Comparing the Accuracy of Diagnostic Tests in Studies with Extreme Verification Bias: A Bayesian Model and Gibbs Sampling Computing Algorithm

Ngoc Ty Nguyen*[1,2], Gene Pennello[1]

Disclaimer: This poster reflects the views of the authors and should not be construed to represent the FDA's views or policies. [1]Food and Drug Administration, Division of Imaging, Diagnostic and Software Reliability; [2]University of Central Florida, Department of Statistics and Data Science *ngocty.nguyen@fda.hhs.gov

Abstract

Background: Diagnostic test accuracy is based on agreement of the test with the reference standard for verifying disease status. However, if the reference standard is not administered to all subjects, verification bias is introduced into test evaluation. Extreme verification bias (EVB) is when no one in a subset is verified for disease status. For example, women who are test negative for human papillomavirus (HPV) are ordinarily not referred to colposcopy to establish whether they have cervical cancer, rendering HPV test sensitivity and negative predictive value inestimable by standard means.

Purpose: Build a user-friendly interface to perform Bayesian analysis of diagnostic test accuracy in EVB studies.

Methodology: For studies comparing two diagnostic tests for a low prevalence disease in an EVB study in which everyone who is negative on both tests is unverified for disease status, we developed a Bayesian model that utilized weak prior information on disease prevalence and positive correlation between the tests. We developed a Gibbs sampling computing algorithm to obtain the posterior distribution of test accuracy parameters. We applied our model to HPV data, pretending that double test negatives were unverified for cervical cancer. We compared our Bayesian estimates based on the incomplete data with those based on the complete data.

Results: Remarkably, for the HPV dataset Bayesian estimates based on incomplete EVB data were unbiased and for some parameters (specificity, PPV) nearly as precise as the complete data estimates.

Conclusion: Heretofore, no statistical method has been available for comparing the accuracy of diagnostic tests in EVB studies. Our Bayesian model and computational algorithm is one such method and has the potential to impact regulatory science applications.

EVB Data Structure

- An index test is compared with a comparator test.
- If either the index or comparator test result is positive, then subject is verified for disease status with a reference method, otherwise not.
- Data consist of a 2 × 2 × 2 table with some cells counts missing:

$D-$			$D+$			$Total$		
Test	$S-$	$S+$	Test	$S-$	$S+$	Test	$S-$	$S+$
$T-$	$[n_{000}]$	n_{010}	$T-$	$[n_{001}]$	n_{011}	$T-$	$n_{00\bullet}$	$n_{01\bullet}$
$T+$	n_{100}	n_{110}	$T+$	n_{101}	n_{111}	$T+$	$n_{10\bullet}$	$n_{11\bullet}$

- n_{tsd} = cell count for new test result $T = t$, comparator test result $S = s$,disease status $D = d$, for $t, s, d = 0,1$ or $t, s, d = -, +$:
- $[n_{tsd}]$ denotes count is missing (disease status unverified)
- The corresponding table of cell probabilities is listed:

- Cell Probabilities

$D-$			$D+$			$Total$		
Test	$S-$	$S+$	Test	$S-$	$S+$	Test	$S-$	$S+$
$T-$	$[b_{00}q_{00}]$	$b_{01}q_{01}$	$T-$	$[b_{00}p_{00}]$	$b_{01}p_{01}$	$T-$	b_{00}	b_{01}
$T+$	$b_{10}q_{10}$	$b_{11}q_{11}$	$T+$	$b_{10}p_{10}$	$b_{11}p_{11}$	$T+$	b_{10}	b_{11}

- $\boldsymbol{b_{ts}} = \Pr(T = t, S = s)$ = probability of $T = t$ and, $S = s$
- $\boldsymbol{p_{ts}} = \Pr(D = 1|T = t, S = s)$ = predictive value of $T = t, S = s$ for $D = 1$, $\boldsymbol{q_{ts}} = \boldsymbol{1} - \boldsymbol{p_{ts}}$.
- Red denotes cells with disease status missing (unverified)

Gibbs Sampler

Prior Distribution

- $\underline{\mathbf{b}} \sim Dir\left(\underline{\gamma}\right), \underline{\gamma} = (\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11}) = (0.25,0.25,0.25,0.25)$
- $\boldsymbol{p_{00}}, \boldsymbol{p_{01}}, \boldsymbol{p_{10}}, \boldsymbol{p_{11}} \sim Beta\left(\underline{\alpha}\right),\ \ \underline{\alpha} = (0.5,0.5)$

Prior Information

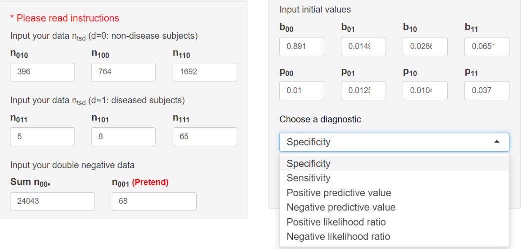
- $p_{00} < min(p_{10}, p_{01}) < p_{11}$ *(Tests are Informative)*
- $\frac{p_{01}p_{10}}{p_{11}o} < \ p_{00} < 1 - \frac{q_{01}q_{10}}{q_{11}o},\ \ o = \frac{b_{00}b_{11}}{b_{10}b_{01}}$ *(Tests Positively Correlated)*

Gibbs Sampling Algorithm

1. $\boldsymbol{n_{001}^{(i)}}|\underline{p}^{(i-1)} \sim Bin\left(n_{00\bullet}, p_{00}^{(i-1)}\right)$ *(Data Augmentation Step)*
2. $\boldsymbol{p_{ts}^{(i)}}|\underline{n}^{(i)} \sim Beta(\alpha_1 + n_{ts1}, \alpha_0 + n_{ts0})$ for $(t, s) = (0,1), (1,0)$, or $(1,1)$
3. $\boldsymbol{p_{00}^{(i)}}|\underline{p}^{(i)} \sim Beta(a + n_{001}^{(i)}, b + n_{000}^{(i)})$
4. $\underline{\boldsymbol{b}}^{(i)}|\underline{n}^{(i)} \sim Dir(\underline{\gamma} + \underline{n}_{\bullet}^{(i)})$

Regulatory Science Tool Interface

- Input 1: Counts for verified for disease status and sum of double negatives
- Input 2: Initial values for Gibbs algorithm (optional)
- Output: specificity, sensitivity, positive predictive value, negative predictive value, positive likelihood ratio, negative likelihood ratio.



* Please read instructions

Input your data n_{tsd} (d=0: non-disease subjects)

n_{010}	n_{100}	n_{110}
396	764	1692

Input your data n_{tsd} (d=1: diseased subjects)

n_{011}	n_{101}	n_{111}
5	8	65

Input your double negative data

Sum $n_{00\bullet}$	n_{001} (Pretend)
24043	68

Input initial values

b_{00}	b_{01}	b_{10}	b_{11}
0.891	0.0149	0.0286	0.065'
p_{00}	p_{01}	p_{10}	p_{11}
0.01	0.0125	0.0104	0.037

Choose a diagnostic

Specificity

Specificity
Sensitivity
Positive predictive value
Negative predictive value
Positive likelihood ratio
Negative likelihood ratio

Numerical Studies

Quantities to be Estimated

$Pr(T = 1|D = 1) =$ True positive fraction = Sensitivity (Se)

$Pr(T = 1|D = 0) =$ False positive fraction = 1– Specificity (1 – Sp)

$Pr(D = 1|T = 1) =$ Positive Predictive Value (PPV)

$Pr(D = 0|T = 0) =$ Negative Predictive Value (NPV)

Human Papilloma Virus (HPV) Tests and EVB Data

HPV tests screen for HPV genotypes, which are precursors to cervical cancer or cervical squamous intraepithelial neoplasia stage 3 (CIN3+), verified by histology of biopsy taken during colposcopy.

Verify-the-Positive (VTP) Design. Reference procedure verifying disease status (e.g., colposcopy for HPV tests) performed only when study subject tests positive for disease by 1 of 2 tests being compared (Schatzkin et al, 1987).

Study Data. Women with NILM Pap Smear Cytology

	< CIN3+		CIN3+	
	$S-$	$S+$	$S-$	$S+$
$T-$	[23975]	396	[68]	5
$T+$	764	1692	8	65

[•] denotes missing CIN3+ status (pretend VTP data).

Bayesian Analysis Results

PPV was estimated precisely with EVB data (Figure 3) because it is function of verified test positives. **Specificity** was also estimated precisely with EVB data (Figure 1) despite being a function of unverified true negatives as well as verified false positives. Evidently, the prior information of informative and positively correlated tests improved precision. **NPV** and **sensitivity** were imprecisely estimated with EVB data (Figures 2 and 4), yet remarkably had posterior medians nearly equal to those for complete data.

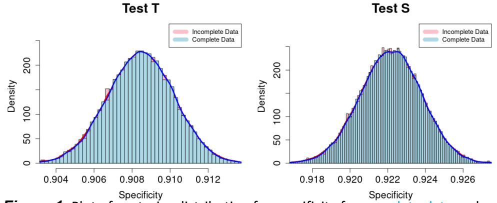


Figure 1. Plot of posterior distribution for specificity for complete data and pretend incomplete VTP data.

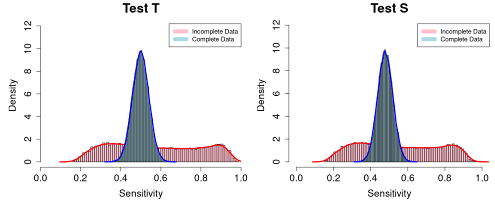


Figure 2. Plot of posterior distribution for sensitivity from Gibbs sampler for for complete data and pretend incomplete VTP data.

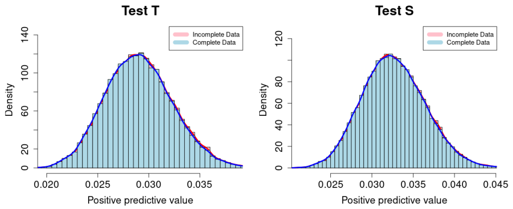


Figure 3. Plot of posterior distribution for positive predictive value for complete data and pretend incomplete VTP data.

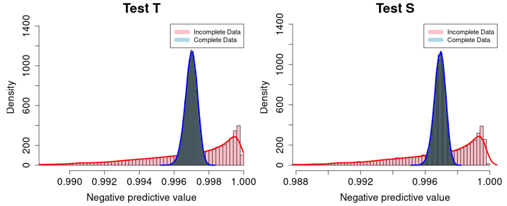


Figure 4. Plot of posterior distribution for negative predictive value for complete data and pretend incomplete VTP data.

Discussion

- PPV and 1 - Specificity are two types of false positive fractions, which may explain why they were estimated precisely with EVB data. NPV and 1 – sensitivity are two types of false negative fractions, which may explain why they were estimated imprecisely with EVB data.
- We are not aware of any statistical method for comparing the accuracy of diagnostic tests in EVB studies. Our Bayesian model and computational algorithm is the first and has the potential to impact regulatory science applications. Our easy-to-use regulatory science tool app implements the Bayesian analysis. For the HPV data, the Bayesian posterior medians of accuracy were unbiased, but more experience with other datasets is needed to evaluate robustness of the method.

References

- Schatzkin A, Connor RJ, Taylor PR, Bunnag B. Comparing new and old screening tests when a reference procedure cannot be performed on all screenees. Example of automated cytometry for early detection of cervical cancer. *Am J Epidemiol.* 1987 Apr; 125(4) : 672-8.