Random-effects meta-analysis of combined outcomes based on reconstructions of individual patient data

Yue Song[1], **Feng Sun**[2], **Susan Redline**[3], **Rui Wang**[1,4]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health

[2]Department of Epidemiology and Biostatistics, Peking University Health Science Center

[3]Division of Sleep Medicine and Circadian Disorders, Department of Medicine, Brigham and Women's Hospital

[4]Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School

Abstract

Meta-analyses of clinical trials typically focus on one outcome at a time. However, treatment decision-making depends on an overall assessment of outcomes balancing benefit in various domains and potential risks. This calls for meta-analysis methods for combined outcomes that encompass information from different domains. When individual patient data (IPD) are available from all studies, combined outcomes can be calculated for each individual and standard meta-analysis methods would apply. However, IPD are usually difficult to obtain. We propose a method to estimate the overall treatment effect for combined outcomes based on first reconstructing pseudo IPD from available summary statistics and then pooling estimates from multiple reconstructed datasets. We focus on combined outcomes constructed from two continuous original outcomes. The reconstruction step requires the specification of the joint distribution of these two original outcomes, including the correlation which is often unknown. For outcomes that are combined in a linear fashion, misspecifications of this correlation affect efficiency, but not consistency, of the resulting treatment effect estimator. For other combined outcomes, an accurate estimate of the correlation is necessary to ensure the consistency of treatment effect estimates. To this end, we propose several ways to estimate this correlation under different data availability scenarios. We evaluate the performance of the proposed methods through simulation studies and apply these to two examples: (1) a meta-analysis of dipeptidyl peptidase-4 inhibitors versus control on treating type 2 diabetes; (2) a meta-analysis of positive airway pressure therapy versus control on lowering blood pressure among patients with obstructive sleep apnea.

1 Introduction

Meta-analysis is widely used in evidence-based clinical research for intervention evaluations. In clinical practice, treatment decision-making depends on an overall assessment of the patient-level outcomes balancing benefit in various domains and potential risks. Oftentimes a variety of outcome measures are reported in randomized clinical trials (RCTs), which naturally elicits the question of how to synthesize treatment effects across studies on multiple outcomes simultaneously. This motivates us to consider a single measure that is carefully constructed to reflect the effectiveness of an intervention across different outcomes. We henceforth refer to this type of measures as *combined outcomes*, and the reported outcomes that contribute to it as *original outcomes*.

Use of combined outcomes is commonplace in clinical assessment, such as the Apgar Score, [1] the Glasgow Coma Scale,[2] the Elixhauser Comorbidity Index,[3] etc. These three examples are all constructed by summing up the component scores, which is a conventional way of constructing combined outcomes. However, the construction of combined outcomes can take on more complex forms than summation and can pool information from outcomes of different types. For instance, Evans et al[4] proposed a way for the evaluation of antibiotic use strategies that was based on the desirability of outcome ranking (DOOR) as a unified measure of both an overall clinical outcome (an ordinal variable with 5 mutually exclusive levels of desirability based on clinical benefit and adverse effects (AE) ranging from clinical benefit without AE to death) and the duration of antibiotic use (a continuous variable). As the name suggests, the construction of DOOR involves ranking all trial participants with respect to the desirability of their clinical outcomes. When two participants have the same clinical outcome, the one with a shorter duration of antibiotic use receives a higher rank. We differentiate the term *combined outcome* from *composite outcome*. The latter is commonly seen in the survival analysis literature, where it refers to a collection of different outcomes and the interest lies in the time to any one of these outcomes, for example, time to major adverse cardiac events or death.[5] In contrast, a combined outcome is designed to represent a clinically meaningful synthesis of multiple outcomes and can be a function of variables of different types including continuous, binary, or ordinal.

When IPD are available, combined outcomes can be easily obtained for each individual, and standard meta-analysis methods apply. However, in practice IPD are rarely available to meta-analysts. Instead, aggregate data (AD), such as group means and the associated standard errors of the original outcomes are readily accessible through published literature. These AD carry important distributional information about components of combined outcomes. The question now becomes whether one can make inferences about treatment effects on combined outcomes by exploiting AD of the original outcomes with limited or no IPD.

When more than one outcome is of interest, multivariate meta-analysis (MVMA) methods can be employed to jointly estimate treatment effects on several outcomes. MVMA extends the two-stage estimation procedure from univariate meta-analysis, where each study supplies a vector of treatment effect estimates together with an estimated variance-covariance matrix. Based on the hierarchical multivariate normality for study-specific treatment effect estimates on multiple outcomes, parameters of interest can be estimated through maximum likelihood,[6,7] restricted maximum likelihood,[6,8] profile likelihood[7] or Bayesian methods.[9–11] Other estimating procedures based on the method of moments[12–14] and the use of U-statistics[15] have also been developed.

One main advantage of MVMA over separate univariate meta-analyses is the improvement in efficiency resulting from joint modeling multiple outcomes and accounting for their correlations.[16,17] Correctly specifying the correlations facilitates the "borrowing of information" across multiple outcomes and leads to efficiency gain.[16,18] It has been demonstrated through both analytical and simulation studies that ignoring within-study correlations can result in increased mean squared errors and standard errors of the pooled estimates, especially when the between- and within-study covariances are of comparable scale.[18,19] However, within-study correlations are typically assumed known without error in MVMA and concerns have been raised about the unavailability of within-study correlations from study reports.[6,16,18,20] Wei and Higgins[21] derived analytical formulae for estimating within-study correlations between study-specific treatment effect estimates on bivariate outcomes. Their method is able to handle various types of effect measures including the standard difference in means, log odds/risk ratio, etc, but requires knowledge of the correlation between the original outcome measures. Alternative options for MVMA when within-study correlations are unknown[18] include using approximation techniques,[22] obtaining IPD,[23,24] narrowing the range of possible values,[25] conducting sensitivity analyses over the entire parameter space,[26] and employing an alternative model that does not require the specification of within-study correlations.[27,28]

Although MVMA makes use of correlations among multiple outcomes, inference is typically made for each original outcome separately. It is possible to make inference about a smooth function of the treatment effect vector through the use of the Delta method.[29] In some settings MVMA can be used to make inference about combined outcomes. For example, in a bivariate meta-analysis of two continuous outcomes U and V, suppose that treatment effects on both outcomes are estimated using mean differences between intervention groups, and the pooled estimates are $\left(\hat{\Delta}_U, \hat{\Delta}_V\right)$ with an estimated variance-covariance matrix $\widehat{\Sigma}$. Based on the results of MVMA, one can make inference about a combined outcome of the form $aU + bV$ where a, b are constants.[13,16,27] However, settings like this are limited. One key difference is that we consider a function of the original outcomes, while the MVMA methods work with the vector of study-specific treatment effects and the functions thereof. It is not obvious how the existing MVMA methods would handle outcomes that combine information from two original outcomes in a non-linear fashion. Take our data application in Section 4.2 for example, where the original outcomes are systolic and diastolic blood pressures (SBP and DBP). We are interested in estimating the treatment effect on the combined outcome which is a binary indicator of whether a

patient's SBP is over 130 mmHg or DBP is over 80 mmHg, i.e., $\mathbf{1}$(SBP > 130 or DBP > 80), where we use $\mathbf{1}(\cdot)$ to denote the indicator function. This combined outcome signifies whether a patient has elevated blood pressure. While standard MVMA can provide estimates of treatment effects on SBP, DBP, or more generally, a linear combination of SBP and DBP, i.e., aSBP+bDBP where a, b are fixed constants, taking into account the correlation between SBP and DBP, for outcomes such as 1(SBP > 130 or DBP > 80), they no longer apply.

In this paper, we develop meta-analysis methods for combined outcomes that are based on the reconstruction of IPD. If we view the unavailable IPD as missing values, this approach can be considered as an imputation approach. For each set of pseudo IPD, we can calculate the values of combined outcomes and apply standard random-effects methods to obtain the pooled treatment effect estimates. We then follow the multiple imputation approach[30] and take the average of the overall treatment effect estimates from multiple reconstructed IPD to be our final treatment effect estimate.

The proposed approach complements existing MVMA methods by expanding the variety of combined outcomes that could be meta-analyzed based on AD and/or partial IPD. Because once complete (pseudo) IPD have been generated, they can be used to construct any types of combined outcomes. Having IPD also permits the use of one-stage approaches that analyze data from all studies simultaneously under a hierarchical model without first obtaining study-specific estimates.[31]

The idea of reconstructing IPD has been previously explored in other settings. Papadimitropoulou et al[32] found that fitting linear mixed effect models to reconstructed IPD led to treatment effect estimates that were comparable to those obtained through traditional aggregate data meta-analyses. Analysis of reconstructed IPD has also been shown to facilitate the detection of treatment-covariate interactions.[33] In meta-analysis of survival data, an algorithm has been developed to map digitised Kaplan-Meier (KM) curves back to the data that generate these plots using information on numbers of events and numbers at risk. The algorithm demonstrated considerable accuracy in reproducing sample statistics including survival probabilities, medians and hazard ratios.[34]

The remainder of the article is organized as follows. In Section 2, we describe the proposed reconstruction and estimation procedure. Section 3 presents results from our simulation study. In Section 4, we illustrate the proposed method with two data examples: the first is a meta-analysis of dipeptidyl peptidase-4 inhibitors (DPP4-I) versus control in treating patients with type 2 diabetes, where the combined outcome of interest depends on both change in hemoglobin A1c (HbA1c) level and change in body weight; and the second involves a meta-analysis of the effect of positive airway pressure (PAP) in reducing blood pressure among patients with obstructive sleep apnea (OSA). Section 5 offers some discussion and concluding remarks.

2 Methods

2.1 Notation and setting

We consider a meta-analysis of J independent two-arm RCTs with two continuous original outcomes U and V. Let $i = 1$ or 2 denote the treatment or the control arm respectively. We use $j = 1, \cdots, J$ to index studies and $k = 1, \ldots, n_{ij}$ to index patients, where n_{ij} is the number of subjects assigned to arm i of study j. We assume that the studies are randomly sampled from an underlying population where the grand means of responses for individuals in intervention arm i are μ_i^U and μ_i^V respectively. Let $\gamma_{ij}^U, \gamma_{ij}^V$ denote the mean responses for individuals in arm i of study j. We assume that the joint distribution of $\left(\gamma_{ij}^U, \gamma_{ij}^V\right)$ has finite first and second moments that are given below:

$$
\begin{aligned}
E\left[\gamma_{ij}^U\right] &= \mu_i^U, \\
E\left[\gamma_{ij}^V\right] &= \mu_i^V, \\
Var\left(\gamma_{ij}^U\right) &= \left(\tau_i^U\right)^2, \\
Var\left(\gamma_{ij}^V\right) &= \left(\tau_i^V\right)^2, \\
Cov\left(\gamma_{ij}^U, \gamma_{ij}^V\right) &= \kappa \tau_i^U \tau_i^V \, .
\end{aligned}
$$

Following convention, we use $E[X]$ and $Var(X)$ to represent the marginal expectation and variance of a random variable X respectively, and $Cov(X_1, X_2)$ to represent the covariance between random variables X_1 and X_2. We assume a random-effects model that permits heterogeneous treatment effects across studies. Such between-study heterogeneity often exists due to differences in study populations, or in the design and execution of the studies.[35] We further assume that, conditioning on intervention arm i and study membership j, the joint distribution of two original outcomes of the same individual satisfies the following:

$$
\begin{aligned}
E\left[U_{ijk} \mid \gamma_{ij}^U, \gamma_{ij}^V\right] &= \gamma_{ij}^U, \\
E\left[V_{ijk} \mid \gamma_{ij}^U, \gamma_{ij}^V\right] &= \gamma_{ij}^V, \\
Var\left(U_{ijk} \mid \gamma_{ij}^U, \gamma_{ij}^V\right) &= \left(\sigma_{ij}^U\right)^2, \\
Var\left(V_{ijk} \mid \gamma_{ij}^U, \gamma_{ij}^V\right) &= \left(\sigma_{ij}^V\right)^2, \\
Cov\left(U_{ijk}, V_{ijk} \mid \gamma_{ij}^U, \gamma_{ij}^V\right) &= \rho \sigma_{ij}^U \sigma_{ij}^V \, .
\end{aligned}
$$

Lastly we assume that the marginal distribution of each original outcome conditioning on intervention arm i and study membership j, i.e. the distributions of $U_{ijk} \mid \gamma_{ij}^U, \gamma_{ij}^V$ and $V_{ijk} \mid \gamma_{ij}^U, \gamma_{ij}^V$, are functionally independent of ρ. The above hierarchical model allows for both the within-individual correlation (ρ) and the correlation induced by study-level characteristics (κ).

Let $g(\cdot, \cdot)$ be a user-specified known function, and $Y = g(U, V)$ is a combined outcome constructed from the original outcomes U and V. For example, $g(U, V) = U + V$ leads to a continuous outcome Y which is the sum of the original outcomes; and $g(U, V) = \mathbf{1}(U > a, V > b)$ yields a binary outcome Y that indicates whether or not original outcomes U and V are greater than some constants a and b. For simplicity of notation, let $E_j[X]$ denote the

expectation of a random variable X conditional on study membership j (i.e. conditional on $\gamma_{ij}^U, \gamma_{ij}^V$). Let $\theta_j = h\big(E_j[Y_{1jk}]\big) - h\big(E_j[Y_{2jk}]\big)$ denote the treatment effect on the combined outcome for study j, where $h(\cdot)$ is the link function that specifies the metric on which the comparison between two intervention arms is made. For example, when Y_{ijk} is a continuous outcome, we may choose $h(\cdot)$ as the identity function, then θ_j represents the difference in means. When Y_{ijk} is a binary outcome, we may choose $h(\cdot)$ to be the logit function, that is, $h(x) = \log\!\left(\frac{x}{1-x}\right)$, then θ_j represents the log odds ratio.

As in conventional random-effects meta-analyses, we assume that the study-specific effect sizes θ_j, $j = 1, \cdots J$, are a random sample from some underlying distribution $G(\cdot)$. Commonly-used random-effects meta-analysis techniques, for example, the DerSimonian-Laird method,[36] focus on making inference about the mean of this distribution. Similarly, our objective is to estimate the overall effect size defined as:

$$\theta_0 := E_G\big[\theta_j\big].$$

To achieve this goal, conventional two-stage meta-analysis methods require a treatment effect estimate $\hat{\theta}_j$ together with an estimated standard error from each study. These estimates are straightforward to calculate if IPD are available from all studies. Therefore, we focus on the settings where IPD are not available for some studies. We divide participating studies into two categories based on the type of data they contribute. Let $\mathcal{A}$ denote the collection of studies for which only AD are available. That is, for $j \in \mathcal{A}$, only summary statistics of the original outcomes $\left\{\big(n_{ij}, \bar{U}_{ij}, \bar{V}_{ij}, s_{ij}^U, s_{ij}^V\big) : i = 1, 2\right\}$ are available, where

$$\bar{U}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} U_{ijk},$$

$$\bar{V}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} V_{ijk},$$

$$s_{ij}^U = \sqrt{\frac{\sum_{k=1}^{n_{ij}}\big(U_{ijk} - \bar{U}_{ij}\big)^2}{n_{ij}(n_{ij}-1)}},$$

$$s_{ij}^V = \sqrt{\frac{\sum_{k=1}^{n_{ij}}\big(V_{ijk} - \bar{V}_{ij}\big)^2}{n_{ij}(n_{ij}-1)}}.$$

Let $\mathcal{I}$ denote the set of studies with available IPD, that is, for $j \in \mathcal{I}$, all individual-level outcomes $\left\{\big(U_{ijk}, V_{ijk}\big) : k = 1, \ldots, n_{ij}, i = 1, 2\right\}$ are available.

Although our model specification shares similar hierarchical structures as the standard random-effects MVMA model, they differ substantially. We impose our modeling assumptions directly on the original individual outcomes, whereas the MVMA approaches typically postulate modeling assumptions on individual study treatment effect estimates. Furthermore, although both models involve within- and between-study correlations, these parameters refer to different quantities under the two models. For MVMA, the within-study

correlation refers to correlation between treatment effect estimates on different outcomes conditional on study membership, and the between-study correlation refers to the correlation between the true treatment effect sizes on different outcomes across studies (see, for example Jackson et al[12]). Moreover, it is typical in MVMA to assume that both the study-specific treatment effect estimates and the true treatment effect parameters follow a multivariate normal distribution (or are approximately normal).[19] In contrast, in the current setting, the within-study correlation refers to the correlation between original outcomes of the same individual, and the between-study correlation refers to the correlation among the underlying group-specific mean responses. We present this contrast in Table 1 for the setting with two continuous outcomes and where treatment effects are estimated by the differences in sample means.

We make the comparison more explicit by writing out the model assumptions for a standard random-effects bivariate meta-analysis of two continuous outcomes using our notations:

$$
\begin{bmatrix} \bar{U}_{1j} - \bar{U}_{2j} \\ \bar{V}_{1j} - \bar{V}_{2j} \end{bmatrix} \mid \left(\gamma_{1j}^{U}, \gamma_{1j}^{V}, \gamma_{2j}^{U}, \gamma_{2j}^{V} \right) \sim MVN \left(\begin{bmatrix} \gamma_{1j}^{U} - \gamma_{2j}^{U} \\ \gamma_{1j}^{V} - \gamma_{2j}^{V} \end{bmatrix}, \mathbf{D}_{1j} \right),
$$

$$
\begin{bmatrix} \gamma_{1j}^{U} - \gamma_{2j}^{U} \\ \gamma_{1j}^{V} - \gamma_{2j}^{V} \end{bmatrix} \sim MVN \left(\begin{bmatrix} \mu_{1}^{U} - \mu_{2}^{U} \\ \mu_{1}^{V} - \mu_{2}^{V} \end{bmatrix}, \mathbf{D}_{2} \right),
$$

where

$$
\mathbf{D}_{1j} = \begin{bmatrix} \dfrac{\left(\sigma_{1j}^{U}\right)^{2}}{n_{1j}} + \dfrac{\left(\sigma_{2j}^{U}\right)^{2}}{n_{2j}} & \rho_{j}^{*} \sqrt{\dfrac{\left(\sigma_{1j}^{U}\right)^{2}}{n_{1j}} + \dfrac{\left(\sigma_{2j}^{U}\right)^{2}}{n_{2j}}} \sqrt{\dfrac{\left(\sigma_{1j}^{V}\right)^{2}}{n_{1j}} + \dfrac{\left(\sigma_{2j}^{V}\right)^{2}}{n_{2j}}} \\ \rho_{j}^{*} \sqrt{\dfrac{\left(\sigma_{1j}^{U}\right)^{2}}{n_{1j}} + \dfrac{\left(\sigma_{2j}^{U}\right)^{2}}{n_{2j}}} \sqrt{\dfrac{\left(\sigma_{1j}^{V}\right)^{2}}{n_{1j}} + \dfrac{\left(\sigma_{2j}^{V}\right)^{2}}{n_{2j}}} & \dfrac{\left(\sigma_{1j}^{V}\right)^{2}}{n_{1j}} + \dfrac{\left(\sigma_{2j}^{V}\right)^{2}}{n_{2j}} \end{bmatrix},
$$

$$
\rho_{j}^{*} = \frac{\dfrac{\rho \sigma_{1j}^{U} \sigma_{1j}^{V}}{n_{1j}} + \dfrac{\rho \sigma_{2j}^{U} \sigma_{2j}^{V}}{n_{2j}}}{\sqrt{\dfrac{\left(\sigma_{1j}^{U}\right)^{2}}{n_{1j}} + \dfrac{\left(\sigma_{2j}^{U}\right)^{2}}{n_{2j}}} \sqrt{\dfrac{\left(\sigma_{1j}^{V}\right)^{2}}{n_{1j}} + \dfrac{\left(\sigma_{2j}^{V}\right)^{2}}{n_{2j}}}},
$$

$$
\mathbf{D}_{2} = \begin{bmatrix} \left(\tau_{1}^{U}\right)^{2} + \left(\tau_{2}^{U}\right)^{2} & \kappa^{*} \sqrt{\left(\tau_{1}^{U}\right)^{2} + \left(\tau_{2}^{U}\right)^{2}} \sqrt{\left(\tau_{1}^{V}\right)^{2} + \left(\tau_{2}^{V}\right)^{2}} \\ \kappa^{*} \sqrt{\left(\tau_{1}^{U}\right)^{2} + \left(\tau_{2}^{U}\right)^{2}} \sqrt{\left(\tau_{1}^{V}\right)^{2} + \left(\tau_{2}^{V}\right)^{2}} & \left(\tau_{1}^{V}\right)^{2} + \left(\tau_{2}^{V}\right)^{2} \end{bmatrix},
$$

$$
\kappa^{*} = \frac{\kappa \tau_{1}^{U} \tau_{1}^{V} + \kappa \tau_{2}^{U} \tau_{2}^{V}}{\sqrt{\left(\tau_{1}^{U}\right)^{2} + \left(\tau_{2}^{U}\right)^{2}} \sqrt{\left(\tau_{1}^{V}\right)^{2} + \left(\tau_{2}^{V}\right)^{2}}}.
$$

ρ_{j}^{*} and κ^{*} correspond to the within- and between-study correlation in the MVMA setting respectively.

2.2 The proposed method

As mentioned before, when complete IPD are available from all studies, values of combined outcomes can be computed for every individual and standard two-stage meta-analysis or IPD meta-analysis methods apply. In this section, we present our approach based on multiple

reconstructed IPD for studies that provide AD only. Once pseudo IPD for every study in $\mathscr{A}$ are generated, they are combined with the observed IPD from studies in $\mathscr{I}$ to form one set of complete IPD, and standard two-stage meta-analytic methods can be carried out to obtain an overall treatment effect estimate on the combined outcome of interest. In what follows, we first present our method for the reconstruction of IPD, then describe how results based on multiple reconstructed IPD are pooled to arrive at the final treatment effect estimates.

2.2.1   Reconstructions of IPD—Suppose $j \in \mathscr{A}$ then for the k-th participant in arm i of study j, the values of original outcomes are imputed from a bivariate normal distribution with mean and variance set to be the observed sample summary statistics, and correlation set to be $\hat{\rho}$, which is an estimate of the within-individual correlation ρ, i.e.

$$
\begin{bmatrix} U_{ijk}^* \\ V_{ijk}^* \end{bmatrix} \sim MVN \left(\begin{bmatrix} \overline{U}_{ij} \\ \overline{V}_{ij} \end{bmatrix}, n_{ij} \begin{bmatrix} \left(s_{ij}^U\right)^2 & \hat{\rho} s_{ij}^U s_{ij}^V \\ \hat{\rho} s_{ij}^U s_{ij}^V & \left(s_{ij}^V\right)^2 \end{bmatrix} \right).
$$

While summary statistics such as sample average $\left(\overline{U}_{ij}, \overline{V}_{ij}\right)$ and the associated standard errors $\left(s_{ij}^U, s_{ij}^U\right)$ are often reported by individual studies, estimates of the within-study correlation ρ are not commonly provided. The within-study correlation ρ we consider here is the correlation between the original study outcomes. As we clarify in Section 2.1, this ρ is different from the within-study correlation commonly considered in MVMA, which is the correlation between the two treatment effect estimates within each study. Nonetheless, the problem of estimating within-study correlations also arises in MVMA and some of the approaches can be adapted to our setting, such as estimating the within-study correlation from IPD and conducting sensitivity analysis over a plausible range of values.

The impact of $\hat{\rho}$ on the final treatment effect estimator depends on how the original outcomes are combined. The estimation of ρ depends on data availability. In the remainder of this subsection, we will first define linear versus non-linear combined outcomes, contrast the impact of $\hat{\rho}$ on treatment effect estimates for these two categories of outcomes, and then propose estimators of ρ that are suitable under different data availability scenarios.

Linear versus non-linear combined outcomes: We introduce the notion of a *linear* combined outcome: a combined outcome is linear if it can be written in the following form:

$$
Y_{ijk} = g\left(U_{ijk}, V_{ijk}\right) = g_1\left(U_{ijk}\right) + g_2\left(V_{ijk}\right),
$$

where $g_1(\cdot)$, $g_2(\cdot)$ are some user-specified functions. For example, $Y_{ijk} = U_{ijk} + V_{ijk}$ and $Y_{ijk} = \mathbf{1}(U_{ijk} > u) + \mathbf{1}(V_{ijk} > v)$, where u and v are user-specified constants representing some meaningful thresholds, are both linear combined outcomes. Note that linear combined outcomes include, but not limited to, linear combinations of the original outcomes. We term combined outcomes that are not linear as non-linear combined outcomes. For example,

$$
Y_{ijk} = \mathbf{1}\left(U_{ijk} > u, V_{ijk} > v\right).
$$

Suppose Y_{ijk} is linear and we are interested in estimating the parameter $\theta_0 = E_G[\theta_j]$, where

$$
\begin{aligned}
\theta_j &= h\big(E_j[Y_{1jk}]\big) - h\big(E_j[Y_{2jk}]\big) \\
&= h\big(E_j[g_1(U_{1jk})] + E_j[g_2(V_{1jk})]\big) - h\big(E_j[g_1(U_{2jk})] + E_j[g_2(V_{2jk})]\big).
\end{aligned}
$$

Note that θ_j depends on the original outcomes only through their marginal distributions. Thus, by the assumption made in Section 2.1, consistent estimates of θ_0 can be obtained as long as the marginal distributions of the reconstructed outcomes U_{ijk}^* and V_{ijk}^* match those of the original data in each study. Therefore, the specification of $\hat{\rho}$ in the reconstruction step does not affect the consistency of the final estimator. It does, however, affects its efficiency. Below we illustrate this dependence through examining the variance of the final treatment effect estimator under a simple setting.

We consider the setting where all studies provide AD only. We assume hierarchical normality for the original outcomes and let $\big(\sigma_{ij}^U\big)^2 = \big(\sigma_{ij}^V\big)^2 = 1$, $i = 1, 2, j = 1, \ldots, J$ and $n_{1j} = n_{2j} = n_j$ for the ease of notation. We assume that all the variance parameters (between-and within-study) are known. As before, let ρ be the true underlying correlation between the two original outcomes and $\hat{\rho}$ be the value used in IPD reconstruction. Let $\bar{U}_{ij}$, $\bar{V}_{ij}$ represent the sample means of the real (unavailable) IPD and $\bar{U}_{ij}^*$, $\bar{V}_{ij}^*$ are sample means of the reconstructed pseudo IPD. We have

$$
\begin{bmatrix} \bar{U}_{ij} \\ \bar{V}_{ij} \end{bmatrix} \mid \big(\gamma_{ij}^U, \gamma_{ij}^V\big) \sim MVN\left(\begin{bmatrix} \gamma_{ij}^U \\ \gamma_{ij}^V \end{bmatrix}, \begin{bmatrix} \dfrac{1}{n_j} & \dfrac{\rho}{n_j} \\ \dfrac{\rho}{n_j} & \dfrac{1}{n_j} \end{bmatrix}\right),
$$

$$
\begin{bmatrix} \bar{U}_{ij}^* \\ \bar{V}_{ij}^* \end{bmatrix} \mid \big(\bar{U}_{ij}, \bar{V}_{ij}\big) \sim MVN\left(\begin{bmatrix} \bar{U}_{ij} \\ \bar{V}_{ij} \end{bmatrix}, \begin{bmatrix} \dfrac{1}{n_j} & \dfrac{\hat{\rho}}{n_j} \\ \dfrac{\hat{\rho}}{n_j} & \dfrac{1}{n_j} \end{bmatrix}\right).
$$

We consider the simple linear combined outcome $Y_{1,ijk} = U_{ijk} + V_{ijk}$. Let $\theta_j = \big(\gamma_{1j}^U + \gamma_{1j}^V\big) - \big(\gamma_{2j}^U + \gamma_{2j}^V\big)$ and $\theta_0 = E_{G_1}[\theta_j]$, where G_1 denotes the underlying distribution of the study-specific effect size θ_j. Furthermore, let $\hat{\theta}_j = \big(\bar{U}_{1j}^* + \bar{V}_{1j}^*\big) - \big(\bar{U}_{2j}^* + \bar{V}_{2j}^*\big)$ denote the study j's treatment effect estimate on combined outcome Y_1 calculated from a one set of reconstructed IPD, it can be verified that

$$
E[\hat{\theta}_j] = E\Big[E\big[\hat{\theta}_j \mid \gamma_{1j}^U, \gamma_{1j}^V, \gamma_{2j}^U, \gamma_{2j}^V\big]\Big] = E\Big[\big(\gamma_{1j}^U + \gamma_{1j}^V\big) - \big(\gamma_{2j}^U + \gamma_{2j}^V\big)\Big] = E[\theta_j] = \theta_0,
$$

$$
Var\big(\hat{\theta}_j \mid \gamma_{1j}^U, \gamma_{1j}^V, \gamma_{2j}^U, \gamma_{2j}^V\big) = Var\big(\bar{U}_{1j}^* + \bar{V}_{1j}^* \mid \gamma_{1j}^U, \gamma_{1j}^V\big) + Var\big(\bar{U}_{2j}^* + \bar{V}_{2j}^* \mid \gamma_{2j}^U, \gamma_{2j}^V\big) = \frac{4}{n_j}(2 + \rho + \hat{\rho}),
$$

$$
Var\big(\hat{\theta}_j\big) = \frac{4}{n_j}(2 + \rho + \hat{\rho}) + \xi^2, \text{ where } \xi^2 = \big(\tau_1^U\big)^2 + \big(\tau_1^V\big)^2 + \big(\tau_2^U\big)^2 + \big(\tau_2^V\big)^2 + 2\kappa\big(\tau_1^U \tau_1^V + \tau_2^U \tau_2^V\big).
$$

Consider the inverse-variance weighted pooled treatment effect estimate based one reconstruction of IPD:

$$\hat{\theta} = \sum_{j=1}^{J} w_j \hat{\theta}_j, \text{ where } w_j = \frac{1/Var(\hat{\theta}_j)}{\sum_{j'=1}^{J} 1/Var(\hat{\theta}_{j'})}.$$

We have

$$E[\hat{\theta}] = \theta_0, \qquad (1)$$

$$Var(\hat{\theta}) = \frac{1}{\sum_{j=1}^{J} 1/Var(\hat{\theta}_j)} = \frac{1}{\sum_{j=1}^{J} \frac{1}{\frac{4}{n_j}(2 + \rho + \hat{\rho}) + \xi^2}}. \qquad (2)$$

As can be seen from (1) & (2), the expectation of the final treatment effect estimate does not depend on $\hat{\rho}$, the value of within-study correlation used in the reconstruction, whereas its variance does. In contrast, for non-linear combined outcomes, θ_j depends on the joint distribution of U and V and θ_0 varies across different values of ρ. Therefore, accurate characterization of ρ is essential and misspecification of $\hat{\rho}$ could lead to substantial bias. In what follows, we describe the estimation approaches in details when some studies provide IPD, and when all studies provide AD only respectively.

Some studies provide IPD: We first consider settings where IPD are available for at least one of the studies, from which an estimate of the within-individual correlation coefficient ρ can be obtained. There are various methods to estimate ρ using IPD. In our simulation study, we used a weighted average of the sample Pearson correlation coefficients with weights proportional to sample sizes. That is,

$$\hat{\rho}_{IPD} = \frac{1}{\sum_{i=1}^{2} \sum_{j \in \mathscr{I}} n_{ij}} \left(\sum_{i=1}^{2} \sum_{j \in \mathscr{I}} \frac{n_{ij} \sum_{k=1}^{n_{ij}} (U_{ijk} - \overline{U}_{ij})(V_{ijk} - \overline{V}_{ij})}{\sqrt{\sum_{k=1}^{n_{ij}} (U_{ijk} - \overline{U}_{ij})^2} \sqrt{\sum_{k=1}^{n_{ij}} (V_{ijk} - \overline{V}_{ij})^2}} \right). \qquad (3)$$

If one is willing to make fully parametric assumptions on the distribution of (U_{ijk}, V_{ijk}), maximum likelihood estimators or Bayesian estimators are viable options as well.

All studies provide AD only: When none of the studies provides IPD, we can only estimate ρ from AD. The data generating process described in Section 2.1 implies the following properties for the marginal distribution of the observed sample means:

$$E[\overline{U}_{ij}] = \mu_i^U,$$
$$E[\overline{V}_{ij}] = \mu_i^V,$$
$$Var(\overline{U}_{ij}) = \frac{(\sigma_{ij}^U)^2}{n_{ij}} + (\tau_i^U)^2,$$
$$Var(\overline{V}_{ij}) = \frac{(\sigma_{ij}^V)^2}{n_{ij}} + (\tau_i^V)^2,$$
$$Cov(\overline{U}_{ij}, \overline{V}_{ij}) = \frac{\rho \sigma_{ij}^U \sigma_{ij}^V}{n_{ij}} + \kappa \tau_i^U \tau_i^V .$$

(4)

Hence, ρ and κ are not identifiable from AD only. Assuming $\kappa = \kappa_0$, we can derive the following estimator for ρ based on the method of moments:

$$\hat{\rho}_{MM} = \frac{1}{2J} \sum_{i=1}^{2} \sum_{j=1}^{J} \frac{n_{ij}\left[(\overline{U}_{ij} - \hat{\mu}_i^U)(\overline{V}_{ij} - \hat{\mu}_i^V) - \kappa_0 \hat{\tau}_i^U \hat{\tau}_i^V\right]}{\hat{\sigma}_{ij}^U \hat{\sigma}_{ij}^V},$$

(5)

where $\hat{\mu}_i^U$, $\hat{\mu}_i^V$, $\hat{\sigma}_{ij}^U$, $\hat{\sigma}_{ij}^V$ are consistent estimators for μ_i^U, μ_i^V, σ_{ij}^U, σ_{ij}^V. One set of estimators is given below: $\hat{\sigma}_{ij}^U = n_{ij}^{1/2} s_{ij}^U, \hat{\sigma}_{ij}^V = n_{ij}^{1/2} s_{ij}^V$, and $\hat{\mu}_i^U$, $\hat{\mu}_i^V$ are the inverse-variance weighted averages of the observed group means $\overline{U}_{ij}$ and $\overline{V}_{ij}$ respectively. The weights were estimated using s_{ij}^V, s_{ij}^U and the DerSimonian and Laird estimators[36] of $(\tau_i^U)^2$ and $(\tau_i^V)^2$:

$$\hat{\mu}_i^U = \frac{\sum_{j=1}^{J} \hat{w}_{ij}^U \overline{U}_{ij}}{\sum_{j=1}^{J} \hat{w}_{ij}^U},$$

with

$$w_{ij}^U = \left\{(s_{ij}^U)^2 + (\hat{\tau}_i^U)^2\right\}^{-1},$$

$$(\hat{\tau}_i^U)^2 = \max\left\{0, \frac{Q_i^U - (J-1)}{\sum_{j=1}^{J}(s_{ij}^U)^{-2} - \frac{\sum_{j=1}^{J}(s_{ij}^U)^{-4}}{\sum_{j=1}^{J}(s_{ij}^U)^{-2}}}\right\},$$

$$Q_i^U = \sum_{j=1}^{J}(s_{ij}^U)^{-2}(\overline{U}_{ij} - \overline{\overline{U}}_i)^2,$$

$$\overline{\overline{U}}_i = \frac{\sum_{j=1}^{J}(s_{ij}^U)^{-2}\overline{U}_{ij}}{\sum_{j=1}^{J}(s_{ij}^U)^{-2}}.$$

The estimation for $\hat{\mu}_i^V$ can be done similarly. Note that the resulting $\hat{\rho}_{MM}$ is not guaranteed to be bounded between −1 and 1. One option is to truncate it at −1 or 1 accordingly if the estimate is out of range. Alternative truncation points can be established based on prior information. For instance, if two outcomes are believed to be positively correlated, such as

the systolic and diabolic blood pressure, it would be more reasonable to truncate the correlation estimates at 0 and 1.

A likelihood-based estimator can also be derived. Under mild regularity conditions, the sample means are approximately normally distributed when sample sizes are sufficiently large by the central limit theorem. Therefore we can obtain an estimate of the correlation through maximizing the approximate likelihood function $\widetilde{L}$ on the interval $[-1, 1]$

$$\hat{\rho}_{ML} = \underset{\rho \in [-1, 1]}{\arg\max} \widetilde{L}(\rho)$$

$$= \underset{\rho \in [-1, 1]}{\arg\max} \prod_{i=1}^{2} \prod_{j=1}^{J} \frac{1}{\sqrt{1 - \psi_{ij}^2}} \exp\left(-\frac{1}{2(1 - \psi_{ij}^2)}\left[\frac{\left(\overline{U}_{ij} - \hat{\mu}_i^U\right)^2}{\frac{\left(\hat{\sigma}_{ij}^U\right)^2}{n_{ij}} + \left(\hat{\tau}_i^U\right)^2} + \frac{\left(\overline{V}_{ij} - \hat{\mu}_i^V\right)^2}{\frac{\left(\hat{\sigma}_{ij}^V\right)^2}{n_{ij}} + \left(\hat{\tau}_i^V\right)^2} - \frac{2\psi_{ij}\left(\overline{U}_{ij} - \hat{\mu}_i^U\right)\left(\overline{V}_{ij} - \hat{\mu}_i^V\right)}{\sqrt{\left[\frac{\left(\hat{\sigma}_{ij}^U\right)^2}{n_{ij}} + \left(\hat{\tau}_i^U\right)^2\right]\left[\frac{\left(\hat{\sigma}_{ij}^V\right)^2}{n_{ij}} + \left(\hat{\tau}_i^V\right)^2\right]}}\right]\right),$$

with

$$\psi_{ij} = \frac{\frac{\rho\hat{\sigma}_{ij}^U\hat{\sigma}_{ij}^V}{n_{ij}} + \kappa_0\hat{\tau}_i^U\hat{\tau}_i^V}{\sqrt{\left[\frac{\left(\hat{\sigma}_{ij}^U\right)^2}{n_{ij}} + \left(\hat{\tau}_i^U\right)^2\right]\left[\frac{\left(\hat{\sigma}_{ij}^V\right)^2}{n_{ij}} + \left(\hat{\tau}_i^V\right)^2\right]}}.$$

The maximize $\hat{\rho}_{ML}$ is also the posterior mode under a uniform prior distribution of ρ over the interval $[-1, 1]$. Alternative choices of priors can be adopted to incorporate prior beliefs about the correlation (see, for example Bujkiewicz et al[11]).

The value of κ is required for the calculation of $\hat{\rho}_{MM}$ or $\hat{\rho}_{ML}$, which is often unavailable in practice. In MVMA, the between-study correlation, as a parameter in the variance-covariance matrix for study-specific treatment effects, can be estimated using various methods including the maximum likelihood,[6,7] the restricted maximum likelihood[6,8] and the method of moments,[12] provided with knowledge of the within-study covariance matrices. In a Bayesian setting, Bujkiewicz et al[11] used an informative prior distribution of the between-study correlation based on a separate collection of studies. As mentioned in Section 2.1, the κ in our setting is different from that in the MVMA setting. For the IPD reconstruction, we only require information on ρ. Knowledge of κ, however, facilitates the estimation of ρ. When there is no external information about κ, applying the method of moments to the study-level summaries ignoring the within-study sampling variability provides a crude estimator for κ, which can be subsequently used to obtain $\hat{\rho}$, especially when the individual study sizes are large. Lastly, because the value of κ affects the final results only through the value of $\hat{\rho}$ used in the reconstruction, we recommend conducting sensitivity analyses for a range of plausible values of $\hat{\rho}$ when a precise estimate of ρ is unavailable.

2.2.2 Obtain final estimates from multiple reconstructed IPD—The

reconstruction procedure is repeated M times. Let $\left(\hat{\theta}_l, V_l\right)$, $l = 1, \ldots, M$ denote the point and

variance estimates obtained from the l-th reconstructed dataset. To combine estimates from multiple reconstructed IPD, we follow the Rubin's rule for combining estimates from multiple imputations to obtain the final pooled point estimate of treatment effect on the combined outcome[30]:

$$\hat{\theta}_{MR} = \frac{1}{M} \sum_{l=1}^{M} \hat{\theta}_l.$$

In the context of multiple imputation, Rubin[30] also proposed the following formula for obtaining the standard error estimates for the pooled estimator across multiple imputed datasets:

$$\widehat{\text{s.e.}}(\hat{\theta}_{MR}) = \sqrt{\frac{1}{M} \sum_{l=1}^{M} V_l + \left(\frac{M+1}{M}\right) \frac{\sum_{l=1}^{M} (\hat{\theta}_l - \hat{\theta}_{MR})^2}{M-1}}. \tag{6}$$

A direct naive application of this estimator does not perform well in our setting for a few reasons. First, when $\hat{\rho}$ is estimated true from the available IPD or AD, $\widehat{s.e.}(\hat{\theta}_{MR})$ is likely to un-derestimate the sampling variation because it does not take into account the uncertainty in estimating $\hat{\rho}$. Second, the accuracy of $\widehat{s.e.}(\hat{\theta}_{MR})$ depends on the variance V_l of $\widehat{\theta}_l$ obtained from each pseudo IPD. The accuracy of V_l is sensitive to mis-specifications of $\hat{\rho}$ used in the reconstruction.

We therefore recommend the use of a bootstrap procedure to estimate the standard errors.[37] Since the observed data $\mathscr{D}$ consists of two parts, IPD: $\{(U_{ijk}, V_{ijk}): k = 1, ..., n_{ij}, j \in \mathscr{I}, i = 1, 2\}$, and AD: $\{(n_{ij}, \overline{U}_{ij}, \overline{V}_{ij}, s_{ij}^U, s_{ij}^V): j \in \mathscr{A}, i = 1, 2\}$, we generate bootstrap replicates of each part separately. We first apply non-parametric bootstrap to generate replicates of the observed IPD. The same technique has been used by Daniels and Hughes[38] to estimate the correlation between treatment effects on a clinical outcome and its potential surrogate marker. We then generate replicates of AD by applying non-parametric bootstrap to the study summary statistics, where we treat $(n_{ij}, \overline{U}_{ij}, \overline{V}_{ij}, s_{ij}^U, s_{ij}^V), j \in \mathscr{A}$ as a random sample from some underlying joint distribution. The specific steps for calculating the bootstrap standard errors are given below:

Step 1: Generate a bootstrap replicate of IPD. For each $j \in \mathscr{I}$, $i \in \{1, 2\}$, take a sample with replacement of size n_{ij} from $\{(U_{ijk}, V_{ijk}): k = 1, ..., n_{ij}\}$ to obtain $\{(U_{ijk}^\dagger, V_{ijk}^\dagger): k = 1, ..., n_{ij}\}$;

Step 2: Generate a bootstrap replicate of AD. Sample with replacement from $\mathscr{A}$ to get a new set of study labels $\mathscr{L} = \{j_1, ..., j_{|\mathscr{A}|}\}$, then form the bootstrap replicate of AD correspondingly, i.e. $\{(n_{ij_r}, \overline{U}_{ij_r}^\dagger, \overline{V}_{ij_r}^\dagger, s_{ij_r}^{U\dagger}, s_{ij_r}^{V\dagger}): j_r \in \{j_1, ..., j_{|\mathscr{A}|}\}, i = 1, 2\}$;

Step 3: Repeat the proposed reconstruction and estimation procedures on the

bootstrap replicate $\mathscr{D}^{\dagger} = \left\{ \left(U_{ijk}^{\dagger}, V_{ijk}^{\dagger} \right) : k = 1, \ldots, n_{ij}, j \in \mathscr{I}, i = 1, 2 \right\}$, and obtain a

$\cup \left\{ \left(n_{ij_r}, \overline{U}_{ij_r}^{\dagger}, \overline{V}_{ij_r}^{\dagger}, s_{ij_r}^{U\,\dagger}, s_{ij_r}^{V\,\dagger} \right) : j_r \in \left\{ j_1, \ldots, j_{|\mathscr{A}|} \right\}, i = 1, 2 \right\}$

new treatment effect estimate $\hat{\theta}_{MI}^{b}$;

Step 4: Repeat *Step 1–3 B* times and obtain $\hat{\theta}_{MR}^{b}$, $b = 1, \ldots, B$. Compute the

bootstrap standard error estimate as

$$\widehat{s.e.}_{boot}\left(\hat{\theta}_{MR}\right) = \sqrt{\dfrac{\sum_{b=1}^{B}\left(\hat{\theta}_{MR}^{b} - \dfrac{\sum_{b'=1}^{B}\hat{\theta}_{MR}^{b'}}{B}\right)^{2}}{B-1}} . \tag{7}$$

3 Simulation studies

We evaluated the finite sample performance of the proposed method through simulation studies. Data from *J* two-arm RCTs with two continuous original outcomes *U* and *V* were generated according to the following hierarchical model.

$$\begin{bmatrix} \gamma_{ij}^{U} \\ \gamma_{ij}^{V} \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mu_{i}^{U} \\ \mu_{i}^{V} \end{bmatrix}, \begin{bmatrix} \left(\tau_{i}^{U}\right)^{2} & \kappa\tau_{i}^{U}\tau_{i}^{V} \\ \kappa\tau_{i}^{U}\tau_{i}^{V} & \left(\tau_{i}^{V}\right)^{2} \end{bmatrix} \right),$$

$$\begin{bmatrix} U_{ijk} \\ V_{ijk} \end{bmatrix} \mid \left(\gamma_{ij}^{U}, \gamma_{ij}^{V}\right) \sim MVN\left(\begin{bmatrix} \gamma_{ij}^{U} \\ \gamma_{ij}^{V} \end{bmatrix}, \begin{bmatrix} \left(\sigma_{ij}^{U}\right)^{2} & \rho\sigma_{ij}^{U}\sigma_{ij}^{V} \\ \rho\sigma_{ij}^{U}\sigma_{ij}^{V} & \left(\sigma_{ij}^{V}\right)^{2} \end{bmatrix} \right),$$

$$i = 1, 2, \quad j = 1, \ldots, J, \quad k = 1, \ldots, n_{ij}.$$

We set $\mu_{1}^{U} = 2$, $\mu_{1}^{V} = 1$, $\mu_{2}^{U} = \mu_{2}^{V} = 0$, $\sigma_{ij}^{U} = \sigma_{ij}^{V} = 1$, for all *i, j*, and considered a range of values for the within-individual correlation: $\rho \in \{0, 0.2, 0.5, 0.8\}$. The sample size of each study ranged from 20 to 200 in steps of 10 with equal number of patients in the treatment and control arms.

We applied the proposed method to analyze the following four combined outcomes:

$$Y_{1, ijk} = U_{ijk} + V_{ijk},$$
$$Y_{2, ijk} = \mathbf{1}\left(U_{ijk} > 1\right) + \mathbf{1}\left(U_{ijk} > 1.5\right) + \mathbf{1}\left(V_{ijk} < 1\right) + \mathbf{1}\left(V_{ijk} < 2\right),$$
$$Y_{3, ijk} = \mathbf{1}\left(U_{ijk} > 1.5, V_{ijk} < 1\right) + \mathbf{1}\left(U_{ijk} > 1.5, V_{ijk} < 2\right),$$
$$Y_{4, ijk} = \mathbf{1}\left(U_{ijk} > 1.5, V_{ijk} > 0.5\right).$$

Here Y_1 and Y_2 are linear combined outcomes while Y_3 and Y_4 are non-linear combined outcomes. The corresponding parameters of interest are given by:

$$\theta_{10}: = E_{G_1}[\theta_{1j}], \quad \text{with } \theta_{1j}: = E_j[Y_{1,1jk}] - E_j[Y_{1,2jk}],$$
$$\theta_{20}: = E_{G_2}[\theta_{2j}], \quad \text{with } \theta_{2j}: = E_j[Y_{2,1jk}] - E_j[Y_{2,2jk}],$$
$$\theta_{30}: = E_{G_3}[\theta_{3j}], \quad \text{with } \theta_{3j}: = E_j[Y_{3,1jk}] - E_j[Y_{3,2jk}],$$
$$\theta_{40}: = E_{G_4}[\theta_{4j}], \quad \text{with } \theta_{4j}: = \log\left(\frac{E_j[Y_{4,1jk}]\left(1 - E_j[Y_{4,2jk}]\right)}{E_j[Y_{4,2jk}]\left(1 - E_j[Y_{4,1jk}]\right)}\right),$$

where G_l, $l = 1, 2, 3, 4$ are the underlying distributions of θ_{lj}, $l = 1, 2, 3, 4$, respectively. We investigated 20 simulation scenarios with varying parameter values. Results under scenarios 1 to 8 (see Table 2) are presented in this section. Results under scenarios 9 to 20 are included in the web supplementary materials.

3.1 One study provides IPD

We first investigated the case where one of the studies provided complete IPD and the rest provided AD. Data were generated under scenario 1 to 4 (see Table 2) with different configurations of the number of studies, J, and levels of the between-study heterogeneity, $\left(\tau_i^U\right)^2$, $\left(\tau_i^V\right)^2$, $i = 1,2$. The sample size of the study that provides IPD was fixed at 20 and the value of κ was set at 0.5.

We set the number of multiple reconstructions M to be 10 and the number of bootstrap replicates B to be 500, and performed 500 experiments under each parameter configuration. Study-specific treatment effect estimates on combined outcomes (calculated from the observed or reconstructed IPD) were pooled by applying the DerSimonian and Laird random-effects method.

We compared point estimates computed from the multiple reconstructed datasets when $\hat{\rho}$ was estimated with Eq.(3), set to the true value of ρ, or set to 0, against the estimates obtained from applying the DerSimonian and Laird random-effects model to complete IPD of all studies, which corresponded to the ideal situation. The naive standard error estimates were calculated according to Eq.(6) with $\left(\hat{\theta}_l, V_l\right)$ obtained from applying the DerSimonian and Laird random-effects model to the l-th reconstructed dataset. We also calculated the average bootstrap standard error estimates (Eq.(7)).

Figure 1 displays the average estimates of treatment effects for the four outcomes in simulation scenario 2. Results from scenario 1, 3 and 4 were similar and were included in the web supplementary materials. Not surprisingly, for linear outcomes Y_1 and Y_2, the proposed method produced results almost identical to those based on complete IPD regardless of the value of $\hat{\rho}$ used in the reconstruction. For non-linear outcomes Y_3 and Y_4, substantial biases were observed when $\hat{\rho}$ was set to 0 (purple lines) but the true correlation was nonzero. In contrast, the average treatment effect estimates computed from the multiple reconstructed datasets using $\hat{\rho}_{IPD}$ (blue lines) or the true value ρ (green lines) were very close to the ones obtained from complete data (red lines) across all values of ρ, even though the sample size for the study with available IPD was only 20.

Figure 2 compares the average naive and bootstrap standard error estimates to their empirical counterparts, which were calculated as standard deviations of the treatment effect estimates across 500 iterations. For each panel, the tabs on the right-hand side (SCN 1–4) denote simulation scenarios, the upper tabs (0, 0.2, 0.5, 0.8) represent the true values of ρ used in the data generation, and the labels on the bottom x-axis ($\rho, \hat{\rho}_{IPD}, 0$) indicate the values taken by $\hat{\rho}$ in the reconstruction. The naive standard error estimates (blue squares) were conservative under most settings, especially when the between-study variances ($\tau^U)^2$, ($\tau^V)^2$ were small. The bootstrap standard error estimates (red dots) in general performed well, except when the number of studies was very small ($J = 7$) and the between-study variances were relatively large (scenario 2). In such settings, resampling from a small number of studies was unlikely to capture the level of variability in the underlying study population, leading to the observed downward biases in the bootstrap standard error estimates.

3.2   All studies provide AD

We evaluated the performance of our proposed methods when only AD were available for all studies. We simulated data under scenario 5 to 8 (see Table 2 with κ set to be 0). For the $\hat{\rho}$ required in the reconstruction step, we computed both the moment-based estimator ($\hat{\rho}_{MM}$) and the likelihood-based estimator ($\hat{\rho}_{ML}$) using formulae provided in Section 2.2.1 with κ_0 = 0. The truncation points for $\hat{\rho}_{MM}$ were set to be −1 and 1.

Figure 3 presents the treatment effect estimates. Without any IPD, using the true correlation ρ (tan lines) in the reconstruction step yielded results virtually identical to those that would have been obtained if complete IPD were available (red lines) in all settings. Results based on $\hat{\rho}_{MM}$ (the method of moments estimator; green lines) and $\hat{\rho}_{ML}$ (the maximum likelihood estimator; blue lines) were also indistinguishable. For linear combined outcomes Y_1 and Y_2, all lines coincided regardless the values of $\hat{\rho}$ used in the reconstruction. For non-linear combined outcomes Y_3 and Y_4, the results were sensitive to the misspecification of $\hat{\rho}$: substantial biases were associated with setting $\hat{\rho} = 0$ when the true ρ deviated from 0 (purple lines). Using $\hat{\rho}_{MM}$ or $\hat{\rho}_{ML}$ in general performed well except when the number of studies was small ($J = 13$ under scenario 6). When only AD were available, accurate estimation of ρ was difficult when the number of studies was too small or when the between-study heterogeneity was large. For instance, under scenario 6 with $\rho = 0$, $\hat{\rho}_{MM}$ was out of range and truncated at 1 or −1 in more than 40% of the 500 experiments. In the current setting, we set $\kappa_0 = 0$, based on Eq.(5)):

$$\hat{\rho}_{MM} = \frac{1}{2J} \sum_{i=1}^{2} \sum_{j=1}^{J} \frac{n_{ij}\left(\bar{U}_{ij} - \hat{\mu}_i^U\right)\left(\bar{V}_{ij} - \hat{\mu}_i^V\right)}{\hat{\sigma}_{ij}^U \hat{\sigma}_{ij}^V}.$$

Therefore, $Var(\hat{\rho}_{MM})$ depends on the variance of $\bar{U}_{ij}$, $\bar{V}_{ij}$, $\hat{\mu}_i^U$, $\hat{\mu}_i^V$, as well as $\hat{\sigma}_{ij}^U$ and $\hat{\sigma}_{ij}^V$. As can be seen from Eq.(4), when the between-study heterogeneity τ_i^U increases, $Var(\bar{U}_{ij})$

increases, and the variance of $\hat{\mu}_i^U$ also increases. Same holds true for outcome V. Therefore, $\hat{\rho}_{MM}$ can be highly variable when the between-study heterogeneity is large.

Results for standard error estimates are presented in Figure 4. For linear combined outcomes (top panels), standard error estimates based on the bootstrap approach (red dots) closely matched the true empirical standard errors (green dots). For non-linear combined outcomes, the bootstrap performed well in general with modest underestimation when the number of studies was small. The standard error estimates obtained by a naive application of the Rubin's rule (blue dots) were generally biased and were further away from the empirical ones compared the bootstrap standard error estimates. When the number of studies was small, $\hat{\rho}_{MM}$ or $\hat{\rho}_{ML}$ can be highly unstable, and estimators using $\hat{\rho}_{MM}$ or $\hat{\rho}_{ML}$ in the reconstruction step were associated with larger variability compared to using the true correlation ρ or a fixed value, e.g., 0. In such settings, it would be useful to consider sensitivity analyses over a range of plausible values for $\hat{\rho}$. We illustrate this with a real-world data example in Section 4.1.

As discussed in Section 2.2.1, treatment effect estimates for linear combined outcomes remain consistent regardless of the value of $\hat{\rho}$ used in the reconstruction whereas the efficiency can be affected. For outcome Y_1, the variance of the treatment effect estimates increases as $\hat{\rho}$ increases (see Eq.(3))). However, this trend was not observed in Figure 4. Because the final estimate was an average over multiple sets of reconstructed IPD (here $M = 10$), this dependence may be diluted by averaging and became harder to detect. We conducted another set of simulations with the same set-up as scenario 18 (see Table 1 in web supplementary materials) except that we reduced M from 10 to 1. In the reconstruction step, we varied the value of $\hat{\rho}$ from -1 to 1. The results are displayed in Figure 5. Based on a single reconstruction of IPD, the empirical standard errors of treatment effect estimates for Y_1 increased as $\hat{\rho}$ increased, as expected from Eq.(3)). In contrast, the empirical standard errors of treatment effect estimates for Y_2 showed an downward trend as $\hat{\rho}$ increased.

3.3 The proposed approach vs. MVMA

As discussed previously, if the parameter representing the treatment effect for the combined outcomes can be written as a smooth function of treatment effect parameters for the original outcomes, MVMA can be applied to obtain treatment effect estimates on the combined outcome without the reconstruction of IPD. An example is the outcome Y_1. We conducted a simulation study to compare the finite-sample performance of our method versus the standard random-effects MVMA. We applied two estimation approaches for MVMA: the restricted maximum likelihood (REML)[6,8] and the method of moments (MM).[12] The latter is an extension of the DerSimonian and Laird method to the multivariate setting. Data were simulated under the same parameter set-up as in scenario 18. For the ease of comparison, we assumed that within-study correlations were known. That is, following the notations in Section 2.1, we used $\hat{\rho} = \rho$ during the reconstruction of IPD for the proposed approach, and used ρ_j^* for MVMA. For both methods, the within-study variances σ_{ij}^U, σ_{ij}^V were estimated using the sample standard deviations. Since Y_1 is just a linear combination of the two original outcomes, the true overall treatment effect for Y_1 equals $\theta_{10} = 3$. Table 3 and 4

present the estimated biases and the empirical standard errors. Both the proposed method and MVMA produced minimally biased point estimates with similar efficiency.

4 Illustrative examples

4.1 Effect of DPP4-I on HbA1c and weight

A common side effect of anti-diabetic medications such as insulin is weight gain,[39] which is undesirable in most cases as over 90% of patients with type 2 diabetes are already overweight or obese.[40] However, an increasing number of therapeutic options are available that are weight neutral or even lead to weight loss in addition to their glycemic benefits.[39] Hence, a combined outcome that characterizes both change in glycated hemoglobin (HbA1c) and change in body weight would be useful when comparing different interventions.

Wu et al[41] conducted a network meta-analysis of 360 RCTs examining the effects of incretin-based therapies on β-cell function and insulin resistance in type 2 diabetes. From their study pool, we identified 13 RCTs that compared the effect of dipeptidyl peptidase-4 inhibitors (DPP4-I, or treatment) to control and reported both change in HbA1c level (U) and change in body weight (V). Group means, standard errors and sample sizes of participating studies are summarized in Table 5. Univariate meta-analyses using the DerSimonian and Laird random-effects model revealed that patients in the treatment arms experienced a greater decline in HbA1c level by 0.67% (95% CI : [−0.90, −0.44], between-study variance estimate $\hat{\tau}^2 = 0.16$), but a smaller reduction in body weight by 0.36 kg (95% CI : [0.05, 0.67], $\hat{\tau}^2 = 0.11$) compared to those in the control arms. We applied the proposed method to analyzing the following non-linear combined outcome:

$$Y_{ijk} = 2 \times \mathbf{1}(U_{ijk} < 0) + \mathbf{1}(V_{ijk} < 0) + \mathbf{1}(U_{ijk} < 0, V_{ijk} < 0).$$

The combined outcome assigns a score in {0,1,2,4} to each of the four possible combinations of increased/decreased HbA1c level and increased/decreased body weight, with higher score being more desirable. The most desirable response is a decrease in both HbA1c and weight (score = 4), followed by a decrease in HbA1c only (score = 2), a decrease in weight only (score = 1) and an increase in both outcomes (score = 0).

The parameter of interest θ_0 was the difference in means of the combined outcome Y between the DPP4-I and control arm, defined as follows:

$$\theta_j := E_j[Y_{1jk}] - E_j[Y_{2jk}],$$
$$\theta_0 := E_G[\theta_j].$$

Since only AD were available and prior knowledge on κ was lacking, we calculated $\hat{\rho}_{MM}$ and $\hat{\rho}_{ML}$ as described in Section 2.3.2 by taking $\kappa_0 = 0$ and imputed pseudo IPD accordingly. The resulting $\hat{\rho}_{MM}$ and $\hat{\rho}_{ML}$ were both −1. When we plugged in a crude estimate of κ obtained by ignoring the within-study sampling variation ($\hat{\kappa} = -0.756$),

$\hat{\rho}_{ML} = 1$ and $\hat{\rho}_{MM}$ was out of range and close to the boundary value 1. In the current setting with only 13 studies, it was not surprising that the resulting estimates for ρ were highly variable, we therefore conducted sensitivity analyses by plugging in $\hat{\rho}$ ranging from a very strong negative correlation ($\hat{\rho} = -0.8$) to a perfect positive correlation ($\hat{\rho} = 1$). For each value of $\hat{\rho}$, we reconstructed $M = 50$ sets of pseudo IPD. Table 6 shows the treatment effect estimates, the bootstrap standard error estimates, and the 95% confidence intervals. The 95% CIs were constructed assuming a normal distribution for $\hat{\theta}$. The bootstrap standard error estimates are calculated from $B = 1000$ bootstrap replicates. Overall, using different $\hat{\rho}$ yielded similar results with effect sizes ranging from 0.60 to 0.71.

Although univariate analysis suggested that the treatment was superior in terms of lowering patients' HbA1c level but less effective in weight reduction compared to the control intervention, meta-analysis of the HbA1c-weight combined outcome revealed that participants in the treatment arms had better outcomes when both were taken into account simultaneously. We also note that here we gave more weights to HbA1c reduction than to weight reduction in the combined outcome, out of consideration that the primary goal of diabetes treatment was glycemic control. But the choice of weights can be tailored to personal preference. If one is more concerned with weight reduction, then the weights can be adjusted accordingly. The proposed approach provides one way to assess treatment effects across multiple outcomes simultaneously.

4.2 Effect of PAP on systolic and diastolic blood pressure

Obstructive Sleep Apnea (OSA) is a sleep disorder where breathing is briefly and repeatedly interrupted during sleep, resulting in intermittent hypoxemia and sympathetic nervous system activation. It is a highly prevalent chronic condition affecting approximately 27% of men and 9% of women.[42] Its prevalence among individuals with cardiovascular disease (CVD) is even higher.[43,44] Multiple epidemiological studies have demonstrated that OSA is a significant risk factor for CVD, including stroke, coronary artery disease, heart failure, and atrial fibrillation.[45–48] Moreover, it has been shown that OSA is an independent risk factor for the development of hypertension, a well-recognized risk factor for CVD.[49] Over 50% of OSA patients have hypertension and approximately 70% of patients with resistant hypertension have OSA.[43] Therefore, there is compelling evidence to identify the role of OSA treatment as a means for primary or secondary prevention of CVD.

Positive Airway Pressure (PAP) therapy is the standard first-line treatment for OSA. A PAP machine provides pressurized air to the naso-pharyngeal airway, preventing airway collapse that may occur in susceptible individuals with inspiratory breathing efforts, thus stabilizing breathing during sleep. While OSA is recognized as a cause of secondary hypertension, the extent to which PAP improves blood pressure profile has been an area of active investigation. Some research has suggested that OSA differs in its effects on diastolic versus systolic blood pressure,[50] which has implications on potential mechanisms and intervention targets. Given that adverse physiological effects of sleep apnea are most acute during sleep, there is also interest in the effect of treatment on 24-hour blood pressure profile.[51,52] Schein et al[53] performed a systematic review of RCTs that compared active PAP versus placebo PAP, sham PAP or subtherapeutic PAP in lowering blood pressure among patients with OSA.

Six studies (620 patients in total) were identified for a meta-analysis of treatment effects on post-intervention 24-hour systolic blood pressure (SBP) and diastolic blood pressure (DBP). The comparison of PAP to control groups revealed a non-significant reduction in 24-hour SBP (−3.57 mmHg, 95% CI: [−8.58, 1.44]; between-study variance estimate $\hat{\tau}^2 = 22.86$) but a significant reduction in 24-hour DBP (−3.46 mmHg, 95% CI: [−6.75, −0.17]; $\hat{\tau}^2 = 10.64$). We extracted group-level aggregate data of those studies. In addition, we included individual patient data from the Best Apnea Interventions for Research (BestAIR) trial,[54] which led to a total of 7 studies (see Table 7). The BestAIR trial data are available upon reasonable request at https://sleepdata.org. Univariate meta-analyses based on the DL random-effects model from all 7 studies gave conclusions consistent with Schein et al[53]: Compared to control arms, PAP resulted in a significant reduction in 24-hour DBP by 3.14 mmHg (95% CI: [−5.67, −0.61]; $\hat{\tau}^2 = 6.59$), but the effect of PAP on 24-hour SBP did not reach statistical significance (treatment effect −4.01 mmHg, 95% CI: [−8.07, 0.04]; $\hat{\tau}^2 = 16.81$).

Although prior research has considered SBP and DBP separately, hypertensive burden is typically examined by considering levels of both systolic and diastolic pressures. Therefore we considered the following two combined outcomes. The first one was a binary indicator of elevated blood pressure that equals 1 if either the post-intervention 24-hour SBP is greater than 130 mmHg or DBP is greater than 80 mmHg and 0 otherwise; The second one was an ordinal outcome that equals the summation of the two indicators $\mathbf{1}(\text{SBP}_{ijk} > 130)$ and $\mathbf{1}(\text{DBP}_{ijk} > 80)$, i.e.

$$Y_{1,ijk} = \mathbf{1}\big(\text{SBP}_{ijk} > 130 \text{ or } \text{DBP}_{ijk} > 80\big),$$
$$Y_{2,ijk} = \mathbf{1}\big(\text{SBP}_{ijk} > 130\big) + \mathbf{1}\big(\text{DBP}_{ijk} > 80\big).$$

The threshold values were selected based on the guidelines published by American College of Cardiology (ACC) and American Heart Association (AHA) in 2017.[55] The guidelines classify blood pressures into 5 categories that range from normal to severe hypertension, and 130/80 mmHg is the cutoff for normal versus elevated blood pressure. Since the first outcome was binary, we chose the target parameter to be the mean log odds ratio of elevated blood pressure comparing patients treated with PAP versus control:

$$\theta_{1j} := \log\left(\frac{E_j[Y_{1,1jk}]\big(1 - E_j[Y_{1,2jk}]\big)}{E_j[Y_{1,2jk}]\big(1 - E_j[Y_{1,1jk}]\big)}\right),$$
$$\theta_{10} := E_{G_1}[\theta_{1j}].$$

And for the second outcome, the parameter of interest was the difference in means:

$$\theta_{2j} := E_j[Y_{2,1jk}] - E_j[Y_{2,2jk}],$$
$$\theta_{20} := E_{G_2}[\theta_{2j}].$$

Our literature search revealed a positive correlation between SBP and DBP of magnitudes varying from 0.61 to 0.74.[56,57] The available IPD from the BestAIR study yielded $\hat{\rho}_{IPD} = 0.63$, which was consistent with published results. We generated $M = 50$

reconstructed datasets to obtain the treatment effect estimate and $B = 1000$ bootstrap replicates to obtain the bootstrap standard error estimate. For Y_1, the proposed procedure yielded a point estimate of -0.69, corresponding to an odds ratio of 0.50 for post-intervention elevated blood pressure (SBP > 130 mmHg or DBP > 80 mmHg) comparing patients in PAP groups to those in control groups. The protective effect of PAP was significant at 0.05 level, with approximate 95% bootstrap confidence interval [0.29, 0.86]. We also observed a significant treatment effect on combined outcome Y_2, where the mean response in PAP groups was estimated to be 0.26 unit lower compared to that in control groups with approximate 95% bootstrap confidence interval [-0.06, -0.46]. Meta-analyses of both combined outcomes supported the conclusion that PAP can facilitate blood pressure control among OSA patients. While univariate analysis based on DBP only did not reach statistical significance, meta-analysis based on combined outcomes revealed that PAP significantly increased the probability of a reduction in either SBP or DBP.

5 Discussion

In this paper, we develop meta-analysis methods for outcomes combining information across multiple endpoints based on reconstructions of IPD. We focus on combined outcomes that are functions of two continuous outcomes. For linear combined outcomes, treatment effect estimates obtained using the proposed procedure remain consistent regardless of misspecification of the within-study correlation ρ in the reconstruction step. In contrast, for non-linear combined outcomes, accurate estimation of ρ is essential to ensure consistency of the treatment effect estimates. When no previous knowledge of ρ is assumed, a reasonably accurate estimate can oftentimes be obtained from even a small sample of IPD. The estimation problem becomes difficult when only AD are available, in which case one needs to assume a value for the between-study correlation κ in order to estimate ρ and we found through simulations that estimators of ρ that were solely based on AD were highly unstable when the number of studies was small and the between-study heterogeneity was large. In these settings, we recommend conducting sensitivity analyses to examine treatment effect estimates for a plausible range of $\hat{\rho}$.

When there is between-study heterogeneity in treatment effects for the original outcomes, it is reasonable to expect between-study heterogeneity in treatment effects for the combined outcomes. It would be of interest to quantify the magnitude of this between-study heterogeneity. When IPD are available from all studies, we can obtain the IPD for the combined outcomes and the usual measure such as the I^2 statistic[58] can be calculated in a straightforward manner. When IPD are not available from some studies, our approach is based on multiply reconstructing IPD. Although it is straightforward to calculate the I^2 statistic or other measures of heterogeneity for each set of pseudo IPD, pooling the estimates across multiple reconstructed datasets requires careful consideration. For measures that are asymptotically normally distributed, such as the maximum likelihood or restricted maximum likelihood estimators of the between-study variance, we can directly apply Rubin's rule as stated in Section 2.2; For measures that are not asymptotically normally distributed, proper normalization may be used so that a pooled estimate can be computed based on the transformed values and then transformed back to the original scale. Take the I^2 statistic as an example, we may consider the Fisher z transformation to the square root of I^2. The Fisher z

transformation has been suggested for R^2 in linear regression,[59] which is also a measure of proportion of variation. Testing the null hypothesis of homogeneity in such settings is also possible. For example, under the null hypothesis, the Cochran's Q statistics follows a chi-square distribution with $(J-1)$ degrees of freedom[60] (where J is the number of studies), and methods for pooling chi-square distributed test statistics from multiple imputed datasets have been developed by Rubin[30] and Li et al.[61] It would be useful to develop metrics that can sensibly characterize between-study heterogeneity of treatment effects on combined outcomes that are pooled from multiple reconstructed IPD in future research.

Our approach essentially approximates the joint distribution of the original outcomes with a bivariate normal distribution through matching the first and second moments. This is a convenient choice since group means and associated standard errors are the most commonly reported AD, and a normal distribution is completely specified by its first two moments. In addition, it is relatively easy to generate normal samples with specified correlations in standard statistical software. We expect this approximation to work reasonably well for original outcomes with symmetric distributions. When IPD are available for some studies, the reconstruction procedure can be generalized to accommodate skewed distributions through the use of more flexible distributions, for example, the Fleishman distribution, where the first four moments can be matched,[62] or non-parametrically through the use of kernel density estimators.

The proposed method can be generalized to analyzing combined outcomes constructed from more than two original outcomes. It can also be adapted to scenarios where the original outcomes are discrete or a mixture of discrete and continuous outcomes, in which case, estimation of correlation parameter requires consideration of additional constraints; for example, the correlation of two binary variables is bounded above by their marginal success probabilities. Lastly, it would be useful to develop Bayesian meta-analytic methods for combined outcomes through constructing hierarchical models and placing priors on unknown parameters. Pseudo IPD can be reconstructed by sampling from the corresponding posterior distributions.

We illustrated our approach using the DerSimonian-Laird random-effects model as it is one of the most routinely-used methods for random-effects meta-analysis. However, the idea of multiply reconstructing IPD based on reported AD can also be combined with other choices of two-stage meta-analysis methods (for example, see Wang et al[63]) or meta-analysis based on IPD directly to allow for more flexible modeling of the variance structure as well as treatment-covariate interactions. Previous work has demonstrated both feasibility and advantages of fitting regression models to reconstructed IPD when meta-analyzing reported outcomes,[32,33] we expect similar results from extending the method to combined outcomes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Role of the funding source

References

1. Apgar V A proposal for a new method of evaluation of the newborn infant. Anesth Analg 1953; 32: 250–259.

2. Teasdale G, Jennett B. Assessment of coma and impaired consciousness: A practical scale. The Lancet 1974; 304(7872): 81–84.

3. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care 1998; 36(1): 8–27. [PubMed: 9431328]

4. Evans SR, Rubin D, Follmann D, et al. Desirability of outcome ranking (DOOR) and response adjusted for duration of antibiotic risk (RADAR). Clin Infect Dis 2015; 61(5): 800–806. [PubMed: 26113652]

5. Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials greater precision but with greater uncertainty?. JAMA 2003; 289(19): 2554–2559. [PubMed: 12759327]

6. Jackson D, Riley RD, White IR. Multivariate meta-analysis: Potential and promise. Stat Med 2011; 30(20): 2481–2498. [PubMed: 21268052]

7. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. Stat Med 1996; 15(6): 619–629. [PubMed: 8731004]

8. Normand SLT. Meta-analysis: formulating, evaluating, combining, and reporting. Stat Med 1999; 18(3): 321–359. [PubMed: 10070677]

9. Nam IS, Mengersen K, Garthwaite P. Multivariate meta-analysis. Stat Med 2003; 22(14): 2309–2333. [PubMed: 12854095]

10. Wei Y, Higgins JP. Bayesian multivariate meta-analysis with multiple outcomes. Stat Med 2013; 32: 2911–2934. [PubMed: 23386217]

11. Bujkiewicz S, Thompson JR, Sutton AJ, et al. Multivariate meta-analysis of mixed outcomes: a Bayesian approach. Stat Med 2013; 32(22): 3926–3943. [PubMed: 23630081]

12. Jackson D, White IR, Thompson SG. Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. Stat Med 2010; 29(12): 1282–1297. [PubMed: 19408255]

13. Chen Y, Hong C, Riley RD. An alternative pseudolikelihood method for multivariate random-effects meta-analysis. Stat Med 2015; 34(3): 361–380. [PubMed: 25363629]

14. Jackson D, White IR, Riley RD. A matrix-based method of moments for fitting the multivariate random effects model for meta-analysis and meta-regression. Biom J 2013; 55(2): 231–245. [PubMed: 23401213]

15. Ma Y, Mazumdar M. Multivariate meta-analysis: a robust approach based on the theory of U-statistic. Stat Med 2011; 30(24): 2911–2929. [PubMed: 21830230]

16. Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. Stat Med 2007; 26(1): 78–97. [PubMed: 16526010]

17. Jackson D, White IR, Price M, Copas J, Riley RD. Borrowing of strength and study weights in multivariate and network meta-analysis. Stat Methods Med Res 2017; 26(6): 2853–2868. [PubMed: 26546254]

18. Riley RD. Multivariate meta-analysis: the effect of ignoring within-study correlation. J R Stat Soc Ser A Stat Soc 2009; 172(4): 789–811.

19. Ishak KJ, Platt RW, Joseph L, Hanley JA. Impact of approximating or ignoring within-study covariances in multivariate meta-analyses. Stat Med 2008; 27(5): 670–686. [PubMed: 17492826]

20. Mavridis D, Salanti G. A practical introduction to multivariate meta-analysis. Stat Methods Med Res 2013; 22(2): 133–158. [PubMed: 22275379]

21. Wei Y, Higgins JP. Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. Stat Med 2013; 32(7): 1191–1205. [PubMed: 23208849]

22. Arends LR, Hunink MM, Stijnen T. Meta-analysis of summary survival curve data. Stat Med 2008; 27(22): 4381–4396. [PubMed: 18465839]

23. Riley RD, Lambert PC, Staessen JA, et al. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. Stat Med 2007; 27(11): 1870–1893.

24. Riley RD, Price MJ, Jackson D, et al. Multivariate meta-analysis using individual participant data. Res Synth Methods 2015; 6(2): 157–174. [PubMed: 26099484]

25. Berrington A, Cox D. Generalized least squares for the synthesis of correlated information. Biostatistics 2003; 4(3): 423–431. [PubMed: 12925509]

26. Berkey C, Anderson J, Hoaglin D. Multiple-outcome meta-analysis of clinical trials. Stat Med 1996; 15(5): 537–557. [PubMed: 8668877]

27. Riley RD, Thompson JR, Abrams KR. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. Biostatistics 2007; 9(1): 172–186. [PubMed: 17626226]

28. Hong C, Riley RD, Chen Y. An improved method for bivariate meta-analysis when within-study correlations are unknown. Res Synth Methods 2018; 9(1): 73–88. [PubMed: 29055096]

29. Doob JL. The limiting distributions of certain statistics. Ann Math Stat 1935; 6(3): 160–169.

30. Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley . 1987.

31. Crowther MJ, Riley RD, Staessen JA, Wang J, Gueyffier F, Lambert PC. Individual patient data meta-analysis of survival data using Poisson regression models. BMC Med Res Methodol 2012; 12(1): 34. [PubMed: 22443286]

32. Papadimitropoulou K, Stijnen T, Dekkers OM, Cessie lS. One-stage random effects meta-analysis using linear mixed models for aggregate continuous outcome data. Res Synth Methods 2019; 10(3): 360–375. [PubMed: 30523676]

33. Yamaguchi Y, Sakamoto W, Shirahata S, Goto M. A meta-analysis method based on simulated individual patient data. Proc. 58th World Statistical Congress 2011.

34. Guyot P, Ades A, Ouwens M, Welton N. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Med Res Methodol 2012; 12: 9. [PubMed: 22297116]

35. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: An update. Contemp Clin Trials 2007; 28(2): 105–114. [PubMed: 16807131]

36. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986; 7(3): 177–188. [PubMed: 3802833]

37. Efron B Bootstrap methods: Another look at the Jackknife. Ann Stat 1979; 7(1): 1–26.

38. Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers.. Stat Med 1997; 16 17: 1965–82. [PubMed: 9304767]

39. Mavian AA, Miller S, Henry RR. Managing type 2 diabetes: Balancing HbA1c and body weight. Postgrad Med 2010; 122(3): 106–117. [PubMed: 20463420]

40. Type 2 diabetes and obesity: Twin epidemics. American Society for Metabolic and Bariatric Surgery https://asmbs.org/resources/weight-and-type-2-diabetes-after-bariatric-surgery-fact-sheet Published 2013. Accessed November 20, 2019.

41. Wu S, Gao L, Cipriani A, et al. The effects of incretin-based therapies on β-cell function and insulin resistance in type 2 diabetes: A systematic review and network meta-analysis combining 360 trials. Diabetes Obes Metab 2019; 21(4): 975–983. [PubMed: 30536884]

42. Peppard PE, Young T, Barnet JH, Palta M, Hagen EW, Hla KM. Increased prevalence of sleep-disordered breathing in adults. Am J Epidemiol 2013; 177 9: 1006–14. [PubMed: 23589584]

43. Somers VK, White DP, Amin R, et al. Sleep apnea and cardiovascular disease: An American Heart Association/American College of Cardiology Foundation Scientific Statement from the American Heart Association Council for High Blood Pressure Research Professional Education Committee, Council on Clinical Cardiology, Stroke Council, and Council on Cardiovascular Nursing in

collaboration with the National Heart, Lung, and Blood Institute National Center on Sleep Disorders Research (National Institutes of Health). J Am Coll Cardiol 2008; 52(8): 686–717. [PubMed: 18702977]

44. Kasai T, Floras JS, Bradley TD. Sleep apnea and cardiovascular disease: a bidirectional relationship. Circulation 2012; 126(12): 1495–1510. [PubMed: 22988046]

45. Gottlieb D, Yenokyan G, Newman A, et al. Prospective study of obstructive sleep apnea and incident coronary heart disease and heart failure: The sleep heart health study. Circulation 2010; 122: 352–60. [PubMed: 20625114]

46. Yaggi HK, Concato J, Kernan WN, Lichtman JH, Brass LM, Mohsenin V. Obstructive sleep apnea as a risk factor for stroke and death. N Engl J Med 2005; 353(19): 2034–2041. [PubMed: 16282178]

47. Gami AS, Pressman G, Caples SM, et al. Association of atrial fibrillation and obstructive sleep apnea. Circulation 2004; 110(4): 364–367. [PubMed: 15249509]

48. Marin JM, Carrizo SJ, Vicente E, Agusti AG. Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure: an observational study. The Lancet 2005; 365(9464): 1046–1053.

49. Peppard PE, Young T, Palta M, Skatrud J. Prospective study of the association between sleep-disordered breathing and hypertension. N Engl J Med 2000; 342(19): 1378–1384. [PubMed: 10805822]

50. Hu W, Jin X, Chen C, et al. Diastolic blood pressure rises with the exacerbation of obstructive sleep apnea in males. Obesity 2017; 25(11): 1980–1987. [PubMed: 28922573]

51. Seif F, Patel SR, Walia HK, et al. Obstructive sleep apnea and diurnal nondipping hemodynamic indices in patients at increased cardiovascular risk. J Hypertens 2014; 32(2): 267. [PubMed: 24351803]

52. Rückert IM, Baumert J, Schunk M, et al. Blood pressure control has improved in people with and without type 2 diabetes but remains suboptimal: a longitudinal study based on the German DIAB-CORE consortium. PloS One 2015; 10(7): e0133493. [PubMed: 26221962]

53. Schein AS, Kerkhoff AC, Coronel CC, Plentz RD, Sbruzzi G. Continuous positive airway pressure reduces blood pressure in patients with obstructive sleep apnea; a systematic review and meta-analysis with 1000 patients. J Hypertens 2014; 32(9): 1762–1773. [PubMed: 24979300]

54. Zhao Y, Wang R, Gleason K, et al. Effect of continuous positive airway pressure treatment on health-related quality of life and sleepiness in high cardiovascular risk individuals with sleep apnea: Best apnea interventions for research (BestAIR) trial. Sleep 2017; 40: zsx040.

55. Whelton PK, Carey RM, Aronow WS, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol 2018; 71(19): 127–248.

56. Sesso H, Stampfer M, Rosner B, et al. Systolic and diastolic blood pressure, pulse pressure, and mean arterial pressure as predictors of cardiovascular disease risk in men. Hypertension 2000; 36: 801–807. [PubMed: 11082146]

57. Gavish B, Ben-dov I, Bursztyn M. Linear relationship between systolic and diastolic blood pressure monitored over 24 h: Assessment and correlates. J Hypertens 2008; 26: 199–209. [PubMed: 18192832]

58. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002; 21(11): 1539–1558. [PubMed: 12111919]

59. Van Buuren S Flexible Imputation of Missing Data. Chapman and Hall/CRC 2018.

60. Cochran WG. The combination of estimates from different experiments. Biometrics 1954; 10(1): 101–129.

61. Li KH, Meng XL, Raghunathan TE, Rubin DB. Significance levels from repeated p-values with multiply-imputed data. Stat Sin 1991; 1(1): 65–92.

62. Headrick T, Sawilowsky S. Simulating correlated multivariate nonnormal distributions: Extending the Fleishman power method. Psychometrika 1999; 64: 251–251.

63. Wang R, Tian L, Cai T, Wei LJ. Nonparametric inference for percentiles of the random effects distribution in meta-analysis. Ann Appl Stat 2010; 4(1): 520–532. [PubMed: 25678939]

64. Derosa G, Ragonesi PD, Fogari E, et al. Sitagliptin added to previously taken antidiabetic agents on insulin resistance and lipid profile: a 2-year study evaluation. Fundam Clin Pharmacol 2014; 28(2): 221–229. [PubMed: 23039403]

65. Derosa G, Carbone A, Franzetti I, et al. Effects of a combination of sitagliptin plus metformin vs metformin monotherapy on glycemic control, β-cell function and insulin resistance in type 2 diabetic patients. Diabetes Res Clin Pract 2012; 98(1): 51–60. [PubMed: 22682949]

66. Ahrén B, Johnson SL, Stewart M, et al. HARMONY 3: 104-week randomized, double-blind, placebo-and active-controlled trial assessing the efficacy and safety of albiglutide compared with placebo, sitagliptin, and glimepiride in patients with type 2 diabetes taking metformin. Diabetes Care 2014; 37(8): 2141–2148. [PubMed: 24898304]

67. Iwamoto Y, Taniguchi T, Nonaka K, et al. Dose-ranging efficacy of sitagliptin, a dipeptidyl peptidase-4 inhibitor, in Japanese patients with type 2 diabetes mellitus. Endocr J 2010; 57(5): 383–394. [PubMed: 20332588]

68. Mohan V, Yang W, Son HY, et al. Efficacy and safety of sitagliptin in the treatment of patients with type 2 diabetes in China, India, and Korea. Diabetes Res Clin Pract 2009; 83(1): 106–116. [PubMed: 19097665]

69. Nonaka K, Kakikawa T, Sato A, et al. Efficacy and safety of sitagliptin monotherapy in Japanese patients with type 2 diabetes. Diabetes Res Clin Pract 2008; 79(2): 291–298. [PubMed: 17933414]

70. Gül ÖÖ, Kıyıcı S, Ersoy C, et al. Effect of sitagliptin monotherapy on serum total ghrelin levels in people with type 2 diabetes. Diabetes Res Clin Pract 2011; 94(2): 212–216. [PubMed: 21855160]

71. Roden M, Weng J, Eilbracht J, et al. Empagliflozin monotherapy with sitagliptin as an active comparator in patients with type 2 diabetes: a randomised, double-blind, placebo-controlled, phase 3 trial. Lancet Diabetes Endocrinol 2013; 1(3): 208–219. [PubMed: 24622369]

72. Samocha-Bonet D, Chisholm DJ, Gribble FM, et al. Glycemic effects and safety of L-Glutamine supplementation with or without sitagliptin in type 2 diabetes patients—A randomized study. PloS one 2014; 9(11).

73. Tian M, Liang Z, Liu R, et al. Effects of sitagliptin on circulating zinc-a2-glycoprotein levels in newly diagnosed type 2 diabetes patients: a randomized trial. Eur J Endocrinol 2016; 174: 147–155. [PubMed: 26546612]

74. Vilsbøll T, Rosenstock J, Yki-Järvinen H, et al. Efficacy and safety of sitagliptin when added to insulin therapy in patients with type 2 diabetes. Diabetes Obes Metab 2010; 12(2): 167–177. [PubMed: 20092585]

75. Violante R, Oliveira J, Yoon KH, et al. A randomized non-inferiority study comparing the addition of exenatide twice daily to sitagliptin or switching from sitagliptin to exenatide twice daily in patients with type 2 diabetes experiencing inadequate glycaemic control on metformin and sitagliptin. Diabet Med 2012; 29(11): e417–e424. [PubMed: 22375612]

76. Yoon K, Steinberg H, Teng R, et al. Efficacy and safety of initial combination therapy with sitagliptin and pioglitazone in patients with type 2 diabetes: a 54-week study. Diabetes Obes Metab 2012; 14(8): 745–752. [PubMed: 22405352]

77. Becker HF, Jerrentrup A, Ploch T, et al. Effect of nasal continuous positive airway pressure treatment on blood pressure in patients with obstructive sleep apnea. Circulation 2003; 107(1): 68–73. [PubMed: 12515745]

78. Campos-Rodriguez F, Grilo-Reina A, Perez-Ronchel J, et al. Effect of continuous positive airway pressure on ambulatory BP in patients with sleep apnea and hypertension: a placebo-controlled trial. Chest 2006; 129(6): 1459–1467. [PubMed: 16778262]

79. Durán-Cantolla J, Aizpuru F, Montserrat JM, et al. Continuous positive airway pressure as treatment for systemic hypertension in people with obstructive sleep apnea: randomised controlled trial. BMJ 2010; 341: c5991. [PubMed: 21106625]

80. Hui DS, To KW, Ko FW, et al. Nasal CPAP reduces systemic blood pressure in patients with obstructive sleep apnea and mild sleepiness. Thorax 2006; 61(12): 1083–1090. [PubMed: 16928705]

81. Kohler M, Pepperell J, Casadei B, et al. CPAP and measures of cardiovascular risk in males with OSAS. Eur Respir J 2008; 32(6): 1488–1496. [PubMed: 18653654]

82. Robinson G, Smith D, Langford B, Davies R, Stradling J. Continuous positive airway pressure does not reduce blood pressure in nonsleepy hypertensive OSA patients. Eur Respir J 2006; 27(6): 1229–1235. [PubMed: 16455835]

Highlights

In practice, treatment decision-making depends on an overall assessment of patient-level outcomes balancing benefit in various domains and potential risks. This calls for meta-analysis methods for combined outcomes that encompass information from different domains. When individual patient data (IPD) are available from each study, the combined outcomes can be constructed for each individual and standard meta-analysis methods would apply. However, IPD are usually difficult to obtain. In this article, we propose a new method that multiply reconstructs pseudo IPD based on the available summary statistics reported by each study and then obtains pooled treatment effect estimates for the combined outcome. This proposed method is applicable to various settings where meta-analysis of outcomes that integrates information from multiple variables is of interest.
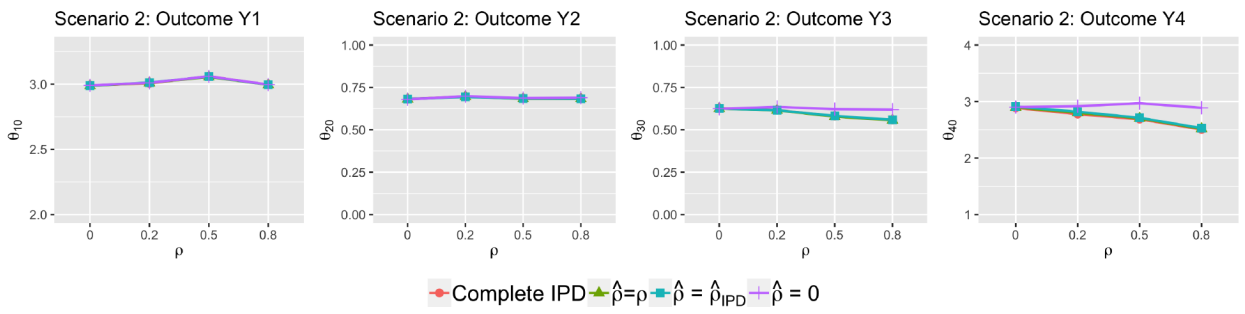
Figure 1:

Treatment effect estimates (scenario 2)

Comparisons of average treatment effect estimates based on various $\hat{\rho}$ versus those obtained from complete IPD. One of the studies provided IPD and the rest provided AD. The simulation parameter values were $J = 7$, $\left(\tau_i^U\right)^2 = \left(\tau_i^V\right)^2 = 0.5$, $i = 1, 2$, $\kappa = 0.5$.

Figure 2:

Standard errors (scenario 1–4)

Comparisons of standard error estimates of treatment effect estimates on combined outcomes for various $\hat{\rho}$. One study provided IPD and the rest provided AD. EMP refers to the empirical standard errors, NAV refers to the average standard error estimates obtained by a naive application of the Rubin's rule; BOOT refers to the average bootstrap standard error estimates.

Figure 3:

Average treatment effect estimates (scenario 5–8)

Comparisons of treatment effect estimates based on various $\hat{\rho}$ to those obtained from complete IPD. All studies provided AD only.
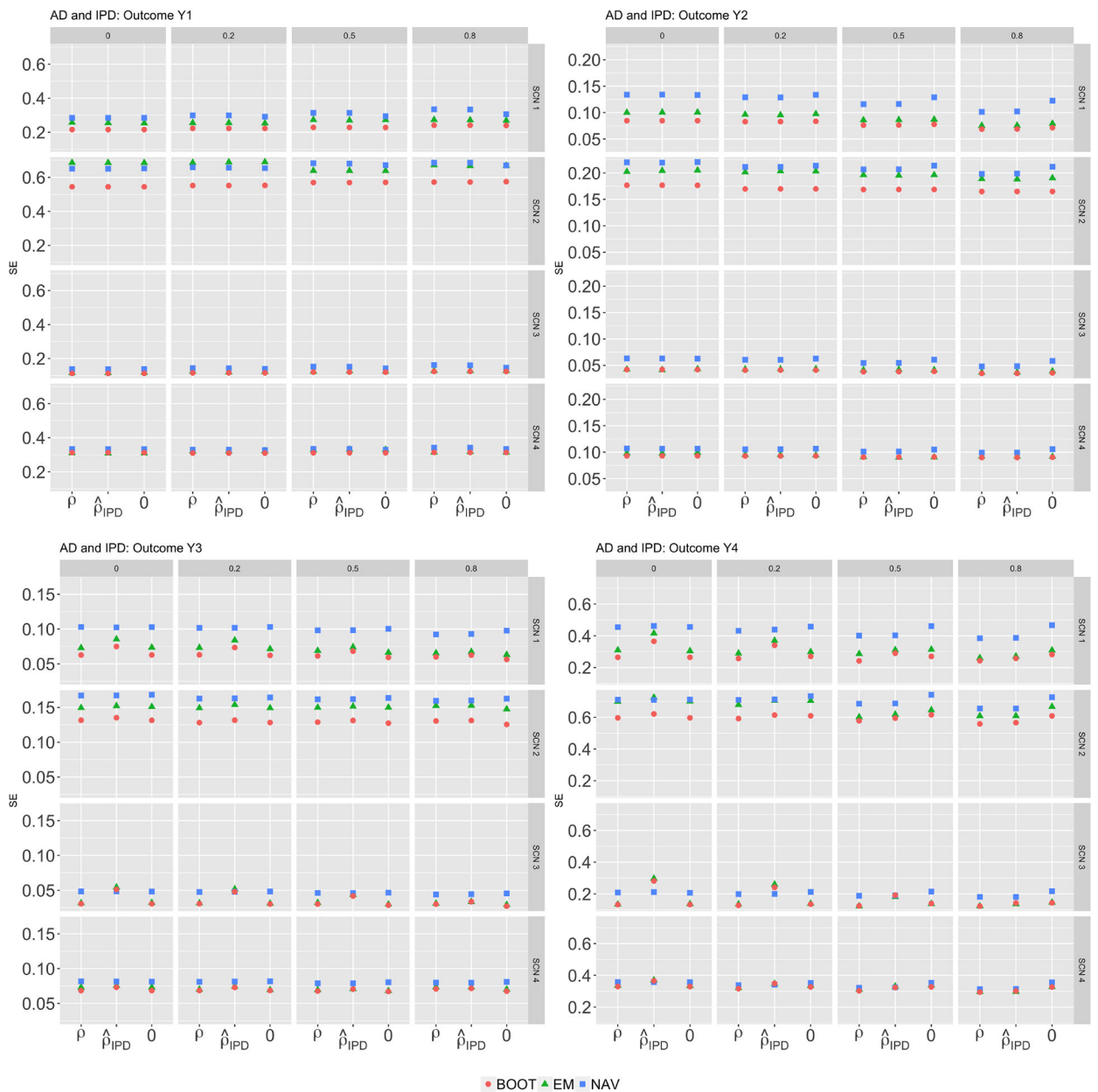
Figure 4:
Standard errors (scenario 5–8)Comparisons of standard error estimates of treatment effect estimates on combined outcomes for various $\hat{\rho}$. All studies provided AD only. EMP refers to the empirical standard errors; NAV refers to the average of standard error estimates obtained by a naive application of the Rubin's rule; BOOT refers to the average bootstrap standard error estimates.
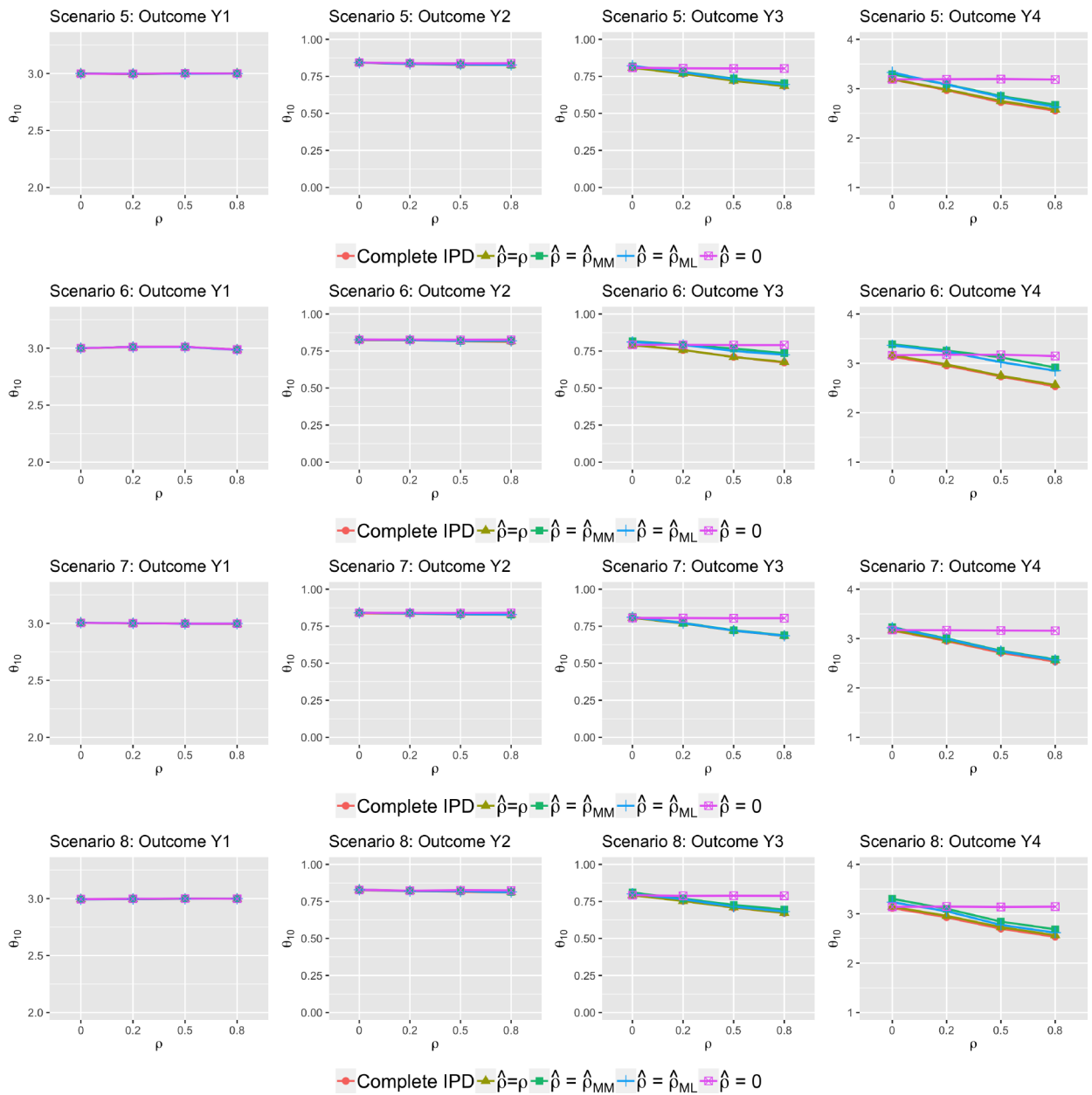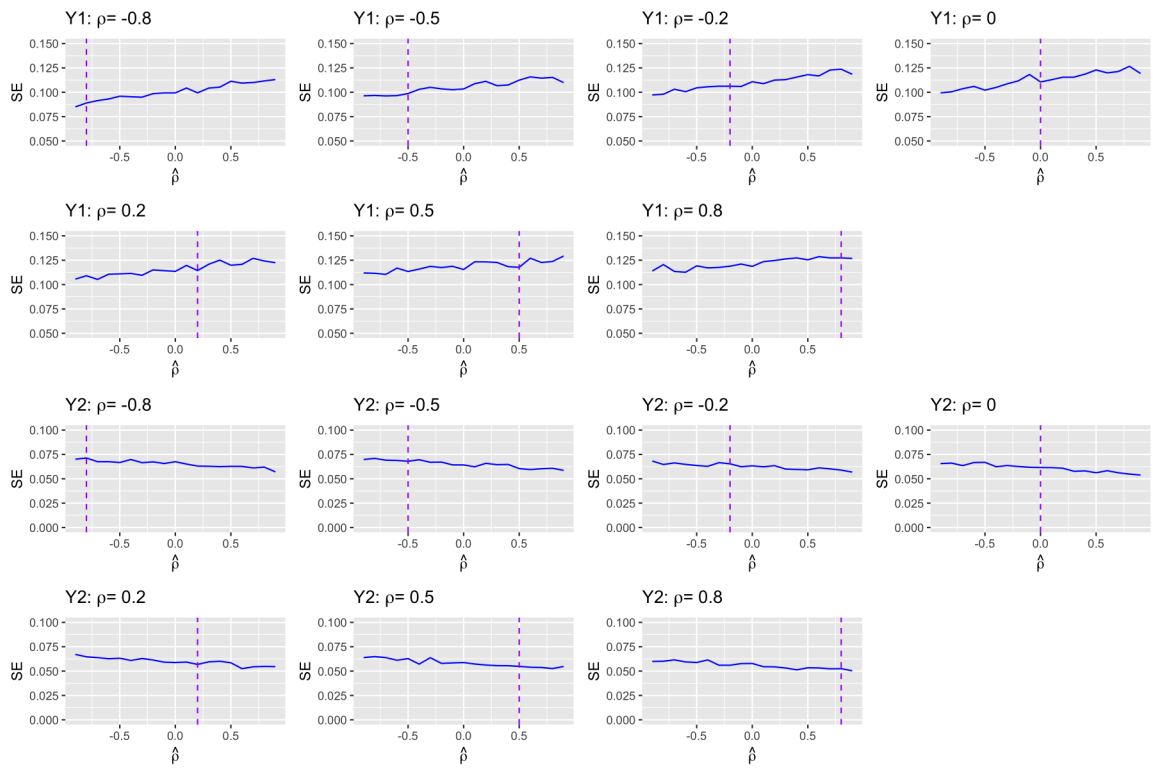
Figure 5:
Empirical standard errors for outcomes Y_1 and Y_2 based on a single reconstruction of pseudo IPD. The vertical purple lines indicate the true ρ.

Table 1:

Comparison of model setups for the proposed approach and the standard MVMA

	The proposed approach	MVMA
Within-study correlation	$corr\left(U_{ijk}, V_{ijk} \mid \gamma_{ij}^{U}, \gamma_{ij}^{V}\right)$	$corr\left(\left(\overline{U}_{1j} - \overline{U}_{2j}\right), \left(\overline{V}_{1j} - \overline{V}_{2j}\right) \mid \gamma_{1j}^{U}, \gamma_{2j}^{U}, \gamma_{1j}^{V}, \gamma_{2j}^{V}\right)$
Between-study correlation	$corr\left(\gamma_{ij}^{U}, \gamma_{ij}^{V}\right)$	$corr\left(\left(\gamma_{1j}^{U} - \gamma_{2j}^{U}\right), \left(\gamma_{1j}^{V} - \gamma_{2j}^{V}\right)\right)$

Table 2:

Simulation scenarios

IPD and AD ($\kappa = 0.5$)			AD only ($\kappa = 0$)		
Scenario	J	$\left(\tau_i^U\right)^2, \left(\tau_i^V\right)^2$	Scenario	J	$\left(\tau_i^U\right)^2, \left(\tau_i^V\right)^2$
1	7	0.05	5	13	0.01
2	7	0.5	6	13	0.05
3	30	0.05	7	50	0.01
4	30	0.5	8	50	0.05

Table 3:

Comparison of biases of treatment effect estimates on Y_1 obtained from multiple reconstructions of IPD (MR) and from MVMA

	MR	MVMA-REML	MVMA-MM
$\rho = -0.8$	0.001	0.005	0.005
$\rho = -0.5$	0.001	−0.002	−0.002
$\rho = -0.2$	0.007	−0.001	−0.001
$\rho = 0$	0.001	−0.008	−0.008
$\rho = 0.2$	−0.007	0.004	0.004
$\rho = 0.5$	0.006	−0.004	−0.004
$\rho = 0.8$	0.002	−0.006	−0.005

Table 4:

Comparison of empirical SEs of treatment effect estimates on Y_1 obtained from multiple reconstructions of IPD (MR) and from MVMA

	MR	MVMA-REML	MVMA-MM
$\rho = -0.8$	0.087	0.087	0.087
$\rho = -0.5$	0.093	0.090	0.090
$\rho = -0.2$	0.098	0.096	0.096
$\rho = 0$	0.101	0.098	0.098
$\rho = 0.2$	0.107	0.104	0.104
$\rho = 0.5$	0.111	0.110	0.110
$\rho = 0.8$	0.115	0.111	0.111

Table 5:

Aggregate data on HbA1c level and body weight

Study	Intervention	N	HbA1c (%)		Weight (kg)	
			Mean	**SE**	**Mean**	**SE**
Derosa, 2014[64]	DPP4-I	101	−1.40	0.07	−3.50	0.56
	Control	96	0.50	0.12	−0.80	0.62
Derosa, 2012[65]	DPP4-I	86	−1.40	0.08	−2.50	0.64
	Control	83	−0.70	0.07	−2.30	0.67
Ahren, 2014[66]	DPP4-I	89	−0.56	0.12	−2.05	0.44
	Control	16	−0.46	0.21	−3.61	0.87
Iwamoto, 2010[67]	DPP4-I	70	−0.69	0.06	0.10	0.18
	Control	73	0.28	0.06	−0.50	0.18
Mohan, 2009[68]	DPP4-I	352	−0.71	0.05	0.60	0.09
	Control	169	0.31	0.10	0	0.19
Nonaka, 2008[69]	DPP4-I	75	−0.65	0.08	−0.10	0.18
	Control	75	0.41	0.08	−0.70	0.15
Gul, 2011[70]	DPP4-I	28	−0.30	0.14	−2.00	3.16
	Control	16	−0.10	0.19	−2.20	2.74
Roden, 2013[71]	DPP4-I	223	−0.53	0.06	0.10	0.20
	Control	228	0.13	0.06	−0.43	0.20
Samocha-Bonet, 2014[72]	DPP4-I	13	−0.20	0.08	−0.20	4.22
	Control	13	−0.10	0.11	0.10	4.34
Tian, 2016[73]	DPP4-I	88	−0.36	0.06	−0.40	0.86
	Control	45	0.02	0.11	−0.40	0.96
Vilsboll, 2010[74]	DPP4-I	322	−0.59	0.05	0.10	0.13
	Control	319	−0.03	0.06	0.10	0.16
Violante, 2012[75]	DPP4-I	111	−0.68	0.08	−2.20	0.24
	Control	97	−0.38	0.09	−2.58	0.25
Yoon, 2012[76]	DPP4-I	161	−2.37	0.09	4.80	0.51
	Control	149	−1.86	0.09	4.10	0.54

Table 6:

Meta-analysis of the HbA1c-weight combined outcome

$\hat{\rho}$	$\hat{\theta}$	Bootstrap SE	95% CI
1	0.595	0.156	[0.290, 0.900]
0.8	0.598	0.157	[0.290, 0.905]
0.5	0.603	0.159	[0.292, 0.914]
0.2	0.608	0.160	[0.295, 0.921]
0	0.618	0.161	[0.301, 0.932]
−0.2	0.627	0.161	[0.310, 0.943]
−0.5	0.649	0.163	[0.327, 0.966]
−0.8	0.661	0.165	[0.338, 0.985]
Assume $\kappa_0 = 0$			
$\hat{\rho}_{ML} = -1$	0.699	0.153	[0.398, 0.999]
$\hat{\rho}_{MM} = -1$	0.705	0.160	[0.392, 1.018]

Table 7:

Aggregate data on post-intervention 24-hour SBP and DBP

Study	Intervention	N	SBP(mmHg)		DBP (mmHg)	
			Mean	SE	Mean	SE
Becker, 2003[77]	PAP	16	126	3.50	73	2.50
	Control	16	137	2.75	82	2.25
Campos-Rodrigues, 2006[78]	PAP	34	131	2.06	77	1.54
	Control	34	130	2.74	77	1.54
Duran-Camtola, 2010[79]	PAP	169	128	1.00	81	0.69
	Control	171	128	0.99	82	0.69
Hui, 2006[80]	PAP	23	123	3.13	80	2.29
	Control	23	120	6.26	81	1.88
Kohler, 2008[81]	PAP	51	128	1.96	81	1.40
	Control	51	139	2.66	89	1.12
Robinson, 2006[82]	PAP	16	137	4.00	84	3.00
	Control	16	139	4.50	87	3.00
Zhao, 2017[54]	PAP	69	123	1.22	71	0.93
	Control	68	128	1.58	74	1.04