

KU Leuven Summer School
Segment 3A
Misclassification

Paul Gustafson

September 15, 2022

A **case-control** study of association between herpes simplex virus and cervical cancer

Women with invasive cervical cancer ($Y = 1$) versus healthy controls ($Y = 0$).

Explanatory variable is presence of HSV ($X = 1$) versus not ($X = 0$).

But, a lab test to definitively determine X for a study participant is very expensive.

A lab test ('western blot') that is less definitive is much cheaper. Let X^* be the result of this test.

The data

```
dim(dta)
```

```
## [1] 2044    3
```

Have (X^*, Y) for all patients:

```
table(dta[, "y"], dta[, "xstr"], dnn=c("y", "xstr"))
```

```
##      xstr
## y      0    1
## 0 750 562
## 1 336 396
```

But distinguish the unvalidated and validated sub-samples

```
unv <- is.na(dta[, "x"])
```

```
vld <- !unv
```

```
c(sum(unv), sum(vld))
```

```
## [1] 1929  115
```

Unvalidated

```
table(dta[unv,"y"],dta[unv,"xstr"], dnn=c("y","xstr"))
```

```
##      xstr  
## y      0    1  
##    0 701 535  
##    1 318 375
```

Validated

```
table(dta[vld,"x"],dta[vld,"xstr"],dta[vld,"y"],  
      dnn=c("x","xstr","y"))
```

```
## , , y = 0  
##  
##      xstr  
## x      0  1  
##    0 33 11  
##    1 16 16  
##  
## , , y = 1  
##  
##      xstr  
## x      0  1  
##    0 13  3  
##    1  5 18
```



Pause, what inference might we draw if we go the simple/naive route

Say the validation exercise had not been carried out, and we weren't aware that western-blot lab assay was error-prone.

```
summary(glm(y~xstr, family=binomial, data=dta))$coef
```

##	Estimate	Std. Error	z value	Pr(> z)
## (Intercept)	-0.803	0.0656	-12.23	2.11e-34
## xstr	0.453	0.0928	4.88	1.06e-06

$$\hat{OR}(x_i^*, y) = e^{.45} = 1.5$$

Technical note: could have determined point-estimate and SE direct from 2 by 2 table. Logistic regression is overkill here.

Or another extreme: Only willing to work with X , treat all X^* measures as worthless

only the 115
subjects with
 X measurements

```
summary(glm(y~x, family=binomial, subset=vld, data=dta))$coef
```

##		Estimate	Std. Error	z value	Pr(> z)
##	(Intercept)	-1.012	0.292	-3.47	0.00053
##	x	0.681	0.400	1.70	0.08845

4 times
larger than
prev. slide

The Bayesian, latent variable approach

Let's build a generative model:

$$f(\gamma_0, \gamma_1, S_n, S_p) \prod_{i=1}^n f(y_i) f(x_i | y_i, \gamma_0, \gamma_1) f(x_i^* | x_i, y_i, S_n, S_p)$$

always observed
leave unmodelled ✓
an assumption

$$Pr(x=1 | y=y) = \begin{cases} \gamma_0 & \text{if } y=0 \\ \gamma_1 & \text{if } y=1 \end{cases}$$

$$Pr(x^*=x | x=x) = \begin{cases} S_n & \text{if } x=1 \\ S_p & \text{if } x=0 \end{cases}$$

sensitivity &
specificity

Pause: Missing at Random assumption here?
Interpretation?

In concept could add

$$R = \begin{cases} 1 & \text{if } X \text{ observed} \\ 0 & \end{cases}$$

but this is ignorable if $R \perp\!\!\!\perp X \mid X^*, Y$

Investigator allowed to decide
on what proportion of
subjects to validate within
each (X^*, Y) stratum

Code-up the generative model

```
genmod.string <- "model{
```

```
### prior distribution
```

```
gamma.0 ~ dunif(0,1)
```

```
gamma.1 ~ dunif(0,1)
```

```
sens ~ dunif(0.5, 1)
```

```
spec ~ dunif(0.5, 1)
```

] rules out X^* being
worse than a coin
flip as
a measurement
of X

```
trgt <- logit(gamma.1)-logit(gamma.0)
```

```
### statistical model
```

```
for (i in 1:n) {
```

```
  x[i] ~ dbern(pr.x[i])
```

```
  pr.x[i] <- (1-y[i])*gamma.0 + y[i]*gamma.1
```

```
  xstr[i] ~ dbern(pr.xstr[i])
```

```
  pr.xstr[i] <- (1-x[i])*(1-spec) + x[i]*sens
```

```
}
```

```
}"
```

Turn the crank

```
### generative model, data go in
mod <- jags.model(textConnection(genmod.string),
  data=list(x=dta$x, y=dta$y, xstr=dta$xstr,
    n=dim(dta)[1]),
  n.chains=3)

update(mod, 2000)  ### burn-in

### MC output comes out
opt.JAGS <- coda.samples(mod, n.iter=10000, thin=1,
  variable.names=c("gamma.0", "gamma.1", "sens", "spec", "trgt"))
```

Our answer

```
MCMCsummary(opt.JAGS)
```

##		mean	sd	2.5%	50%	97.5%	Rhat	n.eff
##	gamma.0	0.418	0.0459	0.326	0.418	0.506	1	743
##	gamma.1	0.653	0.0503	0.554	0.652	0.752	1	1106
##	sens	0.675	0.0388	0.600	0.674	0.752	1	1038
##	spec	0.740	0.0419	0.658	0.739	0.821	1	810
##	trgt	0.975	0.2414	0.539	0.958	1.487	1	1501

1
 $\log \text{OR}(X, Y)$

Computationally frustrating (ess / wall-time) - Collapse?

E.g., for the unvalidated controls:

$$Pr(X^*=1 | Y=0) = Pr(X=1, X^*=1 | Y=0) + Pr(X=0, X^*=1 | Y=0)$$

$$= \gamma_0 s_\eta + (1-\gamma_0)(1-s_p)$$

Code this up

```
genmod.clps.string <- "model{  
  ### prior distribution  
  gamma.0 ~ dunif(0,1)  
  gamma.1 ~ dunif(0,1)  
  sens ~ dunif(0.5, 1)  
  spec ~ dunif(0.5, 1)  
  
  s.0 ~ dbin(gamma.0, nv.0)  
  t.00 ~ dbin(1-spec, nv.0-s.0)  
  t.01 ~ dbin(sens, s.0)  
  
  s.1 ~ dbin(gamma.1, nv.1)  
  t.10 ~ dbin(1-spec, nv.1-s.1)  
  t.11 ~ dbin(sens, s.1)  
  
  u.0 ~ dbinom(pr.0, nu.0)  
  pr.0 <- (1-gamma.0)*(1-spec) + gamma.0*sens  
  
  u.1 ~ dbinom(pr.1, nu.1)  
  pr.1 <- (1-gamma.1)*(1-spec) + gamma.1*sens
```

validated
controls

as per
prev.
slide

Turn the crank

```
### generative model, data go in
mod.clps <- jags.model(textConnection(genmod.clps.string),
  data=list(u.0=535, nu.0=535+701,
            u.1=375, nu.1=375+318,
            s.0=16+16, nv.0=16+16+33+11,
            t.00=11, t.01=16,
            s.1=5+18, nv.1=5+18+13+3,
            t.10=3, t.11=18),
  n.chains=3)
  mod.clps
update(mod, 2000) ### burn-in

### MC output comes out
opt.clps.JAGS <- coda.samples(mod.clps, n.iter=10000, thin=1,
  variable.names=c("gamma.0", "gamma.1", "sens", "spec", "trgt"))
```

Get an answer

```
MCMCsummary(opt.clps.JAGS)
```

##		mean	sd	2.5%	50%	97.5%	Rhat	n.eff
##	gamma.0	0.422	0.0559	0.315	0.422	0.533	1	20595
##	gamma.1	0.642	0.0613	0.518	0.644	0.758	1	6008
##	sens	0.677	0.0409	0.600	0.676	0.760	1	5812
##	spec	0.729	0.0553	0.615	0.730	0.831	1	7086
##	trgt	0.913	0.3566	0.221	0.912	1.621	1	8538

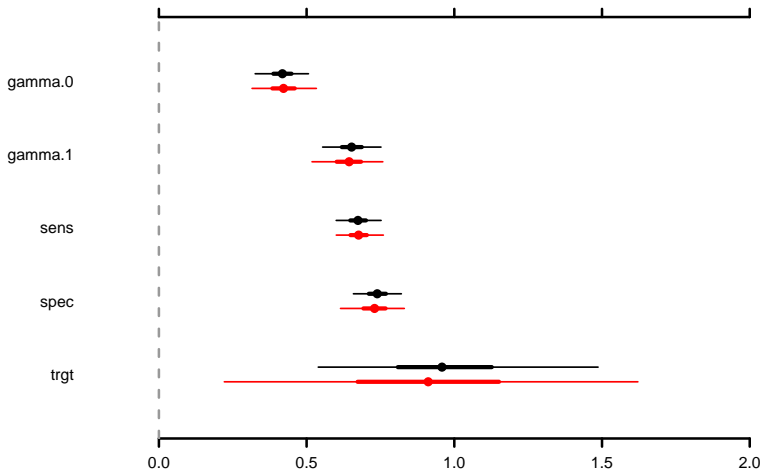
better

Sanity check: Two computational approaches going after **the** posterior distribution

```
MCMCplot(opt.JAGS, opt.clps.JAGS)
```

augmented

collapsed



Putting the inference in context

Have estimated the log odds⁵-ratio describing the (X, Y) association to be 0.91 (posterior mean), with an uncertainty estimate 0.36 (posterior SD).

- Contrast to complete-case analysis?

slide 7

reduced posterior SD ✓

- Contrast to pretending X^* is the gold-standard?

- increased posterior SD ✓*
- point estimate pushed away from the null ✓*

Many things that could be followed up on here

- ▶ Generality of idea: How to make the best use of (X^*, Y) data when the relationship between X and Y is of interest.
- ▶ Computation: Tradeoff in collapsing.
- ▶ Assumptions to be considered: we have invoked $(X^* \perp Y|X)$.
- ▶ Study design: If you were given a budget, how would you trade-off total number of patients versus number validated?