Rev. Sci. Tech. Off. Int. Epiz., 2021, **40** (1), 271-286

Bayesian latent class analysis when the reference test is imperfect

A. Cheung [1, 2], S. Dufour [3], G. Jones [4], P. Kostoulas [5], M.A. Stevenson [1, 2], N.B. Singanallur [2, 6] & S.M. Firestone [1, 2]*

(1) School of Veterinary Science, Faculty of Veterinary and Agricultural Sciences, University of Melbourne, 142 Royal Parade, Parkville, Victoria 3010, Australia
(2) OIE Collaborating Centre for Diagnostic Test Validation in the Asia-Pacific Region, CSIRO, 5 Portarlington Road, East Geelong, Victoria 3219, Australia
(3) Faculty of Veterinary Medicine, University of Montreal, 3200, rue Sicotte, Saint-Hyacinthe, Quebec J2S 2M2, Canada
(4) School of Fundamental Sciences, Massey University, PN461 Private Bag 11222, Palmerston North 4442, New Zealand
(5) School of Health Sciences, University of Thessaly, Trikalon 224, Karditsa 43100, Greece
(6) Australian Centre for Disease Preparedness, CSIRO Health and Biosecurity, 5 Portarlington Road, East Geelong, Victoria 3219, Australia
*Corresponding author: simon.firestone@unimelb.edu.au

Summary
Latent class analysis (LCA) has allowed epidemiologists to overcome the practical constraints faced by traditional diagnostic test evaluation methods, which require both a gold standard diagnostic test and ample numbers of appropriate reference samples. Over the past four decades, LCA methods have expanded to allow epidemiologists to evaluate diagnostic tests and estimate true prevalence using imperfect tests over a variety of complex data structures and scenarios, including during the emergence of novel infectious diseases. The objective of this review is to provide an overview of recent developments in LCA methods, as well as a practical guide to applying Bayesian LCA (BLCA) to the evaluation of diagnostic tests. Before conducting a BLCA, the suitability of BLCA for the pathogen of interest, the availability of appropriate samples, the number of diagnostic tests, and the structure of the data should be carefully considered. While formulating the model, the model's structure and specification of informative priors will affect the likelihood that useful inferences can be drawn. With the growing need for advanced analytical methods to evaluate diagnostic tests for newly emerging diseases, LCA is a promising field of research for both the veterinary and medical disciplines.

Keywords
Bayesian latent class analysis – Diagnostic test evaluation – Gold standard – Imperfect test – Prevalence – Sensitivity – Specificity.

Introduction

In recent years, epidemiologists have increasingly applied the methods of latent class analysis (LCA) to evaluate diagnostic tests and estimate true prevalence with imperfect tests. Historically, the accuracy of a diagnostic test, as measured by the diagnostic sensitivity (DSe) and diagnostic specificity (DSp), was determined by comparison with a reference test, often referred to as a 'gold standard' (1, 2). However, in practice, a reliable reference test with known measures of test accuracy does not always exist. Furthermore, it is difficult to build sufficiently sized collections of specimens from individuals of known disease status that are representative of the individuals that will be tested under field conditions. In such cases, when only an imperfect reference test or

https://doi.org/10.20506/rst.40.1.3224

insufficient numbers of appropriate reference samples of known disease status are available, LCA can be used to draw inferences on test accuracy and disease prevalence (3, 4, 5, 6, 7). The original application of LCA (8) used maximum likelihood estimation. However, subsequently, Bayesian approaches (7) have become increasingly popular.

In a Bayesian LCA (BLCA) approach, an individual's true infection status is assumed to be unobserved, and hence 'latent'. Instead of individuals being explicitly classified as 'infected' or 'uninfected', each individual is assumed to have a probability of infection, given the combination of an observed diagnostic test outcome, knowledge of the accuracy of the tests used (e.g. DSe and DSp) and prior knowledge of disease prevalence in the populations of interest (7). In other words, prior information about the variables to be estimated (e.g. DSe or DSp) is combined with observed data (e.g. diagnostic test outcomes) to obtain a posterior distribution of the variable, which is the updated belief of the variable's true value (3). Thus, knowledge of the true disease status of the sampled individuals is not required to infer measures of test accuracy or prevalence in the populations of interest.

An overview of methods for estimating diagnostic test accuracy with an imperfect reference test was last published in 2005 by Branscum *et al.* (3). Since then, new statistical methods have been developed. In this review, the authors present an updated overview of the literature on the application of LCA methods when using an imperfect reference test, as well as a practitioner's overview of technical considerations when applying a BLCA approach, with reference to online tools and code repositories.

A short literature review on latent class analysis

To illustrate the development and adoption of LCA methodologies when an imperfect reference test is used, a search was conducted with the Web of Science search engine and PubMed database, using 'rwos' (9) and 'pubmedR' (10) with a query targeting the methodology (latent AND class) and intended usage (diagnos* OR test* OR prevalence OR incidence). This search query was applied to titles, abstracts, author keywords, and Keywords Plus® in the Web of Science search engine and to titles and abstracts in the PubMed database. Duplicates were removed from the results which were then analysed, using 'Bibliometrix' (11). All analyses were conducted in the statistical computer language R (12). Because of the inconsistent use of the term 'latent class' during the early period of this field of research, a few key references were added to the search results, based on a manual search of the bibliographies of key reviews.

Hui and Walter (8) proposed the first identifiable LCA model to determine test accuracy (e.g. DSe and DSp) and true prevalence using two imperfect tests applied to two populations with different disease prevalence. A model is considered identifiable when the data are sufficient to identify the unknown parameters, such as DSe, DSp, prevalence of disease in the studied populations, and any correlation between dependent tests. When a model is identifiable without the need for informative prior information, it can, in theory, be used within either a frequentist (8) or Bayesian statistical framework (7). Although Hui and Walter were the first to present an LCA model, the term 'latent class' was not used to describe this type of model until 1985 by Kaldor and Clayton (13).

A key methodological tool that expanded the usage of Bayesian statistics was the introduction of Gibbs sampling (14), which employed simulations to solve the complex integration procedures required for a full Bayesian analysis. The first 'mainstream' Bayesian statistical software package implementing Gibbs sampling, WinBUGS, was released in 1997 and allowed the user to specify the likelihood for the data and the prior distributions for each parameter in the model to be estimated (15). The mathematical form of the posteriors was then derived from the software. With WinBUGS, the Hui–Walter model could be estimated in a Bayesian framework without requiring advanced statistical or computer programming skills. Since the release of WinBUGS, other statistical software packages based on this approach (e.g. JAGS, OpenBUGS) were developed. More recently, integrated nested Laplace approximation (INLA) has been used as an alternative to Gibbs sampling, which can be computationally slow for highly complex models (16).

A key assumption of the Hui–Walter model is that the diagnostic tests under evaluation are independent and conditional on the true disease status of an individual. This means that, given the true status of an individual (infected versus uninfected), knowledge of whether test A yields a positive result provides no information on the likelihood of test B being positive as well. However, this assumption is quite difficult to justify in most diagnostic testing scenarios, especially for tests that detect similar biological mechanisms of change in the body as an indicator of the presence of disease, such as two tests detecting antibodies or two tests targeting nucleic acid sequences. In 2001, two groups proposed LCA models that relaxed the assumption of conditional independence of the diagnostic tests under evaluation by modelling conditional dependence between tests using covariance terms between the tests (17, 18). These more complex BLCA models allowed for more accurate and appropriately precise estimations of diagnostic test accuracy.

After the relaxation of the conditional independence assumption, further studies adapted LCA models to

expanded data structures. For instance, the original Hui–Walter model, which used two populations with distinct prevalence, was expanded to estimate disease prevalence and diagnostic test accuracy in more than two populations, each with distinct prevalence (6). Similarly, Branscum *et al.* (3) presented a model for up to three dichotomous diagnostic tests, including two conditionally dependent tests. Several papers also presented methods for receiver operating characteristic (ROC) curve estimation, and cut-off determination for tests with continuous rather than dichotomous outcomes (19, 20, 21, 22, 23, 24, 25).

The hierarchical model proposed by Hanson *et al.* (26) was another important development in LCA for expanded data structures, especially for animal health diagnostics. This paper proposed a class of LCA models for studies with several exchangeable populations (e.g. several herds), each with a different disease prevalence. With hierarchical LCA models, the cluster- or herd-level disease prevalence distribution was modelled, rather than the disease prevalence within each studied population. Instead of providing an overall prevalence for the whole population, researchers could determine which levels of the hierarchy were infected and, within those levels, the proportion of infected units (Fig. 1). This model was more computationally efficient for complex data structures and also provided useful insights from a disease control perspective.

After the development of the hierarchical model, new studies implemented LCA models with random effect terms to account for complex hierarchical data structures, such as hierarchically clustered samples (27, 28), and mixed models with fixed and random effects influencing prevalence (29) and even DSe (30).

Latent class analysis has also been used to estimate prevalence and diagnostic accuracy using pooled testing. In 1999, Johnson and Pearson (31) presented a Bayesian approach to estimate the diagnostic accuracy of pooled tests. This model was expanded in 2006, using a hierarchical framework to account for varying DSe and DSp when applying the test to individuals rather than pooled samples (32). Finally, Dhand *et al.* (33) presented a model that allowed for not only an imperfect pooled test, but also variations in pool sizes.

As the adoption of LCA continued to grow, the World Organisation for Animal Health (OIE) *Manual of Diagnostic Tests and Vaccines for Terrestrial Animals* endorsed the use of LCA models for evaluating the performance of a new diagnostic assay when the use of a gold standard test was not possible (34). As a result, many publications have implemented BLCA to evaluate imperfect diagnostic tests for emerging and re-emerging infectious diseases, including COVID-19 (35, 36, 37), foot and mouth disease (38, 39), African swine fever (40) and highly pathogenic avian influenza (41). To provide guidance for researchers

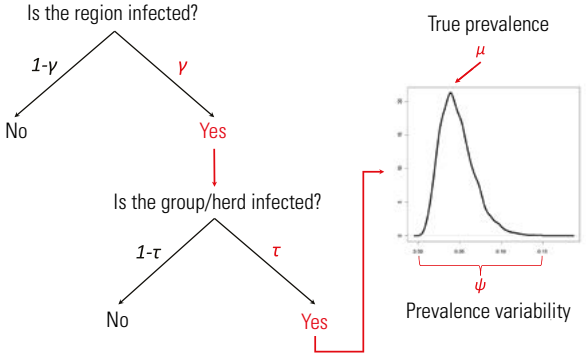


Fig. 1
A hierarchical approach to prevalence estimation, with the probability that each level of hierarchy is infected and an animal-level prevalence (with mean μ and variability ψ given only for infected herds)

implementing these methods, Kostoulas *et al.* (42) proposed a set of guidelines on reporting test accuracy when using BLCA.

These developments have promoted the uptake of LCA methodologies for evaluating diagnostic tests or estimating disease prevalence across various disciplines. A few, mainly methodological, articles were produced annually between 1980 and 2000 (Fig. 2). The annual publication of articles has increased since 2000, with more than 270 papers published each year since 2017. At the time of analysis, articles on LCA were published in a total of 1,914 different journals, with the largest number of publications in *PLOS One* (n = 147), the *Journal of Affective Disorders* (n = 77), and *Preventive Veterinary Medicine* (n = 76) (Fig. 3). Among the 25 journals with the most publications on LCA of diagnostic tests, those appearing most often were journals in the fields of psychiatry, psychology or behavioural science (n = 16 journals), followed by general medical, epidemiological and microbiology journals (n = 6), and, finally, by biostatistical (n = 2) and veterinary journals (n = 1). Biostatistical and veterinary journals were, however, early adopters of the LCA methodology, with substantial numbers of publications in biostatistical journals in the 1990s, and in veterinary journals from 2000 to 2005 (Fig. 3). Despite the common need across disciplines to estimate diagnostic test accuracy and disease prevalence without a gold standard test, this clear differential uptake over time between the medical and veterinary fields may be due to early affiliations between the biostatisticians who developed LCA and veterinary epidemiologists in the application of LCA for diagnostic purposes (1, 5, 6).

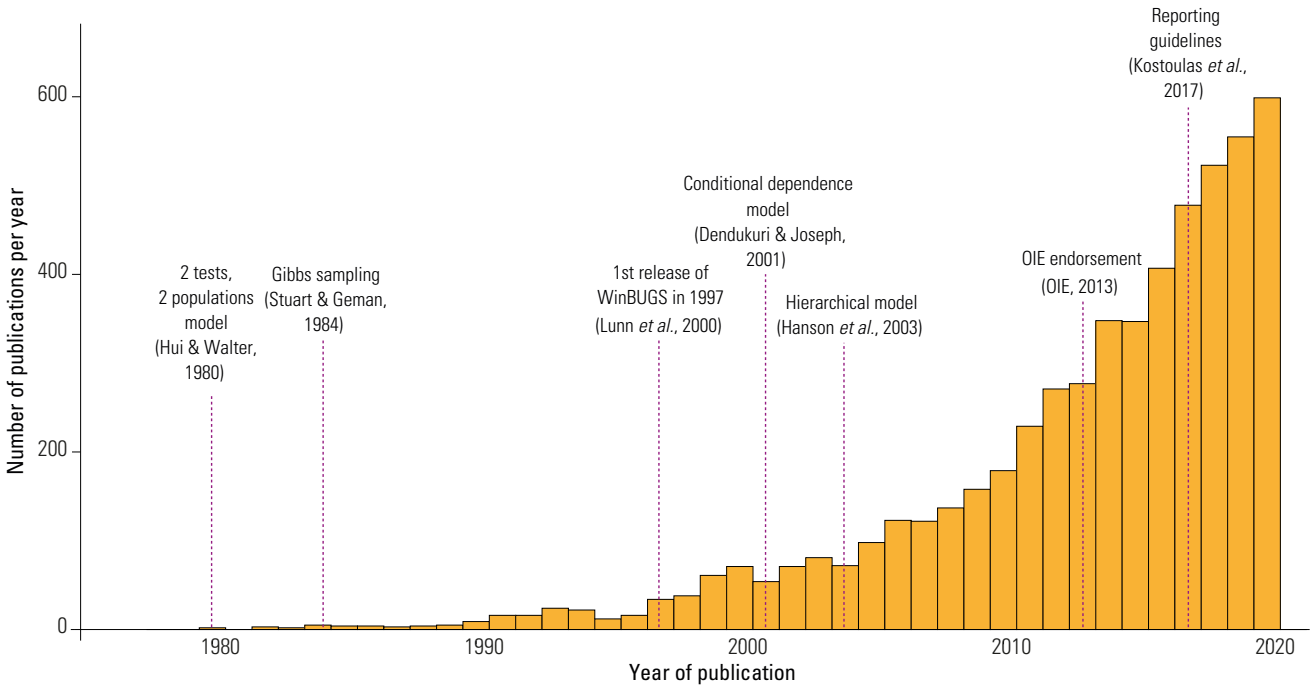


Fig. 2

Frequency histogram of the number of peer-reviewed articles published on latent class analysis when there is an imperfect reference test

These data were based on a Web of Science and PubMed search targeting articles published since 1 January 1980 (Web of Science and PubMed search conducted on 15 January 2021). A total of 5,362 articles was obtained. Select key events and landmark articles for this field of research are indicated, with details given in (8), (14), (15), (17), (26), (34) and (42)

The application of Bayesian latent class analysis when the reference test is imperfect

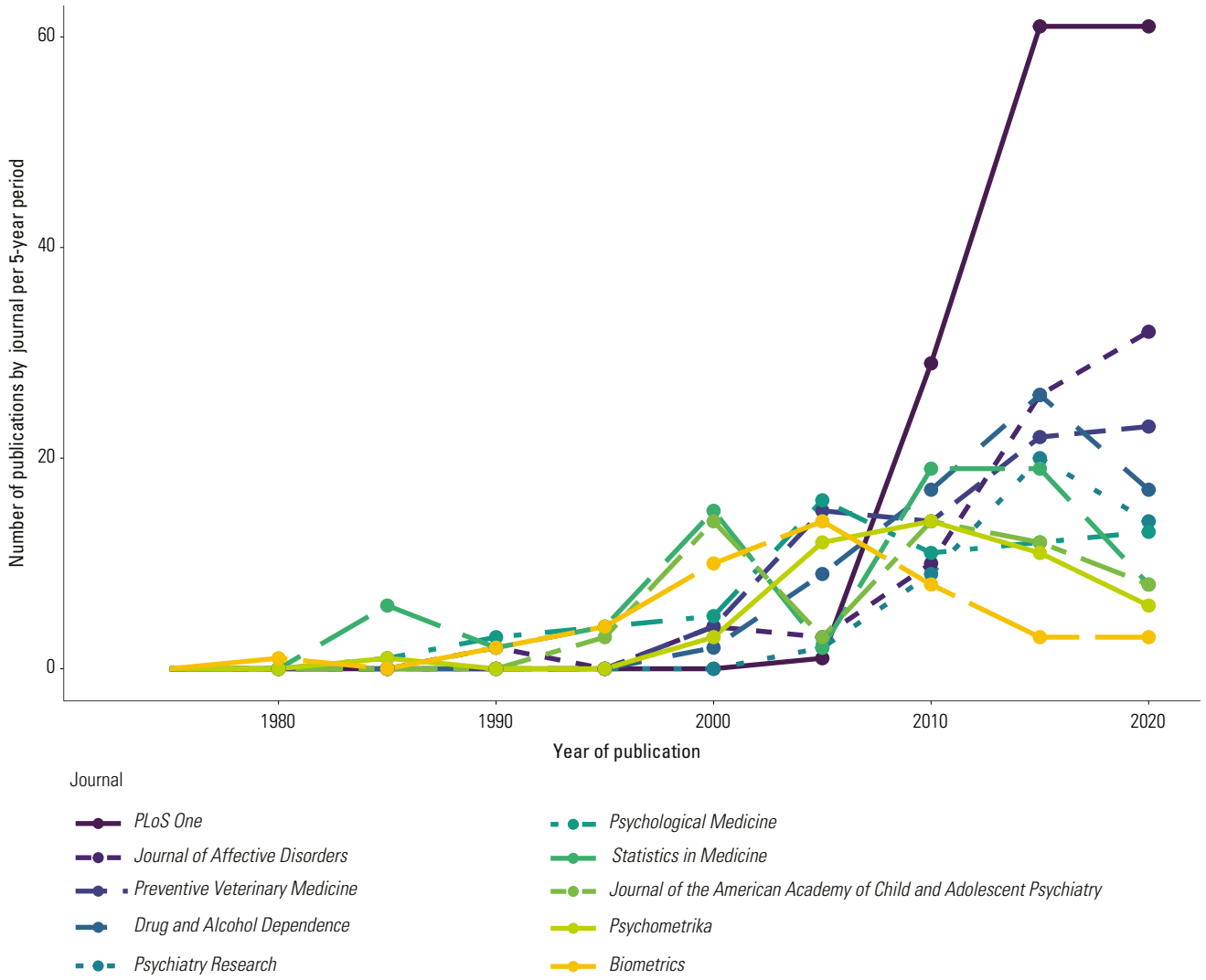
In this section, the key technical considerations in a BLCA are examined, with reference to recently published studies. The aim of this section is not to replicate the statistical background, as this is well covered elsewhere (3, 43), or to advise on reporting and appraisal. For this, readers should refer to the STARD–BLCM (42). Rather, the aim is to present a practitioner's guide to conducting a BLCA with referenced examples. The interested reader is encouraged to refer to tools and tutorials in the 'Supplementary materials' (S1 and S2). Key considerations when planning a BLCA with an imperfect reference test are detailed in the following sections.

Suitability of Bayesian latent class analysis

Certain scenarios are more well suited for BLCA, since the designation of true disease status, as measured by the latent variable, depends on the nature of the tests under evaluation, the timing of shedding of the disease agent, and the qualities of the host immune response. For instance, in studies comparing polymerase chain reaction (PCR) assays with antibody assays, the PCR assay detects nucleic acid, a marker of pathogen shedding, whereas the antibody assays detect immunological responses to pathogen exposure. Therefore, the 'true disease status' defined and inferred by the model corresponds to individuals whose infections have progressed enough for shedding of the agent to be detected by PCR, while also possessing detectable levels of antibodies (44).

Furthermore, diseases characterised by a relatively short or intermittent shedding period may be difficult to detect by PCR, leading to problems in establishing a consistent latent variable definition and a narrowing of the infection window detected under the LCA model (45). Finally, because certain antibodies, particularly immunoglobulin G (IgG), can persist for long periods of time during convalescence, a positive result from an antibody assay does not always reflect 'true infection status' (46).

Latent class analysis models are ideal for infections that either involve long periods of shedding or that elicit immune responses that occur over most of the infection period. On the other hand, with acute infections, selecting the tests to be evaluated and subsequent definition of the latent variable should be undertaken cautiously and



* *PLoS One* was launched in 2006

Fig. 3
Line plot of the number of published articles on latent class analysis when there is an imperfect reference test, by journal, per five-year period
List of articles obtained by a Web of Science and PubMed search, targeting articles published since 1 January 1980 (Web of Science and PubMed search conducted on 15 January 2021)

with appropriate consideration of the purpose of the test, the disease and the sampling context (42). Referencing a diagnostic window figure that compares the variation over time in detection of the agent across different diagnostic tests, as shown in Figure 1 of King *et al.* (47), in Figure 6 of Marmion (48), and in Sethuraman *et al.* (49), may help to determine whether BLCA would suit the diagnostic tests and disease of interest.

Objectives of the analysis

The most common objectives of studies involving BLCA with an imperfect reference test are to evaluate the diagnostic performance of one or more tests (50, 51) or to estimate prevalence in defined populations (52, 53). Such studies rely on a well-characterised typical model structure and a relatively standard data set of samples from individuals in populations tested with one or more diagnostic tests. Pooled samples require a more complex data structure and model. Further objectives may include the evaluation of risk factors for disease (54), estimating attributable fractions (55, 56), and demonstrating freedom from infection in a population (57). The unit of interest may also be the herd or flock rather than the individual animal, especially when using aggregate samples such as bulk tank milk samples, pooled tissue or swab samples (33), or when data are only available in an aggregated form (58).

Availability of appropriate samples and sample size

When evaluating whether tests are fit for purpose, a key consideration is that the sampled animals are representative of those for which the test will be used. Ideally, samples should be collected specifically to meet the study's objectives and derived from populations in which some prior information on disease prevalence is available. If sample sizes are limited, such as during the emergence of a new infectious disease, repeated testing of the same animals may be required. The data may also include other forms of hierarchical clustering, such as multiple animals being sampled from the same farms (59), which will influence model selection, as described in 'Organisational structure of the data and clustering', below. Studies based on convenience samples from animals that do not typify the range of disease states naturally occurring in the target population, such as experimentally infected animals or samples only from animals with severe disease, should be considered cautiously. Further recommendations are available on appropriate sample collection and handling (60).

Sample size considerations for BLCA are similar to those for studies that estimate population proportions (e.g. prevalence) and include:

– the expected proportion(s) (e.g. the estimated true population prevalence)

– the desired degree of precision of the parameters to be estimated (e.g. ±1%, 2% or 5%)

– the level of confidence (typically 95% or 99%)

– any complexities of the study design, such as clustering.

A table has been provided by the OIE to simplify these considerations (2). For studies that intend to estimate both DSe and DSp, the minimum sample size should be calculated separately for each and summed. Numerous online tools are available that provide such estimates for a full range of inputs (https://shiny.vet.unimelb.edu.au/epi/sample.size/). These tools can also be replicated in R using the functions *epi.sssimpleestb* and *epi.ssclus1estb* from the package 'epiR' (61). Bayesian approaches have also been developed to calculate the sample sizes necessary for studies that estimate diagnostic test accuracy (62).

Number of tests to be used

The number of tests to be used depends on the availability and costs of suitable tests and whether the model is identifiable (covered in 'Model structure and selection', below). Where practicable, tests with repeatable, robust, manufacturer-specified protocols should be used. In some instances, previously validated tests may be used alongside tests that are being evaluated in the target population for the first time. In this case, peer-reviewed published information on the specifications of the previously validated tests can inform prior specification. In particular, the prior(s) relating to disease prevalence in the population(s) can be informed by this peer-reviewed information (see 'Prior elicitation and parameterisation', below).

Organisational structure of the data and clustering

Under a typical model structure, when k (tests) are evaluated for p (populations), the data required are 2^k tables per population (e.g. if k = 3 tests are evaluated, a 2×2×2 table is required). The choice and number of populations are informed by considerations of model identifiability, the availability of appropriate prior information and the feasibility of data collection.

When the unit of interest is a herd or other collective entity, such as that described by Su *et al.* (63), data should be in an aggregated 'long format', with one row per herd comprising n tests per herd, 2^k table components in additional columns, and any further herd-level covariates under consideration. Sample code is provided in the supplementary materials of Wood *et al.* (64). This code also incorporates random effects to account for clustering by region. A long-format data structure is also used in models with repeated measures, with one row per observation, with a unique identifier per individual or other unit of interest to fit a random effect.

Model structure and selection

Typically, for a BUGS family statistical package, such as WinBUGS or OpenBUGS, the analysis code includes specification of the following:

- the model form for the data (often multinomial or binomial)

- Hui and Walter (8) equations for the relationship between the latent class and the observed data (apparent prevalence $k \times k$ tables for each population under consideration)

- prior probability distributions, which may incorporate further complexity, such as mixture distributions in the prior specification of true prevalence for certain populations, to allow for disease freedom (i.e. zero prevalence) (65) or a logit function for the true prevalence, which allows for the inclusion of fixed and random effects (45, 64)

- any outputs calculated based on inferred distributions, such as Youden's index, likelihood and odds ratios for any risk factors considered as fixed effects, and predicted probabilities, such as the prevalence within a randomly selected herd or the probability of disease freedom at a specified prevalence (see sample code in Wood *et al.*) (64)

- the data

- a call to execute the BUGS program with control parameters, such as the number of chains, the number of

Markov chain Monte Carlo (MCMC) iterations per chain, parameters to monitor, and initialising values for each unobserved parameter for each MCMC chain

- model diagnostics to check that the model specification, model fit, chain length, convergence, and posterior probability distributions are appropriate to draw inferences from the model.

Several options are available for formulating the BLCA model. The Dendukuri–Joseph (DJ) model is a popular and highly generalisable model structure that accounts for conditional dependence between multiple diagnostic tests, with constraints proposed on covariance terms based on the inferred test sensitivities and specificities (see 'Supplementary materials, S1') (17). A key assumption of the DJ model structure is that the DSe and DSp are constant across the populations under consideration. Examples of how this assumption has been addressed are provided by Kostoulas *et al.* (42).

The DJ model has since been extended to include three or more tests (66, 67, 68). However, a 'fully saturated model', with every covariance term between every test, is rarely justified or feasible. Given diminishing interpretability with higher-order covariance terms, a more parsimonious approach is to consider only pair-wise covariance terms between tests. Thus, a natural starting point for model selection is a 'pair-wise saturated model', followed by a comparison of the model's fit and support for inclusion of each covariance term based on the deviance information criterion (69) and the marginal posterior estimates of each term.

Model identifiability

A model is considered identifiable if it is possible to infer the true values of all of the unknown parameters after obtaining a sufficiently large number of observations from the model. For a BLCA model to be identifiable, the number of degrees of freedom (*df*) provided by the data must be more than or equal to the number of model parameters (*np*). This condition does not guarantee model identifiability, as identifiability also depends on the algebraic structure of the model (43).

If the model is not identifiable, the inclusion of informative priors can enable useful inferences to be drawn. The *df* and *np*, and thus model identifiability and ability to draw useful inferences, depend on the number of tests (*k*) and populations (*p*) under consideration (Table I). Increasing the number of tests or populations may not create any additional advantages in terms of the identifiability of the model but may improve precision around estimates of test accuracy.

Table I

Study design parameters that are likely to result in an identifiable model

For a Bayesian latent class analysis model to be identifiable, the number of degrees of freedom (*df*) provided by the data must be more than or equal to the number of model parameters (*np*). Although this condition does not guarantee model identifiability, the inclusion of informative priors can allow for useful inferences to be drawn even if the model is not identifiable. As the number of tests (*k*) and populations (*p*) increase, fewer informative priors are likely to be required for an identifiable model

Total number of diagnostic tests (*k*)	Number of conditionally dependent diagnostic tests (*d*)	Number of populations (*p*)	Number of unknown parameters (*np*)*	Number of degrees of freedom (*df*)	Number of informative priors needed**
1	0	1	3	1	2
1	0	2	4	2	2
1	0	3	5	3	2
2	0	1	5	3	2
2	0	2	6	6	0
2	2	2	8	6	2
2	0	3	7	9	0
3	0	1	7	7	0
3	0	2	8	14	0
3	2	2	10	14	0
3	3	2	14	14	0
3	0	3	9	21	0
k	0	*p*	$2k + p$	$(2^k - 1)p$	If $df < np$, then $np - df$
k	*d*	*p*	$2k + p + d(d - 1)$	$(2^k - 1)p$	If $df < np$, then $np - df$

* For conditionally dependent tests, only pair-wise covariance terms are considered in this table

** Minimum number of informative priors needed for useful inference

Study designs that are highly likely to provide identifiable models, given sufficient data, include the following:

- two independent tests considered for two populations, each with distinct prevalence. The data for each population's test results can be represented as a 2 × 2 table, each with 3 *df* (total *df* = 6 across both populations). There are 6 *np* in total (two sensitivities, two specificities and two true prevalences, ϖ_1 and ϖ_2);

- three independent tests in one population. The results comprise a 2 × 2 × 2 table, so *df* = 7 and *np* = 7 (three sensitivities, three specificities and one true prevalence, ϖ).

Study designs that are prone to non-identifiable models, and thus require careful specification of informative priors to enable useful inferences to be drawn, include the following:

- one test implemented in *p* populations, where *df* = *p* and *np* = *p* + 2, given that the sensitivity and specificity of the test is to be inferred, along with the true prevalence in each population (ϖ_p);

- two dependent tests applied to three populations. The results comprise three 2 × 2 tables, so *df* = 9 and *np* = 9 (two sensitivities, two specificities, two covariance terms and five prevalences). Although, in this case, *df* = *np*, the model is not identifiable due to the algebraic structure of the model (43).

In a frequentist analysis, an 'identifiable' model will yield two solutions: if one solution is (Se, Sp, ϖ), the other is (1 – Sp, 1 – Se, 1 – ϖ), corresponding to an interchange of the latent classes (infected versus non-infected). In practice, a test is expected to have Se + Sp > 1, thereby identifying which of the two solutions is appropriate.

For non-identifiable models, a frequentist analysis is not possible and would require that the values of some parameters be fixed (e.g. assume a perfect DSe or DSp) to obtain an identifiable model. However, a BLCA with informative priors for at least some of the parameters will allow inferences to be made, even in a non-identifiable model. In this case, since the posteriors depend heavily on the specified priors, a wise choice of priors is imperative, and a thorough sensitivity analysis under various combinations of prior specifications should be performed and presented. Care must also be taken to ensure convergence, since this is typically much slower to be achieved than in identifiable models (see 'Prior elicitation and parameterisation', below, and 'Supplementary materials, S1').

Prior elicitation and parameterisation

An essential part of BLCA is the specification of informative priors, which should be based on expert opinion or the relevant published literature. Priors consist of probability distributions around a specified value that describe one's belief about a variable of interest. When eliciting a prior from an expert, key information to gather includes the most likely value (mode) of the parameter and the probability (typically 95%) that the parameter is greater than or less than another specified value. The source of information from prior elicitation, though relevant to the data at hand, must also be independent of the data.

A further consideration in prior specification is the strength of the priors relative to the data. If highly precise priors are specified, or only a small sample size is available, then the posterior distribution will be mostly informed by the priors rather than the data. For this reason, if the priors provided by experts are too narrow, they can be widened before their inclusion in the model. 'Non-informative' priors, which can also be referred to as 'reference' priors, include flat priors, vague priors or diffuse priors (70). Flat priors are uniform distributions with equal probability density across the range of supported values.

When model identifiability limits the use of flat priors, vague priors centred on the modes of previous prior specifications can be considered with at least double the level of uncertainty (i.e. half the precision) in the prior specification. However, informative priors are preferred because a BLCA with informative priors for at least some of the parameters will allow inferences to be made, even in a non-identifiable model. As a result of the varying strengths of possible priors relative to the data, the posterior distribution of each inferred parameter should always be compared to its prior (see 'Supplementary materials, S1'). A sensitivity analysis of the influence of using different priors is recommended (42).

Several software packages assist in the derivation of prior distributions from existing scientific information or elicitation from experts. The R package 'PriorGen' translates beliefs into prior information formulated as beta and gamma distributions, based on the disease prevalence and the DSe and DSp of the diagnostic tests, and can even be applied to hierarchical models (71). An online implementation of the original BetaBuster (70) tool to parameterise beta distributions with elicited values based on these inputs is also available (https://shiny.vet.unimelb.edu.au/epi/betabuster/), with the corresponding function *epi.betabuster* in the R package 'epiR' (61).

Markov chain Monte Carlo convergence diagnostics and inference

When a BLCA is fitted using MCMC, multiple chains should be run, initiated from distinct and independent plausible values to assess convergence (73). Any inference on the joint posterior distribution should be made only on sections of the chains after convergence is achieved. Convergence can be assessed by using the Gelman and Rubin statistic (74) and by visual inspection of the chains (42). Chain diagnostics such as autocorrelation plots and density distributions for each parameter can be checked for

unimodality (see 'Supplementary materials, S1'), using the R packages 'coda' (75) and 'mcmcplots' (76).

Such chain diagnostics inform how much of the start of each chain should be discarded, which is termed 'burn-in', and whether the number of iterations is sufficient to achieve reasonable, effective sample sizes (ESSs) for each inferred parameter. If there is considerable autocorrelation for certain parameters, then chains may need to run for longer to achieve an appropriate ESS. After the burn-in is discarded, converged chains can be combined and final estimates presented, based on medians and other quantiles of the joint posterior distribution (i.e. the 95% posterior interval or PI). The R package 'HDInterval' enables straightforward estimation of the highest (posterior) density intervals (77).

Concluding recommendations

The purpose of this review was to present an updated overview of the development of LCA methods over the past few decades and to provide a brief guide on how BLCA methods can be applied to diagnostic test evaluation. Although LCA methods were first introduced by Hui and Walter (8), the sheer volume of studies published over the past few years indicates that LCA is a rapidly evolving field of study. Areas of future study in LCA methods include the implementation of LCA within increasingly complex data structures, such as hierarchical structures with repeated measures.

In conclusion, the authors recommend the implementation of LCA to evaluate diagnostic tests either when there is no gold standard reference test or there are not sufficient numbers of appropriate reference samples of known disease status. Taking into account the key considerations detailed in this review, epidemiologists can use imperfect tests not only to draw inferences on test accuracy, but also to estimate the true prevalence of a disease. With the growing need for advanced analytical methods to investigate newly emerging diseases, LCA provides a promising avenue of research in both the veterinary and medical fields. Readers interested in the use of BLCA in animal health can refer to a recent publication by Johnson *et al.* (78), which provides a sequential analysis of the swine toxoplasmosis data referred to in 'Supplementary materials, S1'.

Acknowledgements

The authors would like to thank Professor Cord Heuer of the Epicentre at Massey University, New Zealand, and Professor Ian Gardner of the University of Prince Edward Island, Canada, for their insightful feedback on this article.

Analyse bayésienne à classes latentes dans les situations où le test de référence est imparfait

A. Cheung, S. Dufour, G. Jones, P. Kostoulas, M.A. Stevenson, N.B. Singanallur & S.M. Firestone

Résumé

L'analyse à classes latentes a permis aux épidémiologistes de surmonter les problèmes concrets posés par les méthodes traditionnelles d'évaluation des essais de diagnostic, qui nécessitent à la fois un test de référence absolue (étalon ou *gold standard*) et un grand nombre d'échantillons de référence aux caractéristiques appropriées. Au cours des quatre dernières décennies, les méthodes d'analyse à classes latentes ont acquis de l'ampleur et permettent aux épidémiologistes d'évaluer les essais diagnostiques et d'estimer les taux de prévalence réelle tout en recourant à des tests supposés imparfaits utilisant des structures de données et des scénarios complexes, y compris dans les situations d'émergence de nouvelles maladies infectieuses. Les auteurs font un tour d'horizon des dernières évolutions dans ce domaine et donnent des orientations pratiques concernant la manière d'utiliser l'analyse bayésienne à classes latentes pour évaluer les performances d'un test de diagnostic. Avant de conduire une telle analyse, il convient de déterminer avec soin si elle est adaptée à l'agent

pathogène considéré et si les échantillons disponibles sont appropriés et en nombre suffisant ; il convient également de prendre en compte le nombre de tests de diagnostic à évaluer et la structure des données utilisées. Lors de la conception du modèle, sa structure et la définition préalable des données informatives vont affecter la probabilité que le modèle génère des inférences utiles. Face à la nécessité croissante de disposer de méthodes analytiques sophistiquées pour évaluer les tests de diagnostic utilisés pour les maladies émergentes nouvelles, les analyses à classes latentes offrent des perspectives prometteuses pour la recherche, aussi bien dans le domaine de la santé vétérinaire que de la médecine humaine.

Mots-clés
Analyse bayésienne à classes latentes – Évaluation des tests de diagnostic – Prévalence – Sensibilité – Spécificité – Test de référence absolue (*gold standard*) – Test imparfait.

Análisis bayesiano de clases latentes cuando la prueba de referencia es imperfecta

A. Cheung, S. Dufour, G. Jones, P. Kostoulas, M.A. Stevenson, N.B. Singanallur & S.M. Firestone

Resumen
El análisis de clases latentes ha servido a los epidemiólogos para superar las limitaciones prácticas que imponen los métodos tradicionales de evaluación de pruebas de diagnóstico, que requieren a la vez una prueba de diagnóstico que sirva de patrón de referencia perfecto y un gran número de muestras de referencia adecuadas. En los últimos cuatro decenios, los métodos de análisis de clases latentes se han ido ampliando hasta permitir a los epidemiólogos evaluar pruebas de diagnóstico y calcular la prevalencia real empleando pruebas imperfectas ante muy diversas estructuras de datos y situaciones complejas, incluida la aparición de nuevas enfermedades infecciosas. Los autores, tras presentar a grandes líneas los últimos adelantos en cuanto a métodos de análisis de clases latentes, ofrecen indicaciones prácticas para aplicar el análisis bayesiano de clases latentes a la evaluación de pruebas de diagnóstico. Antes de proceder a un análisis bayesiano de este tipo conviene estudiar con detenimiento la idoneidad del método para el patógeno en cuestión, la disponibilidad de muestras apropiadas, el número de pruebas de diagnóstico y la estructura de los datos. A la hora de formular el modelo, la estructura del propio modelo y la especificación de los elementos informativos previos influirán en la probabilidad de poder extraer conclusiones provechosas. Ante la creciente necesidad de disponer de métodos analíticos avanzados con los que evaluar pruebas de diagnóstico de nuevas enfermedades emergentes, el análisis de clases latentes abre un promisorio campo de investigación para las disciplinas veterinarias y médicas.

Palabras clave
Análisis bayesiano de clases latentes – Especificidad – Evaluación de pruebas de diagnóstico – Patrón de referencia perfecto – Prevalencia – Prueba imperfecta – Sensibilidad.

References

1. Greiner M. & Gardner I.A. (2000). – Epidemiologic issues in the validation of veterinary diagnostic tests. *Prev. Vet. Med.,* **45** (1), 3–22. doi:10.1016/S0167-5877(00)00114-8.

2. World Organisation for Animal Health (OIE) (2019). – Principles and methods of validation of diagnostic assays for infectious diseases. Chapter 1.1.6. *In* Manual of Diagnostic Tests and Vaccines for Terrestrial Animals, 8th Ed. OIE, Paris, France. Available at: www.oie.int/fileadmin/Home/eng/Health_standards/tahm/1.01.06_VALIDATION.pdf (accessed on 11 March 2021).

3. Branscum A.J., Gardner I.A. & Johnson W.O. (2005). – Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. *Prev. Vet. Med.,* **68** (2), 145–163. doi:10.1016/j.prevetmed.2004.12.005.

4. Collins J. & Huynh M. (2014). – Estimation of diagnostic test accuracy without full verification: a review of latent class methods. *Stat. Med.,* **33** (24), 4141–4169. doi:10.1002/sim.6218.

5. Enøe C., Georgiadis M.P. & Johnson W.O. (2000). – Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Prev. Vet. Med.,* **45** (1), 61–81. doi:10.1016/S0167-5877(00)00117-3.

6. Johnson W.O., Gastwirth J.L. & Pearson L.M. (2001). – Screening without a "gold standard": the Hui–Walter paradigm revisited. *Am J. Epidemiol.,* **153** (9), 921–924. doi:10.1093/aje/153.9.921.

7. Joseph L., Gyorkos T.W. & Coupal L. (1995). – Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *Am. J. Epidemiol.,* **141** (3), 263–272. doi:10.1093/oxfordjournals.aje.a117428.

8. Hui S.L. & Walter S.D. (1980). – Estimating the error rates of diagnostic tests. *Biometrics,* **36** (1), 167–171. doi:10.2307/2530508.

9. Barnier J. (2020). – Package 'rwos' [computer software]. R package version 0.0.1.

10. Aria M. (2020). – Package 'pubmedR' [computer software]. R package version 0.0.3.

11. Aria M. & Cuccurullo C. (2017). – Bibliometrix: an R-tool for comprehensive science mapping analysis. *J. Informetrics,* **11** (4), 959–975. doi:10.1016/j.joi.2017.08.007.

12. R Core Team (2019). – R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

13. Kaldor J. & Clayton D. (1985). – Latent class analysis in chronic disease epidemiology. *Stat. Med.,* **4** (3), 327–335. doi:10.1002/sim.4780040312.

14. Geman S. & Geman D. (1984). – Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transact. Pattern Anal. Machine Intell.,* **PAMI-6** (6), 721–741. doi:10.1109/TPAMI.1984.4767596.

15. Lunn D.J., Thomas A., Best N. & Spiegelhalter D. (2000). – WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Stat. Comput.,* **10** (4), 325–337. doi:10.1023/A:1008929526011.

16. Rue H., Martino S. & Chopin N. (2009). – Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. Roy. Stat. Soc., B (Stat. Methodol.),* **71** (2), 319–392. doi:10.1111/j.1467-9868.2008.00700.x.

17. Dendukuri N. & Joseph L. (2001). – Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics,* **57** (1), 158–167. doi:10.1111/j.0006-341X.2001.00158.x.

18. Georgiadis M.P., Johnson W.O., Gardner I.A. & Singh R. (2003). – Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. *J. Roy. Stat. Soc., C (Appl. Stat.),* **52** (1), 63–76. doi:10.1111/1467-9876.00389.

19. Branscum A.J., Johnson W.O., Hanson T.E. & Baron A.T. (2015). – Flexible regression models for ROC and risk analysis, with or without a gold standard. *Stat. Med.,* **34** (30), 3997–4015. doi:10.1002/sim.6610.

20. Branscum A.J., Johnson W.O., Hanson T.E. & Gardner I.A. (2008). – Bayesian semiparametric ROC curve estimation and disease diagnosis. *Stat. Med.,* **27** (13), 2474–2496. doi:10.1002/sim.3250.

21. Choi Y.-K., Johnson W.O., Collins M.T. & Gardner I.A. (2006). – Bayesian inferences for receiver operating characteristic curves in the absence of a gold standard. *J. Agric. Biol. Environ. Stat.,* **11** (2), 210. doi:10.1198/108571106X110883.

22. Erkanli A., Sung M., Costello E.J. & Angold A. (2006). – Bayesian semi-parametric ROC analysis. *Stat. Med.,* **25** (22), 3905–3928. doi:10.1002/sim.2496.

23. Jafarzadeh S.R., Johnson W.O., Utts J.M. & Gardner I.A. (2010). – Bayesian estimation of the receiver operating characteristic curve for a diagnostic test with a limit of detection in the absence of a gold standard. *Stat. Med.,* **29** (20), 2090–2106. doi:10.1002/sim.3975.

24. Müller B., Vounatsou P., Ngandolo B.N.R., Diguimbaye-Djaïbe C., Schiller I., Marg-Haufe B., Oesch B., Schelling E. & Zinsstag J. (2009). – Bayesian receiver operating characteristic estimation of multiple tests for diagnosis of bovine tuberculosis in Chadian cattle. *PLOS ONE,* **4** (12), e8215. doi:10.1371/journal.pone.0008215.

25. Wang C., Turnbull B.W., Gröhn Y.T. & Nielsen S.S. (2006). – Estimating receiver operating characteristic curves with covariates when there is no perfect reference test for diagnosis of Johne's disease. *J. Dairy Sci.,* **89** (8), 3038–3046. doi:10.3168/jds.S0022-0302(06)72577-2.

26. Hanson T., Johnson W.O. & Gardner I.A. (2003). – Hierarchical models for estimating herd prevalence and test accuracy in the absence of a gold standard. *J. Agric. Biol. Environ. Stat.,* **8** (2), 223. doi:10.1198/1085711031526.

27. Stringer L.A., Jones G., Jewell C.P., Noble A.D., Heuer C., Wilson P.R. & Johnson W.O. (2013). – Bayesian estimation of the sensitivity and specificity of individual fecal culture and Paralisa to detect *Mycobacterium avium* subspecies *paratuberculosis* infection in young farmed deer. *J. Vet. Diag. Invest.,* **25** (6), 759–764. doi:10.1177/1040638713505587.

28. Wood C., Muleme M., Tan T., Bosward K., Gibson J., Alawneh J., McGowan M., Barnes T.S., Stenos J., Perkins N., Firestone S.M. & Tozer S. (2019). – Validation of an indirect immunofluorescence assay (IFA) for the detection of IgG antibodies against *Coxiella burnetii* in bovine serum. *Prev. Vet. Med.,* **169**, 104698. doi:10.1016/j.prevetmed.2019.104698.

29. Paul S., Agger J.F., Agerholm J.S. & Markussen B. (2014). – Prevalence and risk factors of *Coxiella burnetii* seropositivity in Danish beef and dairy cattle at slaughter adjusted for test uncertainty. *Prev. Vet. Med.,* **113** (4), 504–511. doi:10.1016/j.prevetmed.2014.01.018.

30. Okura H., Nielsen S.S. & Toft N. (2010). – Prevalence of *Mycobacterium avium* subsp. *paratuberculosis* infection in adult Danish non-dairy cattle sampled at slaughter. *Prev. Vet. Med.,* **94** (3), 185–190. doi:10.1016/j.prevetmed.2010.01.014.

31. Johnson W.O. & Pearson L.M. (1999). – Dual screening. *Biometrics,* **55** (3), 867–873. doi:10.1111/j.0006-341x.1999.00867.x.

32. Hanson T.E., Johnson W.O. & Gastwirth J.L. (2006). – Bayesian inference for prevalence and diagnostic test accuracy based on dual-pooled screening. *Biostatistics,* **7** (1), 41–57. doi:10.1093/biostatistics/kxi039.

33. Dhand N.K., Johnson W.O. & Toribio J.-A.L.M.L. (2010). – A Bayesian approach to estimate OJD prevalence from pooled fecal samples of variable pool size. *J. Agric. Biol. Environ. Stat.,* **15** (4), 452–473. doi: 10.1007/s13253-010-0032-8.

34. World Organisation for Animal Health (OIE). (2013). – Chapter 1.1.6. Principles and methods of validation of diagnostic assays for infectious diseases. *In* Manual of Diagnostic Tests and Vaccines for Terrestrial Animals. OIE, Paris, France, 8th Ed. Available at: www.oie.int/fileadmin/Home/eng/Health_standards/tahm/1.01.06_VALIDATION.pdf (accessed on 24 May 2021).

35. Bisoffi Z., Pomari E. […] & Silva R. (2020). – Sensitivity, specificity and predictive values of molecular and serological tests for COVID-19: a longitudinal study in Emergency Room. *Diagnostics,* **10** (9), 669. doi:10.3390/diagnostics10090669.

36. Kostoulas P., Eusebi P. & Hartnack S. (2020). – Diagnostic accuracy estimates for COVID-19 RT–PCR and lateral flow immunoassay tests with Bayesian latent class models. Preprint (version 1). *Res. Square.* doi:10.21203/rs.3.rs-33243/v1.

37. Symons R., Beath K., Dangis A., Lefever S., Smismans A., De Bruecker Y. & Frans J. (2021). – A statistical framework to estimate diagnostic test performance for COVID-19. *Clin. Radiol.,* **76** (1), 75. doi:10.1016/j.crad.2020.10.004.

38. Bronsvoort B.M. de C., Anderson J., Corteyn A., Hamblin P., Kitching R.P., Nfon C., Tanya V.N. & Morgan K.L. (2006). – Geographical and age-stratified distributions of foot-and-mouth disease virus-seropositive and probang-positive cattle herds in the Adamawa province of Cameroon. *Vet. Rec.,* **159** (10), 299. doi:10.1136/vr.159.10.299.

39. Van Dreumel A.K., Michalski W.P., McNabb L.M., Shiell B.J., Singanallur N.B. & Peck G.R. (2015). – Pan-serotype diagnostic for foot-and-mouth disease using the consensus antigen of nonstructural protein 3B. *J. Clin. Microbiol.,* **53** (6), 1797. doi:10.1128/JCM.03491-14.

40. Cappai S., Sanna G., Loi F., Coccollone A., Marrocu E., Oggiano A., Brundu D., Rolesu S. & Bandino E. (2018). – African swine fever detection on field with antigen rapid kit test. *J. Anim. Sci. Res.,* **2** (3), 8 pp. doi:10.16966/2576-6457.118.

41. Comin A., Toft N., Stegeman A., Klinkenberg D. & Marangon S. (2013). – Serological diagnosis of avian influenza in poultry: is the haemagglutination inhibition test really the 'gold standard'? *Influenza Other Respir. Viruses,* **7** (3), 257–264. doi:10.1111/j.1750-2659.2012.00391.x.

42. Kostoulas P., Nielsen S.S., Branscum A.J., Johnson W.O., Dendukuri N., Dhand N.K., Toft N. & Gardner I.A. (2017). – STARD–BLCM: standards for the reporting of diagnostic accuracy studies that use Bayesian latent class models. *Prev. Vet. Med.,* **138**, 37–47. doi:10.1016/j.prevetmed.2017.01.006.

43. Jones G., Johnson W.O., Hanson T.E. & Christensen R. (2010). – Identifiability of models for multiple diagnostic testing in the absence of a gold standard. *Biometrics,* **66** (3), 855–863. doi:10.1111/j.1541-0420.2009.01330.x.

44. Mahmmod Y.S., Toft N., Katholm J., Grønbæk C. & Klaas I.C. (2013). – Bayesian estimation of test characteristics of real-time PCR, bacteriological culture and California mastitis test for diagnosis of intramammary infections with *Staphylococcus aureus* in dairy cattle at routine milk recordings. *Prev. Vet. Med.,* **112** (3), 309–317. doi:10.1016/j.prevetmed.2013.07.021.

45. Mathevon Y., Foucras G., Falguières R. & Corbiere F. (2017). – Estimation of the sensitivity and specificity of two serum ELISAs and one fecal qPCR for diagnosis of paratuberculosis in sub-clinically infected young-adult French sheep using latent class Bayesian modeling. *BMC Vet. Res.,* **13** (1), 230. doi:10.1186/s12917-017-1145-x.

46. Rahman A.K.M.A., Saegerman C. & Berkvens D. (2016). – Latent class evaluation of three serological tests for the diagnosis of human brucellosis in Bangladesh. *Trop. Med. Hlth,* **44** (1), 32. doi:10.1186/s41182-016-0031-8.

47. King D.P., Madi M., Mioulet V., Wadsworth J., Wright C.F., Valdazo-Gonzalez B., Ferris N.P., Knowles N.J. & Hammond J. (2012). – New technologies to diagnose and monitor infectious diseases of livestock: challenges for sub-Saharan Africa. *Onderstepoort J. Vet. Res.*, **79** (2), 456. doi:10.4102/ojvr.v79i2.456.

48. Marmion B. (2009). – A guide to Q fever and Q fever vaccination. CSL Biotherapies, Parkville, Australia, 122 pp.

49. Sethuraman N., Jeremiah S.S. & Ryo A. (2020). – Interpreting diagnostic tests for SARS-CoV-2. *JAMA,* **323** (22), 2249–2251. doi:10.1001/jama.2020.8259.

50. Lahuerta-Marin A., Milne M.G., McNair J., Skuce R.A., McBride S.H., Menzies F.D., McDowell S.J.W., Byrne A.W., Handel I.G. & Bronsvoort B.M. de C. (2018). – Bayesian latent class estimation of sensitivity and specificity parameters of diagnostic tests for bovine tuberculosis in chronically infected herds in Northern Ireland. *Vet. J.,* **238**, 15–21. doi:10.1016/j.tvjl.2018.04.019.

51. Matope G., Muma J.B., Toft N., Gori E., Lund A., Nielsen K. & Skjerve E. (2011). – Evaluation of sensitivity and specificity of RBT, c-ELISA and fluorescence polarisation assay for diagnosis of brucellosis in cattle using latent class analysis. *Vet. Immunol. Immunopathol.,* **141** (1), 58–63. doi:10.1016/j.vetimm.2011.02.005.

52. Bauman C.A., Jones-Bitton A., Menzies P., Toft N., Jansen J. & Kelton D. (2016). – Prevalence of paratuberculosis in the dairy goat and dairy sheep industries in Ontario, Canada. *Can. Vet. J.,* **57** (2), 169–175. Available at: www.ncbi.nlm.nih.gov/pmc/articles/PMC4712996/ (accessed on 11 March 2021).

53. Fikru R., Andualem Y., Getachew T., Menten J., Hasker E., Merga B., Goddeeris B.M. & Büscher P. (2015). – Trypanosome infection in dromedary camels in eastern Ethiopia: prevalence, relative performance of diagnostic tools and host related risk factors. *Vet. Parasitol.,* **211** (3), 175–181. doi:10.1016/j.vetpar.2015.04.008.

54. Campe A., Abernethy D., Menzies F. & Greiner M. (2016). – Latent class regression models for simultaneously estimating test accuracy, true prevalence and risk factors for *Brucella abortus*. *Epidemiol. Infect.,* **144** (9), 1845–1856. doi:10.1017/S0950268816000157.

55. Blake A., Njanpop-Lafourcade B.M., Telles J.N., Rajoharison A., Makawa M.S., Agbenoko K., Tamekloe S., Mueller J.E., Tall H., Gessner B.D., Paranhos-Baccalà G. & Moïsi J.C. (2017). – Evaluation of chest radiography, lytA real-time PCR, and other routine tests for diagnosis of community-acquired pneumonia and estimation of possible attributable fraction of pneumococcus in northern Togo. *Epidemiol. Infect.,* **145** (3), 583–594. doi:10.1017/S0950268816002211.

56. Plucinski M.M., Candrinho B., Dimene M., Smith T., Thwing J., Colborn J., Rogier E. & Zulliger R. (2020). – Estimation of malaria-attributable fever in malaria test-positive febrile outpatients in three provinces of Mozambique, 2018. *Am. J. Trop. Med. Hyg.,* **102** (1), 151–155. doi:10.4269/ajtmh.19-0537.

57. Caraguel C., Stryhn H., Gagné N., Dohoo I. & Hammell L. (2012). – Use of a third class in latent class modelling for the diagnostic evaluation of five infectious salmon anaemia virus detection tests. *Prev. Vet. Med.,* **104** (1), 165–173. doi:10.1016/j.prevetmed.2011.10.006.

58. Nielsen P.K., Petersen M.B., Nielsen L.R., Halasa T. & Toft N. (2015). – Latent class analysis of bulk tank milk PCR and ELISA testing for herd level diagnosis of *Mycoplasma bovis*. *Prev. Vet. Med.,* **121** (3), 338–342. doi:10.1016/j.prevetmed.2015.08.009.

59. Bauman C.A., Jones-Bitton A., Jansen J., Kelton D. & Menzies P. (2016). – Evaluation of fecal culture and fecal RT–PCR to detect *Mycobacterium avium* ssp. *paratuberculosis* fecal shedding in dairy goats and dairy sheep using latent class Bayesian modeling. *BMC Vet. Res.,* **12** (1), 212. doi:10.1186/s12917-016-0814-5.

60. World Organisation for Animal Health (OIE) (2019). – Chapter 1.1.2. Collection, submission and storage of diagnostic specimens. *In* Manual of Diagnostic Tests and Vaccines for Terrestrial Animals. OIE, Paris, France, 8th Ed. Available at: www.oie.int/fileadmin/Home/eng/Health_standards/tahm/1.01.02_COLLECTION_DIAG_SPECIMENS.pdf (accessed on 11 March 2021).

61. Stevenson M., Nunes T., Sanchez J., Thornton R., Reiczigel J., Robison-Cox J. & Sebastiani P. (2013). – epiR: an R package for the analysis of epidemiological data [computer software]. R package version 0.9-43.

62. Branscum A.J., Johnson W.O. & Gardner I.A. (2007). – Sample size calculations for studies designed to evaluate diagnostic test accuracy. *J. Agric. Biol. Environ. Stat.,* **12** (1), 112. doi:10.1198/108571107X177519.

63. Su C.-L., Gardner I.A. & Johnson W.O. (2007). – Bayesian estimation of cluster-level test accuracy based on different sampling schemes. *J. Agric. Biol. Environ. Stat.,* **12** (2), 250. doi:10.1198/108571107X198895.

64. Wood C., Perkins N.R., Tozer S.J., Johnson W., Barnes T.S., McGowan M., Gibson J.S., Alawneh J., Firestone S.M. & Woldeyohannes S.M. (2021). – Prevalence and spatial distribution of *Coxiella burnetii* seropositivity in northern Australian beef cattle adjusted for diagnostic test uncertainty. *Prev. Vet. Med*., **189**, 105282. doi:10.1016/j.prevetmed.2021.105282.

65. Branscum A.J., Gardner I.A. & Johnson W.O. (2004). – Bayesian modeling of animal- and herd-level prevalences. *Prev. Vet. Med.,* **66** (1), 101–112. doi:10.1016/j.prevetmed.2004.09.009.

66. Dendukuri N., Hadgu A. & Wang L. (2009). – Modeling conditional dependence between diagnostic tests: a multiple latent variable model. *Stat. Med.,* **28** (3), 441–461. doi:10.1002/sim.3470.

67. Ron-Román J., Ron-Garrido L., Abatih E., Celi-Erazo M., Vizcaíno-Ordóñez L., Calva-Pacheco J., González-Andrade P., Berkvens D., Benítez-Ortíz W., Brandt J., Fretin D. & Saegerman C. (2019). – Bayesian evaluation of three serological tests for detecting antibodies against *Brucella* spp. among humans in the northwestern part of Ecuador. *Am. J. Trop. Med. Hyg.,* **100** (6), 1312–1320. doi:10.4269/ajtmh.18-0622.

68. Wang Z., Dendukuri N., Zar H.J. & Joseph L. (2017). – Modeling conditional dependence among multiple diagnostic tests. *Stat. Med.,* **36** (30), 4843–4859. doi:10.1002/sim.7449.

69. Spiegelhalter D.J., Best N.G., Carlin B.P. & Van Der Linde A. (2002). – Bayesian measures of model complexity and fit. *J. Roy. Stat. Soc., B (Stat. Methodol.),* **64** (4), 583–639. doi:10.1111/1467-9868.00353.

70. Lunn D., Jackson C., Best N., Thomas A. & Spiegelhalter D. (2012). – The BUGS book: a practical introduction to Bayesian analysis. CRC Press, Boca Raton, United States of America, 399 pp.

71. Kostoulas P. (2018). – PriorGen: generates prior distributions for proportions [computer software]. R 364 Package.

72. Christensen R., Johnson W., Branscum A. & Hanson T.E. (2011). – Bayesian ideas and data analysis: an introduction for scientists and statisticians. CRC Press, Boca Raton, United States of America, 516 pp.

73. Muma J.B., Toft N., Oloya J., Lund A., Nielsen K., Samui K. & Skjerve E. (2007). – Evaluation of three serological tests for brucellosis in naturally infected cattle using latent class analysis. *Vet. Microbiol.,* **125** (1), 187–192. doi:10.1016/j.vetmic.2007.05.012.

74. Gelman A. & Rubin D.B. (1992). – Inference from iterative simulation using multiple sequences. *Stat. Sci.,* **7** (4), 457–472. doi:10.1214/ss/1177011136.

75. Plummer M., Best N., Cowles K. & Vines K. (2006). – Package 'coda' [computer software]. R package version 0.19-4.

76. Curtis S.M., Goldin I. & Evangelou E. (2018). – Package 'mcmcplots' [computer software]. R package version 0.4.3.

77. Meredith M. & Kruschke J. (2020). – 'HDInterval: highest (posterior) density intervals' [computer software]. R package version 0.2.2.

78. Johnson W., Jones G. & Gardner I. (2019). – Gold standards are out and Bayes is in: implementing the cure for imperfect reference tests in diagnostic accuracy studies. *Prev. Vet. Med.,* **167**, 113–127. doi:10.1016/j.prevetmed.2019.01.010.

79. University of California Davis Graduate Group in Epidemiology (2000–2006). – Glossary of epidemiological terms. Available at: epi.vet.unimelb.edu.au/Courses/EITV_Yarra_Dx_tests_Mar-2019/.

Appendix 1

Glossary

Apparent (test) prevalence (AP)

The probability that a randomly selected unit of analysis has a positive test result. The apparent prevalence is a function of the true prevalence (TP, defined below), diagnostic sensitivity (DSe), and diagnostic specificity (DSp) as shown in the following equation:

$$AP = TP \times DSe + (1 - TP)(1 - DSp).$$

Bayesian analysis

The process by which prior uncertainty about a quantity or quantities is formally described and, through the application of Bayes' Theorem, updated after observing data. The Bayesian method reflects that all uncertainty must be described by probability and that probability laws must be obeyed in order to produce coherent statistical inferences.

Conditional independence of tests

Two tests are conditionally independent when the sensitivity (or specificity) of one test does not depend on whether the results of another test are positive or negative among infected (or non-infected) individuals.

Frequentist analysis

A type of statistical analysis that draws conclusions based only on the sampling distribution of the observed data. For binomial data, this is the frequency or proportion of positive (negative) test results.

Latent class (mixture) model

A statistical model that does not rely on the assumption of a perfect reference (gold standard) test, but instead estimates the accuracy of candidate test(s) based on the combination of observed test outcome data, prior knowledge of test accuracy and prior knowledge of disease prevalence in the population(s) of interest. In a Bayesian latent class model, prior knowledge about the performance of the reference test and the candidate test can be incorporated into the analysis.

Likelihood

The joint probability or density of observing the data that were actually seen, regarded as a function of all the unknown parameters.

Model identifiability

A model is considered identifiable if it is possible to infer the true values of all of the unknown parameters after obtaining a sufficiently large number of observations from the model. Lack of model identifiability can be resolved by inclusion of informative priors about one or more parameters in the model.

Posterior distribution

A probability distribution that reflects uncertainty about a parameter or parameters of interest, after combining scientific information with observed data, using Bayes' theorem.

Prevalence or true prevalence

Estimate of the proportion of infected animals in a population at one given point in time, not to be confused with incidence.

Prior (prior distribution)

A probability distribution reflecting previous experimental data or scientific judgement that provides the basis for a Bayesian statistical model. When appropriately combined with the observed data, the prior is 'updated' to provide the posterior distribution, which is used to make inferences and draw conclusions.

Glossary definitions were adapted from the OIE Manual of Diagnostic Tests and Vaccines for Terrestrial Animals *(2) and from online resources from the University of California, Davis, Department of Medicine and Epidemiology (79).*

Supplementary materials

S1: Example R code using 'R2OpenBUGS', 'epiR' & 'mcmcplots'

See https://doi.org/10.26188/14274947.v1

S2: Resources, summary of key websites, software and books

PriorGen: https://cran.r-project.org/web/packages/PriorGen/index.html

Videos explaining the use of the PriorGen functions: https://youtu.be/M_5WwBGU95c and https://youtu.be/H7YI2bEEA08

'epiR' R package: https://cran.r-project.org/web/packages/epiR/index.html

Key functions: epi.betabuster, epi.sssimpleestb, epi.ssclus1estb and epi.ssdxtest.

Betabuster online: https://shiny.vet.unimelb.edu.au/epi/beta.buster/

OIE CC for Diagnostic Test Validation Science 'Interpretation and Validation of Diagnostic Tests in Veterinary Science' Workshop materials: https://epi.vet.unimelb.edu.au/Courses/EITV_Yarra_Dx_tests_Mar-2019/

COST-Harmony network materials: https://harmony-net.eu/training-material/

Hotline project: https://sites.google.com/view/the-hotline-project/home