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
  - $2 \times 2$ 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.

A combined result

Statistical Simulation
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)

\[
\log\left(\frac{q_{ij}}{1-q_{ij}}\right) = \alpha_i + \beta_1 x_{ij} + \beta_2 z_{ij} + \gamma_B x_{ij} \bar{z}_i + \gamma_W x_{ij}(z_{ij} - \bar{z}_i)
\]

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

\[
\log\Theta R_i = a + b \bar{z}_i + \varepsilon_i, \quad \varepsilon_i \sim \text{N}(0, \sigma_i^2)
\]

\[
\Theta R_i = \frac{m_{1i}(n_{10i} + n_{00i} - m_{0i})}{m_{0i}(n_{11i} + n_{01i} - m_{1i})}, \quad \sigma_i^2 = \text{Var}\left(\log\Theta R_i\right)
\]

<table>
<thead>
<tr>
<th>Trial ( i )</th>
<th>( X = 0 )</th>
<th>( X = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Z = 0 )</td>
<td>( n_{00i} )</td>
<td>( n_{01i} )</td>
</tr>
<tr>
<td>( Z = 1 )</td>
<td>( n_{10i} )</td>
<td>( n_{11i} )</td>
</tr>
<tr>
<td>( Y = 0 )</td>
<td>( n_{10i} )</td>
<td>( n_{11i} )</td>
</tr>
<tr>
<td>( Y = 1 )</td>
<td>( m_{0i} )</td>
<td>( m_{1i} )</td>
</tr>
</tbody>
</table>
Notation and problem

- For group $k$ in $i$-th trial

<table>
<thead>
<tr>
<th>$Z = 0$</th>
<th>$Y = 0$</th>
<th>$m_{0ki}$</th>
<th>$n_{0ki}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z = 1$</td>
<td>$Y = 0$</td>
<td>$m_{1ki}$</td>
<td>$n_{1ki}$</td>
</tr>
<tr>
<td></td>
<td>$Y = 1$</td>
<td>$n_{ki} - m_{ki}$</td>
<td>$m_{ki}$</td>
</tr>
</tbody>
</table>

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

**Synthesis for $2 \times 2$ contingency tables in which margins only observed**
Overview of SIPD method

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

- Observed-data likelihood function: parameter $\theta$
  \[
p(Y_{AD} \mid \theta) = \int_{Y_{AD} = h(Y_{NA-IPD})} p(Y_{NA-IPD} \mid \theta) dY_{NA-IPD}
  \]

- Bayesian model specification (King, 1999; Wakefield, 2004)
  \[
p(Y_{NA-IPD}, \theta \mid Y_{AD}) = p(Y_{NA-IPD} \mid Y_{AD}, \theta) \pi(\theta \mid Y_{AD})
  \]
  \[
  \pi(\theta \mid Y_{AD}) \propto p(Y_{AD} \mid \theta) \pi(\theta)
  \]
Implementing procedure

STEP 1. Modeling and simulation
- Observed AD trials
- SIPD trials
- SIPD trials
- SIPD trials

STEP 2. Analysis
- estimat
- estimat
- estimat

STEP 3. Combining
- combined result
STEP 1. Observed-data likelihood function

- Underlying distribution

\[
\begin{align*}
M_{0ki} & \mid p_{0ki} \overset{\text{Bin}}{\sim} \text{Bin}(p_{0ki}, n_{0ki}) \\
M_{1ki} & \mid p_{1ki} \overset{\text{Bin}}{\sim} \text{Bin}(p_{1ki}, n_{1ki}) \\
\left(p_{0ki} = \Pr(Y = 1 \mid Z = 0, k, i)\right) \\
\left(p_{1ki} = \Pr(Y = 1 \mid Z = 1, k, i)\right)
\end{align*}
\]

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

\[
p(m_{ki}, n_{0ki}, n_{1ki} \mid p_{0ki}, p_{1ki}) = \sum_{m_{0ki} = l_{ki}}^{u_{ki}} \binom{n_{0ki}}{m_{0ki}} \binom{n_{1ki}}{m_{ki} - m_{0ki}} p_{0ki}^{m_{0ki}} (1 - p_{0ki})^{n_{0ki} - m_{0ki}} p_{1ki}^{m_{ki} - m_{0ki}} (1 - p_{1ki})^{m_{ki} - m_{0ki} + m_{0ki}}\]

\[
l_{ki} = \max(0, m_{ki} - n_{1ki}), \quad u_{ki} = \min(n_{0ki}, m_{ki})
\]
STEP 1. Bayesian model specification

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

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

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

- Noncentral-hypergeometric

$$\begin{pmatrix} \frac{p_{0ki}(1 - p_{1ki})}{p_{1ki}(1 - p_{0ki})}, Y_{AD(ki)} \end{pmatrix}$$

$$\log p_{0ki} / (1 - p_{0ki}) = \alpha + \beta_1 k + (\gamma_B - \gamma_W)k\bar{z}_i$$

$$\log p_{1ki} / (1 - p_{1ki}) = \alpha + \beta_1 k + \beta_2 + (\gamma_B - \gamma_W)k\bar{z}_i + \gamma_W k$$

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

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

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

\[
: \quad \gamma_W^{[t]} \overset{\text{from}}{\sim} \pi(\gamma_W \mid Y_{AD}, \alpha^{[t]}, \beta_1^{[t]}, \beta_2^{[t]}, \gamma_B^{[t]})
\]

\( \rightarrow \) Get \( T \) sets of parameters

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

\[\Leftrightarrow \{(p_{0ki}^{[t]}, p_{1ki}^{[t]}): i = 1, \ldots, I, k = 0, 1\}, t = 1, \ldots, T\]
STEP 1. Generate SIPD

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

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

⇒ Get \( R \) sets of SIPD \( (m_{1ki} = m_{ki} - m_{0ki}) \)

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

\[
\Leftrightarrow \{(y_{ij}^{[r]}, x_{ij}^{[r]}, z_{ij}^{[r]}): i = 1, \ldots, I; j = 1, \ldots, n_i\}, r = 1, \ldots, R
\]
STEP 2. Fit a standard model to SIPD

- Fixed effect logistic regression model

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

\[ \log\left(\frac{q_{ij}}{1-q_{ij}}\right) = \alpha_i + \beta_1 x_{ij}^{[r]} + \beta_2 z_{ij}^{[r]} + \gamma_B x_{ij}^{[r]} z_i + \gamma_W x_{ij}^{[r]} (z_{ij}^{[r]} - \bar{z}_i) \]

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

e.g. \( (\hat{\gamma}_W^{[r]}, \text{SE}(\hat{\gamma}_W^{[r]})), r = 1, \ldots, R \)
STEP 3. Combine the results from each SIPD

- Rubin’s (1987) combining rule (e.g. inference on $\gamma_W$)
  - Point estimate and standard error for $\gamma_W$
    \[
    \hat{\gamma}_W = \frac{1}{R} \sum_{r=1}^{R} \hat{\gamma}^{[r]}_W
    \]
    \[
    SE(\hat{\gamma}_W) = \left\{ \frac{1 + R^{-1}}{R - 1} \sum_{r=1}^{R} (\hat{\gamma}^{[r]}_W - \hat{\gamma}_W)^2 + \frac{1}{R} \sum_{r=1}^{R} \text{Var}(\hat{\gamma}^{[r]}_W) \right\}^{1/2}
    \]
  - Reference distribution
    \[
    (\hat{\gamma}_W - \gamma_W) / SE(\hat{\gamma}_W) \sim t_\nu
    \]
    \[
    \nu = (R - 1) \left( 1 + \frac{\hat{\gamma}_W}{(R - 1)^{-1} \sum_{r=1}^{R} (\hat{\gamma}^{[r]}_W - \hat{\gamma}_W)^2} \right)
    \]
Simulation study

- **Objective**
  - To compare the estimates of treatment-covariate interaction effect \( b \) or \( \gamma_w \), 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 \( (R = 100 \text{ sets of SIPD}) \)
Designs

- True IPD model (Lambert et al, 2002)

<table>
<thead>
<tr>
<th></th>
<th>$Z = 0$</th>
<th></th>
<th>$Z = 1$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$Pr(Y = 1 \mid Z = 0)$</td>
<td>odds</td>
<td>$Pr(Y = 1 \mid Z = 1)$</td>
<td>odds</td>
<td></td>
</tr>
<tr>
<td>$k = 0$</td>
<td>0.1</td>
<td>-</td>
<td>0.3</td>
<td>-</td>
</tr>
<tr>
<td>$k = 1$</td>
<td>0.1</td>
<td>1.00</td>
<td>0.2</td>
<td>0.58</td>
</tr>
</tbody>
</table>

- 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.

- $(I, n_{ki}) \subseteq \{(10,100), (20,100), (30,100)\}$, $i = 1, \ldots, I$, $k = 0,1$
  
  $\rightarrow$ Generate 100 sets of original IPD for each scenario
## Result

<table>
<thead>
<tr>
<th>No. of trials</th>
<th>No. of patients</th>
<th>p&lt;0.05</th>
<th>Mean (SD) of estimates</th>
<th>p&lt;0.05</th>
<th>Mean (SD) of estimates</th>
<th>p&lt;0.05</th>
<th>Mean (SD) of estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>IPD</strong></td>
<td><strong>γ_w</strong></td>
<td><strong>SIPD</strong></td>
<td><strong>γ_w</strong></td>
<td><strong>MR</strong></td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>52%</td>
<td>-0.56 (0.26)</td>
<td>29%</td>
<td>-0.60 (0.33)</td>
<td>9%</td>
<td>-0.42 (0.84)</td>
</tr>
<tr>
<td>20</td>
<td>100</td>
<td>80%</td>
<td>-0.53 (0.21)</td>
<td>52%</td>
<td>-0.55 (0.25)</td>
<td>12%</td>
<td>-0.43 (0.63)</td>
</tr>
<tr>
<td>30</td>
<td>100</td>
<td>96%</td>
<td>-0.56 (0.15)</td>
<td>82%</td>
<td>-0.58 (0.18)</td>
<td>16%</td>
<td>-0.47 (0.47)</td>
</tr>
</tbody>
</table>

*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

References

Thank you very much for your attention

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