Discussion on “Testing small study effects in multivariate meta-analysis” by Chuan Hong, Georgia Salanti, Sally Morton, Richard Riley, Haitao Chu, Stephen E. Kimmel and Yong Chen

Hans C. van Houwelingen

Department of Biomedical Data Sciences (retired), Leiden University Medical Center, Leiden, The Netherlands

Correspondence
Hans C. van Houwelingen, Department of Biomedical Data Sciences (retired), Leiden University Medical Center, Leiden, The Netherlands.
Email: hansvanhouwelingen@outlook.com

1 | INTRODUCTION

The paper under discussion (described as “the paper” in this discussion) introduces models for the effect of small studies in multivariate meta analysis. This is done under the working hypothesis that this can be analyzed by studying the effect of the study-specific standard error on the outcomes of interest. That is fine with me. However, I am concerned about the way the data are modeled in the paper and the lack of transparency of the way the data are analyzed. In this discussion, I propose a modification of the model and a simplification of the analysis. I present a more detailed analysis of the two case studies. The P-values obtained in that analysis are close to the ones obtained in the paper. Finally, a sharper correction for multiple testing is presented and the lack of interpretation of “just P-values” is commented.

2 | THE MODEL

2.1 | Standardized effect size

The authors use the standardized effect size \( \text{SND} = Y / \text{sd}(Y) \), where \( Y \) is the outcome of interest and \( \text{sd}(Y) \) is its standard deviation. The concept comes from the social sciences and is used to correct for differences in scale between studies. Outcomes in different studies are all scaled to standard deviation = 1. The effect of an intervention is then given by the differences between the means of the treated group and untreated group. This is intrinsically different from the standardization used here. I would prefer not to apply the standardization and to use the more direct model of the outcome \( Y \). The univariate models used in the paper are the meta-regression models (Thompson and Higgins, 2002; Van Houwelingen et al., 2002).

Egger:

\[ Y_i \sim N(\alpha + \beta s_i, \tau^2 + s^2_i) \]  \hspace{1cm} (1)

Lin-Chu:

\[ Y_i \sim N\left(\alpha + \beta \sqrt{\tau^2 + s^2_i}, \tau^2 + s^2_i\right) \]  \hspace{1cm} (2)

Presenting the models this way does not really change them, but it helps to see them in the broader perspective of the meta-regression models. If the relation between \( \sqrt{\tau^2 + s^2_i} \) and \( s_i \) is approximately linear, the regression parts of the models are even approximately equivalent, but the variances are different if \( \tau^2 > 0 \). The confusing complication might be that the \( \beta \) in (2) pops up as \( \alpha \) in Equation (4) of the paper. So, there is no discussion about what to test, but a difference of opinion about the way of presenting the data.
2.2 Comments on the models: alternative model

In his discussion of Lin and Chu (2018), Jackson (2018) showed dislike of the Lin-Chu model. I share his concern. I want to add three points:

- If the model is meant to quantify the effect of selection, $s_i$ can be relevant but $\tau^2$ will not be known at the moment of selection and cannot act as a modifier of the study-specific treatment effect. It can only be used retrospectively and not prospectively.
- If a parameter plays a role in both expected value and variance, the estimate of the expected value can be biased. (van Houwelingen, 1988)
- The Lin-Chu model cannot be analyzed by GEE-type methodology, unless the two $\tau^2$’s are disconnected.

The drawback of the Egger model is that it does not contain a random effect term. A more suitable model would be a hybrid model

**Hybrid model:**

$$Y_i \sim \mathcal{N}(\alpha + \beta s_i, \tau^2 + s_i^2).$$ (3)

This variance term in this model is flexible enough if there is a substantial random effect ($\tau^2 >> 0$), but it is quite rigid if $\tau^2 = 0$. The authors suggest a two-step approach in which $\tau^2$ is estimated from the marginal model $Y_i \sim \mathcal{N}(\alpha, \tau^2 + s_i^2)$. This leads to the hybrid model $Y_i \sim \mathcal{N}(\alpha + \beta s_i, \tau^2 + s_i^2)$. In this model, the variance is completely specified. I would suggest a more flexible variant:

**Flexible Hybrid model:**

$$Y_i \sim \mathcal{N}(\alpha + \beta s_i, \sigma^2(\hat{\tau}^2 + s_i^2)).$$ (4)

The advantage of this model over the Lin-Chu model is that it can be fitted simply by standard models via the transformation

$$Y^*_i = Y_i / \sqrt{(\hat{\tau}^2 + s_i^2)}, \quad x_{0i} = 1 / \sqrt{(\hat{\tau}^2 + s_i^2)}$$
$$x_{1i} = s_i / \sqrt{(\hat{\tau}^2 + s_i^2)}$$ (5)

The transformed model can be written as

$$Y^*_i \sim \mathcal{N}(x_i \theta, \sigma^2)$$ with $x_i = (x_{0i}, x_{1i})$ and $\theta = (\alpha, \beta)^T$ (6)

It looks very much like the SND of the paper, but it is introduced for computational convenience only. The $\sigma$ parameter is missing in the pseudo-likelihood approach of the paper. Moreover model (6) gives more information on the data than the score test for $\beta$ in the paper. Both $\alpha$ and $\beta$ are estimated and the estimate does not depend on $\sigma$. This approach will be explored in Section 5 for the two case studies of the paper.

3 MULTIVARIATE DATA

The interest of the paper is in the situation of multiple outcome denoted by $Y_1, \ldots, Y_J$ with corresponding $\sigma_1, \ldots, \sigma_J$ and the development of a global test for the $\beta$’s, that is a test for $\beta_j = 0$ for $j = 1, \ldots, J$. As in the paper, we concentrate on $J = 2$. The analysis suggested in Section 3 will yield estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ with corresponding (estimated) covariate covariance matrices $C_1 = \text{cov} (\hat{\theta}_1)$ and $C_2 = \text{cov} (\hat{\theta}_2)$. The outcomes of interest need not be present for all studies. Lacking is the cross-covariance matrix $C_{12} = \text{cov} (\hat{\theta}_1, \hat{\theta}_2)$. As in the GEE with a working independence model, that cross-covariance matrix can be estimated from the set $A$ of studies that have both outcomes, as

$$C_{12} = (x_{1i}^T x_{1i})^{-1} \sum_{i \in A} (x_{1i}^T (Y^*_1 - \hat{Y}^*_1)(Y^*_2 - \hat{Y}^*_2)x_{2i}) \times (x_{2i}^T x_{2i})^{-1}.$$ (7)

Here $x_{1i}$ and $x_{2i}$ are the $i$th row-vectors of the $x$-matrices $x_1$ and $x_2$ for outcomes $Y_1$ and $Y_2$, respectively, as defined in (5) and (6). Furthermore, $Y^*_1 = x_1 \hat{\theta}_1$ and $Y^*_2 = x_2 \hat{\theta}_2$. The estimate depends on the covariance of $Y^*_1$ and $Y^*_2$ and the amount of overlap of the two outcomes. From the large covariance matrix $C_0 = \text{cov} (\hat{\theta})$, $C_\beta = \text{cov} (\hat{\theta}_1, \hat{\theta}_2)$ can be extracted and from that a bivariate $\chi^2$-test for $H_0 : \beta_1 = \beta_2 = 0$ ($x_{2i}^T \hat{\beta} C_\beta^{-1} \hat{\beta}$).

If there is little overlap, the test boils down to adding the two univariate $X^2_1$’s. This procedure must have some relation with the one in the paper, but I found that quite hard to read. The supplementary material was not very helpful. It is even not clear how outcomes are handled that are present in all studies.

The approach suggested here can be extended to $J > 2$, but the large covariance matrix for $\hat{\beta}$ might fail to be non-negative definite.

4 CASE STUDIES REVISITED

4.1 Case study 1: Heart failure

The two outcomes considered in the paper are $Y_1 =$ “all-cause mortality” and $Y_2 =$ “Mental quality of Life.” The related standard errors are denoted by $s_1$ and $s_2$. The data are shown in Figure 1 of the paper. The paper only reports the $P$-values of the univariate tests: $P = 0.06$ for $Y_1$ and
Table 1: Basic data for case study 1: \( n \), mean, standard deviation, and correlations

| \( n \) | Mean | SD  | \( Y_1 \) | \( s_1 \) | \( Y_2 \) | \( s_2 \) |
|--------|------|-----|--------|-------|--------|-------|
| \( Y_1 \) | 34   | -0.32 | 0.60   | 1     |
| \( s_1 \) | 0.51 | 0.49 | -0.154 | 1     |
| \( Y_2 \) | 12   | 0.18 | 0.24   | 0.034 | 0.608  | 1     |
| \( s_2 \) | 0.17 | 0.11 | -0.168 | 0.874 | 0.445  | 1     |

Table 2: Regression of \( Y \) on \( s \) for case study 1. \( p^* \) is the \( P \)-value if \( \sigma \) is fixed at \( \sigma = 1 \)

| \( \tau^2 \) | \( \hat{\alpha} \) (se) | \( \hat{\beta} \) (se) | \( \sigma \) | \( \chi^2 \) | df | \( P \) | \( p^* \) |
|--------|----------------|----------------|--------|-------|-----|------|------|
| \( Y_1 \) | 0.72 | 0.89 (0.34) | 1.34 (0.51) | 0.85 | 6.89 | 1 | 0.009 | 0.025 |
| \( Y_2 \) | 0.45 | 0.78 (0.87) | 0.87 (1.43) | 0.96 | 0.37 | 1 | 0.54 | 0.562 |
| both    | (covariance of the \( \hat{\beta} \)’s equals 0.051) | | | | | | 7.07 | 2 | 0.029 |

\( P = 0.09 \) for \( Y_2 \) and the multivariate \( P = 0.03 \). (Unfortunately the \( P \)-values are presented in the wrong order, which caused serious confusion). The data were analyzed with model (4) using the transformation of (5) for ease of computation.

Table 1 shows the basic information of mean, standard deviation, and correlation. The most striking feature is the high correlation between \( s_1 \) and \( s_2 \).

Table 2 shows the results of my analysis sketched above.

The first observation is that despite the different approaches (score test vs Wald-test) the \( P \)-values for \( \sigma = 1 \) agree with the ones given in the paper. Moreover the \( P \)-value for the multivariate test agrees with the one given in the paper. Finally, it also shows the potential of the multivariate approach. The multivariate \( \chi^2 \) is larger than the sum of the two \( \chi^2 \)'s.

4.2 Case study 2: Neuroblastoma

The data are shown in Figure 1 of the paper. As there are much more observations for Disease-Free Survival (DFS) than for Overall Survival (OS), DFS is taken as \( Y_1 \) and overall OS as \( Y_2 \). Table 3 shows again a very high correlation between \( s_1 \) and \( s_2 \) and a bit confusing pattern of correlations. One should realize that there are only 10 studies with OS outcome and only eight studies with both DFS and OLS and all correlations involving \( Y_2 \) or \( s_2 \) are very shaky.

The results in Table 4 are quite similar to those in Table 2. The difference is that the \( \chi^2 \) is smaller than the sum of the two \( \chi^2 \)'s. The role of global testing is discussed in the next section.

The high correlations between \( s_1 \) and \( s_2 \) in both examples suggest a common measure like sample size that could be used for all outcomes.

5 Multiple Testing

Some clarification of the role of global (multivariate) testing could be helpful. The question is what could be said about the univariate tests if the multivariate test is significant at \( \alpha = 0.10 \). There is some interesting theory in Shaffer (1986) and Goeman and Solari (2011). It gives the following improvement over Bonferoni-Holmes:

Table 3: Basic data for case study 2: \( n \), mean, standard deviation, and correlations

| \( n \) | Mean | SD  | \( Y_1 \) | \( s_1 \) | \( Y_2 \) | \( s_2 \) |
|--------|------|-----|--------|-------|--------|-------|
| \( Y_1 \) | 50   | 1.85 | 1.14   | 1     |
| \( s_1 \) | 0.70 | 0.29 | 0.475  | 1     |
| \( Y_2 \) | 10   | 1.36 | 0.97   | 0.639 | -0.614 | 1     |
| \( s_2 \) | 0.63 | 0.31 | -0.451 | 0.972 | 0.483  | 1     |

Table 4: Regression of \( Y \) on \( s \) for case study 1. \( p^* \) is the \( P \)-value if \( \sigma \) is fixed at \( \sigma = 1 \)

| \( \tau^2 \) | \( \hat{\alpha} \) (se) | \( \hat{\beta} \) (se) | \( \sigma \) | \( \chi^2 \) | df | \( P \) | \( p^* \) |
|--------|----------------|----------------|--------|-------|-----|------|------|
| \( Y_1 \) | 0.72 | 0.89 (0.34) | 1.34 (0.51) | 0.85 | 6.89 | 1 | 0.009 | 0.025 |
| \( Y_2 \) | 0.45 | 0.78 (0.87) | 0.87 (1.43) | 0.96 | 0.37 | 1 | 0.54 | 0.562 |
| both    | (covariance of the \( \hat{\beta} \)’s equals 0.051) | | | | | | 7.07 | 2 | 0.029 |
Let the ordered $p$-values of the $J$ outcomes be $p_1, p_2, \ldots, p_J$.

The corresponding univariate null hypotheses can be rejected sequentially as long as $p(j) \leq \alpha / t_j$ with $t = (J-1, J-1, J-2, \ldots, 1)$.

In the examples with $J = 2$, the $t_j$’s are given by $t = (1, 1)$. This implies that the null hypotheses in the examples can be rejected at $\alpha = 0.10$ if the multivariate test is significant.

Actually, I think there is too much emphasis in the paper on $P$-values. That is the weakness of score tests: no effect measures are given, only $P$-values. I strongly suggest to give estimates of the $\beta$ parameters and some help in interpreting them.

**ORCID**

Hans C. van Houwelingen \(\text{https://orcid.org/0000-0002-0494-4637}\)

**REFERENCES**

Goeman, J.J. and Solari, A. (2011) The sequential rejection principle of familywise error control. *Annals of Statistics*, 38, 3782–3810.

Jackson, D. (2018) Discussion on Quantifying publication bias in meta-analysis. *Biometrics*, 74, 795–796.

Lin, L. and Chu, H. (2018) Quantifying publication bias in meta-analysis. *Biometrics*, 74, 785–794.

Shaffer, J.P. (1986) Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, 81, 826–831.

Thompson, S.G. and Higgins, J.P.T. (2002) How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, 21, 1559–1573.

van Houwelingen, J.C. (1988) Use and abuse of variance models in regression. *Biometrics*, 44, 1073–1081.

van Houwelingen, H.C., Arends, L.R. and Stijnen, T. (2002) Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*, 21, 589–624.

**How to cite this article**: van Houwelingen HC. Discussion on “Testing small study effects in multivariate meta-analysis” by Chuan Hong, Georgia Salanti, Sally Morton, Richard Riley, Haitao Chu, Stephen E. Kimmel and Yong Chen. *Biometrics*. 2020;76:1251–1254. \(\text{https://doi.org/10.1111/biom.13345}\)