Received: 23 October 2015 | Revised: 19 March 2017 | Accepted: 23 March 2017

DOI: 10.1002/jrsm.1243

ORIGINAL ARTICLE

WILEY Research Synthesis Methods

Evaluating the accuracy and economic value of a new test in the absence of a perfect reference test

Xuanqian Xie[1,2] | Alison Sinclair[1] | Nandini Dendukuri[1,3]

[1] Technology Assessment Unit, McGill University Health Centre, Montréal, Quebec, Canada

[2] Toronto Health Economics and Technology Assessment Collaborative, Leslie Dan Pharmacy, University of Toronto, Toronto, Ontario, Canada

[3] Departments of Medicine and Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Quebec, Canada

Correspondence
Nandini Dendukuri, Technology Assessment Unit, McGill University Health Centre, Department of Medicine, McGill University, 5252 Maisonneuve West, 3F.50 Montreal, Quebec H4A 3S5
Email: nandini.dendukuri@mcgill.ca

Background: *Streptococcus pneumoniae* (SP) pneumonia is often treated empirically as diagnosis is challenging because of the lack of a perfect test. Using BinaxNOW-SP, a urinary antigen test, as an add-on to standard cultures may not only increase diagnostic yield but also increase costs.

Objective: To estimate the sensitivity and specificity of BinaxNOW-SP and subsequently estimate the cost-effectiveness of adding BinaxNOW-SP to the diagnostic work-up.

Design: We fit a Bayesian latent-class meta-analysis model to obtain estimates of BinaxNOW-SP accuracy that adjust for the imperfect accuracy of culture. Meta-analysis results were combined with information on prevalence of SP pneumonia to estimate the number of patients who are correctly classified under competing diagnostic strategies. Taking into consideration the cost of antibiotics, we determined the incremental cost of adding BinaxNOW-SP to the work-up per case correctly diagnosed.

Results: The BinaxNOW-SP test had a pooled sensitivity of 0.74 (95% credible interval [CrI], 0.67-0.83) and a pooled specificity of 0.96 (95% CrI, 0.92-0.99). An overall increase in diagnostic accuracy of 6.2% due to the addition of BinaxNOW-SP corresponded to an incremental cost per case correctly classified of $582 Canadian dollars.

Conclusions: The methods we have described allow us to evaluate the accuracy and economic value of a new test in the absence of a perfect reference test using an evidence-based approach.

KEYWORDS
Bayesian latent class meta-analysis model, composite decision rule, conditional dependence, cost-effectiveness, diagnostic accuracy

1 | INTRODUCTION

For a number of medical conditions, a perfect diagnostic test (ie, one that is able to determine the disease status without error) does not exist. In the case of *Streptococcus pneumoniae* (SP), the most commonly identified causative organism in cases of community acquired pneumonia (CAP), etiologic diagnosis depends on culture from blood or respiratory samples. While these tests have near perfect specificity, they are believed to give false negative results for as many as 70% of cases,[1-3] meaning that most patients suspected of SP never receive a confirmed diagnosis. As a result, they are treated empirically with broad spectrum antibiotics, which increase the risk of antibiotic resistance and

Financial support: None.

nosocomial *Clostridium difficile* diarrhea. In an effort to improve the poor diagnostic yield of culture, it has been suggested that a urinary antigen test (BinaxNOW-SP), which has higher sensitivity, be performed simultaneously, and that a patient should be classified as positive for SP pneumonia if either test is positive. In principle, early identification of a case of SP pneumonia could result in antibiotic therapy being better targeted, ie, use of narrow spectrum antibiotics.

Evaluation of diagnostic accuracy or cost-effectiveness of new tests for CAP is complicated because of the lack of a perfect reference test. In particular, if the test being evaluated has higher sensitivity and/or specificity than the reference test, and the standard test is naively assumed to be a perfect reference test, then the additional patients correctly classified by the new test would be erroneously treated as false positive or false negative when the standard test fails. If the new test is also more expensive than the reference, as is often the case for recently developed tests, then a cost-effectiveness analysis to compare the new test with the standard test would always favour the imperfect reference.[4,5] For this reason, for diseases where there is no perfect reference test, cost-effectiveness analyses comparing a new test to an imperfect standard test use subjective estimates for the sensitivity and specificity of the reference test.[6,7] Typically, both the point estimates of these parameters and any uncertainty around them are determined subjectively.

In order to be valid, inputs to a cost-effectiveness analysis should be drawn from the best possible evidence, which is widely considered to be provided by a systematic review of efficacy or effectiveness of health interventions.[8] In the context of diagnostic test evaluation where a perfect reference test is available, meta-analysis results can be directly used as inputs into a cost-effectiveness analysis via a Bayesian approach.[8-11] In this article, we illustrate how a similar approach can be taken even in the absence of a perfect test, via the use of a latent class meta-analysis model.

Latent class models have been used to estimate the sensitivity and specificity of a new test in the absence of a gold-standard reference test in individual studies,[12] their advantage being that they avoid the unrealistic assumption that the standard test has perfect sensitivity and specificity, while at the same time avoiding reliance on fixed guess values for sensitivity and specificity of the standard test. Recently, latent class models have also been used in the context of a meta-analysis.[13-16] These models are particularly relevant in a meta-analysis setting when different reference tests are used for each included study. In addition to estimating the sensitivity and specificity of the test under evaluation, these models can also be used to obtain estimates of the sensitivity and specificity of the reference test(s) and the true disease prevalence.

Our objective is to illustrate the statistical methods involved in estimating the diagnostic accuracy and cost-effectiveness of a new diagnostic test in the absence of a perfect reference test. We illustrate our methods via application to the problem of determining the cost-effectiveness of a urinary pneumococcal antigen test for streptococcus pneumonia.

2 | METHODS

We will begin by describing the Bayesian latent class meta-analysis model. We then present the economic decision model describing the diagnostic testing strategies to be compared. Next we describe how the incremental value of adding BinaxNOW-SP to the work-up can be estimated based on the results from the meta-analysis. Finally, we provide the inputs for the economic analysis model.

2.1 | Diagnostic meta-analysis model in the absence of a gold-standard reference

A recent systematic review identified 27 studies that had reported 2-by-2 tables comparing BinaxNOW-SP to different culture tests.[17] The sensitivity and specificity of culture tests, which are typically developed in-house, are expected to vary across studies. Therefore, we felt a random effects model was suitable to pool these parameters. As BinaxNOW-SP is a commercial test, we expect a certain level of standardization across studies. None the less, because the accuracy of the test may be affected by patient characteristics (eg, prior antibiotic use) and study design, we anticipate that it may vary across studies. Therefore, for the index test (ie, the test under evaluation) as well, a random effects structure for the sensitivity and specificity is preferred. We also report on the results of fitting a fixed effects model to examine the extent of the improvement in model fit due to allowing for between-study variation in BinaxNOW-SP's performance.

2.1.1 | Notation

We denote the dichotomous index test as T_{1j} and the dichotomous reference test as T_{2j}, in the jth study, $j = 1,\ldots,J$. A positive result on either test is denoted by 1 and negative result by 0. We denote the observed result in the jth study by the vector $T_{12j} = (t_{11j}, t_{10j}, t_{01j}, t_{00j})$, whose elements t_{uvj} $(u, v = 0, 1)$ denote the number of patients with results $T_{1j} = u$ and $T_{2j} = v$. The sample size of the jth study is given by $N_j = t_{11j} + t_{10j} + t_{01j} + t_{00j}$. Both the index test and reference test are assumed to be imperfect measures of a common underlying dichotomous latent variable D, the true disease status. The reference test may be composite in some studies, as in our motivating example.[14] Let D denote the true disease status such that $D = 1$ and $D = 0$ denote disease positive and disease negative, respectively. The sensitivities of the tests in

the jth study is given by $S_{pj} = P(T_{pj} = 1 \mid D = 1)$ and specificities by $C_{pj} = P(T_{pj} = 1 \mid D = 1)$, $p = 1,2$.

The data in each study are assumed to arise from a hierarchical Bayesian meta-analysis model as described below. The meta-analysis model allowed for both within-study and between-study variability in the accuracy parameters.

2.1.2 | Within-study level

Within each study, the vector T_{12j} follows a multinomial distribution:

$$\left(t_{11j}, t_{10j}, t_{01j}, t_{00j}\right) \sim Multinomial\left(\left(p_{11j}, p_{10j}, p_{01j}, p_{00j}\right), N_j\right), j = 1, \ldots, J, \text{ where}$$

$$p_{uvj} = \pi_j \, {S_{1j}}^u \left(1-S_{1j}\right)^{1-u} \, {S_{2j}}^v \left(1-S_{2j}\right)^{1-v} + \left(1-\pi_j\right) \left(1-C_{1j}\right)^u {C_{1j}}^{1-u} \left(1-C_{2j}\right)^v {C_{2j}}^{1-v}, u, v = 0, 1$$

2.1.3 | Between-study level

The sensitivities and specificities in each study are assumed to come from a common population distribution. For both the index and the reference test, we used the structure commonly referred to in diagnostic meta-analysis as the bivariate model.[18,19]

$$\begin{pmatrix} \log\left(\frac{S_{pj}}{1-S_{pj}}\right) \\ \log\left(\frac{C_{pj}}{1-C_{pj}}\right) \end{pmatrix} \sim \left(\begin{pmatrix} \mu_{Sp} \\ \mu_{Cp} \end{pmatrix}, \begin{pmatrix} \sigma^2_{Sp} & \rho_p \sigma_{Sp} \sigma_{Cp} \\ \rho_p \sigma_{Sp} \sigma_{Cp} & \sigma^2_{Cp} \end{pmatrix} \right),$$

$$p = 1, 2, j = 1, \ldots, J,$$

where μ_{Sp} is the pooled logit sensitivity of T_p, μ_{Cp} is the pooled logit specificity of T_p, σ_{Sp} is the between-study variance in the logit sensitivity of T_p, σ_{Cp} is the between-study variance in the logit specificity of T_p, and ρ_p is the correlation between the logit sensitivity and logit specificity of T_p across studies. To study the impact of ignoring the between-study variance in accuracy of BinaxNOW-SP, a fixed effects model for index test accuracy was fit by setting the parameters σ_{S1} and σ_{C1} to zero.

To keep the notation simple, we let $p = 2$ denote the reference test. In fact, in our application, we had 3 separate reference tests: (1) reference standard type A, a composite of blood culture, sputum (smear or culture), and culture of any other respiratory sample; (2) reference standard type B, a composite of a blood culture and sputum (smear or culture); (3) reference standard type C, a blood culture alone. Accordingly, separate hierarchical prior distributions were used for the accuracy of each of the 3 reference tests. For all 3 reference tests, we set the correlation (ρ_2) between the logit sensitivity and logit specificity across studies equal to 0 as the specificity of culture tests is believed to be consistently high in all studies. The bivariate meta-analysis model structure has been shown to be algebraically equivalent to the hierarchical summary receiver operating characteristic curve structure,[20] which was used in our earlier application to the same data.[17]

2.1.4 | Prior distributions

To estimate the model using a Bayesian approach, we used low information prior distributions for all parameters so as to allow the observed data to dominate the final inferences. The prevalence in each study is estimated separately using prior distributions $\pi_j \sim$ Uniform(0,1). The parameters μ_{Sp} and μ_{Cp} follow normal prior distributions with mean 0 and variance 100. The parameters σ_{Sp} and σ_{Cp} follow gamma prior distributions with a shape parameter of 2 and a rate parameter of 0.5. The parameter ρ_1 follows a U(−1,1) distribution. If historical information or subjective knowledge is available about some parameters, eg, the sensitivity and specificity of a reference test, then these may be incorporated via informative prior distributions for μ_{Sp} and μ_{Cp}.

2.1.5 | Estimation

The parameters of the meta-analysis model were estimated using a Markov Chain Monte Carlo method implemented via OpenBUGS software.[21] The OpenBUGS program is provided in Appendix. Three separate chains were run with different initial values. Initial values were dispersed but limited to a plausible region when possible. For example, the specificities of both index and reference tests in our application are believed to be very high. Therefore, we used initial values for the pooled specificity parameters that were dispersed in the region from 0.8 to 1. The posterior distribution in a latent class model is in fact multimodal, but there is only 1 of these modes that is meaningful while the remaining are referred to as "mirror solutions." By selecting initial values in a plausible range, we were able to ensure that convergence of the Markov Chain Monte Carlo process to the "mirror solutions" was unlikely.[14] The different initial values we used all converged to the same meaningful solution for our dataset. Convergence was evaluated using the Brooks-Gelman-Rubin statistic that is available within OpenBUGS. When we say below that we sampled from the posterior distribution, we mean we sampled from the posterior mode that provided the solution of interest.

2.1.6 | Model comparison

The fixed and random effects models were compared in terms of the residual deviance and the deviance information criterion (DIC) statistic.[22] A DIC that is 5 to 10 points less would be indicative of a substantially better model. The residual deviance, DIC, and pD statistics were calculated based on

the multinomial probabilities rather than on the sensitivities, specificities, and prevalence parameters. This approach is more appropriate for the type of mixture model we have here.[23] We also calculated the contribution to the residual deviance from individual studies to identify any studies that are outliers.

2.1.7 | Statistics reported from the meta-analysis

For each of the pooled sensitivities and specificities, we extracted the following statistics from the joint posterior distribution: mean, median, equal-tailed 95% credible interval, 95% prediction interval. The posterior means and prediction intervals were used in the cost-effectiveness analyses as explained later.

2.1.8 | Adjusting for conditional dependence

The meta-analysis model can be extended to adjust for possible conditional dependence (ie, unexplained correlation between the index and reference tests) within each latent class in the jth study. Separate covariance parameters would be needed for every study within the 2 disease groups. The multinomial probability in each study would be modified as follows:

$$\begin{aligned} p_{uvj} = {} & \pi_j \left(S_{1j}{}^{u}\left(1-S_{1j}\right)^{1-u} S_{2j}{}^{v}\left(1-S_{2j}\right)^{1-v}\right. \\ & \left.+(-1)^{((1-u)v+u(1-v))}\,\mathrm{covp}_j\right) \\ & +\left(1-\pi_j\right)\left(\left(1-C_{1j}\right)^{u}C_{1j}{}^{1-u}\left(1-C_{2j}\right)^{v}C_{2j}{}^{1\text{-}v}\right. \\ & \left.+(-1)^{((1-u)v+u(1-v))}\,\mathrm{covn}_j\right), \end{aligned}$$

$$\text{where} -\left(1-S_{1j}\right)\left(1-S_{2j}\right) \le \mathrm{covp}_{ij} \le \min\left(S_{1j}, S_{2j}\right) - S_{1j}S_{2j} \text{ and}$$

$$-\left(1-C_{1j}\right)\left(1-C_{2j}\right) \le \mathrm{covn}_{ij} \le \min\left(C_{1j}, C_{2j}\right) - C_{1j}C_{2j}.$$

However, the addition of these covariance parameters would render the model nonidentifiable, and a meaningful solution cannot be obtained unless informative prior distributions were used for at least 4 of the parameters in each study, eg, the sensitivity and specificity of the index and reference tests. This is because each study provides only 3 degrees of freedom, but the number of parameters to be estimated in each study exceeds this.[24]

Because we did not wish to use informative priors, particularly on the accuracy parameters of the index test, we had to work with a simplified, constrained model. First, we reasoned that the impact of the conditional dependence between the specificities would be negligible as the specificities of both tests are high. Therefore, we constrained the covariance between the specificities (covn_j) to 0. Then, we set covp_j to a fixed percentage of its maximum value. This way, we were able to gauge the impact of conditional dependence indirectly without increasing the number of parameters in the model (see program in the Appendix). We compared competing meta-analysis models with different percentages of conditional dependence between the sensitivities using the residual deviance and DIC.

2.2 | Decision analytical model and inputs

When comparing BinaxNOW-SP to culture, it must be noted that BinaxNOW-SP cannot completely replace culture because culture results, when available, also provide information on the sensitivity of the cultured organism to antibiotics. Therefore, we compared the strategy of using a composite test of BinaxNOW-SP and culture versus the strategy of using culture alone for the detection of *S pneumonia*.

Figure 1 shows the economic decision model. Suspected CAP patients are tested by BinaxNOW-SP and have cultures drawn as soon as they present. We assume culture results would be available 48 hours later, while urine antigen results would be available 1 hour after sampling. By culture, we mean the equivalent of reference test type A in our meta-analysis, ie, a composite of blood culture, sputum (smear or culture), and culture of any other respiratory sample such that at least 1 positive is considered test positive. Based on our meta-analysis, the prediction intervals for the accuracies of all 3 tests were very similar (data not shown). Therefore, we did not compare strategies based on different culture tests in our economic analysis.

In the culture alone arm (ie, the single test decision rule arm), all patients are treated by empirical therapy for the first 2 days. From day 3, culture positive patients switch to targeted treatment, and culture negative patients continue on empirical treatment. In the BinaxNOW-SP and culture arm (ie, the composite test decision rule arm), patients who test positive for SP by either BinaxNOW-SP or culture are treated with targeted therapy. The literature on the influence of BinaxNOW-SP on clinical outcomes is sparse. Therefore, we defined cost-effectiveness in terms of the incremental cost per case correctly classified by adding BinaxNOW-SP to cultures as our primary outcome. We considered the incremental cost per *C difficile* infection (CDI) avoided as a secondary outcome. The time horizon was 3 days for the number of correctly classified patients and 1 month for CDI avoided. Given the short time frame of both outcomes, we did not consider discounting of costs.

2.2.1 | Probability of classification based on each decision rule

The effectiveness of each diagnostic testing strategy, ie, the probability of correct classification, is determined using the

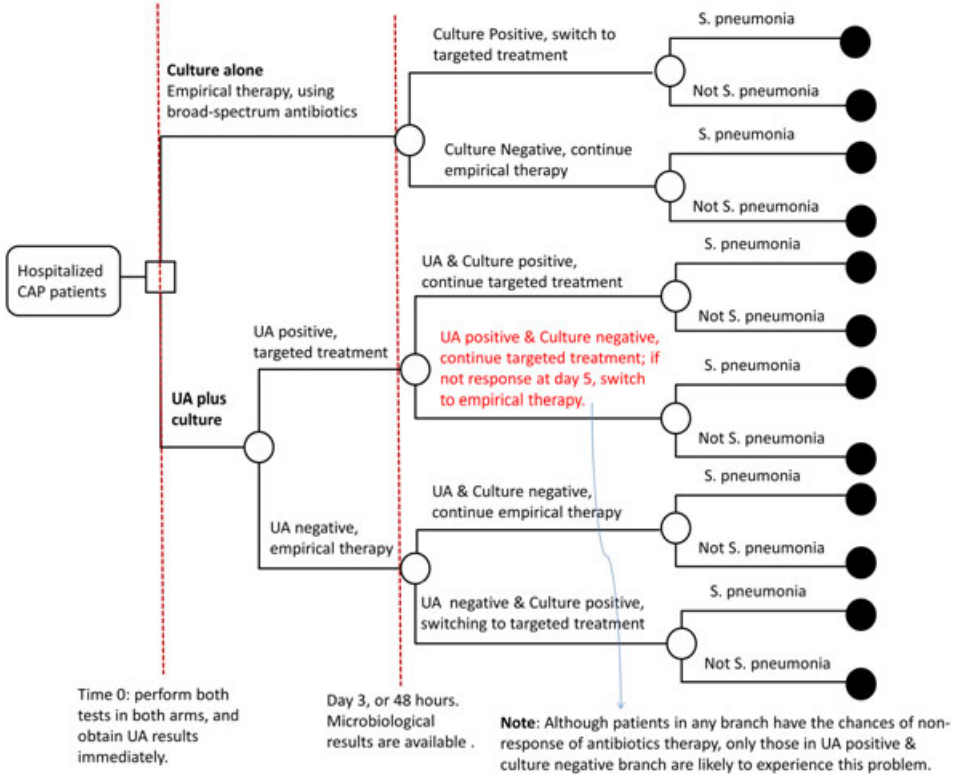

FIGURE 1 Decision analytical model for BinaxNOW-SP plus cultures versus cultures alone. Abbreviations: CAP, community acquired pneumonia; UA, urine antigen test [Colour figure can be viewed at wileyonlinelibrary.com]

pooled estimates of sensitivity (denoted S_2) and specificity (C_2) of the reference test (T_2, culture in our application), the sensitivity (S_1) and specificity (C_1) of the index test (T_1, BinaxNOW-SP in our application) together with the prevalence (π). The expressions for the probabilities of falling in each terminal node in Figure 1 is given in the Appendix in Table A1. Values of these probabilities are calculated using the meta-analysis output by sampling from the 95% posterior prediction intervals of the sensitivities and specificities. We chose to sample from the prediction interval rather than the credible interval as our cost-effectiveness analyses apply to an individual study setting in the future.

The results of BinaxNOW-SP and culture will be combined using a disjunctive positivity criterion (ie, a rule by which a patient is classified as disease positive when either test is positive). This composite rule will gain sensitivity but lose specificity compared to a decision rule based on culture alone. In Appendix Table A2, we present the expressions for the sensitivity and specificity of the composite decision rule and the incremental proportion of correct classification of using the composite rule versus the reference test alone. Expressions are also provided for the conjunctive positivity criterion (ie, a rule by which a patient is classified as disease positive only when both tests are positive) for the interested reader.

2.2.2 | Other inputs to the economic model

Estimates of disease prevalence varied considerably in the studies included in our meta-analysis with the posterior means distributed roughly uniformly from 0.1 to 0.66. Therefore, we did not feel the results of our meta-analysis could provide a useful estimate for the prevalence. Based on subjective input from a clinical expert at our hospital centre, we assumed that the true prevalence of SP pneumonia among inpatients was 30%. However, given the lack of a definitive diagnostic method, this estimate cannot be verified. Therefore, in sensitivity analyses, we considered other values of the prevalence as explained below. In the primary cost-effectiveness analysis, we considered only costs for antibiotic treatment of pneumonia and cost of BinaxNOW-SP. Please see the Appendix for details of the secondary analysis where we considered *C difficile* diarrhea as an outcome.

The cost of antibiotics (in Canadian dollars) were obtained from the Department of Pharmacy and Therapeutics at the McGill University Hospital Centre (MUHC), a university-affiliated tertiary care centre in Montreal, Canada (Table 1). Following MUHC guidelines for use of antibiotics for pneumonia, we assumed that ceftriaxone plus azithromycin are used for empirical treatment. Following the Infectious Diseases Society of America/American

TABLE 1 Inputs to the economic model

	Value	Reference
Effectiveness		
Sensitivity of BinaxNOW-SP test, %	Mean = 72.9 (prediction interval 67.1, 82.7)	Latent-class meta-analysis
Specificity of BinaxNOW-SP test, %	Mean = 95.7 (prediction interval 84.8. 99.6)	Latent-class meta-analysis
Sensitivity of culture test, %	Mean = 61.9 (prediction interval 17.8, 91.5)	Latent-class meta-analysis
Specificity of culture test, %	Mean = 97.4 (prediction interval 86.4, 99.9)	Latent-class meta-analysis
Covariance between sensitivities	Fixed percentage of maximum covariance	Latent-class meta-analysis
True prevalence of *S Pneumonia*	0.3 (sensitivity analysis 0.02, 0.8)	Estimate
Nonresponse rate of antibiotics used in UA positive and culture negative branch	Uniform distribution (0.05, 0.15)	Estimate
Days of taking antibiotics	Uniform distribution (7, 14)	MUHC guidelines
Cost (Canadian dollar in 2011)		
BinaxNOW-SP test	38 per test (fixed value)	MUHC
Ceftriaxone (IV 2 g q 24 h)	2.60 per day (fixed value)	MUHC
Moxifloxacin (IV 400 mg q 24 h)	25.13 per day (fixed value)	MUHC
Penicillin G (IV 8-12 million units/day)	2.11 per day (fixed value)	MUHC
Azithromycin (oral 500 mg q day × 7 days)	1.20 per day (fixed value)	MUHC

Abbreviations: CDI, *Clostridium Difficile* infection; MUHC, McGill University Health Centre; SD, standard deviation; UA, urine antigen.

Thoracic Society Consensus Guidelines (2007),[2] we assumed that penicillin G is the preferred antibiotic for targeted treatment of SP. Because the model is limited to inpatients, both decision arms were assumed to receive the same routine care. Thus, for simplicity, we do not account for other costs, such as the nursing time for intravenous administration of antibiotics, which would be the same in both arms.

2.2.3 | Cost-effectiveness analysis

To compare the 2 testing strategies—testing with culture alone (strategy 1) vs testing with culture and BinaxNOW-SP (strategy 2)—we estimated the incremental cost-effectiveness ratio (ICER) for our 2 outcomes of interest as follows:

$$ICER = \frac{E(Cost\ of\ strategy\ 2) - E(Cost\ of\ strategy\ 1)}{E(Effectiveness\ of\ strategy\ 2) - E(Effectiveness\ of\ strategy\ 1)},$$

where E() denotes the expected value or mean value, which is the preferred statistic for cost-effectiveness analyses.

We used Monte Carlo simulations to capture the uncertainty of inputs to the model. We randomly drew 10 000 sets of the predictive sensitivity and specificity values of the BinaxNOW-SP and culture tests from the best fitting meta-analysis model, as well as 10 000 sets from the distributions of other parameters given in Table 1. The ICERs were calculated by estimating the expected costs and effectiveness values across these 10 000 sets.

To study the robustness of our estimates of effectiveness and cost-effectiveness, we performed sensitivity analyses using alternative estimates for the fixed prevalence of SP pneumonia ranging from 0.02 to 0.8. First, we estimated the change in the incremental probability of correct classification with the change in prevalence of SP. We then used a net monetary benefit (NMB) analysis to determine cost-effectiveness at the different fixed values of the prevalence. The net monetary benefit was calculated using the following formula: NMB = WTP × ΔP −ΔCost, where WTP denotes the "willingness to pay per percentage of correct classification," ΔP denotes the difference in percentage of correct classification between the 2 strategies and ΔCost denotes the difference in cost.

A positive NMB indicates that the addition of BinaxNOW-SP to culture is cost-effective; if NMB is a negative value, the addition of BinaxNOW-SP to culture is considered not cost-effective. Because there is no generally accepted threshold of willingness to pay for one additional correct classification, we used thresholds of $200, $500, and $1,000 in this analysis.

We also performed a sensitivity analysis to study the impact of the conditional dependence between the sensitivities of the 2 tests. This was done by drawing 10 000 sets of sensitivity and specificity estimates obtained from different meta-analysis models that adjusted for different percentages of conditional dependence (see Appendix Table A3).

The economic evaluation was conducted from the perspective of the MUHC. All costs are expressed in 2011 Canadian dollars. We conducted the economic analysis using SAS 9.4.[25]

3 | RESULTS

3.1 | Meta-analysis

The data from the 27 studies included in the meta-analysis are displayed in Table 2, and the results of the fixed and random effects meta-analysis models that assume conditional independence between the tests appear in Table 3. The total residual deviance (Table 3) is clearly lower for the random effects model as is the DIC, leading us to select this model. Based on this model, we can see that the pooled sensitivity of BinaxNOW-SP is clearly improved compared to those of the 3 reference tests. However, its pooled specificity is slightly lower. Figure 2 summarizes the pooled sensitivity and specificity of BinaxNOW-SP together with 95% credible and prediction regions. This figure shows that there is some heterogeneity in the sensitivity of BinaxNOW-SP across the studies, but the specificity was generally high in all studies. It is interesting to note from Table 3 that the reference type A, which was based on a composite of a greater number of culture tests, had a higher pooled sensitivity as expected, followed by reference type B and type C. Also, the pooled specificity of reference type A is also the lowest, followed by reference type B and type C.

The individual contributions of each study to the residual deviance also show that the random effects model decreases these contributions for many studies (Table 2). Because each study contributes 3 independent data points, we can expect

TABLE 2 Data from the 27 studies included in the meta-analysis and residual deviance contributions of each study

Study	Reference Standard	Binax + Reference+	Binax + Reference−	Binax − Reference+	Binax − Reference−	Contribution to Residual Deviance	
						Fixed Effects Model	Random Effects Model
1	A	55	81	23	224	2.61	2.80
2	A	48	30	11	179	3.79	3.09
3	A	11	24	5	118	2.48	2.52
4	A	16	15	9	207	2.41	2.49
5	A	23	11	9	65	2.19	2.49
6	A	63	52	20	214	2.58	2.80
7	A	44	8	38	125	4.36	3.79
8	A	27	41	14	138	2.76	2.72
9	A	25	42	15	65	3.63	2.98
10	A	75	34	45	244	3.36	2.93
11	A	23	14	5	49	2.96	2.80
12	A	21	56	12	658	2.74	2.75
13	B	37	58	28	762	2.82	2.84
14	B	26	26	17	193	2.87	2.70
15	B	3	6	0	50	5.14	4.95
16	B	20	24	3	109	5.13	4.16
17	B	8	1	3	47	4.27	3.98
18	B	10	5	10	67	3.12	2.89
19	B	9	11	5	42	2.31	2.35
20	B	7	25	17	103	3.59	3.87
21	B	14	1	4	85	7.43	6.08
22	B	33	87	39	325	3.19	3.15
23	B	3	15	5	103	3.04	2.79
24	C	17	16	10	126	2.89	2.61
25	C	27	54	9	239	2.56	2.84
26	C	11	27	4	54	2.22	2.50
27	C	51	23	8	77	5.98	3.68

TABLE 3 Results of meta-analysis

	Fixed Effects Model		Random Effects Model	
	Median	**95% Credible Interval**	**Median**	**95% Credible Interval**
BinaxNOW-SP				
Pooled sensitivity	0.78	(0.72, 0.86)	0.74	(0.67, 0.83)
Pooled specificity	0.96	(0.93, 0.99)	0.96	(0.92, 0.99)
Between-study standard deviation				
In logit sensitivity		–	0.47	(0.30, 0.77)
In logit specificity		–	0.55	(0.30, 1.10)
Correlation[a]		–	0.11	(−0.88, 0.91)
Reference test A				
Pooled sensitivity	0.60	(0.45, 0.76)	0.62	(0.45, 0.80)
Pooled specificity	0.97	(0.93, 0.99)	0.99	(0.95, 0.998)
Reference test B				
Pooled sensitivity	0.56	(0.36, 0.79)	0.56	(0.37, 0.86)
Pooled specificity	0.96	(0.93, 0.99)	0.97	(0.94, 0.99)
Reference test C				
Pooled sensitivity	0.50	(0.25, 0.75)	0.51	(0.26, 0.79)
Pooled specificity	0.98	(0.91, 0.998)	0.98	(0.91, 0.998)
Model comparison statistics				
Dbar[b]	92.35		85.49	
DIC	516.4		511.5	
pD	65.2		67.9	

[a]Between logit sensitivity and logit specificity.

[b]Mean residual deviance compared to 81 data points.

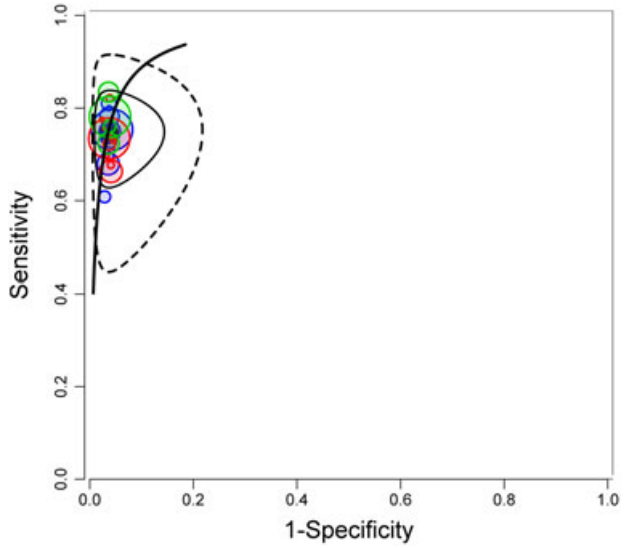

FIGURE 2 Pooled sensitivity and specificity of BinaxNOW-SP together with 95% credible and prediction regions, and a summary receiver-operating characteristic curve. Circles represent individual studies. Radius proportional to sample size. Blue: reference type A, red: reference type B, green: reference type C

the contribution of each study to be around 3 and consider values above 6 to be high.[23] There was only 1 such study (study 21). We can see that those studies that had fewer BinaxNOW-SP positive than culture positive subjects and those studies with small numbers of subjects in one or more cells of the 2-by-2 table had higher residuals.

The comparison of meta-analysis models allowing for different percentages of conditional dependence is summarized in Appendix Table A3. The addition of a constrained covariance term did not improve the DIC or residual deviance. Therefore, we retained the random effects model assuming conditional independence between the tests for the primary economic analysis.

3.2 | Economic analysis

Table 4 shows the results of the primary cost-effectiveness analysis in terms of the cost per patient correctly classified. Compared with culture alone, BinaxNOW-SP plus culture significantly improves the overall sensitivity by identifying an additional 915 of 3002 SP patients. Conversely, because a

TABLE 4 Cost-effectiveness analysis comparing diagnosis based on BinaxNOW-SP plus culture versus culture alone (base case)

	SP Patients: N = 3002	Non-SP Patients: N = 6998	All CAP Patients: N = 10 000	Incremental N (%) of correct classification	Cost per Patient, $	Incremental Cost per Patient, $	Incremental Cost per Case Correctly Classified
	N/% correctly classified	N/% correctly classified	N/% correctly classified	N/% correctly classified			
Culture alone	1746/58.2	6815/97.4	8562/85.6	–	33.8	–	–
BinaxNOW-SP plus culture tests	2662/88.7	6519/93.2	9181/91.8	619/6.2	69.8	36	582

Abbreviations: CAP, community acquired pneumonia; SP, *Streptococcus pneumoniae*.

positive result on either test is considered to be SP, addition of BinaxNOW-SP reduces the specificity because of the false-positive classification of 296 of 6998 non-SP patients. Thus, the overall accuracy (patients correctly classified as either SP or non-SP) increased by a mean (SD) value of 6.2%. The increased accuracy corresponds to an incremental cost per patient of $36, and the incremental cost per case correctly classified of $582.

We estimated that the addition of BinaxNOW-SP would help avoid roughly 12 CDI cases per 10 000 inpatients, with an incremental cost of $33.9 per patient. The total incremental cost per CDI avoided (ICER) is $29 234. See Table A5 in the Appendix for details.

3.3 | Sensitivity analyses

3.3.1 | Prevalence of SP and cost of empirical antibiotics therapy

We found that the incremental probability of correct classification increases with the prevalence of SP over the range explored, a trend resulting from the greater sensitivity of BinaxNOW-SP plus culture over culture (see Figure 3A). However, the prevalence of SP has a negligible impact on incremental cost, as the whole sale price of purchasing generic antibiotics for both targeted therapy and empirical therapy at the MUHC was low and relatively similar, meaning the incremental cost is largely driven by the cost of BinaxNOW of $38 per test.

Figure 3B presents the results of the net monetary benefit analysis. In brief, the NMB increases with the greater prevalence of SP. The 3 willingness-to-pay (WTP) lines intersected at a prevalence of SP of 12%. At this prevalence, the incremental probability of correct classification was zero, and only the cost of BinaxNOW-SP contributes to the estimates of NMB. Compared with culture alone, the combination of BinaxNOW-SP plus culture would be cost-effective (ie, NMB values would be positive) when the prevalence of SP is greater than 62%, 33%, and 23% at the WTP per correct classification of $200, $500, and $1000, respectively.

The cost of antibiotics impacts the incremental cost significantly. If the cost of empirical therapy increases to $28 per day from $3.8 per day, while the cost of targeted therapy remains constant, then the overall cost of both testing strategies becomes the same.

3.3.2 | Conditional dependence model

The results of the economic analysis when using the sensitivity and specificity estimates from meta-analysis models with different covariance between the sensitivities are given in Appendix Table A4. In brief, when including the covariances between the 2 tests in the meta-analysis models, the predicted sensitivities of both tests were reduced. The overall incremental proportion correctly classified by adding BinaxNow Test ranged from 5.8% to 7.3% for models with different levels of covariance, while the incremental cost was almost the same in all models. There was no clear tendency of the ICER to change with the level of conditional dependence. The incremental cost per case correctly classified ranged from $497 to $631.

4 | DISCUSSION

4.1 | Diagnostic accuracy and economic evaluation of a new test in the absence of a gold standard

Evaluation of a new diagnostic test in the absence of a gold-standard reference remains a challenging problem in health technology assessment. We have described an application integrating the results of a latent class meta-analysis model into a cost-effectiveness analysis. The advantage of this approach is that instead of using subjective plug-in values for the sensitivity and specificity of the new test and the standard test, we are able to account for the uncertainty in these values as determined by a meta-analysis. Our latent class meta-analysis model resulted in a significantly higher estimate of BinaxNOW-SP specificity, compared with a meta-analysis model that treated culture as perfect (specificity

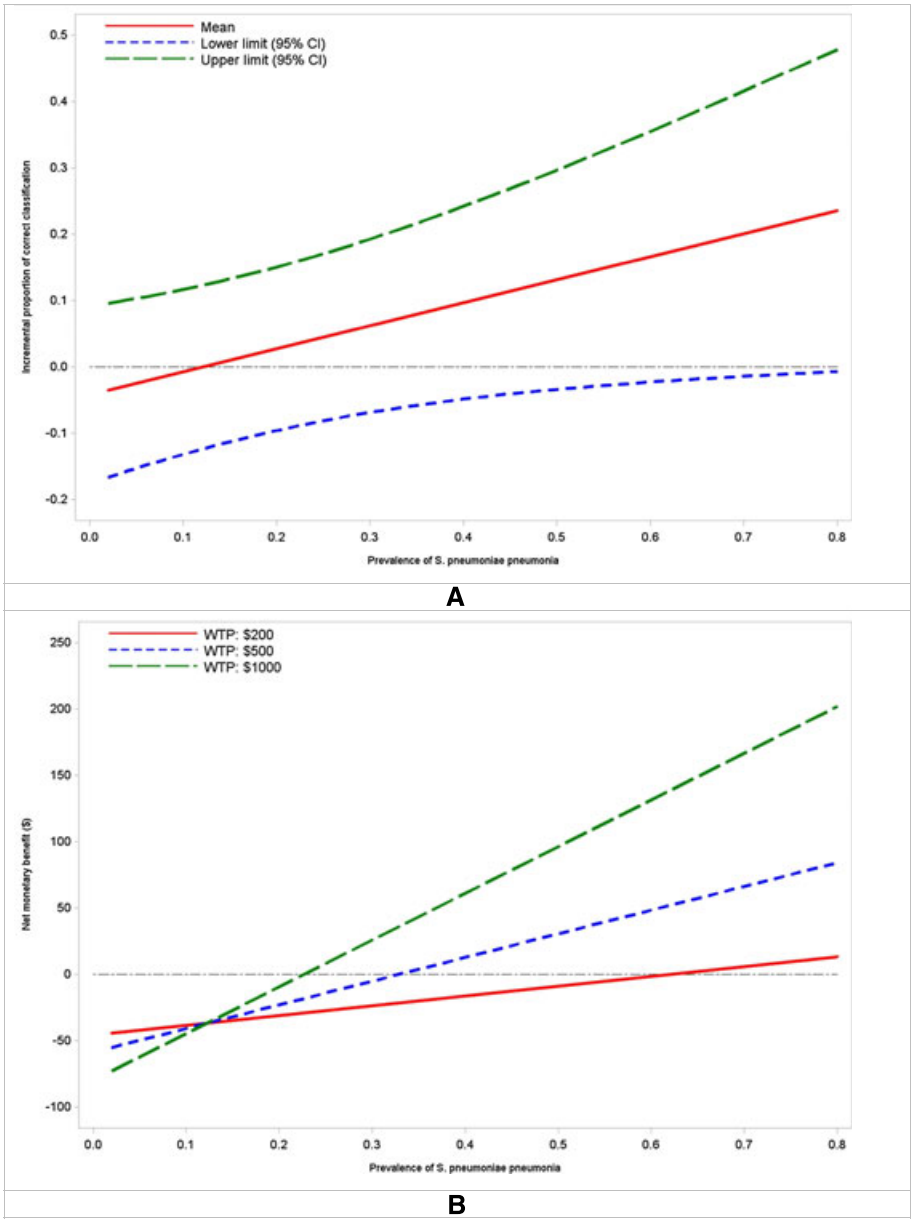

FIGURE 3 A, Relation between incremental proportion of correctly classified patients and prevalence of SP pneumonia. B, Relation between net monetary benefit and prevalence of SP pneumonia. Abbreviations: CI, confidence interval; WTP, willingness-to-pay per percentage of patients correctly classified [Colour figure can be viewed at wileyonlinelibrary.com]

[95% CrI]: 85.6% [81.7%, 88.9%]).[17] This is because patients previously classified as false positives (BinaxNOW-SP+, Culture–) under the standard analysis had the chance of being classified as true positives under the latent class model.

Compared with the classic economic model to determine treatment strategies and corresponding clinical and economic outcomes following true positive (TP), false negative (FN), false positive (FP), and true negative (TN) results,[8,26] we modeled the patient pathway considering the timing of testing and treatment adjustment. All patients are treated empirically in the absence of any test results. Treatment may then be adjusted as the first and second test's results become available, and further adjustment may be made by monitoring the patients' responses.

4.2 | Percentage of correct classification versus clinical outcomes

Clinical outcomes, frequently represented by quality adjusted life years, are preferred in health economic evaluations of therapeutic interventions. However, it is not evident that they should be the outcome of choice in the economic evaluation of a diagnostic test. Randomized clinical trials (RCTs) have shown that improved accuracy of diagnostic tests is not always consistent with better clinical outcomes.[27] This may be in part because of the difficulty in performing such RCTs because of the very high sample size needed to achieve adequate power. It may also be due to unsuccessful

implementation of the treatment protocol despite improved diagnostic classification. Two RCTs have studied the efficacy of empirical treatment versus treatment adjusted according to the results of BinaxNOW-SP in CAP.[28,29] Neither found a significant difference in clinical outcomes or adverse events between groups.

Given the difficulty in performing an RCT to study the impact of BinaxNOW-SP on clinical outcomes, a modeling-based approach relying on indirect information may be useful in projecting the possible outcome of such an RCT. Therefore, we performed a secondary analysis to estimate the cost-effectiveness in terms of cost per CDI avoided. We found that targeted therapy can lead to slightly reduced CDI risk. However, we found that the incremental cost of a testing strategy involving BinaxNOW-SP would be very high resulting in a very high ICER unless there is a significant background risk of CDI.

Ultimately, the interest in curbing antibiotic use is to lessen risk of *C difficile* diarrhea and antibiotic resistance at the institutional level. However, we did not attempt to model these outcomes as it would have required a very complex model and a number of unverifiable assumptions.

4.3 | Trading off gain vs loss in a composite decision rule

As illustrated in our application, compared with decision rules based on a single imperfect test, composite decision rules based on multiple imperfect tests do not necessarily improve overall diagnostic accuracy. The composite test strategy may have advantages in specific situations, such as when we wish to have high sensitivity to maximize case detection, or alternatively when we want to maximize positive predictive value to avoid harm from unnecessary treatment. However, as we have shown, this gain in sensitivity is accompanied by a loss of specificity. The methods we have illustrated can help to quantify the gain and loss of a composite versus a single test alone to support patient management. This is crucial for this particular problem where a gain in identification of SP cases is the desired target, but it must be traded off with the loss of treating a non-SP case too conservatively with a targeted treatment.

4.4 | Incremental value of a diagnostic test

In recent decades, there has been considerable interest in estimating the incremental value of a diagnostic or screening test rather than a limited focus on its sensitivity and specificity.[30] As our application illustrates, when considering the impact of adding a test to the work up, its incremental value is ultimately what determines its cost-effectiveness. Whereas we have described a composite decision rule based on the test results themselves, in other situations, it is possible that predictive values obtained from the latent class model may be used.[31] In other settings where multiple tests are used, an additional uncertainty may be the optimal sequence of testing needed to minimize the cost of testing.

5 | CONCLUSIONS

In conclusion, we have illustrated how to evaluate the diagnostic accuracy and economic value of a new test in the absence of a gold-standard test using an evidence-based approach. This model may be further extended to include covariates, alternative outcomes, and additional imperfect tests.

REFERENCE

1. Ishida T, Hashimoto T, Arita M, Tojo Y, Tachibana H, Jinnai M. A 3-year prospective study of a urinary antigen-detection test for Streptococcus pneumoniae in community-acquired pneumonia: utility and clinical impact on the reported etiology. *J Infect Chemother*. 2004;10:359-363.
2. Mandell LA, Wunderink RG, Anzueto A, et al. Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. *Clin Infect Dis*. 2007;44:S27-S72.
3. Selickman J, Paxos M, File TM, Seltzer R, Bonilla H. Performance measure of urinary antigen in patients with Streptococcus pneumoniae bacteremia. *Diagn Microbiol Infect Dis*. 2010;67:129-133.
4. Dowdy DW, O'Brien MA, Bishai D. Cost-effectiveness of novel diagnostic tools for the diagnosis of tuberculosis. *Int J Tuberc Lung Dis*. 2008;12:1021-1029.
5. Kondagunta GV. Balancing innovation with cost in diagnostic testing. *Am J Manag Care*. 2015;21:SP425
6. Felder S, Mayrhofer T. *Medical Decision Making: A Health Economic Primer*. 1st ed. Verlag Berlin Heidelberg: Springer; 2011.
7. Oxlade O, Schwartzman K, Menzies D. Interferon-gamma release assays and TB screening in high-income countries: a cost-effectiveness analysis. *Int J Tuberc Lung Dis*. 2007;11:16-26.
8. Sutton AJ, Cooper NJ, Goodacre S, Stevenson M. Integration of meta-analysis and economic decision modeling for evaluating diagnostic tests. *Med Decis Making*. 2008;28:650-667.
9. Dendukuri N, Khetani K, McIsaac M, Brophy J. Testing for HER2-positive breast cancer: a systematic review and cost-effectiveness analysis. *Can Med Assoc J*. 2007;176:1429-1434.
10. Novielli N, Cooper NJ, Sutton AJ. Evaluating the cost-effectiveness of diagnostic tests in combination: is it important to allow for performance dependency? *Value Health*. 2013a;16:536-541.
11. Novielli N, Sutton AJ, Cooper NJ. Meta-analysis of the accuracy of two diagnostic tests used in combination: application to the ddimer test and the wells score for the diagnosis of deep vein thrombosis. *Value Health*. 2013b;16:619-628.
12. Hadgu A, Dendukuri N, Hilden J. Evaluation of nucleic acid amplification tests in the absence of a perfect gold-standard test: a review of the statistical and epidemiologic issues. *Epidemiology*. 2005;16:604-612.

13. Chu H, Chen S, Louis TA. Random effects models in a meta-analysis of the accuracy of two diagnostic tests without a gold standard. *J Am Stat Assoc*. 2009;104:512-523.

14. Dendukuri N, Schiller I, Joseph L, Pai M. Bayesian meta-analysis of the accuracy of a test for tuberculous pleuritis in the absence of a gold standard reference. *Biometrics*. 2012;68:1285-1293.

15. Menten J, Boelaert M, Lesaffre E. Bayesian meta-analysis of diagnostic tests allowing for imperfect reference standards. *Stat Med*. 2013;32:5398-5413.

16. Sadatsafavi M, Shahidi N, Marra F, et al. A statistical method was used for the meta-analysis of tests for latent TB in the absence of a gold standard, combining random-effect and latent-class methods to estimate test accuracy. *J Clin Epidemiol*. 2010;63:257-269.

17. Sinclair A, Xie X, Teltscher M, Dendukuri N. Systematic review and meta-analysis of a urine-based pneumococcal antigen test for diagnosis of community-acquired pneumonia caused by Streptococcus pneumoniae. *J Clin Microbiol*. 2013;51:2303-2310.

18. Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: analysing and presenting results. In: Deeks J, Bossuyt PM, Gatsonis C, eds. *Handbook for Diagnsotic Test Accuracy Reviews Version 1.0*. Oxford, UK: The Cochrane Collaboration; 2010:1-47 Available from http://srdta.cochrane.org/.

19. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol*. 2005;58:982-990.

20. Liu Y, Chen Y, Chu H. A unification of models for meta-analysis of diagnostic accuracy studies without a gold standard. *Biometrics*. 2015;71(2):538-547.

21. OpenBUGS Foundation. OpenBUGS v3.2.3. 2014. http://www.openbugs.net

22. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. *J R Stat Soc Ser B*. 2002;64:583-639.

23. Dias S, Welton NJ, Sutton AJ, Ades AE. last updated September 2016. *Technical Support Document 2: A Generalised Linear Modelling Framework for Pairwise and Network Meta-Analysis of Randomised Controlled Trials*. Sheffield (UK): Decision Support Unit (DSU) of the National Institute for Health and Clinical Excellence (NICE), 2011. Available from: http://www.nicedsu.org.uk/TSD2%20General%20meta%20analysis%20corrected%202Sep2016v2.pdf [8 December 2016].

24. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics*. 2001;57:158-167.

25. SAS Institute Inc. *Base SAS® 9.4*. Cary, NC: SAS Institute Inc.; 2015.

26. Koffijberg H, van Zaane B, Moons KG. From accuracy to patient outcome and cost-effectiveness evaluations of diagnostic tests and biomarkers: an exemplary modelling study. *BMC Med Res Methodol*. 2013;13:12

27. Lord SJ, Staub LP, Bossuyt PM, Irwig LM. Target practice: choosing target conditions for test accuracy studies that are relevant to clinical practice. *Br Med J*. 2011;343:d4684

28. Falguera M, Ruiz-Gonzalez A, Schoenenberger JA, et al. Prospective, randomised study to compare empirical treatment versus targeted treatment on the basis of the urine antigen results in hospitalised patients with community-acquired pneumonia. *Thorax*. 2010;65:101-106.

29. Van Der Eerden MM, Vlaspolder F, De Graaff CS, Groot T, Jansen HM, Boersma WG. Value of intensive diagnostic microbiological investigation in low- and high-risk patients with community-acquired pneumonia. *Eur J Clin Microbiol Infect Dis*. 2005;24:241-249.

30. Moons KG, de Groot JA, Linnet K, Reitsma JB, Bossuyt PM. Quantifying the added value of a diagnostic test or marker. *Clin Chem*. 2012;58:1408-1417.

31. Ling DI, Pai M, Schiller I, Dendukuri N. A Bayesian framework for estimating the incremental value of a diagnostic test in the absence of a gold standard. *BMC Med Res Methodol*. 2014;14:67

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Xie X, Sinclair A, Dendukuri N. Evaluating the accuracy and economic value of a new test in the absence of a perfect reference test. *Res Syn Meth*. 2017. https://doi.org/10.1002/jrsm.1243