# Adjusting for Exposure Misclassification in Bayesian Hypothesis Testing in Case-Control Studies

Ehsan. Karim and Paul Gustafson

Department of Statistics
University of British Columbia

December 28, 2010

# Outline

- Motivation
- Problem Settings
- Frequentist and Bayesian Approach for Adjustment
- Simulation Settings and Results
- Application to Epidemiologic Data
- Summary

# Motivation > Origin of the Problem

- Goal is to find relationship between an disease outcome variable ($Y$) and the exposure ($V$)
- Precise quantification of exposure variable ($V$) is not possible due to various practical reasons
- Cruder measurement used or surrogate variable ($V^*$) value collected
- In context of inference, this suffers several consequences

## Motivation > Consequences in Inference

Measurement error in the exposure variable can have adverse effects

- on the power of a hypothesis test in detecting the impact of an exposure variable in the development of a disease.
- As it distorts the structure of data, more uncertainty is associated with the inferential procedure.

In the current work, we will try to find a way to adjust for misclassification error (discrete part) while applying hypothesis testing procedures.

# Problem Setting > Basic Setup

- Retrospective case-control scenario.
- Correctly measured binary response

$$Y = \begin{cases} \text{Diseased or} \\ \text{Non-diseased,} \end{cases}$$

- Binary exposure variable

$$V = \begin{cases} \text{Truly Exposed or} \\ \text{Truly Unexposed,} \end{cases}$$

- Surrogate binary exposure variable

$$V^* = \begin{cases} \text{Apparently Exposed or} \\ \text{Apparently Unexposed,} \end{cases}$$

- Under non-differential misclassification (pattern of error $V^*|V, Y$ does not depend on $Y$).

# Problem Setting > Adjustment Techniques

For the correction of measurement error, we go through the

1. Replicated Measurement
2. Validation study
    - the validated sub-sample is derived from the same population under investigation and
    - superior method of exposure assessment is implemented on each under the sub-sample.

# Problem Setting > Main Part of the Data

| | Main (unvalidated) part of the data | | | |
|---|---|---|---|---|
| $Y$ | $Y = 1$ | | $Y = 0$ | |
| $V$ / $V^*$ | $V^* = 1$ | $V^* = 0$ | $V^* = 1$ | $V^* = 0$ |
| $V = 1$ | $u_{11}$ | $u_{12}$ | $u_{01}$ | $u_{02}$ |
| $V = 0$ | $u_{13}$ | $u_{14}$ | $u_{03}$ | $u_{04}$ |
| Total | $n_{15}$ | $n_{16}$ | $n_{05}$ | $n_{06}$ |

We can calculate $\theta_0$ and $\theta_1$ (apparent exposure prevalence rates) from the whole data.

# Problem Setting > Validation Part of the Data

| | Validation part of the data | | | |
|---|---|---|---|---|
| $Y$ | $Y = 1$ | | $Y = 0$ | |
| $V$ / $V^*$ | $V^* = 1$ | $V^* = 0$ | $V^* = 1$ | $V^* = 0$ |
| $V = 1$ | $n_{11}$ | $n_{12}$ | $n_{01}$ | $n_{02}$ |
| $V = 0$ | $n_{13}$ | $n_{14}$ | $n_{03}$ | $n_{04}$ |
| Total | $n_{11} + n_{13}$ | $n_{12} + n_{14}$ | $n_{01} + n_{03}$ | $n_{02} + n_{04}$ |

We can calculate $r_0$, $r_1$ (exposure prevalence rates), $SN$ (sensitivity) and $SP$ (specificity) from the whole data.

# Problem Setting > Epidemiologic Example

## Cervical Cancer and Herpes Simplex Virus Study

Table: Validation sub-study from HSV-2 study

| $Y$ | Cases ($Y = 1$) | | Controls ($Y = 0$) | |
|---|---|---|---|---|
| Validated Part | $V^* = 1$ | $V^* = 0$ | $V^* = 1$ | $V^* = 0$ |
| $V = 1$ | 18 | 5 | 16 | 16 |
| $V = 0$ | 3 | 13 | 11 | 33 |
| Unvalidated (main) | 375 | 318 | 535 | 701 |
| Total | 396 | 336 | 562 | 750 |

Discarsing $V$, we can calculate $\theta_0$ and $\theta_1$.
Considering $V$, we can calculate $r_0$, $r_1$, $SN$ and $SP$.

# Adjustment > Hypothesis Formation

1. $\theta_i = P(V^* = 1 | Y = i)$
2. $r_i = P(V = 1 | Y = i)$
3. $SN_i = P(V^* = 1 | V = 1, Y = i)$
4. $SP_i = P(V^* = 0 | V = 0, Y = i)$

# Adjustment > Hypothesis Formation

- OR with Validation Data

$$\Psi = \frac{r_1/(1-r_1)}{r_0/(1-r_0)}$$

- OR without Validation Data

$$\Psi^* = \frac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)}$$

$$\theta_i = SNr_i + (1-SP)(1-r_i)$$

i.e., $\theta_i$ is a function of $r_i$; $H_0 : \theta_0 = \theta_1 \equiv H_0 : r_0 = r_1$.

# Adjustment > Frequentist Likelihoods

Without Validation Data

$$\begin{aligned} & L(\theta_0, \theta_1 | V^*, Y) \\ \propto \quad & \theta_0^{(n_{01}+n_{03}+n_{05})} \times \left\{1-\theta_0\right\}^{(n_{02}+n_{04}+n_{06})} \times \\ & \theta_1^{(n_{11}+n_{13}+n_{15})} \times \left\{1-\theta_1\right\}^{(n_{12}+n_{14}+n_{16})}. \end{aligned}$$

$$\begin{aligned} \hat{\theta}_0 &= \frac{n_{01}+n_{03}+n_{05}}{n_{01}+n_{02}+n_{03}+n_{04}+n_{05}+n_{06}}, \\ \hat{\theta}_1 &= \frac{n_{11}+n_{13}+n_{15}}{n_{11}+n_{12}+n_{13}+n_{14}+n_{15}+n_{16}}. \end{aligned}$$

Under $H_0 : \theta_0 = \theta_1 = \theta$,

$$\hat{\theta} = \frac{n_{01}+n_{03}+n_{05}+n_{11}+n_{13}+n_{15}}{n_{01}+n_{02}+n_{03}+n_{04}+n_{05}+n_{06}+n_{11}+n_{12}+n_{13}+n_{14}+n_{15}+n_{16}}.$$

# Adjustment > Frequentist Likelihoods

With Validation Data

$$
\begin{aligned}
&L(r_0, r_1, SN, SP|V^*, V, Y) \\
\propto \quad & \{r_0 SN\}^{n_{01}} \{r_0(1-SN)\}^{n_{02}} \{(1-r_0)(1-SP)\}^{n_{03}} \times \\
& \{(1-r_0)SP\}^{n_{04}} \{r_1 SN\}^{n_{11}} \{r_1(1-SN)\}^{n_{12}} \times \\
& \{(1-r_1)(1-SP)\}^{n_{13}} \{(1-r_1)SP\}^{n_{14}} \times \\
& \{r_0 SN + (1-r_0)(1-SP)\}^{n_{05}} \times \\
& \{1-(r_0 SN + (1-r_0)(1-SP))\}^{n_{06}} \times \\
& \{r_1 SN + (1-r_1)(1-SP)\}^{n_{15}} \times \\
& \{1-(r_1 SN + (1-r_1)(1-SP))\}^{n_{16}}.
\end{aligned}
$$

No close form MLE available under nondifferential misclassification.

# Adjustment > Bayesian Likelihoods

Without Validation Data

$$
\begin{aligned}
L(\tilde{\Omega} = \{\theta_0, \theta_1\} | Y_n, Y_u) &\propto \prod_{i=0}^{1} \theta_i^{n_{i1}+n_{i3}+n_{i5}} (1-\theta_i)^{n_{i2}+n_{i4}+n_{i6}} \\
&= \theta_0^{n_{01}+n_{03}+n_{05}} (1-\theta_0)^{n_{02}+n_{04}+n_{06}} \times \\
&\quad \theta_1^{n_{11}+n_{13}+n_{15}} (1-\theta_0)^{n_{12}+n_{14}+n_{16}}
\end{aligned}
$$

$$
\begin{pmatrix} \Theta_0 \\ \Theta_1 \end{pmatrix} \equiv \begin{pmatrix} log \frac{\theta_0}{1-\theta_0} \\ log \frac{\theta_1}{1-\theta_1} \end{pmatrix} \sim N\left( \begin{pmatrix} \tilde{\mu}_0 \\ \tilde{\mu}_1 \end{pmatrix}, \begin{pmatrix} \tilde{\sigma}_0^2 & \tilde{\rho}\tilde{\sigma}_0\tilde{\sigma}_1 \\ \tilde{\rho}\tilde{\sigma}_0\tilde{\sigma}_1 & \tilde{\sigma}_1^2 \end{pmatrix} \right),
$$

where $\Theta_0$, $\Theta_1$ are just the logit transformed versions of $\theta_0$, $\theta_1$ respectively.

# Adjustment > Bayesian Likelihoods

## With Validation Data

$$
\begin{aligned}
f(Y_n, Y_u|\Omega) &= L(r_0, r_1, SN, SP|Y_n, Y_u) \\
&\propto \quad \prod_{i=0}^{1} \Big[ r_i^{n_{i1}+n_{i2}+u_{i1}+u_{i2}} \times (1-r_i)^{n_{i3}+n_{i4}+u_{i3}+u_{i4}} \times SN_i^{n_{i1}+u_{i1}} \\
&\qquad \times (1-SN_i)^{n_{i2}+u_{i2}} \times (1-SP_i)^{n_{i3}+u_{i3}} \times SP_i^{n_{i4}+u_{i4}} \Big].
\end{aligned}
$$

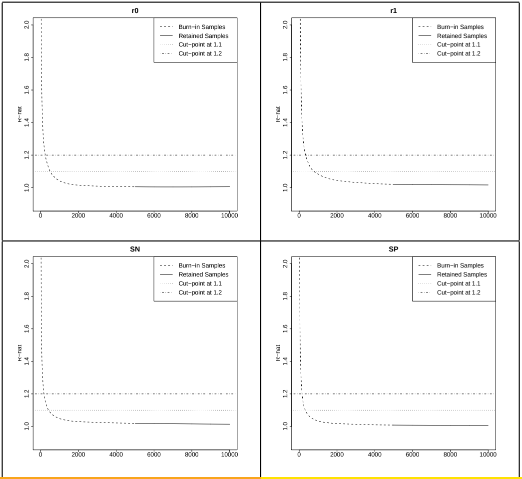
$$
\begin{aligned}
\begin{pmatrix} \Pi_0 \\ \Pi_1 \end{pmatrix} &\equiv \begin{pmatrix} log\,\frac{r_0}{1-r_0} \\ log\,\frac{r_1}{1-r_1} \end{pmatrix} \sim N\left( \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix} \right), \\
\Gamma &\equiv \left( log\,\frac{SN}{1-SN} \right) \sim N\left(\mu_2, \sigma_2^2\right), \\
\Upsilon &\equiv \left( log\,\frac{SP}{1-SP} \right) \sim N\left(\mu_3, \sigma_3^2\right),
\end{aligned}
$$

where $\Pi_0$, $\Pi_1$, $\Gamma$, $\Upsilon$ are just the logit transformed versions of $r_0$, $r_1$, $SN$, $SP$ respectively.

# Adjustment > Frequentist and Bayesian Approach

| | Without Validation | With Validation |
|---|---|---|
| Parameters in LF | $\theta_0$, $\theta_1$ | $r_0$, $r_1$, $SN$, $SP$ |
| Null Hypothesis | $H_0 : \theta_0 = \theta_1$ | $H_0 : r_0 = r_1$ |
| **Frequentist Solution** | Closed form MLE | Optimization (BFGS) |
| > Tool of Comparison | Power Curve (10,000 simulations) | |
| **Bayesian Solution** | MCMC (10,000 chain, $\frac{1}{2}$ burn-in) | |
| > Prior | Normal (hyperparameters selected reasonably) | |
| > Posterior | Beta (based on form of LF) | |
| > Tool of Comparison | Proportion of C.I. excluded $H_0$ value (2,000) | |
| > Convergence Monitoring | Gelman-Rubin ($\hat{R} << 1.1$ after burn-in) | |

# Adjustment > Convergence



# Simulation Setup and Results

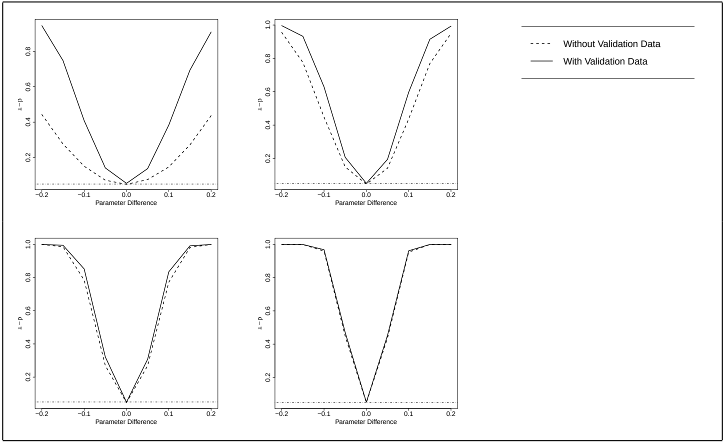
| Factor changed | SN, SP | | | | Total no. of subjects | | | |
|---|---|---|---|---|---|---|---|---|
| Scenarios | *A* | *B* | *C* | *D* | *E* | *F* | *G* | *H* |
| Validated data | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| Unvalidated data | 900 | 900 | 900 | 900 | 100 | 200 | 400 | 900 |
| $r_0/r_1$ | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 |
| $SN/SP$ | 0.6 | 0.7 | 0.8 | 0.9 | 0.7 | 0.7 | 0.7 | 0.7 |

| Factor changed | Exposure Prevalence | | | | Proportion of data | | | |
|---|---|---|---|---|---|---|---|---|
| Scenarios | *I* | *J* | *K* | *L* | *M* | *N* | *O* | *P* |
| Validated data | 100 | 100 | 100 | 100 | 100 | 250 | 500 | 750 |
| Unvalidated data | 900 | 900 | 900 | 900 | 900 | 750 | 500 | 250 |
| $r_0/r_1$ | 0.25 | 0.30 | 0.35 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 |
| $SN/SP$ | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 |

For all the scenarios, both frequentist and Bayesian methods reach to same conclusions. For the next graphs, the only difference is the vertical axis labels.

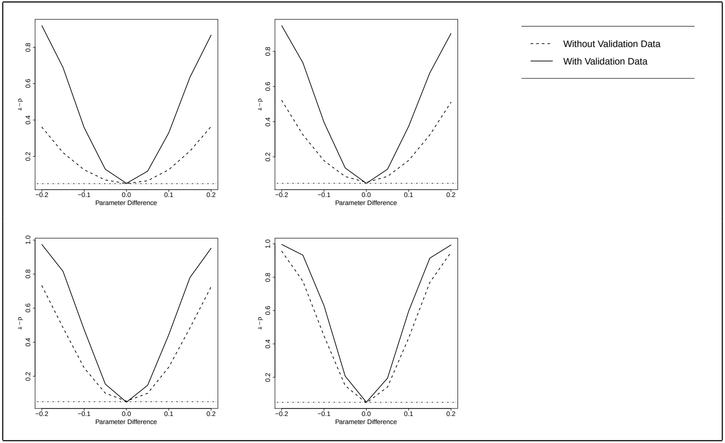
# Simulation Setup and Results > Sensitivity & Specificity

## Curves under different sensitivity and specificity values: 0.6, 0.7, 0.8 and 0.9



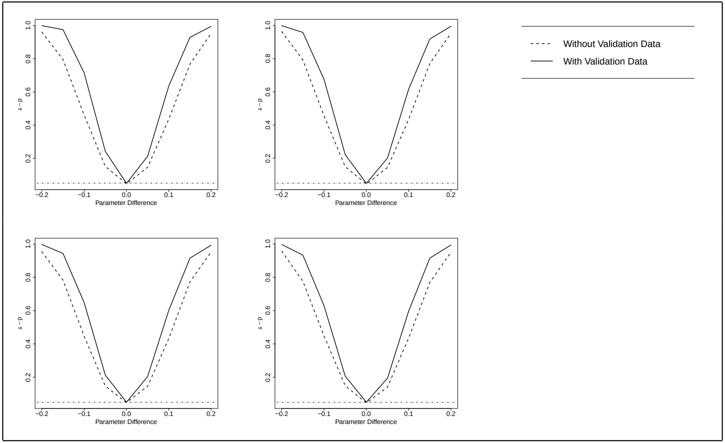
# Simulation Setup and Results > Sample Size

## Curves under different sample sizes: 200, 300, 500 and 1000



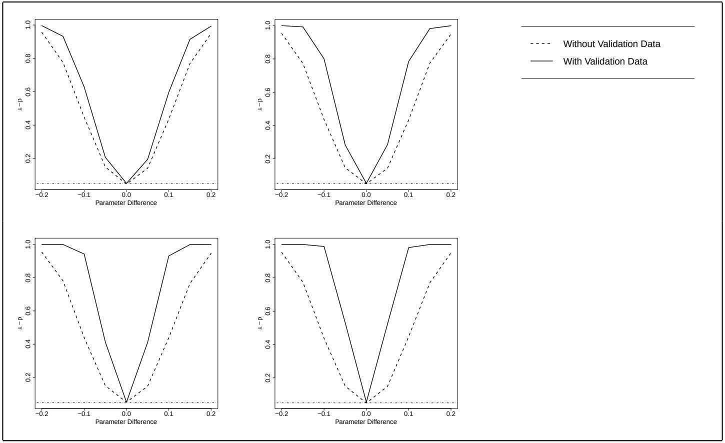
# Simulation Setup and Results > Exposure Prevalence

## Curves under different Exposure Prevalence: 0.25, 0.3, 0.35 and 0.4



# Simulation Setup and Results > Proportion of Data

## Curves under different proportions validation and main data: 1 : 9, 1 : 3, 1 : 1 and 3 : 1

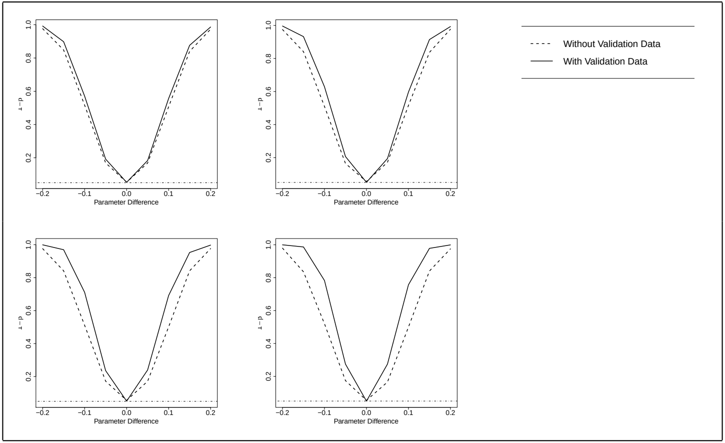


# Simulation Setup and Results > Fixed Cost \$1200

| Setting | Cost times | Validated | Unvalidated | Cost |
|---|---|---|---|---|
| A | 3 | 50 | 1050 | $3 \times 50 + 1050 = 1200$ |
| | 3 | 100 | 900 | $3 \times 100 + 900 = 1200$ |
| | 3 | 200 | 600 | $3 \times 200 + 600 = 1200$ |
| | 3 | 300 | 300 | $3 \times 300 + 300 = 1200$ |
| B | 5 | 50 | 950 | $5 \times 50 + 950 = 1200$ |
| | 5 | 100 | 700 | $5 \times 100 + 700 = 1200$ |
| | 5 | 150 | 450 | $5 \times 150 + 450 = 1200$ |
| | 5 | 200 | 200 | $5 \times 200 + 200 = 1200$ |
| C | 10 | 25 | 950 | $10 \times 25 + 950 = 1200$ |
| | 10 | 50 | 700 | $10 \times 50 + 700 = 1200$ |
| | 10 | 75 | 450 | $10 \times 75 + 450 = 1200$ |
| | 10 | 100 | 200 | $10 \times 100 + 200 = 1200$ |

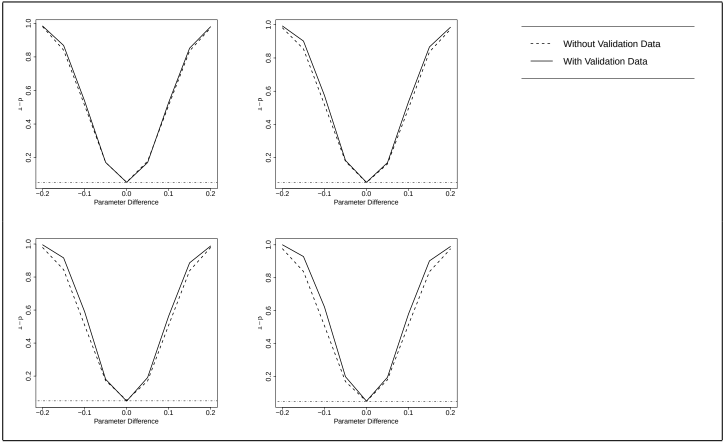
# Simulation Setup and Results > Fixed Cost $1200 > A

## Validated data 3 times costlier than unvalidated data



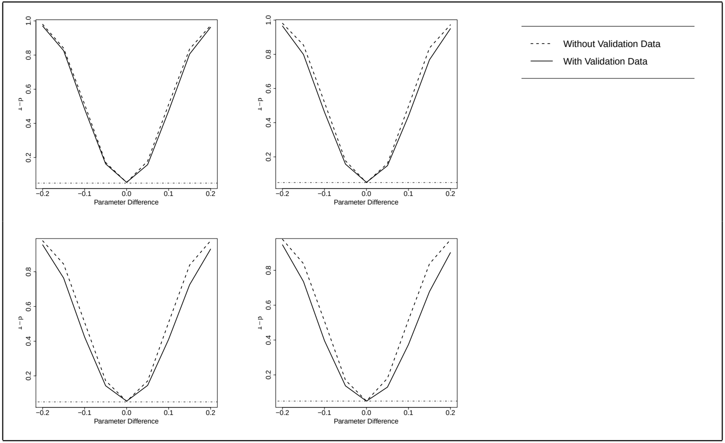
# Simulation Setup and Results > Fixed Cost $1200 > B

## Validated data 5 times costlier than unvalidated data



# Simulation Setup and Results > Fixed Cost \$1200 > C

## Validated data 10 times costlier than unvalidated data



# Application

## Cervical Cancer and Herpes Simplex Virus Study

Table: Validation sub-study from HSV-2 study

| $Y$ | Cases ($Y=1$) | | Controls ($Y=0$) | |
|---|---|---|---|---|
| Validated Part | $V^*=1$ | $V^*=0$ | $V^*=1$ | $V^*=0$ |
| $V=1$ | 18 | 5 | 16 | 16 |
| $V=0$ | 3 | 13 | 11 | 33 |
| Unvalidated (main) | 375 | 318 | 535 | 701 |
| Total | 396 | 336 | 562 | 750 |

# Application > Frequentist Results

## Cervical Cancer and Herpes Simplex Virus Study

| Not considering Validation setting | | | Considering Validation setting | | |
|---|---|---|---|---|---|
| Parameters | Estimate | SD | Parameters | Estimate | SD |
| $\theta_0$ | 0.428 | 0.014 | $r_0$ | 0.418 | 0.046 |
| $\theta_1$ | 0.541 | 0.018 | $r_1$ | 0.652 | 0.053 |
| | | | $SN$ | 0.679 | 0.041 |
| | | | $SP$ | 0.743 | 0.043 |
| $\log(OR)$ | 0.453 | 0.093 | $\log(OR)$ | 0.958 | 0.237 |
| P-value | $9.966 \times 10^{-7}$ | | P-value | $1.482 \times 10^{-6}$ | |

# Application > Bayesian Results

## Cervical Cancer and Herpes Simplex Virus Study

| Not considering Validation setting | | | Considering Validation setting | | |
|---|---|---|---|---|---|
| Parameters | Estimate | SD | Parameters | Estimate | SD |
| $\theta_0$ | 0.427 | 0.0138 | $r_0$ | 0.385 | 0.046 |
| $\theta_1$ | 0.537 | 0.0181 | $r_1$ | 0.609 | 0.052 |
| | | | $SN$ | 0.695 | 0.0393 |
| | | | $SP$ | 0.731 | 0.0398 |
| $\log(OR)$ | 0.445 | 0.0914 | $\log(OR)$ | 0.917 | 0.228 |
| 95%C.I. (OR) | Does not include $H_0$ value (1.308, 1.867) | | 95%C.I. (OR) | Does not include $H_0$ value (1.664, 3.963) | |

# Summary

Frequentist and Bayesian techniques both yield the same conclusion in the scenarios under consideration.

| Scenarios | Without validation | With validation |
|---|---|---|
| Less *SN* / *SP* | | ✓ |
| Less Sample Size | | ✓ |
| Any exposure prevalence rates | | ✓ |
| Few / More Validation data | | ✓ |
| Very Costly Validation Data | ✓ | |

Thank You!