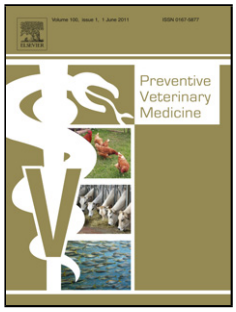
Evaluation of the methodological quality of studies of the performance of diagnostic tests for bovine tuberculosis using QUADAS



Title: Evaluation of the methodological quality of studies of the performance of diagnostic tests for bovine tuberculosis using QUADAS adapted for the veterinary field

Authors: Sara H. Downs, Simon J. More, Anthony V. Goodchild, Adam O. Whelan, Darrell A. Abernethy, Jennifer M. Broughan, Angus Cameron, Alasdair J. Cook, R. Ricardo de la Rua-Domenech, Matthias Greiner, Jane Gunn, Javier Nuñez-Garcia, Shelley Rhodes, Simon Rolfe, Michael Sharp, Paul Upton, Eamon Watson, Michael Welsh, John A. Woolliams, Richard S. Clifton-Hadley, Jessica E. Parry

PII: S0167-5877(16)30297-5
DOI: http://dx.doi.org/doi:10.1016/j.prevetmed.2017.03.006
Reference: PREVET 4217

To appear in: *PREVET*

Received date: 27-8-2016
Accepted date: 18-3-2017

Please cite this article as: Downs, Sara H., More, Simon J., Goodchild, Anthony V., Whelan, Adam O., Abernethy, Darrell A., Broughan, Jennifer M., Cameron, Angus, Cook, Alasdair J., Ricardo de la Rua-Domenech, R., Greiner, Matthias, Gunn, Jane, Nuñez-Garcia, Javier, Rhodes, Shelley, Rolfe, Simon, Sharp, Michael, Upton, Paul, Watson, Eamon, Welsh, Michael, Woolliams, John A., Clifton-Hadley, Richard S., Parry, Jessica E., Evaluation of the methodological quality of studies of the performance of diagnostic tests for bovine tuberculosis using QUADAS adapted for the veterinary field.Preventive Veterinary Medicine http://dx.doi.org/10.1016/j.prevetmed.2017.03.006

Evaluation of the methodological quality of studies of the performance of diagnostic tests for bovine tuberculosis using QUADAS adapted for the veterinary field

Sara H. Downs[a], Simon J. More[b], Anthony V. Goodchild[a], Adam O. Whelan[a,c], Darrell A. Abernethy[d,e], Jennifer M. Broughan[a], Angus Cameron[f], Alasdair J. Cook[a,g], Ricardo de la Rua-Domenech R.[h], Matthias Greiner[i], Jane Gunn[a], Javier Nuñez-Garcia[a], Shelley Rhodes[a], Simon Rolfe[j], Michael Sharp[a], Paul Upton[a], Eamon Watson[a,k], Michael Welsh[l,m], John A. Woolliams[n], Richard S. Clifton-Hadley[a], Jessica E. Parry[a]

[a]Department of Epidemiological Sciences, Animal and Plant Health Agency (APHA), Weybridge, Surrey KT15 3NB, United Kingdom

[b]Centre for Veterinary Epidemiology and Risk Analysis, UCD School of Veterinary Medicine, Belfield, Dublin 4, Ireland

[c]Microbiology, Dstl, Porton Down, SP4 0JQ, United Kingdom

[d]Veterinary Service, Department of Agriculture and Rural Development, Belfast BT4 3SB, United Kingdom

[e]Faculty of Veterinary Science, University of Pretoria, South Africa

[f]AusVet Animal Health Services Pty Ltd, PO Box 3180, South Brisbane, Qld 4101, Australia

[g]Department of Veterinary Epidemiology, School of Veterinary Medicine, University of Surrey, GU2 7AL, United Kingdom

[h]Advice Services, APHA, and Bovine Tuberculosis Programme, Department for Environment, Food and Rural Affairs, London SW1P 3JR, United Kingdom

[i]Federal Institute for Risk assessment (BfR), D-14195 Berlin and

University of Veterinary Medicine, Hannover, Germany

[j]Office of the Chief Veterinary Officer, Welsh Assembly Government, Cardiff CF10 3NQ, United Kingdom

[k]National Milk Laboratories, Wiltshire SN15 1BN, United Kingdom

[l]Veterinary Sciences Division, Agri-Food & Biosciences Institute (AFBI), Belfast BT4 3SD, United Kingdom

[m]CSO SISAF Ltd, Northern Ireland Science Park, Unit 15A The Innovation Centre, Belfast BT3 9DT, United Kingdom

[n]The Roslin Institute, Roslin Biocentre, Roslin, Midlothian EH25 9PS, United Kingdom

Corresponding author: Sara Downs

Email: sara.downs@apha.gsi.gov.uk

Postal address:

Dr Sara Downs, Animal and Plant Health Agency (APHA) Weybridge, Woodham Lane, New Haw, Addlestone, Surrey, KT15 3NB, UK

Key points

Manuscript 3. VETQUADAS

- Most studies of diagnostic test performance had methodological deficiencies
- Similar patterns of methodological deficiencies observed for different test types
- Lack of blinding between assessment of reference test and index test was common

Abstract

There has been little assessment of the methodological quality of studies measuring the performance (sensitivity and/or specificity) of diagnostic tests for animal diseases. In a systematic review, 190 studies of tests for bovine tuberculosis (bTB) in cattle (published 1934 -2009) were assessed by at least one of 18 reviewers using the QUADAS (Quality Assessment of Diagnostic Accuracy Studies) checklist adapted for animal disease tests. VETQUADAS (VQ) included items measuring clarity in reporting (n=3), internal validity (n=9) and external validity (n=2). A similar pattern for compliance was observed in studies of different diagnostic test types. Compliance significantly improved with year of publication for all items measuring clarity in reporting and external validity but only improved in four of the nine items measuring internal validity ($p<0.05$). 107 references, of which 83 had performance data eligible for inclusion in a meta-analysis were reviewed by two reviewers. In these references, agreement between reviewers' responses was 71% for compliance, 32% for unsure and 29% for non-compliance. Mean compliance with reporting items was 2, 5.2 for internal validity and 1.5 for external validity. The index test result was described in sufficient detail in 80.1% of studies and was interpreted without knowledge of the reference standard test result in only 33.1%. Loss to follow-up was adequately explained in only 31.1% of studies. The prevalence of deficiencies observed may be due to inadequate reporting but may also reflect lack of attention to methodological issues that could bias the results of diagnostic test performance estimates. QUADAS was a useful tool for assessing and comparing the quality of studies measuring the performance of diagnostic tests but might be improved further by including explicit assessment of population sampling strategy.

Keywords: bovine tuberculosis, diagnostic tests, evaluation, external validity, internal validity, quality

Introduction

There has been little assessment of the methodological quality of studies that have evaluated the performance or accuracy of tests used to detect bovine tuberculosis (bTB) in cattle, despite the importance of these tests to national surveillance and disease control schemes. Test performance is most commonly estimated by comparing the results from the test being evaluated, referred to as the index test, to the results from another test, referred to as the reference standard, in the same population. The reference standard is a test considered by the study investigators to be the best available method for establishing the presence or absence of infection or disease. Bias in the measurement of the index test or reference standard or bias in the selection of the study population leads to inaccurate estimation of test performance. This in turn compromises the effectiveness of disease control strategies. Over-estimation of Sensitivity (Se) may lead to an ineffective control strategy because the test would give false confidence in the number of infected animals missed. Over-estimation of Specificity (Sp) may lead to inefficient allocation of resources and more test-reactors will be false positives than is apparent.

STARD (Standards for Reporting Diagnostic Accuracy) was developed to provide guidance on reporting to scientists conducting studies measuring the performance of diagnostic tests (Bossuyt et al., 2003). It was hoped that STARD would lead to improvements in the accuracy, completeness and transparency in reporting thereby enabling readers to assess potential for bias and to be better able to evaluate

generalizability of results (Gardner, 2010, Bossuyt et al., 2004, Bossuyt et al., 2015). QUADAS (Quality Assessment of Diagnostic Accuracy Studies) was developed as a tool to assess the methodological quality of studies (Whiting, 2003). The QUADAS tool is structured as a list of questions that assess internal validity (the degree to which estimates of diagnostic accuracy have not been biased in the study population) and external validity (the degree to which the results of the study can be applied to the population for which they have been developed) as well as clarity in reporting. QUADAS, rather than STARD, was therefore a suitable instrument in the context of systematic review for assessing the quality of the primary studies. Both STARD and QUADAS were developed for evaluation studies in human populations, and there has been considerably more evaluation of the methodological quality of studies estimating diagnostic test accuracy in the human health than in the animal health field.

The evaluation reported here was part of a broader study, incorporating a systematic review and meta-analyses of bTB diagnostic tests in cattle with stochastic herd-level modelling of freedom from bTB infection, for testing strategies applied to differing risk scenarios in Great Britain (VLA, 2011). The systematic review was conducted to identify studies that had measured the Se and Sp of diagnostic tests for bTB in cattle. Data were extracted to evaluate the performance of tests, measured by Se and Sp, and also to evaluate the methodological quality of the studies using a version of QUADAS adapted for veterinary use designated VETQUADAS (VQ).

Materials and methods

The methodology of the systematic review to identify studies measuring the performance of diagnostic tests for bTB and the meta-analysis of Se and/or Sp are

reported in detail elsewhere (Downs et al., 2017, Nunez-Garcia et al, 2017). The systematic review was conducted by 18 reviewers who were members of the study expert Working Group (WG), including ten epidemiologists (eight of whom were also veterinarians), four immunologists specialising in the development of diagnostic tests, one veterinary pathologist, one bacteriologist, one bioinformatician and one livestock geneticist. The review included two stages, stage 1 was a review of abstracts of references identified in searches of electronic databases and stage 2 was a review of entire references and also included a quality assessment.

The decision to include an assessment of the methodological quality of references was taken at the first WG meeting for the study. It was agreed that the quality assessment would be based on QUADAS which is an instrument, developed by Whiting and colleagues for reviewing the methodological quality of studies that measure the performance of diagnostic tests in human populations (Whiting, 2003, Whiting et al., 2005, 2006). QUADAS contains 14 items in total including three (items 2, 9 and 10) that measure 'clarity in reporting', nine (items 3-7, 10, 11, 13 and 14) that measure 'internal validity' and issues relating to bias and two (items 1 and 12) that measure representativeness or 'external validity' of the population in the study. The WG took the decision that the QUADAS instrument would have to be adapted to take account of differences in terminology used in veterinary and human populations and to clarify the interpretation with respect to diagnostic tests for bTB in cattle.

Each of the QUADAS items and associated guidance was reviewed and adapted for the study by a panel of five epidemiologists from the WG (including two veterinarians). The revised tool (VETQUADAS, VQ) and guidance for each item was

then circulated to the entire WG for assessment of clarity and 'face validity' (a test can be said to have face validity if it 'looks like' it will measure what it is supposed to measure). Based on the comments received back, the guidance was modified further and re-circulated once more for review and comment. Two new items were introduced. Item 15 was introduced to measure the source of funding for the research since other work has demonstrated an association between source of funding and reported results (Bekelman et al., 2003, Huss et al., 2007). Item 16 was introduced to measure whether the VQ responses were representative of all tests reported in the reference because some references reported the results of studies of the performance of more than one test. The VQ items are shown in Table 1. The guidance provided to each reviewer for the interpretation of each item is reproduced in the online supplement.

The stages of the review including the VQ assessment and number of reviewers at each stage are shown in Figure 1. During stage 1, 9782 abstracts and titles that potentially contained test performance data were identified through searches of electronic databases using a search string and each was reviewed by two members of the WG (Downs et al, 2017). During stage 2, 261 entire references identified as potentially containing estimates of Se and/or Sp at the end of stage 1 underwent a review by one or two reviewers. The number of reviewers to which each reference was assigned at stage 2 depended on the language in which the reference was written. References written in English (82% (215/261)) were reviewed by two reviewers in the WG, assigned at random. There were two native Spanish speakers in the WG, and all references written in Spanish were assigned to them to review. There was one native German speaker in the WG and all references written in German were assigned to that

member. References that were written languages other than English, German or Spanish were reviewed by a native speaker who worked at the Animal and Plant Health Agency (APHA, previously the Veterinary Laboratories Agency, VLA), but were not part of the WG. During the stage 2 review, references only underwent a VQ review where a WG reviewer, concluded that the reference had Se and/or Sp data eligible for the meta-analysis (Nunez-Garcia et al., 2017). After the stage 2 reviews, disagreements between reviewers with regard to performance data extracted from the reference were resolved during a resolution procedure described in Downs et al., 2017. However, the differences between VQ reviews did not go through a resolution procedure.

Agreement between reviewers' responses to the VQ items was assessed for those references that had been reviewed by two reviewers and passed though stage 2, i.e. both reviewers had independently decided that the reference contained eligible data during their first review of the reference and conducted VQ prior to any joint discussion of the reference. The 18 reviewers were also allocated to one of three groups based on their scientific training and background: veterinarians (n=9), laboratory scientists (n=5) and quantitative scientists (two epidemiologists, one biostatistician and one livestock geneticist) (n=4). Agreement was assessed for references reviewed by pairs of reviewers whose scientific background had been allocated to the same professional group (n=27 references), for veterinarians as a group (n=20 references) as well as between reviewers within the entire WG. Preliminary analyses of the joint distribution of results by two reviewers showed that the marginal totals of contingency tables of responses were highly unbalanced and counts across rows (k1, k2, k3), and columns (j1, j2, and j3) were not evenly

distributed (see Table 2) because the predominant response for all items was 'yes'. For this reason, separate measures of agreement were calculated for 'no', 'yes' and 'unsure', as opposed to calculating kappa for the whole table (Feinstein and Cicchetti, 1990).

Frequencies of responses (e.g. yes, no, unclear) were summed for each item on the VQ instrument, with the mean calculated for references reviewed by two reviewers. Distributions of responses to the items were calculated for the different categories of the references including those with one VQ review, with two VQ reviews, with eligible Se and/or Sp data for the meta-analysis, without eligible Se and/or Sp data, by profession of reviewer, for test types where there were at least five references with eligible Se and/or Se data and by year of reference publication. Differences between observed and expected frequencies were tested using chi squared (χ^2) tests or Fisher's exact test if the frequency of observations in any cell was less than 5. In the situation where there were two responses to a VQ item because the reference had been reviewed by two reviewers, a random number generator was used to randomly select one of the two responses before conducting χ^2 or Fisher's exact tests. A non-parametric test was used to test for trend in VQ item compliance across year of reference publication.

Bespoke study databases for stages 1 and 2 of the systematic review and for VQ were built in Microsoft Access 2003. Statistical analyses were conducted using Stata release 12.1 (StataCorp) or Microsoft Excel 2011.

Results

Of the 261 entire references that were reviewed at stage 2 of the systematic review, 190 had a VQ review (Figure 1). Of these, 107 references had two VQ reviews and 83 references had one VQ review because two reviewers and one reviewer respectively considered that the reference had eligible Se and/or Sp data at their initial review of the reference at stage 2. Of the 119 references that, after the resolution procedure, reviewers agreed had Se and/or Sp data eligible for the meta-analysis, 83 references had two VQ reviews and 31 had one VQ review. Five studies with eligible test performance data had not undergone a VQ review because the reference was written in a language other than English, Spanish or German and had not been reviewed by a WG member. Of the references that had at least one VQ review, 85.8% (163/190) were written in English, 6.8% (13/190) in Spanish and 7.4% (14/190) in German.

In references with at least one VQ review, reported funding for the research was from public or charity sources in 53.2% (101/190), from industry in 1.2% (2/190), from mixed sources in 2.6% (5/190) and not reported or unclear in 43.2% (82/190). Type of funding was not a predictor for references that had eligible data for the meta-analysis compared to those that did not (Fisher's exact test p=0.121). Reviewers were more likely to classify their responses to the VQ instrument as being representative of all tests described in a reference in references with eligible data for the meta-analysis than in references without eligible data (90.4% (103/114) versus 73.7% (56/76), χ^2 test p=0.002).

Levels of agreement between reviewers with VQ items for the 83 references with two reviews and data judged eligible for the meta-analysis are shown in Table 3. Overall agreement between reviewers that an item had been complied with was 71.0%, but less for an unsure response (32.3%) and less for non-compliance (29.0%). Agreement for the 'yes' response was highest for item 8 *(execution of the index test described in sufficient detail).* Agreement for item 10 *(index test interpreted without knowledge of reference standard results)* and item 14 *(withdrawals from the study explained)* was less than 50% for the 'yes' response, and even lower for 'no' and 'unsure' responses. Agreement in a no response was less than 10% for item 7 *(reference standard independent of the index test),* and item 12 *(same clinical data would be available when test results were used in practice).* Agreement in an unsure response was highest for item 11 *(reference standard was interpreted without knowledge of the index test results)* and lowest for item 2 *(selection criteria clearly described).* There was no evidence of differences in level of agreement between reviewer pairs based on professional grouping.

Compliance with VQ items 1 to 14 for the 83 references reviewed by two reviewers and judged as having Se and/or Sp data eligible for the meta-analysis is shown in Figure 2.

- *With respect to clarity in reporting:* The mean number of items complied with over all references was 2.0 out of a maximum of 3.0. Almost 22% (18/83) of the references were judged as having complied with all items by two reviewers. On average, item 2 *(selection criteria of the animals clearly described)* and item 8 *(execution of the index test described in sufficient detail)* were assessed as having been met in 65.1 and 80.1% of the references,

respectively. Just over 60% of the references were assessed as having met item 9 *(execution of the reference test described in sufficient detail)*

- *With respect to internal validity:* The mean number of items complied with over all references was 5.2 (SD1.7) out of a maximum of 9. No references were judged as having complied with all items by two reviewers. Over 80% of references were judged as having complied with item 7 *(reference standard independent of index test)* and 5 *(whole or random sample of animal population verified with reference standard)*. Over 75% of references were judged as having complied with item 6 *(all animals received the same reference standard)*. Item 3 *(reference standard will correctly classify target condition)* and item 4 *(time period between the reference standard and the index test short enough)* were judged as having been complied with in over 65% of references. However, over 30% of reviewers responded unclear to item 4 *(time period between the reference standard and the index test short enough)*. Between 30 and 45% of studies in the references were judged as having complied with item 10 *(index test results interpreted without knowledge of the results of the reference standard)*, item 11 *(reference standard interpreted without knowledge of index test results)* and item 13 *(un-interpretable/ intermediate test results reported)*, item 14 *(withdrawals explained)*. Between 27% and 45% of the reviewers' responses to items 10, 11, 13 and 14 were 'unsure'. Compliance with items 10, 11, 13 and 14 was low even where reporting was good. For example in the 18 references that complied with all items measuring clarity in reporting; compliance with items 10, 11, 13 and 14 was 44.4%, 30.6%, 36.1% and 38.9% respectively.

- *With respect to external validity:* The mean number of items complied with was 1.5 out of a maximum of 2. Over 32% (27/83) of references complied with both items measuring external validity according to two reviewers. Twenty-seven percent of references complied with item 1 *(animals in the study representative of animals who will receive the test)*. This increased to 74% when responses that the animals were partially representative were coded to yes. Over 70% of references were assessed as having complied with item 12 *(same clinical data available when test used in practice)*.

Having had two VQ reviews and having had data eligible for the meta-analysis was a predictor for better compliance with VQ items (Table 4). Compliance was statistically significantly better for items 2, 8 and 9 (clarity in reporting), items 1 and 12 (external validity) and items 4, 5, 6, 7 and 10 (internal validity). Lowest compliance was observed in references without eligible data and with one VQ review. However, overall, the different categories of references showed a similar pattern of compliance with VQ. For example, compliance with internal validity items 10, 11, 13 and 14 was worse than for items 3, 4 ,5, 6, and 7 and clarity in reporting was better for item 8 than for item 9.

Item compliance for test types with five or more references assessed by two reviewers is presented in Figure 3. All test-types showed low compliance with items 10, 11, 13 and 14 which measure aspects of internal validity relating to interpretation of the index tests, interpretation of reference standards and loss to follow-up. Compliance with item 6 (*all animals subject to the same reference standard*) and item 8 (*adequate description of the index test*) was better for studies of laboratory tests than of field

tests. Compliance with item 7 (*reference standard independent of index test*), was over 80% for studies measuring the Se of ante-mortem tests and over 70% for studies measuring the Se of post-mortem tests.

The proportion of references that complied with items measuring clarity in reporting and external validity was positively associated with year of publication (Table 5). Compliance with items 5, 6 and 7 measuring aspects of internal validity also increased with year of publication. However there was no evidence for improvement in compliance with 'internal validity' items 10, 11, 13 and 14 by year of reference publication and only 25.0% (95% CI 13.2, 40.3), 43.2% (95% CI 41.0, 75.7) 27.3% (95% CI 15.0, 42.8) and 18.2% (95% CI 8.2, 32.8) of references published between 2005 and 2009 complied with items 10, 11, 13 and 14 respectively.

Discussion

This aim of this study was to describe the quality of methodology used in studies measuring the performance (Se and Sp) of diagnostic tests for bTB in cattle. It is the first reported analysis of the quality of studies measuring the performance of veterinary diagnostic tests and highlights, at least for diagnostic tests for bovine tuberculosis, probable shortcomings in internal validity and also in reporting of studies. The studies assessed were identified through a systematic review that encompassed a large number of data sources (Downs et al, 2017) and included studies published between 1934 and 2009. Shortcomings observed in reported methodology could have biased performance estimates and reduced the accuracy of test results in individual cattle and herds. These may have also affected the efficacy and cost

effectiveness of bTB control strategies. Current developers of diagnostic tests may find it helpful to consider how their research addresses the methodology deficiencies identified in this sample of published studies.

The VQ instrument originated in an instrument called QUADAS (Whiting, 2003). QUADAS was developed for measuring the quality of studies designed to measure the performance of diagnostic tests in humans. QUADAS has been evaluated and has been recognised as a useful instrument for measuring methodological quality in studies measuring diagnostic test performance, including for veterinary diseases (Whiting et al., 2005, Whiting et al., 2006, Gardner et al., 2010).

Using the expertise of our WG, which included veterinarians and scientists, QUADAS was adapted for use for the review of diagnostic tests for bTB in cattle. The adapted instrument included all the items included in original QUADAS but terminology was changed to refer to animals rather than patients. Additionally the user's guide to QUADAS was adapted so that the explanations for interpretation of each item comprised examples relevant to bTB in cattle and did not include examples specific to human populations e.g. diseases such as appendicitis, reference to clinical data collected from human populations (see online supplement for further detail). A methodological investigation of level agreement between reviewers of scientific abstracts reporting studies of diagnostic tests had shown variation between reviewers (Downs et al., 2017). Based on this evidence the WG decided that the guidance provided by QUADAS should be specific for tests for bTB in cattle in an attempt to reduce variability due to reviewer interpretation of an item rather than variability in compliance with VQ items. Each assessment of a reference by a reviewer, of

compliance with VG items, was recorded in a bespoke database. Each reviewer was blind to assessments by another reviewer who had been allocated the same reference.

VQ items 3-7, 10, 11, 13 and 14 all attempt to measure the internal validity of the study. These are key items in terms of determining if the results from a study could be biased. Most recognised sources of bias lead to an over-estimation of effects (Leeflang et al., 2008), which in the present study is test accuracy. The finding that the conduct of the index test was better described than the reference test was unsurprising (compliance with reporting items 8 compared to 9) but implies a lack of appreciation of how Se and Sp are calculated. Compliance with other items measuring other aspects relating to use of the reference standard (items 3, 4, 5 and 7) was reasonably high; over 65% on average for the references reviewed by two reviewers.

Compliance with items 10, 11, 13 and 14 was less than 40%, and similarly low across studies of all test-types. Non-compliance was unlikely to be an artefact related to reporting because compliance was also low in studies that showed good compliance in 'clarity in reporting' items. Items 10 and 11 relate to concealment of the index test results from those assessing the reference standard, and the reverse situation. Knowledge of the results of the index test at the time of the assessment of the reference test (incorporation bias) is known to be strongly and positively associated with estimated test performance (Westwood, 2005). There are various forms of verification or detection bias where the reference test is evaluated with knowledge of the results of the index test (Begg and Greenes, 1983). Inadequate blinding in clinical trials has been associated with changes in measured treatment effect sizes of up to 40% (Schulz et al., 1995), and in the current scenario may result in over-estimation of

Se or Sp. Although concealment of the reference standard sets more practicable challenges when measuring Sp compared to measuring Se because Sp is often measured on population already known to be infection free, it should be possible to design studies that prevent these biases. Loss of subjects or animals due to inadequate follow-up of withdrawals (item 14) or from not reporting uninterpretable/intermediate test results (item 13) can lead to attrition bias (Begg and Greenes., 1983). 'Lost' individuals and the data they might have provided often differ from the rest of the study population.

VQ items 1 and 12 address the representativeness or external validity of the cattle population and the diagnostic test as it may be used in practice. In the current assessment, reviewers were asked to consider these items in the context of testing conditions in the UK and Republic of Ireland (see online supplement). To address these items the reviewer needs information about the cattle population sampled in the study, details of the sampling procedure, and some knowledge of cattle populations and the circumstances under which the test will be used. Census or probability-based sampling frames were used to select cattle in only 6.4% and 24.3 % respectively of references with eligible Se and/or Sp estimates (Downs et al., 2017) which was small proportion of the total. Unfortunately, if the sampling of animals is in any way associated with the performance of the diagnostic test, e.g. weak and diseased animals are preferentially sampled compared to other animals, test results are likely to be biased. Agreement between reviewers was over 70% when the sampling was reported to be random but was less when items had not been met, possibly because of poor reporting in the reference or paucity of reviewer knowledge. Explicit measurement of

procedure used to select the population sample through the inclusion of an additional VQ item might improve reliability in the assessment of this aspect of external validity.

We did not calculate the kappa coefficient to measure agreement between reviewers. The advantage of the kappa coefficient is its adjustment for the amount of agreement that can be expected by chance alone. However, like the indices of positive and negative predictive values and overall accuracy that are calculated from the performance of diagnostic tests (Alberg et al., 2004), the kappa statistic is affected by prevalence (Feinstein and Cicchetti, 1990; Viera and Garrett, 2005). If the marginal totals for a contingency table are unbalanced, the kappa statistic may provide an unreliable estimate of association. In this study, 'Yes' was the predominant response to virtually all items in VQ. Consequently, the marginal totals were highly unbalanced, with the 'No' and 'unclear' responses comprising as little as 1% and 5%, respectively, for some items in the 83 references reviewed by two reviewers. As an alternative, three separate indexes of agreement, for each of three possible responses (yes, no and unclear), were calculated based on the method described by Cichetti and Feinstein (1990).

Whiting et al., (2006) found that agreement between reviewers and the consensus rating to be worst for item 2 *(selection criteria were clearly described)*, item 12 *(same clinical data would be available when the test is used in practice)*, item 13 *(uninterpretable/intermediate test results reported)* and item 14 *(withdrawals explained)*. We had similar findings. However, overall levels of agreement between reviewers in this study were lower than those reported by Whiting et al., 2006. However, these authors used a consensual rating to measure agreement after

reviewers had discussed their scores with one another whereas we undertook a statistical comparison of the independent views of each reviewer prior to any discussion of scores and resolution of differences.

Compliance with each VQ item was calculated independently of other items. A total score was not calculated since appropriate weighting across different study designs is difficult to define. The importance of individual items is likely to vary by context and diagnostic test-type. A combined score may mask important variations in compliance between individual items (Whiting et al., 2005). Two additional items were added to the original instrument. One of the additional items included inVQ, not present within the original version of QUADAS was 'answers representative of all tests analysed within this paper'. Reviewers reported that the quality assessment of all tests described within a reference would be the same in around 90% of references with data eligible for the meta-analysis but lower in references without eligible data, which is consistent with the lower compliance with all quality items observed in this sample of references group compared to sample with eligible data. The other new item related to funding for the research conducted in the reference. Earlier work has indicated that industry funding may favour source of funding associated with reported outcomes that favour the sponsor's products (Bekelman et al, 2003; Huss et al., 2007). Based on this work we had anticipated that there might be a bias towards better estimates of Se and Sp in references where test development was funded by industry. In fact reviewers found that less than 2% of references reported that funding was from industry. However, funding source was either not reported or not clearly reported in over 40% of references.

It is possible that prior knowledge by reviewers about the subject area may influence assessment (Whiting et al., 2006). All WG reviewers in this study have worked in bTB research or eradication programmes. Furthermore, the development of the VQ instrument was discussed in detail by the WG. There was no evidence that agreement between reviewers with similar professional backgrounds was higher than between reviewers whose background differed but the sample size was small.

There was strong evidence that clarity in reporting and external validity of studies measuring the performance of diagnostic tests has improved between 1934 and 2009. Similarly compliance with many of the items measuring internal validity had improved. However, there was no evidence that compliance with items 10, 11, 13 and 14 had changed or even improved over the 75 year period. The proportion of references published in the most recent time period (2005-2009) that complied with these items ranged between 18% and 43.2%. This suggests that there may be a lack of understanding amongst scientists conducting studies of diagnostic test performance of the importance of blinding between index and reference test results and accounting for losses to follow-up in order to reduce bias.

In conclusion, critical appraisal of published diagnostic accuracy studies is essential because biases in study design may lead to overly optimistic estimates of accuracy, which could have important implications for the design of and subsequent effectiveness of disease control strategies. This analysis of the quality of studies of diagnostic tests for bTB in cattle revealed some common deficiencies in design and conduct that could lead to known biases. Probable awareness of the index test results when performing the reference test and the reverse was common in the studies

reviewed and would be unacceptable in human drug trials. Absence of information about animal withdrawals and un-interpretable study results was another common problem. Some of the poor quality scores assigned to references may have arisen because of poor reporting as opposed to methodological deficiencies in the studies. QUADAS modified for veterinary use was a useful tool for assessing and comparing study quality but might be improved further by including explicit assessment of population sampling strategy. Better education of best practice in the methodology and reporting of studies measuring diagnostic test performance is also recommended.

Funding: The SE3238 project "Meta-analysis of diagnostic tests and modelling to identify appropriate testing strategies to reduce *M. bovis* infection in GB herds" was funded by the UK Department for Environment, Food and Rural Affairs (Defra).

Authors' contributions

All authors contributed to discussion of the adaption of the QUADAS instrument, the design of the study, interpretation of the results and read, commented on and approved the final manuscript. SM, AG, AW, DA, JB, AC, RdlR, JG, JNG, SR, SR, MS, MV, EW, MW, JW RCH and SD reviewed references against VETQUADAS criteria. PA created the database for data entry with advice from JP and SD. SD conducted the analysis of the VETQUADAS data, and drafted the first version of the reference circulated to co-authors. SD was project leader.

Acknowledgements

QUADAS was developed by Whiting (2003) and VETQUADAS is a version of this instrument adapted for use in a veterinary context. Whiting and colleagues did not

play any role in the adaption and modification of their instrument. We thank James Tiller at the APHA for his help with the bespoke databases. We thank Dirk Pfeiffer and the Preventive Veterinary Medicine reviewers for helpful comments on drafts of the paper.

Additional Material

A user's guide to VETQUADAS containing the list of items with a description of how the item should be scored is provided in the online supplement.

References

Alberg, A.J., Park, J.W., Hager, B.W., Brock, M.V., Diener-West, M., 2004. The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. J. Gen. Intern. Med. 19, 460-465.

Begg, C.B., Greenes, R.A., 1983. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics. 39, 207-215.

Bekelman JE, Li Y, Gross CP. Scope and impact of financial conflicts of interest in biomedical research: a systematic review. JAMA. 2003;289:454–465.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., Lijmer, J.G., Moher, D., Rennie, D., de Vet, H.C., 2003. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Brit. Med. J. 326, 41-44.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., Lijmer, J.G., Moher, D., Rennie, D., de Vet, H.C., 2004. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Fam. Pract. 21, 4-10.

Bossuyt P.M., Reitsma, J.B., Bruns, D.E., Gatsonis C.A., Glasziou, Irwig L., Lijmer J.G., Moher, D., Rennie, D., de Vet, H.C., Kressel H.Y., Rifai, N., Golub, R.M., Altman, D.G., Hooft, L., Korevaar, D.A., Cohen, J.F. for the STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Brit. Med. J. 351:h5527.

Cicchetti, D.V., Feinstein, A.R., 1990. High agreement but low kappa: II. Resolving the paradoxes. J. Clin. Epidemiol. 43, 551-558.

Downs, S.H., Parry, J. E., Broughan, J.M., Goodchild, A.V., Upton, P.A., Nunez-Garcia, J., Abernethy, D.A., Cameron, A.R., Cook, A.J., de la Rua-Domench, R., Greiner, M., Gunn, J., Pritchard, E., Rhodes, S., Rolfe, S., Sharp, M., Vordermeier, H. M., Watson, E., Welsh, M., Whelan, A.O., Woolliams, J.A., More, S.J., Clifton-Hadley, R.S. Systematic review to identify primary research estimating the performance of ante-mortem and post-mortem diagnostic tests for bovine tuberculosis in cattle. Submitted to Prev. Vet. Med.

Feinstein, A.R., Cicchetti, D.V., 1990. High agreement but low kappa: I. The problems of two paradoxes. J. Clin. Epidemiol. 43, 543-549.

Gardner, I.A. 2010. Quality standards are needed for reporting of test accuracy studies for animal diseases. Prev. Vet. Med. 97, 136-143.

Huss A., Egger M., Hug K., Huwiler-Müntener K., Röösli M., 2007. Source of funding and results of studies of health effects of mobile phone use: systematic review of experimental studies. Environ Health Perspect. 115, 1-4.

Leeflang, M.M., Deeks, J.J., Gatsonis, C., Bossuyt, P.M., 2008. Systematic reviews of diagnostic test accuracy. Ann. Intern. Med. 149, 889-897.

Nunez-Garcia, J., Downs, S.H., Parry, J.E., Abernethy, D.A., Broughan, J.M., Cameron, A.R., Cook, A.J., de la Rua-Domenech, R., Goodchild, A.V., Gunn, J., More, S.J., Rhodes, S., Rolfe, S., Sharp, M., Upton, P.A., Vordermeier, H.M., Watson, E., Welsh, M., Whelan, A.O., Woolliams, J.A., Clifton-Hadley, R.S., Greiner, M., Meta-analysis of the sensitivity and specificity of a range of ante-mortem and post-mortem diagnostic tests for bovine tuberculosis in the UK and Ireland. Submitted to Prev. Vet. Med.

Schulz, K.F., Chalmers, I., Hayes, R.J., Altman, D.G., 1995. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA. 273, 408-412.

Viera, A.J., Garrett, J.M., 2005. Understanding interobserver agreement: the kappa statistic. Fam. Med. 37, 360-363.

VLA (Veterinary Laboratories Agency) 2010. Meta-analysis of diagnostic tests and modeling to identify appropriate testing strategies to reduce *M. bovis* infection GB herds. Final Report.

Westwood, M.E., 2005. How does study quality affect the results of a diagnostic meta-analysis? BMC Med. Res. Methodol. 5, 1-16.

Whiting, P., 2003. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med. Res. Methodol. 3, 1-13.

Whiting, P., Harbord, R., Kleijnen, J., 2005. No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol 5, 19.

Whiting, P.F., Westwood, M.E., Rutjes, A.W., Reitsma, J.B., Bossuyt, P.N., Kleijnen, J., 2006. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol. 6, 9.

Table 1

The sixteen VETQUADAS quality items, and the possible responses shown on the drop-down list

QC	No.	Question	Possible responses
E	1	Is the spectrum of animals in the study representative of the animals who will receive the test in practice?	Very representative/ Partially representative/ Not at all/ Unclear
R	2	Were selection criteria clearly described?	Yes/ No/ Unclear
I	3	Is the reference standard likely to correctly classify the target condition?	Yes/ No/ Unclear
I	4	Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?	Yes/ No/ Unclear
I	5	Did the whole sample or a random selection of the population sample, receive verification using a reference standard?	Yes/ No/ Unclear
I	6	Did the animals receive the same reference standard regardless of the index test result?	Yes/ No/ Unclear
I	7	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	Yes/ No/ Unclear
R	8	Was the execution of the index test described in sufficient detail to permit replication of the test?	Yes/ No/ Unclear
R	9	Was the execution of the reference standard described in sufficient detail to permit its replication?	Yes/ No/ Unclear
I	10	Were the index test results interpreted without knowledge of the results of the reference standard?	Yes/ No/ Unclear
I	11	Were the reference standard results interpreted without knowledge of the results of the index test?	Yes/ No/ Unclear
E	12	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	Yes/ No/ Unclear
I	13	Were un-interpretable/ intermediate test results reported?	Yes/ No/ Unclear

I	14	Were withdrawals from the study explained?	Yes/ No/ Unclear
O	15	What was the source of funding for the study?	Industry/ Public Charity/ Mixed/ Not reported or unclear
O	16	Do you consider that the above answers are representative of all tests analysed within this paper?	Yes/ No/ Unclear

Footnote

QC: Quality Category, E: External validity, I: Internal validity, R: Clarity in

Reporting, O: Other

Table 2

Contingency table layout showing the joint distribution of responses by two reviewers

Responses by reviewer B	Responses by reviewer A			
	NO	YES	UNSURE	Totals
NO	a	B	c	k1
YES	d	E	f	k2
UNSURE	g	H	i	k3
Totals	j1	j2	j3	

Table 3 Percentage agreement in possible responses to each VETQUADAS quality item for 83 references with two reviews and eligible data for the meta-analysis

VETQUADAS Items	Evaluation category	Percent Agreement between reviewers (%) Yes	No	Unsure	Overall
All	all	71.0	29.0	32.3	55.8
2	R	66.7	27.8	9.1	50.6
8	R	85.7	34.8	nd	73.5
9	R	63.4	11.1	27.6	45.7
3	I	71.6	37.8	20.0	57.8
4	I	67.3	nd	32.0	55.4
5	I	79.7	20.0	nd	66.3
6	I	82.5	45.2	nd	71.1
7	I	82.6	9.5	28.6	71.1
10	I	46.2	31.1	40.6	39.8
11	I	51.4	35.3	58.7	53.0
13	I	54.0	37.9	31.1	42.2
14	I	40.0	30.0	45.1	39.8
1	E	78.0	29.6	nd	62.7
12	E	71.2	10.0	nd	51.8
Reviewers from same professional group					
All		73.0	24.4	36.5	57.4
Both reviewers were veterinarians					
All		72.1	25.0	28.6	55.7
Reviewers not from same professional group					
All		70.0	31.1	30.4	55.0

Footnote to Table 3:

'Partial' was coded to 'yes' for the analysis of item 1.

R=Clarity in reporting, I=Internal validity, E=External validity

nd: not possible to determine because of low number of responses.

Professional group: The 18 reviewers were allocated to one of three groups based on their scientific training and background: veterinarians (n=9), laboratory scientists (n=5) or quantitative scientists (n=4).

Table 4 Percentage of references compliant with 14 VETQUADAS quality items, by number of reviewers and eligibility of performance estimates for the meta-analysis

	Eligible data for meta-analysis		No eligible data for meta-analysis		
Item no.	Two* reviewers n=83	One reviewer n=31	Two* reviewers n=24	One reviewer n=52	P value for Difference
Percent compliance with VQ items measuring clarity in reporting					
2	72.3	54.8	50.0	26.9	<0.001
8	83.1	67.7	75.0	53.9	0.003
9	66.3	45.2	58.3	36.5	0.006
Percent compliance with VQ items measuring interval validity					
3	57.8	70.9	41.7	42.3	0.038
4	75.3	45.2	50.0	40.4	0.001
5	80.7	77.4	70.8	48.1	0.001
6	74.7	80.7	75.0	55.8	0.048
7	86.8	83.9	79.2	57.7	0.001
10	36.1	22.6	25.0	11.5	0.016
11	49.4	41.9	33.3	38.5	0.432
13	37.4	41.9	33.3	21.2	0.163
14	31.3	29.0	37.5	17.3	0.212
Percent compliance with VQ items measuring external validity					
1	78.3	64.5	58.3	55.8	0.033
12	68.7	64.5	54.2	26.9	<0.001

Footnote to Table 4:

'Partial' was coded to 'yes' for the analysis of item 1.

*A random number generator was used to randomly select one reviewers responses for each item for the calculation of reported percentages and conduct of χ^2 tests.

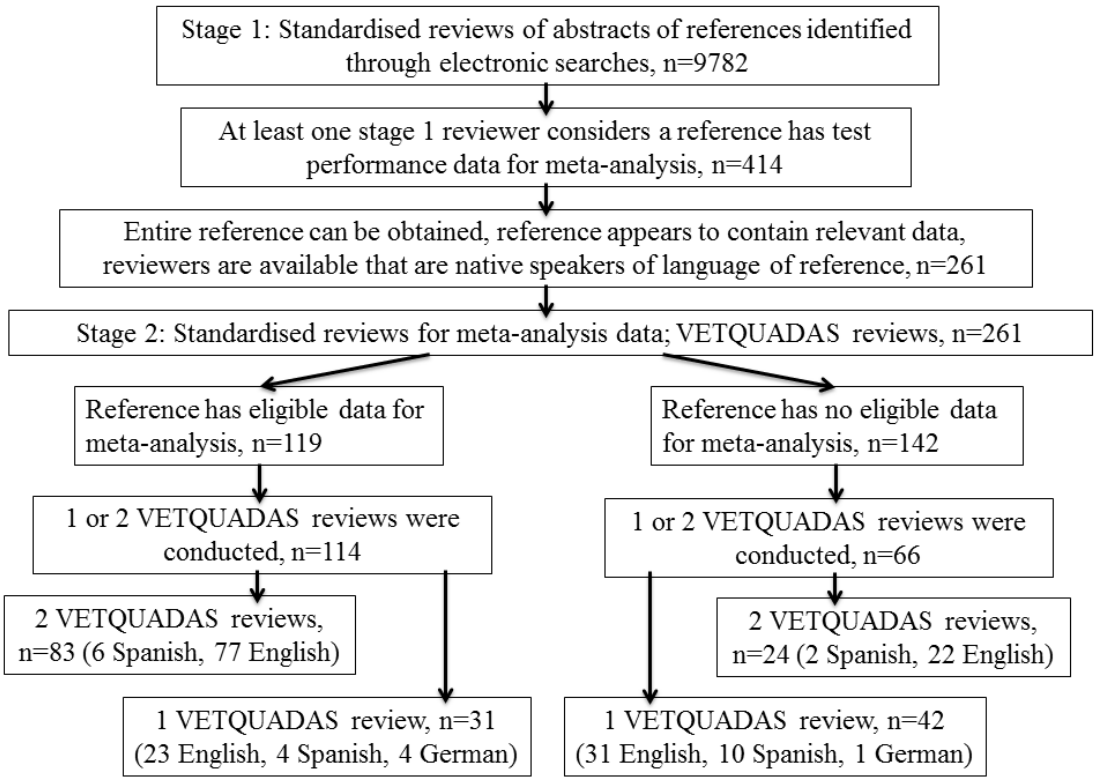
Table 5 Percentage compliance with VETQUADAS items by year of reference publication

Item no.	1934-1959 n=11	1960-1969 n=11	1970-1979 n=19	1980-1989 n=17	1990-1994 n=14	1995-1999 n=34	2000-2004 n=40	2005-2009 n=44	Test for trend P value
Percent compliance with VQ items measuring clarity in reporting									
2	27.3	27.3	57.9	35.3	50.0	55.9	72.5	56.8	0.008
8	36.4	36.4	73.7	70.6	64.3	76.5	87.5	72.7	0.001
9	9.1	27.3	63.2	47.1	42.9	61.8	72.5	50.0	0.002
Percent compliance with VQ items measuring interval validity									
3	45.5	9.1	68.4	58.8	57.4	58.8	50.0	56.8	0.293
4	45.6	36.4	63.2	58.8	35.7	52.9	67.5	59.1	0.302
5	45.5	27.3	84.2	70.6	64.3	67.7	80.0	75.0	0.012
6	36.4	45.5	73.7	82.4	64.3	61.8	85.0	75.0	0.004
7	45.5	54.6	84.2	70.6	78.6	88.2	80.0	79.6	0.013
10	9.1	18.2	36.8	41.2	14.3	26.5	25.0	25.0	0.606
11	27.3	18.2	26.3	29.4	71.4	52.9	50.0	43.2	0.036
13	18.2	45.5	57.9	23.5	35.7	26.5	37.5	27.3	0.567
14	18.2	27.3	47.4	23.5	42.8	23.5	32.5	18.2	0.484
Percent compliance with VQ items measuring external validity									
1	45.5	45.5	57.9	58.8	64.3	73.5	75.0	75.0	0.009
12	36.6	18.2	63.2	58.8	57.1	41.2	62.5	65.9	0.031

Footnote to Table 5

Non-parametric test for trend over year
A random number generator was used to randomly select one reviewer's responses for each item for references with two VETQUADAS reviews.

Fig.1. Stages of systematic review and numbers of VETQUADAS (VQ) reviews



Footnote to Fig. 1.

References written in Spanish were assigned for review to the two native Spanish-speakers in the WG. References written in German were reviewed by the one native German speaker in the WG. For other non-English language references, data were extracted through structured interviews with native speakers who were scientists at APHA who were native speakers but were not part of the WG. These references did not undergo a VETQUADAS review. Further detail of the methodology of the systematic review can be found in Downs et al., (2017).

Fig 2. Mean response to each of the 14 VETQUADAS (VQ) quality items for 83 references with eligible performance data and had two VQ reviews

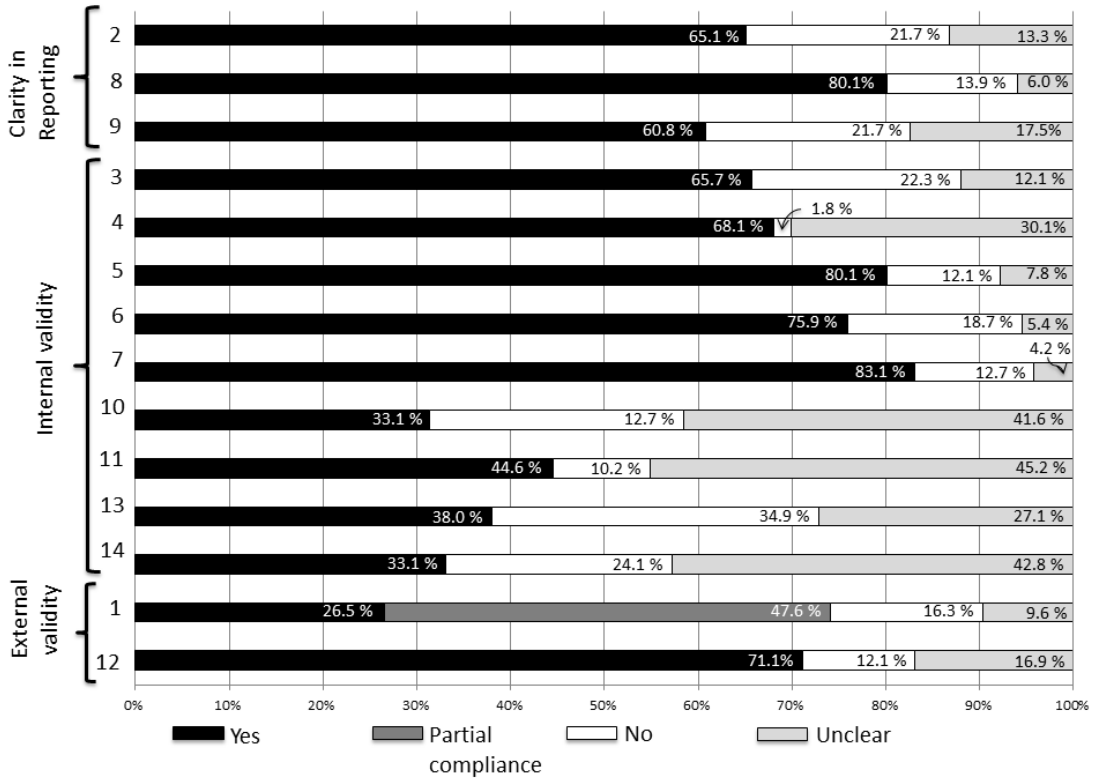
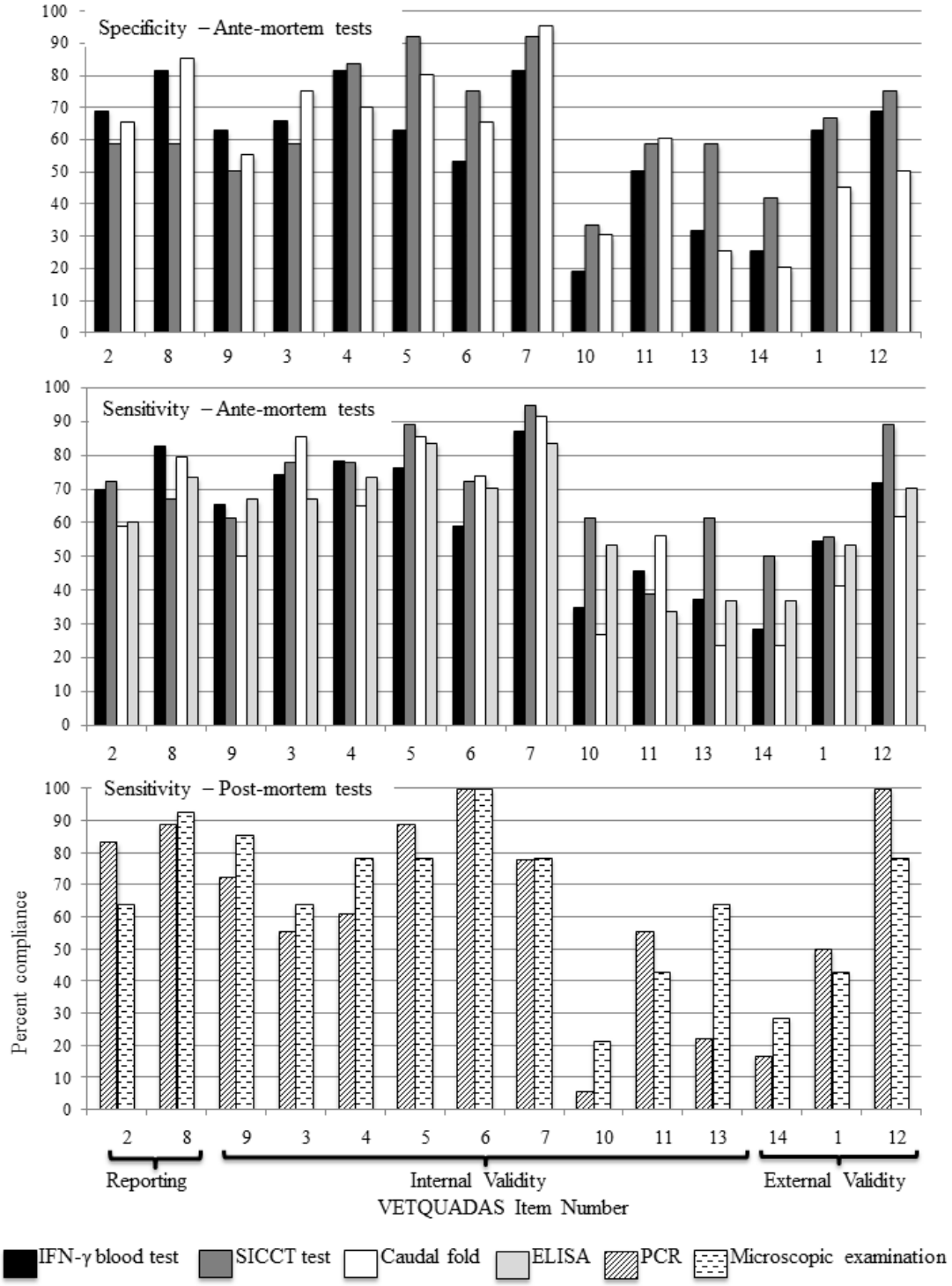


Fig. 3. Percentage of references within six different diagnostic test types that complied with VETQUADAS items measuring study quality.



Footnote to Fig. 3.

Test-types had to have five or more references each reviewed by two Working Group members for inclusion. The years of publication of references were IFN-γ blood test (1991-2009), Single Intradermal Comparative Cervical Tuberculin (SICCT) test (Se:1953-2006, Sp: 1975-2006), ELISA (Se: 1981-2007, Sp: 1981-2004), Caudal fold tuberculin skin test (se: 1934-2007), Polymerase Chain Reaction (PCR) (Se: 1995-2008), Microscopic examination (Se: 1940-2008).