Gold standards are out and Bayes is in: Implementing the cure for imperfect reference tests in diagnostic accuracy studies

Wesley Johnson
UC Irvine

Bayesian Statistics 2018

Diagnostic Screening

- Test for disease or infection, $D+$, in humans or animals
- Test outcomes are $T+$ and $T-$ for dichotomous tests
- Test accuracies are

  Sensitivity: $Se = Pr(T+ \mid D+)$ Specificity: $Sp = Pr(T- \mid D-)$
- Gold Standard tests ($Se = Sp = 1$) are often unavailable or relatively expensive
- Less expensive but imperfect diagnostic outcome measures are available eg serology, fecal culture, microscopy etc.
- General goals are to assess test accuracy, estimate prevalence(s) and diagnose individual subjects

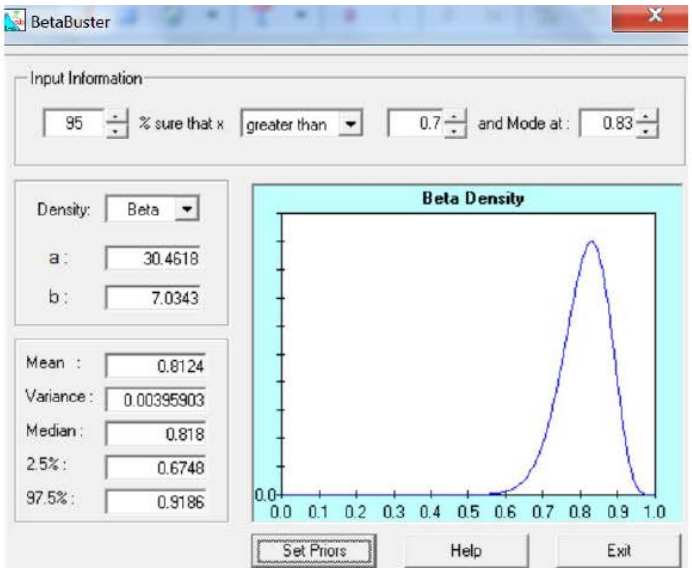
Testing for Toxoplasmosis with MAT and ELISA

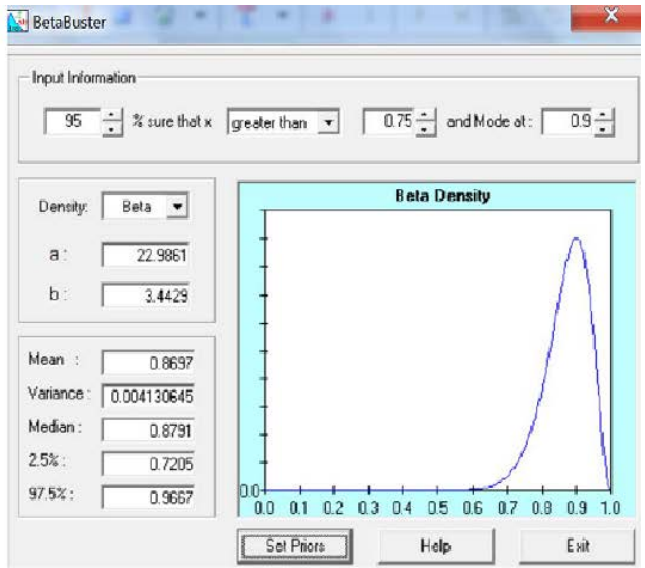
- Data from evaluation of serologic tests for T.gondii in 998 naturally-infected sows, Dubey et al. (1995) AJVR 56: 1030-1036
- Tests considered were MAT and ELISA
- Move to replace MAT with ELISA because ELISA can be automated, yields results more rapidly and is amenable to use in mass screening of pigs

Prior Specification for *SeMAT* and *SpMAT*

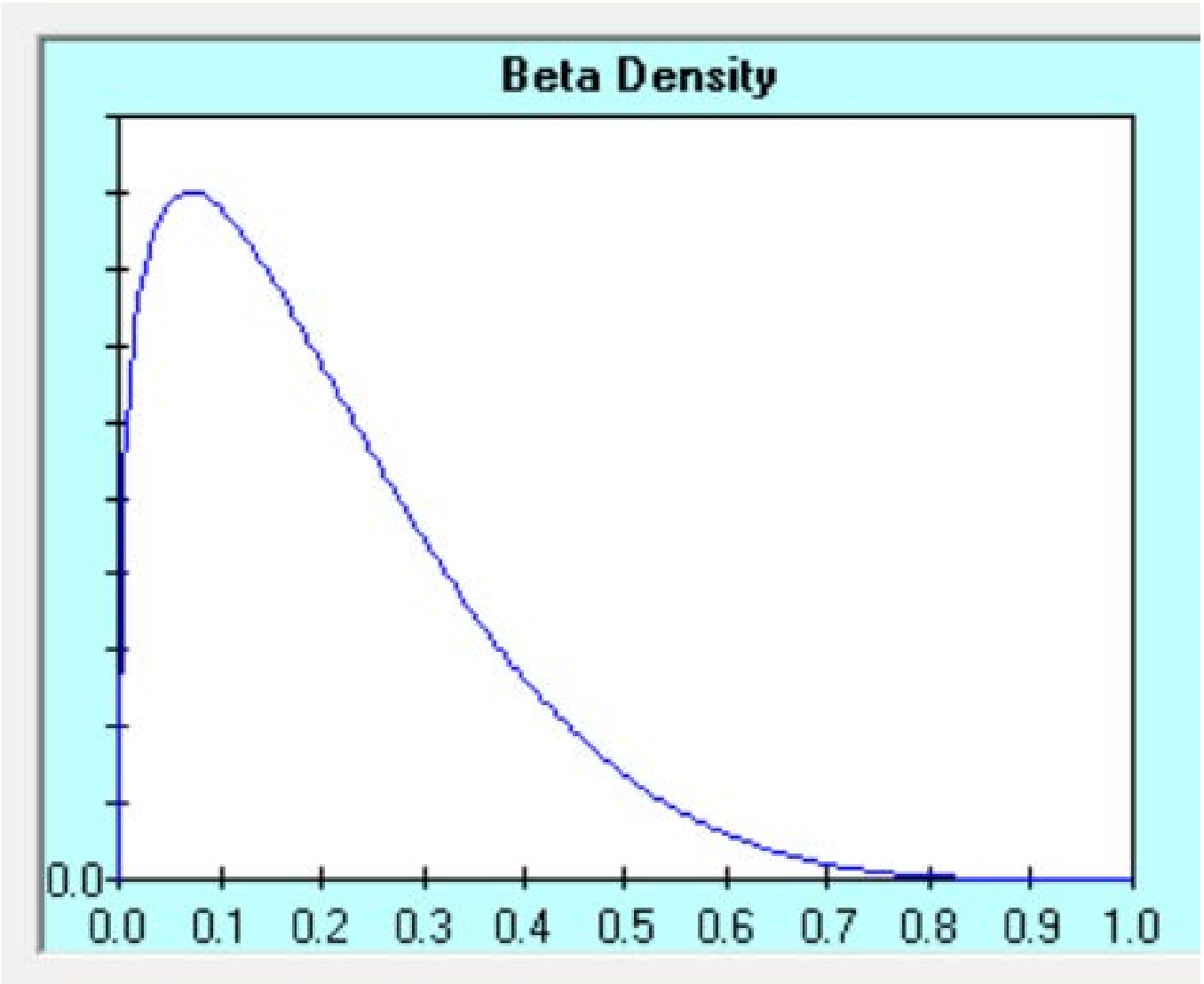
- Need prior information about test accuracy for at least one test eg. MAT
- Best estimate/guess: $SeMAT = 0.83$
- How uncertain/certain? $Pr(SeMAT > 0.70) = 0.95$
- Best estimate/guess $SpMAT = 0.90$
- How uncertain/certain? $Pr(SpMAT > 0.75) = 0.95$

Priors for *SeMAT* and *SpMAT*





Prior for *TPrev* of Toxoplasmosis



Testing for Toxoplasmosis with MAT and ELISA

- Data from Dubey et. al. (1995); analyzed in Georgiadis et. al. (2003)

	$ELISA+$	$ELISA-$	
$MAT+$	164	58	222
$MAT-$	77	699	776
	241	757	998

- We first just considered the solo data on MAT;

  222 $MAT+$ out of 998 tested

- Then solo data on *ELISA*

  241 $ELISA+$ out of 998 tested

- Each count, y, is Binomial($n = 998, Ap$) with

$$Ap = Tprev * Se + (1 - Tprev) * (1 - Sp)$$

- The natural estimate of Ap is y/n. So we could equate

$$222/998 = 0.22 \doteq Tprev * SeMAT + (1 - Tprev) * (1 - SpMAT)$$

Testing for Toxoplasmosis with MAT and ELISA

- So if $SeMAT = 0.83$, $SpMAT = 0.90$, we have

  $0.22 \doteq Tprev * 0.83 + (1 - Tprev) * (0.1) = 0.73 * Tprev + 0.1,$

  which implies that $Tprev \doteq 0.16$

- On the other hand if our estimate of $SpMAT < 0.78$, the (non-Bayesian) estimate of prevalence is negative
  eg. prior information is inconsistent with the data

- The Bayesian approach has the advantage that probability laws cannot be violated when proper probabilities are specified for prior inputs

- A major constraint at this point is due to the fact that the data only have information for one parameter, the *Ap*

- Yet we currently have 3 parameters.
  The simple binomial model lacks identifiability

Testing for Toxoplasmosis with MAT and ELISA

- The Bayesian approach combines information in the data, y, with scientific (prior) information,

$$p(Tprev, Se, Sp) = p(Tprev)\, p(Se)\, p(Sp)$$

  which we express as the product of independent beta functions (see previous betabuster pictures)

- The information in the data is expressed through the Likelihood function, which in this instance is $Lik(Tprev, Se, Sp) =$

$$[Tprev * Se + (1 - Tprev) * (1 - Sp)]^{y} [1 - Ap]^{n-y}$$

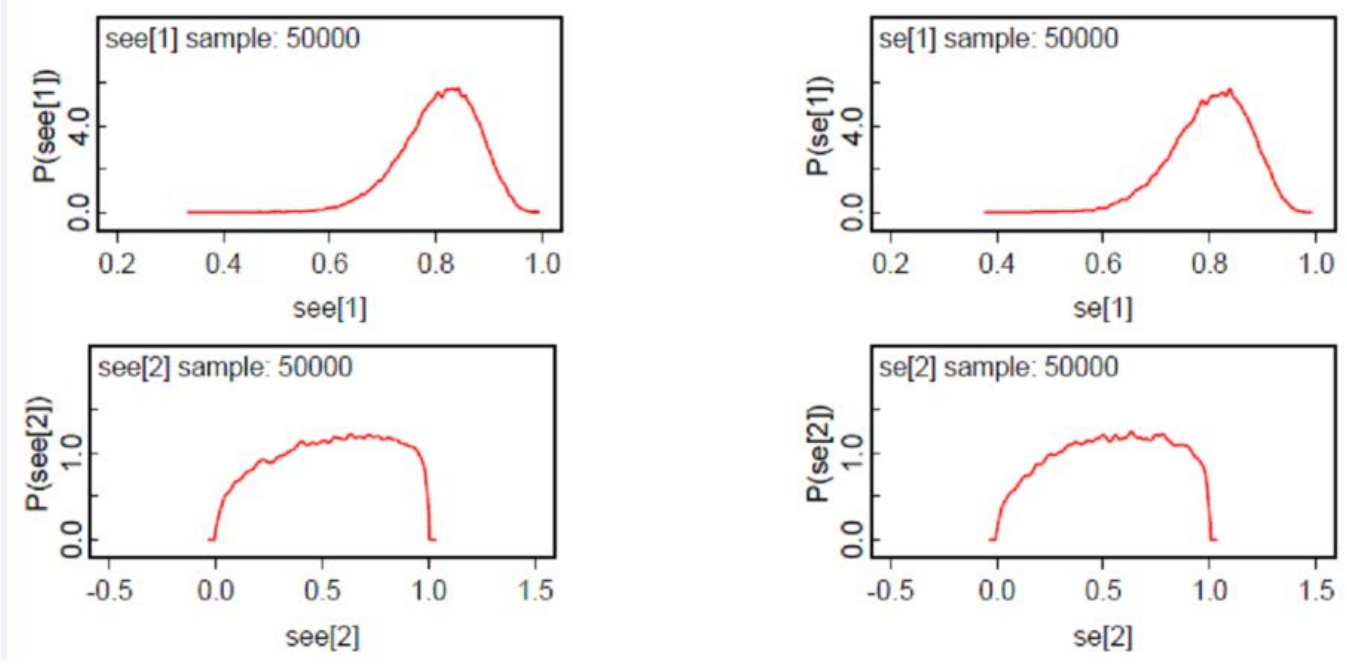
- Using Bayes Theorem, Inferences are based on the *posterior*

$$p(Tprev, Se, Sp \mid data) \propto p(Tprev, Se, Sp) Lik(Tprev, Se, Sp)$$

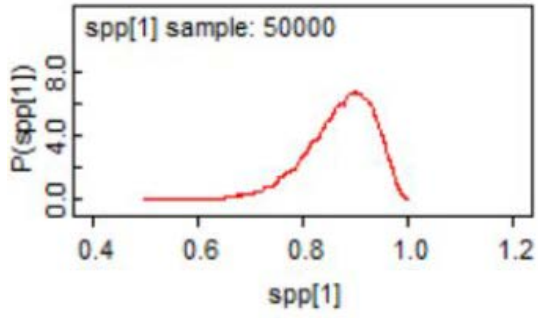
Priors for *SeELISA* and *SpELISA*

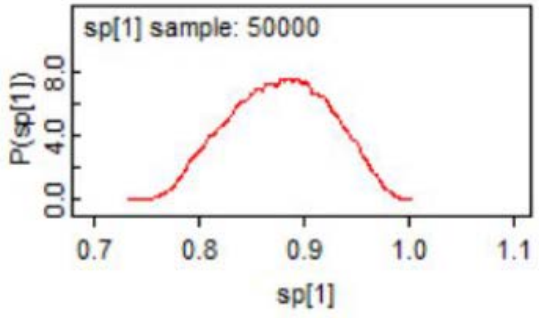
- We specified very diffuse distributions
- *SeELISA*: 95% Prior probability interval (0.06, 0.97)
- *SpELISA*: 95% Prior probability interval (0.14, 0.98)

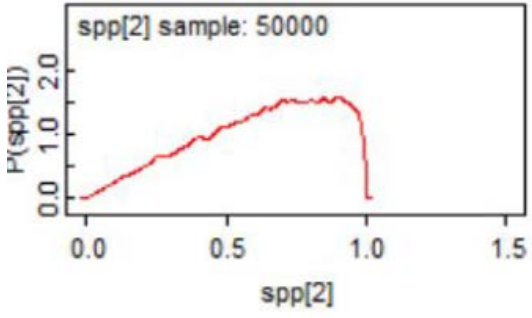
Prior (left) and Posterior (right) for SeMAT and SeELISA: Solo

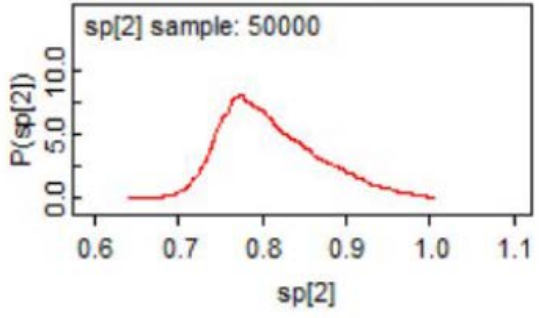


Prior (left) and Posterior (right) for SpMAT and SpELISA: Solo









Testing for Toxoplasmosis with MAT and ELISA: 2×2 data: Conditional Independence (CI) Model

- We now use the full 2×2 table of data

	ELISA+	*ELISA*−	
MAT+	164	58	222
MAT−	77	699	776
	241	757	998

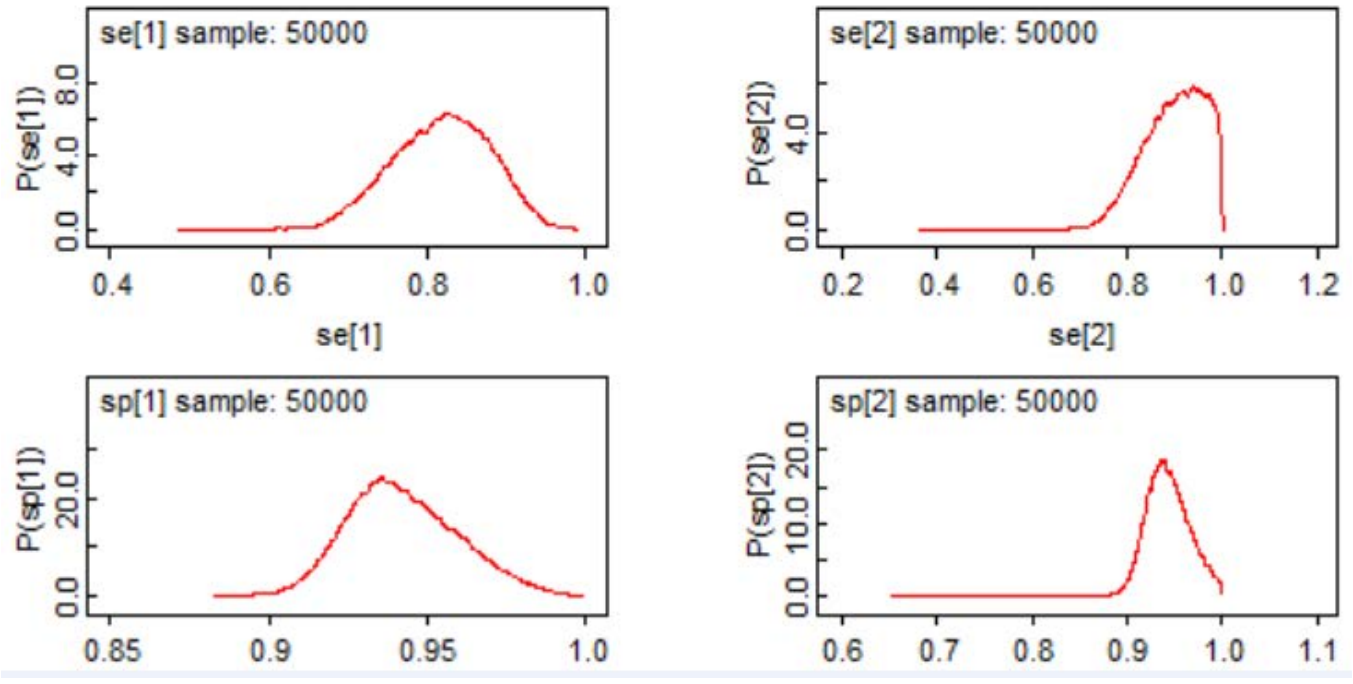
- The data table is multinomial with probabilities for each cell that are analogous to the simple binomial case. The cell probabilities add to one, and the cell counts add to n
- The CI model involves the assumption that

$$Pr(ELISA+ \mid MAT[\pm], Toxo+) = Pr(ELISA+ \mid Toxo+) = SeE$$

$$Pr(ELISA- \mid MAT[\pm], Toxo-) = Pr(ELISA- \mid Toxo-) = SpE$$

- Tests have similar biological bases so CI is probably false

Posteriors for Se-SpMAT and Se-SpELISA: 2×2 Table Data: Cond Indep Model



Testing for Toxoplasmosis with MAT and ELISA: 2×2 data: CI Model

- The 2×2 table has 3 degrees of freedom for estimation eg. we can estimate 3 complicated functions of all of the parameters
- The number of parameters we have to estimate is 5;
  2 *Se*, 2 *Sp*, and 1 *Tprev*
  We already have priors for everything from doing the Solo analyses
- The model still lacks identifiability. We are short 2 dof for estimating all 5 parameters
- But if the model is appropriate (eg. if CI assumption is ok), and if our prior/scientific information is accurate, our inferences are perfectly valid
- But, we really don't believe the CI assumption

Testing for Toxoplasmosis with MAT and ELISA: 2×2 data: Dependence Model

- We now model

$$Pr(ELISA+ \mid MAT+, Toxo+) = SeE_{M+}$$
$$Pr(ELISA+ \mid MAT-, Toxo+) = SeE_{M-}$$
$$Pr(ELISA- \mid MAT+, Toxo-) = SpE_{M+}$$
$$Pr(ELISA+ \mid MAT-, Toxo-) = SpE_{M-}$$

- The number of parameters we have to estimate is now 7, and we only have 3 dof
- The model still lacks identifiability. We are short 4 dof for estimating all 7 parameters.
- We need 4 priors to replace priors for *SeE* and *SpE*

Testing for Toxoplasmosis with MAT and ELISA: 2×2 data: Dependence Model

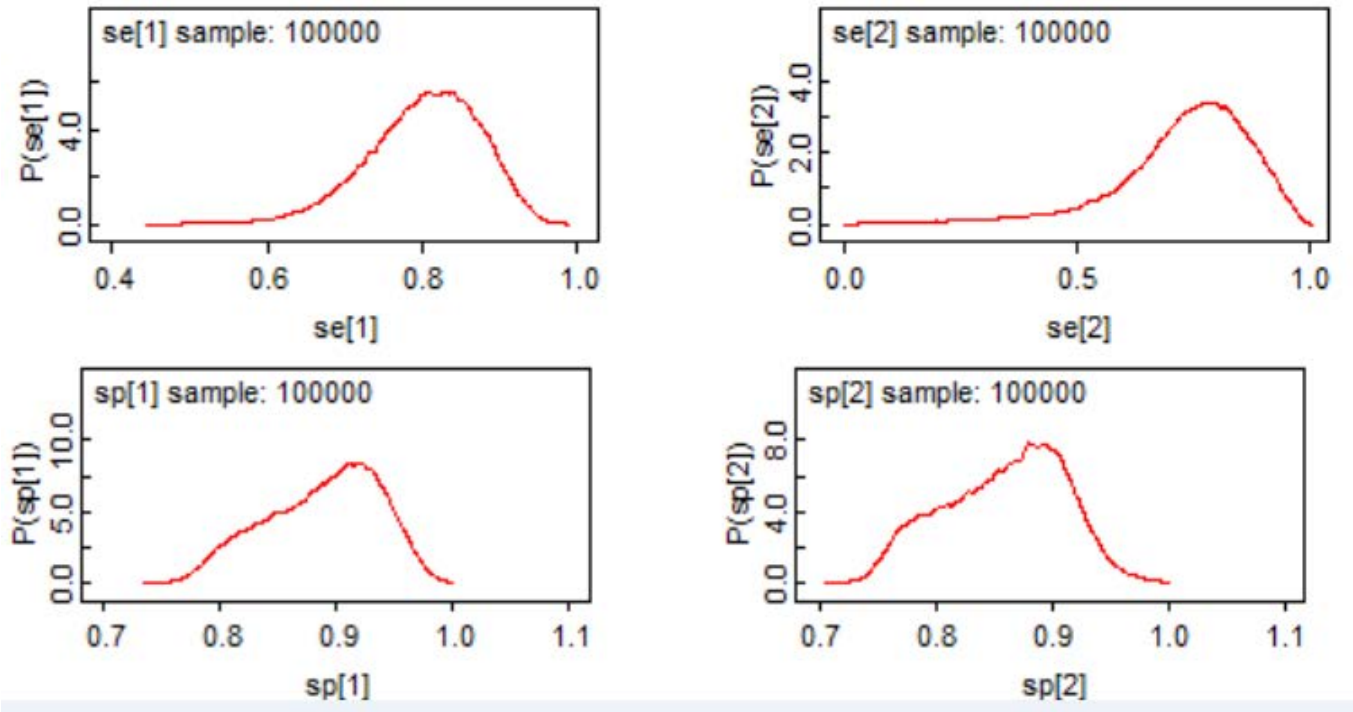
- Since we knew little about the ELISA *Se* and *Sp*, we knew even less about the 4 conditional ELISA *Se* and *Sp* values
- Thus we used exactly the same priors as before eg.

$$SeE_{M+} \sim SeE_{M-} \sim SeE$$

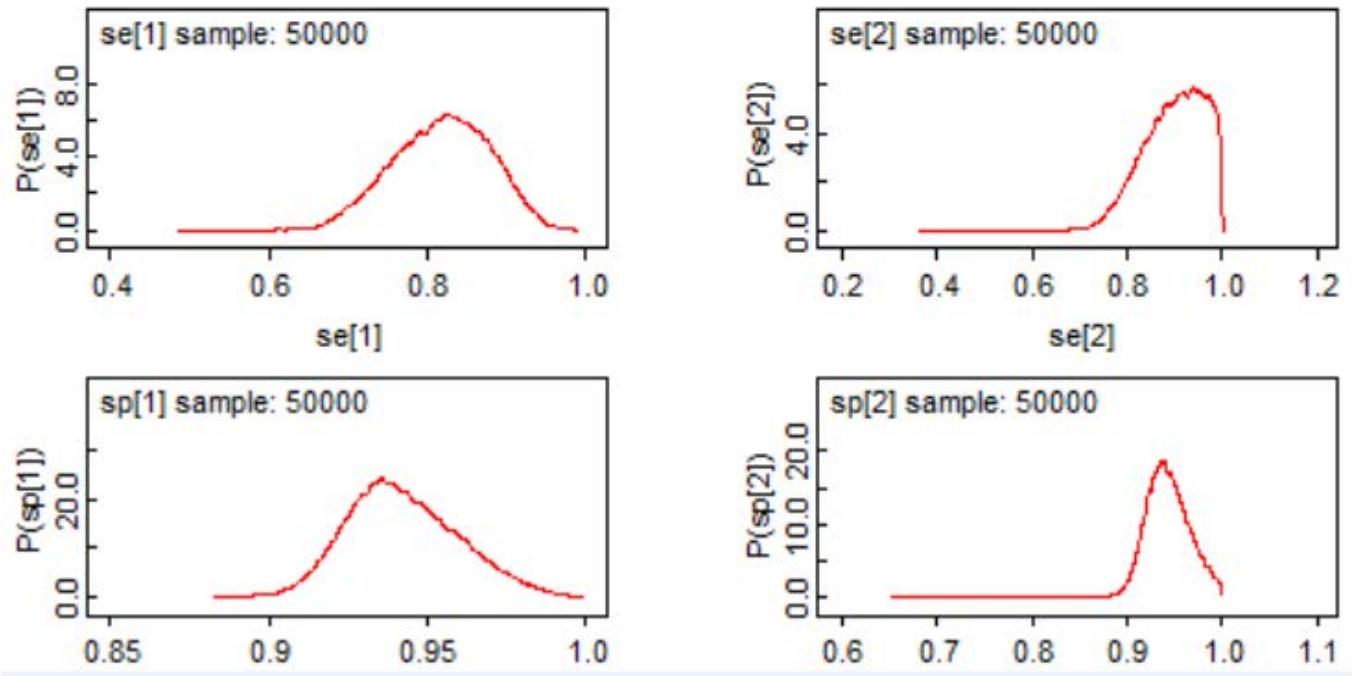
$$SpE_{M+} \sim SpE_{M-} \sim SpE$$

- This prior spec reflects a minor belief in the CI model as we specify the same beliefs about the conditional test accuracies as the non-cond test accuracies
- Specifically, we don't input any information that would infer that $SeE_{M+} > SeE_{M-}$, for example
- But the model does allow for Dep or CI as the data might suggest, in conjunction with all of the prior information

Posteriors for SeMAT, SpMAT and SeELISA, SpELISA: 2 × 2 Table: Dependence Model



Posteriors for SeMAT, SpMAT and SeELISA, SpELISA: 2×2 Table: Conditional Independence Model



Posterior Inferences: (i) Solo (ii) Two Cond Indep (iii) Two Dependent

[$Pr(rhoD > 0) = 0.79$ $Pr(RhoDc > 0 = 0.97)$]

	Solo Med (95% PI)	2 CI Tests Med (95% PI)	2 Dep Tests Med (95% PI)
SeMAT	0.80 (0.65, 0.92)	0.82 (0.70, 0.93)	0.81 (0.65, 0.92)
SeELISA	0.55 (0.06, 0.97)	0.91 (0.77, 0.994)	0.76 (0.36, 0.94)
SpMAT	0.88 (0.79, 0.96)	0.94 (0.91, 0.979)	0.89 (0.79, 0.97)
SpELISA	0.80 (0.72, 0.94)	0.94 (0.91, 0.989)	0.86 (0.76, 0.95)
PDiff(*Se*)	0.78	0.18	0.66
PDiff(*Sp*)	0.90	0.49	0.89

Testing with MAT and ELISA: 2×2 data: Dependence Model

- The two models give different results. Deviance Information Criterion for Dep model is 22.2 and for CI model is 36.4
- Substantial positive correlation between tests
- So we prefer the Dependence model
- But it likely still difficult to believe in a model that requires *so much* prior information
- Clearly more information from data is required

Three Tests, MAT-ELISA and MB: $2 \times 2 \times 2$ Data

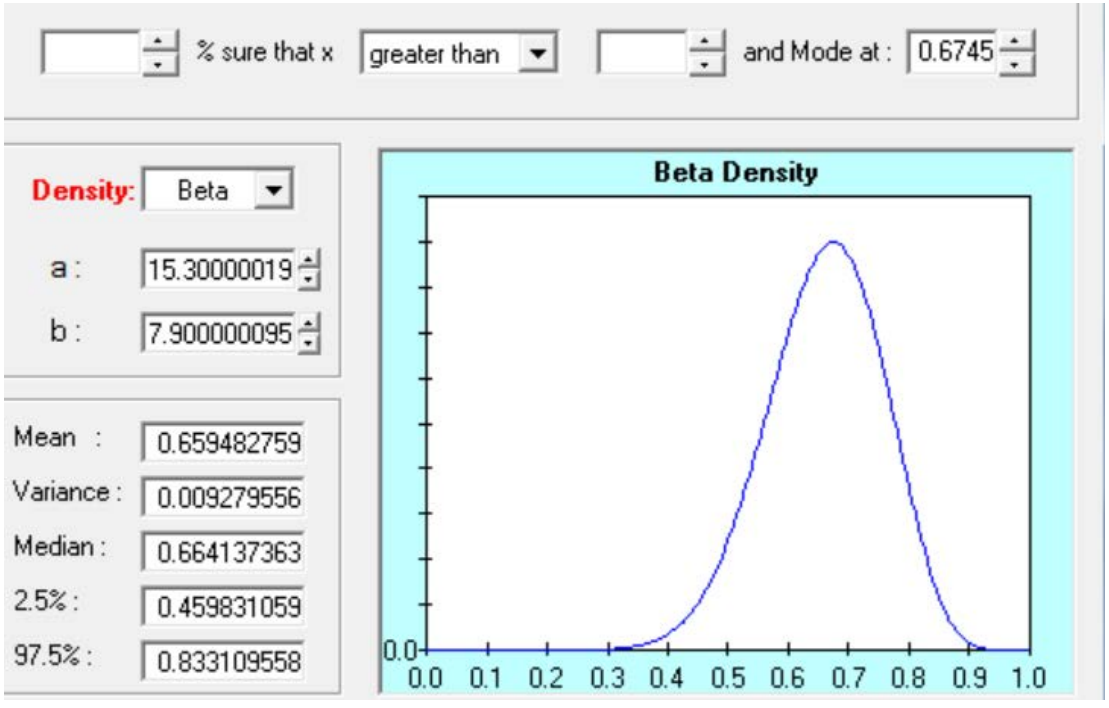
- We add a third test, Mouse Bioassay (MB)
- With three tests, we have a $2 \times 2 \times 2$ table of multinomial data with $n = 998$.
- There are thus $2^3 - 1 = 7$ degrees of freedom

	Mouse+		*Mouse*−	
	ELISA+	*ELISA*−	*ELISA*+	*ELISA*−
MAT+	73	17	91	41
MAT−	4	13	73	686

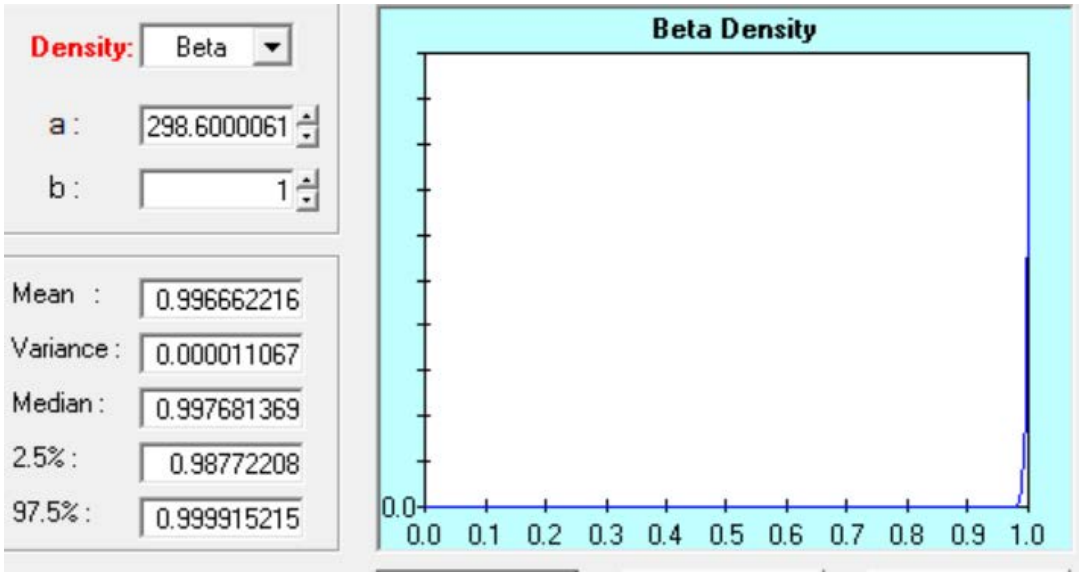
Three Tests, MAT-ELISA and MB: $2 \times 2 \times 2$ data: MAT-ELISA Dep, CI of MB

- We believe that MB will be CI of MAT and ELISA
- Thus we don't require any dependence parameters between MB and the other two
- The number of parameters to be estimated is thus: 1 *Tprev*, 2 (*SeMAT*, *SpMAT*), 4 conditional for (*SeE*, *SpE*), and two (*SeMB*, *SpMB*) for a total of 9 parameters; model lacks identifiability
- Need a prior for (*SeMB*, *SpMB*)
- If all three tests were CI, the model would be identifiable (Jones et al. 2010)

Prior for SeMB



Prior for SpMB



Posterior Inferences: (i) Two-tests, MAT-ELISA, Dep; (ii) Three-tests, MAT-ELISA, MB, Inf Prior; (iii) Three-tests, MAT-ELISA, MB, Mostly Diffuse Prior

	2 Dep Med (95% PI)	3 w/ 2 Dep; Inf Med (95% PI)	3 w/ 2 Dep; Diffuse Med (95% PI)
SeMAT	0.81 (0.65, 0.92)	0.85 (0.78, 0.92)	0.86 (0.77, 0.96)
SeELISA	0.76 (0.36, 0.94)	0.73 (0.63, 0.82)	0.73 (0.63, 0.83)
SpMAT	0.89 (0.79, 0.97)	0.91 (0.86, 0.96)	0.93 (0.85, 0.99)
SpELISA	0.86 (0.76, 0.95)	0.86 (0.82, 0.91)	0.88 (0.81, 0.92)
Diff(*Se*)	0.05 (-0.17, 0.45)	0.12 (0.04, 0.20)	0.13 (0.05, 0.22)
PDiff(*Se*)	0.66	0.9982	0.9984
Diff(*Sp*)	0.03 (-0.02, 0.08)	0.05 (0.02, 0.07)	0.05 (0.02, 0.09)
PDiff(*Sp*)	0.89	0.9999	0.9999
SeMB	-	0.62 (0.47, 0.81)	0.53 (0.39, 0.95)
SpMB	-	0.998 (0.989, 1.0)	0.997 (0.986, 1.0)

Three Tests, One Population

- For “fun” I ran a model with all Unif(0,1) priors
- This is a really stupid prior since we know that all values between 0 and 1 are not equally plausible for test accuracy parameters, and especially not for *SpMB*
- When I ran the model, I got nonsense results since the model lacks identifiability in the absence of constraints
- The “mostly diffuse” prior in the previous table has all Unif(0,1) priors except for our informative prior on *SpMB*.

Two Population Version: $2 \times 2 \times 2 \times 2$ Data

- The data were split into two groups

Popn 2	*Mouse*+		*Mouse*−	
	ELISA+	*ELISA*−	*ELISA*+	*ELISA*−
MAT+	22	6	45	19
MAT−	2	1	39	327

Popn 1	*Mouse*+		*Mouse*−	
	ELISA+	*ELISA*−	*ELISA*+	*ELISA*−
MAT+	51	11	46	22
MAT−	2	12	34	359

Two Population Version: $2 \times 2 \times 2 \times 2$ Data

- For illustration, we assume that two independent samples of size 461 and 537 were taken
- This would result in two independent tables that are Multinomial($n_i, p(\{M\pm, E\pm, MB\pm\} \mid \text{Popn } i)$) $i = 1, 2$
- We assume test accuracies are the same in the two populations
- Assuming distinct prevalences, we have $2 \times (2^3 - 1) = 14$ degrees of freedom and we have $9 + 1 = 10$ parameters to estimate
- Jones et al (2010) showed that the model is locally identifiable provided $Se > 1 - Sp$ for all tests
- The model is still valid if the prevalences are equal, but it is not identifiable

Posterior Inferences: (i) MAT-ELISA, MB, (ii) TwoPopn Version of (i), (iii) Two-Popn-Two-Test; Informative Prior

	One Popn Med (95% PI)	Two Popns Med (95% PI)	2-Pop-2-Test Med (95% PI)
SeMAT	0.85 (0.78, 0.92)	0.84 (0.77, 0.90)	0.81 (0.66, 0.92)
SeELISA	0.73 (0.63, 0.82)	0.72 (0.63, 0.80)	0.75 (0.49, 0.93)
SpMAT	0.91 (0.86, 0.96)	0.89 (0.86,0.94)	0.90 (0.81,0.97)
SpELISA	0.86 (0.82, 0.91)	0.85 (0.81, 0.88)	0.87 (0.78, 0.94)
DiffSe	0.13 (0.05, 0.22)	0.12 (0.04, 0.20)	0.06 (-0.14, 0.31)
ProbSe	0.9982	0.9984	0.72
DiffSp	0.05 (0.02, 0.08)	0.04 (0.02, 0.07)	0.03 (-0.01, 0.08)
ProbSp	0.9999	0.9998	0.92
SeMB	0.62 (0.47, 0.81)	0.69 (0.53, 0.84)	-
SpMB	0.998 (0.989, 1.0)	0.998 (0.991, 1.0)	-

Posterior Inferences: Two-Popns-Three-Tests; MAT-ELISA, MB, Informative Prior

- We see few differences (except for *SeMB*)
- We monitored the difference in prevalences ($Prev2 - Prev1$) 0.09 (0.04, 0.14)
- $Pr(Prev2 > Prev1 \mid \text{data}) = 0.9995$
- There is nothing wrong with splitting the data in two, even if $Prev1 \doteq Prev2$; less parsimonious model
- However, if $Prev1 \doteq Prev2$, the model is virtually identical to the model for the One-Popn case, and so there are no additional dof gained by splitting

Posterior Inferences: Two-Popns-Two-Tests; MAT-ELISA Informative Prior

- With $Prev1 \neq Prev2$, we only have 6 dof here and 8 parameters to estimate
- In the Table, we see considerably wider probability intervals, but similar point estimates as for 3 test case
- $Pr(Prev2 > Prev1 \mid \text{data}) = 0.91$; ambiguous
- Still nothing wrong with splitting the data in two, even if $Prev1 \doteq Prev2$; less parsimonious model; but only 3 dof and still 8 parameters

The Hui-Walter Model

- With two 2×2 tables of independent multinomial data, and assuming conditional independence and equal accuracy across populations, the model described above is "weakly" identifiable
- Equal accuracy across populations means

$$Pr(T+ \mid D+, \text{Popn i}) = Se \qquad Pr(T- \mid D-, \text{Popn i}) = Sp$$

  for all i
- Here, we have 6 dof and 6 parameters to estimate; the model is identifiable
- In theory, one could place all uniform priors on the 6 parameters and if the sample sizes were to grow, posterior estimates would tend to the "truth"
- There is a caveat here, in that, the "weak" identifiability can wreak havoc on estimates. This is easily fixed by forcing any Se or Sp to be above 0.5.

The Hui-Walter Model

- If either of the assumptions of CI and/or equal accuracy are very far from true, test accuracy estimates can be very biased (shown by many authors)
- Georgiadis et al. (2003) showed that if cond correlations between the two tests were modest (≤ 0.2), or if tests were highly accurate ($Se, Sp > 0.99$), then there was little bias
- They also showed that if correlations were on the order of 0.6, inferences would be very biased
- Adding a third popn to the two-dep-test scenario adds 3 dof at the expense of one additional parameter
- With 3 sampled populations, there are now $3 \times 3 = 9$ dof for estimation, and there are 9 parameters to estimate (4 Se/Sp, 2 correlations, 3 prevalences)
- Hanson and Johnson (2005) and Jones et al. (2010) showed that adding populations does not buy identifiability

Issues for Discussion: The Interplay Between Design and Prior Specification

- A basic principle is that prior specification should be based on data that are independent of current data, but somehow similar to it, and/or based expert scientific knowledge about tests in use and/or populations being studied
- When models lack identifiability, at a minimum it is important to identify the difference in the number of parameters to be estimated and the number of dof available for estimation
- When CI between two tests is reasonable, then the first best design attempt would be to apply both tests to independent samples of individuals from populations with distinct prevalences. If it is reasonable to also assume that test accuracies are constant between populations, then the HW model is appropriate.

Issues for Discussion

- Recall that the HW model was not appropriate for the *MAT* and *ELISA* tests due to lack of CI.
- In the case of 2 dependent tests, and if very little or no scientific knowledge is available, two-test models lack identifiability, and adding populations cannot help. This is a "quit and go home" scenario
- However, adding a third test that is conditionally independent of the two tests under consideration does result in an identifiable model
- We of course believe that incorporating scientifically relevant information is still important
- Why would one specify, through the use of a uniform prior, that the specificity of a test is equally likely to be above or below 0.5, when it is known for a virtual fact that this cannot be true?

Issues for Discussion

- Should papers reference their experts and their credentials for the specific disease under study? How important is it to do a sensitivity analysis? What if there is disagreement in the form of distinctly different scientific information for Se and Sp and/or prevalence? Does it make sense to combine prior information, or to simply report different analyses?

- Design: How many tests? Conditional Independence vs Dependence? How many populations? To split or not to split? How to split? When does adding a distinct population help and when not? How should we approach the trade off between getting more dof by splitting and in doing so, possibly losing the assumption of equal accuracy across populations? How to know if a model is identifiable or not

Other Applications of Bayesian Methods in Diagnostic Testing

- A Bayesian Approach to Modeling Bivariate Longitudinal Diagnostic Outcome Data
- Bayesian Nonparametric Receiver Operating Characteristic Curve Estimation

Assessing the Temporal Quality of Diagnostic Biomarkers in the Presence of Variable Biologic Response after Infection: A Case Study for Diagnosing Johne's Disease

Statistics and its Interface (2014)

Michelle Norris

California State University, Sacramento

Wesley Johnson

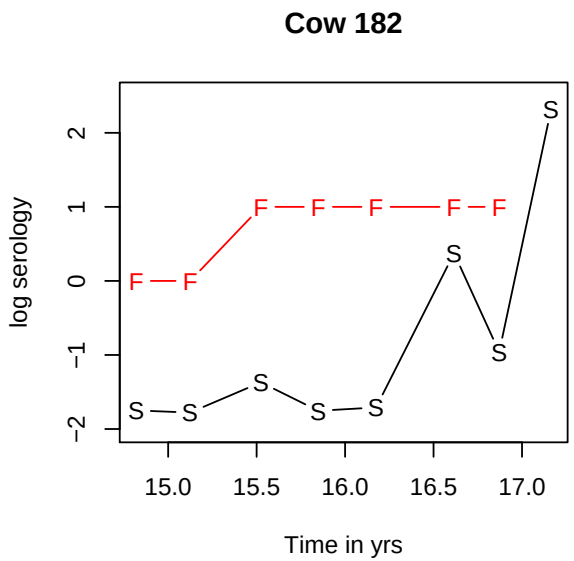
UC Irvine

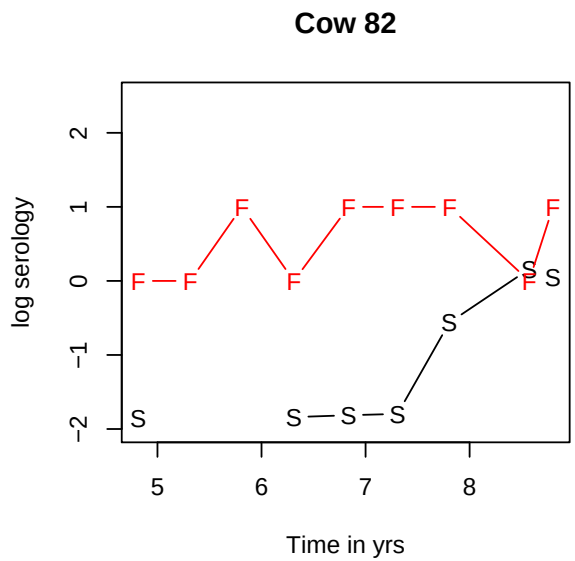
and

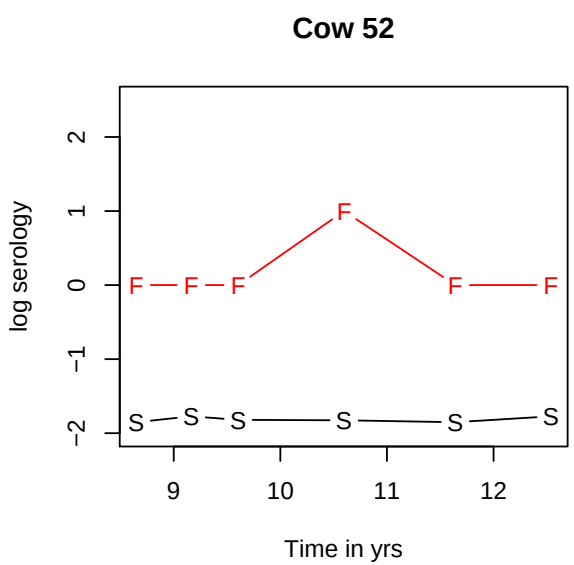
Ian Gardner

University of Prince Edward Island

The Data


Cow 182
log serology
Time in yrs
15.0
15.5
16.0
16.5
17.0


Cow 82
log serology
Time in yrs
5
6
7
8


Cow 52
log serology
Time in yrs
9
10
11
12

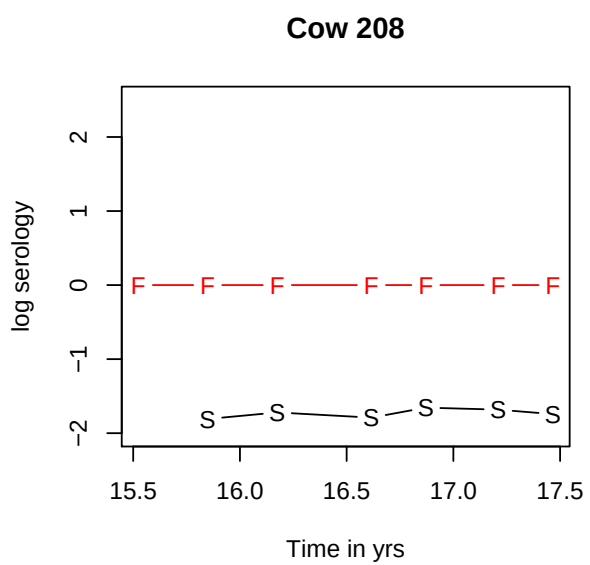
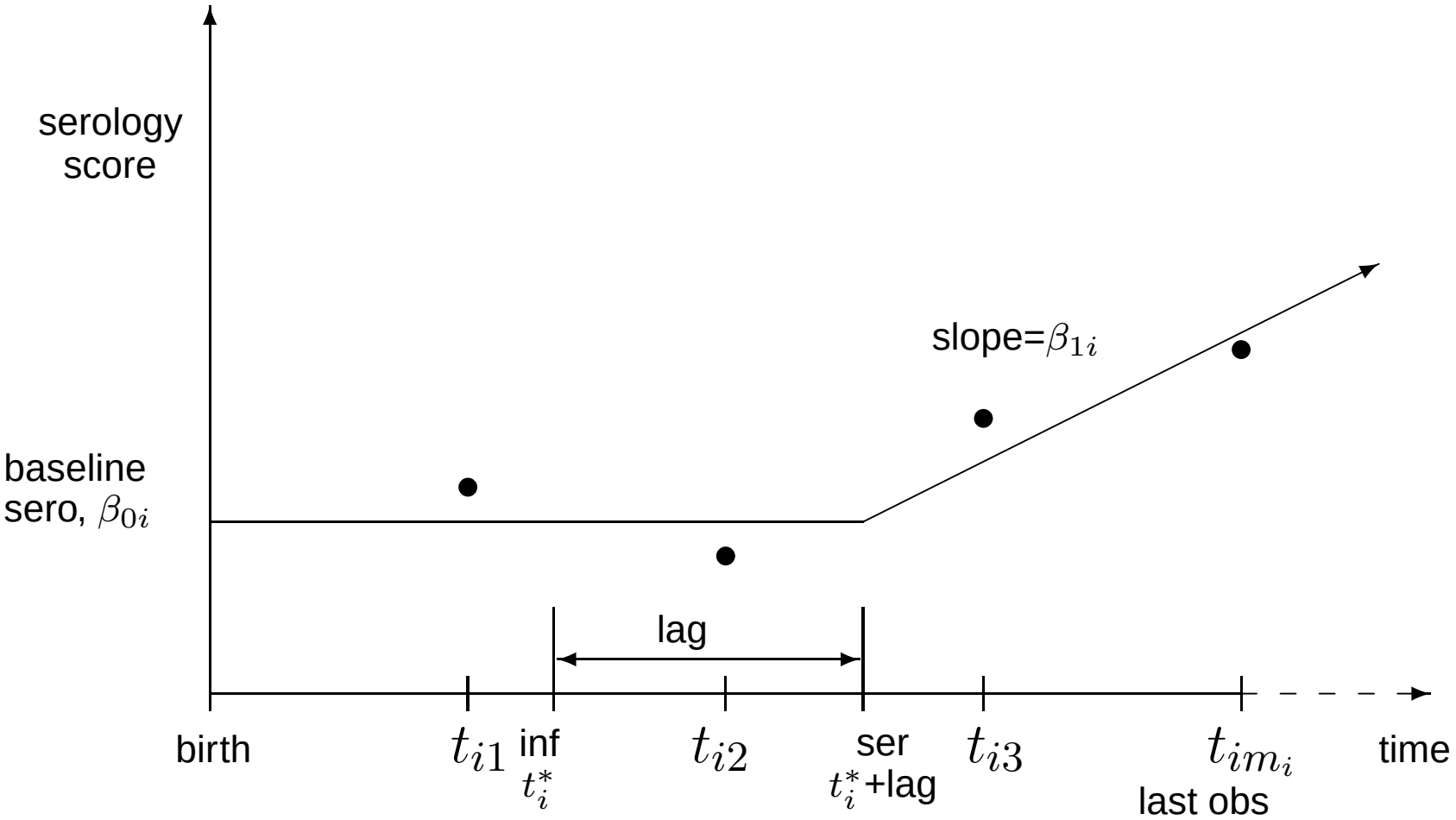

Cow 208
log serology
Time in yrs
15.5
16.0
16.5
17.0
17.5

Figure: Serology Trajectory



Semiparametric Model

- The joint model is the same except we now specify a DPM model for log-slopes:

$$\log(\beta_{1i}) = \gamma_i \mid \mu_i, \tau_i \overset{\perp}{\sim} N(\mu_i, \tau_i) \quad \text{for } i : k_i = 3$$

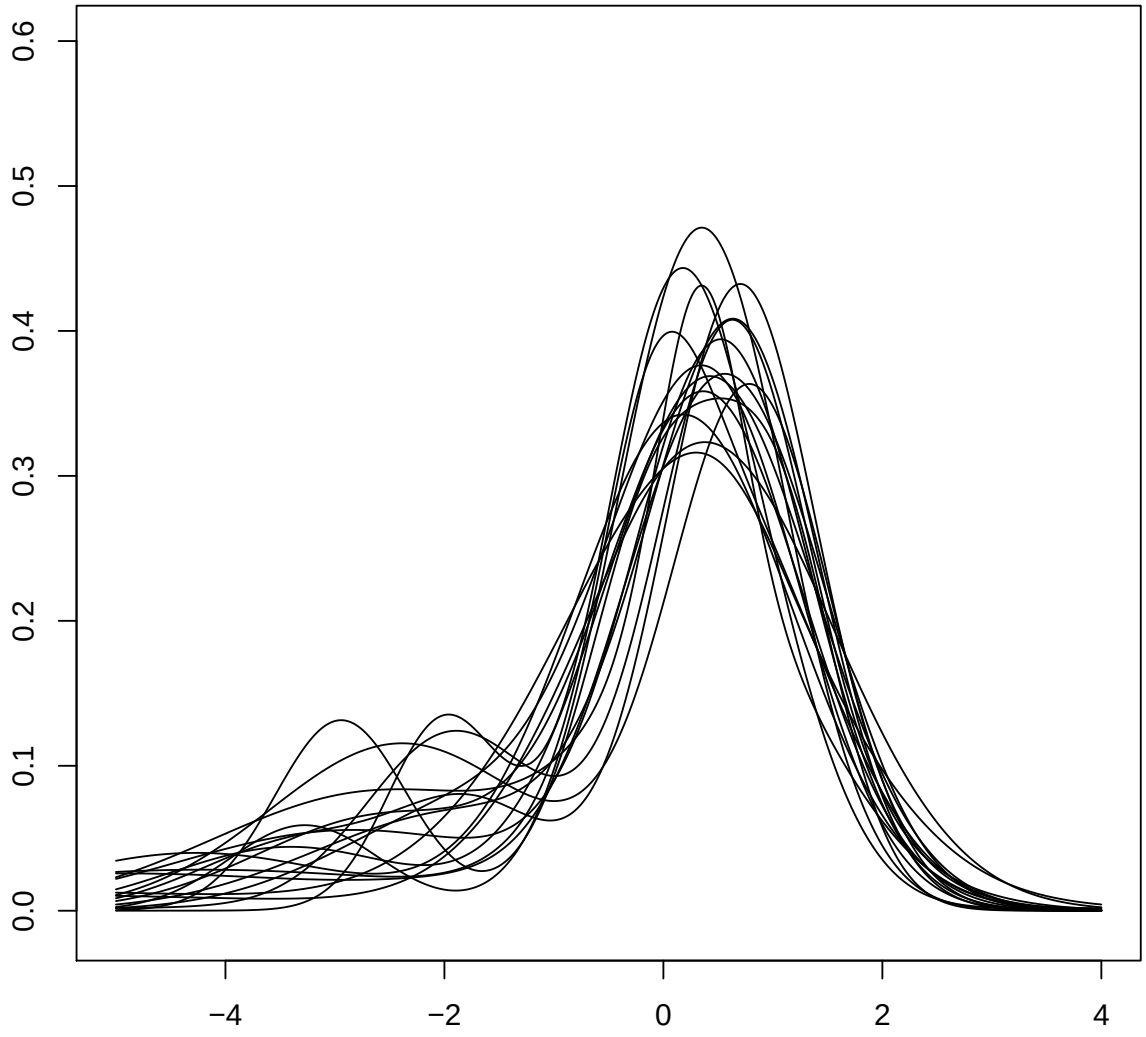
$$(\mu_i, \tau_i) \mid G \overset{\perp}{\sim} G$$

$$G \mid \alpha, G_0 \sim DP(\alpha, G_0)$$

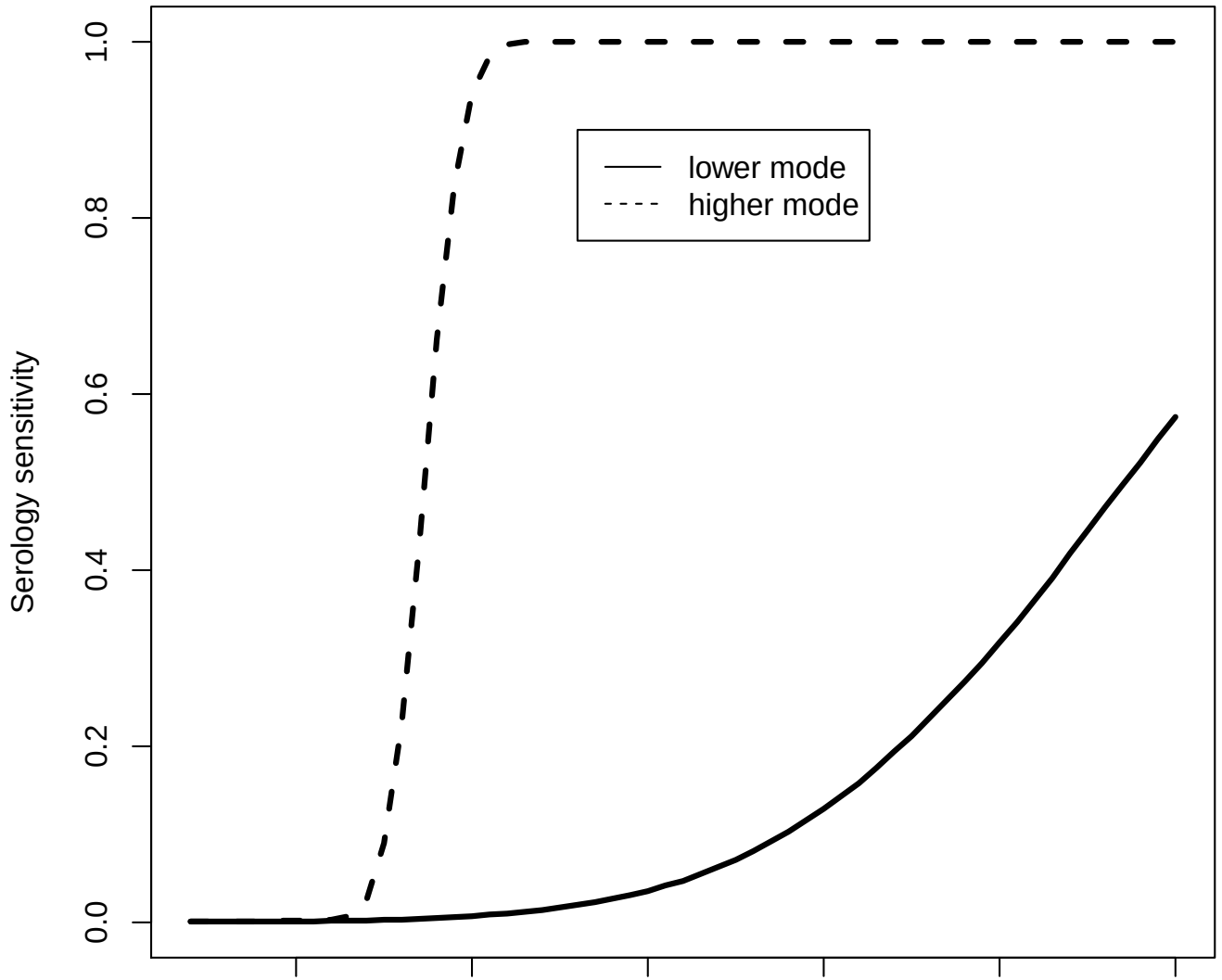
- We choose G_0 to be normal/inverse gamma conjugate:

$$\tau_i \sim \Gamma(\frac{s}{2}, \frac{S}{2}), \qquad \mu_i \mid \tau_i \sim N(m, \frac{d}{\tau_i})$$

JD: Posterior Distn of Slopes



Sens as a Fn of Time by “Clusters"



Flexible Regression Models for ROC and Risk Analysis, with or without a Gold Standard

Adam Branscum Oregon State University

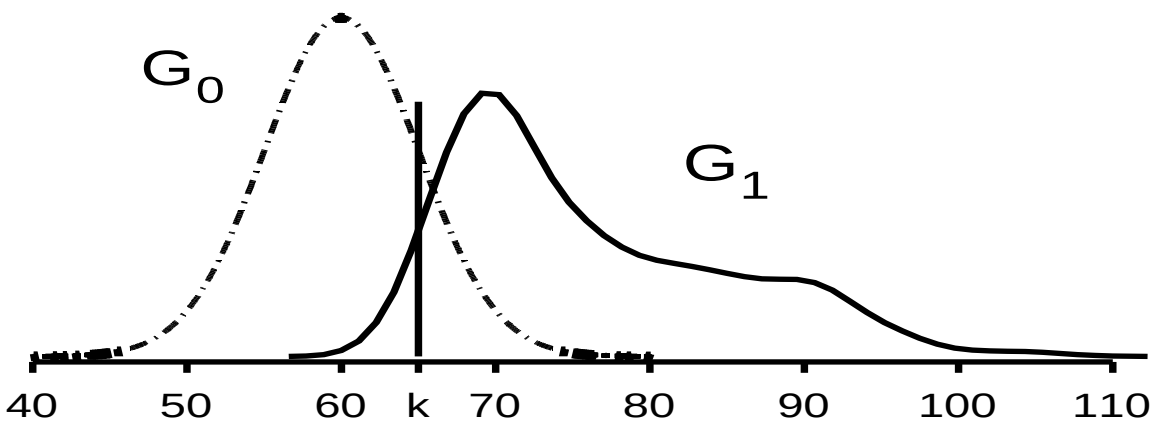
Wesley Johnson UC Irvine

Tim Hanson University of South Carolina

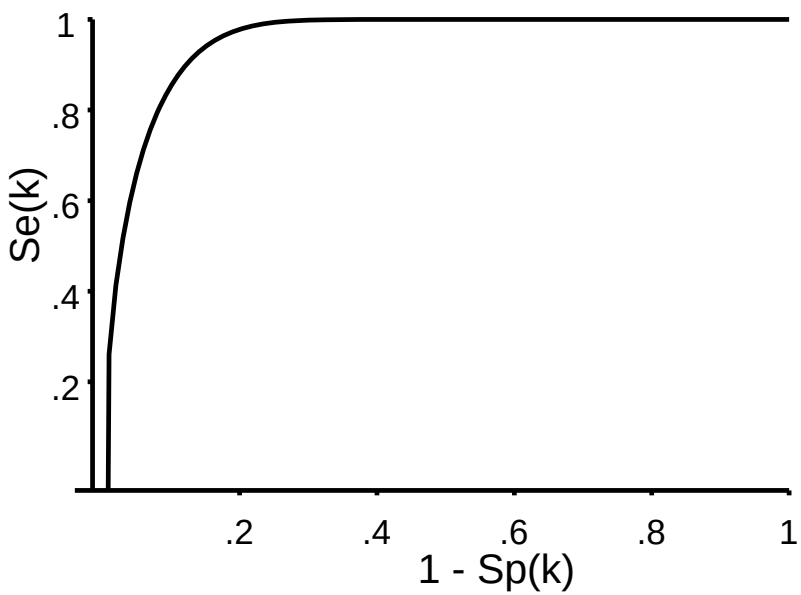
SSC

Standard Approach

- Dichotomize the serology scores using a cutoff value k

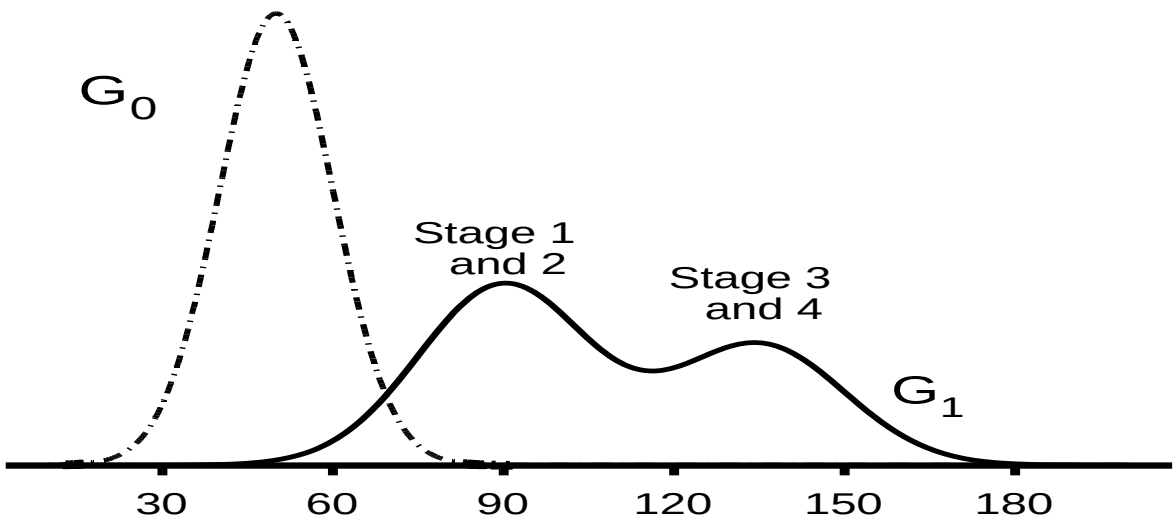


ROC Curve



Motivation for NP Modeling

- Serology scores for diseased individuals are generally coming from a mixture distribution based on staging



Lung Cancer Data

- Investigated the potential of sEGFR (soluble epidermal growth factor receptor) to be a biomarker for lung cancer in men
- Data obtained from a case-control study, with 139 cases and 88 controls
- Analyzed the data as if disease status were unknown, using age as both disease and test covariates
- Made comparisons to a gold-standard data analysis
- Ages ranged from 35 to 88 for cases and from 24 to 79 for controls

ROC Ests; GS (dash), NGS (sol)

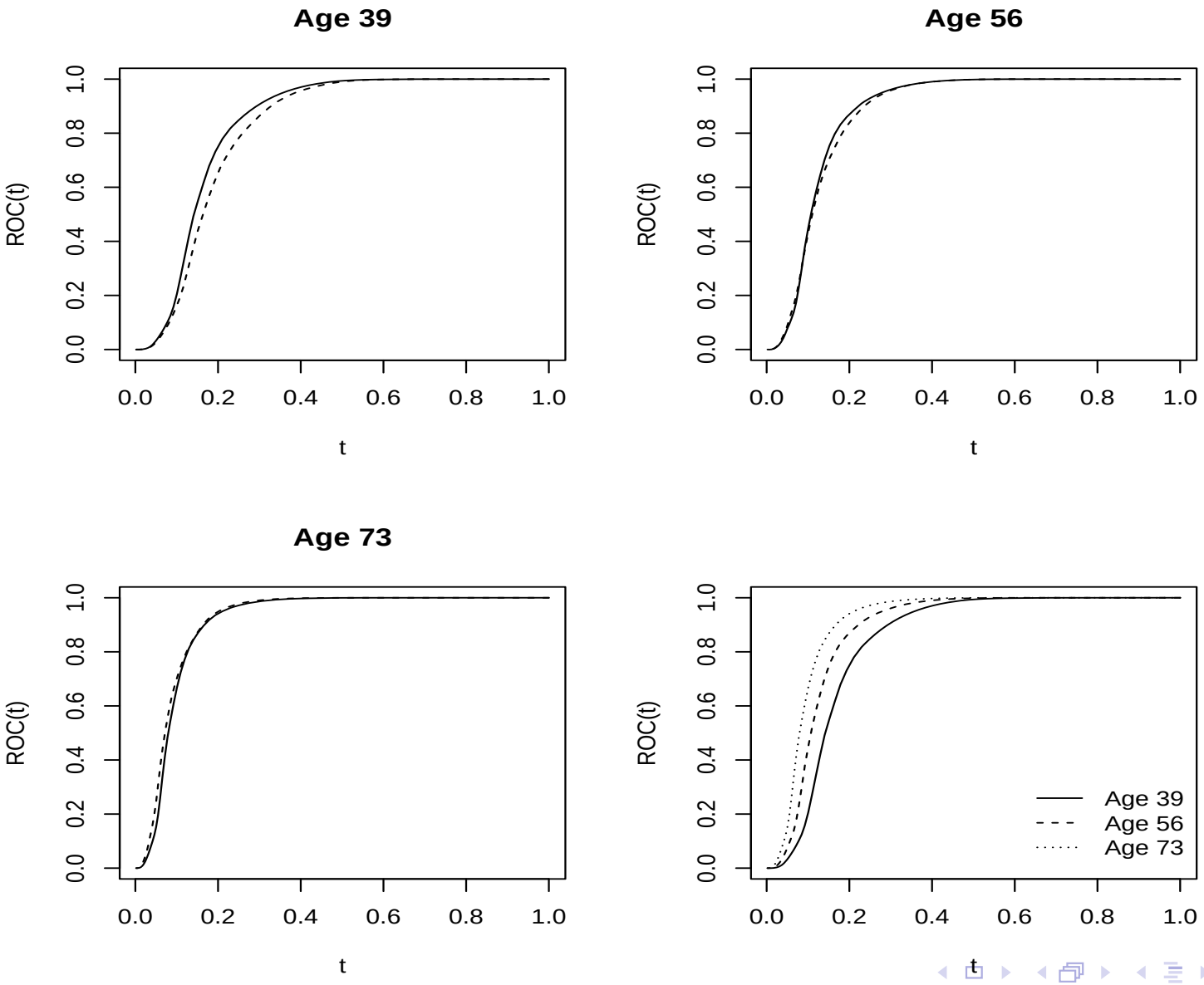


Table : Post medians and 95% PIs

Parameter	Gold standard	No gold standard
AUC_{39}	0.81 (0.75, 0.86)	0.83 (0.75, 0.90)
AUC_{56}	0.87 (0.81, 0.91)	0.88 (0.79, 0.93)
AUC_{73}	0.91 (0.85, 0.95)	0.91 (0.83, 0.95)
$AUC_{73} - AUC_{39}$	0.09 (0.05, 0.14)	0.07 (0.03, 0.12)
$AUC_{73} - AUC_{56}$	0.04 (0.02, 0.06)	0.03 (0.01, 0.05)
$AUC_{56} - AUC_{39}$	0.05 (0.03, 0.08)	0.04 (0.02, 0.07)