Comparing Diagnostic Tests in Studies with Extreme Verification Bias
-- application to HPV testing

Gene Pennello[1], Tingting Hu[1], Ngoc Ty Nguyen[2]

FDA; [1]CDRH/OSEL, [2]CBER/OBPV

This presentation reflects the views of the author and should not be construed to represent FDA's views or policies.

…

References

- Black MA, Craig BA. Estimating disease prevalence in the absence of a gold standard. *Statist Med*. 2002; 21:2653–2669 (DOI:
- 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.
- Pennello GA. Bayesian analysis of diagnostic test accuracy when disease state is unverified for some subjects. *J Biopharmaceutical Statistics*. 2011;21:954–970.

Disclaimer

- The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services.

Outline

- Background
- HPV Example
- Problem Definition and Methodology
- Application to HPV Example
- Discussion (Tool Demo)

Background

- **Medical tests** are used to diagnose or predict presence or absence of a target condition (e.g., a disease) now or in the future. (e.g., *screening* and *confirmatory* tests).

- **Problem of Interest**: Compare two **medical test**s under the **Verify-The-Positive (VTP)** design, which introduces *extreme verification bias*.

- **VTP Design**: Subject receives reference standard procedure to verify disease status only if s/he tests positive by at least one of the tests being compared (Schatzkin et al 1987).

Verification Bias

- **Verification Bias**: Introduced when disease status is unverified by reference standard (i.e., is missing) in a non-random subset of subjects.
- ***Extreme* verification bias (EVB):** Test result determines who is verified for disease status by reference standard.

Pepe M. *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford, 2003, Chapter 7.

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.

EVB is Common

- **Physician practice**
  - Physicians often use tests when they are unsure whether to refer a subject to the reference standard.
- **Reference standard is invasive, costly, or time-consuming**
  - When the reference standard for verifying disease status is invasive, it may be unethical to perform on test negative subjects.
  - Microbiology Devices Advisory Committee meeting, 03/2019:
    - *Benefit*s of colposcopy referral for assessment of verification bias did *not* outweigh *risk*s associated with the procedure and potential overtreatment for the 'double negative' population (negative test results by two HPV tests). (https://www.fda.gov/media/122803/download)

Hypothetical example: Human papilloma virus (HPV)

- **Goal**: Compare two cervical cancer screening tests:
  T vs. S (e.g., Onclarity HPV test vs. Hybrid Capture 2 (HC2) test)
- **Disease of Interest:** CIN3+: cervical cancer, or Cervical Intraepithelial Neoplasia grade 3 (CIN3)
- **Study Data:** 26,873 patients ~ each screened by both test T and test S.
- **Verification Process (VTP design)**
  - Double negative patients (negative results on both tests) unverified.
  - Patients with at least one positive test: Verified by colposcopy.

Hypothetical Example: HPV Dataset

overall

	$S-$	$S+$
$T-$	24043	401
$T+$	772	1757

< CIN3+

	$S-$	$S+$
$T-$	[•]	396
$T+$	764	1692

CIN3+

	$S-$	$S+$
$T-$	[•]	5
$T+$	8	65

[•] denotes missing CIN3+ status.

- $T = t =$ new test result
- $S = s =$ comparator test result
- [missing] - missing count (disease status unverified)

- **Goal:** Compare tests T and S on accuracy in classifying and predicting CIN3+, the disease of interest.

Cell Count Data

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

- *n_{001} and n_{000} are missing.*

Assume

- Joint count ~ multinomial
- Diseased count | joint count~ binomial

Cell Probabilities

Total			$D-$			$D+$		
	$S-$	$S+$		$S-$	$S+$		$S-$	$S+$
$T-$	b_{00}	b_{01}	$T-$	$b_{00}q_{00}$	$b_{01}\,q_{01}$	$T-$	$b_{00}p_{00}$	$b_{01}\,p_{01}$
$T+$	b_{10}	b_{11}	$T+$	$b_{10}q_{10}$	$b_{11}\,q_{11}$	$T+$	$b_{10}p_{10}$	$b_{11}\,p_{11}$

- $\boldsymbol{b_{ts}} = \Pr(T = t, S = s) =$ *joint probability of test results* $(T, S) = (t, s)$.
- $\boldsymbol{p_{ts}} = \Pr(D+ | T = t, S = s) =$ *predictive value* of $(T, S) = (t, s)$ for $D+$,
- $\boldsymbol{q_{st}} = 1 - p_{st}$
- Once $\{\boldsymbol{p_{ts}}\}$ **and** $\{\boldsymbol{b_{ts}}\}$ are estimated, estimates of Se, Sp, PPV, NPV, PLR, NLR follow.
- $\boldsymbol{p_{00}}$ is not (directly) estimable as the subset of test results $(T, S) = (0,0)$ is unverified for disease status.

Bayesian Model

- Joint Test Result Probabilities

Data	Prior
$\underline{n}_{\bullet}\|n_{\bullet\bullet\bullet}\sim Mult\left(n_{\bullet\bullet\bullet}, \underline{\boldsymbol{b}}\right),$	$\underline{b}=\{b_{ts}\}\sim Dir\left(\underline{\gamma}\right),$ $b_{ts}=\Pr(T=t,S=s)$
$\underline{n}_{\bullet}=\{n_{ts\bullet}\}, n_{\bullet\bullet\bullet}=\Sigma_{t=0}^{1}\Sigma_{s=0}^{1}n_{ts\bullet}$	$\underline{\gamma}=(a,a,a,a), a=0.25$

- Predictive Values of Test Results

Data	Prior
$n_{ts1}\|n_{ts\bullet}\sim Bin(n_{ts\bullet}, p_{ts})$	$p_{ts}\sim Beta\left(\underline{\alpha}\right),$ $p_{ts}=\Pr(D=1\|T=t,S=s)$
In VTP design, $\boldsymbol{n_{001}}$ **is missing.**	$\underline{\alpha}=(c,c), c=0.5$

Challenge

- Double negative patients $(T, S) = (0,0)$ are unverified (n_{001} is missing).
- Thus, the data provide no information on

$$p_{00} = Pr(D = 1 | T = 0, S = 0)$$

Idea: Obtain information on p_{00} by imposing plausible constraints.

Prior Information Constraints

- Partial Ordering of Predictive Values

$$p_{00} < \min(p_{10}, p_{01})$$
$$< \max(p_{10}, p_{01}) < p_{11}$$

- Conditional Positive Dependence of Test Results

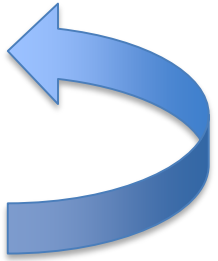
$$Pr(T = t, S = t | D = d) > \Pr(T = t | D = d) \times \Pr(S = t | D = d)$$
$$\Leftrightarrow \frac{p_{10}p_{01}}{p_{11}o} < p_{00} < 1 - \frac{q_{10}q_{01}}{q_{11}o},$$
$$o = \frac{b_{00}b_{11}}{b_{10}b_{01}}, q_{ts} = 1 - p_{ts}$$

Gibbs Sampler, Full Data (n_{001} known)

1) $\underline{b}^{(i)}|\underline{n}_{\bullet} \sim Dir\left(\underline{\gamma} + \underline{n}_{\bullet}\right), \underline{n}_{\bullet} = (n_{00\bullet}, n_{01\bullet}, n_{10\bullet}, n_{11\bullet})$

2) $p_{ts}^{(i)}|\underline{n}_{ts} \sim Beta(\alpha_1 + n_{ts1}, \alpha_0 + n_{ts0}),$

$(t, s) = (0,0), (0,1), (1,0), (1,1)$



Gibbs Sampler, VTP Data (n_{001} unknown)

0) Initialize $p_{00}^{(0)}$.

1) $\underline{b}^{(i)}|\underline{n}_{\bullet} \sim Dir\left(\underline{\gamma} + \underline{n}_{\bullet}\right), \underline{n}_{\bullet} = (n_{00\bullet}, n_{01\bullet}, n_{10\bullet}, n_{11\bullet})$

2) $p_{ts}^{(i)}|\underline{n}_{ts} \sim Beta(\alpha_1 + n_{ts1}, \alpha_0 + n_{ts0})$,
$(t,s) = (0,1),(1,0),(1,1)$

3) $n_{001}^{(i)}|p_{00}^{(i-1)} \sim Bin\left(n_{00\bullet}, p_{00}^{(i-1)}\right), n_{000}^{(i)} = n_{00\bullet} - n_{001}^{(i)}$ ***(Data augmentation)***

4) $p_{00}^{(i)}|\underline{n}_{00}{}^{(i)} \sim Beta\left(\alpha_0 + n_{001}^{(i)}, \beta_0 + n_{000}^{(i)}\right)$

5) If $\left\{p_{ts}^{(i)}\right\}$ meets prior information constraints,
then accept samples,
else reject samples in steps 2)-4) and resample them until accepted.

Hypothetical Example: HPV Dataset

overall

	$S-$	$S+$
$T-$	24043	401
$T+$	772	1757

< CIN3+

	$S-$	$S+$
$T-$	[26]	396
$T+$	764	1692

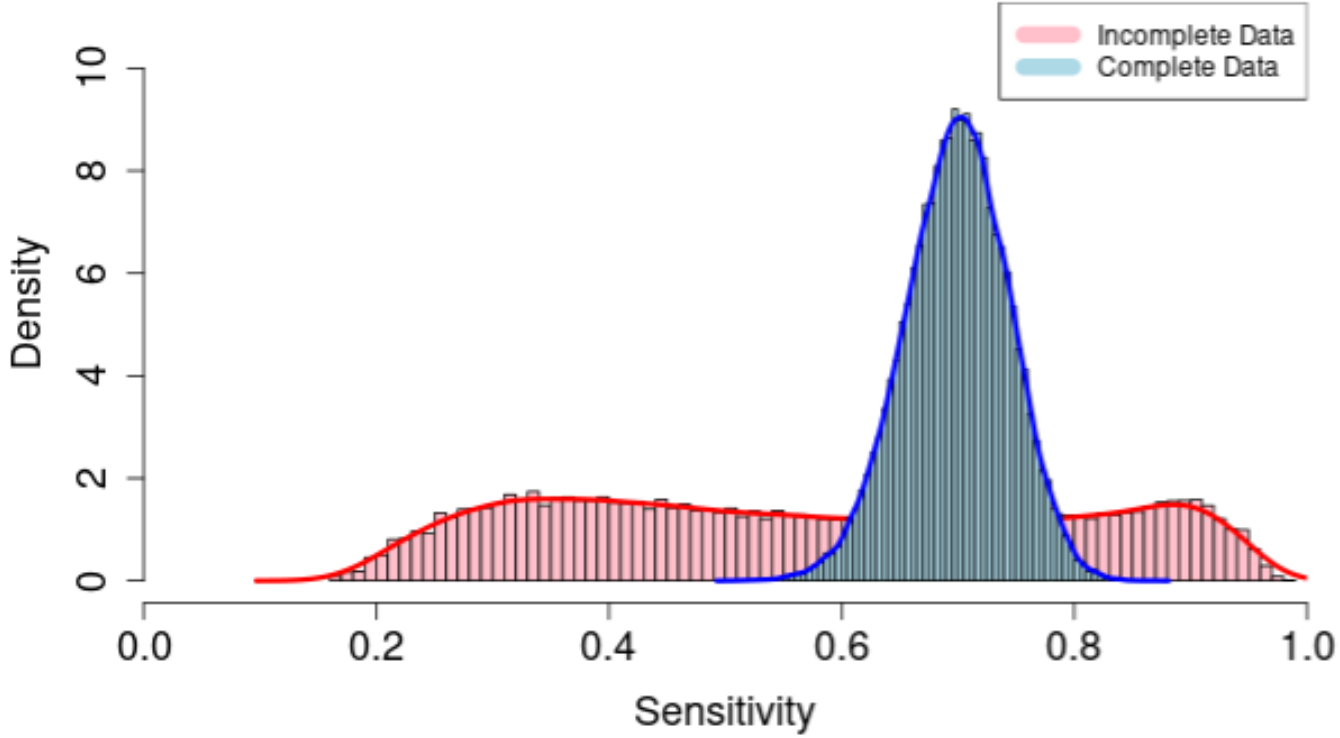
CIN3+

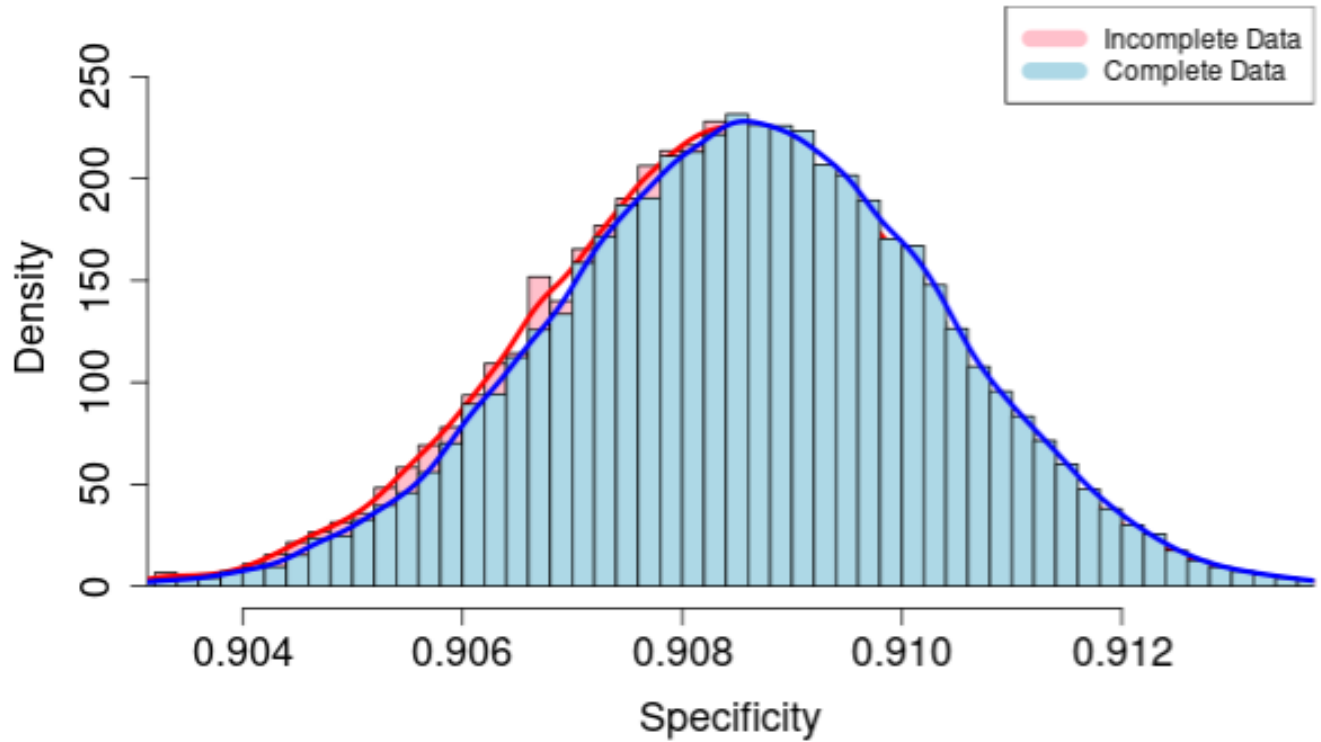
	$S-$	$S+$
$T-$	[24017]	5
$T+$	8	65

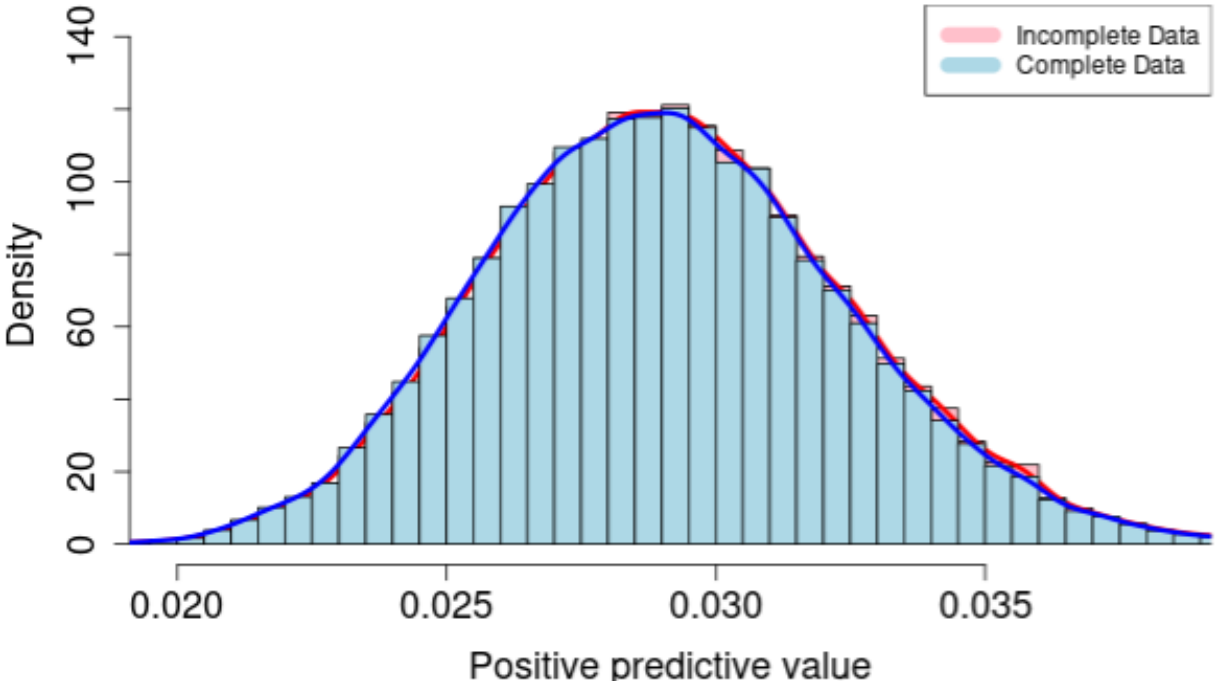
[•] denotes missing CIN3+ status.

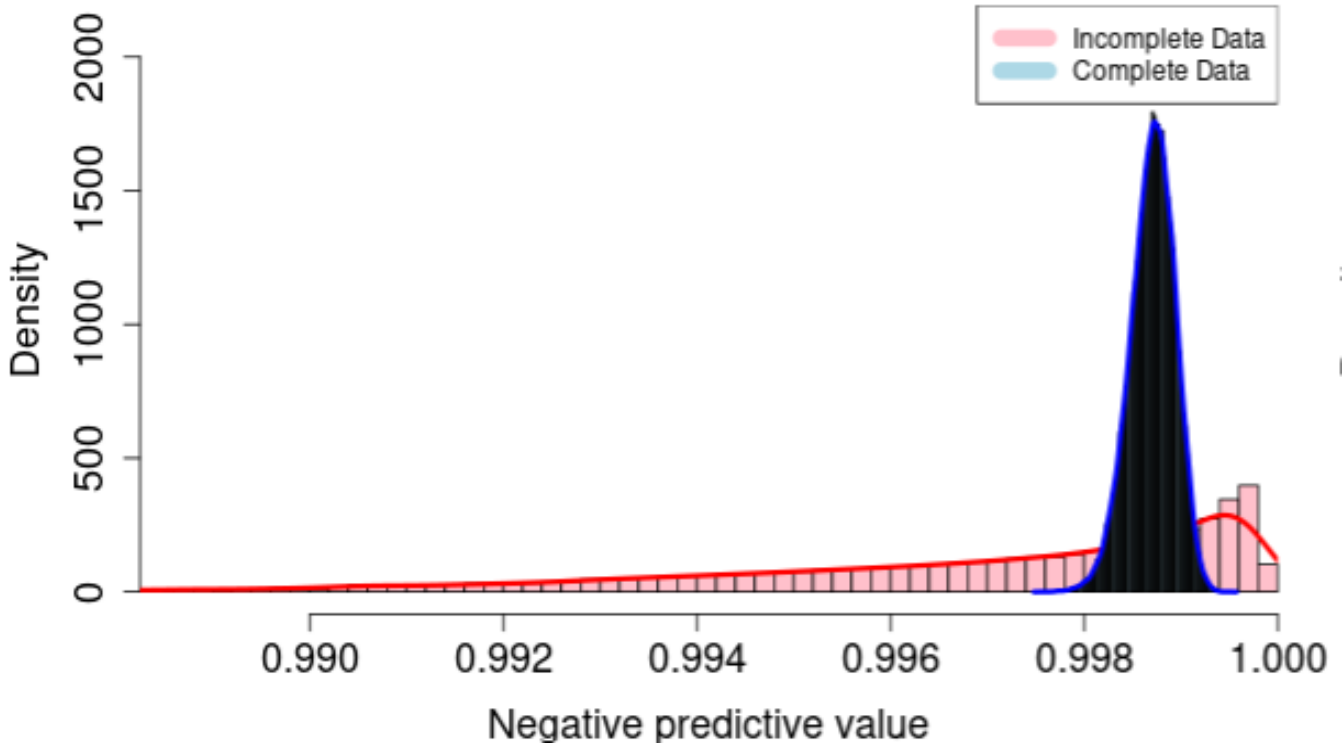
- $T = t =$ new test result
- $S = s =$ comparator test result
- [missing] - missing count (disease status unverified)

- **Goal:** Compare tests T and S on accuracy in classifying and predicting CIN3+, the disease of interest.


Incomplete Data
Complete Data
Density
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8
1.0
Sensitivity


Incomplete Data
Complete Data
Density
0
50
100
150
200
250
0.904
0.906
0.908
0.910
0.912
Specificity


Incomplete Data
Complete Data
Density
0
20
60
100
140
0.020
0.025
0.030
0.035
Positive predictive value


Incomplete Data
Complete Data
Density
0
500
1000
1500
2000
0.990
0.992
0.994
0.996
0.998
1.000
Negative predictive value

Diagnostic Accyracy for Test T

	Complete Data	Poster. Median	Lower bound	Upper bound
Sp	0.909	0.908	0.905	0.912
Se	0.700	0.559	0.132	0.986
PPV	0.029	0.029	0.022	0.036
NPV	0.999	0.998	0.992	1.000
PLR	7.656	6.113	1.415	10.812
NLR	0.330	0.486	0.015	0.956

Diagnostic Accyracy for Test S

	Complete Data	Poster. Median	Lower bound	Upper bound
Sp	0.922	0.922	0.919	0.925
Se	0.671	0.536	0.125	0.947
PPV	0.032	0.033	0.025	0.040
NPV	0.999	0.998	0.992	1.000
PLR	8.632	6.889	1.567	12.211
NLR	0.357	0.504	0.058	0.950

Complete Data: Posterior median of Gibbs samples from complete data

Poster Median: Posterior median of Gibbs samples from incomplete data

Lower bound: Lower bound of 95% CI of Gibbs samples from incomplete data

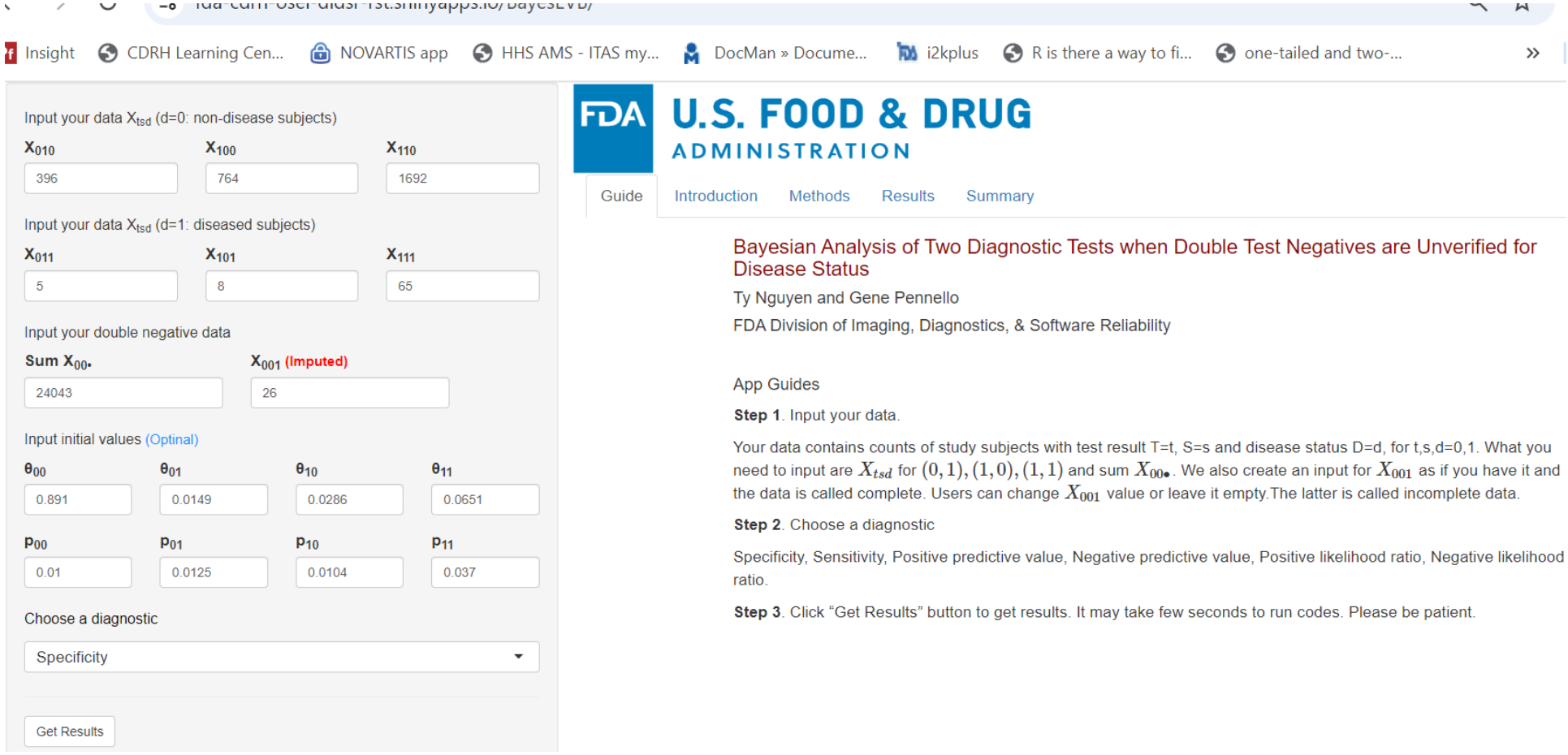
Upper bound: Upper bound of 95% CI of Gibbs samples from incomplete data

Discussion: Potential Regulatory Science Tool

- Implemented the method as a web-based tool (to release later)
- Regulatory Science Tool (RST)
  - tools produced by CDRH's Office of Science and Engineering Labs (https://www.fda.gov/medical-devices/science-and-research-medical-devices/catalog-regulatory-science-tools-help-assess-new-medical-devices#sciencetools)
    - support medical device development and patient access to safe and effective medical devices,
    - help in the assessment of new medical devices.
  - Example tool types: Phantoms, Methods, Computational models and simulations.
  - The list is expanding with new tools available…
- What RS Tools do not do
  - These tools do not replace FDA-recognized standards or MDDTs.
  - These tools have not been qualified as Medical Device Development Tools and the FDA has not evaluated the suitability of these tools within any specific context of use. S*ponsors* considering using a tool from this catalog in marketing submissions may request feedback or meetings for medical device submissions as part of the Q-Submission Program.

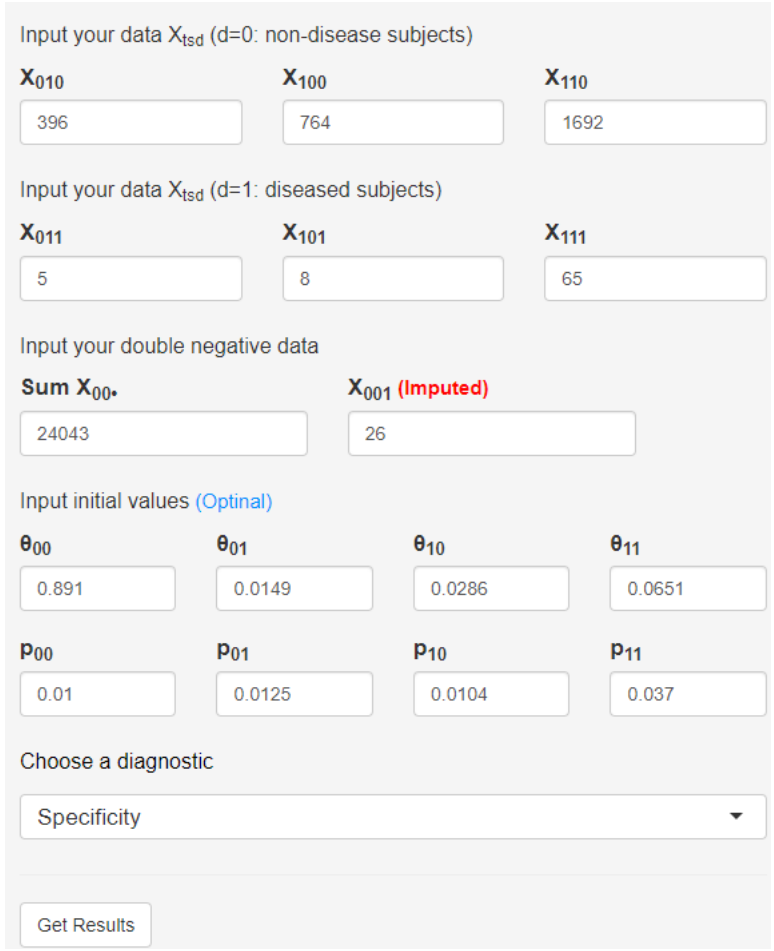
Discussion: Tool Demo

- Potential RST (to release later)
- Homepage



Discussion: Tool Demo

- Potential RST (to release later)
- Tool - input



Given Data

	$D-$	
	$S-$	$S+$
$T-$	$[n_{000}]$	n_{010}
$T+$	n_{100}	n_{110}

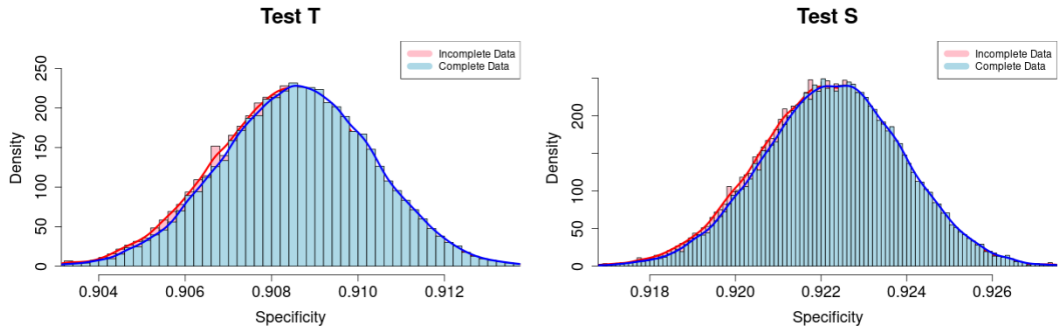
	$D+$	
	$S-$	$S+$
$T-$	$[n_{001}]$	n_{011}
$T+$	n_{101}	n_{111}

Total		
	$S-$	$S+$
$T-$	$n_{00\bullet}$	$n_{01\bullet}$
$T+$	$n_{10\bullet}$	$n_{11\bullet}$

Discussion: Tool Demo

- Potential RST (to release later)
- Tool output - “Results” tab

Guide | Introduction | Methods | Results | Summary



Specificity Analysis for Test T
Estimation from Gibbs sampler with complete data 0.909
Estimation from Gibbs sampler with incomplete data 0.908 with 95% CI (0.905 , 0.912)

Specificity Analysis for Test S
Estimation from Gibbs sampler with complete data 0.922
Estimation from Gibbs sampler with incomplete data 0.922 with 95% CI (0.919 , 0.925)

Tool output - “Summary” tab

Guide | Introduction | Methods | Results | Summary

Diagnostic Accyracy for Test T

	Complete Data	Poster. Median	Lower bound	Upper bound
Sp	0.909	0.908	0.905	0.912
Se	0.700	0.559	0.132	0.986
PPV	0.029	0.029	0.022	0.036
NPV	0.999	0.998	0.992	1.000
PLR	7.656	6.113	1.415	10.812
NLR	0.330	0.486	0.015	0.956

Diagnostic Accyracy for Test S

	Complete Data	Poster. Median	Lower bound	Upper bound
Sp	0.922	0.922	0.919	0.925
Se	0.671	0.536	0.125	0.947
PPV	0.032	0.033	0.025	0.040
NPV	0.999	0.998	0.992	1.000
PLR	8.632	6.889	1.567	12.211
NLR	0.357	0.504	0.058	0.950

Complete Data: Posterior median of Gibbs samples from complete data
Poster Median: Posterior median of Gibbs samples from incomplete data
Lower bound: Lower bound of 95% CI of Gibbs samples from incomplete data
Upper bound: Upper bound of 95% CI of Gibbs samples from incomplete data

Discussion: Estimation

How can the Bayesian model estimate quantities that are not supposed be estimable? Speculation:

- Although 89% (24043/26973) of subjects are unverified (the double negatives), the dataset has 53.4% (78/146) of the events, i.e., much of the information.
- The constraints on the predictive values and classification probabilities impose a lot of structure even though separately they appear weak.

Discussion: Computation

- Starting values in Gibbs Sampler
  - For p_{00}: $p_{00} < \min(\hat{p}_{01}, \hat{p}_{10})$
  - Bayesian estimates of estimable quantities – $rTPF, rFPF, PPV, PPV^*$ – should agree with sample estimates

Future Directions

- Evaluate operating characteristics of Bayesian model in simulated positive dependence datasets:
  - Coverage of Bayesian CI compared with nominal level
  - Average length of Bayesian CI
  - Average absolute difference between posterior mean, median, or mode and full data estimate
  - Average effective sample size of estimate
- Bayesian model average over several models that impose different structures on the tests.
- In any given application, consult domain experts determine structure on the tests that is plausible based on biology and technology.

THANK YOU!

Questions?

Gene.Pennello@fda.hhs.gov