World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:8, No:8, 2014

On Pooling Different Levels of Data in Estimating Parameters of Continuous Meta-Analysis

N. R. N. Idris, S. Baharom

***Abstract*—**A meta-analysis may be performed using aggregate data (AD) or an individual patient data (IPD). In practice, studies may be available at both IPD and AD level. In this situation, both the IPD and AD should be utilised in order to maximize the available information. Statistical advantages of combining the studies from different level have not been fully explored. This study aims to quantify the statistical benefits of including available IPD when conducting a conventional summary-level meta-analysis. Simulated meta-analysis were used to assess the influence of the levels of data on overall meta-analysis estimates based on IPD-only, AD-only and the combination of IPD and AD (mixed data, MD), under different study scenario. The percentage relative bias (PRB), root mean-square-error (RMSE) and coverage probability were used to assess the efficiency of the overall estimates. The results demonstrate that available IPD should always be included in a conventional meta-analysis using summary level data as they would significantly increased the accuracy of the estimates. On the other hand, if more than 80% of the available data are at IPD level, including the AD does not provide significant differences in terms of accuracy of the estimates. Additionally, combining the IPD and AD has moderating effects on the biasness of the estimates of the treatment effects as the IPD tends to overestimate the treatment effects, while the AD has the tendency to produce underestimated effect estimates. These results may provide some guide in deciding if significant benefit is gained by pooling the two levels of data when conducting meta-analysis.

***Keywords*—**Aggregate data, combined-level data, Individual patient data, meta analysis.

I. Introduction

META-ANALYSIS is a statistical technique for integrating quantitative results from several sources and thus provides results based on the whole body of research. [1]. A traditional meta-analysis involves integration of aggregate data (AD) which is extracted from the individual study publications. Typical AD includes a mean difference for continuous outcomes or the number of events and participants for binary outcomes. The overall treatment effect is computed by taking the weighted average of the effects across the trials using methods such as the inverse variance method [2] or the Mantel-Haenszel method [3] for the binary data. Alternatively, meta-analysis may be performed based on individual patient data (IPD), where raw data from individual study is obtained and synthesized directly. Although it has numerous advantages compared to the traditional meta-analysis, particularly in terms of type of analyses that can be done, IPD meta-analysis is usually relatively costly and time consuming [4], [5]. Another potential problem for IPD analysis is that IPD are seldom or may not be available from all the individual studies.

Combining the available IPD with the AD has been advocated [6], [7] in order to maximized the available information and reduce the potential bias. Additionally, combined data allows larger number of patients and greater part of the evidence-based to be included. Currently, combining the IPD and AD in meta-analysis has been increasingly common [8]. A review of 199 meta-analyses [9] which has both the IPD and AD available, found 33 combined the data in their analysis and 166 did not. The review noted that the articles that combined the IPD and the AD in their studies had, on average, IPD available in 64% of the studies, while articles which did not combine the studies had an average of 90% IPD available. As there has been no assessment on the statistical value of combining the AD and IPD in a meta analysis, the implications of these ratios on the overall meta-analysis estimates are yet to be explored. Another study [10] examined the models used for combining the IPD and AD in meta analysis of continuous outcomes. In part of their study, estimates based on IPD only and the combinations of both IPD and AD were compared using real data from a hypertension study [11]. The results showed that two-staged models provide comparable estimates to the one-staged model.

This study extends part of the work by Riley et al [10]. We use simulated meta-analysis to empirically assess and compare statistical properties of estimates based on AD-only, IPD-only and the combination of both IPD and AD (which will be referred to as the mixed data, MD). The study aims to quantify the statistical value of including available IPD into conventional aggregate level meta-analysis. Additionally, for the combined data, we examined the influence of the ratio of AD:IPD on the accuracy and precision of the overall treatment effects estimates. The following scenario for data characteristics are considered; (1) five level of ratio of available IPD and AD (2) four levels of the number of studies included in the meta-analysis, N and (3) two levels of the average size of the studies, n. Each scenario was replicated 500 times, and the estimates of the treatment effects and SEs were averaged across the number of simulations. The statistical properties of interest are the percentage relative bias (PRB), the root mean square error (RMSE) and the coverage probability.

Idris NRN is with the International Islamic University Malaysia, Department of Computational and Theoretical Sciences, Faculty of Science, Kuantan Campus, 25200 Pahang Malaysia. (phone: 00609-6196 4000 ; e-mail: ruzni@iium.edu.my).

Baharom S is attached to the same institution, under the Centre of Foundation Studies (e-mail: suhaila_b@iium.edu.my).

II. Material and Method

A. Generation of Data

This study uses simulation approach to generate continuous response data comprising individual patient level treatment effects using R software. y_{ij} which denote patient j from study i were simulated using the following random effects model

$$y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \epsilon_{ij} \quad (1)$$

where β_{0i} is the random study effect, t_{ij} represents the dummy covariate for treatment which takes two values, namely 0 for the control and 1 for the treatment arm, β_{1i} is the random treatment effect, and ϵ_{ij} are the random error terms. We assume that β_{0i}, β_{1i} and ϵ_{ij} are independent and normally distributed, with $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, $\beta_{0i} \sim N(\beta_0, \sigma_{\nu 0}^2)$, and $\beta_{1i} \sim N(\beta_1, \tau^2)$. For simplicity, each study is assumed to have an equal number of patients in each treatment arm, i.e. $n_{0i} = n_{1i} = \frac{1}{2} n_i$ for $i = 1, 2, \ldots, N$. The following values are assigned to the fixed and random effects; $\beta_0 = 0$,$\beta_1 = 3$ for the fixed effects and $\sigma_\epsilon^2 = 1$, $\sigma_{\nu 0}^2 = 1$ and $\tau^2 = 2$. The quantities that are varied are the number of studies included in meta analysis, (N = 10, 20, 30, 90) and the size of the samples (n = 60, 200). These specifications generated 8 meta-analysis with different combinations of N and n. To create aggregate level data, the IPD were summarized by taking the differences of the means of each treatment arms in each individual study. For each meta-analysis of different combination of N and n , 5 ratios of AD: IPD were be created, as follows (1) 0:100, (i.e. an IPD met-analysis) (2) 20:80 (3) 60:40 (4) 80:20 (5) 100:0 (i.e. AD meta-analysis). The AD and IPD were combined using the conventional two-staged method. These scenarios generated 40 meta-analysis, each of which were replicated 500 times. The average treatment effects and their corresponding SE were computed for estimation of the PRB, RMSE and coverage probability.

B. A Two-Stage Method

The AD and IPD were combined using the conventional two-staged method. The two-stage method was the most common method for combining the AD and IPD in practice [9]. A recent study found that of 33 meta-analyses which combine the AD and IPD reviewed, 27 had used this method [9]. The method is appropriate for this study as the subject of interest is the overall pool (treatment) effect where the covariates are identical for each patient in the study. Using this method, the estimates of $\hat{\theta}_i$ and $v(\hat{\theta})_i$ are first obtained for each IPD by fitting (1). These estimates are then combined with those from existing AD using the inverse-variance approach [2].

III. Results

A. All IPD and All AD Studies

Table I gives the estimates of treatment effects and their corresponding SE for all ratio of AD:IPD for studies with average size of n = 60, and for selected values of N (N = 10, 20, 30, 60). The PRB and RMSE for AD and IPD studies are shown in Fig. 1. These quantities were computed as follows;

$$PRB = \frac{\sum_{i=1}^{k} (\theta - \hat{\theta}_i)}{k} \text{ and } RMSE = \sqrt{\frac{\sum_{i=1}^{k} (\theta - \hat{\theta}_i)^2}{k}},$$

where θ is the true effect size, $\hat{\theta}_i$ is the study specific effect estimate and k is the number of simulations. IPD provides up 3 times smaller PRB compared to AD. Similarly the RMSE is much smaller in IPD. The number of studies included in meta-analysis, N, has little effect on the PRB and the RMSE, particularly for the AD. The majority of the estimates based on IPD were overestimated as evidenced by the negative bias, while AD produces underestimated effects.

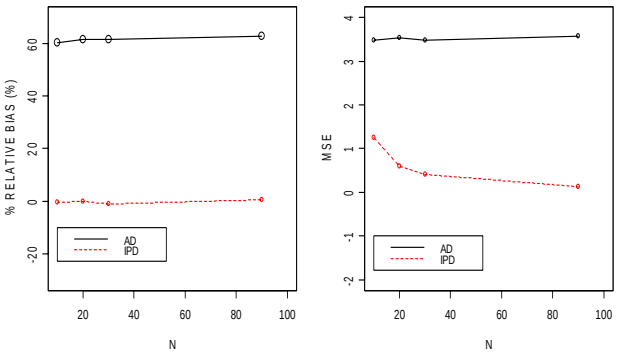

Fig. 1 PRB and MSE for the IPD and AD

B. Combined IPD and AD Studies

1. Percentage Relative Bias

In all three combinations of AD: IPD, the largest PRB occurred when AD-only is used (from 60% - 63%) (Fig. 2). The MD showed relatively smaller bias (from 13% to 54%), and increases with increasing ratio of AD. The trend showed that the bias in MD is larger when majority of the studies comprised of AD level. IPD-only studies seemed to be the most accurate with PRB ranges from -0.3% to 0.7% for the assigned values of N under consideration. The majority of the biases from the IPD-only studies are small negative values suggesting slight overestimation, while those from AD-only studies were large positive bias, suggesting grossly underestimated effects. Estimates of treatment effects from the MD studies were also underestimated, although less severe compared to those from AD.

2. Root Mean-Squared-Error

Similar trends were observed for the RMSE (Fig. 3). Although the SE of estimates from the AD-only data are smaller, their RMSE are consistently larger compared to the MD. The MD is expected to produce larger SE (Table I) as combining the two levels of data will generally induced larger variations in the effects size.

TABLE I
ESTIMATES OF TREATMENT EFFECTS AND THE CORRESPONDING SE $(\theta = 3)$

Ratio AD:IPD	N	10	20	30	90
0:100	IPD-only	3.01 (0.146)	3.02 (0.105)	3.06 (0.086)	2.99 (0.050)
	AD-only	-	-	-	-
	Mixed data	-	-	-	-
20:80	IPD-only	3.01 (0.163)	3.02 (0.117)	3.06 (0.097)	2.99 (0.069)
	AD-only	1.19 (0.205)	1.16 (0.141)	1.15 (0.114)	1.12 (0.081)
	Mixed data	2.60 (0.318)	2.60 (0.223)	2.63 (0.181)	2.56 (0.129)
60:40	IPD-only	3.01 (0.230)	3.03 (0.165)	3.06 (0.136)	2.99 (0.097)
	AD-only	1.19 (0.116)	1.16 (0.081)	1.16 (0.066)	1.12 (0.046)
	Mixed data	1.84 (0.264)	1.81 (0.180)	1.82 (0.144)	1.76 (0.102)
80:20	IPD-only	3.01 (0.325)	3.02 (0.234)	3.05 (0.193)	2.99 (0.138)
	AD-only	1.19 (0.100)	1.16 (0.070)	1.16 (0.057)	1.121 0.040)
	Mixed data	1.47 (0.190)	1.44 (0.127)	1.43 (0.102)	1.38 (0.071)
100:0	IPD-only	-	-	-	-
	AD-only	1.19 (0.090)	1.16 (0.062)	1.16 (0.051)	1.12 (0.029)
	Mixed data	-	-	-	-

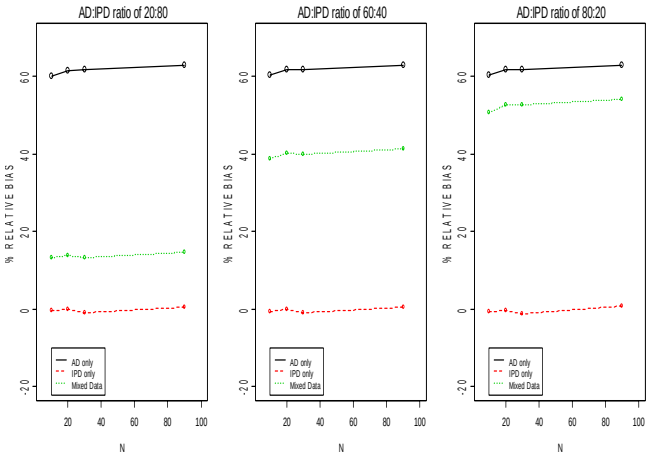

Fig. 2 The PRB for combined level data with different composition of AD:IPD ratio

3. Coverage Probability

The coverage probability was estimated by taking the proportion of times the estimated 95% confidence interval included the true value of θ over the number of simulations. The results showed poor coverage probability for AD-only studies. This is expected as the AD-only studies tend to grossly underestimate the parameters (PRB: 60% to 63%). Combining the AD with the available IPD has moderating effects on the biasness of the estimates and thus increases the coverage probability. The coverage for MD is at best when 80% of the data are IPD and 20% AD and the coverage is close to zero when the ratio is reversed

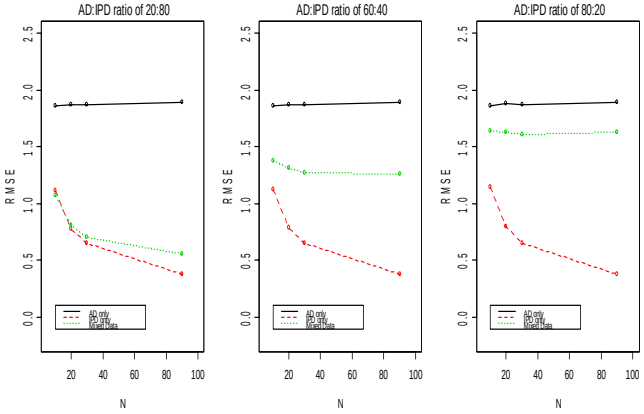

Fig. 3 The RMSE for combined level data with selected AD:IPD ratio

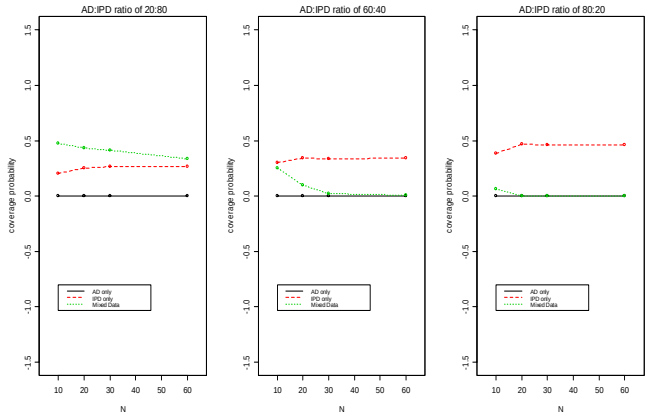

Fig 4 The coverage probability for combined level studies

IV. CONCLUSION

The results of this study confirm earlier findings which suggest that IPD should always be utilized over AD, whenever both types of data are available. Our simulation study showed that IPD produces more accurate results, up to 3 times less bias than those produced by AD. The coverage, for the values assigned in this study, is very low in most of the cases for AD data because the effect estimates were grossly underestimated in this data.

The results also suggest significant advantages in combining the available AD and IPD under certain scenarios. The findings show that if more than 80% of the available data are at IPD level, including the AD does not provide significant differences in terms of accuracy. Inclusion of AD in this case, would increase the SE, resulting in higher RMSE. On the other hand IPD should always be included if majority of the data are at AD level. As expected, the SE based on AD-only data are relatively smaller compared those from the other two types of data. This is expected as average values tend to be more centralized and less variable [12]. Despite larger SE, MD consistently produces smaller RMSE compared to AD-only studies, suggesting better overall estimates and confirming significant advantages in utilizing the available IPD when conducting summary level meta-analysis. Additional benefit of combining the two levels of data is noted when our

simulation study reveals the mixed data adjusts for the overestimation in IPD-only and as well as the underestimation in AD-only data to produce better coverage, performing at its optimum at 20:80 ratio.

Researchers may alternatively view the differences in estimates produced by the three levels of data as a useful sensitivity analysis to gauge the robustness of meta-analytic results to the level of data utilized. The results suggest that available IPD should be routinely included in summary level meta-analysis whenever it is available. It may serve as guidance for practitioners to gauge the efficacy of their estimates given particular level and characteristics of the data used.

ACKNOWLEDGMENT

This work was supported by the Fundamental Research Grant, Ministry of Higher Education, Government of Malaysia. (Grant No: EDW B-11-025-0503).

REFERENCES

[1] Whitehead A (2002). Meta-analysis of controlled clinical trials. London:John Wiley . 215 - 237.
[2] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986, 7, 177-188.
[3] Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst, 1959, 22, 719-748.
[4] Stewart LA, Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. Eval Health Prof, 2002, 25, 76-97.
[5] Simmonds MC, Higgins JPT, Stewart LA, Tierney JF, Clarke MJ, Thompson SG. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. Clin Trials 2005, 2, 209-217.
[6] Jones AP, Riley RD, Williamson PR, Whitehead A. Meta-analysis of individual patient data versus aggregate data from longitudinal clinical trials, Clin Trials., 2009, 6(1), 16-27.
[7] Cooper H, Patall EA. The relative benefits of meta-analysis conducted with individual participant data versus aggregated data, Psychol Methods. 2009, Jun, 14(2), 165-76.
[8] Lambert PC, Sutton AJ, Abrams KR, Jones RD. A comparison of summary patient-level covariates in metaregression with individual patient data meta-analysis, Journal of Clinical Epidemiology, 2001, 55, 86–94.
[9] Riley RD, Simmond MC, Look MP. Evidence synthesis combining individual patient data and aggregate data:a systematic review identified current practice and possible methods. Journal of Clinical Epidemiology, 2007, 60, 431-439.
[10] Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L and Boutitie F, Meta-analysis of continuous outcomes combining individual patient data and aggregate data, Statist. Med., 2008 27, 1870–1893.
[11] Wang JG, Staessen JA, Franklin SS, Fagard R, Gueyffier F. Systolic and diastolic blood pressure lowering as determinants of cardiovascular Hypertension, 2005, 45, 907–913.
[12] Idris, NRN., Robertson C, (2009). The effects of imputing the missing standard deviations on the standard error of the meta analysis estimates. Comm in stats – Sim and Comp ; 38:513-526.