Journal of Clinical Epidemiology ■ (2015) ■

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

# Problems in detecting misfit of latent class models in diagnostic research without a gold standard were shown

Maarten van Smeden[a,*], Daniel L. Oberski[b], Johannes B. Reitsma[a], Jeroen K. Vermunt[b], Karel G.M. Moons[a], Joris A.H. de Groot[a]

[a]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands
[b]Department of Methodology and Statistics, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands

Accepted 20 November 2015; Published online xxxx

**Abstract**

**Objectives:** The objective of this study was to evaluate the performance of goodness-of-fit testing to detect relevant violations of the assumptions underlying the criticized ''standard'' two-class latent class model. Often used to obtain sensitivity and specificity estimates for diagnostic tests in the absence of a gold reference standard, this model relies on assuming that diagnostic test errors are independent. When this assumption is violated, accuracy estimates may be biased: goodness-of-fit testing is often used to evaluate the assumption and prevent bias.

**Study Design and Setting:** We investigate the performance of goodness-of-fit testing by Monte Carlo simulation. The simulation scenarios are based on three empirical examples.

**Results:** Goodness-of-fit tests lack power to detect relevant misfit of the standard two-class latent class model at sample sizes that are typically found in empirical diagnostic studies. The goodness-of-fit tests that are based on asymptotic theory are not robust to the sparseness of data. A parametric bootstrap procedure improves the evaluation of goodness of fit in the case of sparse data.

**Conclusion:** Our simulation study suggests that relevant violation of the local independence assumption underlying the standard two-class latent class model may remain undetected in empirical diagnostic studies, potentially leading to biased estimates of sensitivity and specificity. 

*Keywords:* Latent class analysis; Local independence assumption; Goodness of fit; Simulation; No gold standard; Sensitivity and specificity

## 1. Introduction

A key step in the evaluation of a diagnostic test (e.g., imaging test, electrophysiology, or biomarker test) is the assessment of its accuracy, commonly measured in terms of sensitivity and specificity. To assess the accuracy of the diagnostic test under study, it is necessary to obtain information on the true target disease status of study subjects that is preferably obtained from a reliable source with perfect accuracy: a gold reference standard. Often, however, the best available reference standard is not completely free of error [1,2]. Using such a reference standard while disregarding these problems leads to a biased assessment of accuracy of the diagnostic (index) test [3−5].

Latent class analysis has been proposed to circumvent this bias [6−9]. The latent class model combines the information from multiple, generally three or more, imperfect diagnostic tests to uncover the unobserved disease structure. This approach has, for example, been used to study the diagnostic value of immunohistochemical assays of bladder tumors [10], to evaluate diagnostic tests to detect visceral leishmaniasis [11,12], to estimate diagnostic accuracy of test for acute maxillary sinusitis [13], and the accuracy of surgeons' classifications of bone fracture types [14].

The standard two-class latent class model that accounts for most applications in diagnostic accuracy and disease prevalence studies [15] relies on making two interrelated assumptions: (1) existence of two classes representing groups of true target disease-positive subjects and true target disease-negative subjects and (2) local independence with respect to the imperfect diagnostic test used in the latent class analysis [16]: the outcomes of the diagnostic test are stochastically independent conditional on class membership. Together, these assumptions have been

Funding: Authors M.v.S., J.B.R., K.G.M.M., and J.A.H.d.G. were supported by NWO, grant number 918.10.615. D.L.O. was supported by NWO, grant number 451.14.017. J.K.V. was supported by NWO, grant number 453-10-002.

* Corresponding author. Tel.: +31-(0)88-756-77-08; fax: +31-(0)88-756-90-28.

*E-mail address*: M.vanSmeden@umcutrecht.nl (M. van Smeden).

http://dx.doi.org/10.1016/j.jclinepi.2015.11.012
0895-4356/

**What is new?**

- Sparseness of data and lack of power of goodness-of-fit tests can hamper the evaluation of latent class model assumptions in realistic diagnostic research scenarios.

**Key findings**

- Relevant violation of the local independence assumption underlying the standard two-class latent class model may remain undetected in empirical diagnostic studies.
- A parametric bootstrap procedure improves the evaluation of goodness of fit in the case of sparse data.

**What this adds to what was known?**

- Our study re-emphasizes the relevance of obtaining an adequate sample size when using latent class analysis.

criticized for being unrealistic for most diagnostic studies (e.g., see [17–20]), potentially leading to severely biased assessments of sensitivity, specificity, and disease prevalence [21–24]. Suggested alternative latent class models that prevent this bias by accounting for dependence in diagnostic test errors have been developed and have found application in more recent literature (for reviews and mathematical underpinnings, see [16,25–27]).

In practice, a justification for the locally independent latent class model is often sought in testing its goodness of fit. No studies to date have examined, however, whether this approach yields sufficient power to detect local dependence and prevent bias at sample sizes typical of diagnostic studies. A recent systematic review [15] of latent class applications in diagnostic accuracy and prevalence studies estimated a median sample size of approximately 350 subjects in such studies. One may therefore question whether the sample sizes of these studies are indeed sufficient to detect relevant deviations from assumptions. Second, although commonly used measures of latent class model fit approach a chi-square distribution under the null hypothesis as the sample size increases, in finite samples, this distribution may not be chi-square [28]. Especially when there is large agreement between diagnostic tests, leading to some combinations of diagnostic test outcomes to be observed only rarely, these test statistics may not approach their theoretical distribution. It is therefore paramount to study the behavior of the model fit test under realistic sample size conditions.

In this article, we study the performance of testing the goodness of fit of latent class models based on asymptotic theory and parametric bootstrap procedures [29]. We study power to detect misfit of the standard two-class latent class model in scenarios where there is a relevant violation of the local independence assumption. We will also study the false rejection rates (type-I error) for scenarios where diagnostic test outcomes are locally dependent. First, we consider the basic theory and assumptions of latent class analysis. We subsequently describe three (large sample) case studies obtained from literature that have presented latent class models for dependent diagnostic test outcomes. The reported results from these publications will be used as the data-generating mechanisms in a Monte Carlo simulation study to evaluate the performance of the goodness-of-fit tests in realistic settings.

## 2. Latent class model

The latent class model for the joint density of diagnostic test outcomes *f(x)* can be written as

$$f(x) = \sum_{d} \pi_d \, g(x \mid d),$$

where $\pi_d = \Pr(D = d)$ is an estimator of the prevalence of disease stratum $d$, and $g(x \mid d)$ is a model for the joint density of diagnostic test outcomes within stratum $d$. In the following, we shall limit our discussion to the common case in which diagnostic binary test data are available on $N$ subjects, taking on the values $x_j = 1$ for a positive test result on test $j$, and $x_j = 0$ when negative.

The two-class latent class model that has become the standard in applications in the field of diagnostic research is based on the assumption that the outcomes of the diagnostic tests, $j = 1,\ldots,J$, are mutually independent given the latent variable. This latent variable is assumed to have two classes, here denoted by $d = 0,1$. Hereafter, we refer to this model by two-class local independence model (in short: two-class LI model) that can be written as,

$$f(x) = \sum_{d=0}^{1} \pi_d \, g(x \mid d),$$

$$g(x \mid d) = \prod_{j=1}^{J} \pi_{x_j|d}^{\;x_j} \left(1 - \pi_{x_j|d}\right)^{1-x_j}.$$

These parameters are estimators of the sensitivities of $J$ diagnostic tests $\pi_{x_j|d=1} = Pr(x_j{=}1 \mid d{=}1)$, the specificities of $J$ diagnostic tests $1 - \pi_{x_j|d=0} = Pr(x_j{=}0 \mid d{=}0)$ and the prevalence of the target disease $\pi_{d=1} = 1 - \pi_{d=0} = Pr(D{=}1)$.

Crucially, the parameters of the latent class model must be identifiable to obtain meaningful estimates [30,31]. For estimating the two-class LI model, data must be available on at least three binary diagnostic tests. The other latent class models we consider require data on at least four (models described in case studies II and III) or five diagnostic tests (model described in case study I).