Evidence synthesis

Aims of Session 7

- Understand that all Bayesian models are evidence syntheses
- Practise building evidence synthesis models
- Practise criticising evidence synthesis models
- Understand that different evidence sources can be inconsistent/conflicting
- Understand that detecting conflict is only one step of the model development and criticism cycle: resolving conflict is important
- Practise techniques for conflict resolution, such as bias adjustment and robustifying inference by introducing more flexibility in a model

Overview of Session 7

[30 mins lecture + 1hr practical]

- What is evidence synthesis?
  - Broad definition
  - Illustrative example
  - Key features
  - Formal definition
- Practical 1: HIV example

[30 mins lecture + 1hr practical]

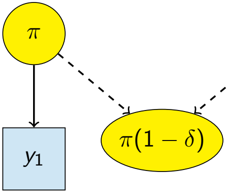
- Model criticism for evidence syntheses
  - Conflict resolution via bias modelling
  - Robustifying inference using over-dispersion
  - Cross-validatory mixed-predictive checks
  - Systematic bias adjustment
- Practical 2: HIV example
- Practical 3: Sepsis example

What is evidence synthesis?

- Any Bayesian analysis is an *evidence synthesis*: combining prior information with new data
- But more generally, we think of evidence synthesis as the generalisation of hierarchical models to *multi-parameter* models where inference is based on *multiple* data sources
- For example, the *meta-analysis* you have already seen in Session 5 on hierarchical models is an evidence synthesis, combining prior information with data from multiple studies, all measuring the same quantity, to obtain increased precision in the estimate of that quantity
- And even more broadly, we can think of *generalised evidence synthesis* as generalising meta-analysis to the combination of prior information with data of *multiple types*, all measuring *different* quantities

Spiegelhalter et al 2004, Ades & Sutton 2006

Simple example: HIV prevalence



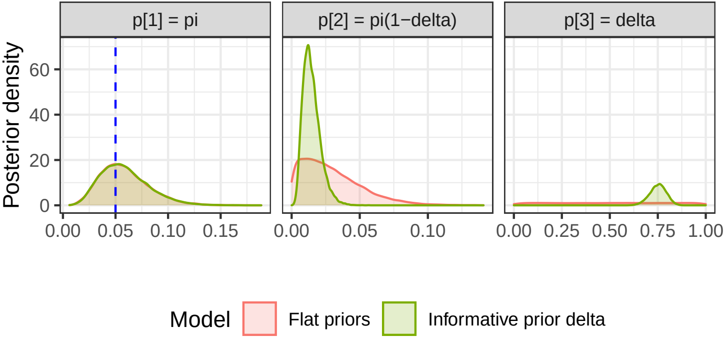
$$\begin{aligned} p(y_1, \pi, \delta) &= p(\pi)p(\delta)p(y_1 \mid \pi, \delta) \\ &= p(\pi)p(\delta)p(y_1 \mid \pi) \end{aligned}$$

- Suppose we are interested in HIV *prevalence* π and the proportion of infections that are *diagnosed* δ.
- The *prevalence of undiagnosed infection* can then be defined as $\pi(1-\delta)$.
- Initially, suppose we only have data from a study measuring π, where y_1 out of n_1 HIV tests return a positive result in a population.
- Then:
  - δ is only *weakly identifiable*, i.e. requires informative prior for identification (since have only 1 data point to inform two parameters)
  - y_1 is independent of δ

Model code

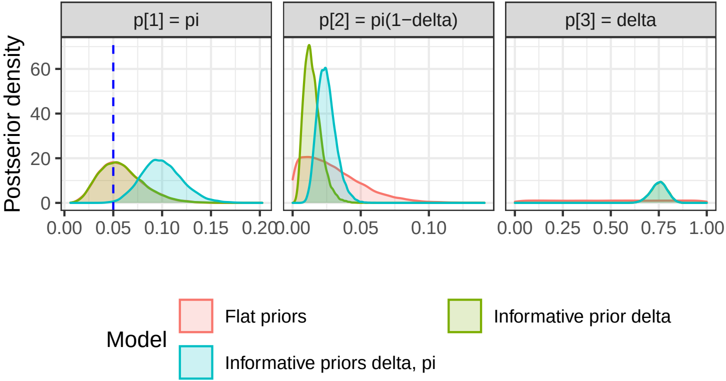
```
## Flat priors
pi    ~ dbeta(1,1)     # or informative implying 15% (9-23%)
delta ~ dbeta(1,1)     # or informative implying 75% (66-83%)

## Likelihood: prevalence data
for(i in 1:1)
{
  y[i] ~ dbin(p[i], n[i])     # (y1,n1) = (5,100)
}

# Proportions in terms of basic and functional parameters
p[1] <- pi
p[2] <- pi * (1 - delta)
p[3] <- delta
```

Simple example: HIV prevalence



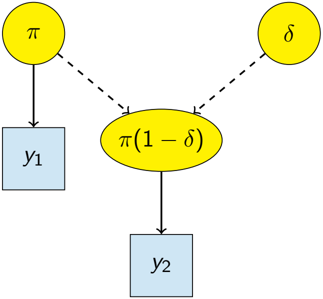
- $\pi(1-\delta)$ *weakly identified* with flat priors, since $\delta \in [0,1]$
- δ identified when use informative prior, additional information *increases precision* of $\pi(1-\delta)$

Simple example: HIV prevalence



- informative prior for π inconsistent with likelihood, so posterior is a *compromise*
- conflict *reduces precision* of $\pi(1-\delta)$ again

Simple example: HIV prevalence



Now suppose we add in a second study measuring the number of individuals living with *previously undiagnosed* infection y_2 out of a total population of size n_2. Then:

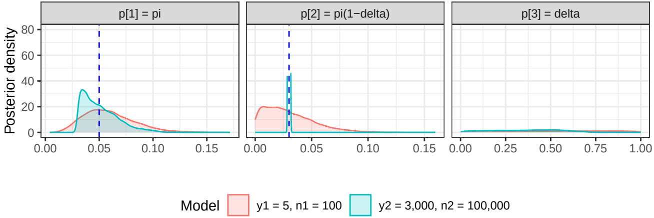
- y_1, y_2 *conditionally* independent, given π, δ
- π, δ both now identifiable *without* informative priors (2 data points, 2 parameters)
- but there is potential for *conflicting evidence* if priors are informative

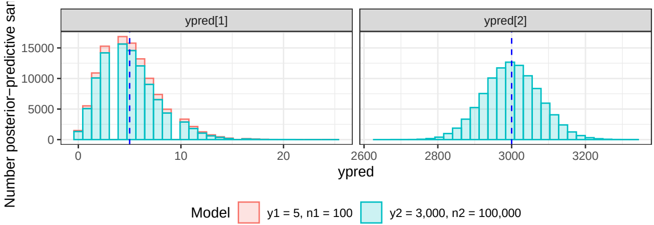
$$
\begin{aligned}
&p(y_1, y_2, \pi, \delta) \\
&= p(\pi)p(\delta)p(y_1, y_2 \mid \pi, \delta) \\
&= p(\pi)p(\delta)p(y_1 \mid \pi)p(y_2 \mid \pi, \delta)
\end{aligned}
$$

Model code

```
...

# Likelihood: prevalence data
# Likelihood: prevalence of undiagnosed infection
for(i in 1:2)
{
  y[i] ˜ dbin(p[i], n[i])     # (y1,n1) = (5,100),
                              # (y2,n2) = (3000,100000)
}

# Proportions in terms of basic and functional parameters
p[1] <- pi
p[2] <- pi * (1 - delta)
p[3] <- delta
```

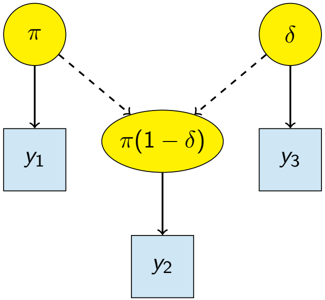
Simple example: HIV prevalence





- both basic and functional parameters now *identified*, even with flat priors
- although δ still relatively uncertain (95% credible interval 0.05 – 0.70), despite large sample size informing $\pi(1-\delta)$, since information *indirect* and small sample size informing π

Simple example: HIV prevalence



Finally we add in a third study measuring the number of individuals living with *diagnosed* infection y_3 out of a population living with HIV of size n_3. Then:

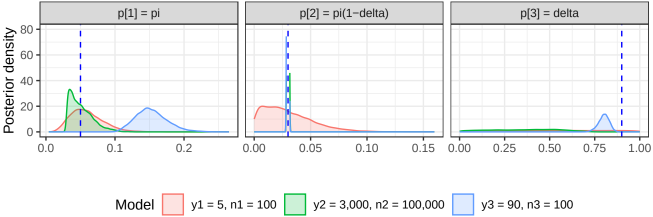
- in theory, more data ⇒ more *precise* estimates
- even with uninformative priors, potential for *conflict*: 3 data items informing 2 parameters

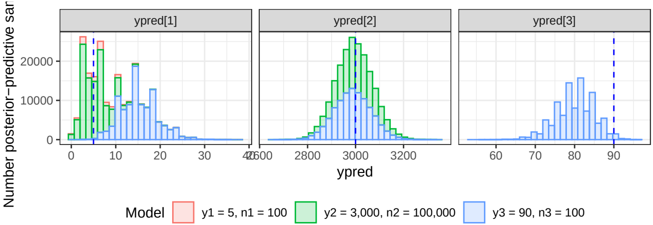
$$
\begin{aligned}
&p(y_1, y_2, y_3, \pi, \delta) \\
&= p(\pi)p(\delta)p(y_1, y_2, y_3 \mid \pi, \delta) \\
&= p(\pi)p(\delta)p(y_1 \mid \pi)p(y_2 \mid \pi, \delta)p(y_3 \mid \delta)
\end{aligned}
$$

Model code

```
...

# Likelihood: prevalence data
# Likelihood: prevalence of undiagnosed infection
# Likelihood: proportion diagnosed
for(i in 1:3)
{
  y[i] ~ dbin(p[i], n[i])   # (y1,n1) = (5,100),
                            # (y2,n2) = (3000,100000),
                            # (y3,n3) = (90,100)
}

# Proportions in terms of basic and functional parameters
p[1] <- pi
p[2] <- pi * (1 - delta)
p[3] <- delta
```

Simple example: HIV prevalence





- δ is now better identified (*more peaked* posterior)
- but *conflict* between the three data points leads to a larger, more uncertain estimate of π
- (y_1, n_1) is the smallest sample size, so estimate of π is closer to the value suggested by the combination of y_2 and y_3 than y_1

Features of evidence synthesis

Evidence synthesis leads to complex probabilistic models

- Combination of all available relevant data sources *ideally* should lead to more *precise* estimates
- Multiple sources informing a single parameter ⇒ potential for *conflicting evidence*
- Sparsity of data ⇒ parameters *unidentifiable* without further model constraints, e.g. *informative priors, exchangeability*

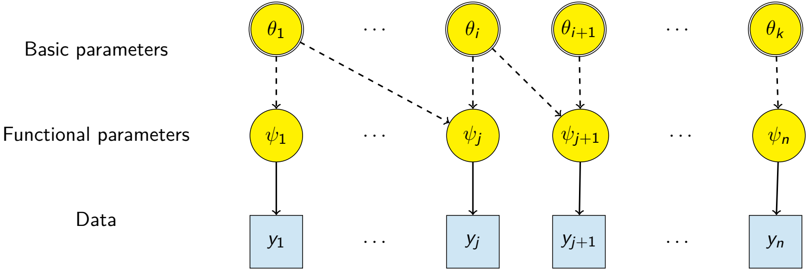
Statistical formulation

- Interest: estimation of $\boldsymbol{\theta} = (\theta_1, \theta_2 \ldots, \theta_k)$ on the basis of a collection of *independent* data sources $\boldsymbol{y} = (y_1, y_2 \ldots, y_n)$
- Each y_i provides information on
  - a *single* component of $\boldsymbol{\theta}$ (*"direct"* data), or
  - a *function* of one or more components, *i.e.* on a quantity $\psi_i = f(\boldsymbol{\theta})$ (*"indirect"* data)

Thus inference is conducted on the basis of both **direct** and **indirect** information.

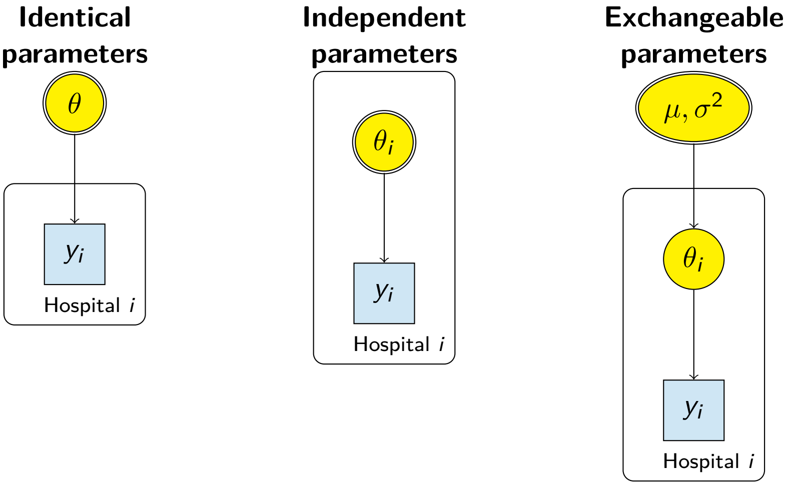
- Likelihood: $L(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} L_i(y_i \mid \boldsymbol{\theta})$
- Posterior: $p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{\theta}) \times L(\boldsymbol{y} \mid \boldsymbol{\theta})$

Graphical representation: DAG



- Basic parameters are *founder nodes* at the top of the DAG.
- In more generality, there could be some *hierarchical* structure above, so that the basic parameters are the hyper-parameters of any hierarchical prior distribution.
- Functional parameters are *deterministic* functions of other parameters.
- Note n does not have to equal k; and indeed, some functions of interest may have no *direct* data informing them.

Recall: DAGs of hierarchical models

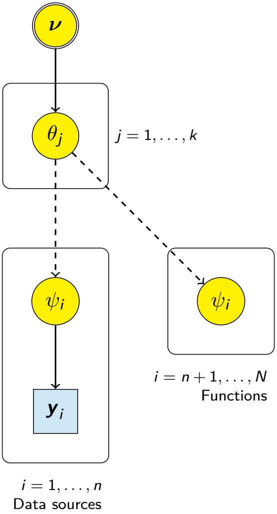


Identical parameters

Independent parameters

Exchangeable parameters

Generalised evidence synthesis: DAG using plates



- We can see how we generalise from meta-analysis to synthesis of multiple data sets informing various quantities, all of which can be expressed as functions of basic parameters.
- The parameters of interest to estimate might be a *mixture* of basic parameters, intermediate parameters and functional parameters.

Realistic examples

- Mixed treatment comparisons/network meta-analysis: Ades, Med. Decision Making (2003); Dias et al, Med. Decision Making (2012)
- Estimating influenza severity: Presanis et al, PloS Med (2009); AoAS (2014); Shubin et al, Epidem & Inf (2013); Wong et al, Epidem (2013); McDonald et al, IORV (2014)
- Estimating HPV disease progression: Jackson et al, Med. Decision Making (2013); and the impact of vaccination on HPV-associated cancers Bogaards et al, BMJ (2015)
- Estimating trends in HIV prevalence: Presanis et al, Lancet Public Health (2021)

Practical 1

We will use the HIV example to

- practise building a simple evidence synthesis;
- understand that synthesising evidence can result in lack-of-fit to the data, if some of the evidence are conflicting;
- practise using posterior-predictive p-values to detect conflict and the DIC to compare models.

Model criticism for evidence syntheses I

Criticising an evidence synthesis is no different from criticising a simpler Bayesian model, and boils down to *comparing two sets of evidence*:

- prior-predictive checks: comparing the prior and the data
- (cross-validatory) posterior-predictive checks: comparing the posterior and the data
- (cross-validatory) mixed-predictive checks: comparing the posterior and the data in hierarchical models, where we replicate *both* parameters and data (Marshall & Spiegelhalter 2007)
- posterior-posterior comparisons for two distinct sub-models with a common parameter to detect conflicting evidence (N.B. not covered in course)

What is common to all model criticism is that *checking* (different aspects of) the model is not the end of the story. Having *detected* conflicting evidence, how do we then *resolve* the conflict? (⇒ model development-criticism cycle.)

Resolving conflict

- We might *exclude* suspect / biased data (a *subjective* judgement);
- We might *robustify* our model, e.g. using heavier-tailed or more flexible distributions (e.g. accounting for over-dispersion) and/or random effects to allow for greater variation: *accommodating* conflicting evidence;
- We might introduce extra parameters to model suspected biases ⇒ bias adjustment or bias modelling:
  - Important to use any possible *external* evidence to derive informative priors for any bias parameters (even if just on *direction* of bias).
  - So that we don't just "mop up" any lack of fit/conflict without *understanding* what the biases may be.

HIV example: bias modelling to resolve conflict I

Recall that the use of three datasets resulted in lack of fit and conflicting evidence:

Parameter	Observation	Post'r median	95% CrI		p-value
π	$5\ /\ 100 = 0.05$	0.151	0.113	0.203	0.996
$\pi(1-\delta)$	$3{,}000\ /\ 100{,}000 = 0.03$	0.0299	0.0288	0.0309	0.436
δ	$90\ /\ 100 = 0.90$	0.802	0.736	0.853	0.0102

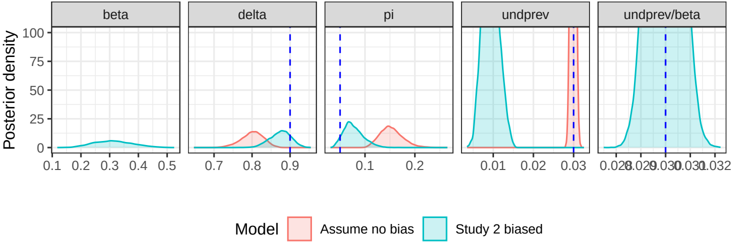
Suppose we had prior expert opinion (possibly based on both subject knowledge and previous studies) that Study 2, measuring $\pi(1-\delta)$, was carried out in a population that was *higher risk* than the general population, such that the true undiagnosed prevalence was likely to be between 20 and 50% of the value measured by Study 2.

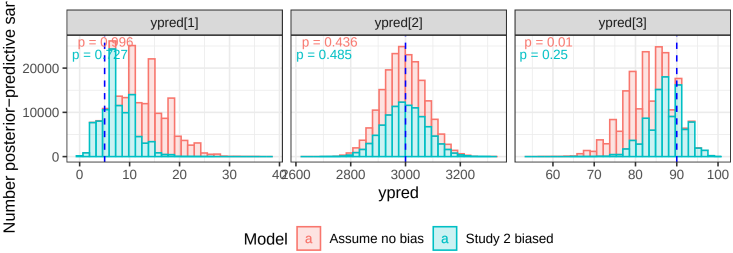
HIV example: bias modelling to resolve conflict II

We could encode this prior knowledge by incorporating a bias parameter β:

```
# Proportions in terms of basic and functional parameters
p[1] <- pi
p[2] <- pi * (1 - delta) / beta
p[3] <- delta

# bias parameter beta, prior suggesting true undiagnosed
# prevalence is lower than that suggested by study 2
beta ~ dbeta(a.beta, b.beta)
```

HIV example: bias modelling to resolve conflict III



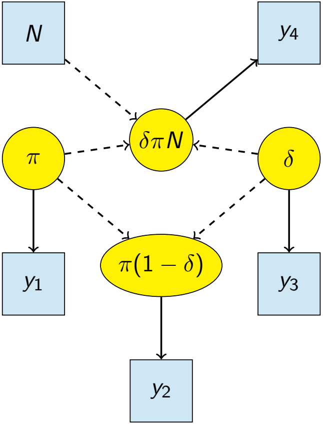


HIV example: bias modelling to resolve conflict IV

- Note the better fit to the observations from Study 1 (informing π) and Study 3 (informing δ) once we account for potential bias in the large-sample Study 3 (informing undiagnosed prevalence $\pi(1-\delta)$).
- *Caution: don't just add in bias parameters to "mop up" the conflict, without having external evidence that a study is biased!*

HIV example: over-dispersion

Suppose now we observe, in a fourth study, the number $y_4 = 400$ of people living with diagnosed HIV, $\delta\pi N$, in a population of size $N = 10,000$:

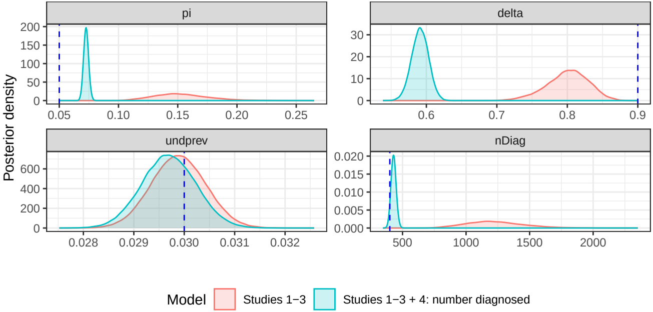


We could consider initially a Poisson sampling distribution

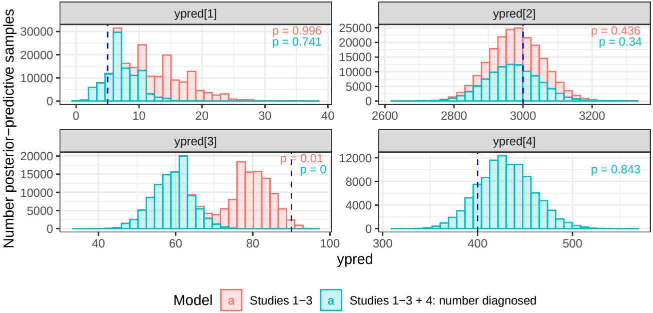
$$y_4 \sim \text{Poisson}(\delta\pi N)$$

and see how the addition of new independent information changes the inference. In this case, the observation y_4 is more or less consistent with $y_1/n_1 = 0.05$ and $y_3/n_3 = 0.9$, but not with the combination of y_1 and $y_2/n_2 = 0.03$; nor with the combination of y_3 and y_2.

HIV example: Poisson model I



HIV example: Poisson model II



Addition of the data brings the posterior estimate of π closer to observation y_1, but there is still substantial conflict (see posterior-predictive p-values).

Over-dispersion: the negative binomial distribution I

Accounting for over-dispersion (variance greater than mean) in observations is another way to *robustify* inference (recall the use of heavier tailed distributions in session 4).

Negative binomial

In `JAGS/rjags`: `y ~ dnegbin(psi,r)`

$$p(Y = y) = \binom{y + r - 1}{y} \psi^r (1 - \psi)^y$$

r is interpreted as a number of failures need to observe y successes, ψ is probability of failure

In the parameterisation we will use, we express r as a function of the mean and ψ:

$$r = \frac{\psi}{1 - \psi} \mathbb{E}(Y)$$

Over-dispersion: the negative binomial distribution II

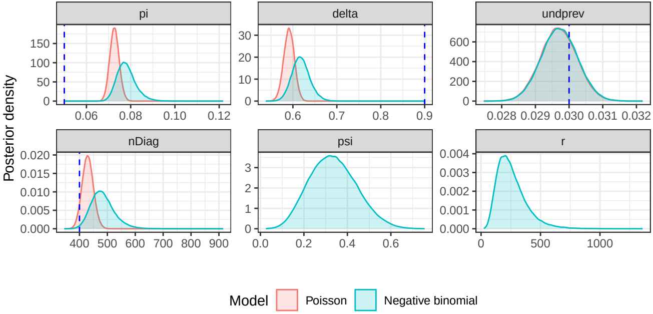
and then place a prior on ψ, e.g. Uniform on some reasonable subset of $[0, 1]$.

This means $Var(Y) = \mathbb{E}(Y)/\psi$, i.e. $1/\psi$ is a measure of the over-dispersion:

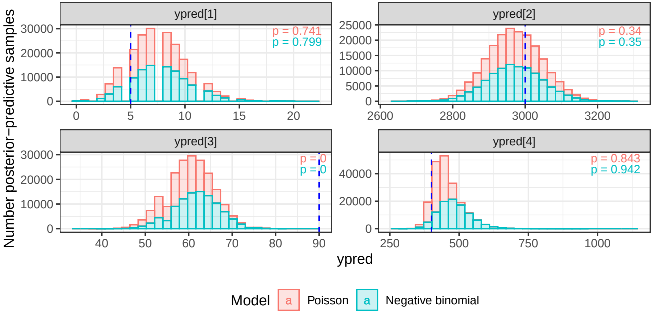
- $\psi = 1 \Rightarrow$ *Poisson*
- $\psi = 0.5 \Rightarrow Var(Y) = 2\mathbb{E}(Y)$
- $\psi = 0.1 \Rightarrow Var(Y) = 10\mathbb{E}(Y)$
- $\psi \to 0$ implies over-dispersion tending to infinity

HIV example: over-dispersion I

For the HIV example, we assume a Beta prior for ψ expressing that it lies between 0.2 and 0.6, representing variances between 1.67 and 5 times the mean. For this particular example, allowing for over-dispersion doesn't help alleviate the conflict.

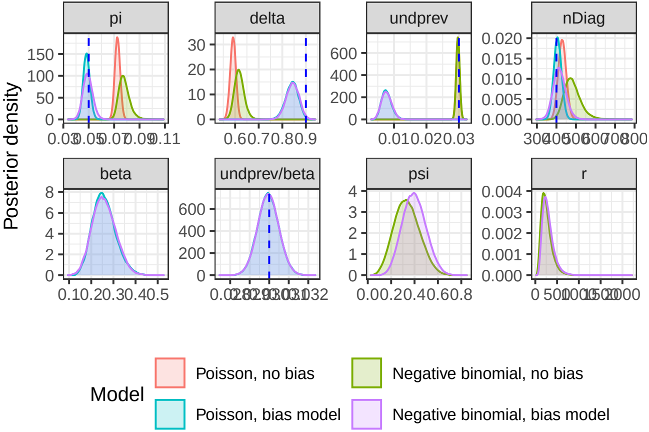


HIV example: over-dispersion II



HIV example: over-dispersion vs bias model

Looking at a version of the HIV model with both a bias parameter and over-dispersion, we see that the bias model makes the most difference to resolving the conflict, the over-dispersion just adds a bit more uncertainty:



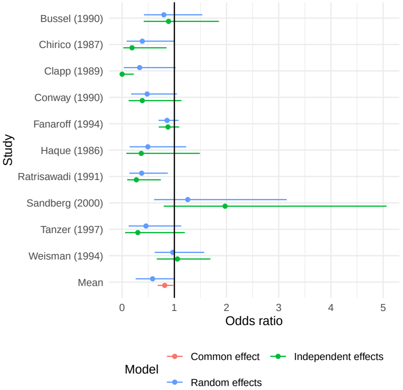
Sepsis example (Ohlsson & Lacy, 2013)

We will use the sepsis example to illustrate cross-validatory mixed-predictive checks and systematic bias adjustment.

- Outcome: Infection (or not) in preterm/low birth weight infants
- Arms: Intravenous immunoglobulin (IVIG) vs placebo
- Question: Does administration of IVIG prevent infection in hospital, compared to placebo? Event = 'sepsis'

	Treatment		Control	
Study	Events	Total	Events	Total
Bussel (1990a)	20	61	23	65
Chirico (1987)	2	43	8	43
Clapp (1989)	0	56	5	59
Conway (1990)	8	34	14	32
Fanaroff (1994)	186	1204	209	1212
Haque (1986)	4	100	5	50
Ratrisawadi (1991)	10	68	13	34
Sandberg (2000)	19	40	13	41
Tanzer (1997)	3	40	8	40
Weisman (1994a)	40	372	39	381

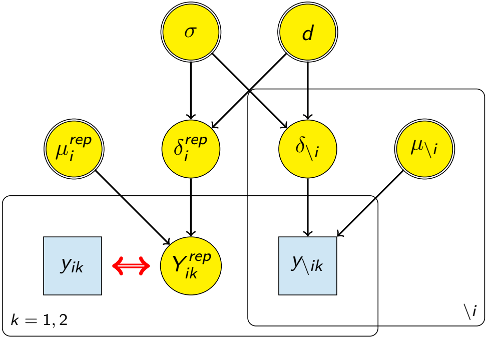
Forest plot:



Sepsis example: cross-validatory mixed-predictive checks

Is unit i *consistent* with other units, i.e. comes from the random effects distribution?

$\Rightarrow$ choice of *appropriate* test statistic? If we compare the data in each arm k with its predictive distribution, we are only assessing the consistency of unit i, arm k with the k'th arm of each other unit, rather than the consistency of unit i with all other units.



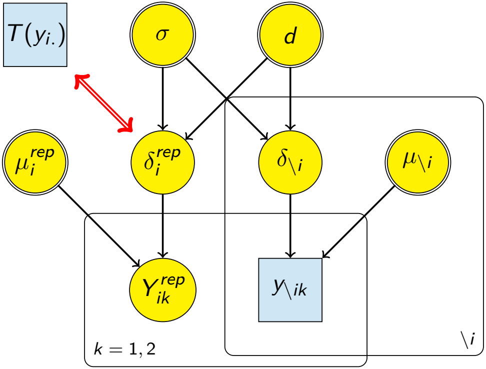
Sepsis example: cross-validatory mixed-predictive checks

Is unit i *consistent* with other units, i.e. comes from the random effects distribution?

$\Rightarrow$ choice of *appropriate* test statistic?

$$T(y_{i.}) = \text{logit}(y_{i2}/n_{i2}) - \text{logit}(y_{i1}/n_{i1})$$

where $\text{logit}(p) = \log(p/(1-p))$ is the *log-odds* of p.



Sepsis example: code outline

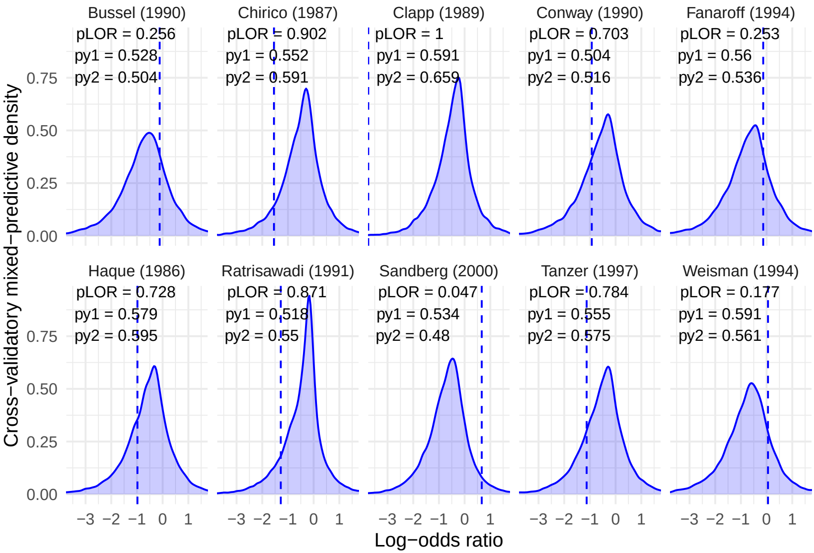
Cross-validation carried out by *repeating* the dataset once for each study being left out:

```
model
{
  # Cross-validation: repeat data set leaving one out each time
  for(j in 1:Ns)
  {
    # For each study, Ns = total number of studies
    for(i in 1:Ns)
    {
      # for each of the two arms
      for(k in 1:Na)
      {
        # Binomial likelihood
        ycv[j,i,k] ~ dbin(p[j,i,k], ncv[j,i,k])
      }
```

Code the study being left out using the `equals` function:

```
      # on logit scale, proportion is probability of success in terms of
      # study baselines mu and study-specific treatment contrasts delta
      # (log odds ratios, relative to study baseline), if not left-out
      # j=i refers to the i'th study being left out, so (1 - equals(j,i))
      # is equal to 0 if j=i, 1 otherwise
      logit(p[j,i,1]) <- ((1 - equals(j,i)) * mu[j,i])
      logit(p[j,i,2]) <- ((1 - equals(j,i)) * (mu[j,i] + delta[j,i]))
```

Sepsis example: choice of test statistic



Systematic bias adjustment

e.g. in meta-analysis of clinical trials, there are a number of recognised issues in poorer quality trials that might bias results:

- unclear/inadequate *sequence generation*
- unclear/inadequate *allocation concealment*
- unclear/inadequate *blinding*
- *incomplete/missing* data

⇒ Cochrane Collaboration's *"Risk of bias"* tool (Turner et al JRSS(A) 2009, Higgins et al BMJ 2011, Savovic et al Ann Int Med 2012) allows systematic assessment/judgement of bias in clinical trials.

Prior judgements on risk of bias - Cochrane

- Systematic *elicitation* of expert opinion on potential biases, each of which is modelled. (Higgins et al BMJ 2011, Ohlssen & Lacy 2013).
- Notice that the outlying Sandberg study is both the most recent study and the study judged at *low risk* of bias in all four criteria (along with Fanaroff).

Study	SeqGen	AllCon	BlindUnspec	IncompData
Bussel (1990a)	Unclear	Low risk	Low risk	High risk
Chirico (1987)	Unclear	Low risk	High risk	Low risk
Clapp (1989)	Unclear	Low risk	Unclear	Low risk
Conway (1990)	Unclear	Low risk	High risk	Low risk
Fanaroff (1994)	Low risk	Low risk	Low risk	Low risk
Haque (1986)	Unclear	Low risk	High risk	Low risk
Ratrisawadi (1991)	Unclear	Unclear	High risk	Unclear
Sandberg (2000)	Low risk	Low risk	Low risk	Low risk
Tanzer (1997)	High risk	High risk	High risk	Low risk
Weisman (1994a)	Unclear	Low risk	Low risk	Low risk

Bias-adjustment model I

(Turner et al JRSS(A) 2009, Savovic et al Ann Int Med 2012)

Expert judgements on the risk of (*internal*) biases are summarised by

$$\beta_{ij} \sim f(\nu_{ij}, \tau_{ij}^2)$$

for some distribution f (e.g. normal) where ν_{ij} is the mean and τ_{ij}^2 the variance, for each bias j in study i.

After the elicitation process, we assume internal biases are *independent*, so that the total internal bias per study i is:

$$\beta_i \sim f\left(\nu_i = \sum_j \nu_{ij}, \tau_i^2 = \sum_j \tau_{ij}^2\right)$$

Bias-adjustment model II

For this sepsis example, we assume an *additive bias* model for the treatment effect in treatment arm $k = 2$ versus the control arm 1:

$$
\begin{aligned}
logit(p_{i1}) &= \mu_i \\
logit(p_{i2}) &= \mu_i + \delta_i^{bias} \\
\delta_i^{bias} &= \delta_i + \beta_i \\
\delta_i &\sim N(d, \sigma^2) \\
\beta_i &\sim N(\nu_i, \tau_i^2)
\end{aligned}
$$

- p_{ik} is the event proportion in study i, arm k;
- μ_i is the log-odds in study i, control arm 1;
- δ_i^{bias} is the biased version of the log-odds ratio for study i;
- δ_i is the log-odds ratio for study i;
- β_i is the bias in study i;
- d is the mean log-odds ratio across studies;
- σ^2 is the between study variance;
- ν_i is the study-specific mean total internal bias;
- τ_i^2 is the study-specific variance of the total internal bias.

Bias-adjustment model III

In the practical, we will assume there is no bias for those studies with *no "high risk"* judgement, i.e. $\beta_i = 0$ for i in the "safe" set of studies: Clapp, Fanaroff, Sandberg, Weisman.

We will assume a common $\nu_i = \nu$ and $\tau_i^2 = \tau^2 \; \forall i$ for the set of "risky" studies, i.e. those that have at least 1 *"high risk"* expert judgement: Bussel, Chirico, Conway, Haque, Ratrisawadi, Tanzer.

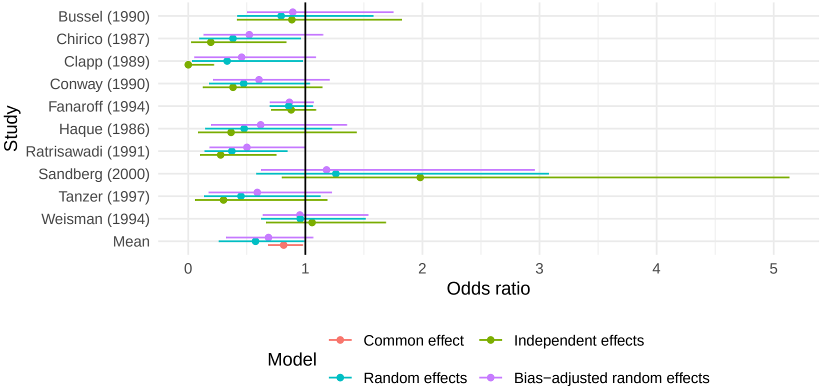
We will choose ν and τ^2 such that there is an average multiplicative effect of 0.82 on the odds ratio for the treatment effect, with an approximate prior 95% interval (0.67, 1), following Savovic et al (2012).

This prior implies that the treatment effects have been *exaggerated* by flawed study conduct.

Bias-adjusted results

Bias adjustment on treatment effect *moderates* estimate of treatment effect towards 1 for *“risky”* studies.

Bias-adjusted mean effect credible interval no longer excludes 1.



Practicals 2 and 3

We will use the HIV example to

- practise detecting lack-of-fit to the data, potentially due to conflicting evidence, using posterior-predictive p-values;
- practise resolving conflict by introducing bias parameters;
- practise accounting for more flexibility in a model by accounting for over-dispersion.

We will use the Sepsis example to

- practise detecting outliers using cross-validatory mixed-predictive checks;
- practise systematic bias adjustment to improve inference.

Summary of Session 7

The key messages of this session:

- Evidence synthesis can lead to more precise inference;
- Provided all the evidence included is consistent with each other;
- Importance of the model building and criticism cycle;
- Importance of external (possibly prior) evidence when introducing bias parameters.