# Title page

Machine learning classification methods informing the management of inconclusive reactors at bovine tuberculosis surveillance tests in England

M. Pilar Romeroa,b\*, Yu-Mei Changb, Lucy A. Bruntonb, Jessica Parrya, Alison Prossera, Paul Uptona, and Julian A. Dreweb

*a Animal and Plant Health Agency, Woodham Lane, Addlestone, Surrey, KT15 3NB, United Kingdom.*

*b Royal Veterinary College, Hawkshead Lane, North Mymms, Hatfield, Hertfordshire, AL9 7TA, United Kingdom.*

\* Corresponding author: APHA, Nobel House, 17 Smith Square, London, SW1P 3JR, United Kingdom. Tel.: +44(0)7900052396;e-mail address: mromero7@rvc.ac.uk.

Word count of main sections including abstract (excluding citations): 6 198.

# Abstract

Bovine tuberculosis (bTB) remains one of the most complex, challenging, and costly animal health problems in England. Identifying and promptly removing all infected cattle from affected herds is key to its eradication strategy; the imperfect sensitivity of the diagnostic testing regime remaining a serious obstacle.

The main diagnostic test for bTB in cattle in England, the Single Intradermal Comparative Cervical Tuberculin Test (SICCT: also known as the skin test), can produce inconclusive results below the reactor threshold. The immediate isolation of inconclusive reactor (IR) animals followed by a 60-day retest may not prevent lateral spread within the herd (if it is substandard, allowing transmission) or transmission to wildlife. Over half of IR-only herds that went on to have a positive skin test result (a bTB herd ‘incident’) in 2020, had it triggered by at least one IR not clearing their 60-day retest, instead of by another test within the previous 15 months.

Machine learning classification algorithms (classification tree analysis and random forest), applied to England’s 2012-2020 IR-only surveillance herd tests, identified at-risk tests for an incident at the IRs’ 60-day retest. In this period, 4 739 out of 22 946 (21%) IR-only surveillance tests disclosing 6 296 out of 42 685 total IRs, had an incident at retest (2 716 IRs became reactors and 3 580 IRs became two-time IRs). Both models showed an AUC above 80% in the 2012-2019 dataset. Classification tree analysis was preferred due to its easy-to-interpret outputs, 70% sensitivity, and 93% specificity in the 20% of 2019-2020 testing dataset.

The paper aimed to identify IR-only surveillance tests at-risk of an incident at the 60-day retest to target them with appropriate measures to mitigate the IRs’ risk. Sixteen percent (341 out of 2 177) of IR-only herd tests were identified as high-risk in the 2020 dataset, with 265 (78%) of these having at least one reactor or IR at retest. Severe-level reinterpretation of the high-risk IR-only disclosing tests identified in this dataset would turn 68 out of the 590 (12%) IRs into reactors, generating 23 incidents, the majority (19 or 83%) part of the 265 incidents that would have been declared at the retest.

Classification tree analysis used to identify IR-only high-risk tests in herds eligible for severe interpretation would enhance the sensitivity of the test-and-slaughter regime, cornerstone of the bTB eradication programme in England, further mitigating the risk of disease spread posed by IRs.

Keywords: *Bovine tuberculosis, Machine learning, Classification tree analysis, Inconclusive reactor management, Surveillance, England.*

1. Introduction

Bovine tuberculosis (bTB), caused by infection of cattle with *Mycobacterium bovis* (*M. bovis*), continues to prove very difficult to tackle, particularly in England, which has the highest incidence in Europe (EFSA/ECDC, 2019), yet which aims to eradicate it by 2038 (Defra, 2014). The control and eradication of *M. bovis* requires the early recognition and prompt removal of subclinically-infected animals to prevent transmission to other animals as well as humans (Álvarez et al., 2012; Cousins, 2001; de la Rua-Domenech et al., 2006; Good and Duignan, 2011). However, currently no diagnostic test allows for a perfect identification of all infected and uninfected cattle (de la Rua-Domenech et al., 2006).

The main diagnostic test for tuberculosis in cattle is the intradermal tuberculin test - or skin test - which elicits a localised delayed hypersensitivity reaction in animals exposed to *M. bovis* (de la Rua-Domenech et al., 2006). Imperfect test sensitivity (the probability that an infected animal is classified as positive) contributes to the persistence of bTB in cattle in parts of Great Britain (Álvarez et al., 2012; Gormley et al., 2006; Karolemeas et al., 2012). Improvements to the detection of infected cattle could have a positive impact on bTB eradication (Nuñez-Garcia et al., 2017).

Due to environmental mycobacteria being responsible for up to 12% of false positive results to the single intradermal test (Pollock et al., 2003), in Great Britain and Ireland the Single Intradermal Comparative Cervical Test (SICCT) is used. The skin thickness is measured and injected at two sites in the neck (Figure S1\_a supplementary materials S1) with avian (*M. avium*) and bovine (*M. bovis*) tuberculins. After 72 hours, the injection sites are inspected for inflammatory or oedematous reactions and the skin thickness remeasured. Cattle infected with *M. bovis* tend to show a greater response to the bovine tuberculin than to the avian tuberculin, as opposed to if exposed to other mycobacteria, like *M. avium* (de la Rua-Domenech et al., 2006; Francis et al., 1973; Pollock et al., 2003). Following the SICCT, an animal can be declared: negative (‘clear’), positive (‘reactor’), or uncertain (‘inconclusive reactor’ [IR]), following the rules for either standard or stricter (severe) interpretation (APHA, 2021) (Figure S1\_b supplementary materials S1).

An IR is an animal showing an SICCT result where the difference in size of reactions to the bovine and avian tuberculins is not large enough to be described as a reactor (positive), nor small enough to be considered clear (negative). Such animals are isolated on farm, having their movements restricted (also in the whole farm, depending on bTB history) pending a second SICCT, 60 days later (IR retest). Before then, they can also be compulsorily removed to slaughter as dangerous contacts (non-reactor cattle at high risk of being infected (APHA, 2020a)), have a follow-up interferon-gamma (IFN-gamma) blood test, or be voluntarily slaughtered by their owner (APHA, 2020a).

Compulsory surveillance for bTB in England (UK Statutory Instruments, 2021) takes place at herd level, yearly or six-monthly in the two worst-affected areas (High-risk area [HRA] and Edge [EA] areas) and four-yearly in the Low-risk area (LRA) (APHA, 2021). The detection of IRs without concurrent test reactors (i.e. ‘IR-only’ herds), following the application of the SICCT in the herd, represents an increased risk of bTB in England at the herd (Brunton et al., 2018) as well as animal (May et al., 2019) levels. Over 60% of IR-only herds and of individual IRs (64% and 68%, respectively) detected in England in 2020, were identified in the HRA (APHA, 2021). Furthermore, 39% of IR-only herds in the HRA went on to have a bTB incident (a positive skin test result) within 15 months (compared to 32% in the EA and 20% in the LRA) (APHA, 2021), in the majority of cases through the detection of reactors at the IRs’ 60-day retest: in 61% of IR-only herds in the HRA and in 86% of cases in the EA and LRA (APHA, 2021).

Since a 60-day isolation and retesting of IRs alone may not be sufficient to eliminate the risk posed by IR-only herds in England and Wales (Brunton et al., 2018), it may be necessary to explore additional control options (Clegg et al., 2011a; May et al., 2019), such as closer follow-up at future tests, immediate slaughter, or lifetime restrictions (Clegg et al., 2011b). Veterinarians and farmers should be made aware of the increased risk associated with resolved IRs (i.e. those with a negative result at the 60-day retest) at future testing events and have access to the testing history of individual animals to monitor their future test results (Clegg et al., 2011b).

The online bTB information portal (TBHub) includes information on the risk posed by IRs but Official Veterinarians (OVs) – private vets performing official duties on behalf of the government - do not have access to records indicating if any animals in a herd are resolved IRs. Nor there is a threshold for removing all IRs in England, as it is done in Ireland (Phelan, 2020). Since November 2017, however, all resolved IRs in the HRA and the EA (and in incident tests in the LRA) have to be restricted to their farm for life, unless moved to slaughter or privately re-tested with a negative result using the IFN-gamma test (Defra, 2020). Resolved IRs that remain on farm could, nonetheless, still pose a risk to other cattle on the same and adjoining farms, or local wildlife, as they do not need to be isolated.

To mitigate against this risk without slaughtering all IRs, we propose a proactive severe reinterpretation of IR-only surveillance tests conducted in certain herds. Applying the severe level interpretation of the SICCT test, as opposed to the default standard interpretation in surveillance tests, is a way to increase its sensitivity (Karolemeas et al., 2012); although it also increases the false positive rate from 1 in 5 000 in to 1 in 1 111 (APHA, 2020a). The lower positive cut-off used under the severe interpretation of the SICCT in England results in some IRs being reclassified as reactors (de la Rua-Domenech et al., 2006) (Figure S1\_b supplementary materials S1).

Severe level reinterpretation of surveillance test results is applied on disclosure of an incident due to the presence of reactors in herds in the HRA and the EA, irrespective of post-mortem and laboratory culture results (Defra, 2014). Severe interpretation is also applied to skin tests performed on individual cattle traced from herds with confirmed bTB incidents in England since April 2017 (Defra, 2020). In addition, reinterpretation at severe level of tests is a measure that can be taken by government vets at their own discretion in high-risk scenarios. Still, a systematic method for identifying the tests targeted for reinterpretation is needed. Machine learning, which has been used to develop sensitive and specific models to assess the likelihood of a bTB incident (Romero et al., 2021, 2020), offers a mechanism to do this.

Decision trees are non-parametric machine learning methods that partition the predictor space successively into regions, using splitting rules that are summarized in a tree (James et al., 2014; Therneau and Atkinson, 2018). The underlying algorithm (Breiman et al., 1984) does not make distributional assumptions, can deal with different levels of measurements in variables (Fei et al., 2017; Lewis, 2000), is robust against extreme biases in observations, and does not have restrictions on the number of predictors (Frisman et al., 2008; Strobl, 2010). Decision trees have been applied to veterinary (Klepp et al., 2019; Ortiz-Pelaez and Pfeiffer, 2008; Romero et al., 2020; Staerk and Pfeiffer, 1999), medical (Afonso et al., 2012; Cheng et al., 2018; Frisman et al., 2008; Imafuku et al., 2018; Kawamura et al., 2012; Kitsantas et al., 2006; Ramezankhani et al., 2014; Scheetz et al., 2009), environmental (Julien et al., 2008; Yang et al., 2016), and economic (Chen et al., 2019; Wałęga and Wałęga, 2021) areas, producing easy-to-interpret models (Kuhnert and Venables, 2005; Saegerman et al., 2011). Random forest (Breiman, 2001; Liaw and Wiener, 2002) is an improved decision-tree method which combines resampled observations and variables from multiple trees producing a single consensus outcome prediction from the de-correlated trees, reducing variability and improving prediction accuracy, although reducing interpretability (Hastie et al., 2017; James et al., 2014). They have also been widely applied (Baca-Garcia et al., 2007; Carranza and Laborte, 2015; Díaz-Uriarte and Alvarez de Andrés, 2006; Kane et al., 2014; Ražanskas et al., 2015), including to the identification of high-risk herds with regards to their probability of having a bTB incident (Romero et al., 2021). Both, classification tree analysis and random forest, have also been compared to solve different health problems (Shaikhina et al., 2019; Speiser et al., 2019). Here, we demonstrate their application to the identification of IR-only bTB surveillance tests that are associated with the detection of an incident at the 60-day IR retest.

Epidemiological analysis of data must be incorporated into control and eradication programmes (Good and Duignan, 2011; White et al., 2008). The aim of this study was to use herd-level data from 2012 to 2019 to build classification tree and random forest models that, after a herd-level assessment, would classify herds with IR-only surveillance tests as high- or low-risk, depending on the likelihood of those IRs failing to clear the 60-day retest. This may allow for the targeted introduction of additional disease control measures to mitigate the risk posed by IRs. Here, we will evaluate the application of the best model to the dataset of IR-only surveillance tests in England in 2020, assessing how many extra IRs would be reclassified as reactors through severe reinterpretation of the disclosing test results in high-risk herds, thus triggering new bTB incidents before the 60-day IR retest. This information could contribute to inform policy changes in the way IR-only herds are managed in England.

1. Methods

2.1 Source datasets

Cattle bTB-testing data held by the Animal and Plant Health Agency (APHA) were used, along with other data on potential herd-level predictors for active herds in England from 2012 to 2020, ranging from demographic herd characteristics and bTB-related variables (e.g. past bTB history from as early as January 2000) from the bTB management system (Sam), to cattle movements from the Cattle Tracing System (from as early as January 2012). An additional purpose-designed dataset comprising IR-only surveillance tests from 2012 to 2020 was populated by the APHA’s Data Systems Group. The outcome variable was *Any reactor or IR (yes/no)* at the 60-day retest of IRs.

2.2 Data eligibility

Non-eligible tests in England were excluded. These were:

* Tests where the number of IRs with all possible outcomes at the 60-day retest (i.e. IR, reactor, clear, dangerous contact, negative, slaughterhouse case, dead, and slaughter) did not match the number of IRs detected at the surveillance test;
* Tests where there was not a 60-day retest (i.e. retest variable is “No”), and
* Tests where the number of IRs with outcomes of interest at the 60-day retest (i.e. the possible results to the skin test: IR, reactor, and clear) did not match the number of IRs detected in the disclosing test (i.e. there were other outcomes apart from the ones of interest).
* IR-only tests in herds that were not present in the demographic, movements, and bTB history variables datasets were also excluded from the analysis. Any tests in herds that were part of multi-herd holdings were also removed, since the movement data is extracted at holding level.

2.3 Data analysis

*2.3.1 Descriptive data analysis*

The initial dataset used for analysis was made up of the outcome variable (i.e. *Any reactor or IR at the retest [yes/no]*) and 93 predictors (supplementary materials S2). Numerical variables were not categorized. Tests in multiple-occurring herds in a given year or across the 2012-2020 study period (supplementary materials S3) were not excluded, since the machine learning algorithms used don’t assume independence of observations (James et al., 2014; Kirasich et al., 2018). The presence of missing values was assessed and dealt with by removing tests with any missing observations (i.e. complete-case analysis) (Hayes et al., 2015; Maimon and Rokach, 2010; Pedersen et al., 2017).

*2.3.2 Feature selection*

To aid the practical management of IR-only surveillance tests in England we purposefully selected only the top 20 variables to be included in the classification tree or random forest analysis. This feature selection process was carried out separately for each algorithm prior to the development of the classification models.

*2.3.3 Classification tree analysis*

Classification tree analysis was carried out using the CART algorithm (Breiman et al., 1984; Therneau and Atkinson, 2018) on the training dataset, resulting from the split of the resulting IR-only surveillance tests dataset from 2012 to 2019 into training and testing sets, using an 80:20 (training: testing) random split (Fei et al., 2017; Kassambara, 2018; Kawamura et al., 2012; Yang et al., 2016) that maintained the proportion of cases and non-cases (Kuhn, 2008). The algorithm repeatedly allocates herds into the most homogeneous groups by outcome class, choosing the best combination of predictor and cut-off point each time (Bruce, P; Bruce, 2017; James et al., 2014). The gini or purity index (G), used as the splitting criteria, is defined as:

$G=\sum\_{k=1}^{K}\hat{p}\_{mk}(1-\hat{p}\_{mk})$,

were $\hat{p}\_{mk}$ represents the proportion of observations in the *m*th region that are in a particular outcome class (*k*th) (James et al., 2014). The terminal nodes after the last split show the majority incident class from the relative frequencies of tests (Song and Lu, 2015; Strobl, 2010), as described (Romero et al., 2020). A five-fold cross-validation strategy was adopted (Kuhn, 2008), pruning the fully-grown trees to prevent overfitting (Breiman et al., 1984). Initially, the mean decrease Gini index was used to select the top 20 most important variables and cut-off points, with the variables included in the list used subsequently to develop the final classification tree prediction algorithm.

*2.3.4 Random forest*

The random forest model (Breiman, 2001), based on an ensemble of classification trees (Breiman et al., 1984; Therneau and Atkinson, 2018), was implemented (Liaw and Wiener, 2002) using the same training and testing datasets as the classification tree analysis. Bootstrapped samples drawn with replacement from the training dataset and a random sample of predictors were selected before each split to create the trees in the ensemble using the Gini index (Genuer et al., 2010; Hastie et al., 2017; Maimon and Rokach, 2010). The final trees were tuned for the number of trees in the forest (500 initially) and the fixed number of input variables chosen at random before each split (eight initially) (Hastie et al., 2017; Liaw and Wiener, 2002). First, the out-of-bag error matrix was accessed to get the number of trees in the forest that minimised the out-of-bag error rate. Then, the number of variables chosen before each split was determined using algorithm tools (Liaw and Wiener, 2002). Mean decrease Gini index was calculated to select the top 20 most important predictors, and these were used to develop the final random forest prediction algorithm. The mean decrease Gini index was calculated to select the top 20 most important variables, and these were used to develop the final random forest prediction algorithm.

*2.3.5 Model selection and evaluation of outputs*

The models’ predictive performance was assessed on the original testing dataset or Test 1 (20% of the 2012-2019 dataset, data not seen by the developed models) (Khun et al., 2014; Kuhn, 2008) using: accuracy (a property of classification models, based on the number of correctly classified observations in the confusion matrix), sensitivity, specificity, positive and negative predictive values, balanced accuracy (i.e. average between sensitivity and specificity), and area under the ROC (AUC) (Fei et al., 2017).

The final model was chosen based on its predictive performance and ease of implementation, having sought feedback from policy makers in meetings, during which model outputs were presented and the pros and cons of the different models were also explained. To further verify the fitting of the chosen model and evaluate the proactive reinterpretation at severe level of the high-risk IR-only surveillance tests identified, the 2020 dataset was used as a new testing dataset (Test 2), with its predictive performance assessed in the same way as before. The number of IRs taken as reactors at their disclosing test through the proactive reinterpretation at severe level of high-risk tests identified, and subsequent incidents created, was compared to the number of IRs that would have yielded a reactor or IR retest at the 60-day retest; also triggering an incident as a result.

*2.3.6 Impact of imbalanced classes data*

Imbalanced classes are a common problem in machine learning where the distribution of samples across the known classes are skewed. This could skew the models towards predicting the majority class of the outcome variable (“no”, with 79% of observations) if the model performance was evaluated based on accuracy, as demonstrated previously (Romero et al., 2021, 2020). To assess the level of impact on the prediction introduced by this imbalance in the data, the analysis was repeated using a down-sampling approach within the training datasets, independent of the cross-validation process. A down-sampled dataset was created by selecting a random sample of non-incident herds matching the number of incident ones (Chawla et al., 2002; Garcia et al., 2009; Kuhn, 2008; Mostafizur Rahman and Davies, 2013). Its predictive performance on the testing dataset was compared against the one from the algorithm developed using the entire training data to evaluate the impact of imbalanced classes data.

Statistical analyses and some reported outcomes, like Sankey graphs (Gandrud et al., 2017), were carried out using the R statistical software (R Core Team, 2021) version 4.1.1.

1. Results

3.1 Summary of data

There were 25 098 IR-only surveillance tests in England in the initial dataset (2012-2020); 53 (0.2%) tests with non-matching retest outcomes compared to IRs detected, 356 (1.4%) tests without a retest, and 801 (3.2%) tests with retest outcomes different from the sum of reactor, IR or clear results were removed (Figure 1). No tests had to be removed due to having taken place in multiple-herd holdings, these representing only 2% (11 396 out of 596 017) of active herds in England and Wales in this study period, nor because of a lack of match in the predictors’ datasets. Tests with missing observations in at least one variable were excluded (942 out of 23 888 or 3.9% of tests removed), totalling 2 152 IR-only surveillance herd tests removed.

In the majority of the 22 946 remaining herd IR-only tests (14 107 or 61%), a single IR was detected and most IRs retested clear in all areas (i.e. an outcome variable result of “No”) (Figure 2), with only 15% (6 296 out of 42 685) of IRs becoming reactors (2 716) or IRs (3 580) at the retest test. A further 2 177 tests were removed as they were carried out in 2020, 17% or 373 of tests having an incident at retest in this group. The 2012-2019 eligible tests remaining were split into training (16 616 tests or 80%) and testing (4 153 tests or 20%) datasets for classification tree and random forest analyses (Figure 1), 21% of tests in these datasets having an incident at retest (3 493 and 873, respectively).

3.2 Data analysis

*3.2.1 Classification tree analysis*

The classification tree using the training dataset (80% observations 2012-2019), with the 16 variables in the top 20 variables with cut-off points list of the original model (Table S4\_a supplementary materials S4), had 25 nodes; 13 of them terminal, and a depth of eight (Figure 3). The root node in the tree (best variable for classification purposes) was *Recurrent (yes/no)*, meaning the herd tested had an incident in the reporting year and a previous one resolved within the previous three years.

The terminal node at the bottom of the tree with the highest proportion of incidents at retest (87% of retests with at least one reactor or IR in 176 tests), was made up of tests in herds without a recurrent incident*,* nora bTB incident resolved in the last year, with a risk score not 5 but 4 (adapted from (Adkin et al., 2016)), with a previous incident not resolved in the last 6-10 years, a surveillance test type low risk (not a back-tracing or a post-incident test), with a previous incident not resolved in the last 1-5 years (either not a previous incident resolved or one resolved over 11 years ago) (Figure 3).

The time (in years) since the last incident was resolved was the most important variable in the tree construction, followed by whether the herd where the IR-only surveillance test took place is recurrent or not (Table S4\_a supplementary materials S4). The classification tree model demonstrated an AUC of 83% on the Test 1 dataset, with a balanced accuracy of 82%, and higher specificity (93%) compared to sensitivity (70%) (Table 1).

3.2.2. Random forest

The final model, using the top 20 variables from the original model (Table S4\_b supplementary materials S4) and trained on the 80% observations of the 2012-2019 dataset, was tuned using 401 trees with eight variables tried at each split, resulting in an estimated out-of-bag (OOB) error rate of 11%. The two most important variables were the same as the classification tree model (Table S4\_b supplementary materials S4). The final model demonstrated an 86% AUC on the Test 1 dataset, a balanced accuracy of 81%, and higher specificity (94%) compared to sensitivity (67%) (Table 2).

3.2.3 Model selection and impact of imbalanced classes data

The classification tree model was chosen over the random forest model, given the similar predictive performance and having the easy-to-interpret tree, which allows for an easier roll out and adoption by operational staff as well as farmers. A simplified version of the tree, showing the variables with cut-off points and whether the terminal node is predominantly “yes” or “no”, could be used by farmers with IR-only surveillance tests to see whether they would be more (“yes” terminal node) or less (“no” terminal node) likely to have additional disease control measures (severe reinterpretation of the disclosing test) applied.

The classification tree model, trained on down-sampled data (3 493 incidents and non-incidents at retest tests) with the same 16 variables, had higher sensitivity (74% vs 70%), lower specificity (89% vs 93%), and the same AUC (83%) compared to the non-downsampled equivalent in the Test 1 dataset (supplementary materials S5).

3.2.4 Evaluation of model outputs against empirical disease control data

The predictive performance was very similar when the 2020 dataset was used as the testing dataset (i.e. Test 2 dataset), using the same training one, with slightly higher specificity (96%), sensitivity (71%), and AUC (85%) (Table 1).

Around one fifth of tests were deemed to be high-risk by the model (incident at retest status or *Any reactor or IR at retest* predicted as “yes”, with predicted probability over 50%) and, therefore, selected for proactive reinterpretation at severe level of the disclosing test: 828 out of 4 153 (20%) tests in the Test 1 dataset and 341 out of 2 177 (16%) in the Test 2 dataset (Table 1). High-risk tests detected two IRs per test, on average, in both the Test 1 (Table 3) and the Test 2 (Table 4) datasets, with the majority IR-only tests and IRs detected in the HRA (Table 4). Incidents were declared at the 60-day retest (i.e. *Any reactor or IR at retest = Yes*) in around three-quarters of these high-risk tests - through the detection of at least one reactor or IR -: 614 out of 828 (74%) in the Test 1 dataset (Table 1) and 265 out of 341 (78%) in the Test 2 dataset (Table 1) and (Figure 4).

However, only a minority of incidents would have been proactively detected at the disclosing test through their reinterpretation at severe level: 12% (74 out of 614) in the Test 1 dataset and 7% (19 out of 265) in the Test 2 dataset (Figure 4). On the other hand, under 10% of high-risk tests that would not have had an incident declared at the 60-day retest would have had incidents declared on disclosure due to severe reinterpretation (17 out of 214 or 8% in the Test 1 dataset and 4 out of 76 or 5% in the Test 2 dataset, respectively) (Figure 4). In the low-risk group of tests (incident at retest or *Any reactor or IR at retest* predicted as “no”), 33 out of 259 (13%) incidents would have been proactively detected through reinterpretation at severe level in the Test 1 dataset and 3 out of 108 (3%) in the Test 2 dataset (Figure 4).

At animal (IR) level, only 22% (1 680 out of 7 655) and 14% (590 out of 4 071) of the IRs in the Test 1 and Test 2 datasets, respectively, were detected in high-risk tests (Figure 5). Around 75% of IRs in high-risk tests were detected in tests where an incident was declared at retest (1 304 in the 2012-19 testing dataset and 463 in the 2020 test dataset). However, only 28% (223 out of 1 304) in the Test 1 dataset and 14% (52 out of 463) in the Test 2 dataset (Figure 5) of IRs disclosed in high-risk tests would have triggered incidents on disclosure, if they were re-interpreted at severe level. Around 15% of IRs disclosed in high-risk tests would have triggered incidents on disclosure, if they were re-interpreted at the severe level, in tests that would not have had incidents declared at retest (55 out of 376 or 15% in the Test 1 dataset and 16 out of 127 or 13% in the Test 2 dataset) (Figure 5). A higher proportion of IRs would have been reactors at severe level in incident-at-retest tests in the low-risk group of tests datasets than in the high-risk group in the Test 1 dataset: 17% or 223 out of 1 304 in the high-risk group vs 20% or 108 out of 542, whereas the opposite is true in the Test 2 dataset: 11% or 52 out of 463 in the high-risk group vs 8% or 14 out of 178 in the low-risk group (Figure 5).

Most of the high-risk tests identified by the model were located in the HRA of England, with Devon having the highest number of IRs in tests in the Test 2 dataset (Figure 6). In 2020, 70% or 238 out of 341 of high-risk tests were located in the HRA, with 70% (413 out of 590) of all IRs and 68% (46 out of 68) of IRs that were reactors at severe level on disclosure detected here (Table 4).

4. Discussion

Machine learning classification algorithms were used to identify high-risk IR-only surveillance tests in England, with regards to having an incident declared at the 60-day retest through the detection of at least one reactor or IR. This triggers an incident in the HRA and the EA (where the majority of IRs are detected), which results in the automatic reinterpretation at severe level not only of the retest test, but also of the disclosing test. Therefore, reinterpretation at severe level of the disclosing test, the measure proposed in high-risk tests, fast-forwards the action taken at the retest in the above scenario, reducing the risk of disease spreading during the 60-day IR isolation period.

This measure will have a higher impact in the HRA, where the highest proportion of IR-only herds (and IRs) were detected in 2020 (64%), although an increasing number are detected in the EA and the LRA (APHA, 2021). This is in part due to the yearly or shorter surveillance tests taking place in the HRA and EA by default, vs four-yearly in the LRA. Demographic and IR-only test datasets show that nearly 60% of active herds had a surveillance test in 2020 in England (78% in the HRA), 8% of those tests revealing IRs in absence of reactors. The immediate positive effect of increasing sensitivity in these tests is therefore likely to be limited, but there is an additional effect of reducing the future bTB risk of IRs in IR-only herds.

The risk of these IRs becoming reactors in the subsequent four years has been estimated as OR 6.85 (95% CI: 5.98-7.86) in the HRA and OR 8.79 (95% CI: 5.92-13.04) in the EA (May et al., 2019) . This risk is substantially reduced through the removal of these cattle by becoming reactors or two-time IRs at the retest (Brunton et al., 2018). Removing them as reactors through reinterpretation at severe level of the disclosing test enhances the reduction in risk. At herd level, IR-only herds have 2.7 times the risk of a bTB incident compared with clear-testing herds in year one (hazard ratio: 2.69; 95% CI: 2.54, 2.84; p<0.001), reducing 63% per year (Brunton et al., 2018).

Although five times more IRs retested as reactors or IRs than would have been reactors at severe (362 vs 68) in high-risk tests in 2020 (Table 4 and Figure 5), over 80% (19 out of 23) of incidents declared at disclosure from this reinterpretation would have been declared at the 60-day retest instead (Figure 4). Therefore, this measure constitutes an initial step in strengthening the mitigating measures against the risk IRs pose. Further measures could be implemented in the high-risk tests identified, if needed, as it is difficult to assess the true bTB infection status of cattle in general and IRs in particular.

IRs could be non-*M. bovis*-infected animals cross-reacting with other pathogens (Clegg et al., 2011a; de la Rua-Domenech et al., 2006; Monaghan et al., 1994). The probability of IRs reinterpreted as reactors being confirmed as bTB-infected through post-mortem examination and/or culture is out of scope in this paper, but it is affected by the lack of a gold standard to accurately differentiate infected from non-infected animals (Álvarez et al., 2012; de la Rua-Domenech et al., 2006). In spite of this, IRs slaughtered before the retest in Ireland and England had higher confirmation rates than clear cattle but lower than reactors (APHA, 2020b; Clegg et al., 2011a).

Only IRs with skin measurements closer to those of reactors, more targeted towards *M. bovis* (de la Rua-Domenech et al., 2006; Francis et al., 1973), will be reclassified as reactors at severe level reinterpretation. However, the reliance on the skin test results is a key limitation of the model since they can be influenced by host and test variables (Álvarez et al., 2012; de la Rua-Domenech et al., 2006), as well as the tester’s operative skills and experience (Clegg et al., 2015), performance, subjective interpretation (Good et al., 2018), and other factors (Clegg et al., 2015) which can affect its sensitivity (Álvarez et al., 2012). The development of a more specific *in vitro* test would overcome some of these obstacles, particularly if blood-based, but it will have to be at least as specific as the skin test to be used widely (EFSA, 2012).

Although severe reinterpretation of SICCT test results enhances the detection of infected cattle (Karolemeas et al., 2012), an alternative would be to place all herds with IR-only tests identified as high-risk under whole-herd movement restrictions, without a chance of having these restrictions lifted if a clear recent bTB history (TB Hub, 2021), pending the re-testing of the still-restricted IRs 60 days later. This measure would prevent transmission to other herds through cattle movements, but it would not prevent spread to other cattle on the farm -if the isolation of IRs was deficient - or to wildlife.

Another option would be to compulsorily slaughter all IRs as dangerous contacts (Clegg et al., 2011b; Godfray et al., 2018). This is done in Ireland, if the number of IRs detected is beyond a specified threshold (currently four) (Phelan, 2020). However, the number of IRs detected in an IR-only test could vary with the farm and other variables, like the animal’s immune response, as well as local cross-reactivity with other mycobacteria such as *M. avium* subsp. paratuberculosis, the causative agent of Johne’s disease (Byrne et al., 2018; de la Rua-Domenech et al., 2006). This measure is therefore likely to attract criticism due to this and to the additional compensation costs (Godfray et al., 2018), since the majority of IRs retest clear (Figure 2).

Finally, the IFN-gamma test, an ancillary blood test, could be used in parallel to the skin test to increase the sensitivity from a median of 66-69% to 95.6% (Álvarez et al., 2012), although the specificity of the IFN-gamma test was estimated as 85.7-90.3%, below the > 99% of the skin test, in the same study. As a laboratory test, its interpretation can be better standardized, although host factors as well as sample quality can affect the results (Álvarez et al., 2012). It is currently only prescribed in certain scenarios due to the increased probability of false positives (de la Rua-Domenech et al., 2006; Defra, 2019, 2014). In fact, this could be higher than the probability of false positives through reinterpretation at severe level of the skin test: from 1 in 5 000 false positives with standard interpretation to 1 in 1 111 at severe (APHA, 2020a) versus 3 in 100 false positive cattle with IFN-gamma (Defra, 2009).

Understanding the reasons why a test is eligible for additional disease control measures is important to harness operational and stakeholder support for implementing the measure, particularly if this leads to a potential increase in the number of reactor cattle slaughtered (true or false positives) and incidents detected. The tree output of the chosen model depicts a set of easy-to-interpret rules showing the combination of variables and cut-off points that explain why a test is declared high-risk, making it a very useful communication tool.

The top two variables for importance in the models’ construction (Table S4\_a supplementary materials S4) and in the highest-risk pathway of the classification tree model (Figure 3), relate to bTB history - one of the most reported risk factors (Broughan et al., 2016) -: an incident in the reporting year and a previous one within three years (i.e. *Recurrent [yes/no]*) and the *Time since the last incident was resolved* [in years]). A version to the latter, restricted to confirmed incidents, has appeared as the top predictive variable in bTB incident models (Romero et al., 2021, 2020).

In terms of the predictive performance of the model, it could be biased by the levels of the outcome variable not being evenly distributed (21% of IR-only surveillance tests had an incident at retest), so even when the model was optimised for sensitivity, its specificity was higher since the majority class of the outcome variable (i.e. the “No” class) was better predicted (Chawla et al., 2002; García et al., 2010; Mostafizur Rahman and Davies, 2013), although the AUC of the down-sampled model was the same. The use of models that incorporate cross-validation mitigates against this and a higher specificity can be adequate in this setting, thanks to the default mitigating measures already in place for all IRs. Still, the application of this model is restricted to IRs detected in surveillance tests (in absence of reactors) and not in other scenarios, such as bTB incidents.

Classification tree models have been used as a guide to patient management (Afonso et al., 2012; Cheng et al., 2018; Fei et al., 2017; Kawamura et al., 2012; Scheetz et al., 2009) and risk identification (Ortiz-Pelaez and Pfeiffer, 2008; Romero et al., 2020; Staerk and Pfeiffer, 1999) following if-then scenarios, similar to deciding which IR-only surveillance tests are subject to additional disease control measures. The final model, with a reduced number of predictors, had over 80% AUC predictive performance without applying prior variable selection procedures (Romero et al., 2021, 2020). In addition, since the balance of sensitivity and specificity was adequate in this scenario, it was not necessary either to manipulate the outputs carrying out ROC analysis (Romero et al., 2021). Together, these make the model easier to adopt and deploy in an operational setting, irrespective of the disease control measure introduced to mitigate the risk of IRs in tests identified as high-risk.

Reinterpretation at severe level is a tool available to case vets to increase the sensitivity of the skin test but, beyond the prescribed scenarios, is appliedunder veterinary discretion. The current lack of consistency has the potential to create resentment in the farming industry, whereby similar high-risk scenarios are treated differently, potentially damaging the eradication effort. Continuing to allow for case veterinarian’s discretion to enhance the management of IRs in certain scenarios (e.g. by removing them as dangerous contacts), can be beneficial. However, the implementation of this methodology means that all high-risk tests would have additional mitigating measures against the risk of IRs consistently applied.

The continuing validity of this methodology, if implemented, is dependent on meeting the required accuracy in identifying high-risk scenarios in future IR-only surveillance tests. In turn, this depends on the quality of the data as well as random elements in the analyses carried out, mitigated by using cross-validated models, among other factors like policy changes that have an impact on surveillance. Although the measure proposed has not translated into many incidents and reactors proactively taken at the disclosing test, the majority of these tests would have yielded an incident at the 60-day retest, preventing potential transmission to cattle and wildlife over these two months. This measure enhances the detection of undetected infected cattle, as recommended in the bTB strategy review (Godfray et al., 2018), although the methodology used to identify high-risk IR-only surveillance tests has a wider application.

1. Conclusion

England needs to accelerate the ongoing reduction in the number of bovine TB incidents if it is to meet its eradication target by the year 2038, particularly in the HRA, where the majority of cases are detected. It is also in this area where over 60% of IRs are revealed and where the highest proportion of IR-only herds sustain an incident in the following 15 months, in the majority of cases at the IRs’ 60-day retest.

We demonstrated the application of classification tree analysis to identify IR-only surveillance tests at higher risk of triggering an incident at the 60-day retest, enabling the rollout of proactive disease control measures at their disclosing test. In particular, we assessed the impact of adopting a severe reinterpretation of the results of this initial test to speed up the detection of infected IRs.

The outputs of the analysis include a tree that provides simple ‘if-then’ scenarios that allow lay people to understand the particular combination of herd-level risk factors that deem a herd test to be high-risk. This facilitates stakeholder and operational engagement with this policy, in addition to harmonising the use of what is currently an *ad hoc* tool, to manage the demonstrated risk of IRs in a consistent way. Severe reinterpretation on disclosure of high-risk IR-only surveillance tests, as proposed, could help accelerate bTB eradication by increasing the early detection of infected cattle.

# Tables

Table 1 a) Classification tree confusion matrix originally calculated on the 20% of 2012-2019 testing dataset (Test 1) showing the reference (Ref) and prediction values (Pred) for outcome “Yes” (“+”) and “No” (“-“). b) Predictive performance of the classification tree analysis in the same testing dataset; c) Confusion matrix on the 2020 dataset used as testing (Test 2), with the same training dataset (80% of 2012-2019 dataset); d) Predictive performance of classification tree analysis on the Test 2 dataset.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **a** |  | Prediction |  | **b** |  |  |  |  |  |  |  |
| Reference |   | Pred+ | Pred- | Total |  | Accuracy | Sensitivity | Specificity | PPV | NPV | Balanced accuracy | AUC |
|  | Ref+ | 614 (70%) | 259 (30%) | 873 (100%) |  | 0.89 | 0.70 | 0.93 | 0.74 | 0.92 | 0.82 | 0.83 |
|  | Ref- | 214 (7%) | 3 066 (93%) | 3 280 (100%) |  |  |  |  |  |  |  |  |
|  | Total | 828 (20%) | 3 325 (80%) | 4 153 (100%) |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| **c** |  | Prediction |  | **d** |  |  |  |  |  |  |  |
| Reference |   | Pred+ | Pred- | Total |  | Accuracy | Sensitivity | Specificity | PPV | NPV | Balanced accuracy | AUC |
|  | Ref+ | 265 (71%) | 108 (29%) | 373 (100%) |  | 0.92 | 0.71 | 0.96 | 0.78 | 0.94 | 0.83 | 0.85 |
|  | Ref- | 76 (4%) | 1 728 (96%) | 1 804 (100%) |  |  |  |  |  |  |  |  |
|  | Total | 341 (16%) | 1 836 (84%) | 2 177 (100%) |  |  |  |  |  |  |  |  |

Table 2 a) Random forest confusion matrix originally calculated on the 20% of 2012-2019 testing dataset (Test 1), showing the reference (Ref) and prediction values (Pred) for outcome “Yes” (“+”) and “No” (“-“). b) Predictive performance of the classification tree analysis in the same Test 1 dataset.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **a** |  | Prediction |  | **b** |  |  |  |  |  |  |  |
| Reference |   | Pred+ | Pred- | Total |  | Accuracy | Sensitivity | Specificity | PPV | NPV | Balanced accuracy | AUC |
|  | Ref+ | 587 (67%) | 286 (33%) | 873 (100%) |  | 0.89 | 0.67 | 0.94 | 0.76 | 0.92 | 0.81 | 0.86 |
|  | Ref- | 188 (6%) | 3 092 (94%) | 3 280 (100%) |  |  |  |  |  |  |  |  |
|  | Total | 775 (15%) | 3 378 (85%) | 4 153 (100%) |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |

Table 3 Evaluation of the impact of proactively reinterpreting IRs in the disclosing test at severe level, compared to the number of IRs that were reactors or IRs at the 60-day retest in high-risk tests predicted by classification tree analysis on the 20% of 2012-2019 testing dataset (Test 1). The “Severe reactor IRs” column shows the number of IRs that would have been reactors at severe level, if the disclosing test had been reinterpreted.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Year | Any Reactor or IR at retest | TotalTests | Total IRs disclosed | Positive IRs | Retest Reactors | RetestIRs | Retest Reactors or IRs |
| 2012 | Yes | 81 | 199 | 47 | 49 | 59 | 108 |
| 2012 | No | 39 | 72 | 11 | 0 | 0 | 0 |
| 2013 | Yes | 101 | 240 | 54 | 46 | 92 | 138 |
| 2013 | No | 37 | 77 | 23 | 0 | 0 | 0 |
| 2014 | Yes | 85 | 166 | 29 | 48 | 61 | 109 |
| 2014 | No | 28 | 54 | 9 | 0 | 0 | 0 |
| 2015 | Yes | 88 | 159 | 12 | 46 | 65 | 111 |
| 2015 | No | 35 | 59 | 9 | 0 | 0 | 0 |
| 2016 | Yes | 79 | 211 | 51 | 51 | 60 | 111 |
| 2016 | No | 20 | 28 | 0 | 0 | 0 | 0 |
| 2017 | Yes | 62 | 117 | 15 | 37 | 39 | 76 |
| 2017 | No | 12 | 25 | 3 | 0 | 0 | 0 |
| 2018 | Yes | 54 | 88 | 3 | 29 | 36 | 65 |
| 2018 | No | 21 | 28 | 0 | 0 | 0 | 0 |
| 2019 | Yes | 64 | 124 | 12 | 30 | 57 | 87 |
| 2019 | No | 22 | 33 | 0 | 0 | 0 | 0 |
|  | Total tests (Pred+) | 828 | 1680 | 278 | 336 | 469 | 805 |

Table 4 Evaluation of the impact of proactively reinterpreting IRs in the disclosing test at severe level, compared to the number of IRs that were reactors or IRs at the 60-day retest in high-risk tests predicted by the classification tree analysis on the 2020 testing dataset (Test 2). The “Severe reactor IRs” column shows the number of IRs that would have been reactors at severe level, if the disclosing test had been reinterpreted.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | Risk area | Any reactor or IR at retest | Total tests | Total IRs disclosed | Severe reactor IRs (%) | Retest reactors (%) | Retest IRs (%) | Retest reactors or IRs (%) |
| 2020 | HRA | Yes | 187 | 332 | 33 | 104 | 161 | 265 |
| 2020 | HRA | No | 51 | 81 | 13 | 0 | 0 | 0 |
| 2020 | EA | Yes | 54 | 72 | 4 | 24 | 39 | 63 |
| 2020 | EA | No | 22 | 40 | 3 | 0 | 0 | 0 |
| 2020 | LRA | Yes | 24 | 59 | 15 | 6 | 28 | 34 |
| 2020 | LRA | No | 3 | 6 | 0 | 0 | 0 | 0 |
|  |  | Total tests (Pred+) | 341 | 590 | 68 (12) | 134 (23) | 228 (39) | 362 (61) |
|  |  | HRA tests (Pred+) | 238 | 413 | 46 | 104 | 161 | 265 |
|  |  | HRA percentage | 70 | 70 | 68 | 78 | 71 | 73 |

# Figures

Figure 1. Data reduction process for the selection of eligible IR-only surveillance tests in England (2012-2020) based on complete 60-day retest results with reactor (R), IR or clear outcomes only. Training and testing datasets split also shown.

Figure 2. a) The majority of eligible IR-only surveillance tests in the study population (2012-2020) detected a single IR; b) the majority of tests did not lead to an incident at retest; c) Sankey graph showing the majority of IRs detected in eligible IR-only surveillance tests retested clear 60 days later in all surveillance risk areas (HRA = High-risk area, EA = Edge area, LRA = Low-risk area).

Figure 3. Classification tree analysis for the probability of any reactors or IRs found at the retest of IR-only surveillance tests, showing the outcome classes’ counts and proportions as well as the overall proportion and number of tests (observations). Nodes in the path leading to the group with the highest proportion of tests with outcome variable = “Yes” are colour-filled in accordance to the scale underneath.

Figure 4. Sankey graphs showing classification tree model outputs’ evaluation at test level using the 20% of 2012-2019 testing dataset (Test 1): a) high-risk tests and b) low-risk tests, and the 2020 testing dataset (Test 2): c) high-risk tests and d) low-risk tests. Outcomes of tests at the 60-day retest, leading to an incident being declared if any reactors or IRs are found, are compared to reinterpreting IRs in the disclosing test at severe level. The latter leads to IRs that either reinterpret as reactors (Positive IRs) triggering an incident on disclosure, or not (Negative IRs).

Figure 5. Sankey graphs showing classification tree model outputs’ evaluation at individual animal or IR level using the 20% of 2012-2019 testing dataset (Test 1): a) high-risk tests and b) low-risk tests), and the 2020 testing dataset (Test 2): c) high-risk tests and d) low risk tests. The number of IRs that retest as reactors, IRs or clear within incident and non-incident at retest tests, is compared to the number of IRs that become reactors at severe level reinterpretation of the disclosing test (Positive IRs).

Figure 6. Distribution of IRs in IR-only surveillance tests deemed to be high-risk according to the classification tree analysis by surveillance risk area a) in the Test 1 dataset (20% of 2012-2019 dataset), b) the Test 2 dataset (2020 dataset), and c) by county in the Test 2 dataset. The majority of IR-only high-risk tests and IRs disclosed are located in the HRA of England in both testing datasets and, at county level, particularly in Devon (Test 2 dataset).



Figure 1



Figure 2



Figure 3



Figure 4



Figure 5



Figure 6

# Conflict of interest declaration

None of the authors of this paper has a financial or personal relationship with other people or organisations that could inappropriately influence or bias the content of the paper.

# Acknowledgements

This research was funded by the Animal and Plant Health Agency and undertaken with the Royal Veterinary College, manuscript approval number 1550275. The authors thank the APHA’s Data Systems Group, which have provided the data as well as useful advice, and to Mr Geoff Jasinski, who read and made relevant comments on drafts. We are particularly grateful to policy and veterinary colleagues in the bTB Programme of Defra, whose input helped us select the final model.

# References

Adkin, A., Brouwer, A., Simons, R.R.L., Smith, R.P., Arnold, M.E., Broughan, J., Kosmider, R., Downs, S.H., 2016. Development of risk-based trading farm scoring system to assist with the control of bovine tuberculosis in cattle in England and Wales. Preventive Veterinary Medicine 123, 32–38. https://doi.org/10.1016/j.prevetmed.2015.11.020

Afonso, A.M., Ebell, M.H., Gonzales, R., Stein, J., Genton, B., Senn, N., 2012. The use of classification and regression trees to predict the likelihood of seasonal influenza. Family Practice 29, 671–677. https://doi.org/10.1093/fampra/cms020

Álvarez, J., Perez, A., Bezos, J., Marqués, S., Grau, A., Saez, J.L., Mínguez, O., de Juan, L., Domínguez, L., 2012. Evaluation of the sensitivity and specificity of bovine tuberculosis diagnostic tests in naturally infected cattle herds using a Bayesian approach. Veterinary Microbiology 155, 38–43. https://doi.org/10.1016/j.vetmic.2011.07.034

APHA, 2021. Bovine tuberculosis in England in 2020: Epidemiological analysis of the 2020 data and historical trends.

APHA, 2020a. Bovine tuberculosis in Great Britain in 2019: Explanatory supplement to the annual reports.

APHA, 2020b. Bovine tuberculosis in England in 2019. Epidemiological analysis of the 2019 data and historical trends.

Baca-Garcia, E., Perez-Rodriguez, M.M., Saiz-Gonzalez, D., Basurte-Villamor, I., Saiz-Ruiz, J., Leiva-Murillo, J.M., de Prado-Cumplido, M., Santiago-Mozos, R., Artés-Rodríguez, A., de Leon, J., 2007. Variables associated with familial suicide attempts in a sample of suicide attempters. Progress in Neuro-Psychopharmacology and Biological Psychiatry 31, 1312–1316. https://doi.org/10.1016/j.pnpbp.2007.05.019

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J.H., Olsen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth Inc.

Broughan, J.M., Judge, J., Ely, E., Delahay, R.J., Wilson, G., Clifton-Hadley, R.S., Goodchild, A. V., Bishop, H., Parry, J.E., Downs, S.H., 2016. A review of risk factors for bovine tuberculosis infection in cattle in the UK and Ireland. Epidemiology and Infection 144, 2899–2926. https://doi.org/10.1017/S095026881600131X

Bruce, P; Bruce, A., 2017. Practical Statistics for Data Scientists, First Edit. ed. O’Reilly Media, Inc.

Brunton, L.A., Prosser, A., Pfeiffer, D.U., Downs, S.H., 2018. Exploring the fate of cattle herds with inconclusive reactors to the tuberculin skin test. Frontiers in Veterinary Science 5, 1–10. https://doi.org/10.3389/fvets.2018.00228

Byrne, A.W., Graham, J., Brown, C., Donaghy, A., Guelbenzu-Gonzalo, M., McNair, J., Skuce, R.A., Allen, A., McDowell, S.W., 2018. Modelling the variation in skin-test tuberculin reactions, post-mortem lesion counts and case pathology in tuberculosis-exposed cattle: Effects of animal characteristics, histories and co-infection. Transboundary and Emerging Diseases 65, 844–858. https://doi.org/10.1111/tbed.12814

Carranza, E.J.M., Laborte, A.G., 2015. Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines). Computers and Geosciences 74, 60–70. https://doi.org/10.1016/j.cageo.2014.10.004

Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W., 2002. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357.

Chen, G., Zhu, Y., Wiedmann, T., Yao, L., Xu, L., Wang, Y., 2019. Urban-rural disparities of household energy requirements and influence factors in China: Classification tree models. Applied Energy 250, 1321–1335. https://doi.org/10.1016/j.apenergy.2019.04.170

Cheng, Z., Nakatsugawa, M., Hu, C., Robertson, S.P., Hui, X., Moore, J.A., Bowers, M.R., Kiess, A.P., Page, B.R., Burns, L., Muse, M., Choflet, A., Sakaue, K., Sugiyama, S., Utsunomiya, K., Wong, J.W., McNutt, T.R., Quon, H., 2018. Evaluation of classification and regression tree (CART) model in weight loss prediction following head and neck cancer radiation therapy 3, 346–355.

Clegg, T.A., Duignan, A., More, S.J., 2015. The relative effectiveness of testers during field surveillance for bovine tuberculosis in unrestricted low-risk herds in Ireland. Preventive Veterinary Medicine 119, 85–89. https://doi.org/10.1016/j.prevetmed.2015.02.005

Clegg, T.A., Good, M., Duignan, A., Doyle, R., More, S.J., 2011a. Shorter-term risk of Mycobacterium bovis in Irish cattle following an inconclusive diagnosis to the single intradermal comparative tuberculin test. Preventive Veterinary Medicine 102, 255–264. https://doi.org/10.1016/j.prevetmed.2011.07.014

Clegg, T.A., Good, M., Duignan, A., Doyle, R., More, S.J., 2011b. Longer-term risk of Mycobacterium bovis in Irish cattle following an inconclusive diagnosis to the single intradermal comparative tuberculin test. Preventive Veterinary Medicine 100, 147–154. https://doi.org/10.1016/j.prevetmed.2011.07.014

Cousins, D. V., 2001. Mycobacterium bovis infection and control in domestic livestock. OIE Revue Scientifique et Technique 20, 71–85. https://doi.org/10.20506/rst.20.1.1263

de la Rua-Domenech, R., Goodchild, A.T., Vordermeier, H.M., Hewinson, R.G., Christiansen, K.H., Clifton-Hadley, R.S., 2006. Ante mortem diagnosis of tuberculosis in cattle: A review of the tuberculin tests, γ-interferon assay and other ancillary diagnostic techniques. Research in Veterinary Science 81, 190–210. https://doi.org/10.1016/j.rvsc.2005.11.005

Defra, 2020. Next steps for the strategy for achieving bovine tuberculosis free status for England - The government’s response to the strategy review, 2018.

Defra, 2019. Wider use of interferon gamma blood testing in the High Risk Area of England [WWW Document]. TB Hub. URL https://tbhub.co.uk/tb-policy/england/wider-use-interferon-gamma-blood-testing-high-risk-area-england/

Defra, 2014. The strategy for achieving Officially Bovine Tuberculosis Free status for England. Defra.

Defra, 2009. Gamma Interferon diagnostic blood test for bovine tuberculosis \_A review of the GB gamma interferon testing policy for tuberculosis in cattle, Defra.

Díaz-Uriarte, R., Alvarez de Andrés, S., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 1–13. https://doi.org/10.1186/1471-2105-7-3

EFSA/ECDC, 2019. The European Union One Health 2018 Zoonoses Report. EFSA Journal 17. https://doi.org/10.2903/j.efsa.2019.5926

EFSA, 2012. Scientific Opinion on the use of a gamma interferon test for the diagnosis of bovine tuberculosis. EFSA Journal 10, 63. https://doi.org/doi:10.2903/j.efsa.2012.2975.

Fei, Y., Gao, K., Hu, J., Tu, J., Li, W., Wang, W., Zong, G., 2017. Predicting the incidence of portosplenomesenteric vein thrombosis in patients with acute pancreatitis using classification and regression tree algorithm. Journal of Critical Care 39, 124–130. https://doi.org/http://dx.doi.org/10.1016/j.jcrc.2017.02.019

Francis, J.D., Choi, C.L., Frost, A.J., 1973. The diagnosis of tuberculosis in cattle with special reference to bovine PPD tuberculin. Australian Veterinary Journal 49, 246–251.

Frisman, L., Prendergast, M., Lin, H.-J., Rodis, E., Greenwell, L., 2008. Applying classification and regression tree analysis to identify prisoners with high HIV risk behaviors. J Psychoactive Drugs 40, 447–458. https://doi.org/doi: 10.1080/02791072.2008.10400651

Gandrud, A.., Gandrud, C.., Russell, K.., Yetma, C.J., 2017. networkD3: JavaScript Network Graphs from R.

Garcia, V.., Mollineda, R.A.., Sanchez, J.S., 2009. Index of balanced accuracy: a performance measure of skewed class distributions. Lecture Notes in Computer Science 5524. https://doi.org/10.1007/978-3-642-02172-5

García, V., Sánchez, J.S., Mollineda, R.A., 2010. Exploring the Performance of Resampling Strategies for the Class Imbalance Problem, in: Trends in Applied Intelligent Systems - 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems. Springer. https://doi.org/10.1007/978-3-642-13022-9\_54

Genuer, R., Poggi, J.-M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recognition Letters 31, 2225–2236.

Godfray, C., Donnelly, C., Hewinson, G., Winter, M., Wood, J., 2018. TB Strategy Review.

Good, M., Bakker, D., Duignan, A., Collins, D.M., 2018. The History of In Vivo Tuberculin Testing in Bovines: Tuberculosis, a “One Health” Issue. Frontiers in Veterinary Science 5. https://doi.org/10.3389/fvets.2018.00059

Good, M., Duignan, A., 2011. Perspectives on the history of bovine TB and the role of tuberculin in bovine TB eradication. Veterinary Medicine International 2011. https://doi.org/10.4061/2011/410470

Gormley, E., Doyle, M.B., Fitzsimons, T., McGill, K., Collins, J.D., 2006. Diagnosis of Mycobacterium bovis infection in cattle by use of the gamma-interferon (Bovigam®) assay. Veterinary Microbiology 112, 171–179. https://doi.org/10.1016/j.vetmic.2005.11.029

Hastie, T., Tibshirani, R., Friedman, J., 2017. The Elements of Statistical Learning, 2nd ed. Springer.

Hayes, T., Usami, S., Jacobucci, R., McArdle, J.J., 2015. Using classification and regression trees (CART) and random forests to analyze attrition: results from two simulations. Psychology and Aging 30, 911–929. https://doi.org/10.1037/pag0000046

Imafuku, A., Sawa, N., Kawada, M., Hiramatsu, R., Hasegawa, E., Yamanouchi, M., Hoshino, J., Ubara, Y., Takaichi, K., 2018. Incidence and risk factors of new-onset hypertrophic pachymeningitis in patients with anti-neutrophil antibody-associated vasculitis: using logistic regression and classification tree analysis. Clinical Rheumatology. https://doi.org/https://doi.org/10.1007/s10067-018-4372-z

James, G., Witten, D., Hastie, T., Tibshirani, R., 2014. An Introduction to Statistical Learning with Applications in R. Springer US. https://doi.org/10.1016/j.peva.2007.06.006

Julien, R., Levy, J.I., Adamkiewicz, G., Hauser, R., Spengler, J.D., Canales, R.A., Hynes, H.P., 2008. Pesticides in Urban Multiunit Dwellings: Hazard Identification Using Classification and Regression Tree (CART) Analysis. Journal of the Air & Waste Management Association 58, 1297–1302. https://doi.org/https://doi.org/10.3155/1047-3289.58.10.1297

Kane, M.J., Price, N., Scotch, M., Rabinowitz, P., 2014. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks, BMC Bioinformatics. https://doi.org/doi:10.1186/1471-2105-15-276

Karolemeas, K.., de la Rua-Domenech, R.., Cooper, R.., Goodchild, A.V.., Clifton-Hadley, R.S.., Conlan, A.J.K.., Mitchell, A.P.., Hewinson, R.G.., Donnelly, C.A.., Wood, J.L.N.., McKinley, T.J., 2012. Estimation of the Relative Sensitivity of the Comparative Tuberculin Skin Test in Tuberculous Cattle Herds Subjected to Depopulation. PLoS ONE 7, e43217. https://doi.org/10.1371/journal.pone.0043217

Kassambara, A., 2018. Machine Learning Essentials: Practical Guide in R. CreateSpace Independent Publishing Platform.

Kawamura, Y., Takasaki, S., Mizokami, M., 2012. Using decision tree learning to predict the responsiveness of hepatitis C patients to drug treatment. FEBS Open Bio 2, 98–102. https://doi.org/http://dx.doi.org/10.1016/j.fob.2012.04.007

Khun, L., Page, K., Ward, J., Worrall-Carter, L., 2014. The process and utility of classification and regression tree methodology in nursing research. Journal of Advanced Nursing 70, 1276–1286. https://doi.org/10.1111/jan.12288

Kirasich, K.., Smith, T.;, Sadler, B., 2018. Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets, SMU Data Science Review.

Kitsantas, P., Hollander M Fau - Li, L., Li, L., 2006. Using classification trees to assess low birth weight outcomes. https://doi.org/doi:10.1016/j.artmed.2006.03.008

Klepp, L.I., Eirin, M.E., Garbaccio, S., Soria, M., Bigi, F., Blanco, F.C., 2019. Identification of bovine tuberculosis biomarkers to detect tuberculin skin test and IFNγ release assay false negative cattle. Research in Veterinary Science 122, 7–14. https://doi.org/10.1016/j.rvsc.2018.10.016

Kuhn, M., 2008. Building predictive models in R using the caret package. Journal of statistical software 28. https://doi.org/10.18637/jss.v028.i05

Kuhnert, P., Venables, B., 2005. An Introduction to R: Software for Statistical Modelling & Computing. Information Sciences 1–364.

Lewis, R.J., 2000. An introduction to classification and regression tree (CART) analysis, in: Annual Meeting of the Society for Academic Emergency Medicine. San Francisco.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2, 18–22. https://doi.org/10.1177/154405910408300516

Maimon, O., Rokach, L., 2010. Data Mining and Knowledge Discovery Handbook, 2nd ed. Springer. https://doi.org/10.1007/978-0-387-09823-4

May, E., Prosser, A., Downs, S.H., Brunton, L.A., 2019. Exploring the risk posed by animals with an inconclusive reaction to the bovine tuberculosis skin test in England and Wales. Veterinary Sciences 6. https://doi.org/10.3390/vetsci6040097

Monaghan, M.L., Doherty, M.L., Collins, J.D., Kazda, J.F., Quinn, P.J., 1994. The tuberculin test. Veterinary Microbiology 40, 111–124. https://doi.org/10.1016/0378-1135(94)90050-7

Mostafizur Rahman, M., Davies, D.N., 2013. Addressing the class imbalance problem in medical datasets. International Journal of Machine Learning and Computing 3, 224–228. https://doi.org/10.7763/IJMLC.2013.V3.307

Nuñez-Garcia, J., Downs, S.H., Parry, J.E., Abernethy, D.A., Broughan, J.M., Cameron, A.R., Cook, A.J., Woolliams, J.A., Greiner, M., Sharp, M., Gunn, J., More, S.J., Welsh, M., Rolfe, S., Clifton-Hadley, R.S., Rhodes, S., Upton, P.A., de la Rua-Domenech, R., Watson, E., Goodchild, A. V., Whelan, A.O., Vordermeier, H.M., 2017. Meta-analyses of the sensitivity and specificity of ante-mortem and post-mortem diagnostic tests for bovine tuberculosis in the UK and Ireland. Preventive Veterinary Medicine 153, 94–107. https://doi.org/10.1016/j.prevetmed.2017.02.017

Ortiz-Pelaez, Á., Pfeiffer, D.U., 2008. Use of data mining techniques to investigate disease risk classification as a proxy for compromised Biosecurity of cattle herds in Wales. BMC Veterinary Research 4, 1–16. https://doi.org/10.1186/1746-6148-4-24

Pedersen, A.B., Mikkelsen, E.M., Cronin-Fenton, D., Kristensen, N.R., Pham, T.M., Pedersen, L., Petersen, I., 2017. Missing data and multiple imputation in clinical epidemiological research. Clinical Epidemiology 9, 157–166.

Phelan, S., 2020. Department confirms TB rule change with clampdown on inconclusive animals. Agriland.

Pollock, J.M., McNair, J., Bassett, H., Cassidy, J.P., Costello, E., Aggerbeck, H., Rosenkrands, I., Andersen, P., 2003. Specific delayed-type hypersensitivity responses to ESAT-6 identify tuberculosis-infected cattle. Journal of Clinical Microbiology 41, 1856–1860. https://doi.org/10.1128/JCM.41.5.1856-1860.2003

R Core Team, 2021. R: A language and environment for statistical computing.

Ramezankhani, A., Pournik, O., Shahrabi, J., Khalili, D., Azizi, F., Hadaegh, F., 2014. Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study 105, 391–398.

Ražanskas, P., Verikas, A., Olsson, C., Viberg, P., 2015. Predicting Blood Lactate Concentration and Oxygen Uptake from sEMG Data during Fatiguing Cycling Exercise. Sensors 98, 20480–20500. https://doi.org/10.3390/s150820480

Romero, M.P., Chang, Y.M., Brunton, L.A., Parry, J., Prosser, A., Upton, P., Rees, E., Tearne, O., Arnold, M., Stevens, K., Drewe, J.A., 2021. A comparison of the value of two machine learning predictive models to support bovine tuberculosis disease control in England. Preventive Veterinary Medicine 188, 105264. https://doi.org/10.1016/j.prevetmed.2021.105264

Romero, M.P., Chang, Y.M., Brunton, L.A., Parry, J., Prosser, A., Upton, P., Rees, E., Tearne, O., Arnold, M., Stevens, K., Drewe, J.A., 2020. Decision tree machine learning applied to bovine tuberculosis risk factors to aid disease control decision making. Preventive Veterinary Medicine 175. https://doi.org/10.1016/j.prevetmed.2019.104860

Saegerman, C., Porter Sr Fau - Humblet, M.F., Humblet, M.F., 2011. The use of modelling to evaluate and adapt strategies for animal disease control. Revue scientifique et technique (International Office of Epizootics) 30, 555–569.

Scheetz, L.J., Zhang, J., Kolassa, J., 2009. Classification tree modeling to identify severe and moderate vehicular injuries in young and middle-aged adults 45, 1–10. https://doi.org/doi:10.1016/j.artmed.2008.11.002

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., Khovanova, N., 2019. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. Biomedical Signal Processing and Control 52, 456–462. https://doi.org/https://doi.org/10.1016/j.bspc.2017.01.012

Song, Y.-Y., Lu, Y., 2015. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry 27, 130–135.

Speiser, J.L., Karvellas, C.J., Wolf, B.J., Chung, D., Koch, D.G., Durkalski, V.L., 2019. Predicting daily outcomes in acetaminophen-induced acute liver failure patients with machine learning techniques. Computer Methods and Programs in Biomedicine 175, 111–120. https://doi.org/10.1016/j.cmpb.2019.04.012

Staerk, K.D.C., Pfeiffer, D.U., 1999. The application of non-parametric techniques to solve classification problems in complex data sets in veterinary epidemiology - An example. Intelligent Data Analysis 3, 23–35.

Strobl, C., 2010. An introduction to recursive partitioning: rationale, application and characteristics of classification. Psychol Methods 14, 323–348. https://doi.org/10.1037/a0016973

TB Hub, 2021. Actions when an inconclusive reactor is found [WWW Document]. URL https://www.tbhub.co.uk/advice-during-a-tb-breakdown/actions-when-an-ir-is-found/

Therneau, T.M., Atkinson, E.J., 2018. An introduction to recursive partitioning using the rpart routines. R package version 4.1-15.

UK Statutory Instruments, 2021. The Tuberculosis in Animals (England) Order 2021.

Wałęga, G., Wałęga, A., 2021. Over-indebted Households in Poland: Classification Tree Analysis. Social Indicators Research 153, 561–584. https://doi.org/10.1007/s11205-020-02505-6

White, P.C.L., Böhm, M., Marion, G., Hutchings, M.R., 2008. Control of bovine tuberculosis in British livestock: there is no “silver bullet.” Trends in Microbiology 16, 420–427. https://doi.org/10.1016/j.tim.2008.06.005

Yang, T., Gao, X., Sorooshian, S., Li, X., 2016. Simulating California reservoir operation using the classification and regression-tree algorithm combined with a shuffled cross-validation scheme. Water Resources Research 52, 1626–1651. https://doi.org/doi:10.1002/ 2015WR017394