LATENT CLASS ANALYSIS:
AN INDISPENSABLE METHOD
FOR DIAGNOSTIC ACCURACY
RESEARCH

Nandini Dendukuri

Depart of Medicine,
Department of Epidemiology, Biostatistics and Occupational Health,
McGill University

Technology Assessment Unit, McGill University Health Centre
Outline

■ What are latent class models? Why are they necessary?

■ Why are they not more widely applied in diagnostic research?

■ How can we make them more accessible?
An example from health technology assessment

■ Should the MUHC approve purchase of a urinary antigen (UA) test to diagnose *streptococcus* pneumonia?

■ Pneumonia commonly suspected in hospitalized patients, but rarely confirmed
  - *Standard culture test has poor sensitivity, takes time*

■ Most cases treated empirically with antibiotics
  - *Concern for increased risk of C. difficile diarrhea, antibiotic resistance*
An example from health technology assessment

- An urinary antigen test (UA) with improved sensitivity, better turn around time could aid in choosing targeted antibiotics

- Questions of interest
  - What is the expected increase in true positives? In false positives?
  - Is the addition of the UA test to the routine work-up cost-effective?

- To answer these questions we carried out a systematic review of studies that estimated the sensitivity and specificity of the urinary antigen test
Results of systematic review

- 27 studies identified

- Statistical analysis involves comparing UA test to assorted reference standards
  - Not possible to compare results across studies.
  - Typical of problems where no perfect reference exists

- Most common reference standard is a composite of culture tests

<table>
<thead>
<tr>
<th>Reference class</th>
<th>Reference definition</th>
<th>Plausible range of sensitivity</th>
<th>Plausible range of specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (12 studies)</td>
<td>Blood OR sputum OR respiratory culture positive</td>
<td>40-70%</td>
<td>80-100%</td>
</tr>
<tr>
<td>B (11 studies)</td>
<td>Blood OR sputum culture positive</td>
<td>30-60%</td>
<td>80-100%</td>
</tr>
<tr>
<td>C (4 studies)</td>
<td>Blood culture positive</td>
<td>10-40%</td>
<td>90-100%</td>
</tr>
</tbody>
</table>
Closer look at one study

- Traditional statistical analysis

UA Sensitivity: \( \frac{55}{78} = 70.5\% \)

UA Specificity: \( \frac{224}{305} = 73.4\% \)

<table>
<thead>
<tr>
<th>Composite Reference Standard</th>
<th>Urinary antigen test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>55</td>
<td>78</td>
</tr>
<tr>
<td>-</td>
<td>81</td>
<td>224</td>
</tr>
<tr>
<td>Total</td>
<td>305</td>
<td></td>
</tr>
</tbody>
</table>

Composite reference standard assumed perfect

⇒ UA cannot improve over it

⇒ Cost-effectiveness analysis would never conclude in favor of UA as it is more expensive and less accurate

From Sordé et al., Archives of Int Med, 2011
Closer look at one study

- Traditional statistical analysis

  UA Sensitivity = \( \frac{55}{78} = 70.5\% \)

  UA Specificity = \( \frac{224}{305} = 73.4\% \)

<table>
<thead>
<tr>
<th>Composite</th>
<th>Reference Standard</th>
<th>Urinary antigen test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td></td>
<td>+ 55</td>
<td>78</td>
</tr>
<tr>
<td>-</td>
<td></td>
<td>- 81</td>
<td>305</td>
</tr>
</tbody>
</table>

- Since culture has poor sensitivity, some of the 81 may be true positives
  
  ⇒ *UA Specificity possibly under-estimated*

- The sensitivity could be over- or under-estimated
  
  ⇒ *We cannot use this estimate as an upper or lower bound*

From Sordé et al., Archives of Int Med, 2011
How can we improve over the traditional analysis?

- Clearly, we need to acknowledge sensitivity and specificity of the composite reference standard are not perfect and need to be estimated,
  - i.e. we need a latent class analysis

- Latent class analysis allows us to
  - estimate increase in true positives detected when using UA
  - compare trade-off between true vs false positives on UA

- It includes the prevalence as an unknown parameter
  - Therefore, it allows for comparison and pooling of results across studies identified by the systematic review
A simple latent class model for two tests

\[
\begin{array}{c|cc}
\text{New test (T1)} & + & - \\
\hline
\text{Reference (T2)} & + & n_{11} & n_{01} \\
& - & n_{10} & n_{00}
\end{array}
\]

Assumes each cell is a mixture of disease positive (D+) and disease negative (D-) patients

- Likelihood: \( L \propto \prod_{i=0}^{1} \prod_{j=0}^{1} p_{ij}^{n_{ij}} \)

- Let D denote the latent disease status. The multinomial probabilities can be expressed as

\[
p_{ij} = P(T_1, T_2) = P(T_1, T_2 \mid D+) P(D+) + P(T_1, T_2 \mid D-) P(D-) \]

The terms \( P(T_1, T_2 \mid D) \) can be expressed in different ways leading to different types of latent class models
Modeling $P(T_1, T_2 | D)$

- **Conditional independence (CI) model:**

Assumes $T_1$ and $T_2$ are independent conditional on $D$, e.g.

$$P(T_1 = 1, T_2 = 1) = P(T_1 = 1, T_2 = 1 | D +) P(D +) + P(T_1 = 1, T_2 = 1 | D -) P(D -)$$

$$= P(T_1 = 1 | D +) P(T_2 = 1 | D +) P(D +) + P(T_1 = 1 | D -) P(T_2 = 1 | D -) P(D -)$$

$$= S_1 S_2 \pi + (1 - C_1)(1 - C_2)(1 - \pi)$$

where $S_1$ and $S_2$ are sensitivities, $C_1$ and $C_2$ are the specificities and $\pi$ is the prevalence.

- **Alternatives:**

As the CI model has been criticized for being unrealistic, different approaches have been proposed for allowing $T_1$ and $T_2$ to be dependent.

  - These approaches add more unknown parameters to the model
Model identifiability

- When tests are dichotomous, it is not uncommon to encounter a situation where we have inadequate degrees of freedom.

- Clearly, modeling dependence means we will encounter non-identifiability even when higher numbers of tests are available.

- When the model is non-identifiable, external information will be needed in terms of constraints or prior information.
  - *This makes Bayesian estimation a natural choice for these models.*

<table>
<thead>
<tr>
<th># of tests</th>
<th># of degrees of freedom</th>
<th># of parameters in CI model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>31</td>
<td>11</td>
</tr>
</tbody>
</table>

CI model: Conditional independence model
Returning to the health technology assessment question

Summary ROC curves for urinary antigen test

- We can see that the specificity estimate is higher under the latent class analysis, the sensitivity lies in between.

- Further, we found that the sensitivities of the imperfect reference standards ranged from about 50-60% and specificities were 98-99%.

- These results permitted us to carry out a cost-effectiveness analysis comparing UA to a composite of culture tests.

# Brief history of the use of latent class (LC) modeling

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1968</td>
<td>First introduced by Lazarsfeld and Henry</td>
</tr>
<tr>
<td>1974</td>
<td>Maximum likelihood solution proposed by Goodman in <em>Biometrika</em></td>
</tr>
<tr>
<td>1980</td>
<td>First application in diagnostic research by Hui and Walter, <em>Biometrics</em>&lt;br&gt;Proposed a method for dealing with non-identifiability</td>
</tr>
<tr>
<td>1985</td>
<td>Model for conditional dependence between a pair of tests proposed by Vacek, <em>Biometrics</em></td>
</tr>
<tr>
<td>1996</td>
<td>Modeling conditional dependence between multiple tests using random effects proposed by Qu et al., <em>Biometrics</em></td>
</tr>
<tr>
<td>2000</td>
<td>Modeling conditional dependence in the presence of non-identifiability using a Bayesian approach, Dendukuri &amp; Joseph, <em>Biometrics</em></td>
</tr>
</tbody>
</table>

1990s onwards<br>• models for conditional dependence<br>• checking model assumptions<br>• sample size estimation<br>• correcting verification bias<br>• meta-analysis
A systematic review of LC models in diagnostic research

- van Smeden et al., *Am J Epi*, 2013 identified
  - 69 theoretical papers
  - 64 applied papers in human research + 47 in veterinary sciences

- Shows that applications of LC models are still not common in human diagnostic research even after 3 decades since the publication by Hui & Walter
Beliefs about LC models in the statistics literature

- “It requires that a minimum of three (imperfect) diagnostic tests be measured on every specimen”,
  

- “… the CI assumption often fails in practice … considerable bias can occur when the CI assumption is violated …”
  
Beliefs about LC models in the statistics literature

- “The approach yields estimates that are derived from a black box and are not intuitively well connected with the data”,
  
  Pepe, *The statistical evaluation of medical tests for classification and prediction*, 2011

- “... even in cases when models are identifiable, their estimators may not be robust to the assumed dependence structure, and it may be impossible to distinguish between competing conditional dependence models”
  
  Albert & Dodd, *Biometrics*, 2004
Beliefs about LC models in the medical literature

- “LCA is not designed for hypothesis testing and therefore cannot estimate differences in performance among the three methods, if any exist. Thus, the use of the PIS for comparison purposes is warranted.”
  
  van der Pol et al., *J Clin Micro*, 2012

- “We recommend using the composite reference standard method [over latent class analysis] both for its statistical properties and its relative ease of use.”

Beliefs about LC models in the medical literature

- "... latent class analysis is unlikely to provide more confidence about our understanding of the effectiveness of Xpert MTB/RIF in identifying the presence or absence of true [childhood] tuberculosis disease."
  Dodd and Wilkinson, Lancet, 2013

- "... there is no consensus on the optimal [statistical] approach to evaluating the performance of NAATs [for Chlamydia trachomatis]."
  Centers for Disease Control (CDC), 2014

- "... latent class models, now allow investigators to liberate themselves from the restrictive assumption of a perfect reference test and estimate the accuracy of the candidate tests and the reference standard with the same data."
  World Organisation for Animal Health, 2014
Anticipated advantages of using a composite reference standard (CRS)

- Increased accuracy in disease classification compared to single imperfect reference test
  - Therefore, decreased bias in estimated accuracy of test under evaluation and estimated prevalence

- Avoid incorporation bias because the CRS is independent of the test under evaluation

- Transparency, simplicity
  - Achieve standardization across studies
Unanswered questions about composite reference standards

- Does increasing the number of component tests improve the CRS?

- What is the impact of ‘conditional dependence’ between the test under evaluation and the CRS?

- How do changes in the underlying prevalence affect estimates?
Bias due to OR-rule composite reference standard

- When component tests have perfect specificity line
  - Estimate of new test’s sensitivity unbiased (i.e. red line falls on the dashed line)
  - Bias in estimate of new test’s specificity decreases with each added test, eventually becoming unbiased
Bias due to OR-rule composite reference standard

- However, if component tests have 98% specificity
  - Sensitivity estimate of new test is biased (yellow line), with bias increasing with every component test
Bias due to OR-rule composite reference standard

- If the component tests have 98% specificity and also make the same errors as the test under evaluation
  
  - The specificity of the new test is overestimated (green line)
Further, sensitivity and specificity estimates can vary across settings because they depend on the disease prevalence
In summary:

- Problems with composite reference standard more apparent when we examine the impact of increasing the number of tests
  
  - Unless specificity of component tests is perfect, new test’s sensitivity is underestimated
  
  - When conditional dependence is present, new test’s specificity is overestimated
  
  - Bias worsens with increasing number of component tests!

- Not what is expected of a sound statistical method
Bias due to composite reference standards

- Other types of composite reference standards (e.g. based on an AND rule) also have similar problems.

- We also found that CRS based estimates are not comparable across studies, because they are functions of the underlying disease prevalence.

- Poor performance of the CRS can be explained by the fact that it makes sub-optimal use of the data:
  - It makes a simplistic classification
  - And then it ignores the uncertainty in that classification
How can we improve over the CRS?

- We need an approach that
  - Uses the complete cross-tabulation between all imperfect tests without simplifying it
  - An approach that models conditional dependence
  - An approach that includes the prevalence as an unknown parameter

i.e. we need latent class analysis!
Returning to latent class models

- What are the challenges in estimating latent class models?
  - Interpreting the latent disease status
  - Selecting the appropriate conditional dependence structure
  - Dealing with non-identifiability
An illustrative example: Childhood Pulmonary Tuberculosis (TB)

- Diagnosis of childhood pulmonary TB relies on multiple tests/signs as no single measure is considered adequate:
  - Microbiological tests (e.g. Culture, Xpert)
  - Symptoms/signs of TB
  - Chest radiograph
  - Immunologic evidence of TB (e.g. tuberculin skin test (TST))
  - Contact with TB patient

- A consequence is that there are no reliable estimates for the burden of childhood pulmonary TB, despite it being a major public health problem
Clinical Case Definitions for Classification of Intrathoracic Tuberculosis in Children: An Update

Goal: “... enhance harmonized classification ... across studies, resulting in greater comparability and the much-needed ability to pool study results.”
Latent class model for childhood pulmonary TB
Latent class model for childhood pulmonary TB

- We had data from a cohort of 749 children hospitalized with suspected pulmonary TB in South Africa.

- A heuristic model was set up to explain how the observed data relate to the latent variables.

- Importantly, both clinicians and methodologists were involved in this exercise.

Heuristic Model
Heuristic Model

- 3 possible latent variables were identified
  - *Active TB disease*
  - *Exposure to TB (latent TB)*
  - *Other respiratory disease*

- Combinations of these latent variables would lead to four possible latent classes. Of these two (Active TB, Not active TB) were considered relevant and distinguishable with the available tests.

- Conditional dependence is anticipated between 4 of the tests.

- Covariates Age, HIV and Malnutrition affect model parameters.

- The preferred model was defined at the outset rather than by relying on statistical criteria.
Modeling conditional dependence

- Culture, Xpert and Smear are all influenced by bacillary load
  - They could all be false negative for the same group of children who have a low bacillary load, leading to a high positive dependence

- The TST test is expected to be negatively correlated with the severity of infection (which is also affected by the bacillary load)

Wang et al, Stats in Med, 2016
Random effect used to model conditional dependence
## Results

<table>
<thead>
<tr>
<th>Test and Parameter</th>
<th>Conditional Independence Model</th>
<th>Model Adjusting for conditional dependence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Posterior Median Estimate</td>
<td>95% CrI</td>
</tr>
<tr>
<td>CPTB Prevalence</td>
<td>16.6</td>
<td>15.6 – 18.0</td>
</tr>
<tr>
<td>Culture</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>96.7</td>
<td>87.8 – 99.8</td>
</tr>
<tr>
<td>Specificity</td>
<td>99.8</td>
<td>98.9 – 100.0</td>
</tr>
<tr>
<td>Xpert</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>74.4</td>
<td>66.0 – 82.2</td>
</tr>
<tr>
<td>Specificity</td>
<td>98.3</td>
<td>97.0 – 99.4</td>
</tr>
<tr>
<td>Microscopy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>33.3</td>
<td>25.3 – 42.1</td>
</tr>
<tr>
<td>Specificity</td>
<td>99.8</td>
<td>99.2 – 100.0</td>
</tr>
<tr>
<td>Radiography</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>65.4</td>
<td>56.5 – 73.8</td>
</tr>
<tr>
<td>Specificity</td>
<td>73.1</td>
<td>69.6 – 76.6</td>
</tr>
<tr>
<td>TST</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>69.0</td>
<td>60.5 – 76.7</td>
</tr>
<tr>
<td>Specificity</td>
<td>62.4</td>
<td>58.5 – 66.1</td>
</tr>
<tr>
<td>Test outcome pattern</td>
<td>Observed frequency</td>
<td>Predicted probability of TB</td>
</tr>
<tr>
<td>----------------------</td>
<td>--------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Cu Xp Mi Ra TS</td>
<td>%</td>
<td>95% Crl</td>
</tr>
<tr>
<td>0 0 0 0 0</td>
<td>296</td>
<td>2 0 – 7</td>
</tr>
<tr>
<td>0 0 0 0 1</td>
<td>149</td>
<td>16 5 – 33</td>
</tr>
<tr>
<td>0 0 0 1 0</td>
<td>87</td>
<td>9 0 – 34</td>
</tr>
<tr>
<td>0 0 0 1 1</td>
<td>78</td>
<td>52 26 – 74</td>
</tr>
<tr>
<td>0 0 1 0 1</td>
<td>1</td>
<td>11 0 – 100</td>
</tr>
<tr>
<td>0 1 0 0 0</td>
<td>5</td>
<td>4 0 – 40</td>
</tr>
<tr>
<td>0 1 0 0 1</td>
<td>7</td>
<td>56 0 – 100</td>
</tr>
<tr>
<td>0 1 0 1 0</td>
<td>2</td>
<td>12 0 – 100</td>
</tr>
<tr>
<td>0 1 0 1 1</td>
<td>2</td>
<td>88 50 – 100</td>
</tr>
<tr>
<td>1 0 0 0 0</td>
<td>3</td>
<td>23 0 – 100</td>
</tr>
<tr>
<td>1 0 0 0 1</td>
<td>8</td>
<td>93 62 – 100</td>
</tr>
<tr>
<td>1 0 0 1 0</td>
<td>1</td>
<td>54 0 – 100</td>
</tr>
<tr>
<td>1 0 0 1 1</td>
<td>20</td>
<td>99 90 – 100</td>
</tr>
<tr>
<td>1 1 0 0 0</td>
<td>1</td>
<td>100 100 – 100</td>
</tr>
<tr>
<td>1 1 0 0 1</td>
<td>17</td>
<td>100 100 – 100</td>
</tr>
<tr>
<td>1 1 0 1 0</td>
<td>4</td>
<td>100 100 – 100</td>
</tr>
<tr>
<td>1 1 0 1 1</td>
<td>27</td>
<td>100 100 – 100</td>
</tr>
<tr>
<td>1 1 1 0 0</td>
<td>8</td>
<td>100 100 – 100</td>
</tr>
<tr>
<td>1 1 1 0 1</td>
<td>5</td>
<td>100 100 – 100</td>
</tr>
<tr>
<td>1 1 1 1 0</td>
<td>21</td>
<td>100 100 – 100</td>
</tr>
<tr>
<td>1 1 1 1 1</td>
<td>7</td>
<td>100 100 – 100</td>
</tr>
</tbody>
</table>
Predicted probabilities

- Examining the predicted probabilities is another way to see how the observed data relates to the latent disease status

- This is another advantage of latent class analysis over descriptive classification methods like cluster analysis
Probability child was treated increased with probability of CPTB

- Without a perfect reference, we can only evaluate the face validity of our latent class model.
- An estimated 95.5% of TB positive children receive anti-TB treatment.
- An estimated 45.8% of TB negative children receive anti-TB treatment.
Future research

- Fit the CPTB model in other datasets drawn from other settings
  - Settings where the prevalence of active TB is lower may lead to other choices for latent classes
  - Use datasets where more variables are recorded

- Develop robust latent class models for other disease areas

- Develop prediction models that can help optimize diagnosis?

- Several interesting methodological questions remain to be answered!
References


