Preventive Veterinary Medicine 92 (2009) 373–381

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

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

Meta-analytical equivalence studies on diagnostic tests for bovine brucellosis allowing assessment of a test against a group of comparative tests☆

Matthias Greiner [a,b,*], Didier Verloo [c], Fabrizio de Massis [c]

[a] Federal Institute for Risk Assessment (BfR), Epidemiology, Biostatistics and Mathematical Modelling, Thielallee 88-92, D-14195 Berlin, Germany
[b] International EpiLab, OIE Collaborative Centre for Research and Training in Population Animal Health Diagnosis and Surveillance Systems, Copenhagen, Denmark
[c] European Food Safety Authority (EFSA), Largo N. Palli 5/A I-43100 Parma, Italy

ARTICLE INFO

Article history:
Received 27 June 2008
Received in revised form 23 July 2009
Accepted 24 July 2009

Keywords:
Meta-analysis
Sensitivity
Specificity
Power
Equivalence
Cattle
Brucellosis

ABSTRACT

In the assessment of diagnostic tests the task may arise to show that a candidate test is non-inferior compared to a comparative (standard) test with regard to the diagnostic sensitivity or specificity. This setting is known as "one-sided equivalence" and has been applied to a single comparison between two diagnostic tests (Chen et al., 2003). Recently, the approach has been extended into a meta-analytical framework (EFSA, 2006), allowing for the difference between the sensitivity (or specificity) of two diagnostic tests to be estimated using information gathered through systematic literature review. Using this approach, confounding factors are adjusted by matching of parameter estimates on study population and preferred levels of the confounding factors. However, the power of this approach was found to be limited and therefore Markov chain Monte Carlo logistic regression (MCMCLR) models that allow adjustment for confounding variables have been developed (EFSA, 2006). We report here a refinement of the statistical inference based on the latter approach. The objective was to generate a posterior distribution of the meta-analytical difference statistic for the candidate test and a set of comparative tests. The algorithm for this purpose uses Monte Carlo sampling from the posterior distributions of sensitivity (or specificity) and, for each iteration, (i) identifies the least performant comparative test, (ii) establishes the difference statistics for this test and the candidate test and (iii) compares the difference statistic with a critical threshold value. The proportion of iterations in which the critical threshold was exceeded is then interpreted as the *P*-value for the one-sided equivalence test for the candidate versus the set of comparative tests. We illustrate and discuss the method using a case study on tests for bovine brucellosis.

1. Introduction

Diagnostic tests, which are applied in the context of international trade in animals and animal products, are subject to particular scientific and regulatory scrutiny. In the "Manual of Diagnostic Tests and Vaccines for Terrestrial Animals" (OIE, 2004) the so-called "prescribed tests" are recommended for each of the OIE-listed diseases for purposes related to international trade. Although many of the prescribed tests have not undergone a formal

☆ This study has been conducted based on the results of a systematic review reported in the Scientific Opinion of the European Authority for Food Safety (EFSA) on the Performance of Brucellosis Diagnostic Methods for Bovines, Sheep, and Goats (EFSA-Q-2005-060).

* Corresponding author. Tel.: +40 30 8412 1958.
E-mail addresses: matthias.greiner@bfr.bund.de (M. Greiner), Didier.VERLOO@efsa.europa.eu (D. Verloo), Fabrizio.DEMASSIS@efsa.europa.eu (F. de Massis).

0167-5877/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.prevetmed.2009.07.014

validation process they are considered sufficiently well validated based on the consensus experience of the international veterinary public health community. Before a new diagnostic test for an OIE-listed disease can be included in the portfolio of veterinary diagnostic laboratories, a thorough validation process is required. In 2005, the OIE launched a new procedure for diagnostic test validation (OIE, 2005), which includes an estimation of the diagnostic sensitivity (*Se*) and specificity (*Sp*) based on consideration of epidemiological principles. However, as the registration process is both time and resource-demanding, these provisions will likely apply only to tests that have a commercial application. Therefore, in the absence of new data generated for the purpose of validation, a meta-analysis based on a systematic review of evidence could be considered an alternative approach to allow a critical assessment of prescribed or otherwise historically accepted tests. A similar situation applies to tests used for international trade within the European Union. For example, the indirect enzyme-linked immunosorbent assay in serum (iELISA), the complement fixation test (CFT), the Rose-Bengal plate test (RBT) and the serum agglutination test (SAT) are approved individual animal tests for bovine brucellosis in the context of intra-Community trade (EC, 1964). In this situation, it is also unlikely that new validation data will be generated with the aim of re-evaluating the approved tests. However, some evidence about the comparative diagnostic performance of approved tests may exist as, for example, the SAT is no longer a prescribed test for bovine brucellosis according to the OIE (OIE, 2004). Recently, a meta-analysis has been conducted in response to a request of the European Commission to the European Food Safety Authority (EFSA) for a scientific opinion concerning brucellosis diagnostic methods for cattle, sheep and goats. The mandate included an assessment of the fluorescence polarisation assay (FPA) and any other new candidate tests that could be suitable for intra-Community trade as well as the EU-approved tests. The full report elaborated by the EFSA working group (EFSA, 2006) has been accepted by the EFSA Panel of Animal Health and Animal Welfare, which on this basis adopted an opinion on the performance of brucellosis diagnostic methods for cattle, sheep and goats (EFSA, 2006).

In the medical arena, meta-analyses (MA) of diagnostic tests are common and guidelines exist for the conduct of both systematic reviews and statistical summary of the results (Irwig et al., 1994). The meta-analytical summary measure of interest depends on the purpose of the study. Previous meta-analyses have been focused on the estimation of the *Se* and *Sp* of a test (Greiner et al., 2003). Considering the situation outlined above one could be interested in slightly different questions such as (a) "*Do any of the approved tests have performance characteristics that are inferior to all of the other approved tests?*" and (b) "*Is a new candidate test inferior to the approved tests?*" A formal procedure to address these questions could provide a scientific basis to recommend that (a) a test looses (or maintains) its status as an approved test because of its inferior (non-inferior) performance or (b) a candidate test should not (or should) be approved or considered for a full validation study depending on the regulatory framework.

A suitable statistical approach to address the above questions is the hypothesis testing for one-sided equivalence (or non-inferiority). Applied to our context, a candidate test is considered non-inferior to a comparative test with regard to a specified performance parameter (*Se* or *Sp* or both) if its performance is better than or equal or if its performance is lower than that of the comparative test but the difference is not greater than a threshold value δ (Chen et al., 2003). Note that in this setting we are not interested in investigating whether the candidate test is better than the comparative test. The equivalence statistical testing for a single pair of parameter estimates (i.e. a single comparative validation study of two tests) has been described (Chen et al., 2003) and extended to a meta-analysis using the inverse-variance-weighting method (EFSA, 2006). However, certain challenges have been reported. Whereas in the ideal situation, each primary study provides one empirical estimate of the difference between the diagnostic parameters of the two tests under comparison, the study showed that many diagnostic studies reported more than one estimate of a *Se* or *Sp*, for example because multiple reference populations were investigated. In this situation the pairing of tests is not straightforward and leads to either data augmentation (when including the same estimate in multiple comparisons) or data diminishing (when selecting only one pair of estimates for equivalence testing). The authors of the cited study were also concerned with discarding valuable evidence from studies that evaluated only one of the tests under comparison, resulting in considerable loss of overall power. The power of the non-inferiority test is the probability of concluding that a test is not inferior when in fact its performance is greater than, equal to or not much lower than that of the comparative test and the difference being not more than δ. It was concluded in the cited report that Markov chain Monte Carlo logistic regression (MCMCLR) was the preferred and more flexible approach to compare meta-analytical summary estimates for *Se* and *Sp* with adjustment for confounding (EFSA, 2006). The approach was focused on estimation rather that hypothesis testing and the interpretation was based on the 95% credibility intervals for the difference statistic for all pairwise comparisons. While this interpretation provided useful detailed information, it is not strictly consistent with a one-sided equivalence testing approach.

Here we describe the statistical approach of generating the MCMCLR models and report a refinement for the generation of the posterior distribution of the difference statistic. The objective of the refinement was to ensure that the difference statistic accommodates the multiple comparison situation (one candidate test versus a set of comparative tests) and can be used to establish a *P*-value for a one-sided equivalence test. For illustration, we use the data for diagnostic tests for brucellosis in cattle from the EFSA study (EFSA, 2006). Non-statistical considerations such as antigen used, conjugate used, availability, test costs and testing throughput capacity are beyond the scope of this assessment, which presupposes that all candidate tests are in principle eligible.