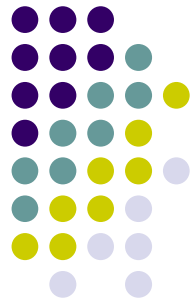
Reconstruction of individual patient data for meta-analysis via Bayesian approach

Yusuke Yamaguchi, Wataru Sakamoto and Shingo Shirahata

Graduate School of Engineering Science, Osaka University

Masashi Goto

Biostatistical Research Association, NPO

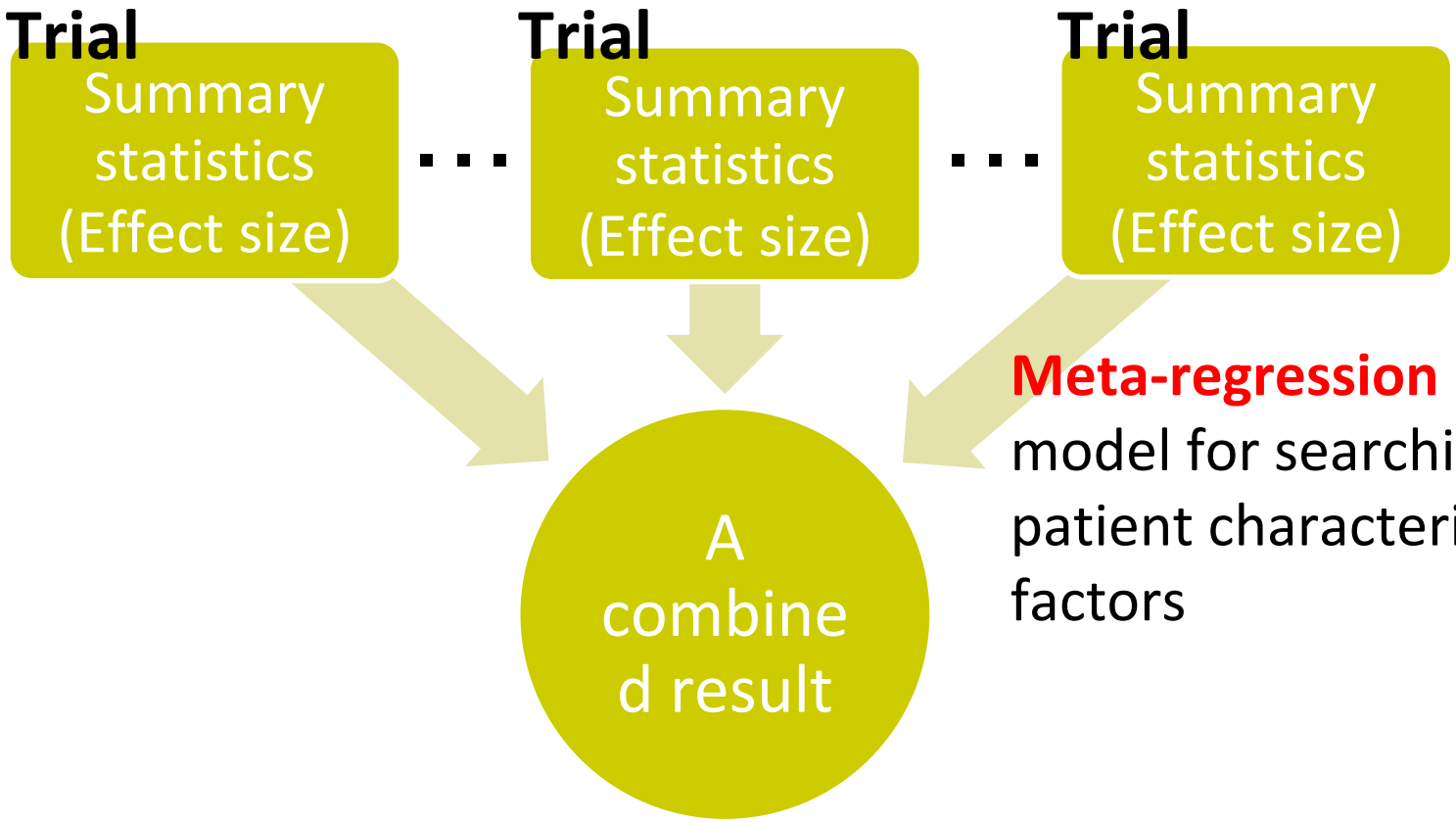


Outline

- Introduction
- Individual patient data and aggregate data
  - 2×2 contingency tables in which margins only are observed
- Proposed method
  - A method based on simulated individual patient data
- Simulation study
- Concluding remarks and future problems

Introduction

- Meta-analysis based on aggregate data (AD)

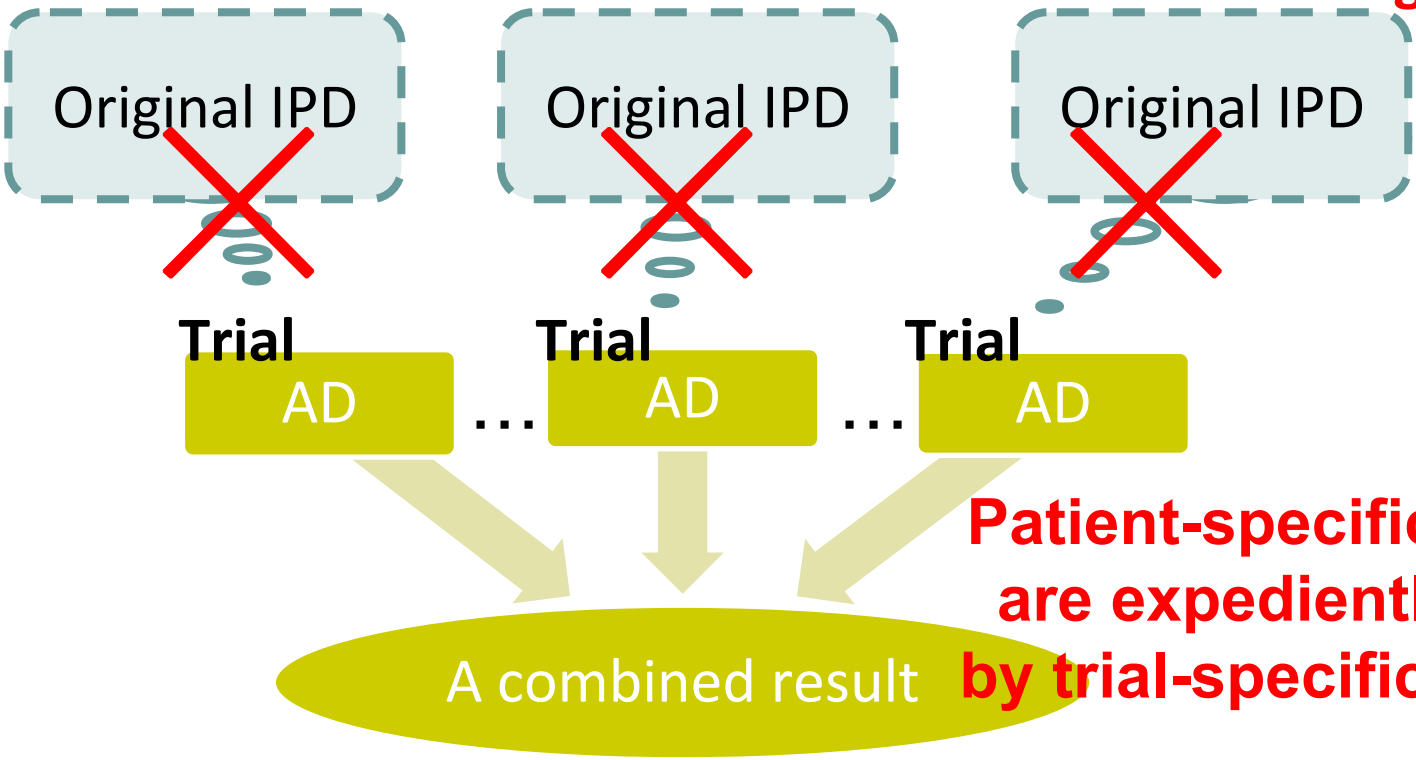


Meta-regression (MR) : model for searching patient characteristic factors

Introduction

- Individual patient data (IPD)

The scheme of sampling IPD is ignored.



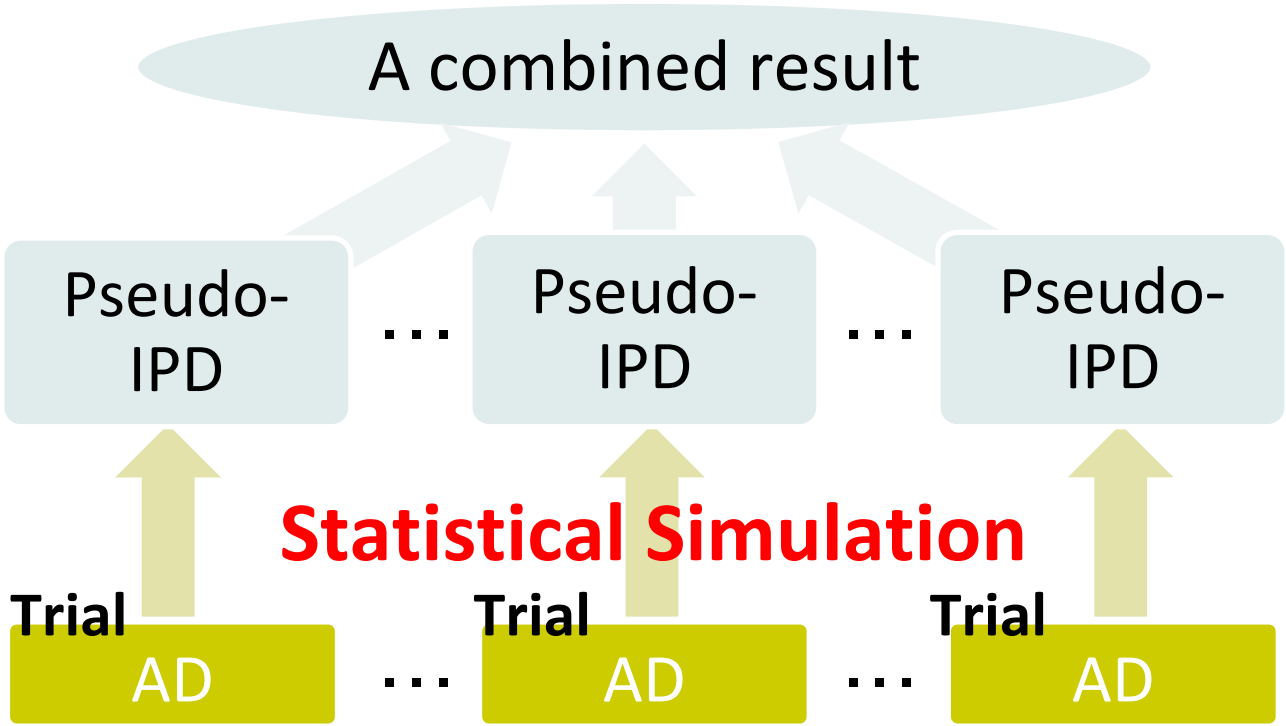
Patient-specific covariates are expediently replaced by trial-specific covariates.

Introduction

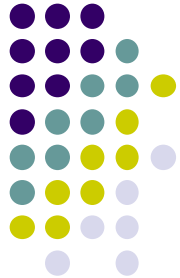
- Criticisms of MR models (Riley & Steyerberg, 2010; Berlin *et al.*, 2002; Thompson & Higgins, 2002)
  - MR assumes that the "across-trial relationship" between summary estimates and mean covariate values reflects the "within-trial relationship" between individual response and covariate values; however, this may not be true in practice.
  - The relationship described by MR is an observational association, so this suffers from the bias by confounding.
  - The power of MR to detect patient-level covariates that truly modify response is generally low.
- IPD should be used to explore the patient characteristic factors. **But, collecting IPD may often be impossible.**

Proposed method:
a method based on simulated IPD (SIPD method)

- Pseudo-IPD from each trial are generated, and then their simulated IPD are combined.



Situation: a binary outcome and a binary covariate (controlled trial with two groups)



- IPD (individual binary variables for each patient)
  - y_{ij}: outcome, x_{ij} : group indicator, z_{ij} : covariate ($\bar{z}_i = n_i^{-1}\sum_{j=1}^{n_i} z_{ij}$)
  - i: the number of trials ($i = 1,\ldots,I$)
  - j: the number of patients ($j = 1,\ldots, n_i$)
  - IPD model: Fixed effect logistic regression model (Riley *et al.*, 2008)

$$y_{ij} \sim \text{Bernoulli}(q_{ij})$$

between-trial effect　　within-trial effect

$$\log(q_{ij}/1-q_{ij}) = \alpha_i + \beta_1 x_{ij} + \beta_2 z_{ij} + \gamma_{\mathrm{B}} x_{ij}\bar{z}_i + \gamma_{\mathrm{W}} x_{ij}(z_{ij} - \bar{z}_i)$$

- AD (number of patients partially classified by each variable)
  - Fixed effect MR model

$$\log \widehat{\mathrm{OR}}_i = a + b\bar{z}_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, \sigma_i^2)$$

$$\widehat{\mathrm{OR}}_i = \frac{m_{1i}(n_{10i} + n_{00i} - m_{0i})}{m_{0i}(n_{11i} + n_{01i} - m_{1i})}, \sigma_i^2 = \mathrm{Var}\left(\log \widehat{\mathrm{OR}}_i\right)$$

Trial *i*	$X=0$	$X=1$
$Z=0$	n_{00i}	n_{01i}
$Z=1$	n_{10i}	n_{11i}
$Y=1$	m_{0i}	m_{1i}

Notation and problem

- For group k in i-th trial

	$Y=0$	$Y=1$	
$Z=0$		m_{0ki}	n_{0ki}
$Z=1$		m_{1ki}	n_{1ki}
	$n_{ki}-m_{ki}$	m_{ki}	n_{ki}

Not available

Observed

- $Y_{\text{NA-IPD}} = \{(m_{0ki}, m_{1ki}, n_{0ki}, n_{1ki}): i = 1,\ldots,I, k = 0,1\}$
- $Y_{\text{AD}} = h(Y_{\text{NA-IPD}}) = \{(m_{ki}, n_{0ki}, n_{1ki}): i=1,\ldots,I, k = 0,1\}$

Synthesis for 2×2 contingency tables in which margins only observed

Overview of SIPD method

- Assumption
  - For all values of $Y_{\text{NA-IPD}}$ that is consistent with the observed data Y_{AD}, the conditional distribution of Y_{AD} given $Y_{\text{NA-IPD}}$ takes the same values (coarsening at random; Heitjan & Rubin, 1991).
- Observed-data likelihood function: parameter θ

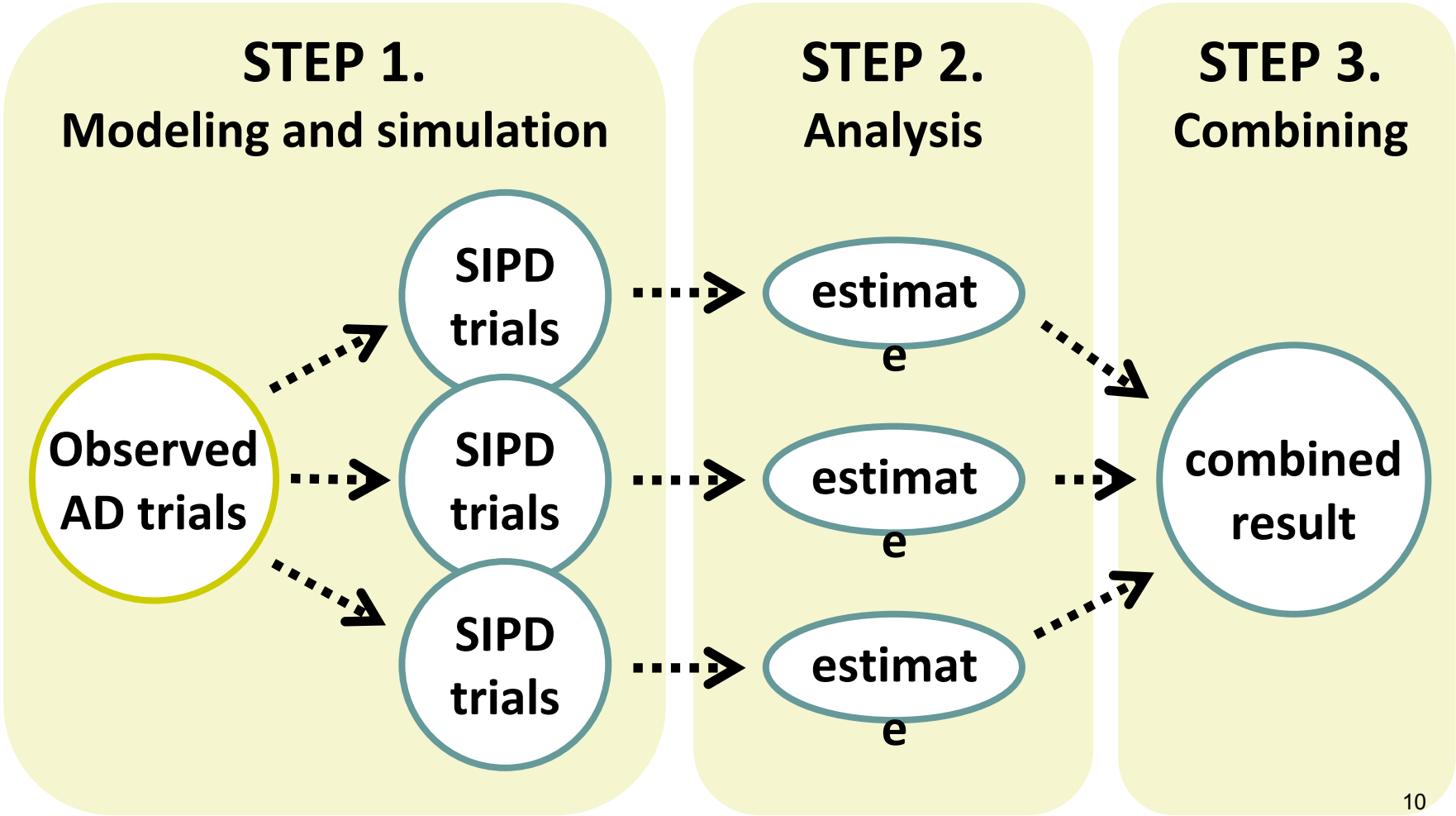
$$p(Y_{\text{AD}} \mid \theta) = \int_{Y_{\text{AD}} = h(Y_{\text{NA-IPD}})} p(Y_{\text{NA-IPD}} \mid \theta) dY_{\text{NA-IPD}}$$

- Bayesian model specification (King, 1999; Wakefield, 2004)

$$p(Y_{\text{NA-IPD}}, \theta \mid Y_{\text{AD}}) = p(Y_{\text{NA-IPD}} \mid Y_{\text{AD}}, \theta)\, \pi(\theta \mid Y_{\text{AD}})$$

$$\pi(\theta \mid Y_{\text{AD}}) \propto p(Y_{\text{AD}} \mid \theta)\, \pi(\theta)$$

Implementing procedure



STEP 1. Observed-data likelihood function

- Underlying distribution

$$M_{0ki} \mid p_{0ki} \sim \mathrm{Bin}(p_{0ki}, n_{0ki})$$
$$M_{1ki} \mid p_{1ki} \sim \mathrm{Bin}(p_{1ki}, n_{1ki})$$
$$\begin{pmatrix} p_{0ki} = \Pr(Y=1 \mid Z=0,k,i) \\ p_{1ki} = \Pr(Y=1 \mid Z=1,k,i) \end{pmatrix}$$

Z	$Y=0$	$Y=1$	
0		M_{0ki}	n_{0ki}
1		M_{1ki}	n_{1ki}
	$n_{ki}-m_{ki}$	m_{ki}	n_{ki}

- Convolution likelihood (Wakefield, 2004; McCullagh & Nelder, 1989)

$$p(m_{ki}, n_{0ki}, n_{1ki} \mid p_{0ki}, p_{1ki})$$
$$= \sum_{m_{0ki}=l_{ki}}^{u_{ki}} \binom{n_{0ki}}{m_{0ki}} \binom{n_{1ki}}{m_{ki}-m_{0ki}} p_{0ki}^{m_{0ki}} (1-p_{0ki})^{n_{0ki}-m_{0ki}} p_{1ki}^{m_{ki}-m_{0ki}} (1-p_{1ki})^{n_{1ki}-m_{ki}+m_{0ki}}$$
$$l_{ki} = \max(0, m_{ki}-n_{1ki}), \quad u_{ki} = \min(n_{0ki}, m_{ki})$$

STEP 1. Bayesian model specification

- Joint posterior distribution of unobserved data and parameters given $Y_{\mathrm{AD}(ki)} = (m_{ki}, n_{0ki}, n_{1ki})$

$$p(M_{0ki}, p_{0ki}, p_{1ki} \mid Y_{\mathrm{AD}(ki)}) = p(M_{0ki} \mid Y_{\mathrm{AD}(ki)}, p_{0ki}, p_{1ki})\, \pi(p_{0ki}, p_{1ki} \mid Y_{\mathrm{AD}(ki)})$$

$$M_{0ki} \mid p_{0ki}, p_{1ki}$$

$$\sim \text{Noncentral-hypergeometric}\left(\frac{p_{0ki}(1-p_{1ki})}{p_{1ki}(1-p_{0ki})}, Y_{\mathrm{AD}(ki)}\right)$$

$$\log p_{0ki}/(1-p_{0ki}) = \alpha + \beta_1 k + (\gamma_{\mathrm{B}} - \gamma_{\mathrm{W}})k\bar{z}_i$$
$$\log p_{1ki}/(1-p_{1ki}) = \alpha + \beta_1 k + \beta_2 + (\gamma_{\mathrm{B}} - \gamma_{\mathrm{W}})k\bar{z}_i + \gamma_{\mathrm{W}} k$$

$$\pi(\theta \mid Y_{\mathrm{AD}}) \propto \pi(\theta) \prod_{i=1}^{I} \prod_{k=0}^{1} p(Y_{\mathrm{AD}(ki)} \mid \theta), \quad \theta = (\alpha, \beta_1, \beta_2, \gamma_{\mathrm{B}}, \gamma_{\mathrm{W}})$$

STEP 1. Draw from posterior distributions

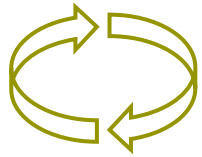
- Random sampling from the posterior distribution of $\theta = (\alpha, \beta_1, \beta_2, \gamma_B, \gamma_W)$ by Markov chain Monte Carlo method (Gelman *et al.*, 1995)
  - Single component Metropolis-Hastings algorithm

$$\alpha^{[t]} \sim \pi(\alpha \mid Y_{AD}, \beta_1^{[t-1]}, \beta_2^{[t-1]}, \gamma_B^{[t-1]}, \gamma_W^{[t-1]})$$

$$\beta_1^{[t]} \sim \pi(\beta_1 \mid Y_{AD}, \alpha^{[t]}, \beta_2^{[t-1]}, \gamma_B^{[t-1]}, \gamma_W^{[t-1]})$$

$$\vdots$$

$$\gamma_W^{[t]} \sim \pi(\gamma_W \mid Y_{AD}, \alpha^{[t]}, \beta_1^{[t]}, \beta_2^{[t]}, \gamma_B^{[t]})$$



iteration

→ Get T sets of parameters

$$\theta^{[t]} = (\alpha^{[t]}, \beta_1^{[t]}, \beta_2^{[t]}, \gamma_B^{[t]}, \gamma_W^{[t]}), t = 1, \ldots, T$$

$$\Leftrightarrow \quad \{(p_{0ki}^{[t]}, p_{1ki}^{[t]}) : i = 1, \ldots, I, k = 0, 1\}, t = 1, \ldots, T$$

STEP 1. Generate SIPD

- Random sampling from the posterior predictive distribution of M_{0ki}

$$p(M_{0ki} \mid Y_{\mathrm{AD}(ki)})$$

$$= \iint p(M_{0ki} \mid Y_{\mathrm{AD}(ki)}, p_{0ki}, p_{1ki})\, \pi(p_{0ki}, p_{1ki} \mid Y_{\mathrm{AD}(ki)}) dp_{0ki} dp_{1ki}$$

$$\approx \frac{1}{T} \sum_{t=1}^{T} p(M_{0ki} \mid Y_{\mathrm{AD}(ki)}, p_{0ki}^{[t]}, p_{1ki}^{[t]})$$

→ Get R sets of SIPD ($m_{1ki} = m_{ki} - m_{0ki}$)

$$\{(m_{0ki}^{[r]}, m_{1ki}^{[r]}) : i = 1, \ldots, I, k = 0, 1\}, r = 1, \ldots, R$$

$$\Leftrightarrow \quad \{(y_{ij}^{[r]}, x_{ij}^{[r]}, z_{ij}^{[r]}) : i = 1, \ldots, I; j = 1, \ldots, n_i\}, r = 1, \ldots, R$$

STEP 2. Fit a standard model to SIPD

- Fixed effect logistic regression model

$$Y_{ij}^{[r]} \square \text{Bernoulli}(q_{ij})$$

$$\log(q_{ij}/1-q_{ij}) = \alpha_i + \beta_1 x_{ij}^{[r]} + \beta_2 z_{ij}^{[r]} + \gamma_{\text{B}} x_{ij}^{[r]} \overline{z}_i + \gamma_{\text{W}} x_{ij}^{[r]} (z_{ij}^{[r]} - \overline{z}_i)$$

→ Get a point estimate of the parameters of interest and its standard error from r-th set of SIPD

e.g. $\left(\hat{\gamma}_{\text{W}}^{[r]}, \text{SE}(\hat{\gamma}_{\text{W}}^{[r]})\right), r = 1, \ldots, R$

STEP 3. Combine the results from each SIPD

- Rubin's (1987) combining rule (e.g. inference on γ_W)
  - Point estimate and standard error for γ_W

$$\hat{\gamma}_W = \frac{1}{R}\sum_{r=1}^{R}\hat{\gamma}_W^{[r]}$$

$$\mathrm{SE}(\hat{\gamma}_W) = \left\{\frac{1+R^{-1}}{R-1}\sum_{r=1}^{R}(\hat{\gamma}_W^{[r]}-\hat{\gamma}_W)^2 + \frac{1}{R}\sum_{r=1}^{R}\mathrm{Var}(\hat{\gamma}_W^{[r]})\right\}^{1/2}$$

  - Reference distribution

$$(\hat{\gamma}_W - \gamma_W)/\mathrm{SE}(\hat{\gamma}_W) \sim t_\nu$$

$$\nu = (R-1)\left(1+\hat{\gamma}_W\Big/(R-1)^{-1}\sum_{r=1}^{R}(\hat{\gamma}_W^{[r]}-\hat{\gamma}_W)^2\right)$$

Simulation study

- Objective
  - To compare the estimates of treatment-covariate interaction effect (b or γ_{W}), which are obtained by three methods.
- Methods
  - Fit a fixed logistic regression model to original IPD
  - Fit an MR model to AD summarized from IPD
  - Apply SIPD method to AD summarized from IPD ($R = 100$ sets of SIPD)

Designs

- True IPD model (Lambert *et al*, 2002)

	$Z=0$		$Z=1$	
	$\Pr(Y=1 \mid Z=0)$	odds	$\Pr(Y=1 \mid Z=1)$	odds
$k=0$	0.1	-	0.3	-
$k=1$	0.1	1.00	0.2	0.58

- Each patient in each trial and group was allocated to $Z=1$ by the probabilities 0.3, 0.4, 0.5, 0.6 or 0.7. These probabilities were used with two, four or six trials, respectively.
- $(I, n_{ki}) \in \{(10,100), (20,100), (30,100)\}$, $i = 1,\ldots, I$, $k = 0,1$

  $\rightarrow$ Generate 100 sets of original IPD for each scenario

Result

		Using original IPD		Using AD only			
		IPD γ_W		SIPD γ_W		MR b	
No. of trials	No. of patients	p<0.05	Mean (SD) of estimates	p<0.05	Mean (SD) of estimates	p<0.05	Mean (SD) of estimates
10	100	52%	-0.56 (0.26)	29%	-0.60 (0.33)	9%	-0.42 (0.84)
20	100	80%	-0.53 (0.21)	52%	-0.55 (0.25)	12%	-0.43 (0.63)
30	100	96%	-0.56 (0.15)	82%	-0.58 (0.18)	16%	-0.47 (0.47)

*SD: Standard Deviation, p<0.05: percentage of significant results

- "IPD" represents the analysis for the original IPD, which cannot be performed in the case of general meta-analysis
- The SIPD method provided the results on treatment-covariate interaction effect closer to those obtained by fitting the IPD model than fitting the MR model

Concluding remarks

- We presented the SIPD method for meta-analyzing binary data from multiple trials.
- The SIPD method provided more appropriate estimates for treatment-covariate interaction effect. Moreover, it was shown that the SIPD method preserved remarkably higher power to detect these effects than those obtained by fitting the MR model.

Future problems

- It will be necessary to verify the benefit of the SIPD method thought various simulation studies and practical case studies. We are particularly interested in whether this improves publication bias which is one of the most important research topics in meta-analysis.
- Although we considered only Bayesian approach for the model specification in SIPD method, we would be able to take a frequentist approach for this.

References

- Berlin J.A., Santanna J., Scmid C.H., Szczech L.A. & Feldman H.I. (2002). Individual patient versus group-level data meta-regression for the investigation of treatment effect modifiers: Ecological bias rears its ugly head. *Statistics in Medicine*, **21**, 371-387.
- Gelman A., Carlin J.B., Stern H.S. & Rubin D.B. (1995). *Bayesian Data Analysis*. Chapman and Hall.
- Heitjan D.F. & Rubin D.B. (1991). Ignorability and coarse data. *The Annals of Statistics*, **19**, 2244-2253.
- King G., Rosen O. & Tanner M.A. (1999). Binomial-beta hierarchical models for ecological inference. *Sociological Methods and Research*, **28**, 61-90.
- Lambert P.C., Sutton A.J., Abrams K.R. & Jones D.R. (2002). A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology*, **55**, 86-94.
- McCullagh H. & Nelder J.A. (1989). *Generalized Linear Models, 2nd edition*, London: Chapman and Hall.
- Riley R.D., Lambert P.C., Staessen J.A., Wang J., Gueyffier F., Thijs L. & Boutitie F. (2008). Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Statistics in Medicine*, **27**, 1870-1893.

References

- Riley R.D. & Steyerberg E.W. (2010). Meta-analysis of a binary outcome using individual participant data and aggregate data. *Research Synthesis Methods*, **1**, 2–19.
- Rubin D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
- Thompson S.G. & Higgins J.P.T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, **21**, 1559–1573.
- Tsiatis A.A. (2006). *Semiparametric Theory and Missing Data*. Springer.
- Wakefield J. (2004). Ecological inference for 2*2 tables. *Journal of the Royal Statistical Society: Series A*, **167**, 385-445.
- Wang N. & Robins J.M. (1998). Large-sample theory for parametric multiple imputation procedure. *Biometrika*, **85**, 935-948.

Thank you very much for your attention

yamaguti@sigmath.es.osaka-u.ac.jp