Title page

A comparison of the value of two machine learning predictive models to support bovine tuberculosis disease control in England

M. Pilar Romero[a,b*], Yu-Mei Chang[b], Lucy A. Brunton[b], Alison Prosser[a], Paul Upton[a], Eleanor Rees[a], Oliver Tearne[a], Mark Arnold[a], Kim Stevens[b] and Julian A. Drewe[b]

[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): 4 774.

Abstract

Nearly a decade into Defra's current eradication strategy (Defra, 2014, 2011), bovine tuberculosis (bTB) remains a serious animal health problem in England, with c.30,000 cattle slaughtered annually in the fight against this insidious disease. There is an urgent need to improve our understanding of bTB risk in order to enhance the current disease control policy. Machine learning approaches applied to big datasets offer a potential way to do this. Regularized regression and random forest machine learning methodologies were implemented using 2016 herd-level data to generate the best possible predictive models for a bTB incident in England and its three surveillance risk areas (High-risk area [HRA], Edge area [EA] and Low-risk area [LRA]). Their predictive performance was compared and the best models in each area were used to characterize herds according to risk.

While all models provided excellent discrimination, random forest models achieved the highest balanced accuracy (i.e. average of sensitivity and specificity) in England, HRA and LRA, whereas the regularized regression LASSO model did so in the Edge (EA). The time since the last confirmed incident was resolved was the only variable in the top-ten ranking in all areas according to both types of models, which highlights the importance of bTB history as a predictor of a new incident.

Risk categorisation based on Receiver Operating Characteristic (ROC) analysis was carried out using the best predictive models in each area setting a 99% threshold value for sensitivity and specificity (97% in the LRA). Thirteen percent of herds in the whole of England as well as in its HRA, 14% in its EA and 31% in its LRA were classified as high-risk. These could be selected for the deployment of additional disease control measures at national or area level. In

this way, low-risk herds within the area considered would not be penalised unnecessarily by blanket control measures and limited resources be used more efficiently. The methodology presented in this paper demonstrates a way to accurately identify high-risk farms to inform a targeted disease control and prevention strategy in England that supplements existing population strategies.

Keywords: *Bovine tuberculosis, Machine learning, Random forest, Regularized regression, ROC analysis, England.*

1. Introduction

Bovine tuberculosis (bTB: infection of cattle with *Mycobacterium bovis*) is a global bacterial zoonosis, reported in 44% of 188 OIE territories from January 2017 to June 2018 (Murai et al., 2019). It can affect nearly all mammals although cattle are the most susceptible (Hamzi, 2014). It represents a serious economic problem globally (Olea-Popelka et al., 2017; Pollock and Neill, 2002) and is one of the most complex (Brooks-Pollock and Keeling, 2009) and pressing animal health problems in the UK (Defra, 2014). Although current bTB prevalence levels of 6 to 14% (Defra, 2019) are below the estimated 20-40% prevalence pre-compulsory controls in the 1940s-50s (Defra, 2006), they do not allow for eradication (Pfeiffer, 2013) in England by the target year of 2038 (Defra, 2014). A successful strategy would require a significant reduction in levels of bTB from the current 7.4% prevalence of confirmed incidents in the High Risk Area (HRA) of England (APHA, 2020) to 0.1% of confirmed incidents over a six-year period (European-Commission, 1964), with interventions targeted according to the risk of infection (Defra, 2014).

The established blanket application of whole-herd test-and-slaughter and abattoir surveillance programmes underpinned by animal identification, tracing and movement control, are cornerstones of the bTB eradication strategy. More refined risk-based strategies, central to eradication programmes in Australia (More et al., 2015), were introduced in 2013 in England with the establishment of the High-risk, Edge and Low-risk areas (Defra, 2014). However, individual risk-based designation of farms, recently recommended (Godfray et al., 2018), has seldomly been attempted (Adkin et al., 2016). This would enable the proactive application of prevention and disease intervention measures in absence of an incident and on incident disclosure, respectively, to further limit disease spread.

Decision-making for disease prevention and control is based on quantitative data analysis and the interpretation and validity of models depend on the epidemiological knowledge about the disease as well as the quality and quantity of data used (Thursfield, 2005). The science of learning from data plays a key role in the fields of statistics, data mining and artificial intelligence applied to multiple disciplines. Non-linear decision tree methods are simple and easy-to-interpret models (James et al., 2014) that account for interactions and non-linear relationships (Afonso et al., 2012; Fei et al., 2017; Schiltz et al., 2018), without making distributional assumptions (Frisman et al., 2008; Kashani and Mohaymany, 2011), without restrictions in predictor numbers (Frisman et al., 2008; Shaikhina et al., 2019; Strobl, 2010) and without the need to transform variables (Fei et al., 2017; Frisman et al., 2008; Lewis, 2000; Shaikhina et al., 2019; Song and Lu, 2015).

This paper builds on previous research (Romero et al., 2020) that used classification tree analysis to provide explanatory models of bTB in England and its three surveillance risk areas (HRA, LRA, EA). Here, we compare two predictive models: random forests and regularized

logistic regression. Random forests (Breiman, 2001; Liaw and Wiener, 2002) are 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 losing interpretability (Hastie et al., 2017; James et al., 2014). Regularized logistic regression, on the other hand, is a method which penalises the number of variables in traditional logistic regression models and selects the ones that contribute more to it (Friedman et al., 2010), trying to improve both accuracy and simplicity (Kassambara, 2018). They reduce the variance of traditional linear models maintaining predictive performance (Hastie et al., 2017) and are preferred to subset approaches in terms of bias (Bielza et al., 2011; Kwok et al., 2014). Like random forest models, they deal well with multi-collinearity, reduce the numerical instability due to overfitting (Pereira et al., 2016) and are useful in relatively high dimension scenarios (Bielza et al., 2011). Both modelling approaches produce estimates of risk on a continuous scale for each farm, whereas we aim to classify herds into risk categories to inform the targeted deployment of disease prevention and control measures.

Receiver Operating Characteristic (ROC) analysis was first used in the late 1960s to select cut-off values for medical diagnostic tests (Greiner et al., 2000) and is now widely accepted for evaluating the discrimination performance of a continuous variable (Fawcett, 2006; Gerds et al., 2008; Gonçalves et al., 2014; van Erkel and Pattynama, 1998). To facilitate the interpretation of the predicted probability of an incident- output of the predictive models- and using it as the continuous variable to be evaluated (Gerds et al., 2008), ROC analysis allows us to classify herds into high-, medium- and low-risk. The selection of the two thresholds or cut-off values needed to separate the herds is determined by the chosen values of sensitivity

and specificity (Schafer (1989) in (Greiner et al., 2000)). This study aimed to demonstrate a methodology that could extend the application of the bTB predictive models to devise risk-based disease control and/or prevention strategies at herd level in England and its differing incidence areas, to supplement the population-level control measures currently applied.

2. Methods

2.1 Source datasets

Animal and Plant Health Agency (APHA)-held and other data on potential herd-level predictors for herds active in England in 2016 were used, ranging from demographic herd characteristics and bTB-related variables (e.g. past bTB history from as early as January 2000, including the status in 2016 as incident or not as the outcome variable) from the Sam bTB management system, to cattle movements from the Cattle Tracing System (from as early as January 2012), badger density (Judge et al., 2017) and land class data (Bunce et al., 2007). UK climate data variables were extracted from the gridded land surface climate observations datasets (Met_Office, 2017): maximum, mean and minimum temperature and rainfall from the daily temperature and precipitation at 5 km resolution datasets (2013-2016); relative humidity (2011-2014) and sunshine (2013-2016) from the monthly climate variables at 5 km resolution datasets.

2.2 Data reduction

Non-eligible herds were excluded:

- Being a government-approved finishing unit (i.e. Approved Finishing Unit, Licensed Finishing Unit and Exempt Finishing Unit). These are biosecure finishing units officially licensed and monitored by the government that can receive cattle from bTB-

restricted premises (first two) (APHA, 2018a, 2017a) and from premises that have not had their required pre-movement test in the latter case (APHA, 2018b), but can only send cattle to slaughter. Movements to these represent a deferred slaughter, possibly beyond the study year;

- Not having a value for herd size, a key predictor based on previous studies (Broughan et al., 2016; Skuce et al., 2012), and
- Not having a chance of an incident being detected in 2016 due to absence of active (disease testing) and passive (slaughterhouse) surveillance.

2.3 Data preparation

Proximity variables to bovine and non-bovine bTB incidents (i.e. from any non-bovine species where bTB has been confirmed on culture from 2008 to 2016), namely the herd's rounded distance as the crow flies to nearest incident occurring in 2015 and to the nearest non-bovine incident, respectively; as well as the land class value, were extracted at herd level using ArcMap™ extraction tool.

2.4 Data analysis

2.4.1 Descriptive data analysis

The initial dataset used for analysis was made up of the outcome variable (i.e. incident or not in 2016) and 141 predictors, which included factors such as number of incidents, number of movements and area incidence (for a full list, see Supplementary materials S1). The presence of missing values was assessed and dealt with by either removing herds with any missing observations (6.12% or 2 461 out of 40 184 herds removed) (complete-case analysis) (Hayes

et al., 2015; Maimon and Rokach, 2010; Pedersen et al., 2017) or by substituting missing observations using multiple imputation (Afifi et al., 2011; Maimon and Rokach, 2010; Pedersen et al., 2017) with chained equations (van Buuren, 2011) to reduce the bias in estimates of missing values, since these are based on the distribution of observed data (White et al., 2011). Numerical variables were not categorized. The proportions of incident and non-incident herds in England, High-risk area (HRA), Edge area (EA) and Low-risk area (LRA) were presented (Figure 1).

2.4.2 Variable selection

To improve the speed and performance of the algorithms, non-predictive variables were identified and removed (Guyon and Elisseeff, 2003; Jain and Singh, 2018; Maimon and Rokach, 2010) in three steps. First, univariable logistic regression analysis was carried out to reveal associations between each individual predictor and the outcome, removing non-significant variables with a relaxed threshold ($p > 0.1$) (Jain and Singh, 2018; Winkler and Mathews, 2015). This relaxed threshold was chosen since non-significant variables could still improve predictive performance in the presence of others (Guyon and Elisseeff, 2003; Hilbe, 2009). Second, the presence of highly-correlated variables was identified by a correlation coefficient (detected using the Spearman test) above 0.79 in absolute value (Campbell and Swinscow, 2009). Among highly-correlated pairs of numerical variables, the one with the lowest mean correlation between that predictor and all other ones was selected and the rest excluded (Kuhn, 2008). Categorical variables were assessed using the Cramer's V test (Cramer, 1946), followed by the manual selection of certain variables within highly-correlated pairs based on practical criteria with the remaining highly-correlated pairs being excluded. Selected highly-correlated and non-highly correlated variables entered the next

step. Third, predictors with near-zero variation (i.e. the ratio of the number of unique values relative to the total number of observations was less than 20% and the ratio of the most frequent value to the second most frequent one was greater than 20) were removed (Kuhn, 2008). A final check for the presence of linear dependencies was also carried out using QR matrix decomposition (Kuhn, 2008).

2.4.3 Random forest

Random forest models (Breiman, 2001), based on an ensemble of classification trees (Breiman et al., 1984; Therneau and Atkinson, 2018), were implemented (Liaw and Wiener, 2002) using training datasets resulting from randomly splitting the original datasets using an 80:20 (training: testing) split (Fei et al., 2017; Kassambara, 2018; Kawamura et al., 2012; Yang et al., 2016). Models were created for England and each surveillance risk area using training datasets with complete-case or with multiple imputation of missing values. Bootstrapped samples of herds drawn with replacement from each 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).

Variable importance was calculated by first recording the OOB prediction accuracy to get an unbiased estimate of the misclassification error (Genuer et al., 2010; Strobl, 2010). This calculation was then repeated after permuting each predictor variable, with the difference between the two accuracies being averaged over all trees and normalized by the standard deviation of the differences (Hastie et al., 2017; Liaw and Wiener, 2002). The predictions for

193 a given herd having an incident or not were assigned by aggregating the results from all trees
194 using majority voting (Boulesteix et al., 2012; Hastie et al., 2017; Maimon and Rokach, 2010)
195 (Figure 2).
196 To alleviate the problem of imbalanced class proportions of the outcome, the analyses were
197 repeated using a down-sampling approach within the training datasets, independent of the
198 cross-validation process. Down-sampled datasets were created by selecting a random sample
199 of non-incident herds matching the number of incident ones (Chawla et al., 2002; Garcia, V.;
200 Mollineda, R.A.; Sanchez, 2009; Kuhn, 2008; Mostafizur Rahman and Davies, 2013).

201 *2.4.4 Regularized logistic regression*
202 To provide an alternative predictive model for comparison, regularized logistic regression was
203 applied to the same data. Three regularized regression models were developed using the same
204 training datasets (Friedman et al., 2010): Ridge regression (Hoerl and Kennard, 1982) that
205 shrinks the predictors' regression coefficients towards zero but keeps all variables in the
206 model, LASSO (Least Absolute Shrinkage and Selection Operator) that shrinks to the point of
207 deselecting some coefficients by reducing them to zero (Tibshirani, 2011) and Elastic net, a
208 combination of the two and generalization of the LASSO (Zou and Hastie, 2005). The models
209 were developed and tuned using leave-one-out cross-validation (James et al., 2014), with the
210 choice of model being set by selecting the mixing parameter (α): "0" for Ridge, "1" for
211 LASSO or unspecified for Elastic net (the best value between 0 and 1 is selected from a grid
212 by the statistical package) (Friedman et al., 2010; Kassambara, 2018). All models were
213 developed using multiple imputation and complete-case, original and down-sampled datasets
214 in each area, as before. The model's predictive performance and best model's selection was

carried out as before. Output variables were chosen from the best models' regression coefficients, excluding predictors that had a null value.

2.4.5 Comparison of model performance

The models' predictive performance was assessed on the testing datasets (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 models with the highest balanced accuracy were chosen for each area. Output variables were chosen from the coefficient ranking excluding variables without a coefficient value (LASSO) and from the variable importance ranking (OOB accuracy) excluding variables with values of zero, negative or positive up to the same value as the negative ones (random forest) (Strobl, 2010).

2.4.6 Receiver Operating Characteristic (ROC) analysis

We carried out an ROC analysis based on these calculated probabilities (Sing et al., 2005) to discriminate between three mutually-exclusive risk categories of herds by selecting two thresholds, informed by defined sensitivity and specificity values. Specificity was first calculated from the false positive ROC analysis outputs, which were ordered in descending value of cut-off values. A requirement of 99% sensitivity and specificity yielded two different cut-offs in England, HRA and EA. In the LRA, 97% sensitivity and specificity values were used due to the lack of incidents. The cut-off values chosen represented the predicted probabilities of an incident in the training datasets for each area that classified herds into low-, medium- and high-risk groups. The predicted probability of an incident was then calculated

for the testing dataset in each area (R_Core_Team, 2020) and the same cut-off values were applied to inform risk-based classification of herds in the testing datasets.

Within each area (i.e. England and each surveillance risk area), complete-case vs imputed data datasets (to evaluate the influence of missing data) and down-sampled vs not down-sampled datasets (to evaluate influence of imbalanced data) were analysed. Statistical analyses were performed using the R statistical software version 3.6.0. and manipulation of spatial data was carried out in ESRI ArcMap 10.6.1.

3. Results

3.1 Summary of data

There were 52 668 active cattle herds in England in 2016; 392 or 0.74% government-approved finishing units, 109 or 0.21% herds without a value of herd size and 11 983 or 22.75% herds without a chance of an incident being detected were removed, leaving 40 184 herds to be included in the analysis. The variable with the highest percentage of missing values was *Prevalence in 100 nearest neighbours* (2.42%). Nine percent of herds (3 561 out of 37 723 in complete-case and 3 639 out of 40 184 in multiple imputation datasets) had had a new incident in 2016: 86% (3 067 and 3 134) in the HRA, 10 % (367 and 374) in the EA and 4 % (127 and 131) in the LRA. These proportions mimic the proportions reported for 2016 in all active herds, although HRA herds are over-represented due to data reduction (APHA, 2017b).

3.2 Variable selection

Sixty-five of the 141 variables remained in the analysis, following removal at different stages (nine after univariable logistic regression analysis; 59 after correlation analysis and eight after near-zero variance analysis). No linear dependencies were detected at the final check and so no additional variables were removed at this stage. The manual selection of categorical variables in highly-correlated pairs was carried out prioritising ease of extraction (*Movement on 2014-2016* chosen over *Movement on 2012-2016; Incident in 2015* chosen over *Reactors at incident disclosure in 2015;* and *Surveillance risk area* chosen over *County*) and information value (*Time since last confirmed incident* chosen over *Previous confirmed incident resolved (yes/no)* and *Previous confirmed incident (yes/no)*). Eight near-zero variance variables were removed, two categorical binary (most frequent class in 99% and 98% of observations, respectively) and six numerical (ratio of most frequent to second most frequent value ranging from 22.69 to 43.60).

3.3. Data analysis

3.3.1 Random forest

The best models were tuned with between 94 trees in the EA and 403 trees in the HRA, with 16 variables tried at each split in all areas except the EA area (32 variables). The estimated OOB error rates were 16% in England, 21% in the HRA and 22% in the EA and LRA areas. The predictor with the highest variable importance in England was *Time since the last confirmed incident was resolved*, in the HRA and EA was the *Number of slaughterhouse destinations*, whereas in the LRA it was *Surveillance tests* (Table 2).

3.3.2. Regularised regression

The best LASSO models had a mixing parameter (λ) of between 0.0013 in England and 0.0327 in the LRA. The predictor with the largest coefficient in absolute value in England, the HRA and the EA was *Time since the last confirmed incident was resolved* (0-2 years), whereas in the LRA it was *Inconclusive reactors only* (yes) (Table 2).

3.3.3 Predictive performance comparison

The best random forest and regularized regression models used down-sampled datasets in all areas, with non-down-sampled equivalents showing a 19-41% and a 22-29% reduction in balanced accuracy, respectively, mainly due to a marked drop in sensitivity (Supplementary materials S2). LASSO models performed best in all areas except England as a whole, but this was chosen over the best one (Ridge) for ease of presentation; having only two centesimal points' lower sensitivity (equal accuracy, balanced accuracy, specificity and AUC). The best models were developed using down-sampled multiple imputation datasets, which had higher balanced accuracy compared to their complete-case equivalents in all areas except the EA. The best random forest models had one centesimal higher balanced accuracy compared to the best LASSO models in all areas except the EA (two centesimal points higher). These models showed excellent discrimination ability in all areas (AUC>=0.80 and <0.90) (Hosmer et al., 2013), with the random forest models in England and the LRA and the LASSO ones in the EA and the LRA being outstanding (AUC>=0.90: Hosmer et al., 2013) in this respect (Table 1).

Time since the last confirmed incident was resolved was the only variable in the top-ten rankings by variable importance in the random forest model and by absolute value of coefficients in the LASSO model (ten unique variables in decreasing order were selected

among the predictors' ranking) in all areas (Table 2). Two further variables were common to all areas in random forest models (*Number of slaughterhouse destinations* and *Prevalence in 100 nearest neighbours*) and three further ones were common to all areas in LASSO models (*Surveillance tests (yes/no)*, *Inconclusive reactors only (yes/no)* (i.e. in the study year or 2016) and *Inconclusive reactors only in 2015 (yes/no)*). Among the full rankings of selected variables, nine were common to both types of models in all areas (nine out of 13 common to all areas in random forest models and all nine common in LASSO models): *Number of slaughterhouse destinations*, *Maximum residence time*, *Prevalence in 100 nearest neighbours*, *Low- and High-risk neighbours in 1 km radius*, *Inconclusive reactors only (yes/no)* and *in 2015 (yes/no)*, *Time since last confirmed incident was resolved and Surveillance tests (yes/no)* (Supplementary materials S3).

3.3.4. Receiver Operating Characteristic analysis

The training datasets from the best models: random forest using down-sampled multiple imputed data in England, HRA and LRA, and LASSO using down-sampled complete-case data in the EA, were used in ROC analysis (Figure 3). The 99% threshold values for sensitivity and specificity chosen in England, HRA and EA resulted in one percent false negative (i.e. incidents in the low-risk group) (29, 26 and three, respectively) and false positive (i.e. non-incidents in the high-risk group) (28, 25 and two, respectively) herds (Table 3). The 97% thresholds chosen in the LRA yielded 4 % (four herds) of false negatives in the low-risk group of herds and 3% (three herds) of false positives in the high-risk group of herds.

In England and the HRA the same threshold values of predicted probabilities in the testing dataset yielded 94 and 84 (13%) high-risk incident herds. Eight and five incident herds (1%)

were missed in the low-risk groups and 62 and 42 non-incident herds (1%) were included in the high-risk group, respectively (Table 3). In the EA testing dataset ten (14%) incident herds were classified as high-risk, whereas five (7%) incident herds were missed in the low-risk group and ten non-incident herds (1%) were included in the high-risk group. In the LRA, eight out of 26 incident herds (31%) were classified as high-risk. No incident herds were missed due to inclusion in the low-risk group but seventy-five non-incidents herds (3%) were present in the high-risk group.

4. Discussion

Random forest and regularized regression predictive models for a bTB incident herd in England were developed and compared, and their outputs used to classify cattle herds within a multiclass system (high, medium and low) according to risk. This was based on several risk factors, hence achieving very good levels of accuracy (McLaren et al., 2010). However, the aim was not to substitute population-level measures (e.g. background surveillance testing regime, default protocol of intervention in incident herds) but to supplement them in a cost-effective way (Rose, 2001).

The best predictive models had even higher AUC values than the classification tree analysis models developed using the same datasets in all areas (Romero et al., 2020). This was expected in the case of random forest -an improved algorithm of the same methodology (Breiman, 2001; Liaw and Wiener, 2002)- but LASSO-informed regression models also performed better, according to this metric.

The ranking of predictors provided by model-specific variable importance measures in the case of random forest -or by a coefficient list in regularized regression LASSO- provided important outputs which may also be used to support disease control decisions (Verikas et al.,

2011). Both predictive models included *Time since last confirmed incident was resolved* in the top-ten ranking of variables in all areas. bTB history is one of the most consistently identified risk factors for bTB in cattle herds (Broughan et al., 2016) and LASSO regression outputs narrowed the timespan of higher risk to 0-2 years.

The detection of *Inconclusive reactors only* (i.e. in absence of reactors) in surveillance tests was among the top-ten variables according to both models in England, the EA and LRA (in England and the LRA inconclusive reactors only in the previous year was also a top-ten variable in both models). These findings support the high-risk status of these animals (Brunton et al., 2018; Clegg et al., 2011a, 2011b; May et al., 2019), which are subject to lifetime movement restrictions in England since 2017 (APHA, 2017c).

Neighbouring cattle herds (high- and low-risk) in a 1 km radius are considered contiguous neighbours and they are among the top-ten variables list of both models in the HRA and EA. In the LRA, the *Prevalence in 100 nearest neighbours* is a top-ten variable according to both models. A review of bTB risk factors found that the occurrence of bTB on contiguous premises and/or the level of bTB in surrounding areas (infection pressure) was one of the most consistently-identified herd-level risk factors (Skuce et al., 2012). The presence of low-risk neighbours decreased the risk (negative LASSO coefficient) whereas the presence of high-risk ones increased it (positive LASSO coefficients) in all areas (Supplementary materials S4). The mechanism by which contiguous neighbours exert their influence is not investigated in this paper but bTB could spread between such holdings via direct (cattle break-ins or nose-to-nose contact over fences) or indirect contact (fomites or infected wildlife reservoir accessing both herds) (Phillips et al., 2003).

An open incident at the end of the previous year was a top-ten variable according to both models in the HRA and found to be protective (negative LASSO coefficient). This has been reported previously (Romero et al., 2020) and could be due to a lack of time to detect a new incident the following year due to the 60-day within-incident testing interval and a six-month interval until the first post-incident test is applied.

The density of reactors where the herd was located and the number of different slaughterhouse destinations were among the top-ten variables according to both models and increased the risk of an incident in the EA. The presence of reactors in the area could be a proxy for proximity of incidents. An increasing risk with the number of different slaughterhouse destinations (positive LASSO coefficient) could be the result of new destinations recorded due to the slaughter of reactors following incident disclosure in APHA-contracted slaughterhouses (if different from the herd's usual one/s). It could also imply an increased probability of detection by passive surveillance in non-incident herds, as different slaughterhouses have different performance (McKinley et al., 2018). In the LRA, the only other variable among the top-ten according to both models is *Surveillance test (Yes/No)* in the study year, increasing the risk of an incident if a surveillance test took place. The relevance of having a surveillance test with regards to the risk of an incident is consistent with the fact that only 19% of incidents were detected using passive surveillance in slaughterhouses in the LRA in 2016; the rest being detected with active surveillance using tests in cattle (APHA, 2017b).

The predicted probability of an incident was calculated using all variables that contributed to the model since taking a few risk factors in isolation would only provide a partial view of the risk profile of herds. There are some limitations to this study, like not elucidating

transmission pathways leading to a bTB incident in different herds, which is the subject of field disease investigation visits (TBhub, 2020).

The thresholds or cut-off values that control how predicted probabilities are converted into risk categories using ROC analysis were chosen arbitrarily to be high but of equal sensitivity and specificity. The same methodology can be applied using different cut-off points to maximise either parameter depending on the relative cost of false positives compared to false negatives. However, the optimal cut-off values for classification would involve clinical and other considerations, like costs, benefits and risks that affect stakeholders (Godfray et al., 2013).

In practice, the predictive model's algorithms and subsequent classification methodology could be automated enabling the deployment of suitable measures in a risk-based targeted approach. For example, non-incident high-risk herds could be subject to prevention measures such as additional advisory visits to the farmer, increased engagement with local vet practices, or spot-check surveillance tests. Seemingly, if a high-risk herd is involved in a bTB incident, enhanced interventions could be introduced proactively to mitigate the extent of the incident. These measures are introduced by default when an incident continues beyond 18 months, at which point it is declared persistent (AHVLA, 2014). The enhanced management of persistent incidents that ensues includes, for example, a more thorough disease investigation visit, drawing individual action plans and allowing more flexibility to carry out additional tests beyond the ones prescribed, to prioritise the detection of infected cattle. Introducing these measures in the small percentage of high-risk incident herds could have a positive effect in the epidemic beyond the benefits to the individual herds.

411 We have demonstrated a methodology to inform a risk-based approach to enhance the bTB
412 disease control in England, supplementing existing population-level or blanket measures.
413 With this information, strategies for deploying adequate prevention and/or disease control
414 interventions can also be designed at both primary (i.e. prior to an incident, based on the
415 herds' risk factors) and secondary (i.e. once an incident is declared in a targeted herd) level
416 (Platt et al., 2017).

417 **4. Conclusion**

418 The application of two of the most well-known machine learning predictive classification
419 algorithms to the prediction a bTB incident in one of the highest-incidence areas in the
420 developed world resulted in high-performing output models. Random forest models were
421 better in terms of balanced accuracy than LASSO equivalents in England, HRA and LRA but,
422 nonetheless, there was a degree of overlap in the most important variables selected by both
423 models; strengthening their relevance as risk factors for the disease. Outputs from the best
424 predictive models in each area were used to classify herds according to risk in a multi-class
425 system (high-, medium- and low-risk). This demonstrated their application to inform the
426 targeted deployment of disease control and prevention measures, supplementing current
427 population-level measures. This methodology can be adapted to a wide variety of disease
428 control scenarios in humans, animals or multi-host systems like bTB, as long as sufficient
429 data on risk factors is available. A single or more predictive models can be used to calculate
430 the predicted probability of a case. A multi-class or an alternative risk classification
431 framework, like the more traditional binary one, is also possible. Finally, the outputs of our
432 predictive models may help identify the likely reduction in risk following the deployment of
433 targeted bTB prevention and control measures.

Tables

Table 1 Predictive performance indicators of the best random forest and LASSO models in England and its surveillance risk areas on their respective testing datasets. The model with the highest balanced accuracy in each area is shaded grey. CC= Complete-case, MI= Multiple Imputation, PPV=Positive Predictive Value, NPV=Negative Predictive Value, AUC=Area Under the ROC, HRA=High-risk area and EA= Edge area and LRA= Low-risk area.

Random forest	Accuracy	Sensitivity	Specificity	PPV	NPV	Balanced accuracy	AUC
England (downsampled MI)	0.81	0.86	0.80	0.30	0.98	0.83	0.91
HRA (downsampled MI)	0.78	0.84	0.77	0.39	0.96	0.80	0.88
EA (downsampled CC)	0.71	0.90	0.70	0.16	0.99	0.80	0.85
LRA (downsampled MI)	0.81	0.92	0.81	0.05	1.00	0.87	0.93
LASSO	Accuracy	Sensitivity	Specificity	PPV	NPV	Balanced accuracy	AUC
England (downsampled MI)	0.81	0.82	0.81	0.30	0.98	0.82	0.89
HRA (downsampled MI)	0.78	0.80	0.77	0.38	0.96	0.79	0.86
EA (downsampled CC)	0.81	0.84	0.80	0.21	0.99	0.82	0.90
LRA (downsampled MI)	0.85	0.88	0.85	0.05	1.00	0.86	0.94

Table 2 Top-ten ranking (R) of selected variables from the best random forest models, and predictors (up to ten distinct variables) from the best LASSO models (variables common to both models for an area appear in bold print). Variables are ranked by variable importance (for random forest) and by coefficient absolute value (for LASSO). A fill colour indicates their distribution, with the England column showing only if a variable is in all areas: All areas, HRA, EAand LRA, HRA and EA, HRA and LRA, EA and LRA. Predictors with duplicated variables in LASSO models are shown in italics.

Model	R	England	HRA	EA	LRA
Random forest	1	**Time since last confirmed incident was resolved**	Number of slaughterhouse destinations	**Number of slaughterhouse destinations**	Surveillance tests
	2	Number of slaughterhouse destinations	**Time since last confirmed incident was resolved**	**Inconclusive reactors only (yes/no)**	**Surveillance tests (yes/no)**
	3	Prevalence in 100 nearest neighbours	Surveillance tests	**Time since last confirmed incident was resolved**	**Prevalence in 100 nearest neighbours**
	4	Surveillance tests	**High-risk neighbours in 1 km radius**	Prevalence in 100 nearest neighbours	**Inconclusive reactors only (yes/no)**
	5	**Inconclusive reactors only (yes/no)**	**Open incident in 2015 (yes/no)**	**Low-risk neighbours in 1 km radius**	**Inconclusive reactors only in 2015 (yes/no)**
	6	High-risk neighbours in 1 km radius	**Low-risk neighbours in 1 km radius**	**High-risk neighbours in 1 km radius**	Reactor density
	7	**Inconclusive reactors only in 2015 (yes/no)**	Number of deaths	Number of deaths	**Time since last confirmed incident was resolved**
	8	Number of deaths	Prevalence in 100 nearest neighbours	36 month-old or over cattle in November	Number of slaughterhouse destinations
	9	Low-risk neighbours in 1 km radius	24-35 month-old cattle in November	Proportion of 6-23 month-old cattle in November	24-35 month-old cattle in November
	10	Surveillance tests in 2015	36 month-old or over cattle in November	**Reactor density**	Number of cattle on
LASSO	1	**Time since last confirmed incident was resolved=1**	**Time since last confirmed incident was resolved=1**	**Time since last confirmed incident was resolved=1**	**Inconclusive reactors only (yes/no)=1**
	2	Surveillance tests (yes/no)=1	**Open incident in 2015 (yes/no)=1**	**Inconclusive reactors only (yes/no)=1**	**Surveillance tests (yes/no)=1**
	3	Land class=22	Inconclusive reactors only in 2015 (yes/no)=1	Inconclusive reactors only in 2015 (yes/no)=1	**Time since last confirmed incident was resolved=1**
	4	Open incident in 2015 (yes/no)=1	Short residence time (yes/no)=1	Land class= 3	**Inconclusive reactors only in 2015 (yes/no)=1**
	5	**Inconclusive reactors only in 2015 (yes/no)=1**	Surveillance tests (yes/no)=1	*Land class=20*	High-risk neighbours in 1 km radius
	6	Reactor density=5	Incident in 2015 (yes/no)= 1	**High-risk neighbours in 1 km radius**	Larger herd size in 2016 than 2015 (yes/no)=1
		Reactor density=4	Movements off (yes/no)=1	Surveillance tests (yes/no)=1	Mean relative humidity 2011-2014
	7	**Inconclusive reactors only (yes/no)=1**	***Time since last confirmed incident was resolved=3***	**Low-risk neighbours in 1 km radius**	**Prevalence in 100 nearest neighbours**
		Reactor density=3	***Time since last confirmed incident was resolved=2***	Recurrent incident in 2015 (yes/no)=1	Land class=10
	8	Movements off (yes/no)=1	**High-risk neighbours in 1 km radius**	**Reactor density=3**	Movements on 2014-2016 (yes/no)=1
		Time since last confirmed incident was resolved=5	Inconclusive reactors only (yes/no)=1	**Number of slaughterhouse destinations**	
	9	Short residence time (yes/no)=1	**Low-risk neighbours in 1 km radius**		

LASSO (additions)	10	Nearest incident in 2015=2

446

447

Table 3 Receiver Operating Characteristic (ROC) analysis outputs classifying herds into high- (H), medium- (M) and low-risk (L) categories in the training and testing datasets within each area considered (HRA=High-risk area and EA= Edge area and LRA= Low-risk area). Cut-off values of 99% sensitivity and specificity were arbitrarily chosen in England, HRA and EA areas, leading to a one percent misclassified herds either as false negatives (low-risk category) or false positives (high-risk category) in the former two areas. In the EA, the proportion of false negatives increased to seven percent in the testing dataset. A 97% cut-off value for sensitivity and specificity was chosen in the LRA, leading to 4% false negatives and 3% false positives in the training dataset but only 3% false positives in the testing dataset (no incident herds were classified in the low-risk group).

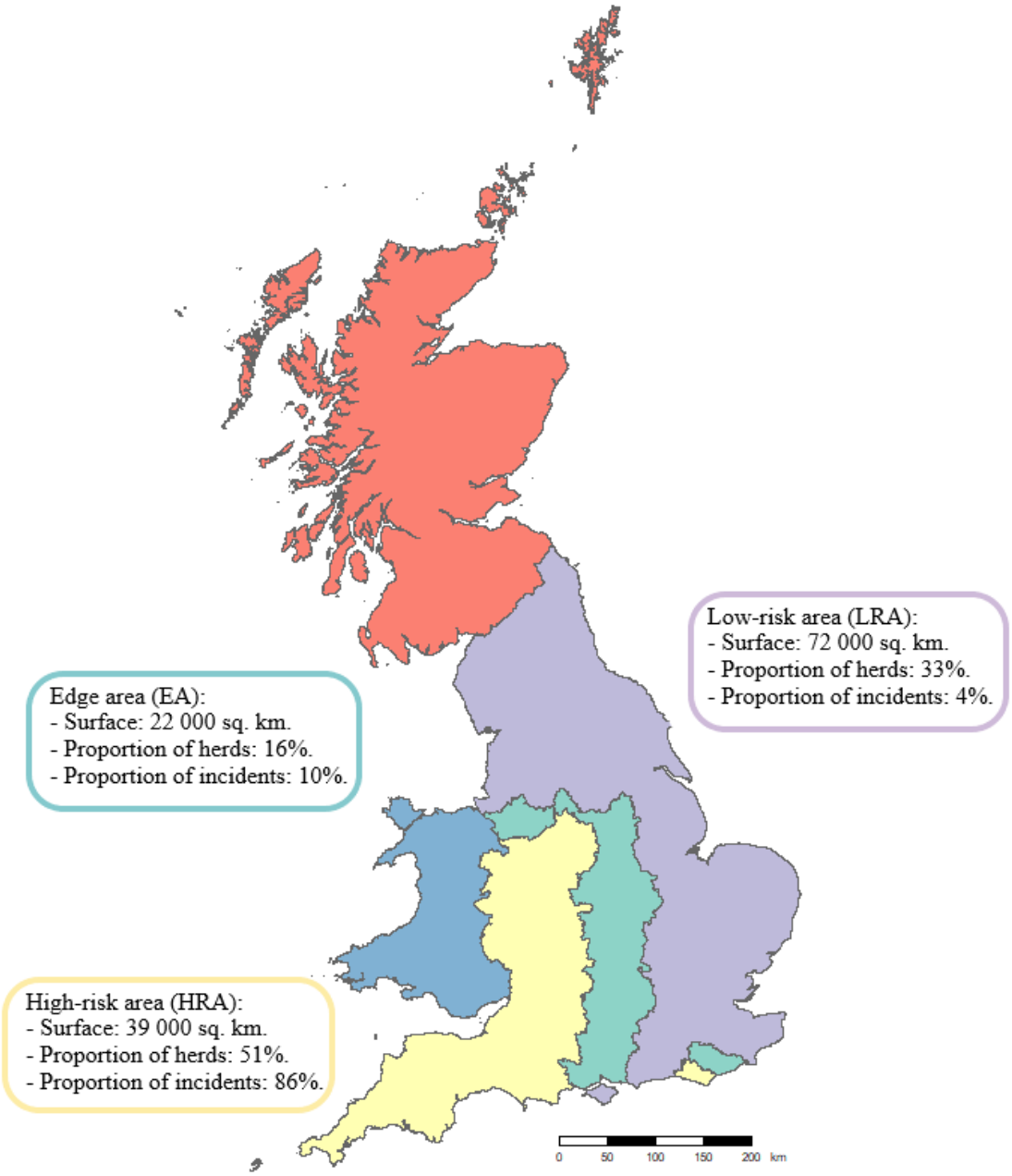
Training dataset	England			HRA			EA			LRA		
Incident (2016)	L	M	H	L	M	H	L	M	H	L	M	H
No	1320	1564	28	784	1699	25	146	146	2	57	45	3
Yes	29	2421	462	26	2125	357	3	262	29	4	74	27
No	0.45	0.54	0.01	0.31	0.68	0.01	0.50	0.50	0.01	0.54	0.43	0.03
Yes	0.01	0.83	0.16	0.01	0.85	0.14	0.01	0.89	0.10	0.04	0.70	0.26
Testing dataset	England			HRA			EA			LRA		
Incident (2016)	L	M	H	L	M	H	L	M	H	L	M	H
No	3397	3850	62	1158	2301	42	552	575	10	1481	1053	75
Yes	8	625	94	5	537	84	5	58	10	0	18	8
No	0.46	0.53	0.01	0.33	0.66	0.01	0.49	0.51	0.01	0.57	0.40	0.03
Yes	0.01	0.86	0.13	0.01	0.86	0.13	0.07	0.79	0.14	0.00	0.69	0.31

Figures

Figure 1. Map of Great Britain (GB) showing bTB surveillance risk areas that applied in England from 2013 to 2017 (inclusive).

Figure 2. Random forest algorithm schematic representation. The algorithm combines random subsets of *n* predictors from many classification trees using bootstrapped samples from the original training dataset. The final output class is derived using majority voting from the committee of classification trees used.

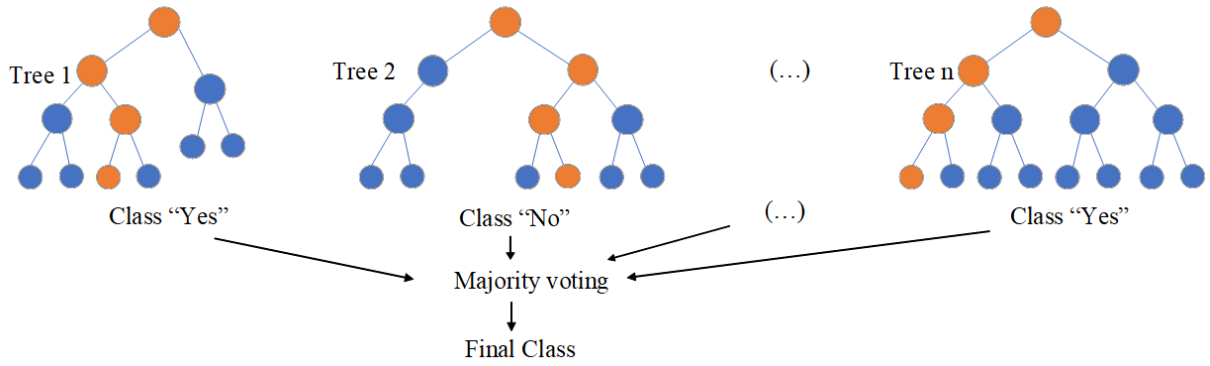
Figure 3. Receiver Operating Characteristic (ROC) analysis outputs in the training dataset by area (HRA=High-risk area, EA= Edge area and LRA= Low-risk area). The primary (left) y-axis represents the values of false positive rate (i.e. 1-specificity) and the secondary (right) y-axis represents the values of true positive rate (i.e. sensitivity); both are plotted against the predicted probability of an incident in the x-axis. This variation of a ROC curve has been presented to illustrate better the process followed for the selection of cut-off points in the predicted probability distribution, together with the resulting low-, medium- and high-risk groups in the bar at the bottom. The different distribution of herds into risk categories is shown in each area, relative to the two cut-offs or thresholds chosen (discontinued vertical lines). This was based on a 99% true positive (dark cyan line) and 1% false positive (dark blue line) thresholds in all areas except the LRA (97% and 3% values chosen, respectively).



477

478 Figure 1

479



480

481 Figure 2

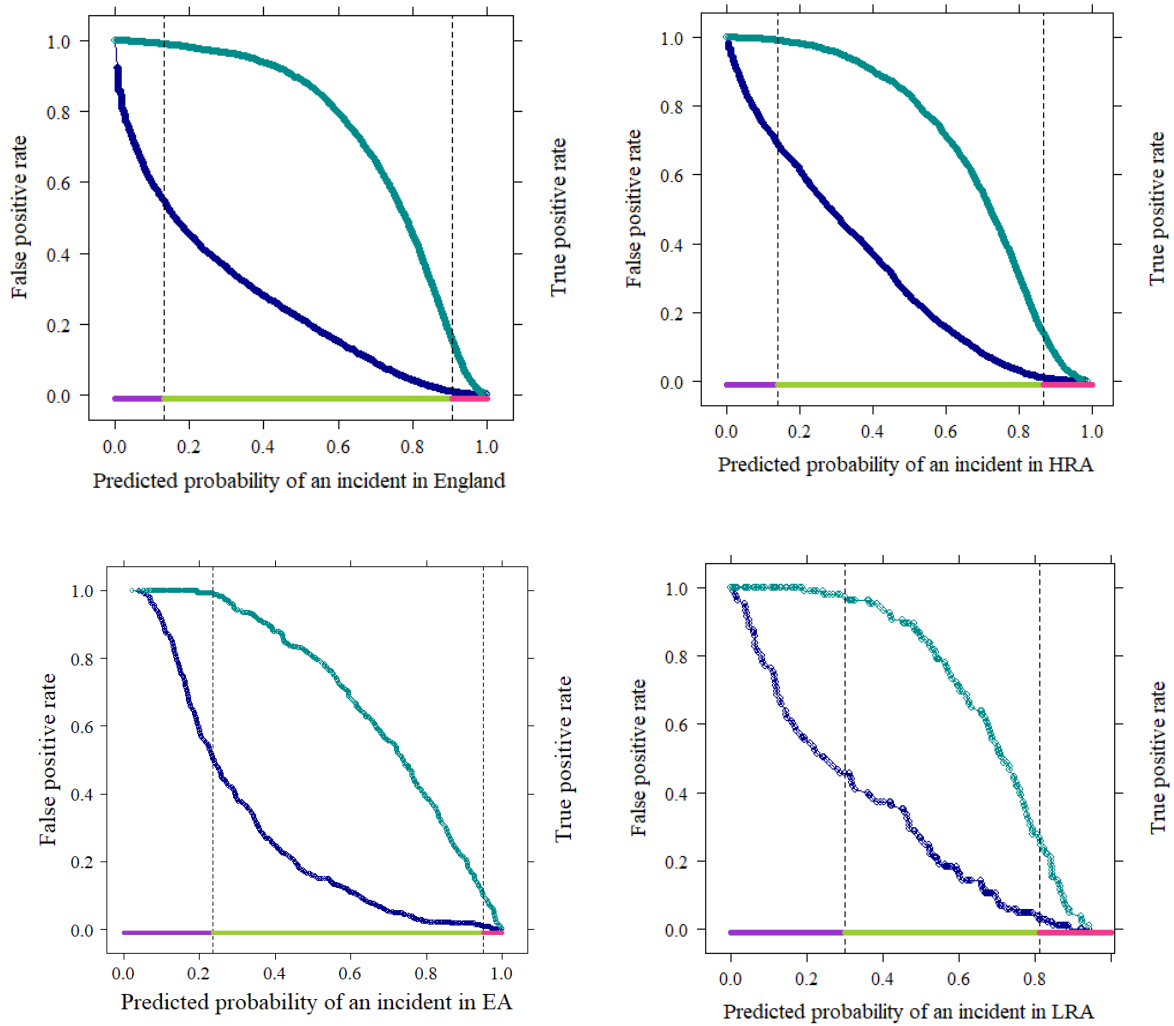

False positive rate
True positive rate
Predicted probability of an incident in England
Predicted probability of an incident in HRA
Predicted probability of an incident in EA
Predicted probability of an incident in LRA

Figure 3

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 PPS_02225. The authors thank specially Stuart Ashfield, Adam Brouwer, Dr Andrew Robertson, the Met Office and the NERC Centre for Ecology & Hydrology for the provision of variable data, to Dr Colin Birch for statistical advice and to Mr Geoff Jasinski, who read and made comments on drafts.

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. Prev. Vet. Med. 123, 32–38. https://doi.org/10.1016/j.prevetmed.2015.11.020

Afifi, A., May, S., Clark, V.A., 2011. Practical Multivariable Analysis, Fifth. ed. Chapman & Hall/CRC.

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. Fam. Pract. 29, 671–677. https://doi.org/10.1093/fampra/cms020

AHVLA, 2014. Enhanced management of persistent TB herds.

APHA, 2020. Bovine tuberculosis in Great Britain Surveillance data for 2019 and historical trends.

APHA, 2018a. Terms and conditions of the approval and operation of a Licensed Finishing Unit. Defra.

APHA, 2018b. Pre-movement and post-movement testing of cattle in Great Britain.

APHA, 2017a. Terms and conditions of the approval and aperation of an Approved Finishing Unit Without Grazing. Defra.

APHA, 2017b. Bovine tuberculosis in England 2016: Epidemiological analysis of the 2016 data and historical trends.

APHA, 2017c. APHA Briefing Note 22 / 17 Bovine TB update – Reducing the risk of resolved inconclusive reactors in England.

Bielza, C., Robles, V., Larrañaga, P., 2011. Regularized logistic regression without a penalty

523 term: an application to cancer classification with microarray data. Expert Syst. Appl. 38,
524 5110–5118. https://doi.org/10.1016/j.eswa.2010.09.140
525 Boulesteix, A., Janitza, S., Kruppa, J., Konig, I., 2012. Overview of random forest
526 methodology and practical guidance with emphasis on computational biology and
527 bioinformatics, Technical Report, Department of Statistics, University of Munich.
528 Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
529 https://doi.org/10.1007/9781441993267_5
530 Breiman, L., Friedman, J.H., Olsen, R.A., Stone, C.J., 1984. Classification and Regression
531 Trees. Wadsworth Inc.
532 Brooks-Pollock, E., Keeling, M., 2009. Herd size and bovine tuberculosis persistence in cattle
533 farms in Great Britain. Prev. Vet. Med. 92, 360–365.
534 https://doi.org/10.1016/j.prevetmed.2009.08.022
535 Broughan, J.M., Judge, J., Ely, E., Delahay, R.J., Wilson, G., Clifton-Hadley, R.S.,
536 Goodchild, A. V., Bishop, H., Parry, J.E., Downs, S.H., 2016. A review of risk factors
537 for bovine tuberculosis infection in cattle in the UK and Ireland. Epidemiol. Infect. 144,
538 2899–2926. https://doi.org/10.1017/S095026881600131X
539 Brunton, L.A., Prosser, A., Pfeiffer, D.U., Downs, S.H., 2018. Exploring the fate of cattle
540 herds with inconclusive reactors to the tuberculin skin test. Front. Vet. Sci. 5, 1–10.
541 https://doi.org/10.3389/fvets.2018.00228
542 Bunce, R.G.H., Barr, C.J., Clarke, R.T., Howard, D.C., Scott, W.A., 2007. ITE Land
543 Classification of Great Britain 2007. https://doi.org/10.5285/5f0605e4-aa2a-48ab-b47c-
544 bf5510823e8f
545 Campbell, M.J., Swinscow, T.D. V, 2009. Statistics at Square One, 11th ed. BMJ Publishing

546 Group Ltd, UK.
547 Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W., 2002. SMOTE: synthetic minority over-
548 sampling technique. J. Artif. Intell. Res. 16, 321–357.
549 Clegg, T.A., Good, M., Duignan, A., Doyle, R., More, S.J., 2011a. Shorter-term risk of
550 Mycobacterium bovis in Irish cattle following an inconclusive diagnosis to the single
551 intradermal comparative tuberculin test. Prev. Vet. Med. 102, 255–264.
552 https://doi.org/10.1016/j.prevetmed.2011.07.014
553 Clegg, T.A., Good, M., Duignan, A., Doyle, R., More, S.J., 2011b. Longer-term risk of
554 Mycobacterium bovis in Irish cattle following an inconclusive diagnosis to the single
555 intradermal comparative tuberculin test. Prev. Vet. Med. 100, 147–154.
556 https://doi.org/10.1016/j.prevetmed.2011.07.014
557 Cramer, H., 1946. Mathematical Methods in Statistics. Princeton University Press.
558 Defra, 2019. TB in cattle in Great Britain: GB by country dataset.
559 Defra, 2014. The strategy for achieving Officially Bovine Tuberculosis Free status for
560 England. Defra.
561 Defra, 2011. Bovine TB Eradication Programme for England. Defra.
562 Defra, 2006. Bovine TB special edition. Gov. Vet. J. 16.
563 European-Commission, 1964. Council Directive 64/432/EEC on animal health problems
564 affecting intra-Community trade in bovine animals and swine. Official Journal.
565 Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27, 861–874.
566 https://doi.org/10.1016/j.patrec.2005.10.010
567 Fei, Y., Gao, K., Hu, J., Tu, J., Li, W., Wang, W., Zong, G., 2017. Predicting the incidence of
568 portosplenomesenteric vein thrombosis in patients with acute pancreatitis using

569 classification and regression tree algorithm. J. Crit. Care 39, 124–130.
570 Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear
571 models via coordinate descent. J. Stat. Softw. 33.
572 Frisman, L., Prendergast, M., Lin, H.-J., Rodis, E., Greenwell, L., 2008. Applying
573 classification and regression tree analysis to identify prisoners with high HIV risk
574 behaviors. J Psychoact. Drugs 40, 447–458.
575 Garcia, V.; Mollineda, R.A.; Sanchez, J.S., 2009. Index of balanced accuracy: a performance
576 measure of skewed class distributions. Lect. Notes Comput. Sci. 5524.
577 https://doi.org/10.1007/978-3-642-02172-5
578 Genuer, R., Poggi, J.-M., Tuleau-Malot, C., 2010. Variable selection using random forests.
579 Pattern Recognit. Lett. 31, 2225–2236.
580 Gerds, T.A., Cai, T., Schumacher, M., 2008. The performance of risk prediction models.
581 Biometrical J. 50, 457–479. https://doi.org/10.1002/bimj.200810443
582 Godfray, C., Donnelly, C., Hewinson, G., Winter, M., Wood, J., 2018. TB Strategy Review.
583 Godfray, H.C.J., Donnelly, C.A., Kao, R.R., Macdonald, D.W., McDonald, R.A.,
584 Petrokofsky, G., Wood, J.L.N., Woodroffe, R., Young, D.B., McLean, A.R., 2013. A
585 restatement of the natural science evidence base relevant to the control of bovine
586 tuberculosis in Great Britain. Proc. R. Soc. B Biol. Sci. 280.
587 https://doi.org/10.1098/rspb.2013.1634
588 Gonçalves, L., Subtil, A., Rosário Oliveira, M., De Zea Bermudez, P., 2014. ROC curve
589 estimation: an overview. Revstat Stat. J. 12, 1–20.
590 Greiner, M., Pfeiffer, D., Smith, R.D., 2000. Principles and practical application of the
591 receiver-operating characteristic analysis for diagnostic tests. Prev. Vet. Med. 45, 23–41.

https://doi.org/10.1016/S0167-5877(00)00115-X

Guyon, I., Elisseeff, A. e, 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182. https://doi.org/10.1016/j.aca.2011.07.027

Hamzi, A., 2014. Bovine TB, in: The Art & Science of Tuberculosis Vaccine Development. FAO.

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. Psychol. Aging 30, 911–929. https://doi.org/10.1037/pag0000046

Hilbe, J.M., 2009. Logistic Regression Models, First. ed. Chapman & Hall/CRC.

Hoerl, A.E., Kennard, R.W., 1982. Ridge regression: biased estimation for nonorthogonal problems. CC/Eng. Tech. Appl. Sci. 35, 18. https://doi.org/10.1080/00401706.1970.10488634

Hosmer, D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied Logistic Regression, 3rd ed. John Wiley & Sons, Inc.

Jain, D., Singh, V., 2018. Feature selection and classification systems for chronic disease prediction: A review. Egypt. Informatics J. 19, 179–189.

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

Judge, J., Wilson, G.J., Macarthur, R., McDonald, R.A., Delahay, R.J., 2017. Abundance of badgers (Meles meles) in England and Wales. Sci. Rep. 7, 1–8. https://doi.org/10.1038/s41598-017-00378-3

615 Kashani, A.T., Mohaymany, A.S., 2011. Analysis of the traffic injury severity on two-lane,
616 two-way rural roads based on classification tree models. Saf. Sci. 49, 1314–1320.
617 Kassambara, A., 2018. Machine Learning Essentials: Practical Guide in R. CreateSpace
618 Independent Publishing Platform.
619 Kawamura, Y., Takasaki, S., Mizokami, M., 2012. Using decision tree learning to predict the
620 responsiveness of hepatitis C patients to drug treatment. FEBS Open Bio 2, 98–102.
621 Khun, L., Page, K., Ward, J., Worrall-Carter, L., 2014. The process and utility of
622 classification and regression tree methodology in nursing research. J. Adv. Nurs. 70,
623 1276–1286. https://doi.org/10.1111/jan.12288
624 Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28.
625 https://doi.org/10.18637/jss.v028.i05
626 Kwok, K.O., Cowling, B.J., Wei, V.W.I., Wu, K.M., Read, J.M., Lessler, J., Cummings,
627 D.A., Malik Peiris, J.S., Riley, S., 2014. Social contacts and the locations in which they
628 occur as risk factors for influenza infection. Proc. R. Soc. B Biol. Sci. 281.
629 https://doi.org/10.1098/rspb.2014.0709
630 Lewis, R.J., 2000. An introduction to classification and regression tree (CART) analysis, in:
631 Annual Meeting of the Society for Academic Emergency Medicine. San Francisco.
632 Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2, 18–
633 22. https://doi.org/10.1177/154405910408300516
634 Maimon, O., Rokach, L., 2010. Data Mining and Knowledge Discovery Handbook, 2nd ed.
635 Springer. https://doi.org/10.1007/978-0-387-09823-4
636 May, E., Prosser, A., Downs, S.H., Brunton, L.A., 2019. Exploring the risk posed by animals
637 with an inconclusive reaction to the bovine tuberculosis skin test in England and Wales.

638 Vet. Sci. 6. https://doi.org/10.3390/vetsci6040097
639 McKinley, T.J., Lipschutz-Powell, D., Mitchell, A.P., Wood, J.L.N., Conlan, A.J.K., 2018.
640 Risk factors and variations in detection of new bovine tuberculosis breakdowns via
641 slaughterhouse surveillance in Great Britain. PLoS One 13, e0198760.
642 https://doi.org/10.1371/journal.pone.0198760
643 McLaren, L., McIntyre, L., Kirkpatrick, S., 2010. Rose's population strategy of prevention
644 need not increase social inequalities in health. Int. J. Epidemiol. 39, 372–377.
645 https://doi.org/10.1093/ije/dyp315
646 Met_Office, 2017. UKCP09: Met Office gridded land surface climate observations - daily
647 temperature and precipitation at 5 km resolution. Centre for Environmental Data, 15th
648 May 2019. [WWW Document]. URL
649 http://catalogue.ceda.ac.uk/uuid/319b3f878c7d4cbfbdb356e19d8061d6
650 More, S.J., Radunz, B., Glanville, R.J., 2015. Review: Lessons learned during the successful
651 eradication of bovine tuberculosis from Australia. Vet. Rec. 177, 224–232.
652 https://doi.org/10.1136/vr.103163
653 Mostafizur Rahman, M., Davies, D.N., 2013. Addressing the class imbalance problem in
654 medical datasets. Int. J. Mach. Learn. Comput. 3, 224–228.
655 https://doi.org/10.7763/IJMLC.2013.V3.307
656 Murai, K., Tizzani, P., Awada, L., Mur, L., Mapitse, N.J., Caceres, P., 2019. Controlling
657 bovine tuberculosis: a One Health challenge, OIE Panorama Bulletin.
658 Olea-Popelka, F., Muwonge, A., Perera, A., Dean, A.S., Mumford, E., Erlacher-Vindel, E.,
659 Forcella, S., Silk, B.J., Ditiu, L., El Idrissi, A., Raviglione, M., Cosivi, O., LoBue, P.,
660 Fujiwara, P.I., 2017. Zoonotic tuberculosis in human beings caused by Mycobacterium

661 bovis—a call for action. Lancet Infect. Dis. 17, e21–e25. https://doi.org/10.1016/S1473-
662 3099(16)30139-6
663 Pedersen, A.B., Mikkelsen, E.M., Cronin-Fenton, D., Kristensen, N.R., Pham, T.M.,
664 Pedersen, L., Petersen, I., 2017. Missing data and multiple imputation in clinical
665 epidemiological research. Clin. Epidemiol. 9, 157–166.
666 Pereira, J.M., Basto, M., Silva, A.F. da, 2016. The logistic lasso and ridge regression in
667 predicting corporate failure. Procedia Econ. Financ. 39, 634–641.
668 https://doi.org/10.1016/s2212-5671(16)30310-0
669 Pfeiffer, D.U., 2013. Epidemiology caught in the causal web of bovine tuberculosis.
670 Transbound. Emerg. Dis. 60, 104–110. https://doi.org/10.1111/tbed.12105
671 Phillips, C.J.C., Foster, C.R.W., Morris, P.A., Teverson, R., 2003. The transmission of
672 Mycobacterium bovis infection to cattle. Res. Vet. Sci. 74, 1–15.
673 https://doi.org/10.1016/S0034-5288(02)00145-5
674 Platt, J.M., Keyes, K.M., Galea, S., 2017. Efficiency or equity? Simulating the impact of
675 high-risk and population intervention strategies for the prevention of disease. SSM -
676 Popul. Heal. 3, 1–8. https://doi.org/10.1016/j.ssmph.2016.11.002
677 Pollock, J.M., Neill, S.D., 2002. Mycobacterium bovis infection and tuberculosis in cattle.
678 Vet. J. 163, 115–127. https://doi.org/10.1053/tvjl.2001.0655
679 R_Core_Team, 2020. R: A language and environment for statistical computing.
680 Romero, M.P., Chang, Y.M., Brunton, L.A., Parry, J., Prosser, A., Upton, P., Rees, E.,
681 Tearne, O., Arnold, M., Stevens, K., Drewe, J.A., 2020. Decision tree machine learning
682 applied to bovine tuberculosis risk factors to aid disease control decision making. Prev.
683 Vet. Med. 175. https://doi.org/10.1016/j.prevetmed.2019.104860

684 Rose, G., 2001. Sick individuals and sick populations: 20 Years later. Int. J. Epidemiol. 30,
685 427–432. https://doi.org/10.1136/jech.2005.042770
686 Schiltz, F., Masci, C., Agasisti, T., Horn, D., 2018. Using regression tree ensembles to model
687 interaction effects: a graphical approach. Appl. Econ. 50, 6341–6354.
688 https://doi.org/10.1080/00036846.2018.1489520
689 Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., Khovanova, N., 2019. Decision
690 tree and random forest models for outcome prediction in antibody incompatible kidney
691 transplantation. Biomed. Signal Process. Control 52, 456–462.
692 https://doi.org/https://doi.org/10.1016/j.bspc.2017.01.012
693 Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., 2005. ROCR: visualizing classifier
694 performance in R. Bioinformatics 21, 3940–3941.
695 https://doi.org/10.1093/bioinformatics/bti623
696 Skuce, R.A., Allen, A.R., McDowell, S.W.J., 2012. Herd-level risk factors for bovine
697 tuberculosis: a literature review. Vet. Med. Int. https://doi.org/10.1155/2012/621210
698 Song, Y.-Y., Lu, Y., 2015. Decision tree methods: applications for classification and
699 prediction. Shanghai Arch. psychiatry 27, 130–135.
700 Strobl, C., 2010. An introduction to recursive partitioning: rationale, application and
701 characteristics of classification. Psychol Methods 14, 323–348.
702 https://doi.org/10.1037/a0016973
703 TBhub, 2020. Other actions taken during a TB breakdown [WWW Document].
704 www.tbhub.co.uk. URL https://tbhub.co.uk/advice-during-a-tb-breakdown/other-actions-
705 taken-during-a-tb-breakdown/ (accessed 3.13.20).
706 Therneau, T.M., Atkinson, E.J., 2018. An introduction to recursive partitioning using the rpart

routines. R package version 4.1-15.

Thursfield, M., 2005. Veterinary Epidemiology, 3rd ed. Blackwell Science Ltd.

Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: a retrospective. J. R. Stat. Soc. Ser. B 73, 273–282.

van Buuren, S., 2011. Mice: multivariate imputation by chained equations in R. J. Stat. Softw. 45. https://doi.org/10.1177/0962280206074463

van Erkel, A.R., Pattynama, P.M.T., 1998. Receiver operating characteristic (ROC) analysis: basic principles and applications in radiology. Eur. J. Radiol. 27, 88–94. https://doi.org/10.2337/diacare.22.12.1988

Verikas, A., Gelzinis, A., Bacauskiene, M., 2011. Mining data with random forests: a survey and results of new tests. Pattern Recognit. 44, 330–349.

White, I.R., Royston, P., Wood, A.M., 2011. Multiple imputation using chained equations: issues and guidance for practice. Stat. Med. 30, 377–399. https://doi.org/doi:10.1002/sim.4067

Winkler, B., Mathews, F., 2015. Environmental risk factors associated with bovine tuberculosis among cattle in high-risk areas. Biol. Lett. 11. https://doi.org/10.1098/rsbl.2015.0536

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 Resour. Res. 52, 1626–1651.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 67, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00527.x