Meta analysis using Sarmanov beta priors

May 2016

Chuck Chan

What is Meta analysis

- Methods that contrast and combine results from different studies.
- Used to identify patterns and sources of disagreement in multiple studies.
- Thorough summary of several studies
- Methods
  - Identifying a common measure (ie. Weighted average)

Advantages

- Generalized results for a larger population
- Precision improves as more data is used
- Inconsistent results can be quantified and analyzed. (ie sampling errors or external influence)
- Hypothesis testing can be applied to estimates
- Allows investigation of publication bias

Disadvantages

- Can be subject to agenda driven biases (ie. Conflict of interest, biased by funding source, abused for personal bias)
- Several small studies does not predict results of a single large study (one paper PMID 9262498)
- Correlation of accuracy characteristics is usually ignored
- Publication bias: negative or insignificant results are not usually published.

Steps

- Formulate problem
- Search literature
- Select studies with chosen criteria
- Decide on dependent variable or summary measures that will be used
  - Differences
  - Means
  - Effect size (R^2, Cohen f^2, Hedges' g)
- Model selection

Approaches

- Evidence used from studies
  - Individual Participant Data (IPD) : Raw data
  - Aggregate Data (AD): risk estimate or ratios
- Approach
  - One stage IPD are clustered
  - IPD converted to AD and weighted

Models

- Fixed Effects Model
  - Use inverse variance as weights
  - Not really realistic due to heterogeneity
- Random Effects Model
  - Inverse variance weight + unweight by applying Random effects variance component (variability of effect size)
  - Also can use Restricted Maximum likelihood.
- Quality Effects
  - Incorporate a relevant component (quality) that differs intra study that is used in fixed effects.
  - Uses tau i where i is a composite based on other studies used to redistribute quality to adjust weights
- Ivhet(inverse variance quasi likelihood based alternative)
  - coverage remains at the nominal level for confidence intervals
  - maintains the inverse variance weights of individual studies with increasing heterogeneity

Epidemiological studies

- Bayesian modeling approach
- comparison between two populations with binary outcomes
- summarized by a single or multiple 2 × 2 tables used for
- 2 × 2 tables used for Inference on the comparative measures of adverse event, or risks

Confidence intervals

- confidence intervals derived from conventional large sample theory(log likelihood expanded by 2 term Taylor series) often have poor coverage probabilities when the risk is rare or the sample size is small
- “zero cell” problem occur
  - Solved with adding an arbitrary positive number to the cells
  - arbitrary positive number makes the interpretation of results difficult and contradicting conclusions
- Confidence interval of odd ratios
  - Inverted Fisher's test
  - Fisher test coverage probabilities always greater than nominal levels and have been criticize of being too conservative.

Proposals for rare events

- Frequentist
  - confidence intervals have been proposed with the primary goal being to obtain the actual coverage probability close to the nominal level
  - Constructed by Inverted likelihood ratio test or inverted two sided test
- Bayesian
  - obtain the posterior distribution of odds ratios that reflects the evidence from the data and the available prior knowledge.
  - No “zero cell” problem because of a prior distribution of risk is assumed and inference based on the posterior distribution of the risk
  - Conjugate beta prior distributions for risks are often used due to simplicity and flexibility

Bayesian inference

- Exact Bayesian inference of a **single** or **multiple** 2×2 tables under a class of independent or correlated priors.
  - Advantages:
    - having closed form formulas for the posterior distributions of odds ratio
    - allowing for between studies heterogeneity and within study correlations.
- The ability to allow for within study correlation

Single 2 × 2 table

posterior distribution of odds ratio derived by Marshall

n_j: be the number of subjects

2 cases

- 1= case
- 2= control

Odds ratio of risk

$$\theta = \{p_2/(1-p_2)\}/\{p_1/(1-p_1)\}.$$

Beta function

$$\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

This derivation assumes case and control are **independent**

posterior distribution of odds ratio

$$f_\theta(\theta;\alpha_1,\beta_1,\alpha_2,\beta_2) = \theta^{-1-\beta_2}\{B(\alpha_1,\beta_1)B(\alpha_2,\beta_2)\}^{-1}B(\alpha_1+\alpha_2,\beta_1+\beta_2)$$
$$\times F(\alpha_2+\beta_2,\beta_1+\beta_2;\alpha_1+\alpha_2+\beta_1+\beta_2;1-\frac{1}{\theta}), \quad \text{for } \theta>0,$$

prior risks p_1 and p_2 are beta random variables with hyperparameters (a_1, b_1) and (a_2, b_2)

The posterior distributions of p_1 and p_2 are beta distributions with parameters (α_1, β_1) and (α_2, β_2) respectively, where $\alpha_j = y_j + a_j$ and $\beta_j = n_j - y_j + b_j\ (j = 1, 2)$.

Hypergeometric function

$$F(\alpha,\beta;\gamma;z) = \frac{1}{B(\beta,\gamma-\beta)}\int_0^1 t^{\beta-1}(1-t)^{\gamma-\beta-1}(1-tz)^{\alpha}dt, \quad \text{for } \gamma>\beta>0.$$

Posterior distribution of odds ratio

Independent priors between risks in cases and controls may be an **over-simplified** assumption because cases and controls within the same study are **likely to share some common factors**.

- Example: genetic association, in multivariate meta analysis multiple correlated outcomes of interests in several studies

The Sarmanov beta priors

Studied a family of bivariate distributions constructed from marginal distributions from PMC5789784

General form of Sarmanov bivariate distribution

p1 & p2 are random variables

Correlation

Note that when the correlation is 0 the product of two independent beta distributions.

specified marginal distributions

$$g(p_1, p_2) = f_1(p_1) f_2(p_2)\{1 + \rho \psi_1(p_1)\psi_2(p_2)\},$$

bounded integrable nonconstant functions that satisfy

$$\int \psi_j(t) f_j(t) dt = 0 \text{ for } j = 1, 2,$$

$$1 + \rho \psi_1(p_1)\, \psi_2(p_2) \geq 0 \quad \text{nonnegative}$$

Example

- Beta margins for p1 and p2 are assumed

$$f_j(p_j) = B(\alpha_j, \beta_j),$$
$$\psi_j(p_j) = (p_j - \mu_j)/\delta_j,$$

$$\text{mean of } p_j \qquad \mu_j = a_j/(a_j + b_j)$$

$$\text{square root of variance of } p_j \quad (j = 1, 2) \qquad \delta_j = \sqrt{\mu_j(1 - \mu_j)/(a_j + b_j + 1)}$$

Advantages of Sarmanov beta priors

- allows for both positive and negative correlations
- only needs specification of marginal distributions and correlation parameter
- can be expressed as linear combinations of independent bivariate beta distributions

exact posterior distribution of odds ratio under Sarmanov beta priors

posterior density function of odds ratio under independent beta priors

$$f_\theta^*(\theta;\alpha_1,\beta_1,\alpha_2,\beta_2,\rho) = \omega_1 f_\theta(\theta;\alpha_1,\beta_1,\alpha_2,\beta_2) + \omega_2 f_\theta(\theta;\alpha_1+1,\beta_1,\alpha_2,\beta_2) + \omega_3 f_\theta(\theta;\alpha_1,\beta_1,\alpha_2+1,\beta_2) + \omega_4 f_\theta(\theta;\alpha_1+1,\beta_1,\alpha_2+1,\beta_2),$$

Weights

ω_k $(k = 1, \ldots, 4)$ are functions of a_1, b_1, a_2, b_2, ρ

$$\omega_1 = \frac{v_1 B(\alpha_1,\beta_1)B(\alpha_2,\beta_2)}{CB(a_1,b_1)B(a_2,b_2)}, \quad \omega_2 = \frac{v_2 B(\alpha_1+1,\beta_1)B(\alpha_2,\beta_2)}{CB(a_1+1,b_1)B(a_2,b_2)},$$

$$\omega_3 = \frac{v_3 B(\alpha_1,\beta_1)B(\alpha_2+1,\beta_2)}{CB(\alpha_1,b_1)B(a_2+1,b_2)}, \quad \text{and} \quad \omega_4 = \frac{v_4 B(\alpha_1+1,\beta_1)B(\alpha_2+1,\beta_2)}{CB(\alpha_1+1,b_1)B(a_2+1,b_2)},$$

Normalizing constant

$$C = \frac{v_1 B(\alpha_1,\beta_1)B(\alpha_2,\beta_2)}{B(a_1,b_1)B(a_2,b_2)} + \frac{v_2 B(\alpha_1+1,\beta_1)B(\alpha_2,\beta_2)}{B(a_1+1,b_1)B(a_2,b_2)} + \frac{v_3 B(\alpha_1,\beta_1)B(\alpha_2+1,\beta_2)}{B(a_1,b_1)B(a_2+1,b_2)} + \frac{v_1 B(\alpha_1+1,\beta_1)B(\alpha_2+1,\beta_2)}{B(a_1+1,b_1)B(a_2+1,b_2)}.$$

exact posterior distribution of odds ratio under Sarmanov beta priors

Correlation is constraint between

$$-c/\max(a_1a_2, b_1b_2) \leq \rho \leq c/\max(a_1b_2, a_2b_1),$$
$$c = \sqrt{a_1a_2b_1b_2}/\sqrt{(a_1+b_1+1)(a_2+b_2+1)}.$$

Conditions sets range narrower than [-1,1]

$$f_\theta^*(\theta;\alpha_1,\beta_1,\alpha_2,\beta_2,\rho) = \omega_1 f_\theta(\theta;\alpha_1,\beta_1,\alpha_2,\beta_2) + \omega_2 f_\theta(\theta;\alpha_1+1,\beta_1,\alpha_2,\beta_2)$$
$$+ \omega_3 f_\theta(\theta;\alpha_1,\beta_1,\alpha_2+1,\beta_2) + \omega_4 f_\theta(\theta;\alpha_1+1,\beta_1,\alpha_2+1,\beta_2),$$

When correlation = 0

$$\omega_1 = 1 \text{ and } \omega_2 = \omega_3 = \omega_4 = 0,$$

Multiple 2 × 2 tables

- Bayesian hierarchical model

Sarmanov beta prior

hyperparameters
$(a_1, b_1, a_2, b_2, \rho)$

$$(p_{1i}, p_{2i}) \mid (a_1, b_1, a_2, b_2, \rho) \overset{i.id.}{\sim} g(p_1, p_2; a_1, b_1, a_2, b_2, \rho),$$

$$(y_{1i}, y_{2i}) \mid (n_{1i}, n_{2i}, p_{1i}, p_{2i}) \overset{ind.}{\sim} \text{Binomial}(y_{1i} \mid n_{1i}, p_{1i}) \times \text{Binomial}(y_{2i} \mid n_{2i}, p_{2i}),$$

number of exposed subjects

number of subjects

risk of being exposed in the *j*th group

(j=1,2 for case and control groups respectively)

Multiple 2 × 2 tables

dispersion parameter $\varphi_j = 1/(a_j + b_j + 1)$

two types of correlations between the exposure status for two subjects from the

same study and the same group φ_j

same study but different groups $\rho\sqrt{\varphi_1\varphi_2}$.

Hyperparameters

Obtained by maximizing the **log marginal likelihood** combining all studies as considered

When $\rho = 0$, the Sarmanov beta-binomial reduces to the independent beta-binomial model

$$\log L(a_1, b_1, a_2, b_2, \rho) =$$

$$\sum_{i=1}^{n} \log \left[P_{BB}(y_{1i}; n_{1i}, a_1, b_1) P_{BB}(y_{2i}; n_{2i}, a_2, b_2) \left\{ 1 + \frac{\rho}{\delta_1 \delta_2} \frac{(y_{1i} - n_{1i}\mu_1)(y_{2i} - n_{2i}\mu_2)}{(a_1 + b_1 + n_{1i})(a_2 + b_2 + n_{2i})} \right\} \right],$$

PMF binomial distribution

PMF beta-binomial distribution

Calculating intervals

- delta method to get the variance of log odds ratio.
- The Wald intervals for log odds ratio is then calculated and transformed to the Wald intervals for odds ratio.

overall odds ratio

- estimated by plugging in the estimates of hyperparameter

$$\theta = \{\mu_2/(1-\mu_2)\}/\{\mu_1/(1-\mu_1)\}$$

$$\mu_j = a_j/(a_j + b_j)$$

- study-specific odds ratio in the ith study

posterior distribution

hyperparameters were known

$$f^*_{\theta_i}(\theta_i; y_{1i} + a_1, n_{1i} - y_{1i} + b_1, y_{2i} + a_2, n_{2i} - y_{2i} + b_2, \rho)$$

inference based

$$f^*_{\theta_i}(\theta_i; y_{1i} + \widehat{a_1}, n_{1i} - y_{1i} + \widehat{b_1}, y_{2i} + \widehat{a_2}, n_{2i} - y_{2i} + \widehat{b_2}, \widehat{\rho})$$

ignores the uncertainty on the hyperparameter estimates

Confidence intervals obtained by bias correction method or bootstrap

Adjust for study level covariates

$$(p_{1i}, p_{2i}) \mid (a_1, b_1, a_2, b_2, \rho) \overset{i.id.}{\sim} g(p_1, p_2; a_1, b_1, a_2, b_2, \rho),$$
$$(y_{1i}, y_{2i}) \mid (n_{1i}, n_{2i}, p_{1i}, p_{2i}) \overset{ind.}{\sim} \text{Binomial}(y_{1i} \mid n_{1i}, p_{1i}) \times \text{Binomial}(y_{2i} \mid n_{2i}, p_{2i}),$$

Extend to regression settings
study-specific risk p_{ji} for $j = 1, 2$

$$p_{ji} \mid (\varphi_j, \mu_{ji}) \text{-Beta}\{p_{ji}; \mu_{ji}/(1/\varphi_j - 1), (1 - \mu_{ji})/(1/\varphi_j - 1)\}$$
$$\text{for } j = 1, 2,$$

Dispersion parameter

mean parameters of Beta Distribution

beta distribution
$\text{Beta}(p; \alpha, \beta)$
defined by $B(\alpha, \beta)^{-1} p^{\alpha-1}(1-p)^{\beta-1}$

$E[p_{ji}|\varphi_j, \mu_{ji}] = \mu_{ji}$
$\text{var}(p_{ji} \mid \varphi_j, \mu_{ji}) = \delta_{ji}^2 = \varphi_j \mu_{ji}(1 - \mu_{ji})$

mean of each Beta distribution is a function of covariates

$$\mu_{ji} = h^{-1}(X_i \eta_j) \text{ for } j = 1, 2,$$

where $h(\cdot)$ is some link function and X_i are the study-specific covariates related to study-specific risks.

Allowing Correlation between risks

bivariate beta-binomial regression model

assume paired study-specific risks (p_{1i}, p_{2i})

follow Sarmanov beta prior distribution

$$(p_{1i}, p_{2i}) \mid (\varphi_1, \mu_{1i}, \varphi_2, \mu_{2i})\text{-Beta}\{p_{1i}; \mu_{1i}/(1/\varphi_1 - 1), (1-\mu_{1i})/(1/\varphi_1 - 1)\}$$

$$\times \text{Beta}\{p_{2i}; \mu_{2i}/(1/\varphi_2 - 1), (1-\mu_{2i})/(1/\varphi_2 - 1)\}\left\{1 + \rho \frac{(p_{1i} - \mu_{1i})}{\delta_{1i}} \frac{(p_{2i} - \mu_{2i})}{\delta_{2i}}\right\}$$

dispersion parameter across different groups

can be fitted by maximizing the log marginal likelihood function

Example of usage from PMC5789784

N-acetyltransferase 2 acetylation status & colorectal cancer risk

Study Details

- N-acetyltransferase 2 (NAT2) gene
  - Metabolize hydrophobic compounds like carcinogens
  - Considered risk for colorectal cancer in various studies. But there are **inconsistent results** due to magnitude of association

Author	Cases		Control	
	no. events	no. observations	no. events	no. observations
Ilett	27	49	10	41
Ilett	27	49	19	45
Wohlleb	23	43	13	41
Ladero	49	109	40	96
Rodriguez	20	44	13	28
Lang	14	34	92	205
Oda	33	36	33	36
Shibuta	112	234	151	329
Bell	96	202	50	112
Spurr	32	103	34	96
Hubbard	100	275	140	343
Welfare	73	174	74	174
Gil	44	114	68	201
Chen	81	212	96	221
Lee	156	216	134	187
Yoshika	99	106	95	100
Potter	228	527	88	200
Slattery	931	1624	807	1963
Agundez	60	120	119	258
Butler	156	200	162	209

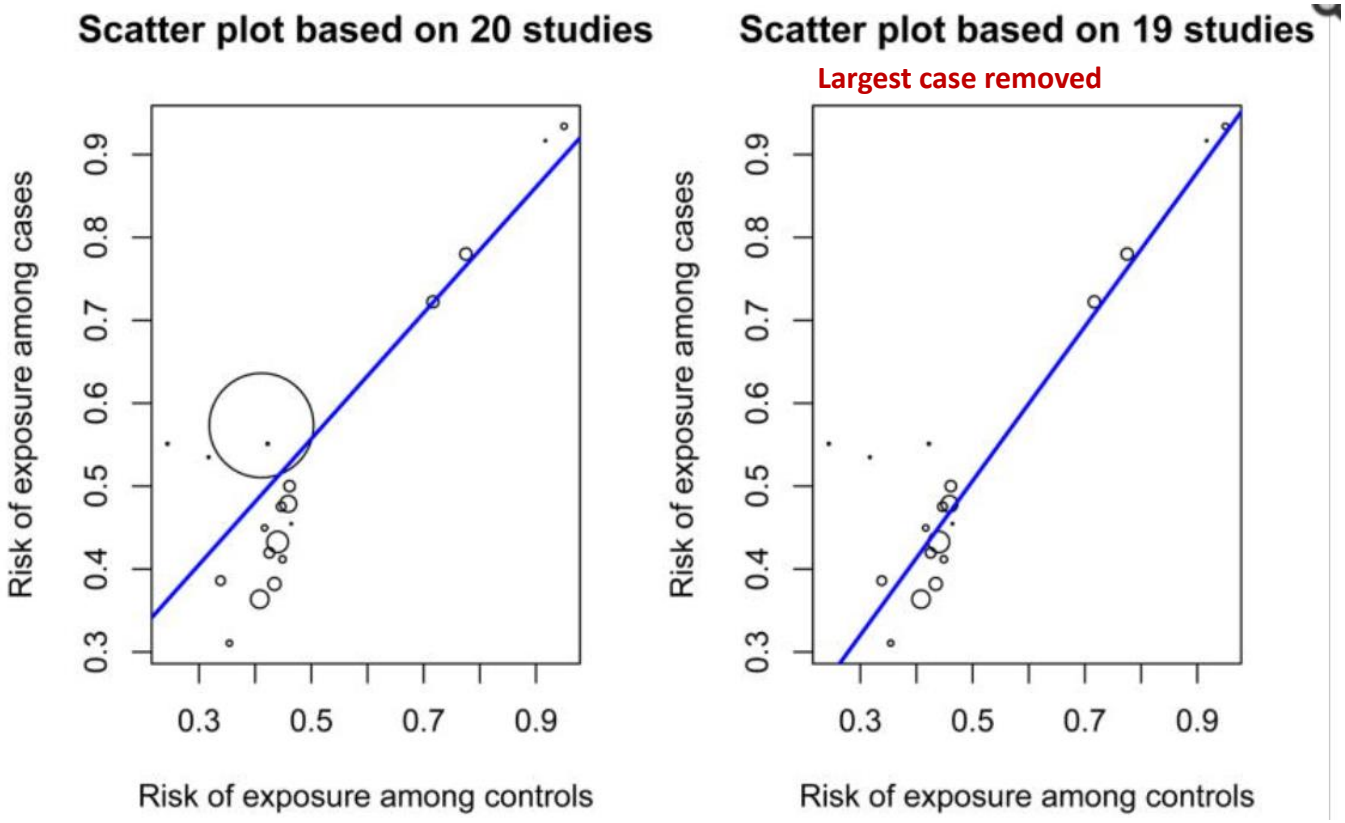
Dataset

- **Number of cases**: 22 cases of N-acetyltransferase 2 acetylation status and colorectal cancer risk.
- **Time**: January 1985 to October 2001
- Different locations including Australia, Japan, Spain, UK and USA
- **heterogeneity between studies**: Cases and controls in the same study are likely to share some common, but possibly unmeasured, factors such as ancestors.
- Probability of exposure in cases and controls within the same study were likely to be correlated
  - consider the consequence of ignoring within study correlation, and extend the current results under independent prior risk assumption to dependent prior assumption.

Correlation

- A strong within study correlation between probabilities of exposure in cases and controls
  - Pearson's 0.87
  - Spearman's rank 0.493
  - Kendall's tau 0.396
- All p-values less than 0.03

Correlation



Circle size = proportion to sample size

Strong positive correlation

Odds ratio

- within study has to be accounted for to ensure valid inference on odds ratio
- **odds ratio** : ratio of odds of having rapid NAT2 acetylator status vs. those with colorectal cancer to those without.
- fit both independent beta-binomial model and Sarmanov beta-binomial model.

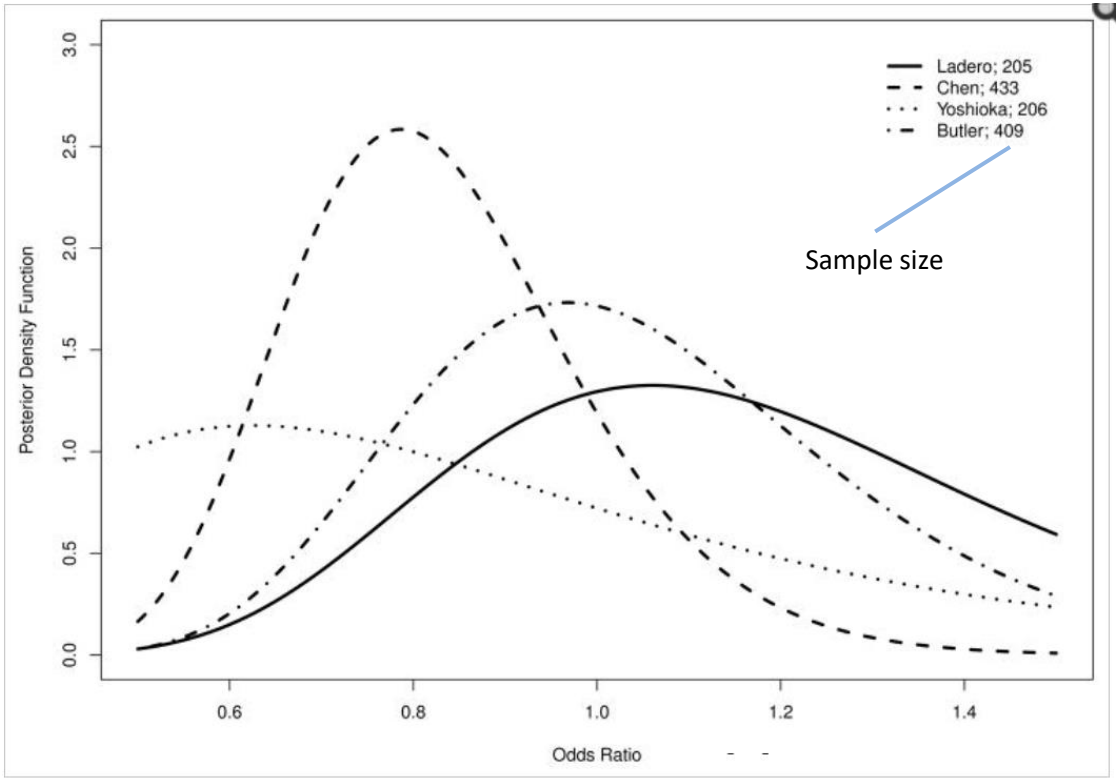
Sarmanov beta-binomial model

- likelihood ratio test yields a p-value of 0.075

obtained the estimates of hyperparameters $(\hat{a}_1, \hat{b_1}, \hat{a}_2, \hat{b_2}, \hat{\rho}) = (3.108, 2.914, 3.942, 3.361, 0.125)$
exact posterior distribution of each study-specific odds ratio using

$$f_\theta^*(\theta; \alpha_1, \beta_1, \alpha_2, \beta_2, \rho) = \omega_1 f_\theta(\theta; \alpha_1, \beta_1, \alpha_2, \beta_2) + \omega_2 f_\theta(\theta; \alpha_1 + 1, \beta_1, \alpha_2, \beta_2)$$
$$+ \omega_3 f_\theta(\theta; \alpha_1, \beta_1, \alpha_2 + 1, \beta_2) + \omega_4 f_\theta(\theta; \alpha_1 + 1, \beta_1, \alpha_2 + 1, \beta_2),$$

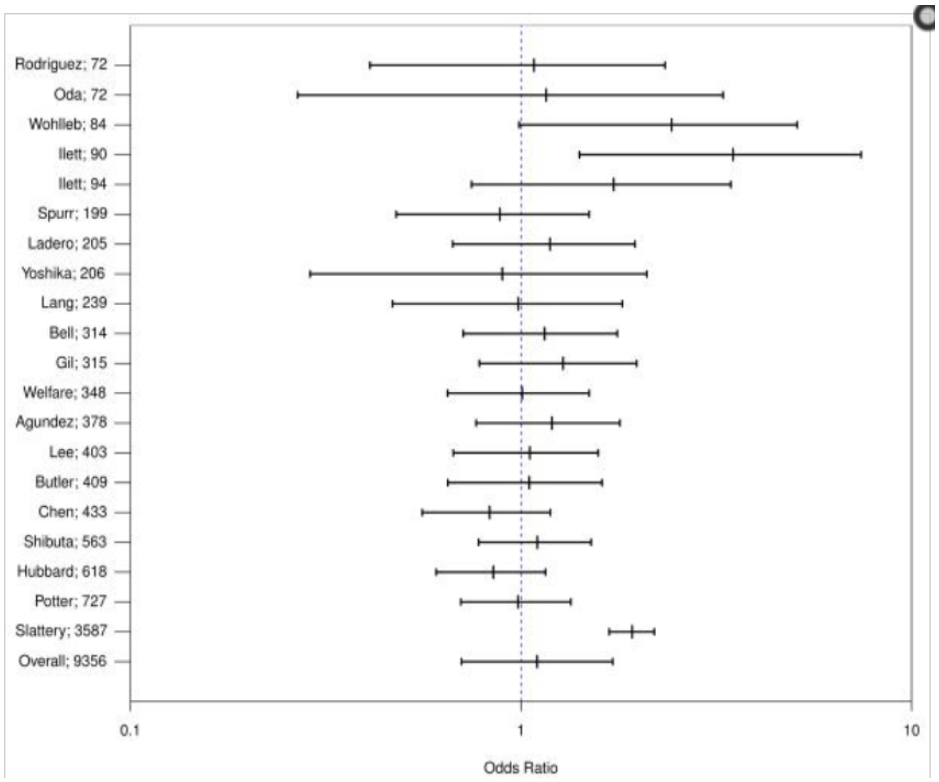
Posterior distributions of study-specific odds ratios for four studies



- Odds ratios are defined as the ratio of odds of having rapid N-acetyltransferase 2 (NAT2) acetylator status comparing those with colorectal cancer to those without

Calculating intervals

forest plot with credible intervals of study-specific odds ratios and confidence interval of overall odds ratio



- Bisection root-finding method to compute the 2.5% and 97.5% quantiles, we constructed the 95% equal-tail credible intervals of each study-specific odds ratio.
- odds ratio is estimated by $(\hat{a}_2\hat{b}_1)/(\hat{a}_1\hat{b}_2)$ and the 95% confidence interval is constructed by exponentiating the Wald's intervals of overall log odds ratio.

Calculating intervals

- The overall odds ratio for rapid NAT2 acetylator status and colorectal cancer risk is 1.100 (95% CI: 0.704, 1.718).
- Overall odds ratio estimated from the independent beta-binomial model is 1.138 (95% CI: 0.717, 1.806)
- Both not significant
- Sarmanov beta-binomial model provides sizable efficiency gain compared to independent beta-binomial model due to its ability of accounting for correlation within studies (relative efficiency is 0.867)