Using Gibbs Sampling and Data Augmentation to Compare Diagnostic Tests in RWE Studies with Extreme Verification Bias

Gene Pennello[2] and Qin Li[1] , FDA/CDRH,
[1]Division of Biostatistics, [2]Division of Imaging Diagnostics & Software Reliability

Abstract

Diagnostic tests are used to detect or predict presence or absence of a disease or a clinical condition now or in the future. Clinical studies are often designed to evaluate the performance of the diagnostic tests with the comparison to the reference method that is used to determine the true status of the subjects. Meanwhile, alternative sources of evidence such as real-world data (RWD) may exist for the diagnostic tests of interest. Verification bias (partial or extreme) is not uncommon to encounter in a clinical study and RWD, when the reference procedure used to verify disease status is invasive or otherwise unethical to perform on everyone. It is difficult to evaluate the diagnostic tests and can be a challenge for regulatory decision making in this situation, especially when the extreme verification bias exists (i.e. no one or more subgroups are verified). In this poster, we develop Bayesian models to make the comparison of two tests possible under the situation of extreme verification bias. A Gibbs sampler-based computation algorithm is developed accordingly for drawing posterior samples and inference. As an example, the proposed method is applied to a Human papilloma virus (HPV) diagnostic device.

Background and motivation

Verification Bias: Estimates of accuracy – sensitivity (Se) and specificity (Sp) are biased if selection of subjects for verification of disease status is non-random.

Extreme Verification Bias: In one or more subsets, no one is verified for disease status. It occurs by design when the reference procedure used to verify disease status is invasive and thus deemed unethical to perform on anyone in particular subsets.

HPV tests are used to screen for HPV genotypes that are precursors to cervical cancer or cervical squamous intraepithelial neoplasia stage 3 (CIN3+ histology).

Verify-the-Positive (VTP) Design A subject is referred to coloposcopy to verify cervical cancer status only if they are test positive by one of two HPV tests being compared (Schatzkin et al, *Biometrics*, 1987).

	<CIN3+		CIN3+	
	T^*-	T^*+	T^*-	T^*+
$T-$	[23975]	396	[68]	5
$T+$	764	1692	8	65
	26827		146	

- NILM: Pap Cytology Result is **N**egative for **I**ntraepithelial **L**esion or **M**alignancy.
- CIN3+ prevalence is 0.54% (146/26973).
- VTP design [ ] is missing

Estimable Quantities: Ratio of TPF (sensitivity), Ratio of FPF (1-specificity), PPVs

Bayesian Model

Data Notation

$D-$			$D+$			Total		
Test	T^*-	T^*+	Test	T^*-	T^*+	Test	T^*-	T^*+
$T-$	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 test results $T=t, T^*=s$, disease status $D=d$, for $t,s,d = 0,1$ or $-,+$

Data Distribution

$$\underline{n} \sim \boldsymbol{Mult}(\boldsymbol{n_{\dots}}, \underline{\boldsymbol{\theta}}),$$
$$\underline{\boldsymbol{n}} = \{n_{ts\bullet}\} = (n_{00\bullet}, n_{01\bullet}, n_{10\bullet}, n_{11\bullet})$$
$$n_{\dots} = \sum_{t=0,1}\sum_{s=0,1} n_{ts\bullet}$$
$$\underline{\boldsymbol{\theta}} = \{\theta_{ts\bullet}\} = (\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11})$$
=joint prob of test results

$$\boldsymbol{n_{ts1}} \sim \boldsymbol{Bin}(\boldsymbol{n_{ts\bullet}}, \boldsymbol{p_{ts}}),$$
$$n_{ts\bullet} = \sum_{d=0,1} n_{tsd}$$
$$\boldsymbol{p_{ts}} = \Pr(D=1|T=t, T^*=s)$$
= predictive value of $T=t, T^*=s$.

Diffuse Priors

$$\underline{\theta} \sim Dir(\underline{\gamma}),$$
$$\underline{\gamma} = (\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11})$$
$$= (0.25, 0.25, 0.25, 0.25)$$
$$p_{ts} \sim Beta(\underline{\alpha}), t, s = 0,1$$
$$\underline{\alpha} = (0.5, 0.5)$$

Computation

- **Markov Chain Monte Carlo (MCMC) Gibbs Sampling**. Parameter values were sampled from their full conditional posterior distributions until Markov Chain converged to samples from the joint posterior distribution of the parameters.
- **Data Augmentation**. Missing disease status for HPV test double negatives was sampled from its full conditional posterior predictive distribution, greatly simplifying the Gibbs sampler.

Gibbs Sampler

$$\underline{\theta}^{(i+1)} | rest \sim Dir(\underline{\gamma} + \underline{n})$$
$$p_{ts}^{(i+1)} | rest \sim Beta(\alpha_0 + n_{ts0}, \alpha_1 + n_{ts1})$$
for $(t,s) = (0,1), (1,0)$, or $(1,1)$
$$n_{001}^{(i+1)} | rest \sim Bin\left(n_{00\bullet}, p_{00}^{(i)}\right)$$
$$p_{00}^{(i+1)} | rest \sim Beta(\alpha_0 + n_{000}^{(i+1)}, \alpha_1 + n_{001}^{(i+1)})$$

Model Constraints

Constraint 1:

None of HPV double negatives is verified.

That is, the data provide no information on

$$p_{00} = \Pr(D=1|T=0, T^*=0)$$

A reasonable constraint is that

$$\boldsymbol{p_{00} < min(p_{10}, p_{01})}$$

Constraint 2:

HPV tests are based on similar technology.

A reasonable assumption: conditional on disease status the two HPV tests are ***positively dependent***, that is, the classification probability

$$Pr(T=t, T^*=s|D=d)$$

is bounded below by conditional independence:

$$Pr(T=t, T^*=s|D=d)$$
$$> \Pr(T=t|D=d) \times \Pr(T^*=s|D=d)$$

No disease table

$D-$	T^*-	T^*+	
$T-$	$\boldsymbol{1-FPF_T-FPF_{T^*}+\theta_0}$	$\boldsymbol{FPF_{T^*}-\theta_0}$	$\boldsymbol{1-FPF_T}$
$T+$	$\boldsymbol{FPF_T-\theta_0}$	$\boldsymbol{\theta_0}$	$\boldsymbol{FPF_T}$
	$\boldsymbol{1-FPF_{T^*}}$	$\boldsymbol{FPF_{T^*}}$	

$$FPF_T \times FPF_{T^*} < \theta_0 < \min(FPF_T, FPF_{T^*})$$

Disease table

$D+$	T^*-	T^*+	
$T-$	$\boldsymbol{1-TPF_T-TPF_{T^*}+\theta_1}$	$\boldsymbol{TPF_{T^*}-\theta_1}$	$\boldsymbol{1-TPF_T}$
$T+$	$\boldsymbol{TPF_T-\theta_1}$	$\boldsymbol{\theta_1}$	$\boldsymbol{TPF_T}$
	$\boldsymbol{1-TPF_{T^*}}$	$\boldsymbol{TPF_{T^*}}$	

$$TPF_T \times TPF_{T^*} < \theta_1 < \min(TPF_T, TPF_{T^*})$$

In the Gibbs sampler, we only accept samples that satisfy these constraints.

HPV Example Results

True Parameter Values
vs.
Sample Estimates for Fully Verified Data
vs.
Bayesian Estimates for Incompletely Verified Data

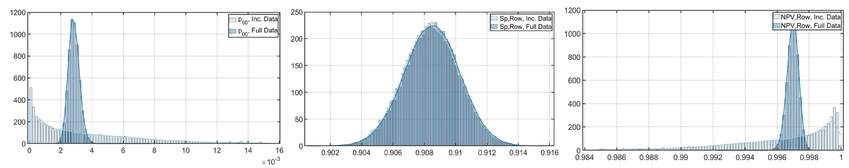
	True Value	Full Data	Analysis of Incomplete Data: Start Value	Poster. Median	95% CI 2.5%	95% CI 97.5%
p_{00}	0.003	0.003	0.008	0.003	0.000	0.012
p_{01}	0.010	0.012	0.012	0.013	0.005	0.027
p_{10}	0.012	0.010	0.010	0.011	0.005	0.020
p_{11}	0.040	0.037	0.037	0.037	0.029	0.047
p	0.005	0.005	0.010	0.005	0.003	0.014

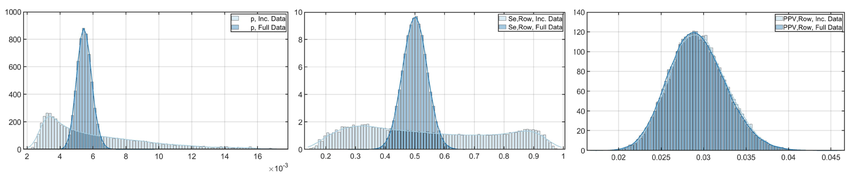
Test T

	True Value	Full Data	Analysis of Incomplete Data: Start Value	Poster. Median	95% CI 2.5%	95% CI 97.5%
Sp	0.908	0.908	0.908	0.908	0.905,	0.912
Se	0.555	0.500	0.276	0.521	0.201,	0.931
NPV	0.997	0.997	0.992	0.997	0.987,	1.000
PPV	0.032	0.029	0.029	0.029	0.023,	0.036
NLR	0.490	0.550	0.798	0.528	0.076,	0.881
PLR	6.033	5.462	2.997	5.685	2.170,	10.222

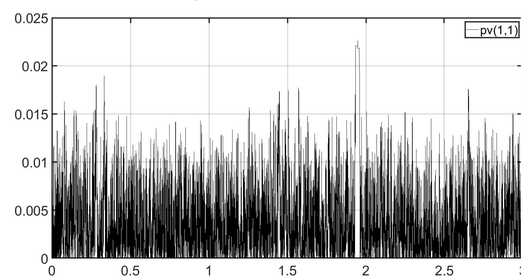
Test T*

	True Value	Full Data	Analysis of Incomplete Data: Start Value	Poster. Median	95% CI 2.5%	95% CI 97.5%
Sp	0.921	0.922	0.922	0.922	0.919,	0.925
Se	0.521	0.479	0.264	0.499	0.188,	0.898
NPV	0.997	0.997	0.992	0.997	0.987,	1.000
PPV	0.035	0.032	0.032	0.033	0.026,	0.041
NLR	0.521	0.564	0.798	0.543	0.111,	0.881
PLR	6.575	6.160	3.381	6.424	2.395,	11.593





Trace Plot for p00



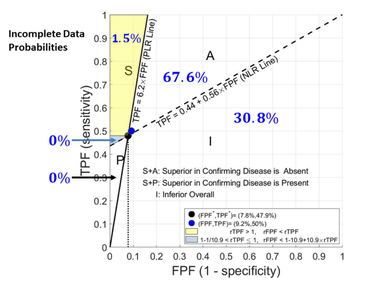


Figure 2. Likelihood ratio graph with regions of comparison S, A, P and I for a test vs. a comparator with $(FPF^*, TPF^*) = (0.078, 0.479)$ and odds ratio $o^+ = 10.9$. [19]

Concluding Remarks

Estimation

- Bayesian posterior medians with unknown disease status for test double negatives (24043/26973 = 89%) agreed surprisingly well with sample estimates when they were known.
  - Majority of CIN3+ disease (78/146, 53.4%) occurred in subjects who were test positive by one of the tests.
  - The two constraints on the predictive values and classification probabilities place a lot of structure on their distribution, increasing the precision of the Bayesian estimates.

Computation

- Starting values for Gibbs Sampler:
  - Use $\hat{p}_{01}, \hat{p}_{10}, \hat{p}_{11}, \hat{\theta}_{00}, \hat{\theta}_{01}, \hat{\theta}_{10}, \hat{\theta}_{11}$ as starting values for $p_{01}, p_{10}, p_{11}, \theta_{00}, \theta_{01}, \theta_{10}, \theta_{11}$
  - For p_{00}: $p_{00} < \min(\hat{p}_{01}, \hat{p}_{10})$
- In VTP studies, estimable quantities are

$$rTPF = \frac{Se}{Se^*}, rFPF = \frac{1-Se}{1-Se^*}, PPV, PPV^*,$$

For these quantities, Bayesian estimates should agree with sample estimates or something is wrong.

References

- Brief Summary of the Microbiology Devices Panel – March 8, 2019, https://www.fda.gov/media/122803/download.
- 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.
- Biggerstaff BJ. Comparing diagnostic tests: a simple graphic using likelihood ratios. Stat Med. 2000 Mar 15;19(5):649-63.
- Black MA, Craig BA. Estimating disease prevalence in the absence of a gold standard. Statist. Med. 2002; 21:2653–2669 (DOI: 10.1002/sim.1178).