Supplementary materials: On the robustness of latent class models for diagnostic testing with no gold standard

Matthew R. Schofield[1]*, Michael J. Maze[2,3], John A. Crump[4], Matthew P. Rubach[5], Renee Galloway[6], and Katrina J. Sharples[1]

[1]Department of Mathematics and Statistics, University of Otago, New Zealand.

[2]Centre for International Health, University of Otago, New Zealand.

[3]Department of Medicine, University of Otago, Christchurch, New Zealand.

[4]Department of Preventive and Social Medicine, University of Otago, New Zealand.

[5]Division of Infectious Diseases, Duke University, Durham, NC, USA.

[6]Bacterial Special Pathogens Branch, US Center for Disease Control and Prevention, Atlanta, GA, USA.

March 1, 2021

1 Model for dependence

To include dependence between the the Test-It *Leptospira* and Leptocheck WB tests we include an interaction within a log-linear model. We describe the model for $M = 2$ classes.

*E-mail: matthew.schofield@otago.ac.nz.

In particular, we assume that the definition of $\boldsymbol{q} = (q_1, \ldots, q_S)$ becomes

$$q_k = \sum_{c=1}^{2} \pi_c p^*_{ck}, \quad k = 1, \ldots, S.$$

The term p^*_{ck} is defined by a multinomial logit model, where we treat the Sth combination of test results $(0 \ldots 0)$ as the reference,

$$\log\left(\frac{p^*_{ck}}{p^*_{cS}}\right) = \sum_{j=1}^{T} \alpha_{cj} x_{kj} + \gamma_c x_{k3} x_{k4}, \quad c = 1, 2, \quad k = 1, \ldots, S-1.$$

The terms γ_1 and γ_2 describe the dependence between tests 3 and 4 (Test-It *Leptospira* and Leptocheck WB). The sensitivities and specificities are shown in Figure 1. The posterior median for π_1, the prevalence of disease, was 0.11.

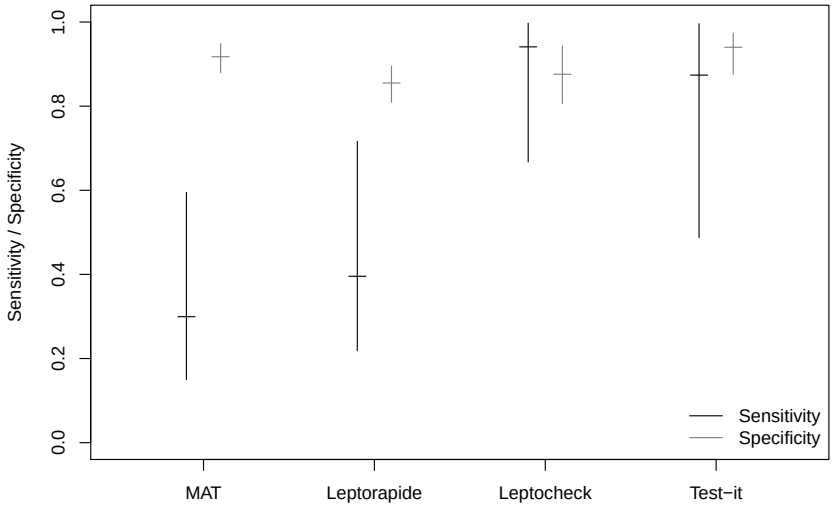

Figure 1: The vertical lines represent the central 90% credible intervals for sensitivity (black) and specificity (grey) for each of the four tests for the Tanzania data using the latent class model, allowing for dependence between Test-It *Leptospira* and Leptocheck WB. The horizontal line represents the median of the posterior distribution.

2 Pairwise posterior predictive checks

To assess lack-of-fit with respect to undiagnosed dependence we compare the observed pairwise counts to their posterior predictive distribution. Pairwise counts are defined as the number of patients that tested positive any given pair of tests. This approach is in the spirit of Qu et al. (1996) who define a correlation residual between any given pair of tests. The observed pairwise counts appear to be consistent to their respective posterior predictive distributions (Figure 2).

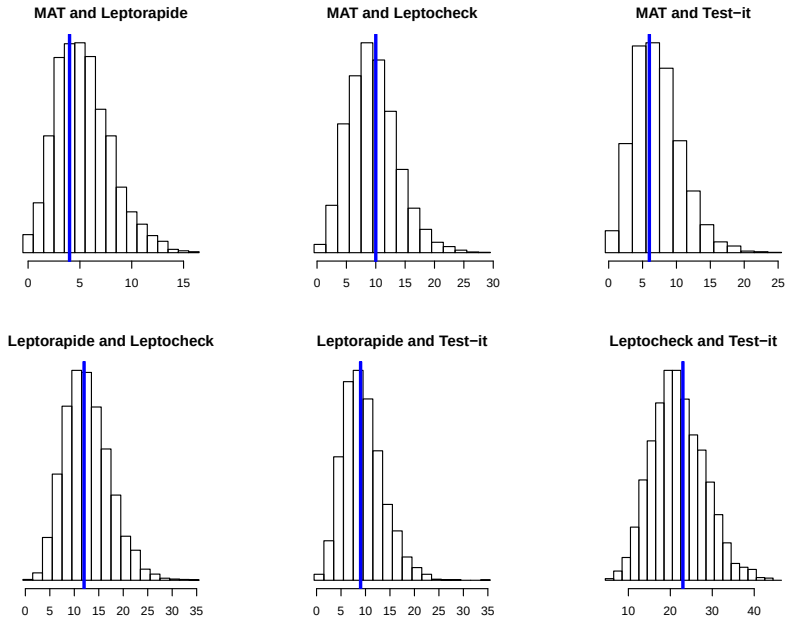

Figure 2: Posterior predictive distributions for the pairwise counts. Pairwise counts are defined as the number of patients that tested positive for both tests specified in the title of the plot. The vertical blue line gives the observed pairwise counts.

3 Label switching

To account for label switching, we use the algorithm of Stephens (2000) as implemented in the R package `label.switching` (Papastamoulis 2016). This algorithm is run after the

MCMC sampling is complete and assigns labels in order to minimize KL-divergence.

For the motivating data, we implement the algorithm on the full 30 000 draw from the posterior (three chains of 10 000 draws).

For the simulations, we thin the posterior samples. This increases the speed of the label switching algorithm and allows the simulation to be run in a reasonable time.

4 Additional simulation

A further simulation was performed at different parameter values (Table 1). They reflect a situation where all four tests have sensitivities ranging 0.50 to 0.90 for state 1. The sensitivities (state 2) range from 0.88 to 0.99. If an individual is in state 3, the probability of a positive test ranges from 0.20 to 0.6. The parameters used for the two-state and three-state model lead to similar expected counts (Table 2).

As with the simulation in the manuscript, we generate 1000 data sets under each the following four scenarios:

1. True model has two-states; sample size is $N = 225$.
2. True model has two-states; sample size is $N = 1000$.
3. True model has three-states; sample size is $N = 225$.
4. True model has three-states; sample size is $N = 1000$.

We follow the same process for fitting and summarizing the simulation results as in the manuscript. In particular, we find (i) the bias of the point estimates (Figure 3 and Table 3); (ii) the coverage of the interval estimates (Table 3); and (iii) the proportion of models that were assessed as ill-fitting (Table 4).

Parameter	$M = 2$	$M = 3$
p_{11}	0.63	0.80
p_{21}	0.57	0.56
p_{31}	0.40	0.50
p_{41}	0.72	0.90
p_{12}	0.01	0.01
p_{22}	0.06	0.02
p_{32}	0.05	0.05
p_{42}	0.12	0.12
p_{13}	–	0.23
p_{23}	–	0.60
p_{33}	–	0.20
p_{43}	–	0.30
π_1	0.14	0.08
π_2	0.86	0.80
π_3	–	0.12

Table 1: True parameter values for two-state ($M = 2$) simulation and the three-state ($M = 3$) simulation. The first two classes represent the disease of interest and no disease, respectively. The third class represents individuals with an alternate disease that triggers a response at a rate higher than the 'no disease' state. The value p_{jc} represents the probability of testing positive for test j when in class c and π_c is the probability of being in class c.

Test 1	Test 2	Test 3	Test 4	Expected count $M = 2$	Expected count $M = 3$
1	1	1	1	3.3	3.9
1	1	1	0	1.3	0.9
1	1	0	1	4.9	4.5
1	1	0	0	2.0	2.5
1	0	1	1	2.5	3.0
1	0	1	0	1.0	0.7
1	0	0	1	3.9	3.6
1	0	0	0	3.0	3.2
0	1	1	1	2.0	1.7
0	1	1	0	1.2	2.0
0	1	0	1	4.2	4.3
0	1	0	0	10.7	10.1
0	0	1	1	2.5	2.3
0	0	1	0	8.5	8.9
0	0	0	1	22.7	22.6
0	0	0	0	151.4	150.7

Table 2: The expected counts for the two true simulation models when $N = 225$. The expected counts are similar between the two-state ($M = 2$) and three-state ($M = 3$) models.

The broad conclusions are the same as with the simulation in the manuscript. The estimates are close to unbiased and coverage near nominal when data are simulated and fitted under a two-state latent class model (Figure 3 and Table 3). In contrast, many of the parameters, including sensitivities, specificities and prevalence are biased with very low coverage when data are simulated under the three-state model where we assume disease corresponds to state 1 (Figure 3 and Table 3).

We are unable to routinely identify the model misspecification with standard goodness-of-assessment approaches (Table 4), with nearly 90% of simulations showing no evidence of lack of fit when $N = 225$. That number is more than 70% when $N = 1000$.

As in the manuscript, we again consider the alternative definition that disease is a combination of state 1 and state 3. The bias is non-negligible and the coverage rates are poor, particularly when $N = 1000$ (Figure 4 and Table 5).

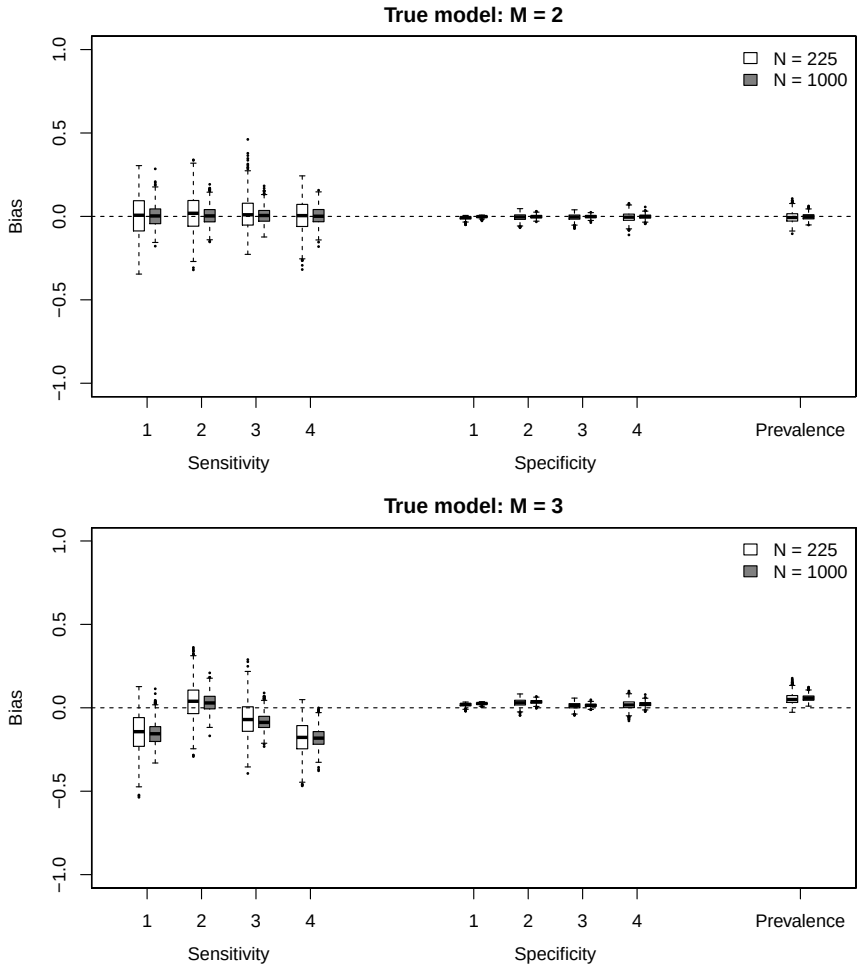

Figure 3: The difference between the true sensitivity/specificity and estimated sensitivity/specificity in each of the 1000 simulations when the true model had two-states ($M = 2$) and three-states ($M = 3$).

	$M = 2$				$M = 3$			
	Bias		**Coverage**		**Bias**		**Coverage**	
	$N = 225$	$N = 1000$	$N = 225$	$N = 1000$	$N = 225$	$N = 1000$	$N = 225$	$N = 1000$
sens_1	0.003	0.004	0.91	0.91	-0.147	-0.155	0.73	0.25
sens_2	0.017	0.005	0.92	0.91	0.038	0.031	0.90	0.86
sens_3	0.018	0.004	0.90	0.90	-0.066	-0.084	0.85	0.49
sens_4	0.003	0.003	0.93	0.91	-0.177	-0.182	0.46	0.03
spec_1	-0.009	-0.002	0.93	0.92	0.018	0.025	0.72	0.04
spec_2	-0.005	-0.001	0.90	0.90	0.030	0.036	0.63	0.07
spec_3	-0.005	-0.001	0.90	0.90	0.012	0.014	0.85	0.53
spec_4	-0.005	-0.001	0.89	0.89	0.017	0.022	0.85	0.50
π_1	-0.005	-0.002	0.92	0.90	0.054	0.059	0.49	0.02

Table 3: The bias and coverage of sensitivity and specificity for each of the four tests across the various simulation strategies. The value of M denotes the number of states in the true model. The labels sens_i and spec_i represent the sensitivity and specificity of test i, respectively. The posterior median was used for estimation and the coverage is based off a 90% central credible interval.

$\boldsymbol{M}$	$\boldsymbol{N}$	$\boldsymbol{\Pr(P_B < 0.05)}$	$\boldsymbol{\Pr(P_B^* < 0.05)}$
2	225	0.01	0.05
2	1000	0.00	0.05
3	225	0.01	0.06
3	1000	0.08	0.28

Table 4: The probability of rejecting the Bayesian p-value (P_B) as well as the calibrated Bayesian p-value (P_B^*) in the simulation. The true model is indexed based on the number of latent states (M) and the sample size is given by N.

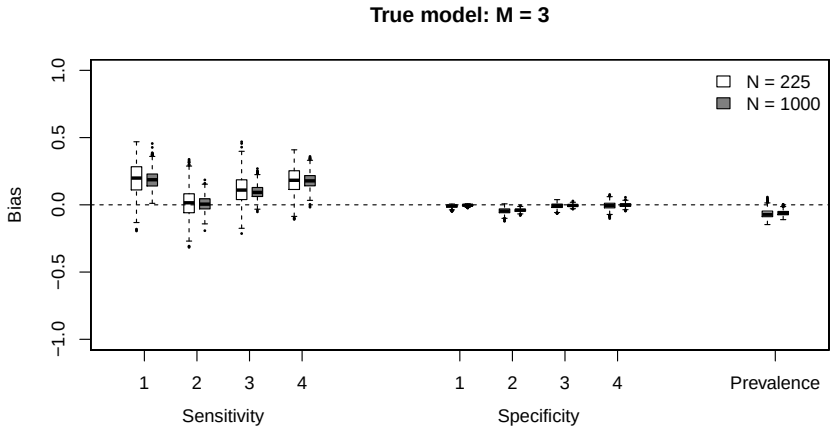

Figure 4: The difference between the true sensitivity/specificity and estimated sensitivity/specificity of the 1000 simulations when the true model has three-states ($M = 3$). We assess the sensitivity of the test to diagnose any disease (either state 1 or state 3).

	Bias		Coverage	
	$N = 225$	$N = 1000$	$N = 225$	$N = 1000$
sens_1	0.195	0.187	0.58	0.11
sens_2	0.014	0.007	0.93	0.90
sens_3	0.114	0.096	0.71	0.39
sens_4	0.183	0.178	0.51	0.07
spec_1	-0.010	-0.003	0.89	0.90
spec_2	-0.045	-0.040	0.26	0.01
spec_3	-0.008	-0.005	0.88	0.85
spec_4	-0.006	-0.002	0.90	0.88
π_1	-0.066	-0.061	0.50	0.10

Table 5: The bias and coverage of sensitivity and specificity when assessing the sensitivity of the test to diagnose any disease (either state 1 or state 3) when the three state model is true. The labels sens_i and spec_i represent the sensitivity and specificity of test i, respectively. The posterior median was used for estimation and the coverage is based off a 90% central credible interval.

References

Papastamoulis, P. (2016), "`label.switching`: An R package for dealing with the label switching problem in MCMC outputs," *Journal of Statistical Software*, 69.

Qu, Y., Tan, M., and Kutner, M. H. (1996), "Random effects models in latent class analysis for evaluating accuracy of diagnostic testss," *Biometrics*, 52, 797–810.

Stephens, M. (2000), "Dealing with label switching in mixture models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 795–809.