Harmonisation of the diagnostic performances of serological ELISA tests for C. burnetii in ruminants in the absence of a gold standard: Optimal cut-offs and performances reassessment

Laureline Rivière, Elodie Rousset, Elsa Jourdain, Marie-Laure Delignette-Muller, Thibaut Lurier

To cite this version:

Laureline Rivière, Elodie Rousset, Elsa Jourdain, Marie-Laure Delignette-Muller, Thibaut Lurier. Harmonisation of the diagnostic performances of serological ELISA tests for C. burnetii in ruminants in the absence of a gold standard: Optimal cut-offs and performances reassessment. Preventive Veterinary Medicine, 2025, 239, pp.106509. ⟨10.1016/j.prevetmed.2025.106509⟩. ⟨hal-05017943⟩

HAL Id: hal-05017943
https://hal.inrae.fr/hal-05017943v1

Submitted on 2 Apr 2025

Preventive Veterinary Medicine 239 (2025) 106509

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

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

Harmonisation of the diagnostic performances of serological ELISA tests for *C. burnetii* in ruminants in the absence of a gold standard: Optimal cut-offs and performances reassessment

Laureline Rivière [a,b,c,d,*], Elodie Rousset [d], Elsa Jourdain [a], Marie-Laure Delignette-Muller [c], Thibaut Lurier [a,b]

[a] *Université Clermont Auvergne, INRAE, VetAgro Sup, UMR EPIA, Saint-Genès-Champanelle 63122, France*
[b] *Université de Lyon, INRAE, VetAgro Sup, UMR EPIA, Marcy l'Etoile 69280, France*
[c] *Université de Lyon, Université Lyon 1, VetAgro Sup, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne 69622, France*
[d] *ANSES, Sophia Antipolis Laboratory, Animal Q fever Unit, Sophia Antipolis, France*

ARTICLE INFO

Keywords:
Q fever
Agreement
Hubert's kappa
Sensitivity
Specificity
Prevalence

ABSTRACT

None of the three ELISA tests currently available in Europe for the serological diagnosis of *Coxiella burnetii* (*C. burnetii*) infections in ruminants can be considered as a gold standard. The difference in their diagnostic performances (e.g., from 39 % to 87 % for sensitivity in sheep – Lurier *et al.* 2021) affects the agreement between the test results obtained in different veterinary laboratories and limits the comparability of prevalence estimates for surveillance. In order to harmonise these tests, we tried to adjust their cut-offs so that there is a maximum of agreement between them. Using the results of the three ELISAs applied to 1258 cattle, 1474 goat and 1432 sheep serum samples collected in France, we identified the cut-offs that maximised Hubert's kappa, an agreement coefficient applicable to more than two tests, using a differential evolution algorithm for optimisation. We then evaluated the effect of changing the cut-offs. In particular, we estimated the sensitivities and specificities of each test at the optimal cut-offs using a latent class model. While the manufacturer's cut-offs of the three tests were 30, 40 and 40 respectively, regardless of the species, the cut-offs that maximise the agreement are 44.7, 30.8 and 84.3 for cattle, 6.6, 18.8 and 49.8 for goats, and 26.2, 50.9 and 88.6 for sheep. By using these cut-offs instead of those provided by the manufacturers, the observed proportion of disagreement between the tests is reduced by approximately half, and the diagnostic performances and apparent prevalence are more consistent from one test to another. The use of these species-specific cut-offs allows for better harmonisation of the tests. However, before implementing these new cut-offs, it is essential to assess the measurement uncertainty around them to ensure that the analytical performances of the tests are maintained.

1. Introduction

Q fever is a zoonotic disease caused by the bacterium *Coxiella burnetii* (*C. burnetii*). In ruminants, it can cause abortions, premature births and dead or weak offspring. In humans, it can cause flu-like syndrome, pneumonia, hepatitis, vascular infections or chronic fatigue. Most human cases are related to an exposure to domestic ruminants (WOAH, 2018). Surveillance of *C. burnetii* infections is essential for monitoring the infection in ruminants and preventing its transmission to humans. Since 2021, the reporting and monitoring of Q fever cases in cattle, goats and sheep is mandatory (as category E) in the European Union (Animal Health Law, Regulation (EU) 2016/429 on transmissible animal diseases).

Most frequently, the epidemiological surveillance of Q fever relies on serology to identify herds infected by or exposed to *C. burnetii*. Although an intracellular pathogen such as *C. burnetii* is typically expected to be countered primarily by a cell-mediated immune response, antibodies have also been shown to play a significant role in combating *C. burnetii* infection. Studies in B-cell-deficient mice have shown that antibodies act in synergy with cellular immunity to combat *C. burnetii* (Schoenlaub et al., 2015). ELISA methods are recommended by the European Food Safety Authority (EFSA, 2010) and by the World Organisation for

---

* Corresponding author at: Université Clermont Auvergne, INRAE, VetAgro Sup, UMR EPIA, Saint-Genès-Champanelle 63122, France.
E-mail address: laureline.riviere@vetagro-sup.fr (L. Rivière).

https://doi.org/10.1016/j.prevetmed.2025.106509
Received 9 July 2024; Received in revised form 4 March 2025; Accepted 6 March 2025

Animal Health (WOAH, 2018) for the serological diagnosis of *C. burnetii* infection in ruminants. Three ELISA tests, commercialised by IDEXX, Thermofisher Scientific and Innovative Diagnostics, are widely used in Europe. These tests are suitable for the serological diagnosis of C. *burnetii* infections in cattle, goats and sheep from serum or milk samples. They are used in a variety of contexts, including surveillance, differential diagnosis of abortion in a ruminant herd or to identify potential sources of human Q fever outbreaks. In the present study, we focus on their use in a surveillance context, by considering a representative sample of sera collected from a given population, in France, with the aim to assess the seroprevalence of *C. burnetii* infections. Their outputs are expressed as optical density ratios. They measure immunoglobulin G (IgG) antibodies targeting a mix of Phase I and Phase II forms of *C. burnetii* (Williams et al., 1981). For now, although distinct serological profiles for Phase I and Phase II antibodies have been observed in ruminants, their interpretation remains challenging due to difficulties in producing well-purified and characterised antigens for each phase. IgG antibodies specific for *C. burnetii* appear in serum approximately three weeks post-infection, with variable concentrations among individuals. For example, an experimental study on goats found that some animals aborted despite antibody levels being below the cut-off of the ELISA tests (Arricau-Bouvery et al., 2005).

None of the ELISA tests currently available can be considered as a gold standard (i.e., a test with a sensitivity and specificity equal to 100 %) nor as a reference test (i.e., the available test considered as the most reliable). The variation in their diagnostic performances (e.g., from 39 % to 87 % for sensitivity in sheep, as estimated by Lurier et al., 2021) affects the agreement between the test results obtained in different veterinary laboratories and limits the comparability of surveillance data. Additionally, while the manufacturers' cut-offs are the same for all species, the diagnostic performances differ among them (e.g., the sensitivity of the IDEXX test was estimated at 39 % in sheep and 72 % in cattle by Lurier et al., 2021).

It is crucial to harmonise the available tests (i.e., to ensure similar results between them) in order to guarantee comparability and consistency of test results, particularly across network laboratories. Such harmonisation will not only facilitate reliable interpretation for surveillance and consistency across epidemiological studies, but it will also improve the monitoring of animal health within herds and strengthen overall disease surveillance. One possible solution to harmonise these tests is to adjust their cut-offs. However, selecting the appropriate cut-offs can be challenging in the absence of a gold standard.

Without a gold standard, the most common approach used to set up the cut-off of a test is to select a cut-off that discriminates individuals of predetermined or known status. Predetermination of the individual's status is performed using one or several other diagnostic tests of the disease, which are necessarily imperfect. Estimates obtained with this approach are then biased and often lead to the replication of the diagnostic error of the imperfect test used as reference. Another approach, based on individuals of known status, also presents important limitations, because the individuals may not be representative of the target population. For example, negative controls may consist of individuals from pathogen-free countries, who may have limited opportunity to be exposed to microorganisms that might produce cross-reacting antibodies in the targeted population, and positive controls may consist of experimentally infected animals sampled at a given stage of infection whose immune response may differ from animals at other stages (Greiner and Gardner, 2000). Thus, obtaining a collection of reference samples representative of the target population and context for evaluating the diagnostic performances of a serological test for Q fever is particularly challenging (Miller and Kersh, 2020; Stellfeld et al., 2020).

An alternative is to choose the cut-offs by maximising the agreement between the three tests commonly used. All tests may show a proportion of concordant results that are due to chance, which can lead to an overestimation of the agreement that is actually related to the detection of the target condition. For example, two coin tosses will show an observed agreement of 50 %, even though their outcomes are completely random and independent. The proportion of chance agreement is higher when the probability of each test being positive or negative is high, which occurs when the cut-offs are very low or very high, respectively.

Several methods exist for evaluating the agreement between tests. The simplest method is the observed agreement, which refers to the proportion of individuals for whom all the tests give the same result. However, the observed agreement is maximal when all individuals are tested either positive or negative with all the tests, which occurs when the selected cut-offs of the tests are either very low or very high. In these cases, the specificities or the sensitivities of the tests (i.e., their ability to detect truly seronegative or seropositive individuals, respectively) are equal to 0. Therefore, selecting the cut-offs that maximise the observed agreement is not a relevant approach, as it would result in poor diagnostic performances. Some agreement coefficients take into account the proportion of concordant cases due to chance and can be interpreted as the proportion of agreement when chance agreement is removed. This is the case of the classical Cohen's kappa (1960), which is usable to compare only two tests. The agreement coefficients suitable for more than two tests include Fleiss's kappa (1971), Light's kappa (1971), Hubert's kappa (1977) and Gwet's AC1 (2008). Fleiss's kappa, Light's kappa and Hubert's kappa are extensions of Cohen's kappa. They calculate the chance agreement as the proportion of agreement expected under the hypothesis that the results of the different tests are mutually independent. When all cut-off values are either very low or very high, most individuals are classified as either positive or negative by all tests, and so the agreement expected under the hypothesis of mutually independent test results is close to 1. As a result, almost all observed agreements will be attributed to chance and these kappa coefficients (that only calculate the proportion of agreement not attributed to chance) are close to 0 for low or high cut-off values, contrary to the observed agreement. On the other hand, Gwet's AC1 hypothesises that a test either classifies an individual with certainty, in which case it is correctly classified, or by chance, in which case it has a 0.5 probability of being positive. Fleiss's kappa, Light's kappa and Gwet's AC1 assess the pairwise agreement, whereas Hubert's kappa assesses the agreement of all tests at once. The use of the Hubert's kappa seems more relevant in our case, as the idea is to maximise the agreement of the three tests at once rather than the pairwise agreement.

Once the optimal cut-offs have been identified, it is important to evaluate the diagnostic performances of the tests at these cut-offs. In the absence of a gold standard, latent class models can be used to estimate the sensitivity and specificity of tests (Hui and Walter, 1980). These models use the cross-classified test outcomes, categorised as positive or negative according to the selected cut-offs, to infer an unobserved latent status, such as the presence or absence of *C. burnetii* specific antibodies in a serum sample (Lurier et al., 2021).

The objective of this study is to identify the cut-offs that maximise Hubert's kappa, in a surveillance context, using a representative sample of a domestic ruminant population. In a second step, the study aims to assess the impact of the cut-off change by reassessing the diagnostic performances of the tests using a latent class model, and by comparing the observed agreement and the apparent prevalence at the manufacturers' and at the new cut-offs, hereafter termed optimal cut-offs.

2. Materials and methods

2.1. Data

The data come from the Kiteval4500 project. It is a subsample of sera from a French serological survey on Q fever (Gache et al., 2017) which were analysed by three ELISA tests commercially available in Europe for the serological diagnosis of *C. burnetii* infections. The sampling plan is described in Lurier et al. (2021). In summary, 96 cattle herds, 106 goat herds and 99 sheep herds were randomly selected from 10 French administrative departments. Between 6 and 15 individuals per herd were then selected. Serum samples collected from the 1258 cattle, 1474 goats and 1432 sheep were analysed by the national Q fever laboratory.

2.2. Diagnostic tests

The three ELISA tests used to detect antibodies specific for *C. burnetii* in ruminant serum were the IDEXX Q Fever Ab Test, the PrioCHECK™ Ruminant Q Fever Ab Plate Kit, and the ID Screen® Q fever Indirect Multi-species. These tests detect IgG, using different types of conjugates (secondary antibodies or protein G). The antigens used in each test are derived from strains isolated from different species (tick, sheep or cow). The specifications of these tests are detailed in Table 1.

The outputs of the tests are expressed in optical density ratio (ODR). They are calculated, according to the manufacturer's instructions using formula (1) for tests 1 and 2, and formula (2) for test 3.

$$ODR_{serum} = \frac{OD_{serum} - OD_{NC}}{OD_{PC} - OD_{NC}} \tag{1}$$

$$ODR_{serum} = \frac{OD_{serum}}{OD_{PC}} \tag{2}$$

where ODR_{serum} is the optical density ratio of the tested serum, OD_{serum} is the optical density of the tested serum, and OD_{NC} and OD_{PC} are the optical density of the negative and positive controls included in the test, respectively.

In this study, we considered that a serum interpreted as doubtful by the manufacturer was positive. Therefore, we considered the manufacturers' cut-offs to be 30 for test 1 and 40 for tests 2 and 3.

2.3. Search for cut-offs that maximise agreement

2.3.1. Method for assessing agreement

The agreement between the tests was assessed using Hubert's kappa (1977), which is calculated as follows:

$$\kappa = (P_o - P_e)/(1 - P_e)$$

where P_o is the observed proportion of agreement, P_e is the expected proportion of agreement under the hypothesis of mutually independent test results, calculated as the sum, for the positive and negative concordant results, of the products of the observed marginal probabilities of each test (which can be multiplied because of the independence hypothesis).

Thus, for three tests: $P_o = \frac{n_{000}+n_{111}}{N}$

and

$$\begin{aligned} P_e &= \frac{N_{T1+}}{N} \times \frac{N_{T2+}}{N} \times \frac{N_{T3+}}{N} + \frac{N_{T1-}}{N} \times \frac{N_{T2-}}{N} \times \frac{N_{T3-}}{N} \\ &= \frac{n_{100} + n_{101} + n_{110} + n_{111}}{N} \times \frac{n_{010} + n_{011} + n_{110} + n_{111}}{N} \\ &\quad \times \frac{n_{001} + n_{011} + n_{101} + n_{111}}{N} + \frac{n_{000} + n_{001} + n_{010} + n_{011}}{N} \\ &\quad \times \frac{n_{000} + n_{001} + n_{100} + n_{101}}{N} \times \frac{n_{000} + n_{010} + n_{100} + n_{110}}{N} \end{aligned}$$

where N is the total number of individuals, N_{Tt+} (respectively N_{Tt-}) is the number of individuals positive (respectively negative) for test t, and each number n_{ijk} (with i, j and k $\in \{0;1\}$, for tests 1, 2 and 3 respectively) corresponds to the number of individuals with the corresponding result of the three tests, with 0 and 1 in the indices coding for a negative or a positive test respectively. For instance, n_{010} is the number of individuals who tested negative for tests 1 and 3, and positive for test 2.

2.3.2. Optimisation

2.3.2.1. Optimisation challenges. Let's consider that we vary the cut-off of one test while keeping the other tests' cut-offs fixed. The cut-off values to be explored for this test correspond to a succession of continuous intervals separated by the observed result values in the data set. Let's consider one of these intervals. Since no individual has an ODR in this interval, the number of positive and negative individuals remains the same whatever the cut-off value taken in the interval. Thus, as the value of the agreement coefficient only depends on the number of positive and negative individuals, it is constant in each interval. Therefore, a function representing an agreement coefficient as a function of cut-offs is expected to be a step function. Maximising any agreement coefficient is therefore a complex problem.

Optimisation methods, needed to minimise or maximise a function (here to maximise an agreement coefficient), are often based on a gradient method. The principle of such methods is to traverse the function by following the steepest slope until the gradient becomes zero, indicating that a plateau has been reached (Vanderplaats, 1988). This is because the gradient is zero at the extremums. But gradient-based methods may only identify a local extremum. The extremum identified thus depends on the starting point of the algorithm. To identify the global extremum, it is therefore necessary to start from a point close to it. In the case of a step function, the gradient is zero almost everywhere, except on the points where the function rises steeply, making gradient-based methods unsuitable.

Stochastic algorithms were also proposed for global optimisation. They search for the global minimum and do not require initial values to be close to it. In addition, since they are not based on the gradient, they do not require the function to be continuous or derivable and can therefore be applied to step functions. Examples of such algorithms include genetic algorithm, differential evolution, particle swarm optimisation and simulated annealing (Venter, 2010), which are algorithms based on biological or physical principles.

2.3.2.2. Choice of the optimisation algorithm. Four stochastic optimisation algorithms were compared to determine the most appropriate algorithm for our problem:

- The DEoptim function from the R package DEoptim (Mullen et al., 2011). This is a differential evolution algorithm inspired by natural selection.

Table 1
ELISA tests used in the study.

Name used in the study	Test 1	Test 2	Test 3
Commercial name	IDEXX Q fever Ab test	PrioCHECK™ Ruminant Q Fever Ab Plate Kit	ID Screen® Q fever indirect multi-species
Manufacturer	IDEXX	Thermofisher Scientific	Innovative Diagnostics
Strain used for antigen production	Isolated from *Dermacentor andersoni* ticks (*Nine Mile* reference strain)	Isolated from an ewe	Isolated from a cow
Conjugate	Secondary antibodies biding to ruminant IgG	Protein G (binding to IgG of diverse mammalian species)	Protein G (binding to IgG of diverse mammalian species)
Interpretation rules according to the manufacturer (ODR: optical density ratio)	$ODR < 30\%$ Negative $30\% < ODR < 40\%$ Doubtful $40\% < ODR$ Positive	$ODR < 40\%$ Negative $40\% < ODR < 100\%$ Positive + $100\% < ODR < 200\%$ Positive ++ $200\% < ODR$ Positive +++	$ODR < 40\%$ Negative $40\% < ODR < 50\%$ Doubtful $50\% < ODR < 80\%$ Positive $80\% < ODR$ Strong positive

- The GA function from the R package GA (Scrucca, 2013). This is a genetic algorithm also inspired by natural selection.
- The GenSA function from the R package GenSA (Xiang et al., 2013). This is a simulating annealing algorithm inspired by a process used in metallurgy, annealing, which consists of gently cooling an object with periods of heating.
- The psoptim function from the R package pso (Clerc, 2012). This is a particle swarm optimisation algorithm inspired by the movement of a flock of birds, where each individual's movement is influenced by their own experience as well as that of the group.

The principles of these algorithms are explained in Supplementary materials I (Appendix B).

At this step, the purpose was to compare the algorithms, not to obtain a precise estimation of the new optimal cut-offs. As these algorithms are stochastic and therefore do not always return the same result and because the computation time could be long, they were run only 50 times on the dataset using their default calibration parameters to obtain the minimum, maximum and median of the Hubert's kappa values returned. The optimisation algorithm that was best suited to our problem was selected, based on the maximal value of Hubert's kappa obtained and the computation time.

2.3.3. Uncertainty

2.3.3.1. Optimisation uncertainty. As explained, the function representing Hubert's kappa as a function of the cut-offs is expected to be a step function. Therefore, there would not be a trio of cut-off values that maximises Hubert's kappa, but a range of trios of cut-offs. Furthermore, because the algorithms previously described are stochastic, they do not always return the same optimal cut-offs when applied to the same data set. Therefore, running a stochastic optimisation algorithm multiple times enables exploration of the range of optimal cut-offs. For each species, the chosen optimisation algorithm was run 1000 times on the data and the 2.5 %, 50 % and 97.5 % quantiles of the estimated cut-offs were calculated.

An R script that can be used for the optimisation of Hubert's kappa is available in Supplementary material II.

2.3.3.2. Sampling uncertainty. The uncertainty associated with the sampling of animals in the herds was assessed by bootstrapping. A total of 1000 data sets with the same design as the original data set (same number of herds and same number of individuals per herd) were generated by drawing with replacement individuals from each herd. The optimal cut-offs were then evaluated once for each of the generated data sets. Finally, the 2.5 %, 50 % and 97.5 % quantiles of the cut-offs were calculated from the 1000 bootstrap estimations to give an idea of the uncertainty due to sampling.

2.4. Impact of the cut-off change

2.4.1. Diagnostic performances

The sensitivities and specificities of the three tests were estimated at the optimal cut-offs using the latent class model described and assessed by Lurier et al. (2021). They were compared to the ones estimated at the manufacturer cut-offs by Lurier et al. (2021), as we used the same data set and model. The model is a hierarchical zero-inflated beta-binomial latent class model for estimating the sensitivities and specificities of multiple tests in the absence of a gold standard taking into account the conditional dependency between the tests. The latent status modeled in this analysis refers to the presence or absence of *C. burnetii*-specific antibodies in the serum. As samples were obtained during an epidemiologic survey conducted on a representative sample of the domestic ruminant population in France, the latent status assessed in this study corresponds to a surveillance context in the general population, without focusing on any specific clinical context.

2.4.2. Observed disagreement and apparent prevalence

In order to see to what extent the use of optimal cut-offs instead of the manufacturers' cut-offs reduces the number of discordant cases, the observed global disagreement between the three tests (i.e., the proportion of individuals for whom one of the test does not give the same result as the other two) and the observed pairwise disagreement (i.e., the proportion of discordant results between each pair of tests) were compared with both the manufacturer's cut-offs and the optimal cut-offs.

To see how changing the cut-offs improves the comparability of surveillance data, the apparent prevalence (proportion of animals tested positive with each test) was calculated at the manufacturers' cut-offs and at the optimal cut-offs.

3. Results

3.1. Data description

The ODR ranges were quite wide and varied between tests and species. Test 2 had the widest ODR ranges, while test 1 had the narrowest (Table 2).

The number of positive individuals differed between tests. For all species, test 3 was the one with the highest number of positive individuals (Table 2).

3.2. Choice of the optimisation algorithm

The maximum values of Hubert's kappa obtained by the DEoptim and GenSA algorithms were similar and higher than those of GA and psoptim. For instance, in goats, DEoptim and GenSA returned Hubert's kappa values of 0.830 for each of the 50 iterations, while GA returned values between 0.759 and 0.815, and psoptim returned values between 0.793 and 0.830 (Supplementary material I: Appendix C). Additionally, DEoptim had a lower computation time than GenSA. The DEoptim algorithm took approximately 6 minutes per data set, while GenSA took 40 minutes. Therefore, we opted to use DEoptim.

In the following, the number of population members n, which is one of the parameters of this algorithm based on natural selection, was set so that the algorithm returned the same value (to within 10^{-7}) of the maximised Hubert's kappa obtained when run 1000 times ($n = 60$ for cattle and goats; $n = 100$ for sheep). The default values for the other calibration parameters have been retained.

Table 2
ODR range and number of individuals tested positive and negative with the manufacturers' cut-offs for each species and each test.

Species	Test	Minimum ODR	Maximum ODR	Number of negatives (%)	Number of positives (%)
Cattle	Test 1	−2.9	230.2	1049 (83.4 %)	209 (16.6 %)
	Test 2	−10.3	339.5	1089 (86.6 %)	169 (13.4 %)
	Test 3	2.7	241.6	995 (79.1 %)	263 (20.9 %)
Goats	Test 1	−2.8	216.9	1126 (76.4 %)	348 (23.6 %)
	Test 2	−12.4	454.6	1034 (70.1 %)	440 (29.9 %)
	Test 3	2.9	290.1	914 (62.0 %)	560 (38.0 %)
Sheep	Test 1	−2.8	140.3	1322 (92.3 %)	110 (7.7 %)
	Test 2	−9.5	339.7	1275 (89.0 %)	157 (11.0 %)
	Test 3	3.4	263.6	1188 (83.0 %)	244 (17.0 %)

3.3. Optimal cut-offs

The cut-offs of tests 1, 2 and 3 that maximise the agreement were respectively 44.7, 30.8 and 84.3 for cattle, 6.6, 18.8 and 49.8 for goats, and 26.2, 50.9 and 88.6 for sheep.

The maximum Hubert's kappa obtained were identical (to within 10^{-7}) for each of the 1000 iterations of DEoptim in each species. Its values were 0.855 for cattle, 0.830 for goats and 0.713 for sheep. In comparison, the Hubert's kappa calculated for the manufacturers' cut-offs were 0.696 for cattle, 0.689 for goats and 0.590 for sheep.

The medians obtained by bootstrapping were similar to those obtained by iterations of DEoptim for cattle and goats, and slightly lower for sheep (Fig. 1). The uncertainty estimated through iterations of DEoptim was low compared to the uncertainty estimated through bootstrapping (Fig. 1).

3.4. Impact of the cut-off change

3.4.1. Diagnostic performances at individual level

When moving from manufacturers' cut-offs to optimal cut-offs, the sensitivity of test 1 increased by 0.175–0.310 depending on the species, the sensitivity of test 2 increased by 0.113–0.288 and the sensitivity of test 3 decreased by 0.062–0.126 (Fig. 2). The specificities of tests 1 and 2 differed by less than 0.01 while the specificity of test 3 increased by 0.010–0.25 depending on the species (Fig. 2).

In addition, the sensitivities and specificities were closer from one test to another (Fig. 2).

3.4.2. Observed disagreement

At the manufacturers' cut-offs, the proportion of global disagreement was 0.132 in cattle, 0.198 in goats and 0.130 in sheep. At the optimal cut-offs, it was 0.054, 0.121 and 0.078 respectively. Thus, using the optimal cut-offs instead of the manufacturer cut-offs reduced the proportion of discordant individuals for the three tests by 61 % in cattle, 39 % in goats, and 40 % in sheep.

The proportion of pairwise disagreement was reduced by 24–51 %, depending on the species and the test pair (Fig. 3). Furthermore, the proportions of discordant cases were more similar from one pair of tests to another (Fig. 3). Test 3 was the most discordant with the other tests, both when the manufacturers' cut-offs and the optimal cut-offs were used (Fig. 3).

3.4.3. Apparent prevalence

The apparent prevalence of each species varied considerably between tests when the manufacturers' cut-offs were used. They ranged from 13.4 % to 20.9 % in cattle, from 23.6 % to 38.0 % in goats, and from 7.7 % to 17 % in sheep (Fig. 4). However, when the optimal cut-offs were used, the apparent prevalence was more consistent between tests. It ranged from 13.4 % to 14.5 % in cattle, from 35.8 to 37.9 in goats, and from 8.7 to 10.3 in sheep (Fig. 4).

4. Discussion

This study identified the cut-offs that maximise the agreement between the ELISA tests for the serological diagnosis of *C. burnetii* in ruminants at individual level in a context of epidemiological surveillance. At the optimal cut-offs, the observed disagreements between the three tests and each pair of tests decrease (Fig. 3), indicating greater consistency among the three tests. The apparent prevalence is also more similar from one test to another (Fig. 4), which improves the comparability of surveillance data obtained with different tests. Additionally, the diagnostic performances of the tests are globally improved (Fig. 2). Overall, the methodological approach used in this study makes it possible to propose an operational solution for defining the cut-offs for several diagnostic tests, especially when the methods described for estimating ROC curves in the absence of a gold standard are not applicable.

4.1. Optimal cut-off values

While the manufacturers' cut-offs are 30, 40 and 40, the optimal cut-offs vary from 6.6 to 88.6 depending on the test and the species. For all species the cut-offs of test 3 increased, which is coherent with the fact that it was the least specific of the three tests (Lurier et al., 2021). Increasing these cut-offs allows the specificity of this test to be improved. The cut-offs of tests 1 and 2 decreased or increased depending on the species. In particular, the cut-off of test 2 decreased in cattle and the cut-off of test 1 decreased in sheep and goats. This is coherent with

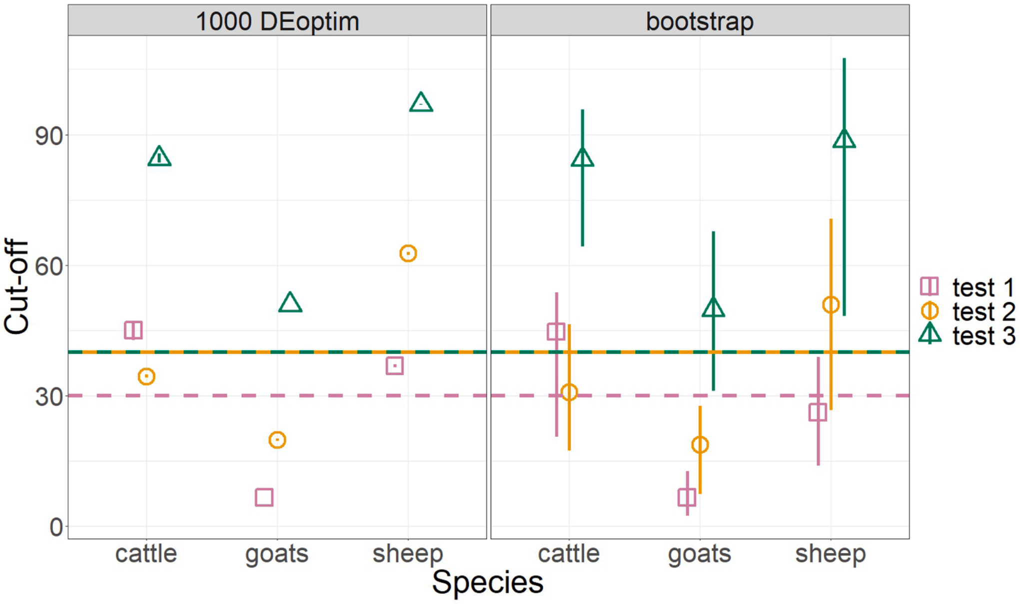

Fig. 1. Cut-offs that maximise Hubert's kappa obtained for 1000 iterations of DEoptim and for 1000 iterations of bootstrapping. The squares, circles and triangles represent the medians for tests 1, 2 and 3 respectively and the vertical lines the 95 % fluctuation intervals. The horizontal dashed lines represent the manufacturer cut-offs.

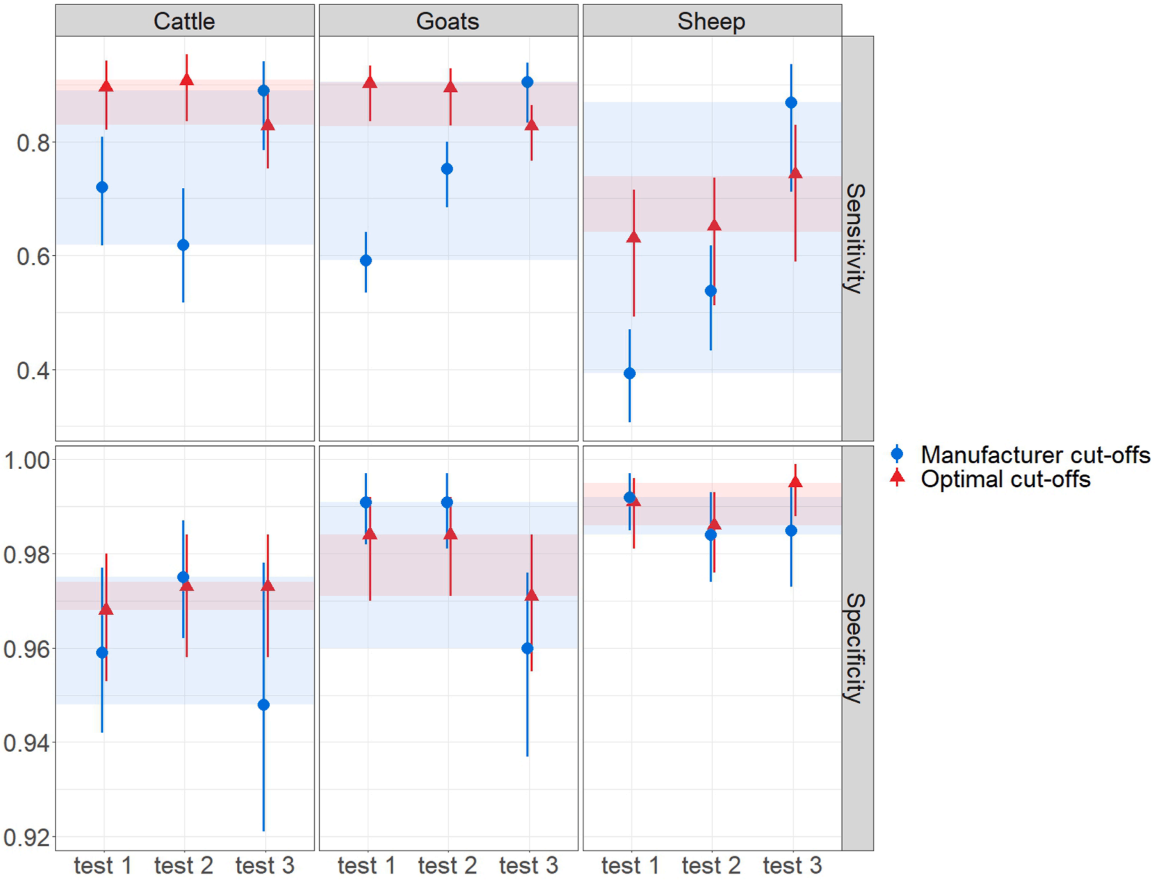

Fig. 2. Estimates of the sensitivities and specificities for diagnosis at individual level of the three tests at the manufacturer cut-offs (30, 40 and 40 for tests 1, 2 and 3 respectively, blue dots) and at the optimal cut-offs (44.7, 30.8 and 84.3 for cattle, 6.6, 18.8 and 49.8 for goats, and 26.2, 50.9 and 88.6 for sheep, red triangles). The triangles and dots represent the median and the lines the 95 % credibility intervals. The ranges of the sensitivities and specificities for each species are represented by blue bands for the manufacturer cut-offs and by red bands for the optimal cut-offs.

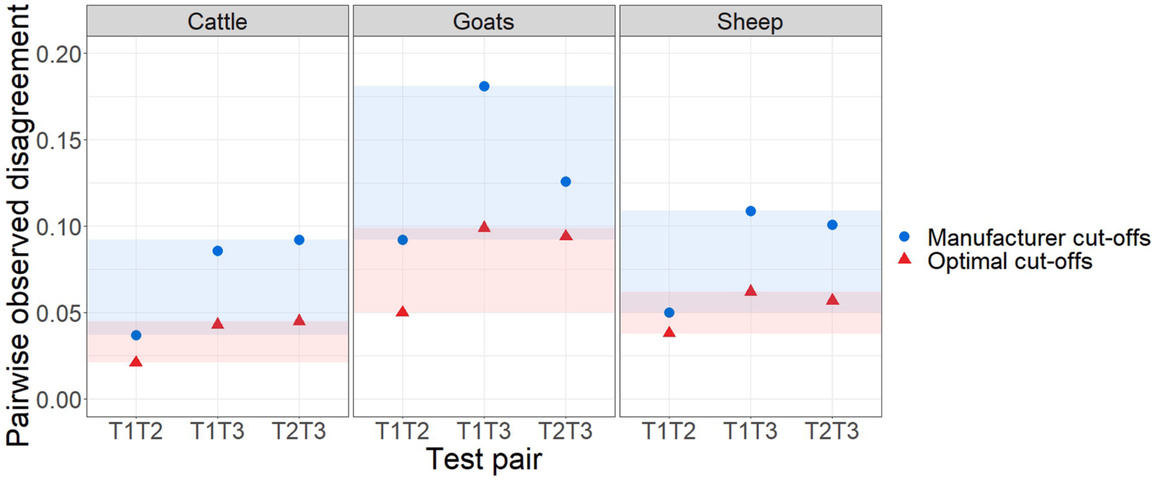

Fig. 3. Pairwise observed disagreement at the manufacturer cut-offs (30, 40 and 40 for tests 1, 2 and 3 respectively, blue dots) and at the optimal cut-offs (44.7, 30.8 and 84.3 for cattle, 6.6, 18.8 and 49.8 for goats, and 26.2, 50.9 and 88.6 for sheep, red triangle). The ranges of variation in the disagreement observed for each species are represented by blue bands for the manufacturer cut-offs and red bands for the optimal cut-offs.

the fact that test 2 is the least sensitive in cattle and that test 1 is the least sensitive in sheep and goats (Lurier et al., 2021). These decreases allow the sensitivity of tests 1 and 2 to be improved.

The pairwise disagreement was approximately divided by two for all pairs of tests and for the three species when using the optimal cut-offs (Fig. 3). However, test 3 remained the most discordant with the other two tests, both when the manufacturer's cut-offs and the optimal cut-offs were used. This is probably due to a high conditional dependence between tests 1 and 2, as this pair of tests was the one with the highest pairwise conditional dependence in the study by Lurier et al. (2021).

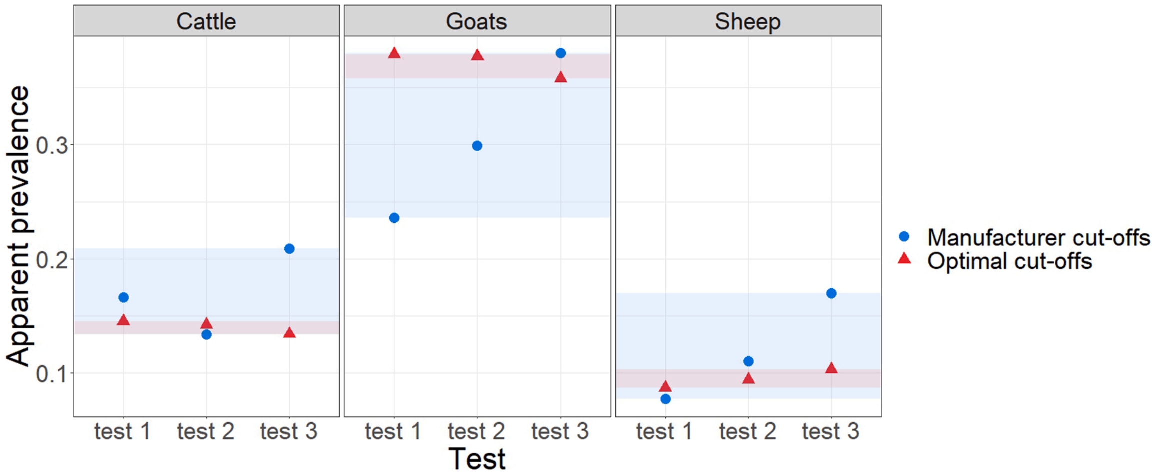

Fig. 4. Apparent prevalence at the manufacturer cut-offs (30, 40 and 40 for tests 1, 2 and 3 respectively, blue dots) and at the optimal cut-offs (44.7, 30.8 and 84.3 for cattle, 6.6, 18.8 and 49.8 for goats, and 26.2, 50.9 and 88.6 for sheep, red triangles). The ranges of variation in the apparent prevalences for each species are represented by blue bands for the manufacturer cut-offs and red bands for the optimal cut-offs.

This stronger agreement between test 1 and test 2 persists even when using the optimal cut-offs. This suggests that the associated conditional dependence may not only be influenced by the respective cut-off of each test, but also by other factors, potentially related to the chemical and physical characteristics of the tests, that we could not identify, such as the reagent concentrations or the components used, particularly the antigen stain and its preparation and purification.

Some optimal cut-offs are much higher or lower than the manufacturers' cut-offs. This raises concerns about the maintaining of the analytical performances of the tests. In particular, the cut-off for test 1 in goats is low (6.6) which may lead to a lack of specificity of this test if the measurement uncertainty around this cut-off is high. Then, before recommending the use of tests at these optimal cut-offs, it will be important to check the measurement uncertainty of each test around its optimal cut-off to ensure that the analytical performance of each test is maintained around the new cut-offs. Further studies are planned to address this issue. These investigations will require detailed analysis of several components involved in analytical variability including intra- and inter-laboratory variations, as well as reproducibility over time and between different batches for each test.

The optimal cut-offs differ between species, unlike the manufacturers' cut-offs, which are common to all three species. These differences could be attributed to variations in the test's composition, the individual factors of the host's immunological response or the biodiversity of circulating *C. burnetii* strains. Although all the tests are indirect ELISAs relying on the same principle, the antigens used in each test are produced from a different strain of *C. burnetii* and the conjugate of test 1 differs from that of tests 2 and 3 (Table 1). In particular, the conjugates used exhibit different affinities for the IgG of different species. Specifically, the test 1 uses a ruminant anti-IgG conjugate whereas the other two employ a protein G conjugate (Table 1). More importantly, the strain used in test 1, the *Nine Mile* reference strain, was isolated from a tick in the USA, whereas the strains used in tests 2 and 3 were isolated in France from an ewe and a cow respectively. The strains circulating in the different ruminant species generally belong to different genotypes of *C. burnetii* (Frangoulidis et al., 2014; Joulié et al., 2017), which could also impact the host antibody production and therefore the performances of each ELISA in each species. One possible approach to harmonising these ELISA tests would be to develop a composite antigen containing a mixture of different *C. burnetii* strains. This mixture would better represent the biodiversity of strain genotypes and thus provide a more comprehensive and effective reference serological antigen.

The use of different cut-offs for each species may be operationally restrictive for veterinary laboratories, which are used to interpreting test results in the same way for all three species. However, by identifying a specific cut-off for each species, the three tests can be more effectively harmonised. Moreover, these tests also hold potential for use in species beyond ruminants, particularly the two tests designed with a protein G (Jeong et al., 2016). For multi-species applications, depending on the performance knowledge needs and application stakes (e.g., import controls for species such as llamas or rhinoceros versus screening for dogs in ruminant herds as sentinels), it will be useful to characterise performances and establish a specific cut-off for each considered species. Thus, the species-specific cut-off definition approach described here appears to offer promising beyond the domestic ruminant species that are classically considered as the main reservoirs of Q fever.

Finally, our study was conducted in a context of surveillance in France and the optimal cut-offs identified may not be optimal for all diagnostic applications and in other areas. For example, in an outbreak context with abortion, infection might be more recent, leading to higher individual antibody responses: the cut-offs that will maximise the agreement between tests may need to be adjusted upwards in such a context. Therefore, studies focusing on other diagnostic contexts should be conducted to verify whether the optimal cut-offs identified in this study, in a surveillance context, remain applicable or whether different cut-offs should be used in each context. The validity of these cut-offs should also be evaluated in other countries to determine whether they are broadly applicable or need to be adapted to specific regional contexts.

4.2. *Impact on the surveillance of Q fever in domestic ruminants*

Since 2021, Q fever is a mandatory surveillance disease at the EU level (Animal Health Law, Regulation (EU) 2016/429 on transmissible animal diseases) but no harmonisation exists in the protocols deployed in the field within the diverse EU members. We may anticipate that part of the surveillance will likely be based on serological surveys that will undoubtedly rely on diverse ELISA tests depending on the country or region. In this study, we showed that the apparent prevalence at the animal level is more similar from one test to another at the new optimal cut-offs (Fig. 4). As ELISA tests are recommended to be used at the herd level instead of the individual level, we also calculated the apparent prevalence at the herd level: they were also more similar from one test to another (Supplementary material I: Appendix D). Overall, using the optimal cut-offs estimated in the current study would allow more comparable surveillance data to be obtained both at the animal and the herd level.

4.3. Kappa values and agreement coefficients

In this study, we chose to focus on Hubert's kappa because it calculates the agreement between the three tests at once rather than the pairwise agreement. We also investigated the cut-offs that maximised Fleiss's kappa, Light's kappa and Gwet's AC1. Fleiss, Light and Hubert's kappas are maximised by similar cut-offs (Supplementary material I: Appendix E). However, Gwet's AC1 is maximal for high values of cut-offs, when almost all the individuals are classified as positive by the three tests. Using cut-offs that maximise Gwet's AC1 would therefore result in poor diagnostic performances with specificities close to 0. Gwet's AC1 is therefore not suitable for our application.

4.4. Optimisation algorithms

We chose the DEoptim algorithm because it was the fastest of the algorithms we tried and was able to achieve the highest values of Hubert's Kappa. When comparing the algorithms, we used their default calibration parameters. However, by changing the calibration parameters, psoptim may have been able to always reach the maximal value of Hubert's kappa and GenSA may have been faster. However, GA does not seem suitable for our study as it was never able to reach the maximal value of Hubert's kappa.

4.5. Uncertainty around the optimal cut-offs

Our study is based on the identification of cut-offs that maximise an agreement coefficient applied to a data set which is a sample of 1258 cattle, 1474 goats and 1432 sheep in France. There are then two potential sources of uncertainty (the "optimisation uncertainty" related to the optimisation process and the "sampling uncertainty" related to the sampling), which should be taken into account in order to assess the precision of the estimation of the optimal cut-offs. The optimisation uncertainty, probably mostly related to the shape of the function to be maximised, seems to be negligible (Fig. 1).

To our knowledge, there is no calibrated method to account for the sampling uncertainty (e.g., to give 95 % confidence intervals) around the cut-offs that maximise an agreement coefficient. Therefore, we assessed the sampling uncertainty by bootstrapping while maintaining the structure of the data (same number of herds and same number of individuals per herd). The fluctuation intervals obtained with this method were wide (Fig. 1), which means that the "true" values of the optimal cut-offs could be far from those presented in this study. However, although these fluctuation intervals reflect the uncertainty associated with sampling variability, they are difficult to interpret because their estimation method has not been calibrated to correspond to 95 % confidence intervals. In particular, a few of the 1000 bootstrapped data sets resulted in outliers estimated optimal cut-offs that were higher than the rest of the cut-offs obtained (Supplementary material I: Appendix F). These particular data sets simulated during the bootstrapping process are likely to have an impact on the calculated fluctuation intervals, which should lead us to interpret them with caution. It would be more appropriate to reassess the optimal cut-offs on another data set, as a way to assess their external validity.

4.6. Impact on the diagnostic performances

In our case, the sensitivities and specificities of the tests are globally better at the optimal cut-offs. However, the increase in sensitivity for test 1 in cattle and test 2 in sheep is not consistent with the increase in the optimal cut-off compared to the manufacturer's cut-off. This inconsistency may be due to differences in the latent status adjusted by the latent class model when applied to the data sets dichotomised at the optimal cut-offs and the manufacturer's cut-offs. Indeed, in parallel with the increase in sensitivity as the cut-off increases, we also observed a decrease in the true prevalence assessed by the model. This observation is similar to the non-monotonous variation of ROC curves assessed in other studies (Frössling et al., 2003; Nielsen et al., 2002) when multiple latent class models were applied independently with multiple cut-offs.

We chose to identify the optimal cut-offs by maximising the agreement between tests. However, maximising the agreement between tests does not guarantee an improvement in their diagnostic performances. Ideally, the choice of the cut-offs should be based on the ROC curves. However, the methods proposed for such a purpose (ordered multinomial model – Wang et al., 2007 – and mixture model – Choi et al., 2006) cannot be applied to our data, due to the conditional dependency of the three tests (Lurier et al., 2021) and to the non-Gaussian distribution of their continuous output (ODR) (Supplementary material I: Appendix G). Another approach to estimating ROC curves in the absence of a gold standard, using the results of a latent class model, has recently been proposed by Olsen et al. (2022) and could have been used for the three ELISAs considered in this study. However, we observed inconsistencies in the variation of the sensitivity and specificity of the tests when changing cut-offs (e.g. increased sensitivity when the cut-off was increased). As this approach calculates the ROC curves on the basis of the probabilities of true positivity of each individual, which in turn is calculated on the basis of the sensitivity, specificity and prevalence estimated at a trio of cut-off points, the ROC curves were dependent on the trio of cut-off points originally used to calculate them. Such dependency makes it challenging to determine a reliable and representative ROC curve. Therefore, we decided not to use this approach in this study.

One perspective for future research would be to develop a method to estimate sensitivities and specificities simultaneously for all cut-offs, taking into account the conditional dependency of the tests.

5. Conclusion

This study determined the optimal cut-offs that maximise the agreement between the ELISA tests used for the serological diagnosis of C. burnetii infections in ruminants in a context of epidemiological surveillance. Specifically, the identified cut-offs for tests 1, 2 and 3 are 44.7, 30.8 and 84.3 in cattle, 6.6, 18.8 and 49.8 in goats, and 26.2, 50.9 and 88.6 in sheep, respectively. The use of these species-specific cut-offs rather than the uniform ones provided by manufacturers significantly enhances the harmonisation of diagnostic performances across the tests and improves the comparability of surveillance data.

Funding sources

This study was funded by the French Agency for Food, Environmental and Occupational Health & Safety (ANSES), the French Directorate General for Food (DGAL) of the Ministry of Agriculture, the French National Animal Health Farmers' Organization (GDS France), the French Institute for Research in Agriculture, Food & the Environment (INRAE) and VetAgro Sup.

CRediT authorship contribution statement

Rousset Elodie: Writing – review & editing, Supervision, Investigation, Conceptualization. **Rivière Laureline:** Writing – original draft, Visualization, Software, Methodology, Formal analysis. **Lurier Thibaut:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Delignette-Muller Marie-Laure:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Jourdain Elsa:** Writing – review & editing, Investigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the French platform for epidemiological surveillance in animal health, the farmers who participated in this study, the veterinarians who collected the samples, the Departmental Veterinary Laboratories that performed the analyses, and the Animal Health Farmers' Organizations that coordinated the study locally.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.prevetmed.2025.106509.

References

Arricau-Bouvery, N., Souriau, A., Bodier, C., Dufour, P., Rousset, E., Rodolakis, A., 2005. Effect of vaccination with phase I and phase II *Coxiella burnetii* vaccines in pregnant goats. Vaccine 23, 4392–4402. https://doi.org/10.1016/j.vaccine.2005.04.010.

Choi, S., Fisman, R., Gale, D., 2006. Substantive and Procedural Rationality in Decisions under Uncertainty.

Clerc, M., 2012. Standard Particle Swarm Optimisation.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46. https://doi.org/10.1177/001316446002000104.

EFSA, 2010. Scientific opinion on Q fever. EFSA J. 8, 1595. https://doi.org/10.2903/j.efsa.2010.1595.

Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychol. Bull. 76, 378–382. https://doi.org/10.1037/h0031619.

Frangoulidis, D., Walter, M.C., Antwerpen, M., Zimmermann, P., Janowetz, B., Alex, M., Böttcher, J., Henning, K., Hilbert, A., Ganter, M., Runge, M., Münsterkötter, M., Splettstoesser, W.D., Hanczaruk, M., 2014. Molecular analysis of *Coxiella burnetii* in Germany reveals evolution of unique clonal clusters. Int. J. Med. Microbiol. One Health 304, 868–876. https://doi.org/10.1016/j.ijmm.2014.06.011.

Frössling, J., Bonnett, B., Lindberg, A., Björkman, C., 2003. Validation of a *Neospora caninum* iscom ELISA without a gold standard. Prev. Vet. Med. 57, 141–153. https://doi.org/10.1016/S0167-5877(02)00216-7.

Gache, K., Rousset, E., Perrin, J.B., Cremoux, R.D., Hosteing, S., Jourdain, E., Guatteo, R., Nicollet, P., Touratier, A., Calavas, D., Sala, C., 2017. Estimation of the frequency of Q fever in sheep, goat and cattle herds in France: results of a 3-year study of the seroprevalence of Q fever and excretion level of Coxiella burnetii in abortive episodes. Epidemiol. Infect. 145, 3131–3142. https://doi.org/10.1017/S0950268817002308.

Greiner, M., Gardner, I.A., 2000. Epidemiologic issues in the validation of veterinary diagnostic tests. Prev. Vet. Med. 45, 3–22. https://doi.org/10.1016/S0167-5877(00)00114-8.

Gwet, K.L., 2008. Computing inter-rater reliability and its variance in the presence of high agreement. Br. J. Math. Stat. Psychol. 61, 29–48. https://doi.org/10.1348/000711006X126600.

Hubert, L., 1977. Kappa revisited. Psychol. Bull. 84, 289–297. https://doi.org/10.1037/0033-2909.84.2.289.

Hui, S.L., Walter, S.D., 1980. Estimating the error rates of diagnostic tests. Biometrics 36, 167–171. https://doi.org/10.2307/2530508.

Jeong, S., Heu, W., Kim, J.-W., Kim, H.-S., 2016. Protein binders specific for immunoglobulin G from different species for immunoassays and multiplex imaging. Anal. Chem. 88, 11938–11945. https://doi.org/10.1021/acs.analchem.6b03851.

Joulié, A., Sidi-Boumedine, K., Bailly, X., Gasqui, P., Barry, S., Jaffrelo, L., Poncet, C., Abrial, D., Yang, E., Leblond, A., Rousset, E., Jourdain, E., 2017. Molecular epidemiology of *Coxiella burnetii* in French livestock reveals the existence of three main genotype clusters and suggests species-specific associations as well as regional stability. Infect. Genet. Evol. 48, 142–149. https://doi.org/10.1016/j.meegid.2016.12.015.

Light, R.J., 1971. Measures of response agreement for qualitative data: some generalizations and alternatives. Psychol. Bull. 76, 365–377. https://doi.org/10.1037/h0031643.

Lurier, T., Rousset, E., Gasqui, P., Sala, C., Claustre, C., Abrial, D., Dufour, P., de Crémoux, R., Gache, K., Delignette-Muller, M.-L., Ayral, F., Jourdain, E., 2021. Evaluation using latent class models of the diagnostic performances of three ELISA tests commercialized for the serological diagnosis of Coxiella burnetii infection in domestic ruminants. Vet. Res. 52, 56. https://doi.org/10.1186/s13567-021-00926-w.

Miller, H.K., Kersh, G.J., 2020. Analysis of recombinant proteins for Q fever diagnostics. Sci. Rep. 10, 20934. https://doi.org/10.1038/s41598-020-77343-0.

Mullen, K., Ardia, D., Gil, D., Windover, D., Cline, J., 2011. DEoptim: an R package for global optimization by differential evolution. J. Stat. Soft. 40. https://doi.org/10.18637/jss.v040.i06.

Nielsen, S.S., Grønbæk, C., Agger, J.F., Houe, H., 2002. Maximum-likelihood estimation of sensitivity and specificity of ELISAs and faecal culture for diagnosis of paratuberculosis. Prev. Vet. Med. 53, 191–204. https://doi.org/10.1016/S0167-5877(01)00280-X.

Olsen, A., Nielsen, H.V., Alban, L., Houe, H., Jensen, T.B., Denwood, M., 2022. Determination of an optimal ELISA cut-off for the diagnosis of Toxoplasma gondii infection in pigs using Bayesian latent class modelling of data from multiple diagnostic tests. Prev. Vet. Med. 201, 105606. https://doi.org/10.1016/j.prevetmed.2022.105606.

Schoenlaub, L., Elliott, A., Freches, D., Mitchell, W.J., Zhang, G., 2015. Role of B cells in host defense against primary Coxiella burnetii infection. Infect. Immun. 83, 4826–4836. https://doi.org/10.1128/iai.01073-15.

Scrucca, L., 2013. GA: a package for genetic algorithms in R. J. Stat. Softw. 53, 1–37. https://doi.org/10.18637/jss.v053.i04.

Stellfeld, M., Gerlach, C., Richter, I.-P., Miethe, P., Fahlbusch, D., Polley, B., Sting, R., Pfeffer, M., Neubauer, H., Mertens-Scholz, K., 2020. Evaluation of the diagnostic potential of recombinant Coxiella burnetii Com1 in an ELISA for the diagnosis of Q fever in sheep, goats and cattle. Microorganisms 8, 1235. https://doi.org/10.3390/microorganisms8081235.

Vanderplaats, G.N., 1988. Multidiscipline design optimization. Appl. Mech. Rev. 41, 257–262. https://doi.org/10.1115/1.3151897.

Venter, G., 2010. Review of optimization techniques. Encyclopedia of Aerospace Engineering.

Wang, C., Turnbull, B.W., Gröhn, Y.T., Nielsen, S.S., 2007. Nonparametric estimation of ROC curves based on Bayesian models when the true disease state is unknown. JABES 12, 128–146. https://doi.org/10.1198/108571107X178095.

Williams, J.C., Peacock, M.G., McCaul, T.F., 1981. Immunological and biological characterization of Coxiella burnetii, phases I and II, separated from host components. Infect. Immun. 32, 840–851. https://doi.org/10.1128/iai.32.2.840-851.1981.

WOAH, 2018. Q fever - Terrestrial Animals (mammals, birds and bees). Contributors: Rousset, E., Niemczuk, K, Sidi-Boumedine, K., in: WOAH Terrestrial Manual 2018. pp. 560–577.

Xiang, Y., Gubian, S., Suomela, B., Hoeng, J., 2013. Generalized simulated annealing for global optimization: the GenSA package. R. J. 5, 13. https://doi.org/10.32614/RJ-2013-002.