Structured approach to design of diagnostic test evaluation studies for chronic progressive infections in animals

Søren Saxmose Nielsen, Nils Toft, Ian Andrew Gardner

To cite this version:

Søren Saxmose Nielsen, Nils Toft, Ian Andrew Gardner. Structured approach to design of diagnostic test evaluation studies for chronic progressive infections in animals. Veterinary Microbiology, 2011, 150 (1-2), pp.115. 10.1016/j.vetmic.2011.01.019. hal-00687332

HAL Id: hal-00687332
https://hal.science/hal-00687332
Submitted on 13 Apr 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Structured approach to design of diagnostic test evaluation studies
for chronic progressive infections in animals

Søren Saxmose Nielsen, Nils Toft, Ian Andrew Gardner

1Department of Large Animal Sciences, Faculty of Life Sciences, University of Copenhagen,
Grønnegårdsvej 8, DK-1870 Frederiksberg C, Denmark

2Department of Medicine and Epidemiology, School of Veterinary Medicine, University of California,
Davis, CA95616, USA

Corresponding author: Søren S. Nielsen, Department of Large Animal Sciences, Grønnegårdsvej 8,
DK-1870 Frederiksberg C, Denmark, e-mail: ssn@life.ku.dk, fax: +4535333022, tel: +4535333096.
ABSTRACT

Diagnostic test evaluations (DTEs) for chronic infections are challenging because a protracted incubation period has to be considered in the design of the DTE, and the adverse effects of infection may be widespread and progressive over an animal’s entire life. Frequently, the specific purpose of the test is not formally considered when a test is evaluated. Therefore, the result is often a DTE where test sensitivity and specificity estimates are biased, either because of problems with establishing the true infection status or because the test detects another aspect of the infection (and analyte) than originally intended.

The objective of this paper is to outline a structured approach to the design and conduct of a DTE for diagnostic tests used for chronic infections in animals, and intended for different purposes. We describe the process from reflections about test purpose and the underlying target condition through considerations of the pathogenesis, and specification of a practical case definition, which can subsequently be used in the DTE for the specific purpose.

The process is illustrated by two examples of Mycobacterium avium subsp. paratuberculosis (MAP) infections in cattle. MAP infections are chronic and can result in different adverse effects at different time points during the incubation period. The description provides input on the process and deductive reasoning which are integral parts to develop a high-quality design of a DTE for chronic infectious diseases.

Key words: chronic infectious diseases; design strategy; diagnostic test evaluation; paratuberculosis
1. Introduction

Chronic progressive infectious diseases such as brucellosis, tuberculosis, leishmaniasis and paratuberculosis pose special challenges in the design and conduct of diagnostic test evaluations (DTEs). Specifically, the long and to some extent unpredictable disease progression can result in large variation in sensitivity and specificity estimates from DTEs of similar tests (Nielsen and Toft, 2008).

Lack of explicit focus on the specific utility of a diagnostic test under evaluation, often termed the index test, is a likely explanation, i.e. the purpose of testing has not been made explicit and therefore test results are interpreted in relation to multiple purposes without really being evaluated specifically in relation to these purposes. The World Organisation for Animal Health (OIE) endorses the use of the “fitness for purpose” criterion, where purposes for example could be to demonstrate freedom from infection in a defined population, to confirm a clinical diagnosis, or to determine the immune status of individuals (OIE, 2009). “Fit for purpose” means that test results must be interpreted to a specific meaning and purpose. This purpose needs to be defined a priori by the decision makers, who must specify what they want the test to detect (an objective) before a DTE can be initiated. For chronic infections, the timing of the diagnosis, which corresponds to stage of disease, needs to be included in the purpose description.

Early diagnosis can for example reduce the risk of transmission of the causative agent and disease progression. However, some animals at risk of progressing to the stage of “disease” or to the stage of “transmission” may never reach these stages, either because they are genetically resistant to developing disease, are protected by vaccination or because they will not live long enough. Therefore, treatment or culling may be a costly alternative to doing nothing. Late-stage disease diagnosis will often result in a high correlation between test result and clinical signs, and thereby increase the decision makers’ confidence in the test-results. However, transmission may already have already occurred without being detected.

Many DTEs of chronic infections are flawed by improper designs and inherent biases both from design issues and from lack of consideration of the close relationship between index test – purpose – target and context of application, exemplified by tuberculosis, HIV and malaria in humans (Fontela et al., 2009), and by Mycobacterium avium subsp. paratuberculosis (MAP) infections in animals (Nielsen and Toft, 2008). We believe that the primary reason for the design flaws stems from not really putting the test and its application into perspective prior to commencing the DTE.
The process and general principles for DTEs are described by TDR Diagnostics Evaluation Expert Panel (2006) and Greiner and Gardner (2000a). However, these generic guidelines do not specifically deal with the challenges of chronic infectious diseases in modern veterinary medicine. Thus, the objective of this paper is to provide readers with a structured approach to address the most important issues involved in designing (and conducting) a DTE for chronic, progressive infections. Specifically, we consider biases imposed by the choice of case definition compared to target condition, biases imposed by the choice of design, and biases occurring from the practical issues in conducting the study.

To illustrate the specific considerations, we use MAP infection in ruminants (paratuberculosis) as our example. MAP infections cause significant economical losses globally and control is most wanted in countries with a major dairy industry (Kennedy and Nielsen, 2007). Diagnosis is pivotal to infectious disease prevention and control. However, a long and partially unpredictable incubation period makes diagnosis of paratuberculosis challenging. Biased test information inevitably arises, when diagnostic tests are evaluated and used. Specifically, we use two examples of a purpose to illustrate how the initial choices influence the remaining study design decisions.

2. Overview of the Diagnostic Test Evaluation (DTE)

A key element in any successful study involving data collection is the planning phase, where the association between the context (including already known features of the problem), the objective of the current study, and the characteristics of the target and study populations are considered. Only when these aspects of the problem and their interrelationships are understood, should planning commence. Thus before considering a study design, many other decisions or choices have to be made. Hence, development of a well-designed DTE includes considerations and understanding of the following elements in the initial planning phase prior to considering a design:

a) Pathogenesis, including immune responses, predilection sites for the organism, effects of the infection, clinical manifestations and occurrence of any pathognomonic signs;

b) Analytical principle of the index test, i.e. what analyte does it detect and how is it detected;

c) Which alternative tests are available, and what are their strengths and limitations, including performance (e.g. sensitivity, specificity and reproducibility) and operational characteristics (e.g. speed of performing the test, ease of sampling, test-costs);
d) Context in which the index test is to be used, i.e. what are you trying to identify and why.

We believe that the logical flow in planning and designing a DTE should follow Fig. 1. A proper DTE is usually preceded by an initial phase in which an assay is developed and optimised. This phase should not be mixed with the actual DTE. Once the index test has been defined, it should be optimised. The starting point must be consideration of the purpose, i.e. why, how and where will the test be used once it has been evaluated. The optimisation may be closely related to the pathogenesis, which should therefore also be considered in the initial phase. Understanding the purpose and considering the relevant components of the pathogenesis is an iterative process, i.e. if the purpose of the test is to detect animals which may become infectious, then it is important to understand the process leading up to or causing shedding of the infectious agent, whereas a purpose such as confirming clinical disease might make other components of the pathogenesis more important. Only after the purpose has been defined, should the target condition be specified. The target condition is the underlying infection condition that the index test should detect. This condition has to reflect the purpose of testing as much as possible, i.e. if we want to optimise economically-important production measures and control infection, there may be little reason to focus on diseased animals, if infected non-diseased animals can also transmit the agent.

After selection of the target condition, one may then proceed to determine the case definition. The case definition is based on the results of the reference standard, against which the index test should be compared. Again, it is imperative that there is a high degree of association between the underlying target condition and the more pragmatic case definition, which may be considered as the practical description of the target condition. The more discordant they are the more problems in terms of biases will be inferred. As an example, the purpose could be to prevent transmission of an infectious agent. The target condition could then be infectious animals (those that currently transmit the agent), and a case definition might be all animals testing positive in a specific agent-detection test. The challenges incurred might be that a test-positive animal excreting low numbers of the infectious agent may not be infectious, or a test-negative animal may become infectious the day after it was tested.

Once the purpose, pathogenesis, target condition and case definition have been defined, the design follows logically. There are of course practical considerations and constraints such as time,
resources and logistics, which may influence the actual design. However, the process should ensure that the link between purpose and DTE design remains intact as shown in Fig. 1.

3. Purpose and target condition

Decision makers should consider a specific purpose for the use of a diagnostic test, for example to maximise profit at the herd-level, to eradicate an infectious agent from a population, or to determine if an animal is infected. With this overall goal in mind, a purpose of testing and a relevant target condition can be selected.

The sensitivity and specificity estimates of the index test express the probabilities being test-positive among animals with the condition and being test-negative among animals without the condition, respectively. Specific stages of a chronic infection are different underlying conditions, and the choice of target condition therefore affects both sensitivity and specificity, and ultimately the interpretation of test results. For example, three target conditions for MAP tests were defined in Nielsen and Toft (2008): 1) “infected”, 2) “infectious”, and 3) “affected”, describing three different underlying conditions where the latter two are subsets of the first. Hence, these target conditions are not mutually exclusive, in fact, they should more be seen as successive stages of the disease. It might, however, be possible to move from “infected” to “affected”, without any discernable “infectious” stage. The choice of a specific target condition ultimately relies on the purpose of testing. MAP infected means that the animal is “infected” with MAP; that is MAP have crossed the animal’s natural barriers (usually the intestinal epithelium) and is persisting in the animal without simply being transported passively through the intestinal tract in digested material. This animal could be of relevance for herd classification, or for evaluating the risk of the animal becoming infectious or diseased in the future.

The MAP infectious animal sheds sufficient bacteria in milk or faeces to infect susceptible animals horizontally, or can transfer sufficient MAP vertically to result in infection of a susceptible foetus. The sensitivity in this case describes the probability of being test positive if the animal is MAP infectious. The specificity expresses the probability of being test negative among non-infectious animals, i.e. an animal can be infected but non-infectious and still be a non-case.

Finally, the MAP affected animal can be of interest because it is adversely affected by the infection, thereby losing financial value because of reduced production performance (milk yield and...
body condition), although this reduced performance may not be clearly visible to an observer in early phases of infection. For other chronic infections, such as bovine tuberculosis, other purposes and target conditions might be relevant. For example in Denmark, *M. bovis* has been eradicated and the purpose and context of current testing is at the slaughterhouse to demonstrate the continued freedom from infection in the national herd.

4. Pathogenesis

Knowledge of the pathogenesis of a chronic infection is a pre-requisite to define a relevant target condition, and thereby the purpose of the study. However, the purpose of the study should usually be the primary motivator for identification of a relevant test, but the purpose can only be specifically described once the pathogenesis is adequately understood. A continuous and iterative update of the purpose - pathogenesis – target condition triad (Fig. 1) is therefore required.

For example, MAP infections are presumed to occur in calfhood (Windsor and Whittington, 2010), although time of natural infection is practically impossible to determine in field settings. Adult infections (after first calving) probably occur in some dairy management systems through contaminated feed, drinking water or ingestion of waste water from holding lagoons that is used to flush freestall alleyways. Unpublished data indicates female replacements introduced from a presumed non-infected herd into a MAP infected herd became infected based on ELISA test results and evidence of faecal shedding about 2 to 3 years post entry (Wells et al., 2010). However, it is likely that new adult MAP infections are a small contributor to the total infection pool and for simplicity, a calfhood infection scenario is the only one considered hereafter.

Chronicity of MAP infections is likely one of the most important factors affecting the ability to correctly diagnose “paratuberculosis”, or specifically the defined target condition in a specific study. The host responses to the infection determine the length of the incubation period. However, the host responses to MAP are still poorly understood which makes interpretation of diagnostic test results a challenge. Current understanding of the host responses is described in detail in Coussens (2004).

Briefly, pro-inflammatory and anti-inflammatory immune responses both play a role in the pathogenesis of MAP infections. The former may indicate that the animal has control of the infection, whereas the latter may suggest that control of the infection is lost (Coussens, 2004). Pro-inflammatory responses include production of cytokines such as interferon-γ and interleukin 10 (IL-10; Nagata et al.,
2010) and IgG2 type antibodies, whereas anti-inflammatory responses usually are characterized by antibodies of the IgG1 type (Koets et al., 2001).

The transition from pro- to anti-inflammatory responses is unpredictable, although animals in the range from 2-5 years of age are more likely to be shedding MAP and have IgG1 antibodies than older and younger animals (Nielsen and Erskbøll, 2006). Whether anti-inflammatory responses can revert back to pro-inflammatory responses is not known. The anti-inflammatory phase is usually accompanied with bacterial shedding (Collins, 2002; Nielsen, 2008), declines in milk production (Nielsen et al., 2009) and reduction in body weight (Kudahl and Nielsen, 2009). Furthermore, end-stage infection can be characterised by extensive bacteraemia (Alonso-Hearn et al., 2009), and the animal can die from the infection if not slaughtered. A summary of the current understanding of the pathogenesis and the effects of MAP infections is shown in Fig. 2.

5. Case definition

Target conditions are underlying entities, which we should define in our DTE. The practical realisation of our target condition is the case definition which can be made both at both the animal and herd level. The case definition must reflect the target condition. If the target condition and case definition differ, misclassification bias occurs.

A clear and rule-based definition of a case is required in most situations. All non-cases can then be considered controls, or some non-cases can be excluded. Still, exclusion should be considered carefully. For example, if a case is a cow infected with agent X and presenting with diarrhoea, then all other cows are non-cases. However, diarrhoea can be caused by both agent X and agent Y. If the diarrhoea is actually caused by agent X, misclassification of the cases occur, but exclusion of the cow can result in spectrum bias (see Section 6.2 for details). One or more diagnostic reference tests can be used to define the case definition, but these tests should be independent of the index test. For example, if a case definition of “affected by agent X” is specified by occurrence of diarrhoea and antibodies to agent X, then we should not evaluate an antibody test against this case definition. Once we have defined a case definition, we can start considering relevant index tests. In some situations, a case is not defined, a priori and latent class statistical methods are used to analyse data. The case definition in these situations is dependent on the joint results of the diagnostic tests, and a “latent” undefinable condition constitutes our case definition. For example, if two antibody tests are used and
evaluated against each other, the latent condition is presence of antibodies rather than infection status. If an agent-detection and an antibody-detection test are used, the latent condition reflects the mutual condition that exists between the presence of pathogen (infection) and immune-response (response to infection). These undefinable conditions may be difficult to interpret, particularly if no prior knowledge exists about how the co-existence of the two analytes relates to pathogenesis.

6. Bias

Biases are systematic errors, which result in under- or overestimation of the parameters of interest, here primarily sensitivity or specificity. Complete random sampling as part of the study design should result in unbiased estimates, whereas bias resulting from host responses may still be present and should be taken into account when assessing generalisability of results.

6.1. Bias resulting from host responses

False-negative and false-positive reactions are highly likely to occur with most tests, resulting in reduced test sensitivity and specificity, respectively. If infection has not reached a specific disease stage, analytes characteristic of this stage cannot be expected and thus will not be detectable. If infection has reached one disease stage and reverted (or been cleared), the analytes may still be present. Furthermore, the virulence of strains of a pathogen may differ and affect the subsequent pathogenesis and immune response.

6.2. Bias resulting from study design

Biases are hard to avoid in DTEs of chronic infections, primarily because multiple infection markers (e.g. antibodies) occur relatively late. These misclassifications resulting from the sequence of biological processes are usually what we want to elucidate, i.e., we want to demonstrate how the index test deviates from the target condition. Estimates of sensitivity and specificity can be used to provide more reliable estimates of prevalence (Greiner and Gardner, 2000b), if the sensitivity and specificity estimates are unbiased. However, biases in sensitivity and specificity estimates resulting from design of the DTE should be avoided.

These design-type biases include different types of selection bias such as “disease progression bias”, “diagnostic access bias”, “work-up bias” or “verification bias”, “incorporation bias”, “loss to follow-
“up bias”, and “spectrum bias”. Other biases can occur in DTEs in general (Kelly et al., 1997), but the above-mentioned frequently occur in veterinary medical studies, and many other types are special cases of selection bias.

a) Disease progression bias can occur, if the time between sampling for the DTE and the reference test is so long that disease has progressed from one stage to another. A biological plausible time-span must be selected to avoid this type of bias, e.g. for a chronic infection like MAP, little disease progression is expected to occur in one week, whereas progression from one stage to another can occur in 3 months. Furthermore, if animals are intermittently excreting a pathogen, and the pathogen is present on one test-day but not the next, infection can have progressed/regressed and therefore the case definition may vary between test-days. One solution for dealing with intermittent excretion is to have a case definition aggregated over multiple testing days, e.g. pathogen was excreted at least once during a week.

b) Diagnostic access bias can occur if part of the study population is not geographically representative of the target population, e.g. a negative reference population is selected from an area considered to be free of the infection. If this area is also free of other infections, which may induce cross-reactions in the index test, diagnostic access bias occurs. In this particular example, we would overestimate the specificity.

c) Work-up bias or verification bias occurs if the target condition is more likely to be verified for some study objects than for others. Assuming that animals subject to verification are a random sample of the index test positive and index test-negative groups and the same verification procedure is used, then the bias is correctable statistically (Begg and Greenes, 1983). However, verification bias correction may be inadequate when the sensitivity of the screening test is close to perfect (Cronin and Vickers, 2008). In contrast, if a cow with clinical disease is subject to several reference tests to verify that infection is present, whereas the non-clinical cow is subject to only one reference test, the induced verification bias is not correctable, unless at least some of the non-cases are verified. Blinding of results of the index test and reference standard to those performing the test should prevent many forms of verification bias.

d) Incorporation bias occurs when the definition of the target condition is partially based on the index test. If one antibody and four bacteriological tests are carried out on samples from the same animal, and the case definition is that at least one should be positive in one of the
bacteriological tests for a sample to be considered a case, then the bacteriological tests should not be included as index tests, because the bacteriological tests are part of the case definition.

e) Loss to follow-up bias occurs when study subjects are more likely to exit a study because of positive compared with negative test results.

f) Spectrum bias (Ransohoff and Feinstein, 1978) is perhaps the most frequent type of bias in DTEs of chronic infections. This type of bias occurs when animals in specific infection stages are more likely to be sampled than a randomly-selected animal from the target population. Furthermore, many DTEs are carried out in high-prevalence herds because these herds have more diseased animals and therefore, fewer herds need to be visited to obtain the required sample size. However, the spectrum of disease is often different in high-prevalence herds compared to low-prevalence herds, for example because they are in a different temporal phase of an epidemic. Disease spectrum in a herd could also be affected by interventions, where the most diseased or infectious animals are removed because of positive test results.

Regardless of its underlying source, bias in individual animal-level sensitivity and specificity estimates can result in substantial bias in herd sensitivity and herd specificity estimates when the latter are obtained by extrapolation of individual values to "herd tests" of varying size (Christensen and Gardner, 2000).

To avoid bias associated with design issues, the above should be considered carefully in the design phase. The QUADAS tool (Whiting et al., 2003) can be used as a check-list to avoid these, but considerations of sampling and case definition are key factors to avoid bias. The best way of avoiding these biases is use of complete random sampling of the animals (and herds) used in the study. A biased sampling frame such as selection of cases and controls from a population of animals visiting a veterinary clinic may occur because these animals do not necessarily represent the target population, for example, young animals are often at a veterinary clinic for vaccination purposes and may be selected as controls, but they are less likely to have a chronic disease because they are young.

7. Study designs
Many study designs can be used for DTEs. Essentially, the classic observational studies (cross-sectional, case-control or cohort types), or derivatives thereof, can be used. An overview is given in Table 1, although other derivatives also can be used. The choice of study design depends on the choices and effects outlined in previous sections and shown in Fig. 2.

Choice of study design often depends on practical issues such as costs and logistics. Detailed pathological examination may not be feasible in some countries and regions, whereas a high-throughput sampling system based on milk samples is easy to implement. Costs of testing may determine, if a study can be funded. However, these testing costs may vary considerably among regions because of variable experience with the assay and frequency of use, and different labour and equipment costs.

8. Examples

Two examples are provided in the following sections and the implications of decisions made in the process are discussed.

8.1. Example 1. Eradication of MAP

8.1.1. Purpose and target condition

Animal health officials may have decided that MAP needs to be eradicated from a defined population such as a herd or a region. To achieve this goal, a diagnostic test is needed to detect all infected herds and animals with removal of test-positive animals. Given this purpose, an objective for the planned DTE could be to evaluate a diagnostic test, which can be used to detect MAP infected animals, because they pose a risk for future spread of MAP. Our target population is given by the definition of the population, i.e. all animals in the specified population that can be or are infected with MAP.

8.1.2. Pathogenesis

The purpose “eradication” requires that we focus on detection of “MAP infected” animals as our target condition. Multiple questions should then be considered in relation to the pathogenesis: How do we define an infected animal? Can infected animals clear the infection? Which test is most likely to detect the infected animal, i.e. which analytes (specific cytokines, antibodies, MAP etc) have the
highest probability of being present. If the animal is truly infected, what will be the most cost-effective
action to take to achieve the purpose? Currently, we do not know if MAP infections can be cleared by
the cell-mediated immune reactions. If not, early detection of infection should target cell-mediated
immune reactions, if specific targets are available. MAP specific interferon-gamma detection has been
proposed in animals >15 months of age, as younger animals may have non-specific circulating
interferon-gamma (Jungersen et al., 2002; Olsen et al., 2005). Detection of interferon-gamma or other
cytokines from the cell-mediated immune responses may therefore be appealing because of the
potential early detection, but has the disadvantage that non-specific analytes may be present without
infection ever being present, and infection may have been cleared. A robust cell-mediated immune-
response may contain the infection for the remainder of the animal’s life, and a farmer may be
reluctant to cull this cow, as the decision may not be cost-effective. However, such reasoning must be
deemed irrelevant considering the choice of target condition. If a cost-effective diagnosis for this
particular farmer’s purpose is required, we might have to change the target condition.

Detection of MAP specific IgG1 antibodies produced during anti-inflammatory immune reactions
may also be appealing. A disadvantage is that the infection should have progressed significantly, and
animals are usually 2 to 5 years of age, or even older, before infection can be detected. An advantage
is that infection may have almost have progressed to the stage, just before the animal starts shedding
infectious doses of MAP. If animals are tested regularly, they may be detected before this occurs.

Furthermore, logistics about sampling of milk or serum to detect IgG1 antibodies is currently
much easier than collection of samples for measurement of interferon-gamma, because the latter
method requires live T-cells in the samples when they arrive at the laboratory. The last option
mentioned here is detection of MAP through culture or PCR techniques. During pro-inflammatory
immune responses, the MAP or MAP DNA may be shed in insufficient amounts to be detected, and
when shed in significant amounts, the detection may be too late. Considerations about possible false-
positive and false-negative test-results should be made for the agent-detection tests.

8.1.3. Case definition

We wish to detect the MAP infected animal, but have to find an appropriate reference method to
define this in practice. Previous research suggests that MAP can be present in the animal without
being detected even if the predilection sites in the intestinal tissues are sampled. Samples from 100
tissues from multiple sites are suggested for culture (Whitlock et al., 1996). However, obtaining this number of tissues may be practically infeasible and too costly to motivate if a large sample size is required to obtain required precise sensitivity and specificity estimates. A feasible sample might include 5 intestinal tissues (e.g. 2 x jejunum, 1 x ileo-cecal valve, 1 x colon and 1 x cecum), and 5 lymph nodes (e.g. 1 x retro-pharyngeal lymph nodes (lnn), 1 x hepatic lnn., 1 x jejunal lnn., 1 x ileo-cecal lnn., 1 mesenteric lnn.). Culturing these 10 samples would yield 10 results, and it should be specified how to aggregate them into an overall result for the animal. For example, if only a single sample is positive, the animal could be considered a case, and all other animals would be considered non-cases.

8.1.4. Bias from host responses

Sensitivities of tests detecting anti-inflammatory responses and bacterial shedding to detect the target condition of MAP infected are likely to be highly affected by chronicity of infection, although age can to some extent be used to improve the diagnostic predictions.

The specificity of tests detecting pro-inflammatory responses can be difficult to interpret, because it is not known if infected animals can eliminate the infection (Mikkelsen et al., 2009). If infections can be eliminated, and persisting indicators of pro-inflammatory responses are present, these will result in false-positive reactions after the elimination.

8.1.5. Bias from study design

Choice of the case and non-case definitions can significantly affect results. A typical situation with spectrum bias from studies of the target condition of MAP infected could be use of banked samples. If a case was defined as an animal with detectable excretion of MAP in 3 g of faeces, and the target condition was “infected with MAP”, all infected animals should be expected to excrete detectable amounts of MAP in 3 g faeces, which is unlikely in early stages of infection. Choosing all cases among the animals with significant shedding of MAP would therefore impact the distribution of animals towards those with more severe infection with resultant higher test sensitivity values.

For specificity studies, it is not uncommon to define as non-infected, herds that have previously been found negative by herd screening using an antibody ELISA. The lack of test-positive reactions may be an indicator of lack of infections caused by MAP-related mycobacteria such as *M. avium*
subsp. *avium* in this herd. If these closely-related mycobacteria are common in other herds of the target population, diagnostic access bias is introduced, and the specificity might be overestimated. Furthermore, if the antibody ELISA is subsequently evaluated in this population, incorporation bias is introduced, because the antibody ELISA is part of the case definition: the non-infected animal is defined as an animal derived from a previously ELISA-negative herd.

### 8.1.6. Sampling

The sample must reflect the population, which is the general dairy population. Thus some element of random sampling is required. The challenge is most likely that the case definition (and not least the conclusive classification of a non-case) will require some choices with respect to targeted sampling. For example, selecting cases might be based on prior knowledge of MAP being confirmed in certain herds and controls might be selected from (closed) herds without history of disease. Thus, at herd level the sampling might be considered purposive to ensure that both cases and controls can be established. However at the individual animal level, random sampling should be applied to avoid further biases.

### 8.1.7. Study design

A random cross-sectional sample of animals would be ideal to avoid spectrum bias. The desired precision of the sensitivity and specificity estimates should be determined from sample size calculations (Jacobson, 1998). The study animals should be subject to pathological examination immediately after the sample for the index test has been obtained. Sampling of animals at a slaughterhouse may result in spectrum bias, because these animals may be more likely to have been culled because of late-stage MAP infection, or simply because they are old, and older animals are more likely to have late-stage infection than younger animals. On the other hand, the most advanced cases i.e. cows dying from infection and those euthanized at the herd are not included in the sample.

An alternative is to obtain a random sample of the animals and conduct a longitudinal study using a reference test and the index test at slaughter to classify the animals. The advantage of this method is that exit status of the animal would show if the animals have actually progressed to the infection stage of MAP infectious, in which the animals transmit MAP. A disadvantage is that the study period can be quite long because some animals initially subjected to the index test may live for many
years before they exit the herd. Furthermore, the actual infection status will not necessarily be
determined. Therefore, the actual target condition should be redefined to “a MAP infected animal,
which does not reach the MAP infectious stage in a life-time”. Animals removed from the herd,
regardless of whether for MAP infection or other reasons, would be expected to do this at the same
rate as in the target population, and therefore reflect this adequately. Furthermore, if the animals are
tested repeatedly, and they are test-positive early but become test-negative later, these animals may
be considered false-positive and therefore the specificity can be assessed. However, a cow that is
test-positive on the last test before she is slaughtered cannot be used in the study, because we would
not know if she is a true positive or a false-positive.

8.2. Example 2: Confirmation of clinical paratuberculosis

8.2.1 Purpose and target condition

A narrower purpose than described in Example 8.1 is confirmation of MAP infection in a clinical
suspect animal. This purpose is relevant to veterinarians and farmers, who suspect that a clinically
diseased animal may be infected with MAP, and that MAP is the cause of the disease observed. The
target population is now only diseased animals, but there is a need to refine this further, e.g. an animal
with pneumonia would normally not be included, because MAP is not known to cause pneumonia (see
section 4 on Pathogenesis). The target population could be described as “cows with chronic or
intermittent diarrhoea”, or “cows with milk production significantly below herd-mates in the same parity
and stage of lactation”. This cow is affected by a disease causing entity, whether this is infectious or
not. Our purpose is to detect MAP infection in clinically-affected animals, and the target condition is
therefore the “MAP affected animal”. Notice that “MAP affected” and “clinical diseased” may differ, in
that “clinical” may imply that we have observed that the cow is diseased, but MAP affected is really an
underlying condition. However, observational skills may be imperfect and therefore we are not capable
of determining if the animal is in fact diseased or not.

8.2.2 Pathogenesis

MAP infection may lead to diarrhoea, weight loss and decreased milk production. However,
multiple conditions may result in these effects. The description of the target condition is therefore of
paramount importance: Are all effects relevant, or just one? If MAP are causing this condition, which
analytes are likely to be correlated to disease progression? Production of IgG1 antibodies and
excretion of large amounts of MAP have both been associated with late-stage infection. However, both
may also be detectable at earlier stages. Furthermore, in the end-stage of infection, the animal may
become anergic with no detectable antibodies. However, antibody and agent detection tests may still
be highly relevant, and logistics and costs may be the next items to consider in the choice of test. We
would have to go back and refine our purpose to determine, if we want to confirm the diagnosis most
accurately, and therefore would have to compare the two tests, or if we want the earliest diagnosis to
be able to take action on the clinical suspect.

8.2.3. Case definition

A case definition for MAP associated clinical disease could be straightforward: 1) chronic
diarrhoea for > 10 days, or 2) a deviation in the milk production of more than 20% of herd-mates in the
same lactation and stage of lactation, or 3) weight loss, should be present along with detection of MAP
in faeces. While we would most often expect both excretion of detectable amounts of MAP and
occurrence of antibodies in the clinical suspect, both definitions could be appropriate, although we
could risk that there are no MAP in the specific part of the faeces collected, or the animal is in the
anergic phase without production of antibodies. Notice that for example the cow with diarrhoea > 10
days but with no detection of MAP in faeces is a non-case and should therefore be included in the
specificity estimation.

8.2.4. Bias from host responses

Detection of the target condition MAP affected is probably the MAP condition least affected by
chronicity, because the occurrence of IgG1 antibodies as well as detection of bacterial shedding in
faeces is well correlated with the occurrence of clinical disease. However, antibodies to MAP can
occur prior to the animal’s detectable shedding of MAP and occurrence of clinical signs (Nielsen,
2008; Nielsen et al., 2009). Therefore, “false-positive” test results prior to occurrence of the target
condition MAP affected are highly likely to occur, simply because the target condition has yet to occur.
False-positive reactions to any immuno-diagnostic test may also occur if antigens in the immuno-
assay bind to cross-reacting antibodies from related organisms.
8.2.5 Bias from study design

Defining a MAP clinical suspect may seem trivial, but because the three clinical signs “milk production deviation”, “chronic diarrhoea” and “weight loss” are not specific to MAP infection, we will likely want to include another parameter such as detection of MAP in faeces in the case definition. In this situation we cannot evaluate MAP detection tests without introducing incorporation bias.

8.2.6 Sampling

Diagnostic laboratories receiving samples to confirm the diagnosis may be an excellent choice to use as a study population. However, it should be noted that these animals may not all meet the case definition. If only samples from cows with clinical diarrhoea are submitted to the laboratory, cows with deviating milk yield are not part of the case definition, thereby introducing spectrum bias. Therefore, either the purpose and target condition should be changed, or a different study population is needed. Hence, at this stage we should reconsider the target condition. Reduced milk production is a non-specific indicator of a health problem and in some systems this is not really a measureable condition except when the monthly milk testing is carried out. Thus, we should consider prior to the study if it will be possible to define in practice, or if the case definition should only contain the intermittent weight loss without loss of appetite and diarrhoea.

8.2.7 Study design

A case-control study can be used and will likely be the most cost-effective and will not necessarily include bias. The most important consideration is the case definition, and controls should include all animals not fulfilling the case definition to avoid spectrum bias.

9. Discussion

Inferences about the performance characteristics of diagnostic test results is the ultimate goal of any DTE study, and reliable estimates of sensitivity and specificity are pivotal to reduce misinterpretation attributable to misclassification bias. The structured approach described herein is intended to aid in the planning of a DTE so that minimally biased estimates of sensitivity and specificity for different target conditions may be obtained.
First, a purpose and a target condition must be defined, and subsequently a case definition must be established to classify test subjects according to the target condition with minimal errors. The most appropriate and cost-effective study design depends on the relevant target condition, the possibilities for creating a useful case definition and the logistical requirements for carrying out the study. A single study design cannot be deemed the only one for a specific purpose, but clearly some designs are more likely than others to provide biased estimates.

We have discussed two examples, in which some choices have been made. The two chosen target conditions "MAP infected" and "MAP affected", and their corresponding target populations "the general dairy population" and "cattle with enteric diseases" reflect the situations about which we can make inferences about results when the index test we evaluate is to be used in practice. Use of the results from the test evaluation is only applicable to the target condition and target population for which they have been evaluated. Extrapolation to other target conditions is generally unwarranted, but extrapolation to other similar target populations may in some instances be reasonable, although tests should usually be evaluated in the target population. The tests cannot be used for the ‘other’ purpose, but may easily be transferred to other similar populations.

Specifying case definitions is also challenging, and it can be difficult to justify some of the choices. Why should 10 tissues be selected rather than 12 tissues? Can 20% deviation in milk production be measured under practical circumstances, and are 15 or 20% more appropriate cut-values? This can be a very practical issue, and the best advice would probably be to replicate definitions used by others to facilitate among-study comparisons. The choices made throughout the process affect the subsequent steps, and can introduce bias. However, the effects imposed by the choices must be considered to differentiate between target conditions, between case definition etc. and therefore it is not possible to provide prescriptive guidelines, which covers all possible scenarios and diseases. We have refrained from providing an “ideal study design”, because we believe that this study design may not really exist, or is practically and logistically impossible to achieve. However, this does not mean that all designs are useful.

Equally important to the design of the test-evaluation is the reporting (Gardner, 2010). All elements discussed above should be mentioned. Detailed guidelines for reporting are proposed in Gardner et al. (2011).
The approach described here is given on a starting point where decision makers present a problem, for which a diagnostic test should be developed and evaluated. An alternative situation could occur where a new test has been developed, and the goal is to determine for which purpose(s) the new test can be used. In principle, the same considerations apply and only the starting point in Fig. 1 is different. However, because Fig. 1 is a triad, the starting point does not impact the reflection steps required.

Throughout the DTE, it is important to continuously consider the purpose and target condition, the pathogenesis and case definition. They are closely interrelated and refining them may be required throughout the process to design the ideal DTE for the relevant situation. Ultimately, the proper use and interpretation of the results of the DTE defines the success of the DTE. If the goal is to compare sensitivities and specificities among tests (and determine which tests are superior in one or the and not to establish absolutes for each, then the effect of some biases may be reduced. This would necessitate a head-to-head comparison of multiple tests on the same set of samples.

References


<table>
<thead>
<tr>
<th>Study design</th>
<th>Description</th>
<th>Target conditions</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-sectional study with detailed pathological examination and tissue culture</td>
<td>Complete random sampling where enrolled animals represent the age-distribution in target population. Ideally, samples for the index test should be obtained simultaneously with collection of tissues for gross and histopathological examination, and culture or PCR.</td>
<td>All, but ideal for MAP infected</td>
<td>Pathological examination and culture/PCR of sufficient tissues is highly sensitive and specific</td>
<td>Costly and often not logistically feasible</td>
</tr>
<tr>
<td>Longitudinal study with follow-up using reference test</td>
<td>A random sample of animals should be followed for at least 4 yr, in herds with different MAP prevalences. Animals are tested frequently with a reference test such as culture of faecal samples and samples for the index test are obtained simultaneously. Animals should be followed until they are culled or to the end of study. At culling, additional tissue samples may be obtained for detection of MAP by culture or PCR.</td>
<td>All</td>
<td>Follow-up period covers typical lifetime experience in infected herds</td>
<td>Loss to follow-up bias; may be costly. Logistically challenging with long study periods</td>
</tr>
<tr>
<td>Cross-sectional design with complete verification: latent class analysis (LCA) of data</td>
<td>Data are collected using a cross-sectional design. At least two index tests with different targets (e.g. detection of agent and antibody responses) are used. Paired samples should be obtained from the same animal either simultaneously or within a short time interval. Use of LCA requires that certain assumptions be met for</td>
<td>All, but specifically relevant to targets of index test</td>
<td>Reference test need not be perfect; no assumptions necessary about its accuracy</td>
<td>Cross-sectional with complete verification more costly than designs with partial verification. For</td>
</tr>
</tbody>
</table>
inferences to be valid (Enøe et al., 2000). Partial verification approaches are possible but require a Bayesian LCA model (Su et al., 2004).

<table>
<thead>
<tr>
<th>Hierarchical and hybrid designs</th>
<th>Hierarchical designs take advantage of pre-existing data about herd MAP infection status (e.g. non-infected) to define animal status. Sometimes hybrid studies are used which combine data from non-infected herds and data obtained from cross-sectional studies with complete verification in infected herds. LCA is used for inferences</th>
<th>Non-infected in a hierarchical design; all targets possible in a hybrid design</th>
<th>Efficient to estimate specificity; in hybrid designs, specificity data from non-infected herds may reduce uncertainty in inferences from LCA models</th>
<th>Might be difficult to find non-infected herds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-control</td>
<td>Random sampling of animals that are categorized as having the target condition of interest (cases) or not (controls) is used. The case-control study might be implemented nested within a longitudinal design. Alternately, a study might be implemented retrospectively using biologic samples that are stored in a repository.</td>
<td>MAP infectious; MAP affected; MAP in end-products (beef, milk etc.)</td>
<td>Easy; Low cost</td>
<td>Multiple biases (verification; spectrum) are likely especially in a retrospective design</td>
</tr>
</tbody>
</table>
Fig. 1. Triads in a structured approach to a DTE for chronic infectious animal disease. The upper triad is the "application triad", i.e. the context in which the test is applied should be used should be considered here, whereas the lower triad is the "evaluation triad", where the actual DTE is created.
Fig. 2. Schematic presentation of hypothetical pathogenesis of *Mycobacterium avium* subsp. *paratuberculosis* (MAP) infection with resulting effects, examples of target conditions and case definitions.

CFU=colony forming units; Ct = cycle threshold; qPCR = quantitative real-time PCR