Meta-analysis of binary or continuous outcomes combining individual patient data and aggregate data

Neha Agarwala[1], Anindya Roy[1], and Junyong Park[1]

[1] *University of Maryland Baltimore, Baltimore County*

Abstract

Often both aggregate or meta-analysis (MA) studies and Individual Patient Data (IPD) studies are available for specific treatments. Combining these two sources of data could improve the overall meta-analytic estimates of treatment effects. We propose a method to combine treatment effects across trials where the response can be binary or continuous. For some studies with MA data, the associated IPD maybe available, albeit at some extra effort or cost to the analyst. We consider the case when treatment effects are fixed and common across studies and evaluate the wisdom of choosing MA when IPD is available by studying the relative efficiency of analyzing all IPD studies versus combining various percentages of MA and IPD studies. For many different models design constraints under which the MA estimators are the IPD estimators, and hence fully efficient, are known. For such models we advocate a selection procedure that chooses MA studies over IPD studies in a manner that force least departure from design constraints and hence ensures a fully efficient combined MA and IPD estimator.

1 Introduction

When both the meta-analysis (MA) and Individual Patient Data (IPD) studies are available, combining these two levels of data could improve the overall meta-analysis estimates, compared to utilizing MA studies alone. One common application of combining studies is estimating the effect of several treatments in a multicenter trial. We consider the cases for both binary and continuous responses and provide the combined treatment effects across trials when the treatment effect is fixed and common across trials. Assuming the observations within and between the studies are independent having a common variance, we investigate the loss of efficiency from using combined estimator with various percentages of MA and IPD studies. When treatments are fixed and trial effects are random, Mathew and Nordstorm (2010) derived the necessary and sufficient condition for the IPD estimator to coincide with meta-analysis estimator for a general within trial covariance matrix. The condition for equality requires that the fraction of observations corresponding to any given treatment is same across trials for the linear model. In practice, it is more likely to have studies with differential allocation to treatments. For such models, we studied the relative efficiency of analyzing IPD versus combining IPD & MA studies under systematic departures from the same allocation proportions condition.

For models with binary outcome, the condition for equality of IPD estimator and meta-analysis estimator involves more than the treatment allocation in different studies.

2 Methods

We will start with the problem of combining a one-dimensional parameter of interest and finally extend it to estimating multidimensional parameter. The general idea is to aggregate information from IPD studies and meta-analysis studies (summary statistics) which includes results from published journals,etc to obtain a combined estimate for our parameter. Individual Patient Data estimator is the gold standard. However getting usually involves extra cost and hence we may have limited IPD resources. Thus we propose a method to select the IPD studies among the available studies so as to get the maximum efficiency in terms of the combined estimator.

2.1 Aggregation

Consider that there is one continuous outcome of interest and assume that there are two groups, namely treatment(T) and control(C) group for all k studies. Let n_{i1} and n_{i2} be the number of persons in each group for study i with $n_{i1}+n_{i2} = n_i$. WLG suppose we can k_1 studies for which we have access to Individual Patient Data and k_2 studies with meta-analysis results so that $k_1 + k_2 = k$. The model for response y_{ij} for the jth patient in study i is

$$\begin{gathered} y_{ij} = \alpha_i + \beta x_{ij} + \epsilon_{ij}, \quad i = 1, ..., k_1 \\ \epsilon_{ij} \sim N(0, \sigma_i^2), \quad \alpha_i \sim N(\alpha, \sigma_\alpha^2) \end{gathered} \tag{1}$$

where α_i and ϵ_{ij} are assumed independent and x_{ij} is 0/1 denoting the treatment and control groups respectively. In meta-analysis literature, we generally assume σ^2 and σ_α^2 are known.

Here, α acts as a common nuisance parameter across studies. When combining information across these k_1 studies to estimate β, the meta-analysis estimator and the IPD estimator doesn't necessarily coincide. For our model, the the two estimators coincide if and only if the vectors $(n_{i1}/n_i, n_{i2}/n_i)$, are all equal for $i = 1, ..., k_1$.

The distribution of $\mathbf{y}$ for the i^{th} study is

$$\begin{aligned} E(\boldsymbol{y_i}) &= \alpha \mathbf{1}_{\boldsymbol{n_i}} + \beta X_i \\ Cov(\boldsymbol{y_i}) &= H_i = \sigma_\alpha^2 \mathbf{1}_{\boldsymbol{n_i}} \mathbf{1}_{\boldsymbol{n_i}}^{\boldsymbol{T}} + \sigma_i^2 I_{n_i} \end{aligned} \tag{2}$$

For the k_2 MA studies, assuming the same model as (1), we have the restricted maximum likelihood estimates (REML) and their estimated variances. Hence the model for meta-analysis study is

$$\begin{aligned}\hat{\beta}_i &\sim N(\beta, \hat{\sigma}_i^2), \quad i = 1, \dots, k_2 \\ \hat{\sigma}_i^2 &= \frac{n_{i1} n_{i2}}{n_i \sigma_i^2} = \frac{n_i \pi_i (1 - \pi_i)}{\sigma_i^2}\end{aligned} \tag{3}$$

where $\pi_i = n_{i1}/n_i$.

With the above model, the combined estimate of (α, β) is

$$\begin{pmatrix} \hat{\alpha}_{comb} \\ \hat{\beta}_{comb} \end{pmatrix} = (U^T \Sigma^{-1} U)^{-1} U^T \Sigma^{-1} Y^* \tag{4}$$

and the variance of the combined estimate is give below:

$$Cov \begin{pmatrix} \hat{\alpha}_{comb} \\ \hat{\beta}_{comb} \end{pmatrix} = (U^T \Sigma^{-1} U)^{-1} \tag{5}$$

where

$$Y^* = \begin{pmatrix} \mathbf{y} \\ \hat{\boldsymbol{\beta}} \end{pmatrix}$$

$$U = \begin{pmatrix} \mathbf{1}_{\boldsymbol{n_{ipd}}} & X \\ 0 & \mathbf{1}_{\boldsymbol{k_2}} \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}$$

$$\Sigma_1 = \begin{pmatrix} H_1 & & \\ & \ddots & \\ & & H_{k_1} \end{pmatrix}$$

$$\Sigma_2 = \begin{pmatrix} \hat{\sigma}_1^2 & & \\ & \ddots & \\ & & \hat{\sigma}_{k_2}^2 \end{pmatrix}$$

and n_{ipd} is the total number of patients for the IPD studies.

The variance for the combined estimate for the parameter of interest β assuming all $n_i = n$ and $\sigma_i^2 = \sigma^2$ to the following:

$$var(\hat{\beta}_{comb}) = \sigma^2 \left[\frac{n^2 b}{a} \sum_{i=1}^{k_1} \pi_i (1 - \pi_i) + n \sum_{i=1}^{k_2} \pi_i (1 - \pi_i) + \frac{n}{a k_1} \Big(\sum_{i=1}^{k_1} \pi_i \Big) \Big(\sum_{i=1}^{k_1} (1 - \pi_i) \Big) \right]^{-1} \tag{6}$$

where $a = 1 + nb, \ b = \frac{\sigma_\alpha^2}{\sigma^2}$ & $\pi_i = \frac{n_{i1}}{n_i}$.

The variance for estimator of β when IPD data is available for all k studies is

$$var\left(\hat{\beta}_{allipd}\right) = \sigma^2\left[\frac{n^2b}{a}\sum_{i=1}^{k}\pi_i(1-\pi_i) + \frac{n}{ak}\left(\sum_{i=1}^{k}\pi_i\right)\left(\sum_{i=1}^{k}(1-\pi_i)\right)\right]^{-1} \tag{7}$$

Note that the above expression is different from the variance of the estimator of β when IPD data is available for all k studies in case of a fixed effects model. In a fixed effects model,

$$var\left(\hat{\beta}_{allipd(FE)}\right) = \sigma^2\left[\frac{n}{k}\left(\sum_{i=1}^{k}\pi_i\right)\left(\sum_{i=1}^{k}(1-\pi_i)\right)\right]^{-1} \tag{8}$$

The variance for estimator of β when summary statistics is available for all k studies is

$$var(\hat{\beta}_{allMA}) = \sigma^2\left[n\sum_{i=1}^{k}\pi_i(1-\pi_i)\right]^{-1} \tag{9}$$

Note that the variance for all meta-analysis studies isn't different for fixed effects model since we get the estimates first for each study and then combine the information across studies whereas for IPD studies we combine the data first and then get the variance.

The relative efficiency of the combined estimator with respect to the estimator with all IPD studies is

$$\begin{aligned} RE\left(\hat{\beta}_{comb}\right) &= \frac{var(\hat{\beta}_{allipd})}{var(\hat{\beta}_{comb})} \\ &= \frac{\left[\frac{n^2b}{a}\sum_{i=1}^{k_1}\pi_i(1-\pi_i) + n\sum_{i=k_1}^{k_2}\pi_i(1-\pi_i) + \frac{n}{ak_1}\left(\sum_{i=1}^{k_1}\pi_i\right)\left(\sum_{i=1}^{k_1}(1-\pi_i)\right)\right]}{\left[\frac{n^2b}{a}\sum_{i=1}^{k}\pi_i(1-\pi_i) + \frac{n}{ak}\left(\sum_{i=1}^{k}\pi_i\right)\left(\sum_{i=1}^{k}(1-\pi_i)\right)\right]} \end{aligned} \tag{10}$$

We see that the expressions for variance and relative efficiency do not involve y and depends only on π_i's, n and σ's.

2.1.1 Properties of Combined estimator

Comment on unbiasedness, consistency, asymptotic efficiency

Table 1. **Relative efficiency (RE) of combined estimator to IPD estimator for different percentage of MA studies in the estimator**

	0 percent	**20 percent**	**40 percent**	**60 percent**	**80 percent**	**100 percent**
Estimate	1.503	1.502	1.503	1.503	1.503	1.503
RE	0.970	0.983	0.986	0.991	0.997	1

Table 2. **Distribution of proportion of treatment across studies**

Study1	Study2	Study3	Study4	Study5	Study6	Study7	Study8	Study9	Study10
0.1	0.2	0.3	0.3	0.3	0.5	0.6	0.6	0.8	0.8

2.2 Selection

k_1 and k_2 has to be chosen based on cost constraint.
Some idea:

We want the $\hat{\beta}_{comb}$ to be efficient. Each of the terms in variance has be large. Given k_1,

- Select IPD from both the ends since the 3rd term in $var(\hat{\beta}_{comb})$ can be rewritten as $\frac{nk_1}{a}\tilde{\pi}(1-\tilde{\pi})$ where $\tilde{\pi} = \frac{\sum_{i=1}^{k_1}\pi_i}{k_1}$. This term is maximum when $\tilde{\pi} = 0.5$. In other words we want to achieve $\sum_{i=1}^{k_1}\pi_i = k_1/2$.

- For the 2nd term in the expression, we want each $\pi \approx 0.5$ for all MA studies.

- Among the first term and 2nd term, the 2nd term has more weight $(= n)$ whereas the first term has weight $< n$.

3 Numerical Results

For simulation studies, we present two main scenarios to show the importance of selection of studies for the combined estimator. We considered both k and n to be small and equal to 10, $\beta = 1.5, \alpha = 0.5, \sigma^2 = 2.5, \sigma_\alpha^2 = 0.025$.

For $\pi_i = 0.3$ for all $i = 1(1)10$ and $\beta = 1.5$, the combined estimator performs quite well as is expected shown in table 1. This is true as the proportion of treatment is equal for all studies. However when the π_i are not equal, it we propose to choose the IPD studies so as the reach a specified efficiency.

3.1 Uniform distribution of proportion of treatment

For this scenario, we used *runif* to generate $k = 10$ random proportion of treatment over the interval $[0, 1]$ in table 2:

If the desired relative efficiency is 0.95, we could achieve that with lots of possible combinations of 60% IPD and 40% MA studies. However 60% IPD may not be cost effective. But if we choose the right combination of 40% IPD

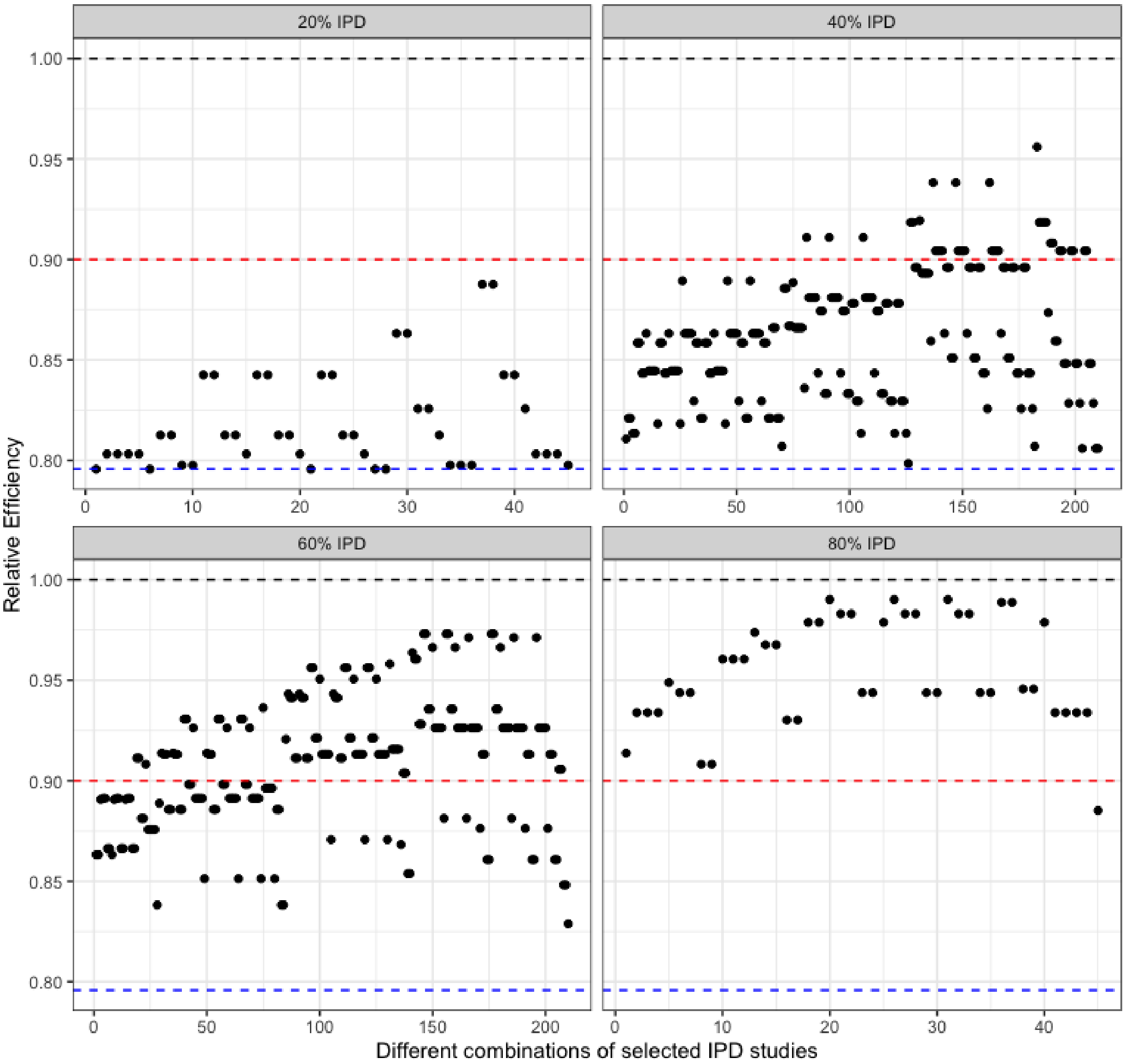

Figure 1. The relative efficiency of all 45 possible combinations for each of 20% IPD and 80% IPD, 210 possible combinations for each of 40% IPD and 60% IPD are plotted. Black dashed line represents the RE for all IPD studies which is 1, blue dashed line is RE for all MA studies which is 0.79 and the red dashed line is the desired RE, say 0.9, for example.

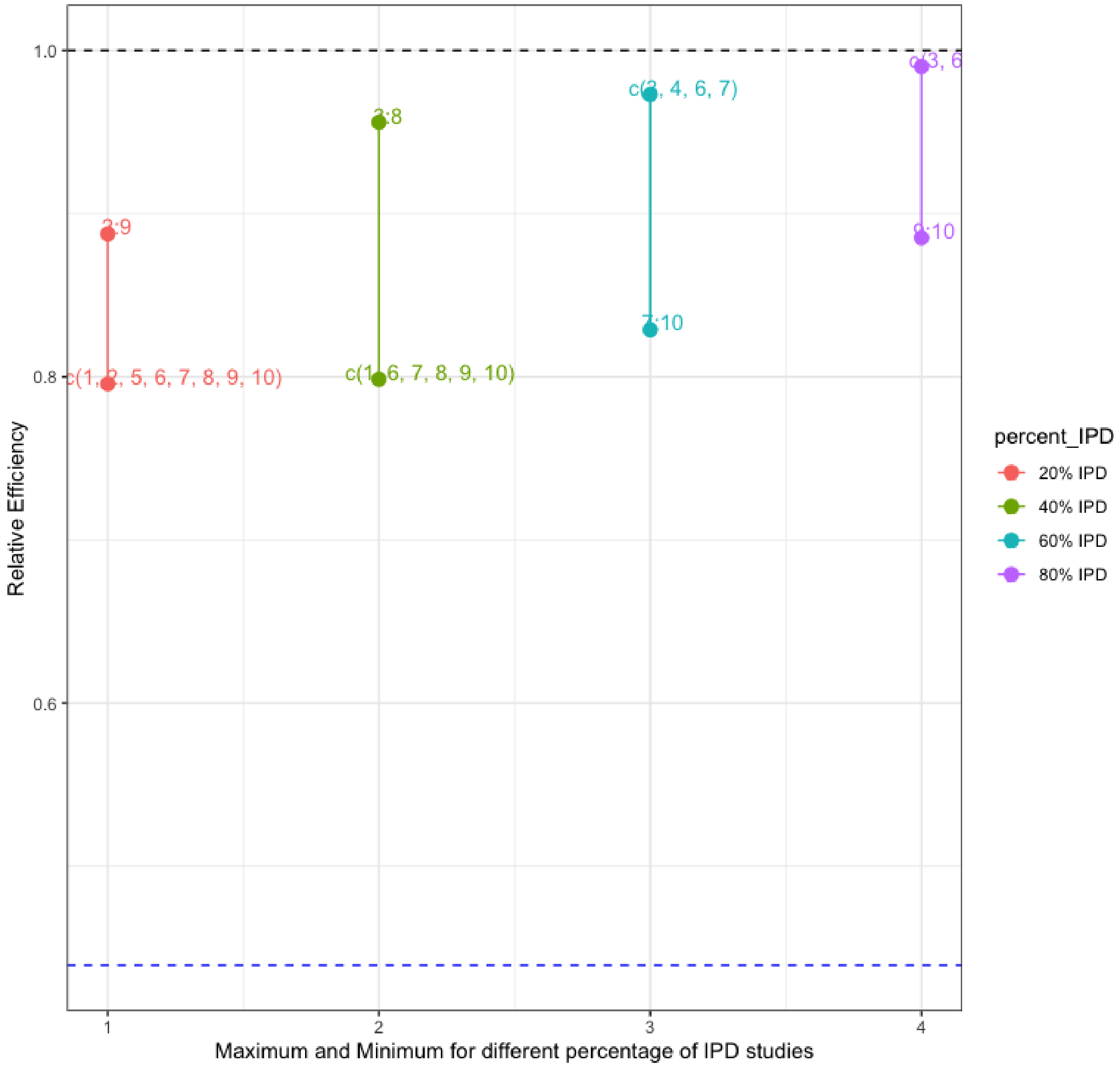

Figure 2. Plot showing the maximum and minimum relative efficiency and the MA combination within each percentage of IPD and MA studies.

Table 3. **Distribution of proportion of treatment across studies**

Study1	Study2	Study3	Study4	Study5	Study6	Study7	Study8	Study9	Study10
0.1	0.1	0.1	0.1	0.2	0.8	0.9	0.9	0.9	0.9

and 60% MA, we can have a RE of 0.956 in this case. The combination which maximizes the RE for 40% IPD is study no 1,2,9 and 10 for IPD and 3:8 for MA.

The combinations 2:9 in 20% IPD, for example, represents that the study no 2:9 accounts for 80% MA studies and study no 1 & 10 account for 20% IPD and this distribution of studies between MA and IPD has the maximum relative efficiency among all combinations in 20% IPD group.

3.2 Bathtub distribution of proportion of treatment

This is an example of unbalanced proportion of treatment in table 3.

The importance of selection of right IPD and MA study for the combined estimate is quite evident here. The all meta estimator has a relative efficiency of 0.44 wrt to the all IPD estimator. Th situation can worsen for severely unbalanced distribution of proportion of treatment. However we need to quantify unbalancedness before we move forward.

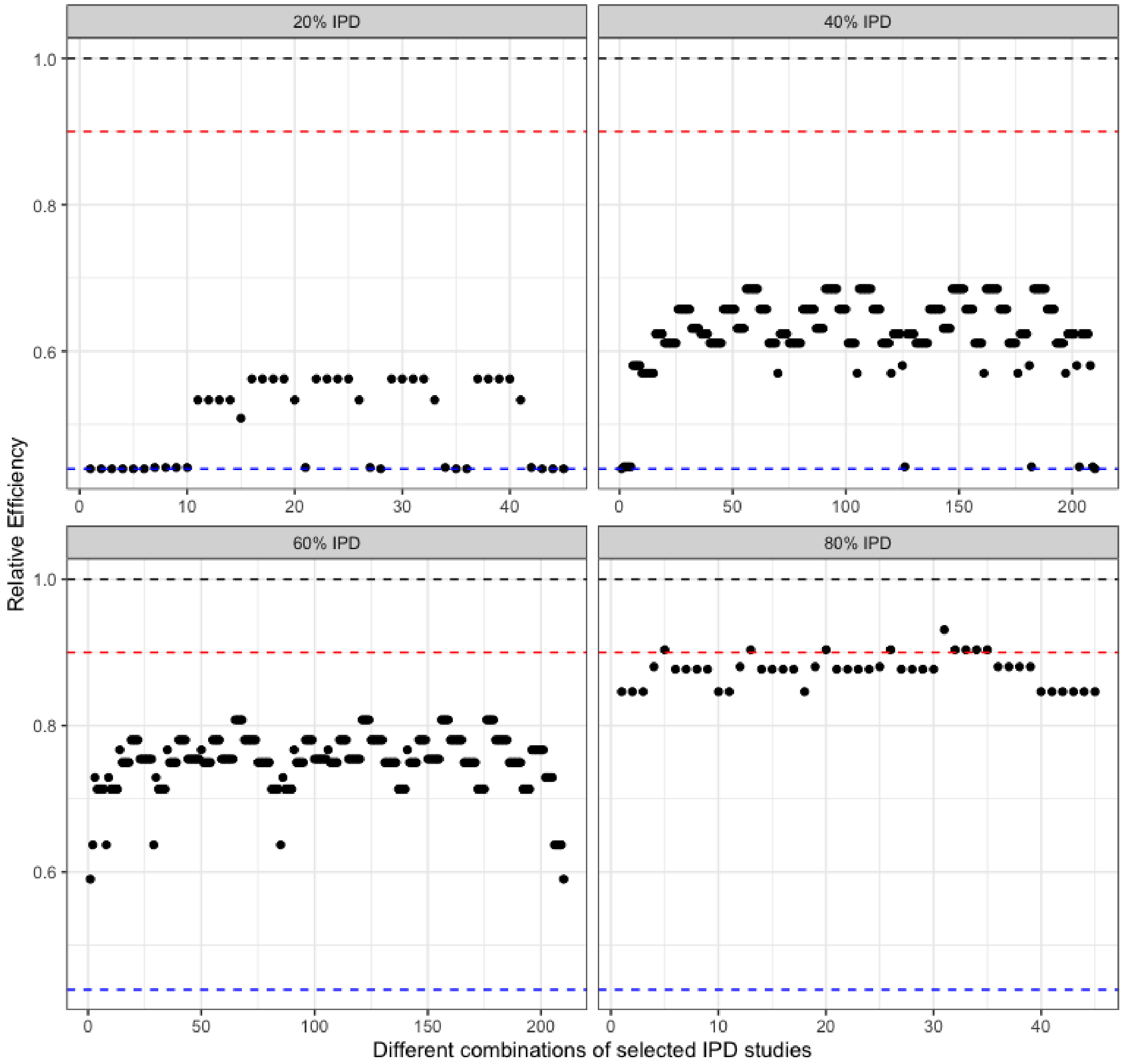

Figure 3. The relative efficiency of all 45 possible combinations for each of 20% IPD and 80% IPD, 210 possible combinations for each of 40% IPD and 60% IPD are plotted. Black dashed line represents the RE for all IPD studies which is 1, blue dashed line is RE for all MA studies which is 0.79 and the red dashed line is the desired RE, say 0.9, for example.

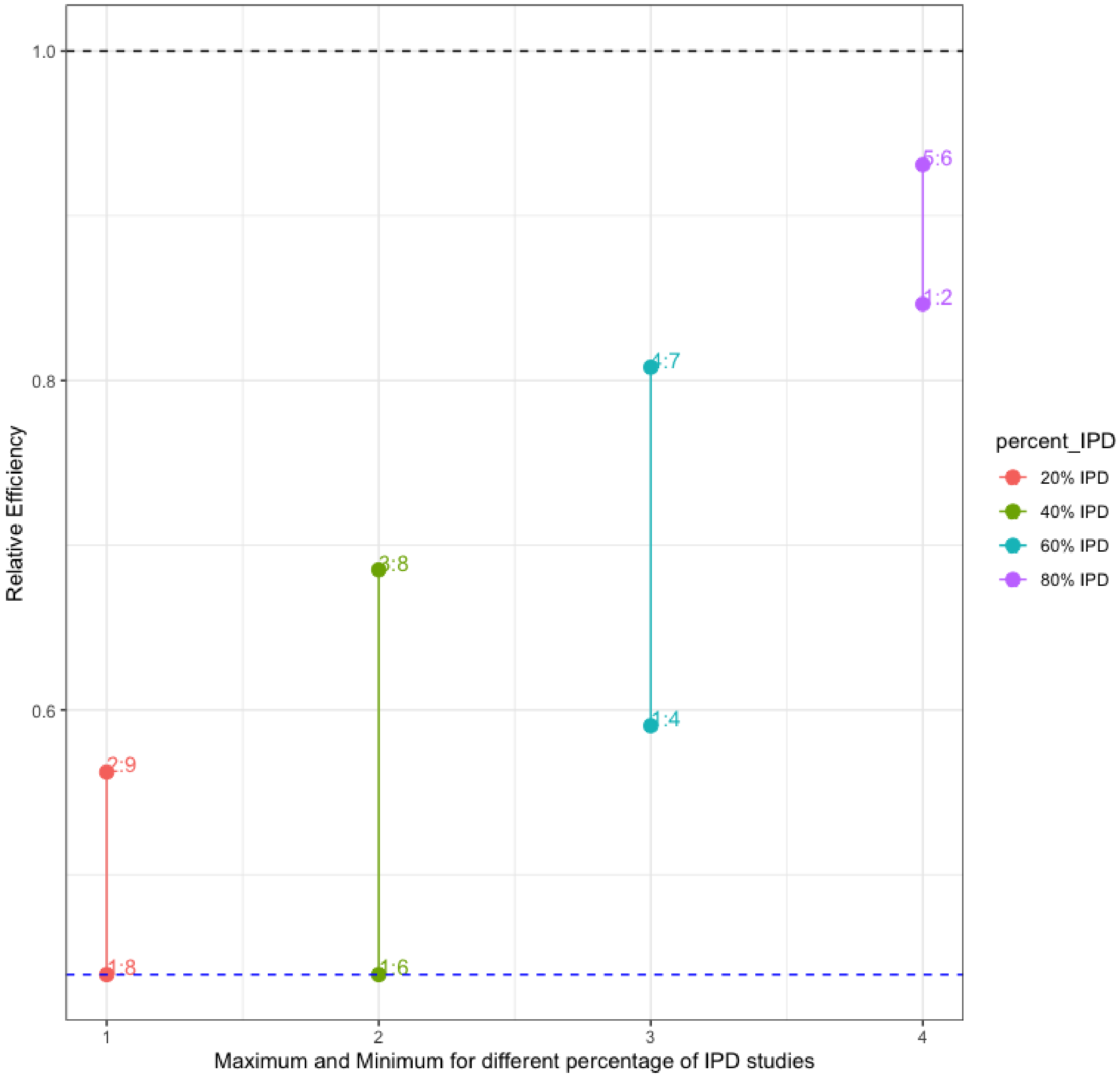

Figure 4. Plot showing the maximum and minimum relative efficiency and the MA combination within each percentage of IPD and MA studies.