KU Leuven Summer School
Segment 4
Misclassification and COVID tests

Paul Gustafson

September 16, 2022

A "fun" research project back Spring 2020

Burstyn *et al. BMC Medical Research Methodology* (2020) 20:146
https://doi.org/10.1186/s12874-020-01037-4

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Towards reduction in bias in epidemic curves due to outcome misclassification through Bayesian analysis of time-series of laboratory test results: case study of COVID-19 in Alberta, Canada and Philadelphia, USA

Igor Burstyn[1,2*], Neal D. Goldstein[2] and Paul Gustafson[3]

(Last datapoint: March 27, 2020)

(First version posted to medRxiv: April 11, 2020)

Paper: doi.org/10.1186/s12874-020-01037-4

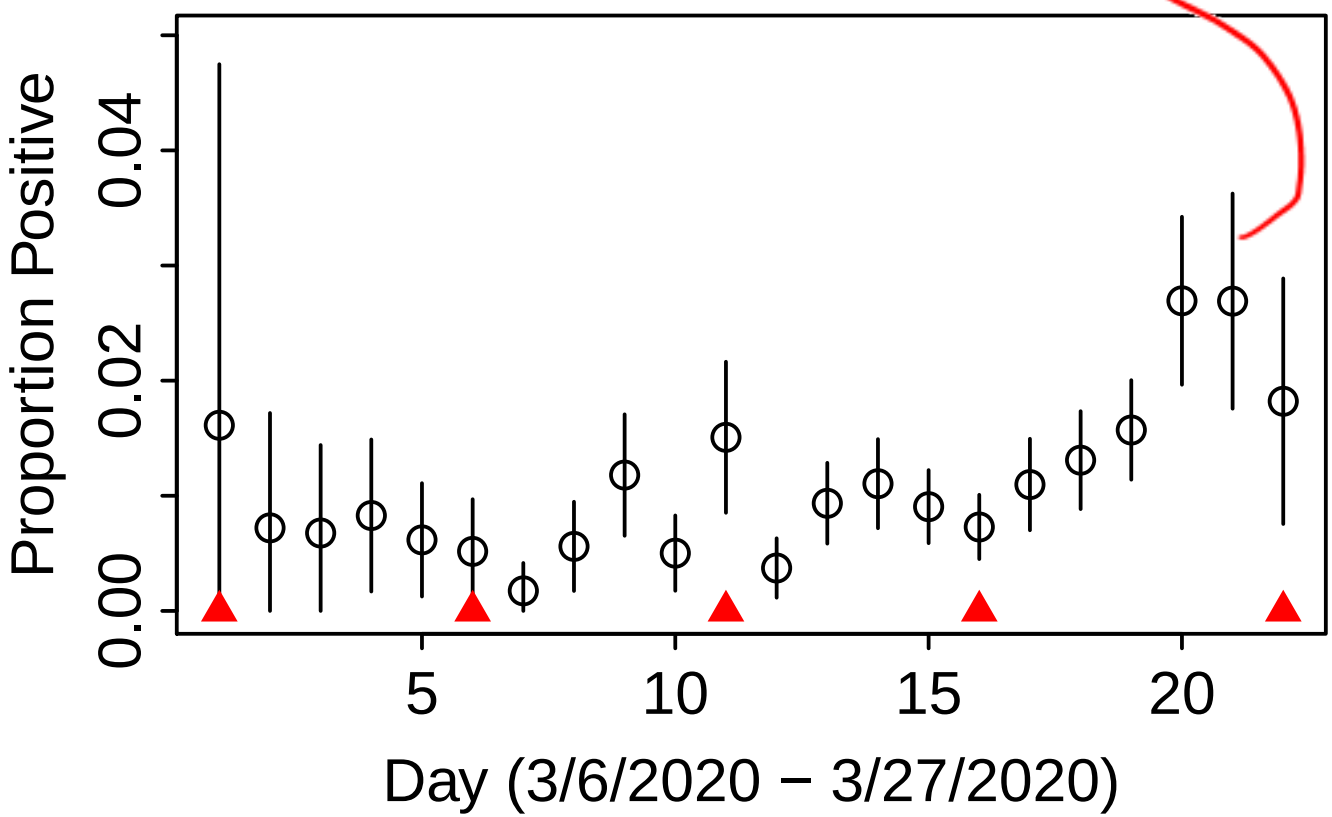
Data and JaGS code:
github.com/paulgstf/misclass-covid-19-testing

Alberta daily testing data

Y_t^* out of n_t tests came back positive on day t.

As a proportion: Y_t^*/n_t: $= \hat{p}_t$

$\pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n_t}}$



PCR test for current covid infection

Test result (-/+) imperfect surrogate for true infection status (-/+)

Assume that amongst the t-th day testing population, the test has *specificity* Sp, but *sensitivity* Sn_t.

Why is it that (perhaps!) static specificity, but time-varying sensitivity, is appropriate (as a prior assertion)

$\Pr\{\text{test negative} \mid \text{uninfected}\}$

$\Pr\{\text{test positive} \mid \text{infected}\}$

Nature of the PCR test (very, very roughly, 1 of 2)

Nasal swab looking for virus particles

How can a false positive arise?

Cross-contamination of the swab (seriously, apparently). So "lab quality" issue, no particular reason to think specificity time-varying.

Nature of the PCR test (very, very roughly, 2 of 2)

Nasal swab looking for virus particles

How can a false negative arise?

There is a little bit of virus in the nasal cavity, but the swab misses it. So maybe...

At one period of time, only people with heavy respiratory symptoms are eligible to get a test. Amongst such people, the true positives were likely infected a while back, hence lots of virus to find, hence higher sensitivity.

At another period of time, testing is used to screen a population that's largely asymptomatic. At least some true positives within this population will be very recently infected, hence less virus to find, hence lower sensitivity.

Statistical model (observables and latents, given params)

$$f(y^*_{1:T}, y_{A,1:T}, y_{B,1:T} y_{1:T} | r_{1:T}, Sn_{1:T}, Sp) =$$

$$\prod_{t=1}^{T} f(y_t|r_t) f(y_{A,t} \mid y_t, Sn_t) f(y_{B,t} \mid y_t, Sp) f(y^*_t | y_{A,t}, y_{B,t})$$

Parameters: Sp, $Sn_{1:T}$ (as already defined), and $r_{1:T}$, where r_t is the population prevalence of infection amongst the day t testing pool.

Latents:

Y_t is the number (out of the n_t tested that day) that are *truly* infected. So $(Y_t|r_t) \sim Bin(n_t, r_t)$.

$Y_{A,t}$ is the number of truly infected who test positive. So $(Y_{A,t}|Y_t, Sn_t) \sim Bin(Y_t, Sn_t)$.

$Y_{B,t}$ is the number of truly uninfected who test positive. So $(Y_{B,t}|Y_t, Sn_t) \sim Bin(n_t - Y_t, 1 - Sp)$.

Statistical model, continued

Observables:

Y_t^* is the number (that day) who test positive. So $(Y_t^* | Y_{A,t}, Y_{B,t})$ is deterministic, simply $Y_t^* \equiv Y_{A,t} + Y_{B,t}$.

And prior distributions for parameters

e.g. knots

very simple spline

- $r_{1:T}$ piecewise-linear in time,
  - treating $(r_1, r_6, r_{11}, r_{16}, r_{22})$ as the five unknown parameters,
  - each of which is, independently, ascribed a $Unif(0, 0.5)$ prior.
- Very roughly (but see upcoming slide for actual), $Sn_{1:T}$ is treated this same way.
  - each of $(Sn_1, Sn_6, Sn_{11}, Sn_{16}, Sn_{22})$ ascribed a $Unif(0.6, 0.9)$ prior.
- $Sp \sim Unif(0.95, 1)$.

wide relative to scientific context

non-crazy representations of knowledge about the test circa March 2020

Example JAGS coding, prior for $r_{1:22}$

```
### prevalence parameterized by value at knots
for (i in 1:num.kn) {
  r.kn[i] ~ dunif(0, r.hi[i])
}

### these imply the daily values
for (i in 1:(num.kn-1)) {
  for (j in 0:(spc.kn[i]-1)) {
      r[knts[i]+j] <- ((spc.kn[i]-j)*r.kn[i]+j*r.kn[i+1])/
                      (spc.kn[i])
  }
}
r[knts[num.kn]] <- r.kn[num.kn]
```

linear interpolation for days in between knots

Example JAGS coding ... latents + observables given parameters

```
for (i in 1:(knts[num.kn])) {
  y[i] ~ dbinom(r[i], n[i])
  ya[i] ~ dbinom(sn[i], y[i])
  yb[i] ~ dbinom(1-sp, n[i]-y[i])
  ystr[i] ~ sum(ya[i], yb[i])
}
```

truly infected

~~true positives~~

correct positives

false positives

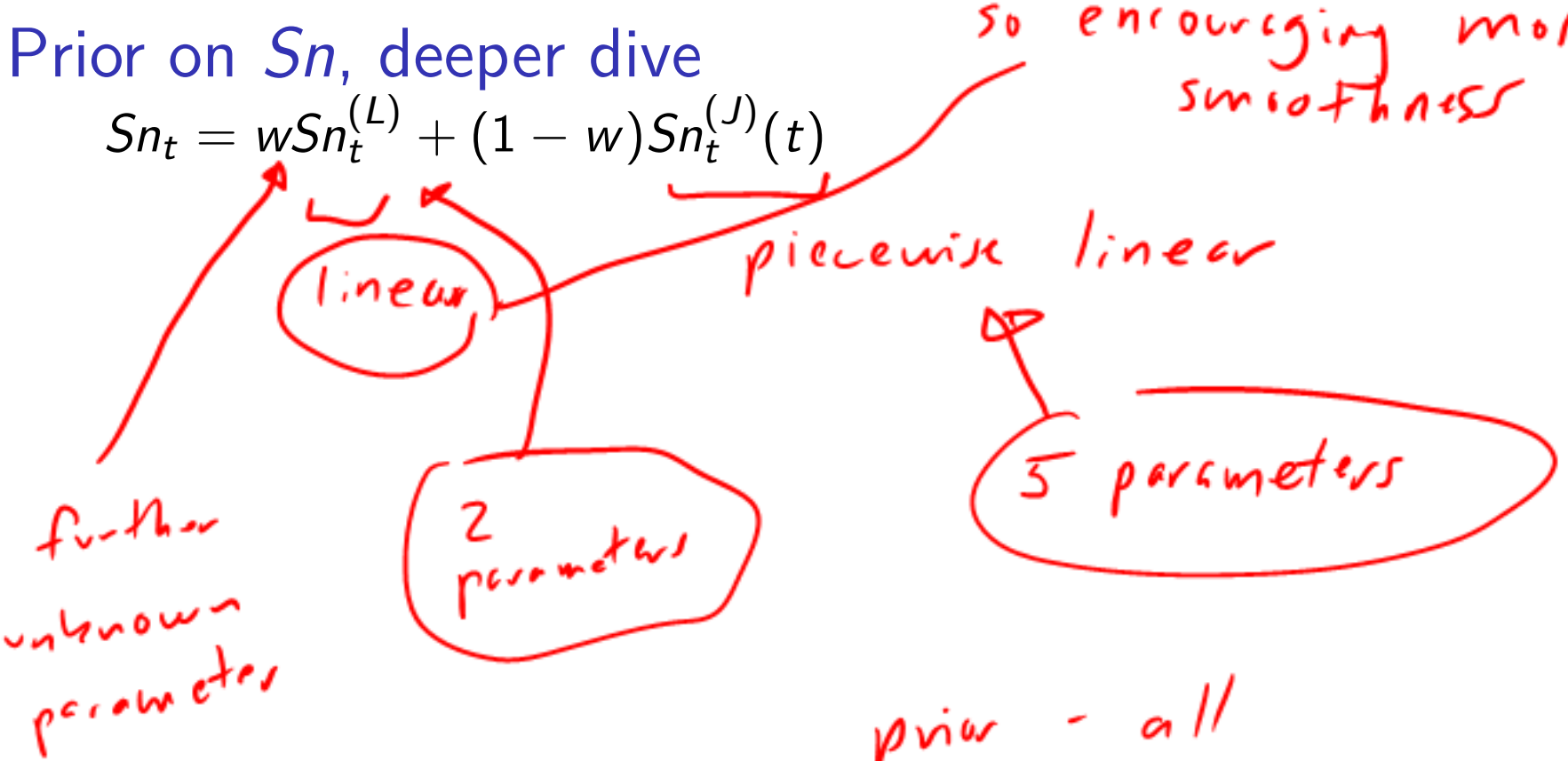
true

JAGS representation

observable = sum of two latents

Prior on Sn, deeper dive

$$Sn_t = wSn_t^{(L)} + (1-w)Sn_t^{(J)}(t)$$



prior - all 8 params indep. of one another

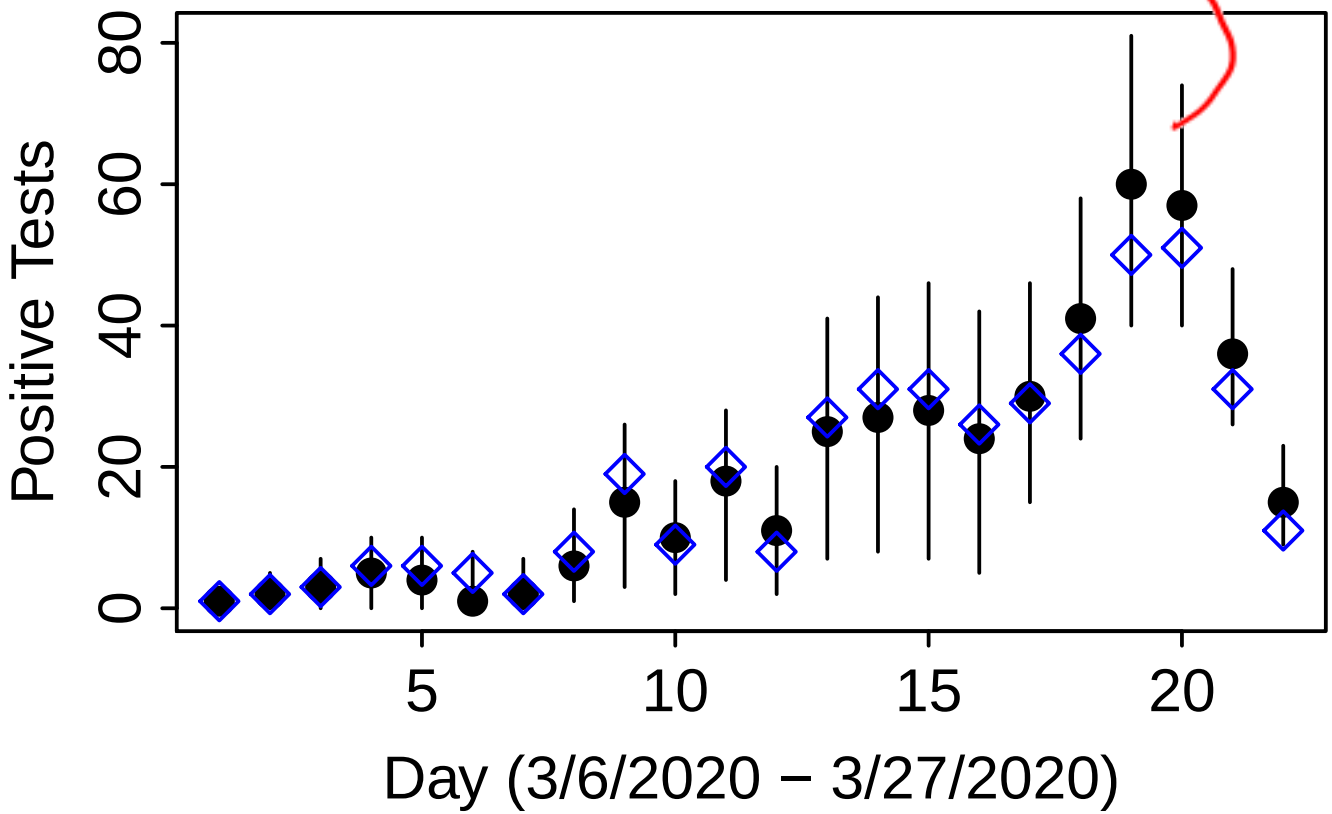
Deeper dive continued

- Not showing you this because it's necessarily exemplary
- Indeed, the works was done in haste, and in hindsight there may have been other (desirable) prior specifications to handle the time-varying prevalence and sensitivity.
- Am showing you this as an example of "Can we build this? Yes we can!"
  - We could dream up a probability distribution for $(r_1, \ldots, r_{22}, Sn_1, \ldots, Sn_{22}, Sp)$.
  - We could express this in JAGS code and press go.

Primary analysis: inference on true number of ~~positives~~ infected amongst those tested per day

E.g., inference about latent $Y_{1:T}$ rather than simply reporting the oberved counts $Y^*_{1:t}$

posterior mean and 95% credible interval for Y_{20}



Sidenote: Augmenting versus Collapsing

could have collapsed

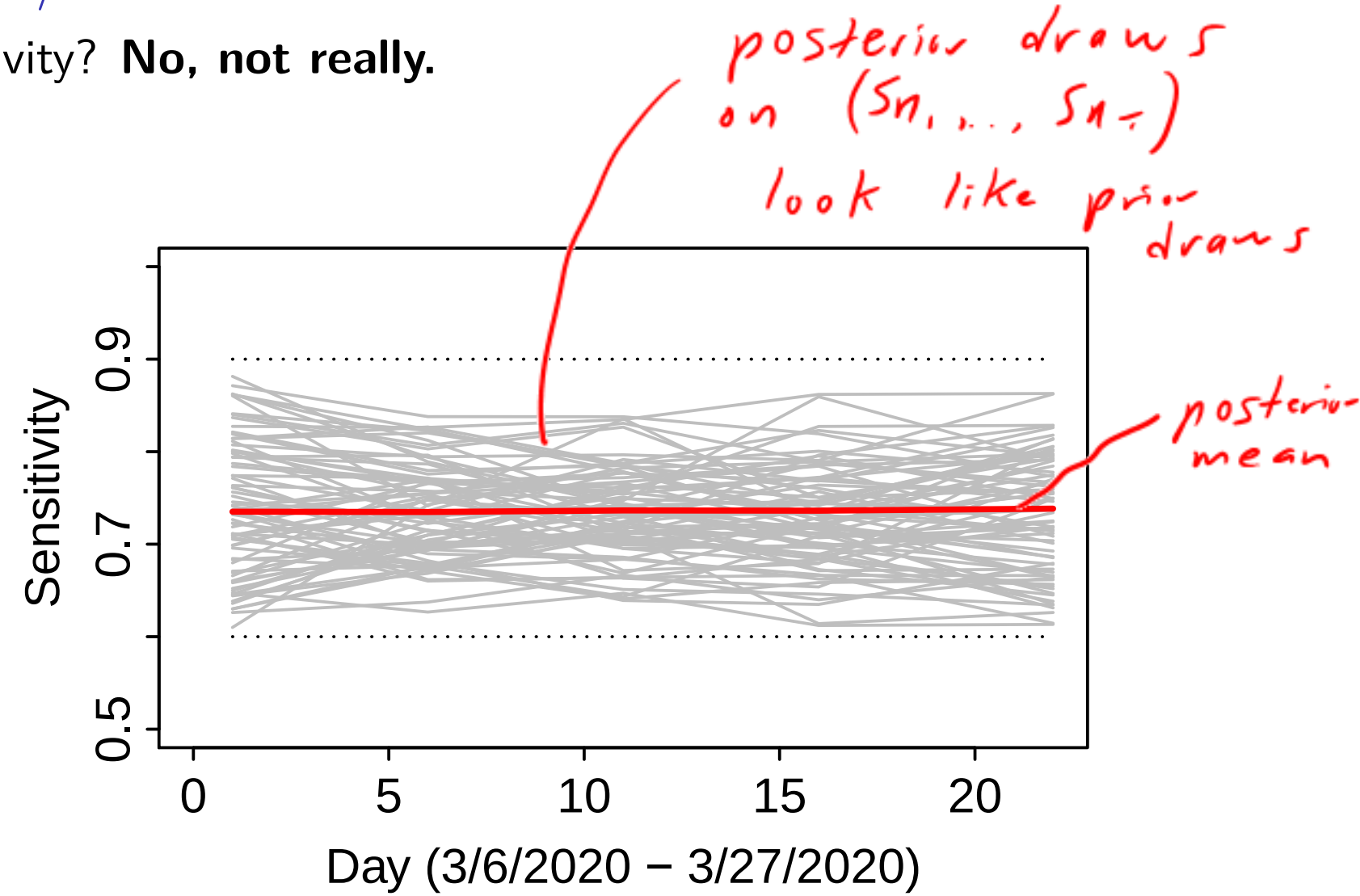
$$y_t^* \sim \text{Bin}\left(n_t, (1-r_t)(1-sp) + r_t s_{h_t}\right)$$

but then couldn't easily access inference about y_t (prev. slide)

whereas that inference is "automatic" with an augmented implementation

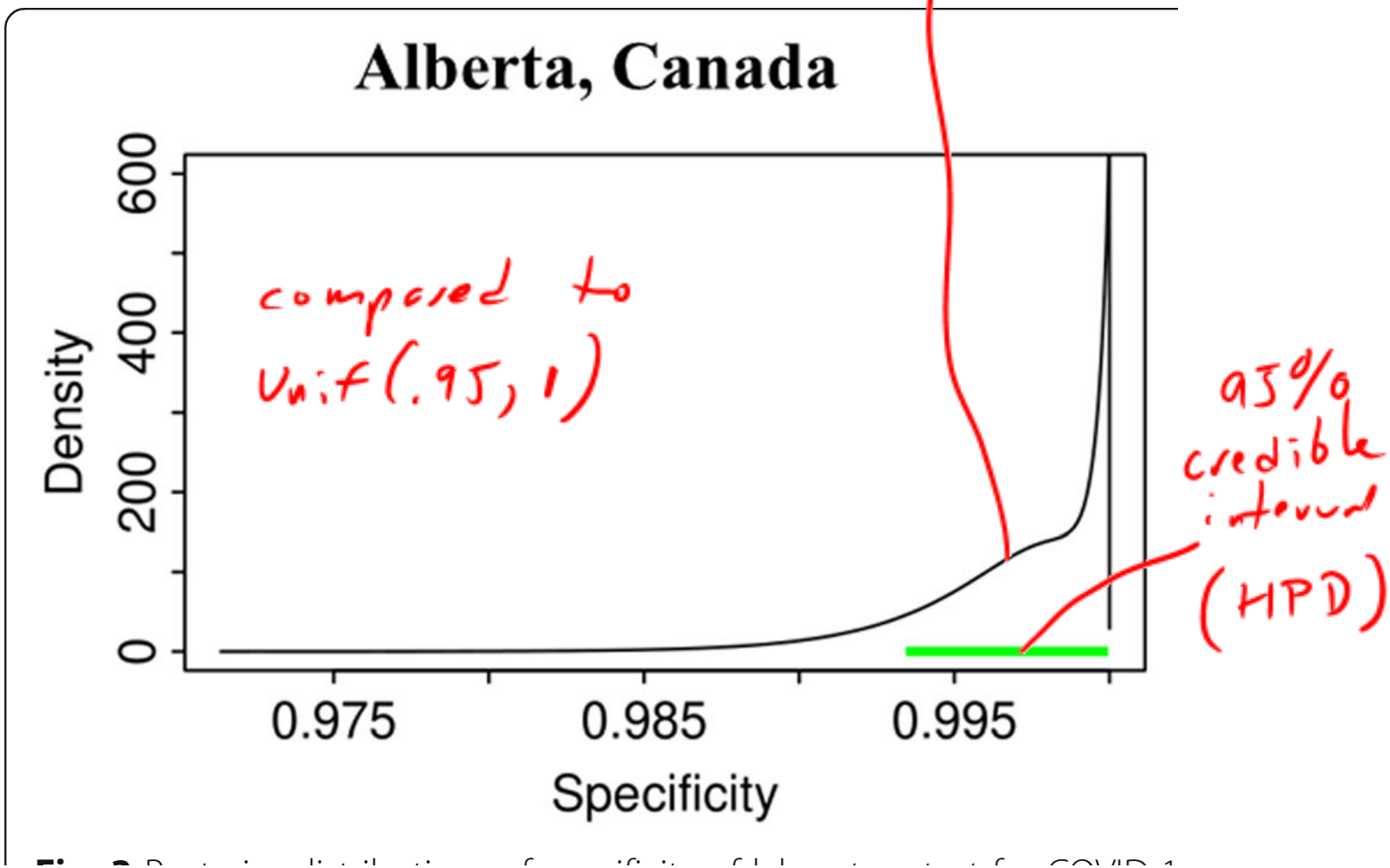
Secondary analysis: Do the data supply any info about how good/bad the PCR test is?

Sensitivity? **No, not really.**



Secondary analysis, continued: Do the data supply any info about how good/bad the PCR test is?

Specificity? **Yes.**



Intuition for why the data are so quiet concerning $Sn_{1:T}$?

Consider slightly simpler situation : $Sp \equiv 1$

$\rightarrow \quad y_t^* \sim Bin\left(n_t, r_t \, Sn_t\right)$

r_t: $\Pr\{\text{truly infected}\}$ $\quad$ Sn_t: $\Pr\{\text{test +} \mid \text{truly infected}\}$

Maybe can estimate this product $r_t Sn_t$ well, but no info to separate out the two terms

More space for intuition?

Intuition for why the data are quite loud concerning Sp?

Let $r_t^* = Pr\{\text{test positive}\}$ ← for an individual from day t testing pool

$= r_t Sn_t + (1-r_t)(1-Sp)$

Hence $\min\{Sn_t, 1-Sp\} < r_t^* < \max\{Sn_t, 1-Sp\}$

and our prior refines this to:

$1-Sp < r_t^* < Sn_t$

So $Sp > 1 - r_t^*$ for all t

Now back to data ...

More space for intuition?

strong evidence that at least on a couple of days (day 7 in particular) r_t^* is very close to zero

Therefore have evidence that Sp is very close to one

- can understand the information flow