American Journal of Epidemiology
© The Author 2013. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

Vol. 179, No. 4
DOI: 10.1093/aje/kwt286
Advance Access publication:
November 21, 2013

Systematic Reviews and Meta- and Pooled Analyses

Latent Class Models in Diagnostic Studies When There is No Reference Standard—A Systematic Review

Maarten van Smeden*, Christiana A. Naaktgeboren, Johannes B. Reitsma, Karel G. M. Moons, and Joris A. H. de Groot

* Correspondence to Maarten van Smeden, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG Utrecht, the Netherlands (e-mail: m.vansmeden@umcutrecht.nl).

Initially submitted April 29, 2013; accepted for publication October 28, 2013.

Latent class models (LCMs) combine the results of multiple diagnostic tests through a statistical model to obtain estimates of disease prevalence and diagnostic test accuracy in situations where there is no single, accurate reference standard. We performed a systematic review of the methodology and reporting of LCMs in diagnostic accuracy studies. This review shows that the use of LCMs in such studies increased sharply in the past decade, notably in the domain of infectious diseases (overall contribution: 59%). The 64 reviewed studies used a range of differently specified parametric latent variable models, applying Bayesian and frequentist methods. The critical assumption underlying the majority of LCM applications (61%) is that the test observations must be independent within 2 classes. Because violations of this assumption can lead to biased estimates of accuracy and prevalence, performing and reporting checks of whether assumptions are met is essential. Unfortunately, our review shows that 28% of the included studies failed to report any information that enables verification of model assumptions or performance. Because of the lack of information on model fit and adequate evidence "external" to the LCMs, it is often difficult for readers to judge the validity of LCM-based inferences and conclusions reached.

diagnostic tests, routine; models, statistical; prevalence; reference standards; review; sensitivity and specificity

Abbreviation: LCM, latent class model.

An essential step in the evaluation of a diagnostic test or biomarker is to obtain valid estimates of its accuracy, that is, its ability to discriminate between patients who have the disease of interest and those who do not (1, 2). Typically, the accuracy of a single or set of diagnostic index tests is analyzed by examining the results of index tests in relation to the outcome of the reference standard in patients suspected of a disease of interest. A single and error-free reference standard is preferred, but for many diseases such a reference standard, also known as a "gold standard," does not exist (2–5). The use of an imperfect reference standard will often lead to misclassification of the disease status in a substantial portion of subjects, which can lead to biased estimates of index test performance and disease prevalence.

One approach to reducing these misclassifications is to combine multiple pieces of diagnostic information to determine the disease status among study patients. Multiple tests may, for example, be used in expert panels in which a group of clinicians reach consensus based on the available test results of patients (1, 5–7). Results from multiple tests can also be combined through a fixed rule as in a composite reference standard (i.e., diagnostic decision rule (8)). Finally, as a probabilistic alternative, a latent variable approach may be adopted by combining multiple diagnostic tests using a latent class model (LCM).

In the past decades, latent class modeling (i.e., latent class analysis) has been applied in medical and veterinary sciences, particularly in test accuracy research (9–13). The use of LCMs appears attractive because it avoids the time-consuming process of reaching consensus diagnoses and the inherent difficulty of defining a diagnostic decision rule a priori in cases where a single reference standard for the target disease is lacking. LCMs can produce valid estimates of accuracy even in the absence of a perfectly accurate disease

Downloaded from https://academic.oup.com/aje/article-abstract/179/4/423/128491 by guest on 09 April 2020

status classification (an accurate reference standard) and can be estimated in popular statistical software packages such as SAS (SAS Institute, Inc., Cary, North Carolina) and *R* (R Foundation for Statistical Computing, Vienna, Austria), as well as specialized software such as Latent GOLD (14).

Latent class modeling refers to a heterogeneous group of statistical models. Differently specified LCMs can be fitted to the same set of test results, which in turn can lead to relevant differences in disease prevalence and test accuracy estimates (15–19). Researchers, therefore, need to inform readers of how their LCMs were specified. Additionally, as with any statistical technique or model, the validity of its results are jeopardized when assumptions are not met. Hence, performing and reporting checks of whether assumptions are met is essential to allow readers to appraise the validity of the reported results when LCMs are used.

To explore the methodology and reporting of LCM applications in diagnostic research, we performed a systematic review of test accuracy studies that applied such a model. This review provides an overview of the history of LCM applications in test accuracy research, reveals variations in the models used across studies, and provides clues on how to improve the reporting and methodology of these studies. Before presenting the results of the review, we will first describe the key characteristics of LCMs.

INTRODUCTION TO LCMs

Diagnostic studies that apply LCMs treat the target disease status as an unmeasured ("latent") categorical variable with K classes, reflecting the levels of the underlying disease. The manifest variables, the outcomes of R (binary) diagnostic tests, are considered to be imperfect classifiers of the disease status. The LCM describes a statistical model relating the manifest variables to the latent disease status. For the mathematical underpinning, see Appendix 1.

When $K = 2$, it is assumed that the 2 latent classes correspond to a class of subjects in which the target disease is present and a class of subjects in which the target disease is absent. Parameter estimates obtained from the 2-class LCM are interpreted as estimates of the sensitivity of each test (i.e., the probability of a positive test result when the target disease is present) and specificity of each test (i.e., the probability of a negative test result when the target disease is absent) and the prevalence of the target disease (i.e., the (prior) probability that the target disease is present). Two important issues that are encountered when applying LCMs are identification of the LCM and the assumption of conditional independence.

Estimation and identifiability of LCMs

Maximum likelihood estimates of the LCM parameters can be obtained by using a variety of optimization methods, including expectation-maximization or Newton-Raphson algorithms (14). However, LCMs may not always be identified, which implies that the maximum likelihood estimates are not "unique"; a different set of parameter estimates exists for which the likelihood (value) is the same.

A necessary condition for the LCM to be identified is that the number of freely estimated parameters does not exceed the number of unique diagnostic test patterns. For example, an unconstrained "basic" 2-class LCM is not identified with $R \leq 2$ diagnostic tests and is "just identified" with $R = 3$, resulting in an LCM with 0 degrees of freedom. Nonnegative degrees of freedom, though, is not a sufficient condition for identification, as is evidenced by an LCM with $R = 4$ and $K = 3$, which has 1 degree of freedom but is not locally identified (20). In practice, local identifiability of LCMs can be explored by examining the rank of the Jacobian matrix (20, 21).

One solution to nonidentifiability is imposing constraints to the parameters (20). For example, if the conditional test outcome probabilities of a particular diagnostic test can be assumed to be known a priori (e.g., based on theory (22)), then degrees of freedom can be gained by fixing the parameters to their "true" values, allowing the remaining LCM parameters to be estimated freely.

Another strategy is to adopt a Bayesian approach. Because the true values of conditional diagnostic test outcome probabilities and the disease prevlance are often not exactly known in advance, the use of fixed parameters may be invalid. Instead of constraining the parameters to a fixed value, "informative" prior distribution can be defined for those parameters for which prior knowledge is available. With substantive prior information, estimates from the posterior densities of unidentified LCMs can be obtained by a Gibbs sampler (23). Detailed discussions on optimization and identification are found elsewhere (20, 24).

Conditional independence

The important assumption that underlies LCM estimation is that of conditional independence (i.e., local independence). In its most basic form, this assumption reduces to independence of observations conditional on the presence or absence of the disease of interest. This assumption is central to the 2-class independence LCM (defined in equation 3 in Appendix 1). It results in the model being identified with only 3 binary diagnostic tests. However, violations of local independence assumptions are known to lead to bias in estimates of accuracy and prevalence, and the assumptions may not be warranted in many practical situations (25–28).

One way to relax the conditional independence assumption is by increasing the number of latent classes to be estimated. We will refer to these LCMs with more than 2 classes as multiclass independence LCMs. Alternatively, other LCMs that do not require the independence assumption of observations conditional on (or "within") the classes have been suggested. For example, when independence among observations in a K-class LCM is not met because of bivariate dependence among a pair or a subset of pairs of tests, these bivariate associations can be modeled directly by defining an additional parameter for each bivariate relation (29–31). Other strategies to account for dependence within classes include defining marginal models (32), estimating a multiple latent variable model (19), and adding random effects (33, 34). We will refer to these models as dependence LCMs. For a detailed discussion of common LCM specifications in diagnostic research, we refer to the report by Hui and Zhou (10).

Downloaded from https://academic.oup.com/aje/article-abstract/179/4/423/128491 by guest on 09 April 2020

Evidently, an increased number of parameters is estimated when the independence assumption is relaxed, at the expense of degrees of freedom. Hence, a higher number of diagnostic tests is needed for extensions to the independence 2-class LCMs. For example, at least 5 binary diagnostic tests are needed for an unconstrained 3-class independence LCM to be identified. Estimating multiclass independence LCMs or dependence LCMs is, therefore, not always feasible when the number of available diagnostic tests (*R*) is limited.

LCM verifications

To verify LCM assumptions, measures of model fit can provide important information. Preferred is an LCM that provides a superior, or at least equivalent, fit to the data compared with alternative LCMs (e.g., with more classes) and one that has an adequate "global" fit (i.e., the differences in observed vs. the expected number of patients with specific patterns of test results should be small). Particularly useful are also residual dependence diagnostics that can pinpoint sources of misfit due to bivariate dependence between diagnostic tests (14, 33, 35).

Clearly, evaluating model fit can provide important information regarding potential misspecification of the LCMs used. The reporting of the model evaluation steps taken and results obtained from alternative models therefore provides valuable information for readers to judge the credibility of the results presented. Nonetheless, because the latent variable is unobserved, LCM assumptions cannot be tested directly. Even with a large sample size, it may be hard to distinguish between LCMs (15).

The credibility of LCM-based inferences may therefore also rely on the use of external data. For example, it is sometimes possible to compare the latent class outcomes with outcomes derived by using a proxy measure for disease status (e.g., disease status measured by using an adequate reference standard in a subset of patients (36)).

SYSTEMATIC REVIEW OF DIAGNOSTIC STUDIES APPLYING LCMs

Our aim was to identify diagnostic studies that reported diagnostic test accuracy or disease prevalence estimates derived directly or as a function of LCM parameter estimates. Hence, techniques that use assigned clusters (e.g., cluster analysis) to derive accuracy or prevalence estimates fall outside of the scope of this review. The EMBASE and PubMed databases were searched for the following free-text search terms: "latent class" OR "latent classes" OR "finite mixture" OR "finite mixtures" on November 8, 2011. Papers published in English, Dutch, or Spanish were considered for inclusion.

Information was extracted on clinical context, study characteristics, model specifics and diagnostics, model comparisons, and software. The extraction form was pilot tested by 4 researchers (M.v.S., C.A.N., J.B.R., and J.A.H.d.G.). One researcher (M.v.S.) screened and evaluated all studies. Parallel screening and data extraction were performed independently by a second researcher (C.A.N., J.B.R., or J.A.H.d.G.). Disagreements were resolved by discussion between the first and second researchers, and in case of remaining doubt, by a research-group discussion.

To evaluate reported evidence of LCM performance in diagnostic studies, we differentiated among the following 4 categories of model fit criteria: 1) assessment of relative fit, which ideally results in the selection of 1 LCM among a set of theoretically plausible LCMs, for example, by significance testing or information criteria (e.g., Akaike information criterion); 2) goodness-of-fit testing, evaluating the fit of an LCM to the observed data by using a significance test (e.g., Pearson χ^2 test (37)); 3) dependence diagnostics to pinpoint residual dependence under an LCM (35), for example, correlation residual plot (20) or bivariate residual statistics (13); and 4) a table of expected versus observed frequencies.

Model performance can also be evaluated by a leave-1-out comparison, in which a manifest variable is excluded and its results are compared with the results obtained from another LCM in which the manifest variable is included. Finally, model performance can be evaluated by using external evidence, for example, by comparing LCM estimates with similar estimates derived using an imperfect reference standard.

RESULTS

Figure 1 depicts the results of the search and inclusion of papers. Of the 1,704 papers whose titles and abstracts were screened for eligibility, 242 met the eligibility criteria. One publication could not be retrieved in full-text format. After full-text reading of 241 publications, we excluded another 91. Reasons for exclusion included the following: falling outside the scope of this paper, reporting LCMs that only estimated (rater) agreement (38), and featuring models that included continuous test results (39) or covariates (40). Screening of reference lists of the included papers yielded an additional 30 publications, resulting in a total of 180 publications, of which 179 were published in English and 1 in Spanish.

The included papers were classified as theoretical ($n = 69$) when focused on LCM methodology or empirical ($n = 111$) when the focused on analyzing original data ("original papers"). The empirical studies were further divided into animal studies ($n = 47$) and human studies ($n = 64$). In the remainder of this paper, we will focus only on the empirical studies involving human subjects. A short description of the veterinary studies is found in Appendix 2.

General characteristics

The first applications of LCMs in diagnostic studies involving human subjects originated around 1990. A steep increase in statistical and methodological publications has occurred since 2000, followed by an increase in empirical studies starting approximately 5 years later (Figure 2). In total, 64 studies were identified, of which approximately half ($n = 34$) were published between 2007 and 2011.

The primary goal of the vast majority of publications ($n = 59$; 92%) was the evaluation of accuracy of diagnostic tests ($n = 51$) or accuracy of a diagnosis made by clinicians ($n = 8$) (Table 1). All of these studies reported test sensitivity and specificity estimates; predictive values of the tests were

Downloaded from https://academic.oup.com/aje/article-abstract/179/4/423/128491 by guest on 09 April 2020

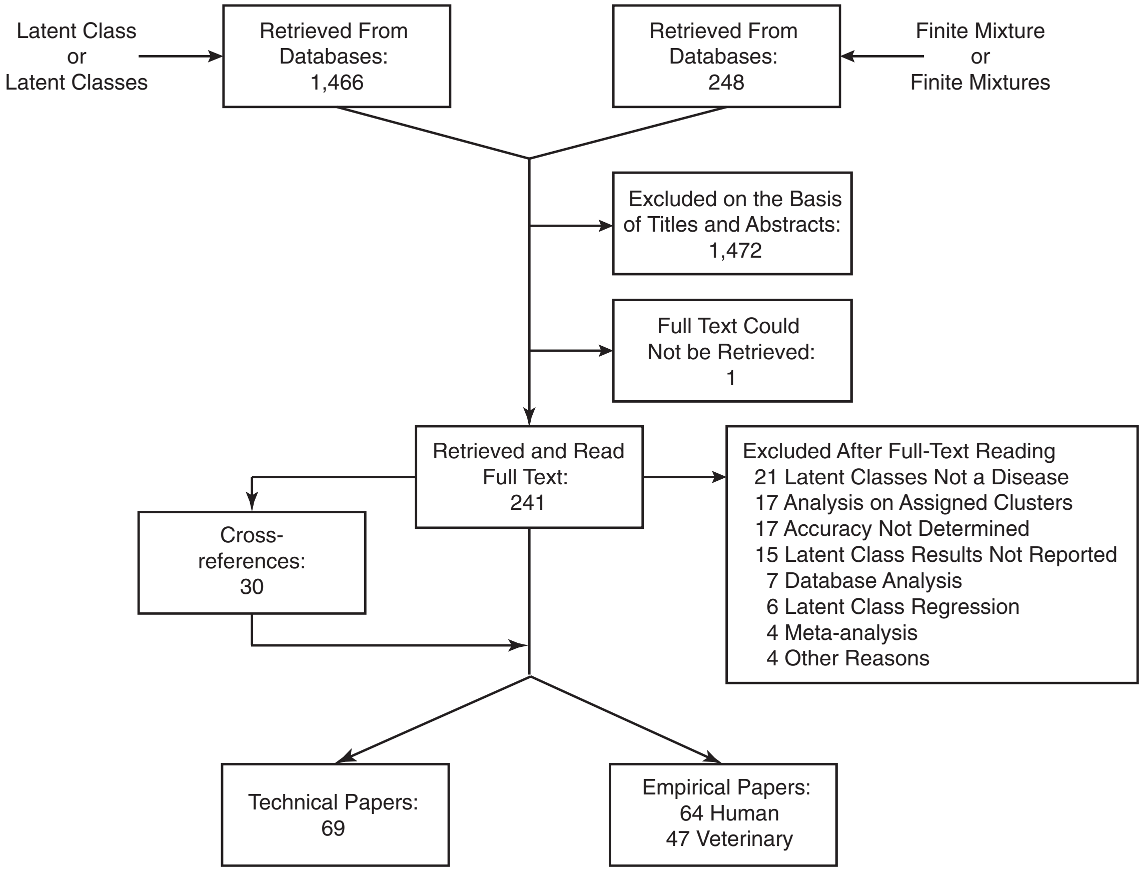

Figure 1. Inclusion of studies for systematic review on latent class methodology in diagnostic studies. Other reasons for exclusion were the use of longitudinal analysis, being identical to an included paper, or the use of nonparametric models.

additionally reported in 11 of these studies (17%). The predictive values of each particular test pattern combination were reported in 7 studies. In 4 studies, the primary goal was to estimate disease prevalence. In 1 study, the primary goal could not be determined.

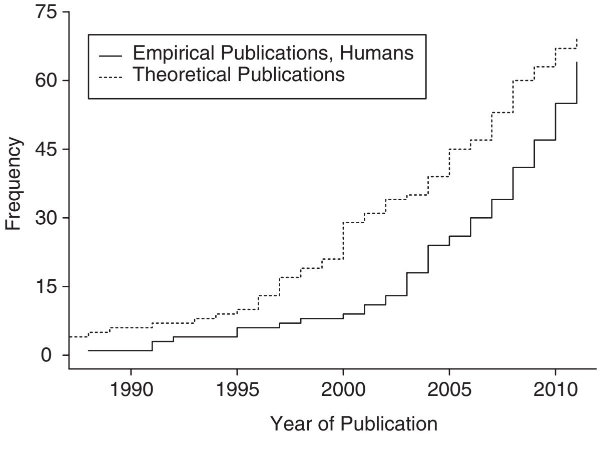

Figure 2. Cumulative number of empirical diagnostic studies in humans using a latent class model and theoretical studies published.

LCM applications were found primarily in the field of infectious disease research. In 38 publications (59%), the disease of interest was an infectious disease. Other diseases were mental or behavioral problems ($n = 6$), diseases of the musculoskeletal system ($n = 4$), diseases of the digestive system ($n = 3$), and neoplasms ($n = 2$).

Thirty-three papers (52%) included either 3 ($n = 19$) or 4 ($n = 14$) diagnostic tests (i.e., manifest variables) in their LCM(s). The median number of included diagnostic tests was 4 (interquartile interval, 3–6). Reported sample sizes ranged from 34 to 4,708 with a median of 315 (interquartile interval, 171–737).

Variety of LCMs

The majority of studies ($n = 39$; 61%) reported analyses based solely on 2-class independence LCMs (Table 2). These studies did not report LCM estimates or model fit statistics from multiclass LCMs or dependence LCMs. These studies also did not report comparisons between LCMs, of which at least 1 was not a 2-class independence LCM.

Downloaded from https://academic.oup.com/aje/article-abstract/179/4/423/128491 by guest on 09 April 2020

Table 1. Characteristics of 64 Studies That Used Latent Class Models to Estimate Accuracy of Diagnostic Tests or Prevalence of a Disease in Humans

Characteristic	No. of Studies
Latent condition of interest[a]	
Infectious and parasitic diseases	38
Mental and behavioral disorders	6
Diseases of the musculoskeletal system	4
Diseases of the digestive system	3
Neoplasms	2
Diseases of the respiratory system	2
Other	9
Year of publication	
2007–2011	34
2002–2006	19
<2002	11
Main goal of publication	
Test accuracy	51
Expert accuracy	8
Disease prevalence	4
Unknown	1

[a] Based on *International Classification of Diseases, Tenth Edition*, codes.

Analyses originating from 3- and 4-class LCMs were reported in 12 papers (19%). LCMs with more than 4 classes were not found in the surveyed studies. Fifteen papers (23%) reported results based on dependence LCMs (e.g., by including random effects). Two papers could not be classified in Table 2 because of insufficient information on the number of classes estimated.

Assessment of relative fit

In 24 studies (38%), all results originated from a single LCM. The other 40 studies reported parameter estimates or model diagnostics of 2 or more differently specified LCMs. In 15 of the 40 studies presenting more than 1 LCM, information on relative fit was reported (e.g., by significance testing or information criteria).

Model fit and leave-1-out comparison

Goodness-of-fit evaluation of the model(s) based on significance testing was reported in 26 studies (41%); a table of observed against expected number of test patterns was present in 8 studies (13%); and residual dependence diagnostics were reported in 4 studies. Thirty-five (55%) studies did not report any of the model fit criteria on relative fit, goodness-of-fit, dependence diagnostics, or expected versus observed test pattern frequencies.

Five studies reported a leave-1-out comparison. These studies evaluated the stability of obtained LCM estimates using varying subsets of the available diagnostic tests in the LCMs (i.e., reported analyses originating from multiple LCMs differing in the diagnostic tests that were included in the model). In total, 33 studies (52%) did not report any information on model fit criteria or an evaluation of estimates stability.

External evidence to support LCM findings

Thirty-three studies (52%) reported additional results that were not directly derived from an LCM, which enables a comparison with the estimates obtained from the LCM(s). For example, by using 1 of the diagnostic tests as a reference standard, we determined apparent accuracy and prevalence estimates for the other tests in 17 studies (27%). Eight studies (13%) applied a fixed rule that combined the results of diagnostic tests to determine the target disease status of patients (i.e., a composite reference standard). In 5 studies (8%), the disease status was determined by a single expert or panel of experts. Three studies used a combination of the above. Eighteen studies (28%) failed to report any information that could be used to verify the validity of LCM estimates.

Estimation

In most studies, optimization was obtained by using maximum likelihood estimation ($n = 49$; 77%); 14 studies (22%) used a Bayesian optimization approach. One study reported estimates obtained from frequentist and Bayesian models. A total of 14 different software programs were reported in 52 studies; Latent GOLD ($n = 10$), WinBUGS (MRC Biostatistics Unit, Cambridge, United Kingdom) ($n = 9$), LEM (Prof. Jeroen Vermunt, Tilburg, the Netherlands) ($n = 8$), LATENT 1 (McMaster University, Hamilton, Ontario, Canada) ($n = 7$), and SAS (SAS Institute, Inc.) ($n = 4$) were most frequently reported.

DISCUSSION

Our systematic review shows that the use of LCMs in diagnostic studies has increased considerably in the past decade. This is probably a reflection of increased awareness that a gold standard does not exist for many conditions (2–5). LCMs may seem appealing because they combine the results of multiple tests to improve the classification of the disease status in an objective way. Our review, however, revealed several problematic issues related to the methodology and reporting of studies using LCMs that deserve further attention.

The majority of studies used a 2-class independence LCM to estimate test sensitivity, specificity, and target disease prevalence. The strong assumption made in these studies is that, conditional on the binary target disease status, test results are independent ("uncorrelated"). This conditional independence assumption can easily be violated, for instance, when some individuals without the disease of interest have another condition in common that increases the likelihood of 2 (or more) tests to render false positives because they are based on a comparable biological principle (19). Another cause for conditional dependence among test results could arise if there is a subgroup of individuals with an early or

Downloaded from https://academic.oup.com/aje/article-abstract/179/4/423/128491 by guest on 09 April 2020

Table 2. Properties of LCMs in Absolute Numbers as Applied in 64 Empirical Diagnostic Accuracy Studies

Property	Total No. of Studies	No. Reporting Single LCM ($n = 23$)			No. Reporting >1 LCM ($n = 39$)			No. Not Classified[a]
		$K = 2$[b]	$K > 2$[c]	Dependent[d]	$K = 2$[b]	$K > 2$[c]	Dependent[d]	
Frequency	64	18	1	4	21	7	11	2
Model fit measures								
Relative model fit	15	NA	NA	NA	3	3	9	0
Goodness-of-fit testing	26	3	0	1	11	4	7	0
Dependence diagnostics	4	1	0	0	1	0	2	0
Expected/observed frequencies	8	1	0	0	1	2	4	0
None of the above	35	14	1	3	10	3	2	2
Leave-1-out comparison[e]	5	NA	NA	NA	4	0	1	0
No model fit nor leave-1-out	33	14	1	3	8	3	2	2
External evidence	33	12	0	2	9	2	7	1
No model fit nor leave-1-out nor external evidence	18	6	1	1	4	3	2	1
Bayesian optimization	15	3	0	4	4	1	3	0

Abbreviations: LCM, latent class model; NA, not applicable.

[a] Number of latent classes modeled could not be extracted from publications. One of these papers reported analyses based on a single LCM, whereas the other reported multiple models of which the number of latent classes could not be determined.

[b] Studies using the 2-class independence LCM.

[c] Studies using the multiclass independence model.

[d] Four publications presenting results of dependence LCMs also presented results from multiclass independence LCMs or multiclass dependence LCMs.

[e] Leave-1-out comparison refers to the situation in which a manifest variable is excluded and its results are compared with the results obtained from another LCM in which the manifest variable is included.

less severe stage of the disease of interest and if these individuals are more likely to be missed by different tests (41).

Examining whether the LCM used for inferences is appropriate for the data at hand is critical, because violating the independence assumptions can lead to biased estimates of accuracy (15, 25–27). Several approaches exist, such as examining residual correlations, comparing with alternative LCMs that assume different dependence structures (dependence LCMs), or evaluating the goodness-of-fit. Unfortunately, more than half of studies (52%) fail to present any information that is related to the fit of the model or stability of the estimates. Because of this absence of model performance information, readers are often left in the dark about the appropriateness of the models and the validity of the results.

One contributing factor for the limited number of studies comparing the results from the 2-class independence model with more complex models is the limited number of diagnostic tests (i.e., manifest variables) available in the surveyed studies. This is especially true for studies that have data available on only 3 diagnostic tests. The possibilities for evaluating model fit and model comparisons are then limited unless parameter constraints can be imposed.

The parameter constraints that could be imposed can take the form of fixing parameters to a "known" value or equality constraints. A Bayesian alternative is defining informative prior distributions on the parameters for which prior information is available, which creates an opportunity to estimate and compare LCMs that, from a frequentist perspective, are not identified (23). Of course, the validity of a Bayesian approach relies on the proper use of prior information. When prior information is lacking, researchers might want to collect data on additional diagnostic tests to verify LCM assumptions.

Performing and reporting on the checks of assumptions is a major step forward in the critical appraisal of the results of LCMs. However, we acknowledge the limited "power" of performance criteria to detect misfit. In the absence of explicit criteria, researchers can incorporate partial information on the disease status into the analysis (e.g., an adequate reference standard measured for a subset of patients (36)) or use model averaging techniques (34). Applications of model averaging or incorporation of partial information on the disease status were, however, not found in our systematic search.

To substantiate the face validity of inferences, some of the studies alternatively reported external evidence to enable comparison of estimates derived using the LCM. For example, some studies compared LCM estimates with estimates derived using a composite reference standard or a panel diagnosis.

More often, though, external evidence was based on apparent disease prevalence and test sensitivity, and specificity estimates were obtained by using 1 of the available tests as the (imperfect) reference standard. A comparison of these estimates with LCM parameter estimates rarely contributes to a satisfying conclusion regarding the credibility of LCM-based accuracy and prevalence estimates because the reason for applying LCMs is the absence of an accurate reference standard. The comparison of the "naïve" estimates obtained from the

Downloaded from https://academic.oup.com/aje/article-abstract/179/4/423/128491 by guest on 09 April 2020

imperfect reference standard with the LCM estimates might not be used for evaluation of the LCM, but rather as a sensitivity analysis for the obtained naïve estimates, taking into account the imperfect nature of the standard. Surely, the validity of this comparison also relies on the assumptions of the LCM.

We recognize that some publications of diagnostic studies that use LCMs may not have been identified in our review. However, because this is a review of reported methods, complete coverage is not as critical as it is, for example, in an intervention review. Our broad search strategy captured a large representative sample, and the risk of missing relevant publications was reduced by checking the reference lists of included publications. Because our goal was to evaluate the LCM reporting and methodology in peer-reviewed journals, research that has not been published falls outside the scope of our review. Therefore, the potential for bias due to unpublished work (i.e., publication bias) is not relevant.

The recent increase in the use of LCMs in test accuracy research has coincided with an increase in attention to problems that can be encountered when facing a reference standard that is imperfect. It should be recognized that, despite all the attention given to the potential problems, valid LCMs can reduce the risk of reference standard bias. However, the reporting of latent class analyses is often insufficient for readers to be able to critically appraise the obtained results. To improve reporting, we suggest that future studies provide detailed information about the exact specifications of their LCMs. Additionally, we suggest that all studies describe in detail how the assumption of conditional independence was verified, as well as report information that can help readers appraise the validity of the obtained results.

ACKNOWLEDGMENTS

Author affiliations: Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands (Maarten van Smeden, Christiana A. Naaktgeboren, Johannes B. Reitsma, Karel G. M. Moons, Joris A. H. de Groot).

We gratefully acknowledge financial contribution by the Netherlands Organisation for Scientific Research (project 918.10.615).

Conflict of interest: none declared.

REFERENCES

1. Knottnerus JA, ed. *The Evidence Base of Clinical Diagnosis*. London, United Kingdom: BMJ Books; 2002.
2. Rutjes AW, Reitsma JB, Coomarasamy A, et al. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess*. 2007;11(50):iii, ix–51.
3. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med*. 1987;6(4):411–423.
4. Reitsma JB, Rutjes AW, Khan KS, et al. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *J Clin Epidemiol*. 2009;62(8):797–806.
5. Moons KG, Grobbee DE. When should we remain blind and when should our eyes remain open in diagnostic studies? *J Clin Epidemiol*. 2002;55(7):633–636.
6. Gagnon R, Charlin B, Coletti M, et al. Assessment in the context of uncertainty: How many members are needed on the panel of reference of a script concordance test? *Med Educ*. 2005;39(3):284–291.
7. Worster A, Carpenter C. Incorporation bias in studies of diagnostic tests: how to avoid being biased about bias. *CJEM*. 2008;10(2):174–175.
8. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Stat Med*. 1999; 18(22):2987–3003.
9. Rindskopf D, Rindskopf W. The value of latent class analysis in medical diagnosis. *Stat Med*. 1986;5(1):21–27.
10. Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. *Stat Methods Med Res*. 1998;7(4):354–370.
11. Formann AK, Kohlmann T. Latent class analysis in medical research. *Stat Methods Med Res*. 1996;5(2):179–211.
12. Garrett ES, Eaton WW, Zeger S. Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: a latent class model approach. *Stat Med*. 2002;21(9): 1289–1307.
13. Young MA. Evaluating diagnostic criteria: a latent class paradigm. *J Psychiatr Res*. 1982;17(3):285–296.
14. Vermunt JK, Magidson J. *Technical Guide for Latent GOLD 4.0: Basic and Advanced*. Belmont, MA: Statistical Innovations Inc; 2005.
15. Albert PS, Dodd LE. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*. 2004;60(2):427–435.
16. Xu H, Craig BA. A probit latent class model with general correlation structures for evaluating accuracy of diagnostic tests. *Biometrics*. 2009;65(4):1145–1155.
17. Chu H, Zhou Y, Cole SR, et al. On the estimation of disease prevalence by latent class models for screening studies using two screening tests with categorical disease status verified in test positives only. *Stat Med*. 2010;29(11):1206–1218.
18. Black MA, Craig BA. Estimating disease prevalence in the absence of a gold standard. *Stat Med*. 2002;21(18):2653–2669.
19. Dendukuri N, Hadgu A, Wang L. Modeling conditional dependence between diagnostic tests: a multiple latent variable model. *Stat Med*. 2009;28(3):441–461.
20. Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*. 1974;61(2): 215–231.
21. McCutcheon AL. *Latent Class Analysis*. Newbury Park, CA: Sage; 1987.
22. Walter SD. Estimation of test sensitivity and specificity when disease confirmation is limited to positive results. *Epidemiology*. 1999;10(1):67–72.
23. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *Am J Epidemiol*. 1995;141(3):263–273.
24. Skrondal AS, Rabe-Hesketh S. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: CRC Press; 2004.
25. Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*. 1985;41(4): 959–968.
26. Torrance-Rynard VL, Walter SD. Effects of dependent errors in the assessment of diagnostic test performance. *Stat Med*. 1997; 16(19):2157–2175.
27. Spencer BD. When do latent class models overstate accuracy for diagnostic and other classifiers in the absence of a gold standard? *Biometrics*. 2012;68(2):559–566.

Downloaded from https://academic.oup.com/aje/article-abstract/179/4/423/128491 by guest on 09 April 2020

28. Walter SD, Macaskill P, Lord SJ, et al. Effect of dependent errors in the assessment of diagnostic or screening test accuracy when the reference standard is imperfect. *Stat Med*. 2012; 31(11-12):1129–1138.
29. Harper D. Local dependence latent structure models. *Psychometrika*. 1972;37(1):53–59.
30. Hagenaars JA, McCutcheon AL, eds. *Applied Latent Class Analysis*. Cambridge, United Kingdom: Cambridge University Press; 2002.
31. Espeland MA, Handelman SL. Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics*. 1989;45(2):587–599.
32. Yang I, Becker MP. Latent variable modeling of diagnostic accuracy. *Biometrics*. 1997;53(3):948–958.
33. Qu Y, Tan M, Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*. 1996;52(3):797–810.
34. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics*. 2001;57(1):158–167.
35. Sepulveda R, Vicente-Villardon JL, Galindo MP. The biplot as a diagnostic tool of local dependence in latent class models. A medical application. *Stat Med*. 2008;27(11):1855–1869.
36. Albert PS, Dodd LE. On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. *J Am Stat Assoc*. 2008;103(481):61–73.
37. Formann AK. Latent class model diagnostics: a review and some proposals. *Comput Stat Data Anal*. 2003;41(3-4): 549–559.
38. Agresti A. Modelling patterns of agreement and disagreement. *Stat Methods Med Res*. 1992;1(2):201–218.
39. Ladouceur M, Rahme E, Belisle P, et al. Modeling continuous diagnostic test data using approximate Dirichlet process distributions. *Stat Med*. 2011;30(21):2648–2662.
40. Lewis FI, Gunn GJ, McKendrick IJ, et al. Bayesian inference for within-herd prevalence of *Leptospira interrogans* serovar Hardjo using bulk milk antibody testing. *Biostatistics*. 2009; 10(4):719–728.
41. Brenner H. How independent are multiple ‘independent’ diagnostic classifications? *Stat Med*. 1996;15(13):1377–1386.
42. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. *Biometrics*. 1980;36(1):167–171.
43. Branscum AJ, Gardner IA, Johnson WO. Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. *Prev Vet Med*. 2005;68(2-4):145–163.

APPENDIX 1

Mathematical Underpinning of LCMs

Suppose that data are collected from a single diagnostic test X in a sample of patients suspected of having a specific target disease. Assuming that patients are either “diseased” ($D = 1$) or “not diseased” ($D = 0$), and that test results can be classified as either “positive” ($X = 1$) or “negative” ($X = 0$), it follows that the probability of observing a positive test result is the sum of probabilities of a true positive and a false positive test result,

$$\begin{aligned} P(X=1) &= P(X=1 \cap D=1) + P(X=1 \cap D=0) \\ &= \theta\alpha + (1-\theta)(1-\beta), \end{aligned} \tag{1a}$$

where θ is the prevalence of the target disease, and α and β are the sensitivity and specificity of the diagnostic test, respectively. Similarly, for the probability of a negative test result,

$$\begin{aligned} P(X=0) &= P(X=0 \cap D=0) + P(X=0 \cap D=1) \\ &= (1-\theta)\beta + \theta(1-\alpha). \end{aligned} \tag{1b}$$

A generalization of equations 1a and 1b can be written as

$$P(X=x) = \theta\alpha^{x}(1-\alpha)^{1-x} + (1-\theta)\beta^{1-x}(1-\beta)^{x}. \tag{2}$$

Assuming that a sample of subjects suspected of a target disease of size N is drawn and every sampled subject receives R diagnostic tests, equation 2 can be generalized by assuming independence of observations conditional on the target disease status as follows:

$$\begin{aligned} &P(X=[x_1,\ldots,x_R]) \\ &\quad = \theta\prod_{r=1}^{R}\alpha_r^{x_r}(1-\alpha_r)^{1-x_r} + (1-\theta)\prod_{r=1}^{R}\beta_r^{1-x_r} \\ &\qquad \times (1-\beta_r)^{x_r}. \end{aligned} \tag{3}$$

Positive and negative predictive values (probability of disease presence (or absence) when the test is positive (or negative)) can be obtained by using Bayes’ theorem.

By letting $\tau_{r|k}$ denote the probability of a positive test r in class k and θ_k denote the probability of a random individual belonging to class k, a generalization of equation 3 to an LCM with K classes is obtained by

$$\begin{aligned} P(X=[x_1,\ldots,x_R]) &= \sum_{k=1}^{K}\theta_k\prod_{r=1}^{R}\tau_{r|k}^{x_r}(1-\tau_{r|k})^{1-x_r}, \\ &\sum_{k=1}^{K}\theta_k = 1. \end{aligned} \tag{4}$$

This model has degrees of freedom, df = $2^R - P - 1$, where P is the number of freely estimated parameters. Note that when $K = 2$, the model in equation 4 is equivalent to the model in equation 3.

Let $n = [n_1, \ldots, n_2R]$ denote the frequencies of 2^R patterns of test results, where $\sum_{s=1}^{2^R} n_s = N$.

The likelihood of the data may be written as

$$\begin{aligned} &L(\boldsymbol{n}|\theta_1,\ldots,\theta_K,\tau_{1|1},\ldots,\tau_{R|1},\ldots,\tau_{R|K}) \\ &\quad \propto \prod_{s=1}^{S}\left[\sum_{k=1}^{K}\theta_k\prod_{r=1}^{R}\tau_{r|k}^{x_r}(1-\tau_{r|k})^{1-x_r}\right]^{n_s}. \end{aligned} \tag{5}$$

Optimization of the likelihood function (e.g., by an expectation-maximization algorithm) yields maximum likelihood estimates of the parameters, and (asymptotic) variances and covariance can be obtained by the inverse of the expected Fisher information matrix. Alternatively, Bayesian optimization techniques (e.g., Gibbs sampling, (23)) and analytical

Downloaded from https://academic.oup.com/aje/article-abstract/179/4/423/128491 by guest on 09 April 2020

solutions have been proposed to obtain the parameter and variance estimates.

APPENDIX 2

Characteristics of Veterinary Publications

We identified 47 applications of LCMs in veterinary publications that evaluated the accuracy of diagnostic tests or disease prevalence. All of these studies targeted infectious diseases, and 34 were published between 2007 and 2011. Interestingly, more than 3 out of 4 ($n = 37$) publications reported Bayesian models. Often, veterinary publications have test observations that are nested in herds varying in infection prevalence. The method introduced by Hui and Walter (42, 43) can be used to estimate herd-level prevalence rates while assuming that the test accuracies are constant across herds. This can be viewed as a special case of a multigroup LCM and was used in 34 of the veterinary studies.

Downloaded from https://academic.oup.com/aje/article-abstract/179/4/423/128491 by guest on 09 April 2020