The risk factor analysis showed that these three variables were associated with the ProbInf, but were more directly available from the Danish Cattle Database, and thus, are more practical than ProbInf to be used in the programme. However, the number of herds and calves to test, and consequently the number of FP herds, appeared by far higher in classification III than in classifications I-II (Table 2), while the number of FN was similar (Table 5). Thus, the practicality-based classification III was not considered further.

In classification II (applied also to all SA scenarios), the allocation of herds into the high and low risk strata completely matched that of classification I (Table 2). In those cases, the number of herds (127) and calves (966) to test in blood samples would be targeted towards very large HR herds. LR herds were approximately half the size of HR herds. Hence, in the latter, disease detection by BTM testing could require longer time, and several animals could be moved/sold during the HRP. Moreover, in classifications I and II, a similar number of TI herds was allocated between the two strata (27 as HR vs. 32 as LR), while the total number of LR herds was approximately 17 times larger than the number of HR herds (Table 2). The different PrP of herds allocated in each stratum, combined with the respective RR of infection, led very different within-stratum effective probabilities of infection (Table 2: $EPI_{HR} > EPI_{LR}$). Consequently the number of (simulated) TI herds related with each EPI value suggested that, if any of these two risk-based classifications (I-II) is implemented in the programme, blood testing could be targeted efficiently to (at least) half of the TI herds present in the country (see TIs in Table 2).

Classification II appeared the best for combining: managerial applicability, costs of blood testing and improved early detection of (several) TI large herds, while minimizing classification errors.

4.2. Combined effects of test strategies and risk-based classifications on surveillance outputs

The high annual NPV did not change remarkably across combinations of testing strategies and risk classifications, because in Eq. 5 a high annual HSe (≥ 95%, Table 4a,b) was combined with a low effective probability of infection, EPI_{HR} or EPI_{LR} (Table 2). This was the case especially within the LR stratum, where the EPI_{LR} was 1.3–1.5% and only BTM testing with high BTMSe (Warnick et al., 2006), were simulated for both strategies (Table 4a,b). The NPV tends to increase if the probability of infection reduces and/or if herd sensitivity is maximized.

As argued by Warnick et al. (2006), high NPV and relatively low PPV, like those estimated in this study (Tables 4a,b), would be consistent with the principal surveillance programme goal of reaching high confidence in a negative test result, to avoid secondary disease spread from FN farms. The PPV differed between strata, even within strategy A, where the same BTMSe and BTMSp were used for all herds. This happened because within the HR stratum, the median EPI was more variable across classifications (Table 2) and was higher than that of the LR stratum, yielding higher PPV for the HR herds than for the LR herds (according to Eq. 6).

In HR herds, strategy B caused a small increase of the HSeParallel (97–98% vs BTMSe 95%), but a remarkable decrease in HSpParallel (35–64 vs BTMSp 97%) compared to strategy A (Table 4a). Accordingly, also the median PPV were lower for strategy B. If this testing strategy is applied, cut-off 50 ODC% could be preferred for individual blood testing, because it can cause less FP herds than using 25 ODC%, while the number of FNs would be similar with the two cut-offs (Table 5; classifications I to III).

The LR stratum contained the highest number of FP herds with both testing strategies (Table 5), because despite the high annual BTMSp (97%), this stratum represented the majority of level 1 TF herds (Table 2). Thus, although the probability of resulting FP (according to Eq. 8) was small (3%), it applied to most of the herds (Table 2).

The deterministic model allowed understanding to what extent the herd status classification errors were influenced by: test performance, testing strategies and risk-based classification.

4.3. Impact of time and test performance

In all four sensitivity analysis scenarios (SA1, SA2, SA3 and SA4), the BTMSe was reduced, while the BTMSp was increased (Table 4b), compared to the inputs used in the basic classification scenarios I to III (Table 4a). Moreover, annual and quarterly HRPs were considered, because within infected herds, combinations of: time from disease introduction, herd size and management; can affect within-herd daily disease transition-states dynamics and related temporal herd sensitivity (Thurmond, 2003; Foddai et al., 2014; Foddai et al., 2016). From a general point of view, the antibody ELISA used for BTM testing is likely to be more sensitive after long than after short HRPs. Usually, the longer the time elapsed from day of disease introduction, the higher the seroprevalence reached within the milking group (Foddai et al., 2014), which can cause increases of the antibody titres in BTM by the day of testing. At the same time, it must be kept in mind that the BTM sample represents (mainly) the epidemiological status of the lactating cows on the day of sampling, and between two or more samplings, the composition of the group can change remarkably. Especially in large herds, a high number of uninfected recently calved cows could be introduced to the milking parlour within a few days. At the same time, antibody positive cows could be moved to the dry-off group and no longer contribute milk to the BTM sample. Any of those within-herd movements can cause sudden antibodies fluctuations in BTM, not least at the beginning of the herd́s infection period, when only a few animals have seroconverted.

Scenarios SA1 and SA2 reflected test performance after annual HRPs and showed that, most herds infected for at least one year were very unlikely to be missed at the 4th BTM testing round, when temporal BTMSe = 92% or 95% were assumed (Table 4a, basic classification scenarios I to III; Table 4b, scenario SA1-2). Paradoxically, if high sensitivity is always assumed for BTM testing, then additional blood testing of calves (*i.e.* changing from strategy A to strategy B) appears disadvantageous; because apart from the additional costs of sampling and testing, the number FPs could increase remarkably (Table 5).

In contrast, in scenarios SA3 and SA4, which only included a single BTM result, the change from strategy A to strategy B, led to an evident reduction in FN high risk herds and a more modest increase in FPs, compared to results of SA1 and SA2 (Table 5). Hence, if low BTMSe is assumed for quarterly testing (*i.e.* by three months from disease introduction into the herd), additional blood testing could improve the chances of detecting large HR herds that have been recently infected, while the chances of TF herds resulting FP could be minimized.

At the same time, it could be argued that if SA4 was assumed as the most realistic scenario (with the lowest quarterly BTMSe and the highest BTMSp), and if strategy B was implemented quarterly, a total of 28 FN high risk herds could be still obtained during a year; because 7 were estimated for a single quarter (Table 5, SA4). However, this might not be the case, because the higher number of FNs "avoided" after each single quarter, by applying strategy B instead of strategy A, should consequently reduce the number of secondary cases in the following quarters, and thus, should decrease the total annual FNs as well. For these reasons, comparisons between testing strategies were made only under the same HRP (*i.e.* SA1 vs. SA2 and SA3 vs. SA4), while comparisons between testing strategies referring to different surveillance periods (e.g. SA2 vs. SA4) should be avoided or made with caution.

4.4. Interpretation of outputs considering epidemiological context and tolerated high-risk periods

For interpreting and using the outputs described above, the current epidemiological context (*i.e.* the actual herd prevalence), the population/herd structure and the tolerated HRP for detecting infected herds with the assumed BTMSe, should be considered. In fact, those factors affect the interpretation of the BTM values, and thus, the related pros and cons of each testing strategy (A vs. B). For the Danish programme, the current main aims are: improving early detection of newly infected herds, stopping re-increase of prevalence (e.g. due to secondary disease spread), and finalizing eradication as soon as possible.

When the iELISA was validated for BTM testing, Warnick et al. (2006) explained that, the probability of an infected herd classifying BTM positive was conditional on a 1-year disease history, because the surveillance programme classifies herds from four sequential measurements, which are taken at 3-months intervals. We followed the same principle when annual HRPs were used. At the same time, it was also clarified that: on one hand, depending on the correlation between measurements from the same herd, applying test criteria based on the average of repeated BTM samples, can reduce the variability of the test parameter and thereby can increase testing accuracy. On the other hand, this benefit must be weighed against the disadvantage of errors, which can result from the influence of past herd́s status and previous test results on current classification (Warnick et al., 2006).

The latter point means that, if strategy A is used as dependent of four tests, an example herd "X" which has been TI for a year and has had BTM values = 10,15,20, and 30 ODC%, would maintain the free status, because both criteria to be classified in level 1 (Section 2.1), would be fulfilled. In contrast, the same herd would lose the free status, if only the 4th BTM value was considered and if cut-off 20 or 25 ODC% was used for the single milk sample classification.

At the beginning of the surveillance programme, situations such as herd "X" were less likely to occur than in the current situation, because increases of BTM values could be noted more promptly, even earlier than a year from disease introduction. In other words, the increase of 20 ODC% above the average of the previous measurements, could happen more easily due to lower antibody dilution. In 2006, when the study by Warnick et al. was carried out, the Danish dairy herds had a median size around 100 cows (Foddai et al., 2015), while in the current situation, higher antibody dilutions could be expected in HR herds, which are approximately three times bigger (Table 2). Hence in the current context, the original testing interpretation based on four consecutive BTM tests, could allow disease detection after relatively long infection (e.g. after one year or even more). Such a duration (or HRP) could be "too long", especially for the very large HR herds, which could spread the disease to other farms meanwhile.

In the initial phases of eradication programmes, when herd prevalence and PPV are high, it might be sufficient restricting a small percentage of TI herds, to show evident decreases of prevalence at country level, whilst limiting the number of herds under restriction. Unnecessary control measures are disruptive for farmers and could affect their willingness to engage in the eradication scheme, whereas showing efficient decreases of prevalence (within short periods) can improve participation. When the iELISA was validated, the herds classification based on four consecutive BTM tests, offered the best compromise between the pros and cons mentioned above. This is evident by the steady reduction in prevalence observed from 2003 to 2015 (SEGES, 2021). Nowadays, with increased herd size and prevalence, a higher quarterly sensitivity (through additional blood testing) combined with an interpretation of BTM values based on a single quarter, may offer improvement of early detection for infected large HR herds. This could improve the early warning attribute (timeliness) of the system, and could minimize secondary disease spread (epidemiological consequences) from this kind of newly infected farms, while tolerating a small increase of FP errors (Table 5, SA4).

4.5. Limitations of the study

Alternative risk-based classifications of herds could have been applied. For example, risk-based classifications that used the 3rd quartile or the 90th percentile ProbInf as cut-off between strata were investigated at the beginning of the study. In those cases, the number of herds classified at HR and the number of calves to test from them, were by far higher than those estimated for classifications I to III (Table 2). The number of HR herds would have been 536 and 318 respectively, and the number of calves to test would have been 4,010 and 2,214. Such classifications would have the advantage that most of the simulated 59 TI herds would be targeted by blood testing (45 and 37, respectively), but costs and number of FPs would be much higher (results not shown).

The number of FN and FP herds could have been estimated in two different ways; i) using the predictive values, or ii) using the HSe and HSp, as we did in this study Eqs. (7) and ((8)). If data on herds tested with both strategies (A and B) had been available, the number of FNs could have been estimated by multiplying the number of herds that tested negative in each strategy, by the complementary probability of the negative predictive value (1-NPV). The number of FP herds could have been estimated by multiplying the number of test-positive herds in each strategy by the complementary probability of the positive predictive value (1-PPV). Since strategy B was a hypothetical sampling strategy, actual testing results from all 2,283 level 1 herds were only available for strategy A. Thus, the number of FNs and FPs were estimated by combining the complementary probability of the median HSe and HSp, with the simulated number of (within-stratum) TI and TF herds Eqs. (7) and ((8))

Regarding the performance of the iELISA in BTM, we could not exclude cross-reactions to *S.* Typhimurium or other serovar in herds actually free of *S.* Dublin. Nevertheless, changes of BTM specificity were simulated as inversely related to: a) the BTMSe and b) the time elapsed between samplings (HRP). As the HRP increased the BTMSp decreased, and a truly *S.* Dublin free herd was (assumed) more likely to be classified as FP due to lowered temporal specificity. This assumption represented the situations where, a dairy herd which is truly free of *S.* Dublin but is infected with *S.* Typhimurium (or others) increases its chances of being "wrongly" classified positive (*i.e.* FP) to *S.* Dublin, with longer HRP; *i.e.* when the antibody prevalence against the alternative serovar of *Salmonella* increases in the milking group.

The uncertainty around the iELISA specificity in individual blood could be considered problematic, because a small increase of animal level specificity, as that simulated passing from the original scenario in classification II (Sp = 95%) to SA1 (Sp = 99%), generated a remarkable decrease of FP high risk herds (-26) in strategy B (Table 5). Therefore, the number of FP estimated in the original classifications I to III (Table 5) could be overestimated, if the Sp = 99% is the closest value to the true Sp between those considered (Nielsen at al., 2004; Nielsen and Ersbøl, 2004). This uncertainty was taken into account in the sensitivity analysis.

The deterministic simulation modelling approach investigated the uncertainty around the final outputs by using different inputs for both BTM and blood testing across scenarios, instead of using different iterations and distributions within the same simulation, as it could be the case with stochastic simulation models, where outputs can be produced with their respective prediction intervals. On the other hand, in models composed of several variables (like ours), the running time of the stochastic tools could be by far longer than in deterministic models and issues in separating impact of variability and uncertainty could arise. Moreover, when outputs of stochastic models are used to inform decision making, focus is usually addressed to the median/mean estimates of the simulations, rather than to their prediction intervals. Hence, the main results (*i.e.* the simulated median sensitivity, specificity, predictive values and classification errors) could be expected similar between the two modelling procedures. Since we expect the main outputs would have been similar between stochastic and deterministic modelling, and because variability and uncertainty were investigated in the sensitivity analysis, we preferred to pursue the model̓s parsimony principle when choosing the deterministic approach.

In addition, it must be noted that in each scenario, the temporal BTMSe and BTMSp were simulated with similar values for HR and LR herds. In reality, further differences could be present between herds of several different sizes, within and between risk strata. From that point of view, a simplification was applied due to missing knowledge on how seroprevalence could vary daily within the milking paddock of herds with different size and risk. This uncertainty could be reduced through disease-spread simulation studies (see perspectives below). However, we would expect that, since HR herds had median number of cows by far bigger than LR herds (Table 2), if the two herd types had been simulated with different BTMSe (higher for low risk herds) and BTMSp (higher for high risk herds), the benefits of changing from strategy A to B could have appeared even clearer.

Furthermore, the variability of herd structure and size were partly included in the model by simulating all 2,283 level 1 dairy herds. Although currently the within-groups animal prevalence could be lower than those estimated by Nielsen, (2013b), we expect the relative risk of infection between age groups (the RRg values in the Appendix B) to be similar, due to similar dairy herd structure.

Regarding the animal level design prevalence WGP and WHP, they were both set to 10%. This value was selected based on our knowledge of the epidemiology of *S.* Dublin in Danish dairy herds, and falls within the general used cut-off between 1 and 10% (Martin et al., 2007b; Cameron, 2014), e.g. when the design prevalence is not defined as an input based standard by legislation. Across the 2,283 herds, the overall median number of old calves (3–6 months old) was 20, which should allow good timeliness of detection, if this group of animals is infected *i.e.* 10% WGP = 2 infected calves (d). In contrast, in large herds with 200 or more milking cows (Table 2), it would take longer to reach the detection limit of 10% antibody positives adults ($d = 20$). For that reason, the performance of the iELISA on BTM samples was challenged in the sensitivity analysis.

4.6. Perspectives

The impact of the FN and of FP herds estimated in this study could be further assessed before deciding, which testing strategy to use in the Danish *Salmonella* Dublin eradication programme. For both strategies, disease spread simulation models could evaluate variability in temporal within-herd disease dynamics and the effects on surveillance outcomes, and could estimate the epidemiological consequences (e.g. in number of secondary cases) due to disease spread from FN herds during the HRP. Cost-benefit analysis is recommended to investigate the economic consequences of secondary cases caused by FN herds and of false alarms in TF herds.

5. Conclusion

Our study found that the Danish *Salmonella* Dublin eradication programme could benefit from adding blood testing of calves in large HR herds to the current BTM testing strategy, if HR herds were classified as those having ≥ 8 neighbours in level 2 and ≥ 200 cows (≈ 530 cattle). This could improve the timeliness of the system by earlier disease detection in these herds (e.g. within three months from infection). Consequently, the number of FN herds would be reduced compared to the current situation, where only BTM testing is used in level 1 herds. Nevertheless, the current strategy will classify a lower number of FP herds, while the alternative testing strategy based on additional blood testing, could result in more falsely restricted HR herds than today. Costs for additional blood testing and extra FPs identified in the alternative strategy, should be balanced against cost-of-error, e.g. the potential "extra" disease spread from FN herds, if the current testing scheme is maintained. This work paves the road for further studies of disease spread simulation modelling and cost-benefit analysis, which could support a final decision on which testing strategy to prioritise, to finalise disease eradication in the shortest time and in the most cost-efficient manner.

CRediT authorship contribution statement

Alessandro Foddai: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Jørgen Nielsen:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – review & editing. **Liza Rosenbaum Nielsen:** Conceptualization, Investigation, Methodology, Supervision, Writing – review & editing. **Erik Rattenborg:** Conceptualization, Funding acquisition, Investigation, Resources, Validation, Writing – review & editing. **Hans Ebbensgaard Murillo:** Conceptualization, Funding acquisition, Investigation, Resources, Validation, Writing – review & editing. **Johanne Ellis-Iversen:** Conceptualization, Funding acquisition, Investigation, Methodology, Resources, Supervision, Validation, Writing – review & editing.

Declaration of Competing Interest

During the study AF and JN were only involved as scientists. LRN, ER, HE, and JEI were further engaged in the programme working group, chaired by the Danish Veterinary and Food Administration, dealing with the continuous evaluation and development of the *S.* Dublin surveillance and eradication programme in Denmark. None of the authors received personal or financial benefits in this function and declare that they have no competing interests related to the presented work

Acknowledgments

The authors are grateful to the Danish Cattle Federation (SEGES) for providing the data on surveillance of *Salmonella* Dublin in Danish dairy herds and to the the Ministry of Environment of Denmark and Ministry of Food, Agriculture and Fisheries of Denmark, for funding the modelling work.

Appendix of: Evaluation of risk-based surveillance strategies for *Salmonella* Dublin in Danish dairy herds by modelling temporal test performance and herd status classification errors

List of abbreviations

Abbreviation	**Meaning**
ARR$_{\text{HR}}$	Adjusted relative risk of herd infection within the high risk stratum
ARR$_{\text{LR}}$	Adjusted relative risk of herd infection within the low risk stratum
BTM	Bulk tank milk
CHR	Cattle herd identification number
d	Number of simulated truly infected animals per group
EPIg	Within-group effective probability of infection
EPI$_{\text{HR}}$	Effective probability of herd infection within the high risk stratum
EPI$_{\text{LR}}$	Effective probability of herd infection within the low risk stratum
FN	False negative herd
FP	False positive herd
GSeOlderCalves	Sensitivity obtained from testing just calves 3-6 months old when testing strategy B is used
HR	Herd at high risk of infection
HRP	High risk period elapsing between day of disease introduction to the herd and day of its detection
HSe	Herd sensitivity
HSeParallel	Overall herd sensitivity (HSe) in high risk herds when testing both calves and BTM, in testing strategy B
HSp	Herd specificity
HSpParallel	Overall herd specificity (HSp) in high risk herds when testing both calves and BTM, in testing strategy B
iELISA	Indirect Enzyme-Linked Immunosorbent Assay
LR	Herd at low risk of infection
NPV	Negative predictive value
ODC%	Background corrected proportion of the test sample optical density (OD) to a known positive reference sample
PPV	Positive predictive value
ProbInf	Probability the herd was infected in reality, estimated through logistic regression
ProbInfAnimal	Individual animal probability to be infected within an age group
PrP$_{\text{HR}}$	Proportion of herds in the high risk stratum
PrP$_{\text{LR}}$	Proportion of herds in the low risk stratum
PrPanimals	Proportion of animals per age group within a herd
RR	Relative risk of herd infection per population stratum
RRg	Relative risk of animal infection within an age group
RR$_{\text{HR}}$	Risk of herd infection within the high risk stratum relative to the risk of infection within the low risk stratum
RR$_{\text{LR}}$	Relative risk of herd infection within the low risk stratum (set as 1, for risk reference category)
***S.* Dublin**	*Salmonella enterica* subsp. *enterica* serovar Dublin
SA1, SA2, SA3 and SA 4	Sensitivity analysis scenarios 1, 2, 3 and 4
SEGES	Danish Agriculture & Food Council SEGES
TF	Truly disease free herd
TI	Truly infected herd
WGP	Within group prevalence
WHP	Within herd prevalence

A. Estimating effective probability of herd infection within each stratum (EPI$_{\text{HR}}$ and EPI$_{\text{LR}}$)

The effective probability of herd infection (Martin et al., 2007a; 2007b), within each herd risk stratum (LR = low risk and HR = high risk), was estimated as:

$$\text{EPI}_{\text{LR}} = \text{ProbInf} * \text{ARR}_{\text{LR}} \tag{A.1}$$

$$EPI_{HR} = ProbInf * ARR_{HR} \tag{A.2}$$

Where, ProbInf represented the national median probability of herd infection (2.6%) reported in Section 2.1.3 and estimated through logistic regression (Section 2.1). Whereas $\mathbf{ARR_{LR}}$ and $\mathbf{ARR_{HR}}$ represented the adjusted relative risk of infection within the LR and the HR stratum respectively and were calculated as:

$$ARR_{LR} = 1 / (PrP_{LR} + PrP_{HR} * RR_{HR}) \tag{A.3}$$

$$ARR_{HR} = ARR_{LR} * RR_{HR} \tag{A.4}$$

The PrP_{LR} and PrP_{HR} were the proportions of cattle herds within each population stratum out of the total 2283 level 1 dairy herds. Whereas, RR_{LR} and RR_{HR} were the relative risk of infection in the LR and HR strata (Table 2). The RR_{LR} was set = 1 because the LR herds had lower risk of infection than HR herds, and thus, the former represented the risk reference category. Whereas the RR_{HR} was calculated as: the median ProbInf within the HR stratum divided by the median ProbInf within the LR stratum.

B. Estimating effective probability of animal infection within herds infected in multiple groups

For the 1783 dairy herds, where the (potential) overall within-herd design prevalence (WHP) was split across different age groups, we applied the same principles used at between-herds level, to estimate EPI_{HR} and EPI_{LR} (Martin et al., 2007a; 2007b), but now considering: the proportion of animals located within each age group (PrPanimals) and the respective individual relative risk of infection (RRg).

Accordingly, the within-group effective probability of infection (EPIg) was firstly estimated for each age group. Thereafter, the number of seropositive animals was simulated for each age group as d = rounded (EPIg * group size), so that the sum of all d values corresponded to the total number of seropositive animals simulated within the herd, according to WHP = 10% and herd size (in total cattle) reported in the data.

Regarding the RRg values, it must be noted that young calves 0–3 months old were considered as the risk reference category with RRg = 1, because in this age group the seroprevalence is usually very low or not detectable (Nielsen, 2013b). Whereas the RRg within the other age groups, was calculated using the mean seasonal prevalence (here called ProbInfAnimal) reported by Nielsen (2013b). This was set at: 28.3%, 27.0% and 31.3%; for old calves, heifers-steers, and cows. For calves younger than three months, the ProbInfAnimal was set at 15.7% (the mid value between 0% and the value used for cows). Then, the RRg inputs used for older calves, heifers-steers and cows were (approximated): (28.3 / 15.7) = 1.8, (27.0 / 15.7) = 1.7 and (31.3/ 15.7) = 2.0, respectively.

References

Alban, L., Häsler, B., van Schaik, G., Ruegg, S., 2020. Risk-based surveillance for meat-borne parasites. Exp. Parasitol. 208, 107808 https://doi.org/10.1016/j.exppara.2019.107808.

Anonymous, 2004. Annual Report on Zoonoses in Denmark 2003. Ministry of Food, Agriculture and Fisheries. Pp 1-32. Available at. https://www.food.dtu.dk/english/publications/disease-causing-microorganisms/zoonosis-annual-reports. Accessed on 10 May 2021.

Anonymous, 2019. Annual Report on Zoonoses in Denmark 2018. National Food Institute, Technical University of Denmark, pp. 1–61. Available at. https://www.food.dtu.dk/english/publications/disease-causing-microorganisms/zoonosis-annual-reports. Accessed on 10 May 2021.

Cameron, A.R., 2012. The consequences of risk-based surveillance: developing output-based standards for surveillance to demonstrate freedom from disease. Prev. Vet. Med. 105, 280–286. https://doi.org/10.1016/j.prevetmed.2012.01.009.

Cameron, A., Njeumi, F., Chibeu, D., Martin, T., 2014. Risk Based Disease Surveillance. A Manual for Veterinarians on the Design and Analysis of Surveillance for Demonstration of Freedom from Disease, 2014. Food and Agriculture Organization of the United Nations, pp. 1–197.

Danish Agriculture & Food Council, 2020. Statistics 2019, dairy, June 2020. Pp. 1-45.

Danish order no. 1326 of 29/11/2017 (in Danish). BEK nr 1326 af 29/11/2017 (Historisk). Lovtidende A 2017. Bekendtgørelse om salmonella hos kvæg m.m. 29 November 2017; No. 1326. Pp. 1-21. https://www.retsinformation.dk/eli/lta/2017/1326. Accessed on 10 May 2021.

Fierer, J., 1983. Invasive *Salmonella* Dublin infections associated with drinking raw milk. West. J. Med. 138 (5), 665–669.

Foddai, A., Enøe, C., Krogh, K., Stockmarr, A., Halasa, T., 2014. Stochastic simulation modeling to determine time to detect bovine viral diarrhea antibodies in bulk tank milk. Prev. Vet. Med. 117, 149–159. https://doi.org/10.1016/j.prevetmed.2014.07.007.

Foddai, A., Enøe, C., Stockmarr, A., Krogh, K., Uttenthal, Å., 2015. Challenges for bovine viral diarrhoea virus antibody detection in bulk milk by antibody enzyme-linked immunosorbent assays due to changes in milk production levels. Acta Vet. Scand. 32 ((2015), 1–8. https://doi.org/10.1186/s13028-015-0125-z.

Foddai, A., Stockmarr, A., Boklund, A., 2016. Evaluation of temporal surveillance system sensitivity and freedom from bovine viral diarrhea in Danish dairy herds using scenario tree modelling. BMC Vet. Res. 118 (2016)), 1–12. https://doi.org/10.1186/s12917-016-0744-2.

Foddai, A., Floyd, T., McGiven, J., Grace, K., Evans, S., 2020. Evaluation of the English bovine brucellosis surveillance system considering probability of disease introduction and non-random sampling. Prev. Vet. Med. 176, 1–14. https://doi.org/10.1016/j.prevetmed.2020.104927.

Hansen, R.K., Nielsen, L.H., El Tholth, M., Haesler, B., Foddai, A., Alban, L., 2018. Comparison of alternative meat inspection regimes for pigs from non-controlled housing - considering the cost of error. Front. Vet. Sci. 5, 92. https://doi.org/10.3389/fvets.2018.00092.

Harvey, R.R., Friedman, C.R., Crim, S.M., Judd, M., Barrett, K.A., Tolar, B., Folster, J.P., Griffin, P.M., Brown, A.C., 2017. Epidemiology of *Salmonella enterica* serotype Dublin infections among humans, United States, 1968–2013. Emerg. Infect. Dis. 23 (9), 1493–1501. https://doi.org/10.3201/eid2309.170136.

Helms, M., Vastrup, P., Gerner-Smidt, P., Mølbak, K., 2003. Short and long term mortality associated with foodborne bacterial gastrointestinal infections: registry based study. BMJ 326, 1–5. https://doi.org/10.1136/bmj.326.7385.357.

Hoorfar, J., Feld, N.C., Schirmer, A.L., Bitsch, V., Lind, P., 1993. Serodiagnosis of *Salmonella* Dublin infection in Danish dairy herds using O-antigen based enzyme-linked immunosorbent assay. Can. J. Vet. Res. 57, 268–274.

Hoorfar, J., Lind, P., Bitsch, V., 1995. Evaluation of an O antigen enzyme-linked immunosorbent assay for screening of milk samples for *Salmonella* Dublin infection in dairy herds. Can. J. Vet. Res. 59, 142–148.

Horst, H.S., Huirne, R.B.M., Dijkhuizen, A.A., 1997. Risks and economic consequences of introducing classical swine fever into the Netherlands by feeding swill to swine. Rev. Sci. Tech. Off. Int. Epiz. 16, 207–214. https://doi.org/10.20506/rst.16.1.1004.

Hosmer, D.W., Lemeshow, S., 1989. Applied Logistic Regression, Chapter 3. Interpretation of Coefficients. Wiley.

Martin, P.A.J., Cameron, A.R., Greiner, M., 2007a. Demonstrating freedom from disease using multiple complex data sources 1: a new methodology based on scenario trees. Prev. Vet. Med. 79, 71–97. https://doi.org/10.1016/j.prevetmed.2006.09.008.

Martin, P.A.J., Cameron, A.R., Barfod, K., Sergeant, E.S.G., Greiner, M., 2007b. Demonstrating freedom from disease using multiple complex data sources 2: case study classical swine fever in Denmark. Prev. Vet. Med. 79, 98–115. https://doi.org/10.1016/j.prevetmed.2006.09.007.

MacDiarmid, S.C., 1988. Future options for brucellosis surveillance in New Zealand beef herds. N. Z. Vet. J. 36, 39–42. https://doi.org/10.1080/00480169.1988.35472.

Nielsen, L.R., Ersbøll, A.K., 2004. Age-stratified validation of an indirect *Salmonella* Dublin serum enzyme-linked immunosorbent assay for individual diagnosis in cattle. J. Vet. Diagn. Invest. 16, 212–218. https://doi.org/10.1177/104063870401600306, 10.1177%2F104063870401600306.

Nielsen, L.R., Toft, N., Ersbøll, A.K., 2004. Evaluation of an indirect serum ELISA and a bacteriological faecal culture test for diagnosis of *Salmonella* serotype Dublin in cattle using latent class models. J. Appl. Microbiol. 96, 311–319. https://doi.org/10.1046/j.1365-2672.2004.02151.x.

Nielsen, L.R., Dohoo, I., 2012. Survival analysis of factors affecting incidence risk of *Salmonella* Dublin in Danish dairy herds during a 7-year surveillance period. Prev. Vet. Med. 107, 160:–169. https://doi.org/10.1016/j.prevetmed.2012.06.002.

Nielsen, T.D., Green, L.E., Kudahl, A.B., Østergaard, S., Nielsen, L.R., 2012. Evaluation of milk yield losses associated with *Salmonella* antibodies in bulk tank milk in bovine dairy herds. J. Dairy Sci. 95, 4873–4885. https://doi.org/10.3168/jds.2011-4332.

Nielsen, L.R., 2013a. *Salmonella* Dublin in cattle. Epidemiology, design and evaluation of surveillance and eradication programmes. Dr. med. vet. thesis 2013 Faculty of Health and Medical Sciences. University of Copenhagen, pp. 1–396.

Nielsen, L.R., 2013b. Within-herd prevalence of *Salmonella* Dublin in endemically infected dairy herds. Epidemiol. Infect. 141, 2074–2082. https://doi.org/10.1017/S0950268812003007.

Noordhuizen, J.P.T.M., Frankena, K., Thrusfield, M.V., Graat, E.A.M., 2001. Application of Quantitative Methods in Veterinary Epidemiology, 2nd ed. Wageningen: Wageningen Pers.

Richardson, A., Watson, W.A., 1971. A contribution to the epidemiology of *Salmonella* Dublin infection in cattle. Br. Vet. J. 127, 173–183. https://doi.org/10.1016/S0007-1935(17)37634-0.

R Core Team, 2013. http://www.R-project.org/ Accessed on 10 May 2021.

SEGES, 2021. http://www.kvaegvet.dk/Dublin/AAHistNivPlot.html. Accessed on 10 May 2021.

Stärk, K.D.C., Regula, G., Hernandez, J., Knopf, L., Fuchs, K., Morris, R.S., Davies, P., 2006. Concepts for risk-based surveillance in the field of veterinary medicine and veterinary public health: review of current approaches. BMC Health Serv. Res. 6 (20), 1–8. https://doi.org/10.1186/1472-6963-6-20, 10.1186%2F1472-6963-6-20.

Thurmond, M.C., 2003. Special article. Conceptual foundations for infectious disease surveillance. J. Vet. Diagn. Invest. 15, 501–514. https://doi.org/10.1177/104063870301500601.

Warnick, L.D., Nielsen, L.R., Nielsen, J., Greiner, M., 2006. Simulation model estimates of test accuracy and predictive values for the Danish *Salmonella* surveillance program in dairy herds. Prev. Vet. Med. 77, 284–303. https://doi.org/10.1016/j.prevetmed.2006.08.001.