RESEARCH BRIEF: Machine learning for herd-level bovine Tuberculosis breakdown prediction
Author: Kajetan Stanski, Samantha Lycett, Thibaud Porphyre, May Fujiwara, Harriet Auty, Lisa Boden, Mark Bronsvoort
Date: 8 April 2020

1. KEY MESSAGE

The application of Machine Learning (ML) to cattle tracing system (CTS) and bovine tuberculosis (bTB) surveillance data allows for a sophisticated interpretation of the herd-level single intradermal comparative cervical tuberculin (SICCT) test results leading to increased testing sensitivity and specificity. This improvement can help to issue better-informed movement restrictions, follow-up tests and other control measures that will result in more rapid identification of infected herds and their removal from the network thus reducing outbreak size. Importantly, the ML model can detect potentially infected but undetected herds (with no reactors).

2. MAJOR FINDINGS

1. After training on 2012-2014 data on bTB at GB level with several ML models, the Neural Network model achieved the best classification on previously unseen 2015 data leading to sensitivity of 67% (increase of 6% comparing to the observed herd-level SICCT test) and specificity of 92% (increase of 2%). These improvements are statistically significant.
2. The model correctly predicted additional 362 bTB breakdowns (out of all 5,504 breakdowns in 2015) which were missed by the herd-level SICCT test. The locations of these farms are presented as a heatmap in Figure 1.
3. The improvement in sensitivity is apparent only in West England and Wales (known high risk areas).

3. OBJECTIVES

To assess the benefit of applying ML methods in bovine tuberculosis surveillance.

4. POLICY IMPLICATIONS

The application of ML to bTB and CTS data allows for a sophisticated interpretation of the SICCT test results at the herd level leading to increased herd-level sensitivity and specificity (see section 5 for more detail). This improvement can help issue better-informed movement restrictions, follow-up tests and other control measures. Implementing the ML approach requires deploying a trained ML model on a server which is connected to up-to-date CTS and bTB surveillance databases. The predictions made by this system can then serve as an additional piece of evidence to decide on appropriate control measures for the farm.

5. IMPORTANT ASSUMPTIONS AND LIMITATIONS

In order to develop predictive models, both ground truth positives (future bTB breakdowns) and negatives (bTB-free farms) had to be defined to train and evaluate the ML algorithms. As such, a herd was defined as a future breakdown if a bTB infection of at least one animal in the herd was confirmed by lesion inspection or *M. bovis* culture within 90 days following the herd-level SICCT test. There is no gold standard diagnostic test for bTB, i.e. test with perfect sensitivity and specificity, so one needs to consider this when interpreting the performance metrics of the models. The models were trained to mimic the ground truth positives/negatives distinction and since they are based on lesion detection 'truth' they may exhibit imperfect specificity.

Classification ML models are mathematical functions which aim to produce predictions (approximation of probability of future breakdown) based on input variables representing a farm, similarly to logistic regression.

The classification is governed by internal parameters of the model which are optimised to match historical breakdown (training data). The trained model can then be used to make predictions for future breakdowns (testing data). ML models are non-linear and they utilise interactions between variables (e.g. high value of variable A may correlate with high risk of breakdown, only when variable B is small). This makes the ML models more powerful than simple statistical models (e.g. logistic regression or decision tree), but unlike simple models, it is very challenging to explain why a ML model classified each farm as it did. It is possible, however, to measure the importance of input variables, i.e. how much every variable contributes to the model overall. This means that one can identify which features of farms make them high-risk farms in general (e.g. herd size and the number of incoming moves strongly correlate with bTB risk in this dataset), but one cannot pinpoint which features of a specific farm makes it high or low risk (e.g. a farm may be classified as high-risk, but we do not know if it is because of its herd size, number of incoming moves or both of them or some other feature). Therefore, it is challenging to tell what actions or measures would best decrease the risk of a bTB breakdown at this farm according to the model. This issue was not covered in this study.

Data related to between-farm cattle movement and bTB breakdown history of farms was extracted from the EPIC database. This data was further extended with the climate and landscape data. The parameters of four Machine Learning models (Neural Network, Random Forest, Boosted Decision Trees and Support Vector Machine) were separately optimised with data from 2012-2014 (totalling 14,119 ground truth positives and 360,947 negatives). Then, the models were evaluated with 2015 data (4,690 ground truth positive herds and 126,255 negative). Improvement of herd-level bTB diagnostic sensitivity was only apparent in high-risk areas (West England and Wales). This is likely the case because of the low number of bTB breakdowns in low-risk areas (Scotland and East England) which makes it hard for the predictive model to recognise them (small number of examples of negatives has been presented to the model).
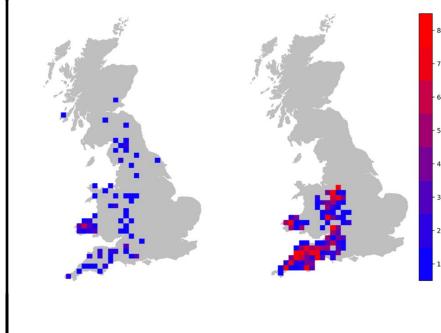
6. FIGURE

Figure 1. Spatial distribution of true positive bTB breakdown predictions. 92 bTB breakdowns were missed by the Neural Net but detected by the herd-level SICCT test (left). Additional 362 bTB breakdowns were predicted by the Neural Net (right). The Neural Net prediction offers an improvement in the interpretation of the SICCT test, but this advantage is only apparent in West England and Wales.

7. LINKS TO EXISTING PUBLICATIONS OR REPORTS

K. Stański, S. Lycett, T. Porphyre and B.M.de C. Bronsvoort (2019) Data-driven modelling for improving herd-level bovine tuberculosis breakdown predictions in GB cattle. In, Society for Veterinary Epidemiology and Preventive Medicine, Proceedings: Utrecht, The Netherlands, March 27-29 2019 (SVEPM Proceedings). ISBN 978-0-948073-50-2

Stański, K., Lycett, S., Porphyre, T. *et al.* Using machine learning improves predictions of herd-level bovine tuberculosis breakdowns in Great Britain. *Sci Rep* **11,** 2208 (2021). https://doi.org/10.1038/s41598-021-81716-4