Published in final edited form as:

Stat (Int Stat Inst). 2022 December ; 11(1): . doi:10.1002/sta4.462.

A hierarchical meta-analysis for settings involving multiple outcomes across multiple cohorts

Tugba Akkaya Hocagil[1], **Louise M Ryan**[2], **Richard J. Cook**[1], **Sandra W. Jacobson**[3], **Gale A. Richardson**[4], **Nancy L. Day**[4], **Claire D. Coles**[5], **Heather Carmichael Olson**[6], **Joseph L. Jacobson**[3]

[1]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, N2L 3G1, Canada

[2]School of Mathematical and Physical Sciences, University of Technology Sydney, Sydney, NSW, 2007, Australia

[3]Department of Psychiatry and Behavioral Neurosciences, Wayne State University, Detroit, Michigan, 48201, USA

[4]Department of Psychiatry, University of Pittsburgh, Pittsburgh, Pennsylvania, 15213, USA

[5]Department of Psychiatry and Behavioral Sciences, Emory University, Atlanta, Georgia, 30322, USA

[6]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington, 98195, USA

Abstract

Evidence from animal models and epidemiological studies has linked prenatal alcohol exposure (PAE) to a broad range of long-term cognitive and behavioural deficits. However, there is a paucity of evidence regarding the nature and levels of PAE associated with increased risk of clinically significant cognitive deficits. To derive robust and efficient estimates of the effects of PAE on cognitive function, we have developed a hierarchical meta-analysis approach to synthesize information regarding the effects of PAE on cognition, integrating data on multiple outcomes from six U.S. Iongitudinal cohort studies. A key assumption of standard methods of meta-analysis, effect sizes are independent, is violated when multiple intercorrelated outcomes are synthesized across studies. Our approach involves estimating the dose–response coefficients for each outcome and then pooling these correlated dose–response coefficients to obtain an estimated "global" effect of exposure on cognition. In the first stage, we use individual participant data to derive estimates of the effects of PAE by fitting regression models that adjust for potential confounding variables using propensity scores. The correlation matrix characterizing the dependence between the outcome-specific dose–response coefficients estimated within each cohort is then run, while accommodating incomplete information on some outcome. We also compare inferences based on the proposed approach to inferences based on a full multivariate analysis.

Correspondence: Tugba Akkaya Hocagil, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada. takkayahocagil@uwaterloo.ca.

Keywords

Cognition; fetal alcohol syndrome; hierarchical model; multiple outcomes; prenatal alcohol exposure; synthesis of evidence; two-stage estimation

1 | INTRODUCTION

Meta-analysis is commonly used to synthesize quantitative evidence across studies to generate a summary exposure or treatment effect that is more precise than estimates obtainable from individual studies alone. Traditionally, meta-analysis is based on estimated effect sizes. Although it is cost-effective and easy to implement, this approach has been criticized on the grounds that it is prone to ecological and confounding bias (Riley & Steyerberg, 2010; Simmonds & Higgins, 2007). Individual participant data (IPD) meta-analysis can help mitigate such biases and accommodate missing data at the participant level (Riley et al., 2010). Moreover, with access to individual level data, a choice can be made between a fully specified multivariate IPD and a two-stage IPD approach. The full multivariate approach generally uses mixed-effects multilevel regressions to model between and within heterogeneity and quantify the effect of interest in a single model. Although this approach is considered flexible, it may be challenging for conducting and communicating the findings, particularly visualization using the hallmark forest plot. The alternative IPD approach involves modelling the data in two stages. In the first step, effect size estimates for each study are obtained using separate regression models. In the second step, standard methods of meta-analysis are used to obtain an overall estimate. A key assumption with standard methods of meta-analysis is that effect sizes are independent. This assumption is violated when multiple outcomes correlated are synthesized across studies. To avoid the dependence of the effect sizes, several ad-hoc methods have been proposed including averaging the effect sizes or selecting one effect size per study. A major disadvantage of these ad-hoc approaches is that they do not make use of all available data to address the relevant research questions (Cheung, 2019).

More principled approaches have been proposed to deal with correlated effects when conducting IPD meta-analysis. These advances include multivariate meta-analysis which has been used to jointly synthesize the outcomes observed across studies to estimate multiple pooled effects simultaneously (Riley et al., 2007). However, using multivariate meta-analysis is less straightforward when studies do not consistently report on the same outcomes (Van den Noortgate et al., 2014). Another approach is the three-level meta analytic model (Cheung, 2013; Konstantopoulos, 2011; Van den Noortgate et al., 2013), which has been used to adjust for dependence of effect sizes within clusters. This approach treats participants within each cluster as contributing only one effect size, so the nonindependence is handled within the nested structure of the effects (Cheung, 2019). An alternative approach is based on a two-stage meta-analysis that uses summary measures. In this approach, dependency among effect sizes is handled via robust variance estimation in which the dependence between the outcomes is not explicitly modelled, but instead the standard errors for the overall treatment effect or meta-regression coefficients are adjusted (Hedges et al., 2010). This approach may require making a reasonable guess about the between-outcome correlations to estimate the between-study variance and to approximate the optimal weights.

In this paper, we propose an innovative approach: a hierarchical meta-analysis for the settings in which each cohort study provides multiple outcomes, resulting in correlated estimated effect sizes. The work is motivated by a project that involves the integration of data from six longitudinal cohorts, each of which used multiple interrelated tests and assessment tools to measure child cognition. Cognition is not directly observable since there is no single measure that can be regarded as a highly reliable indicator of cognition. These six longitudinal cohorts were conducted independently and used different neuropsychological test batteries to assess IQ and the same domains of cognitive function, including learning and memory, executive function, and academic achievement in reading and in mathematics (Jacobson et al., 2021). All these tests provide a comprehensive assessment of the child's cognitive function. A major strength of the proposed approach is that it facilitates the synthesis of data across diverse outcomes within each cohort, and thereby an assessment of consistency of patterns across cohorts. Furthermore, by including multiple correlated responses from each child, the analyses make full use of available data to maximize the efficiency of estimation and enhance power of associated tests for effects. Robust variance estimation ensures valid inferences at each stage of the analysis.

In the proposed approach, we first derive the estimates and the standard errors by fitting regression models for each separate outcome of interest. By contrast to existing methods of two-stage IPD analysis, we account for the correlated effect sizes within each cohort at this stage. Specifically, within-study robust covariance matrices are obtained at this stage to be combined at the second stage. Within each cohort, not all outcomes were observed for all children. This additional complexity was addressed in the estimation of robust covariance matrices. Specifically, we derived a formula for pairwise correlation between the estimated effects using an adjustment that accounts for the fact that we have partially observed outcome measures for some children. In the second stage, we combine the summary measures within each cohort using a random-effects model. In the last stage of our hierarchical meta-analytic approach, we combine the independent, cohort-specific effect size estimates in a random-effects model to obtain a global measure of the effect size across cohorts (Lin & Zeng, 2010; Whitehead, 2002). We compare and contrast the findings from our proposed approach to those obtained using a full multivariate analysis in order to determine the degree to which the results of these two models coincide.

The remainder of the article is organized as follows. In Section 2, we introduce our motivating application which is a meta-analysis of correlated outcomes used to assess the effect of prenatal alcohol exposure (PAE) on cognition in six cohort studies. In Section 3, we introduce notation and describe the two-stage analysis and modelling framework used to combine multiple correlated outcomes within a single cohort. In Section 4, we present the modelling framework used to combine pooled effect size estimates across cohorts. In Section 5, we compare and contrast the proposed approach with the corresponding one-stage approach using simulation studies. In Section 6, we illustrate our method using data from our motivating application. Finally, in Section 7, we discuss the strengths and limitations of our method.

2 | EFFECT OF PRENATAL ALCOHOL EXPOSURE ON COGNITION

Evidence from animal models and epidemiological studies has linked PAE to a broad range of cognitive and behavioural deficits, growth restriction, and physical anomalies, which are known collectively as fetal alcohol syndrome disorders (FASD) (Carter et al., 2016; Jacobson et al., 2004, 2008; Mattson et al., 2019). Fetal alcohol syndrome (FAS), the most severe of the FASD, is characterized by distinctive cranio-facial dysmorphology (small palpebral fissures, flat philtrum, thin vermillion), small head circumference, and growth restriction (Hoyme et al., 2005; Stratton et al., 1996) while partial FAS (PFAS) is diagnosed in the presence of the characteristic alcohol-related facial dysmorphology, a history of PAE and growth restriction, small head circumference, or central nervous system (CNS) impairment. Individuals with PAE who lack the characteristic pattern of dysmorphic features but exhibit cognitive and/or behavioural impairment are diagnosed as having alcohol-related neurodevelopmental disorder (ARND), which is the most prevalent FASD. Although the diagnosis of ARND requires a confirmed history of maternal alcohol consumption during pregnancy, there is no information in the scientific literature regarding the levels of PAE associated with an increased risk of clinically significant adverse effects.

Between 1975 and 1993, the National Institutes of Health (NIH) funded six longitudinal cohort studies in four U.S. cities–Detroit (Jacobson et al., 1993), Pittsburgh (Day et al., 1991; Richardson et al., 1999) (two cohorts), Atlanta (Brown et al., 1998; Coles et al., 2006) (two cohorts), and Seattle (Streissguth et al., 1981); these are described briefly in Appendix A. To enhance efficiency when examining the effects associated with different levels and patterns of PAE, the data are synthesized across the studies. The sample sizes in the individual longitudinal cohort studies range between 138 and 720. Participant retention was good to excellent from childhood to adolescence $(median = 90.3\%;\ range = 86.4\ to\ 96.3\%)$. Retention from adolescence to young adulthood was excellent (≥ 91.5%) in the Atlanta-1, Seattle, and 2 Pittsburgh cohorts. The Detroit young adult follow-up, which focused on neuroimaging, was funded to assess only a subsample (43.6%) of the cohort. In all but one of these studies, mothers were recruited and interviewed prospectively about their alcohol use during pregnancy, and their children were followed longitudinally from infancy through young adulthood; one of the Atlanta cohort studies (Schuetze et al., 2007) recruited the mothers shortly following delivery, interviewed them about their drinking during pregnancy, and followed the children through early childhood. The number of maternal interviews varies by cohort. In these interviews, detailed information regarding quantity and frequency of drinking during pregnancy and dose per occasion were obtained. Data on alcohol consumption during pregnancy from all six cohorts are summarized in terms of ounces of absolute alcohol averaged across pregnancy (oz AA). In all studies, investigators administered a variety of neuropsychological tests to assess IQ and four domains of cognitive function: learning and memory, executive function, and academic achievement in reading and in mathematics.

Although there was some variation in the particular auxiliary covariates collected across the different studies, data on a broad range of covariates were provided by each cohort.

3 | NOTATION AND MODEL FORMULATION

Let Y_{ijk} be the random variable representing response k for individual j in cohort $i, k = 1, \ldots, K_i, j = 1, \ldots, J_i$, where J_i is the number of individuals in cohort $i, i = 1, \ldots, l$. Let A_{ij} be the exposure of interest (i.e., prenatal alcohol exposure) for individual j in study i and S_{ij} be their corresponding propensity score.

$$Y_{ijk} = \alpha_{ik} + B_{ik}A_{ij} + \gamma_{ik}S_{ij} + E_{ijk}, \tag{1}$$

where B_{ik} is the effect of a one-unit increase in A_{ij} (alcohol volume) on the mean for response k in cohort i given the propensity score $S_{ij}, j = 1, \ldots, J, k = 1, \ldots, K, i = 1, \ldots, l$.

Because the sets of covariates measured differ between cohorts, the propensity score is estimated separately for each cohort using the two-part generalized propensity score (Akkaya Hocagil et al., 2021). By using the two-part generalized propensity score, we model the causal effect of a semicontinuous exposure variable A on an outcome Y in the presence of a set of confounding variables $Z = (Z_1, \ldots, Z_p)'$. Specifically, we let $A^+ = I(A > 0)$ indicates a positive value for A, and $\pi(Z) = E(A^+ \mid Z)$. We consider a binary regression model defined by the link function $g(\cdot)$ mapping the interval [0,1] onto the real line and setting $g(\pi(Z; \alpha_1)) = \overline{Z}'\alpha_1$ where $\overline{Z} = (1, Z')'$ and $\alpha_1 = (\alpha_{10}, \alpha_{11}, \ldots, \alpha_{1p})'$.

We also let $P(A \leq a \mid A^+ = 1, Z; \alpha_2) = F^+(a \mid Z; \alpha_2)$ denote the cumulative distribution function for the positive part of A given Z, and $A^+ = 1$ is indexed by a $(p + 1) \times 1$ parameter α_2. The full distribution for A is therefore indexed by $\alpha = (\alpha_1', \alpha_2')'$. A key requirement of the model for $A \mid Z, A^+ = 1$ is that it involves a simple way to compute $E(A \mid Z, A^+ = 1; \alpha_2) = \mu(Z; \alpha_2)$; we adopt a generalized linear model and ultimately compute

$$S = E(A \mid Z; \alpha) = \pi(Z; \alpha_1)\mu(Z; \alpha_2) \tag{2}$$

as the marginal mean for $A \mid Z$ based on the two-part model formulation (Akkaya Hocagil et al., 2021).

After we estimated the propensity score, we assume here that conditioning on the propensity score renders the exposure variable independent of all confounders and so that it is sufficient to condition on S_{ij} in (1) rather than the confounders themselves to mitigate the effect of confounding (Rosenbaum & Rubin, 1983).

The parameter γ_{ik} characterizes the effect of the propensity score on outcome k in study i (for a given level of alcohol exposure) and E_{ijk} is the error term which has mean zero and variance $\sigma_{ik}^2, k = 1, \ldots, K, i = 1, \ldots, l$.

We suppose that the effects B_{ik}, the effect of prenatal alcohol exposure on the mean for response k in cohort i, vary about some average exposure effect in cohort i with

$$B_{ik} \mid \beta_i \sim N(\beta_i, \phi_i), \tag{3}$$

where β_i is the exposure effect for cohort i and ϕ_i represents the heterogeneity of the response-specific exposure effects within cohort i.

We suppose that the average cohort-specific exposure effects are independent and vary about an overall exposure effect β_0 with,

$$\beta_i \sim N\left(\beta_o, \eta^2\right); \tag{4}$$

here, β_o represents the "average effect" of a one-unit increase in the exposure across all cohorts and is our parameter of ultimate interest. The variance η^2 in (4) reflects the extent of heterogeneity of the cohort-specific exposure effects.

In the next two subsections, we describe a two-stage approach to estimation and inference with data from a single cohort, and in Section 4, we show how to synthesize cohort-specific exposure effects to obtain an estimate for the average effect of a one-unit increase in the exposure across all cohorts.

3.1 | Stage I estimation for a single cohort

In this section, we temporarily omit the subscript i and describe a two-stage approach to estimate the average exposure effect for a single cohort where the effects are correlated. Before model fitting, we standardize the responses so that they have the same first two moments as the full-scale IQ variable which has a mean of 100 and a standard deviation of 15. By conducting this standardization, the exposure effects can be expressed in terms of the decrement in IQ associated with a one-unit increase in prenatal alcohol exposure (Axelrad et al., 2007).

For the first stage, we fit separate linear models for each response, assuming

$$Y_{jk} = \alpha_k + B_k A_j + \gamma_k S_j + E_{jk}, \tag{5}$$

where B_k is the effect of a one-unit increase in A_j (alcohol in our application) on the mean for response k, given the propensity score $S_j, j = 1, \ldots, J, k = 1, \ldots, K$. The parameter γ_k characterizes the effect of the propensity score for a given level of alcohol exposure and E_{jk} is the error term that has mean zero and variance $\sigma_k^2, k = 1, \ldots, K$.

We suppose that the effects $B_k, k = 1, \ldots, K$ vary about some average exposure effect, with

$$B_k \sim N(\beta, \phi), \tag{6}$$

independently and identically distributed and β is the average exposure effect. The variance ϕ reflects the extent of heterogeneity of the response-specific exposure effects for a single cohort.

If we let $X_{jk} = (1, A_j, S_j)'$ be the covariate vector, we can write

$$Y_{jk} = X_{jk}^{'}\theta_k + E_{jk}, \tag{7}$$

where $\theta_k = (\alpha_k, B_k, \gamma_k)'$. We assume $E_{jk} \perp\!\!\!\perp (A_j, B_k, S_j)$ with $E_{jk} \sim N\left(0, \sigma_k^2\right)$ are i.i.d. for $j = 1, \ldots, J, k = 1, 2, \ldots, K$. Note that X_{jk} does not vary by the response type because the exposure variable A_j and the propensity score S_j are individual level covariates, but we retain this notation for generality.

We next define $K \times 1$ vectors $Y_j = (Y_{j1}, Y_{j2}, \ldots, Y_{jK})'$, $\alpha = (\alpha_1, \ldots, \alpha_K)'$, $B = (B_1, \ldots, B_K)'$ and $\gamma = (\gamma_1, \ldots, \gamma_K)'$ and a $K \times 3K$ covariate matrix

$$X_j' = \begin{bmatrix} X_{j1}' & 0 & 0 & \ldots & 0 \\ 0 & X_{j2}' & 0 & \ldots & 0 \\ \vdots & & \ddots & & 0 \\ 0 & \ldots & 0 & 0 & X_{jK}' \end{bmatrix}, \tag{8}$$

The model given by (7) can then be represented in a unifying model

$$Y_j = X_j'\theta + E_j \tag{9}$$

where $\theta = \left(\theta_1', \ldots, \theta_K'\right)'$ is a $3K \times 1$ vector of parameters, $E_j = (E_{j1}, \ldots, E_{jK})'$ and $E_j \sim N(0, \Sigma)$, where Σ is a $K \times K$ covariance matrix with diagonal entries $\Sigma_{kk} = \sigma_k^2 = \mathrm{var}(E_{jk})$. The off-diagonal entries $\Sigma_{kl} = \sigma_{kl} = \mathrm{cov}(E_{jk}, E_{je})$ accommodate a conditional dependence (given X_{jk}, X_{jj}, B) between the responses from the same individual.

Following the separate fit of the K linear models at Stage I, we have estimates $\left(\hat{\theta}_k, \hat{\sigma}_k^2\right), k = 1, 2, \ldots, K$. The covariance term characterizing the dependence between errors are then estimated to facilitate estimation of a robust covariance matrix characterizing the dependence between the Stage I estimators $\left(\hat{\theta}_1, \ldots, \hat{\theta}_K\right)$. The elements of interest in $\hat{\theta}$ are the parameter estimates $\hat{B}_1, \ldots, \hat{B}_K$, which are consistent for $B_1, \ldots, B_K$, respectively. In the second stage of estimation, these estimates are pooled over to obtain a single estimate of the global measure of the causal effect denoted by β in (3)

We begin the second stage by estimating the covariance between the errors $\mathrm{cov}(E_{jk}, E_{jl}) = \sigma_{kl}, I \neq k = 1, \ldots, K$, characterizing the dependence between the Stage I estimators $\hat{\theta}_1, \ldots, \hat{\theta}_K$. The challenge in estimating the covariance between the errors $\mathrm{cov}(E_{jk}, E_{jl}) = \sigma_k, I \neq k = 1, \ldots, K$ is that not all responses were observed for all children in the study. We assume that the responses are missing at random (MAR) (Little & Rubin, 2019). To accommodate the fact that not all individuals contribute data for all responses, we introduce the indicators $R_{jk} = I(Y_{jk} \text{ is observed }), k = 1, \ldots, K$.

Specifically, if $S_{jk}(\theta_k) = X_{jk}\left(Y_{jk} - X_{jk}'\theta_k\right)$ is the desired contribution from individual j to the score function for θ_k given B, the observed data score equation for estimating θ_k at Stage I can be written as

$$S_k(\theta_k) = \sum_{j=1}^{J} R_{jk} S_{jk}(\theta_k) = 0, \tag{10}$$

the solutions to which are

$$\hat{\theta}_k = \sum_{j=1}^{J} R_{jk}\left(X_{jk} X_{jk}^{'}\right)^{-1} X_{jk} Y_{jk}, k = 1, \ldots, K. \tag{11}$$

Then we can obtain the maximum likelihood estimate of σ_k^2 in the presence of partially observed outcomes as

$$\widehat{\mathrm{var}}(E_{jk}) = \hat{\sigma}_k^2 = \frac{\sum_{j=1}^{J} R_{jk}\left(Y_{jk} - X_{jk}^{'}\hat{\theta}_k\right)^2}{n_k}, \tag{12}$$

where $n_k = \sum_{j=1}^{J} R_{jk}$ is the number of individuals contributing to the estimation of $\theta_k, k = 1,2, \ldots, K$. Similarly, we also obtain the maximum likelihood covariance estimate as

$$\widehat{\mathrm{cov}}(E_{jk}, E_{jl}) = \hat{\sigma}_{kl} = \frac{\sum_{j=1}^{J} R_{jk} R_{jl}\left(Y_{jk} - X_{jk}^{'}\hat{\theta}_k\right)\left(Y_{jl} - X_{jl}^{'}\hat{\theta}_l\right)}{n_{kl}}, \tag{13}$$

where $n_{\mathrm{kl}} = \sum_{j=1}^{J} R_{\mathrm{jk}} R_{\mathrm{j}}$, which is consistent under a missing at random assumption (Little & Rubin, 2019).

We then let $\hat{\Sigma}$ denote the estimated covariance matrix for the errors where σ_{k}^2 and σ_{kl} are replaced by (12) and (13), respectively.

3.2 | Stage II: Synthesis across responses within a cohort

To consider the synthesis of estimators across all responses, we note that

$$E\left(\hat{B}_k - \beta\right) = E\left\{\left(\hat{B}_k - B_k\right) + (B_k - \beta)\right\} = 0,$$

so $\hat{B}$ is composed of K dependent unbiased estimators of β. Thus,

$$\hat{B} \sim \mathrm{MVN}\left(\mu\left(\beta\right), \Psi\left(\phi\right)\right) \tag{14}$$

asymptotically, where $\mu(\beta)$ is a $K \times 1$ vector with each element equal to β and $\hat{\Psi}\left(\phi\right) = J^{-1}\hat{\Gamma} + \Delta\phi$ denotes the unconditional covariance matrix for $\hat{B}$ where $\Psi_{kk}\left(\phi\right) = \mathrm{var}\left(\hat{B}_k\right) = J^{-1}\Gamma_{kk} + \phi, k = 1, \ldots, K, \Psi_{kl}\left(\phi\right) = \mathrm{cov}\left(\hat{B}_k, \hat{B}_l\right) = J^{-1}\Gamma_{kl} k \neq I = 1, \ldots, K, \Delta$ is a $K \times K$ identity matrix and ϕ reflects the extent of heterogeneity of the response-specific exposure effects for a single cohort. We provided detailed derivation of the covariance matrix for $\hat{B}$ in Appendix B.

Then, we specify a pseudo-likelihood $PL(\beta, \phi)$ for (β, ϕ) given by

$$PL(\beta, \phi) \propto \frac{1}{(2\pi)^{\frac{K}{2}}\sqrt{\hat{\Psi}(\phi)\mid}} \exp\left(-\frac{1}{2}(\hat{B} - \mu(\beta))'\hat{\Psi}^{-1}(\phi)(\hat{B} - \mu(\beta))\right). \tag{15}$$

Note that (15) could be maximized with respect to (β, ϕ), but we proceed in a computationally convenient iterative approach based on (15). Given an estimate $\phi^{(r)}$, we compute an estimate $\beta^{(r)}$ based on a linear combination of $\hat{B}_1, \ldots, \hat{B}_K$. The most efficient linear estimator of β has the form

$$\hat{\beta} = \left[\mathbb{1}'\left[\Psi\left(\phi\right)\right]^{-1}\hat{B}\right] / \left[\mathbb{1}'[\Psi(\phi)]^{-1}\mathbb{1}\right]. \tag{16}$$

Here, we replace $\Psi(\phi)$, the covariance matrix of $\hat{B}$, with an estimator $\hat{\Psi}\left(\phi^{(r)}\right) = J^{-1}\hat{\Gamma} + \Delta\phi^{(r)}$. We could invert $\hat{\Psi}\left(\phi^{(r)}\right)$, but in practice, it may be difficult and while a generalized inverse could be used, the weights resulting from this approach were often found to vary greatly in magnitude and even in sign. A more stable linear estimate using inverse variance weights, whereby we replace $\Psi(\phi)$ with $\mathrm{diag}\left(J^{-1}\hat{\Gamma}_{kk} + \phi^{(r)}, k = 1, \ldots, K\right)$ in (16) to obtain $\hat{\beta}^{(r)}$. We then maximize $PL\left(\hat{\beta}^{r}, \phi\right)$ with respect to ϕ to obtain $\phi^{(r+1)}$, with which we recompute $\hat{\beta}^{(r+1)}$ and repeat iteratively until convergence; we let $\left(\hat{\beta}, \hat{\phi}\right)$ denote the estimates upon convergence. A robust variance estimate is then obtained for $\hat{\beta}$ based on $\hat{\Psi}\left(\hat{\phi}\right)$, which is given by $\widehat{\mathrm{var}}\left(\hat{\beta}\right) = \left[\mathbb{1}'\hat{\Psi}\left(\hat{\phi}\right)\mathbb{1}\right]^{-1}$.

3.3 | An alternative one-stage (fully specified multivariate) approach

The parameters estimated in Sections 3.1 and 3.2 can alternatively be fitted in one step via software for fitting hierarchical mixed effect linear models. To do so, we define $K - 1$ covariates $T_{jkr} = I(k = r), r = 2, \ldots, K$, which indicates the response $k, k = 2, \ldots, K$. We may put these $K - 1$ indicators in vector format and define the $(K - 1) \times 1$ vector $T_{jk} = (T_{jk2}, \ldots, T_{jkk})'$. We consider the first outcome as the reference type and let $T_{ik2} = 1$ for the second outcome, $T_{ik3} = 1$ for the third endpoint, and so on. Then we fit the model

$$Y_{jk} = \alpha + B_k A_j + \gamma S_j + \tau' T_{jk} + E_{jk}, k = 1, \ldots, K,$$

where $\tau = (\tau_2, \ldots, \tau_K)'$ is $(K - 1) \times 1$ vector with $\tau_r = \alpha_r - \alpha_1$ and $r = 2, \ldots, K$. We assume $B_k \sim N(\beta, \phi)$ as specified in (3) where B_k is the effect of a one-unit increase in A_j on the mean of response k given the propensity score S_j, β is the parameter of ultimate interest representing the "average causal effect" of a one-unit increase in the exposure across all responses within the cohort, and ϕ characterizes the degree of heterogeneity in the effect across responses. We also assume $E_j = (E_{j1}, \ldots, E_{jK})'$ is a $K \times 1$ error term with $E_j \sim \mathrm{MVN}(0, \Sigma)$ with Σ a $K \times K$ covariance matrix as in Section 3.1. The one-step approach involves simultaneous estimation of all fixed effects, β, ϕ, and Σ at once. This can be fitted using software for fitting hierarchical linear mixed effects models.

After fitting hierarchical linear mixed effects model for each cohort separately, we let $\widetilde{\beta}_i$ denote the estimate of β_i obtained from fitting the hierarchical model to the data from cohort i and $\widetilde{V}_i(\widetilde{\beta}_i)$ denote the corresponding variance estimate based on the observed information matrix, $i = 1, \ldots, l$.

4 | SYNTHESIS ACROSS COHORTS

In the previous section, we described methods for synthesizing data across multiple outcomes to obtain estimates of the global causal effect using a two-stage approach based on fitting a hierarchical mixed effect model. These methods were based on analysing data from a single cohort. Here, we describe how to combine cohort-specific estimates to obtain an overall estimate of a causal effect while accommodating possible heterogeneity. The approach described in Section 3.2 is an extension of the approach described by Viechtbauer and implemented in the metafor package (Viechtbauer, 2010) which deals with independent estimates; Section 3.2 adapted the methods to deal with dependent effect estimates so what follows is a simplification of the approach for the last stage of the data synthesis. We describe it briefly as follows:

We consider β_i as the global causal effect of exposure in cohort i reflecting the impact of an increment in the volume of prenatal alcohol exposure on the common underlying construct; we let $\hat{\beta}_i$ be the corresponding estimate. Note that the studies draw individuals from different populations and so the composition of the samples varies across cohorts. Moreover, the methods used to measure exposure and the specific outcome measures differ between studies, even though they were measuring the same latent attributes regarding cognition. We therefore wish to accommodate a component of variation between studies (heterogeneity) for the true effects which we accomplished by use of a random effects model of the form

$$\hat{\beta}_i = \beta_i + \epsilon_i, \tag{17}$$

$$\beta_i = \beta_\circ + u_i, \tag{18}$$

where we let $\epsilon_i \sim N\left(0, \hat{V}_i\left(\hat{\beta}_i\right)\right)$ reflect the sampling variation of the estimator from cohort i about the true effect β_i, and $u_i \sim N\left(0, \eta^2\right)$ reflects the heterogeneity of the global cohort-specific causal effects across studies. The parameter $\beta_\circ$ represents the overall global effect, which is the parameter of ultimate interest. Through this variance decomposition then upon introducing the heterogeneity between studies, we have $\widehat{\text{var}}\left(\hat{\beta}_i\right) = \hat{V}_i\left(\hat{\beta}_i\right) + \eta^2$. The synthesis is achieved in a similar spirit to Section 3.2 whereby we consider a pseudo-likelihood of the form

$$PL(\beta_\circ, \eta) \propto \prod_{i=1}^{I}\left\{\frac{1}{(2\pi)^{I/2}\sqrt{\left(\hat{V}_i\left(\hat{\beta}_i\right)+\eta^2\right)}}\exp\left(-\frac{\left(\hat{\beta}_i-\beta_\circ\right)^2}{2\left(\hat{V}_i\left(\hat{\beta}_i\right)+\eta^2\right)}\right)\right\}. \tag{19}$$

The pooled exposure effect estimate $\hat{\beta}_\circ$ is obtained as a weighted average of the $\hat{\beta}_i$ terms with cohort weights equal to the inverse of $\hat{V}\left(\hat{\beta}_i\right) + \hat{\eta}^2$ where $\hat{\eta}^2$ is obtained as the solution to iteratively maximizing (19). The R package “metafor” can be used to carry out this final stage of the data synthesis. If the linear model of Section 3.3 is used for simultaneous estimation of the overall causal effect, then $\tilde{\beta}_i$ and $\tilde{V}_i\left(\tilde{\beta}_i\right)$ can be used in a similar fashion to obtain the estimator $\tilde{\beta}_0$.

5 | SIMULATION STUDIES

For the simulation studies, we consider k correlated continuous outcomes from a single study. We generated outcomes from the following linear regression model:

$$Y_{jk} = \alpha_k + \beta_k X_j + \gamma_k Z_j + E_{jk}, \tag{20}$$

where Y_{jk} is the random variable representing response k for individual $j, k = 1, \ldots, K, j = 1, \ldots, J$, and β_k is the effect of a one-unit increase in X_j on the mean for response k given the covariate Z_j. We let β_k vary about some average exposure effect within a study, with $\beta_k \sim N\left(\beta, \tau^2\right)$. The parameter γ_k characterizes the effect of the covariate for a given level of exposure, X_j. We also assume $E_j = (E_{j1}, \ldots, E_{jK})$ is a $K \times 1$ error term with $E_j \sim MVN(0, \Sigma)$ with Σ a $K \times K$ covariance matrix. The simulations were performed in different scenarios. We generated the effect size for the exposure, β_k, from the normal distribution with mean 3 and variance τ^2. Scenarios were created by manipulating the number of outcomes (k) and varying between-study heterogeneity $\left(\tau^2\right)$. We consider the scenarios in which the number of outcomes is equal to 3, 5, and 10 and τ^2 takes the values 0.10, 0.25, and 0.50. For each combination of the simulation parameters, we generated 1000 datasets with the sample size of 500 for each outcome. For each dataset, we performed the two types of meta-analysis, that is, the one based on the proposed approach versus the full multivariate analysis.

We evaluated the performance of the proposed approach in simulation settings previously described, over 1000 iterations. The estimates of interest were the average exposure effect. To allow for a comprehensive comparison, performance was assessed on a range of metrics: empirical mean bias (EBIAS), average model based standard error (ASE), empirical standard error (ESE), and coverage probability (CP). The results are summarized in Table 1 and some interesting patterns emerge.

Overall, both methods performed well in terms of empirical mean bias. There was a suggestion that one-stage method had lower bias when τ^2 was very low, though both methods had low bias in this case as well. As τ^2 increased, the bias seems to increase for the one-stage method. The similarity between ASE and ESE suggests that inference is working well for the hierarchical meta analysis method, over a wide range of scenarios (including numbers of outcomes and the value of τ). Coverage probability for the proposed method was about the nominal 0.95 for all scenarios considered in this paper. However, the coverage probabilities were less reliable for the one-stage method. There was no particularly clear pattern, though a sense that things became a bit more unstable as τ^2 increased. Empirical standard errors tend to be slightly larger than the average model-based standard errors, especially as τ^2 becomes larger. This most likely reflects the limitations in standard mixed model software.

Overall, the patterns seen in Table 1 reflect the classic variance/bias tradeoff. Our proposed approach has an appealing robustness; however, the cost of this robustness is a slight increase in standard errors in settings (small τ^2) when the one-stage method is working well.

6 | PRENATAL ALCOHOL EXPOSURE AND COGNITIVE FUNCTION IN CHILDREN

We now come back to our motivating application that involves data from six longitudinal cohort studies to assess the effects of PAE on intelligence quotient (IQ), which is a measure of cognitive function. The proposed hierarchical meta-analytic approach is well suited to assess the effect of PAE on IQ measure because it enables us to pool data from diverse, correlated outcomes across cohorts. Table 2 lists the tests used to measure IQ that are considered in this paper along with the summary statistics. As it has been shown in Table 2, these six cohorts used different IQ tests including the Wechsler Intelligence Scale for Children (WISC), Stanford–Binet Intelligence Scales, Kaufman Assessment Battery for Children, and Differential Ability Scales (DAS). Together, all these subtests provide a comprehensive assessment of child’s IQ.

To yield sufficiently precise estimates of effect sizes, we considered a broad set of potential confounders when fitting separate linear models for each outcome. Because each cohort provided a somewhat different set of control variables, we employed a propensity score approach to adjust for potential confounders (Akkaya Hocagil et al., 2021). We estimated the propensity score for each cohort separately and included the propensity score in the linear model as an additional covariate as in model (5).

For each outcome in each study, the effect of alcohol was estimated from model (5). Table 3 lists the estimated effect size and standard errors from the first stage of the hierarchical meta-analytic approach. With the exception of WISC Freedom from Distractibility, and Kaufman ABC simultaneous processing in Detroit and Atlanta Cohort 1, respectively, none of the effects of PAE on IQ were statistically significant. The aim of the second stage of the proposed methods was to pool the estimates of the PAE and estimate the cohort specific overall true mean effect β_i while adjusting for the fact that outcomes are correlated within a cohort and accommodating incomplete information on some outcomes. Table 4 shows the estimated effect sizes and standard errors for each cohort. Table 4 also shows the estimated effect sizes for each cohort obtained from the fully specified multivariate model that was constructed using SAS procedure “proc mixed.” The two methods provided impressively similar estimates for effect sizes and the standard errors. Although the difference was not substantial, these two methods did provide slightly different estimates for the between outcomes heterogeneity $\left(\tau^2\right)$ and there was no observed heterogeneity between outcomes in Seattle and the two Pittsburgh cohorts.

To combine the independent effect size estimates across cohorts and obtain a global effect size estimate of PAE on IQ at age 7 years, we used the R package “metafor” to pool the estimates resulted from our hierarchical meta-analysis and from the fully specified multivariate model (Table 5). For the completeness, we also conducted conventional meta-analysis, which ignores the correlation among outcomes and synthesizes information across cohorts in one step. The resulting global effect sizes from our hierarchical meta-analysis and the full multivariate model were almost identical with similar estimates of the heterogeneity parameter τ^2. Use of the conventional meta-analytic approach (ignoring the dependence across the outcomes within cohorts) led to a larger effect size estimate and a more conservative standard error. Thus, ignoring the dependence across outcomes alters the final pooled estimate and the associated standard error and provides a very different impression of the extent of the between cohort heterogeneity.

7 | DISCUSSION

In this paper, we propose the use of a hierarchical meta-analysis to synthesize data on multiple outcomes from multiple studies. The studies we included were conducted independently and used different neuropsychological test batteries to assess IQ and the same domains of cognitive function. The approach was particularly helpful in terms of synthesizing data across diverse outcomes within each cohort. Furthermore, by including multiple correlated responses from each child, the analyses make full use of available data to maximize the efficiency of estimation and enhance power of associated tests for effects. Robust variance estimation ensures valid inferences at each stage of the analysis.

Our hierarchical meta-analysis consists of three stages. In the first stage, we obtain effect size estimates for each outcome separately, while adjusting for control variables via a propensity score approach. In the second stage, we obtain cohort-specific pooled effect size estimates while adjusting for between-outcome correlation and incomplete data. In the last stage, we combine effect sizes across the cohorts employing a random-effects model. Our procedure follows the same steps as conventional methods for two-stage IPD analysis for making inferences about the effect size but extends these analyses by accounting for the correlation between outcomes and by accommodating incomplete data on some outcomes.

Our approach has several advantages over the one-stage IPD meta-analysis. First, it builds upon the two-stage IPD meta-analysis that practitioners are already familiar with. Second, with our approach, one can create forest plots to visualize the estimated effect sizes for each outcome. Third, our approach is less likely to encounter convergence problems compared with the one-stage IPD meta-analysis. Finally, our approach uses the known within study variances, which helps to provide more precise estimates.

We evaluated and compared our approach with a fully specified multivariate analysis. Previous studies evaluated this question in different settings. Olkin and Sampson (1998) showed that in the case of comparing multiple treatments and a control with respect to a continuous outcome, the traditional meta-analysis based on estimated treatment contrasts is equivalent to the least squares regression analysis of individual patient data if there are no study-by-treatment interactions and the error variances are constant across trials. Mathew and Nordstrom (1999) claimed that the equivalence holds even if the error variances are different across trials. Empirically, meta-analysis using original data has been found to be generally similar but not identical to meta-analysis using summary statistics. Whitehead (2002) and Lin and Zeng (2010) showed that for all commonly used parametric and semiparametric models, there is no asymptotic efficiency gain by analysing original data if the parameter of main interest has a common value across studies, the nuisance parameters have distinct values among studies, and the summary statistics are based on maximum likelihood. More recently, Kontopantelis (2018) conducted a comprehensive simulation study to compare one-stage and two-stage IPD analysis and concluded that a fully specified one-stage model is preferable especially when investigating interactions. We extend the results from these existing studies to the setting in which there are correlated outcomes within multiple cohorts. In simulation scenarios considered in this paper, we observed that the proposed approach can successfully reduce bias relative to the fully specified multivariate approach. Our simulation results suggest that, when the number of outcomes is small and the between outcomes variance is large, our proposed approach outperforms the multivariate analysis.

We illustrated our approach using data on childhood IQ from six cohorts. We included 18 outcomes from these cohorts. We showed how to extend two-stage IPD meta-analysis in the presence of correlated effect size estimates and how to address missing data on outcomes by providing an adjustment formula for the pairwise correlation. With this new approach, one can conduct two-stage IPD meta-analysis in the presence of correlated effect size estimates while taking advantage of visualization via forest plots. Our hierarchical meta-analysis consists of three stages. In the first stage, we obtained effect size estimates for each outcome separately while adjusting for control variables via propensity score approach. In the second stage, we employed our proposed approach to obtain cohort-specific pooled effect size estimates while adjusting for between-outcome correlation and incomplete data. In the last stage, we combined effect sizes across the cohorts employing a random-effects model. We compared the results from our approach with the results from the fully specified multivariate approach and the conventional meta-analysis that ignores the fact that the effect sizes are being combined are dependent. In this comparison, we showed that ignoring within-cohort correlation can markedly alter meta-analysis results in important ways. When we compare our method with the full multivariate approach, our method performed well and thus provides a useful innovative tool for performing and interpreting meta-analyses with the correlated effect sizes. While the proposed approach empirically performs very well in the scenarios we considered in the simulation studies, there may be situations that we did not consider where the performance does not share the simplicity and efficiency as was seen here.

ACKNOWLEDGEMENTS

We thank Neil Dodge, PhD, for his assistance with data management. This research was funded by grants to Sandra W. Jacobson and Joseph L. Jacobson from the National Institutes of Health/National Institute on Alcohol Abuse and Alcoholism (NIH/NIAAA; R01-AA025905) and the Lycaki-Young Fund from the State of Michigan. Louise Ryan received support from the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (Grant No. CE140100049). Richard J. Cook was supported by Natural Sciences and Engineering Research Council of Canada RGPIN 155849 and RGPIN 04207. Data collection for the Detroit cohort was supported by NIAAA/NIH grants R01-AA06966, R01-AA09524, and P50-AA07606 and National Institute on Drug Abuse (NIDA)/NIH grant R21-DA021034; for the Pittsburgh cohorts, by NIAAA/NIH grants R01-AA06390, R01-AA06666, R01-AA14215, and R01-AA18116, National Institute on Child Health and Human Development/NIH grant HD036890, and NIDA/NIH grants R01-DA00090, R01-DA03209, R01-DA03874, and R01-DA17786, R01-DA05460, R01-DA06839, R01-DA08916, and R01-DA12401; for the Atlanta cohorts, by NIAAA/NIH grants R01-AA08105, R01-AA10108, and R01-AA13272, NIDA/NIH grants R01-DA0737362 and R01-DA0737362, and Georgia State Contract DHR 427-93-05050762; and for the Seattle cohort, by NIAAA/NIH grant R01-AA01455.

APPENDIX A: DETAILED DESCRIPTION OF THE SIX COHORT STUDIES

In this appendix, we provide a detailed description of the six cohort studies that we use data from in our application. We specifically provide information on study design, sampling selection, and sample size for each cohort.

Detroit Cohort birth years: 1986–1989

All women ($N > 6400$) enrolling in the antenatal maternity clinic at a large, inner-city hospital were interviewed regarding their alcohol use at their first prenatal visit ($M = 23.4$-week gestation; $SD = 7.9$), using a timeline follow-back interview (Jacobson et al., 2002). Moderate and heavy drinking women were overrepresented in the sample by including all women reporting at least 0.5 oz AA at conception and a random sample of approximately 5% of the lower level drinkers and abstainers. The timeline follow-back interview was repeated at each prenatal clinic visit ($M = 5.4$ visits). To reduce the risk that alcohol might be confounded with cocaine exposure, 78 heavy cocaine (<2 days/week), light alcohol (<7 drinks/week) users were also included in the final sample, which consisted of 480 pregnant women and their children. Participants were followed up at 6.5, 12, and 13 months and 7, 14, and 19 years.

Pittsburgh Cohort 1 birth years: 1983–1986

Participants were recruited from the prenatal clinic at a maternity hospital if they were English speaking, age 18 or older, and in their fourth or fifth gestational month.

The birth sample consisted of 763 live singleton infants. The alcohol, tobacco, and drug use interview was repeated in the seventh gestational month and at delivery, when second and third trimester substance use information was obtained. The cohort consisted of women who were pre-dominantly low income and of fairly equal numbers of Caucasian and African American women. Participants were followed up at 8 and 18 months, and 3, 6, 10, 14, 16, and 22 years.

Pittsburgh Cohort 2 birth years: 1988–1993

English-speaking women in their fourth or fifth month of pregnancy attending the prenatal clinic at a large inner-city hospital who were 18 years old or older were interviewed regarding their usual, maximum, and minimum consumption of cocaine, alcohol, marijuana, tobacco, and other drugs prior to pregnancy and during the first trimester. Every woman who reported any cocaine/crack use during the first trimester was enrolled in the study cohort, as was the next woman interviewed who reported no cocaine or crack use during both the year prior to pregnancy and the first trimester. Although crack/cocaine use was the criterion for recruitment, a large proportion of these women also drank moderate-to-heavy levels of alcohol. The alcohol and drug use interview was repeated at the end of the second and third trimesters, and offsprings were assessed at 1, 3, 7, 10, 15, and 21 years. The birth cohort consisted of 295 women and infants; the women were predominantly of low socio-economic status and were roughly equally divided by Caucasian and African American race.

Atlanta Cohort 1 birth years: 1980–1986

Five hundred twenty-seven low socioeconomic status (SES), pregnant women were recruited at their first prenatal visit at an urban Atlanta hospital serving a primarily African American, low income population. Women who reported drinking at least 1 oz AA/week during pregnancy were recruited. Nondrinkers, who were similar in demographic background, were recruited at the same time to serve as controls. Women were interviewed at recruitment about their alcohol and drug use; the majority reported drinking on weekends in a “binge” pattern. Infants were evaluated following birth. Subsamples were followed up at 6 and 12 months and 7, 14, and 22 years.

Atlanta Cohort 2 birth years: 1992–1994

Three hundred six mothers and their infants were recruited shortly after delivery at an urban Atlanta hospital; 111 reported having drunk alcohol during pregnancy, 71 of whom also had used cocaine (based on self-report or urine screen); 44 used cocaine but no alcohol; 151 did not drink alcohol or use cocaine. All participants were English speaking, 19 years or older, and had singleton births; most were African American and low SES. The infants were assessed at 2 and 8 years.

Seattle Cohort birth years: 1975–1976

All women who were enrolled in prenatal care by the fifth month of pregnancy at two large Seattle hospitals were eligible to participate. To ascertain PAE, participating mothers ($N = 1529$) were administered a Quantity-Frequency-Variability interview (Cahalan & Cisin, 1968) regarding alcohol, tobacco, and drug use for two time periods: during pregnancy and just prior to pregnancy recognition; 462 newborns were selected based on an algorithm derived from maternal absolute alcohol (AA)/day, alcohol use/occasion, volume variability, and frequency of intoxication constructed to overrepresent infants born to heavier drinkers. Controls included both abstainers and light drinkers. Infants were followed up at 8 and 18 months and 4, 7, 11, 14, 21, 25, and 30 years. Although cohort retention was high (e.g., 82%

at 14 years), other children not initially selected whose mothers had been interviewed during pregnancy were added at follow- up assessments to keep the sample size close to 500 at each examination.

APPENDIX B: DERIVATION OF THE COVARIANCE MATRIX FOR $\hat{\boldsymbol{B}}$

In this appendix, we provided detailed derivation a robust covariance matrix characterizing the dependence between Stage I estimators (i.e., $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_K$). This derivation prove the results in Section 3.2 of the main text.

The expression for the covariance between $\hat{\theta}_k$ and $\hat{\theta}_l$ is obtained based on a general formula for robust variance estimation. If $S_{jk}(\theta_k) = X_{jk}\left(Y_{jk} - X_{jk}^{'}\theta_k\right)$ is the desired contribution from individual j to the score function for θ_k given $\boldsymbol{B}$, the observed data score equation for estimating θ_k at Stage I can be written as

$$S_k(\theta_k) = \sum_{j=1}^{J} R_{jk}S_{jk}(\theta_k) = 0, \tag{B1}$$

the solutions to which are

$$\hat{\theta}_k = \sum_{j=1}^{J} R_{jk}\left(X_{jk}X_{jk}^{'}\right)^{-1}X_{jk}Y_{jk}, k = 1, \ldots, K\,. \tag{B2}$$

If we stack the score function in (B1), we obtain $S(\theta) = \left(S_1^{'}(\theta_1), \ldots, S_K^{'}(\theta_K)\right)'$.

Then given $\boldsymbol{B} = (B_1, \ldots, B_K)'$, we note that

$$\sqrt{J}\left(\hat{\theta} - \theta\right)\overset{d}{\sim}\mathrm{MVN}\left(0, \mathscr{A}^{-1}\left(\theta\right)\mathscr{B}\left(\theta\right)\mathscr{A}^{-1}\left(\theta\right)\right) \tag{B3}$$

as $J \to \infty$, where $\mathscr{A}(\theta) = E\{-\partial S(\theta)/\partial\theta'\}$ is a block diagonal $3K \times 3K$ matrix of the form

$$\mathscr{A}(\theta) = \begin{bmatrix} \mathscr{A}_{11}(\theta_1) & 0 & \ldots & 0 \\ 0 & \mathscr{A}_{22}(\theta_2) & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & & \mathscr{A}_{KK}(\theta_K) \end{bmatrix},$$

where the kth 3 × 3 diagonal submatrix is given by

$$\mathscr{A}_{kk}(\theta_k) = E\left\{-\partial S_k(\theta_k)/\partial\theta_k^{'}\right\} = E\left\{\sum_{j=1}^{J} R_{jk}X_{jk}X_{jk}^{'}\right\} = JE\left\{X_{jk}X_{jk}^{'} \mid R_{jk} = 1\right\}P(R_{jk} = 1),$$

$k = 1, \ldots, K$. If we let $\Omega_{kk} = P\left(X_{jk}X_{jk}^{'} \mid R_{jk} = 1\right)$ be a 3 × 3 matrix, we can then write

$$\mathscr{A}_{kk}(\theta_k) = J\Omega_{kk}P(R_{jk} = 1)\,. \tag{B4}$$

Note that $\mathscr{B}(\theta) = E\{S(\theta)S'(\theta)\}$ is also a $3K \times 3K$ matrix. Under the assumption that the response data are missing at random (i.e., $R_{jk} \perp Y_{jk} \mid X_{jk}$), the diagonal elements of $\mathscr{B}(\theta)$ are the covariance matrices of the score functions for θ_k, $\mathscr{B}_{kk}(\theta) = \text{cov}(S_k(\theta_k) \mid B)$, $k = 1, \ldots, K$, where

$$\mathscr{B}_{kk}\left(\theta\right) = E\left\{\sum_{j=1}^{J} R_{jk}S_{jk}(\theta_k)S_{jk}^{'}(\theta_k)\right\} = E\left\{\sum_{j=1}^{J} R_{jk}X_{jk}X_{jk}^{'}\text{var}(E_{jk})\right\},$$

because the error terms are assumed independent of the covariates. This can then be written as

$$\mathscr{B}_{kk}(\theta) = J\Omega_{kk}P(R_{jk} = 1)\sigma_k^2, k = 1, \ldots, K\,. \tag{B5}$$

In a similar fashion, we note that

$$\begin{aligned}\mathscr{B}_{kl}\left(\theta\right) &= \text{cov}(S_k(\theta_k), S_l(\theta_l) \mid B) = E\left\{\sum_{j=1}^{J} R_{jk}R_{jl}E\left\{S_{jk}(\theta_k)S_{jl}^{'}(\theta_l) \mid X_{jk}, X_{jl}, R_{jk} = R_{jl} = 1\right\}\right\}\\ &= E\left\{\sum_{j=1}^{J} R_{jk}R_{jl}X_{jk}X_{jl}^{'}\text{cov}(E_{jk}, E_{jl})\right\}\\ &= JE\left\{X_{jk}X_{jl}^{'} \mid R_{jk} = R_{jl} = 1\right\}P(R_{jk} = R_{jl} = 1)\sigma_{kl}\\ &= J\Omega_{kl}P(R_{jk} = R_{jl} = 1)\sigma_{kl},\end{aligned}$$

where $\Omega_{\text{kl}} = E\left\{X_{\text{jk}}X_{jl}^{'} \mid R_{\text{jk}} = R_{\text{jl}} = 1\right\}$ is a 3 × 3 matrix. If $X_{\text{jk}} = X_{jl}$ as in this setting, this becomes

$$\mathscr{B}_{kl}(\theta) = \text{cov}(S_k(\theta_k), S_l(\theta_l) \mid B) = J\Omega_{kk}P(R_{jk} = R_{jl} = 1)\sigma_{kl} \tag{B6}$$

because $\Omega_{kk} = \Omega_{kl} = \Omega$ for all $k \neq l$.

If we wish to estimate the covariance of $\sqrt{J}\left(\hat{\theta}_k - \theta_k\right)$ and $\sqrt{J}\left(\hat{\theta}_l - \theta_l\right)$ given B, we note that this has the general form

$$\text{cov}\left(\sqrt{J}\left(\hat{\theta}_k - \theta_k\right), \sqrt{J}\left(\hat{\theta}_l - \theta_l\right) \mid B\right) = \mathscr{A}_{kk}^{-1}(\theta)B_{kl}(\theta)\mathscr{A}_{ll}^{-1}(\theta)\,.$$

Inserting the derived expressions gives the (k, l), 3 × 3 submatrix of the full covariance matrix in (B3) as

$$\text{cov}\left(\sqrt{J}\left(\hat{\theta}_k - \theta_k\right), \sqrt{\text{J}}\left(\hat{\theta}_l - \theta_l\right) \mid B\right) = \frac{\sigma_{kl}\Omega^{-1}P(R_{jk} = R_{jl} = 1)}{P(R_{jk} = 1)P(R_{jl} = 1)}\,. \tag{B7}$$

We estimate (B7) as follows. Because $X_{jk} = X_j$ is available for all individuals, we estimate $\Omega = \Omega_{kk} = \Omega_{kl}$simply as $\hat{\Omega} = \sum_{j=1}^{J} \left(X_{jk} X_{jk}^{'}\right)/J$. Moreover, we estimate $P(R_{jk} = R_{jl} = 1)$ empirically as $\hat{P}(R_{jk} = R_{jl} = 1) = n_{kl}/J$ where $n_{kl} = \sum_{j=1}^{J} R_{jk} R_{jl}$, and likewise let $\hat{P}(R_{jk} = 1) = n_k/J$ where $n_k = \sum_{j=1}^{J} R_{jk}, k = 1, \ldots, K$. Replacing unknown quantities with their estimates gives

$$\widehat{\operatorname{cov}}\left(\sqrt{J}\left(\hat{\theta}_k - \theta_k\right), \sqrt{J}\left(\hat{\theta}_l - \theta_l\right) \mid B\right) = \frac{\hat{\sigma}_{kl}}{J^{-1} \sum_{j=1}^{J} \left(X_{jk} X_{jl}^{'}\right)} \frac{J n_{kl}}{n_k n_l}, \tag{B8}$$

where $\hat{\sigma}_{\mathrm{kl}}$ is given by (13).

Let $\mu(\beta) = \mathbb{1}\beta$ where $\mathbb{1}$ is a $K \times 1$ vector of ones and β is a scalar. We then let $\operatorname{cov}\left\{\sqrt{J}\left(\hat{B} - \mu\left(\beta\right)\right) \mid B\right\} = \Gamma$, where Γ is the covariance matrix for $\hat{B}$ obtained by selecting the corresponding elements of (B7) related to the coefficients of the exposure variable in the K marginal least squares estimates. We aim to use $\operatorname{cov}\left(\sqrt{J}\left(\hat{B} - \mu\left(\beta\right)\right) \mid B\right)$ to combine the estimates across all responses, but we note there is an additional component of variation in the estimators of the exposure effects because the B_k terms are themselves independent and identically distributed (i.e., $B_{ik} \sim N(\beta_i, \phi_i)$).

Thus, $\operatorname{cov}\left(\hat{B} \mid B\right) = J^{-1}\Gamma$, where Γ is a $K \times K$ matrix with diagonal elements Γ_{kk} and off-diagonal elements $\Gamma_{kl}, I = 1, \ldots, K, k = 1, \ldots, K,$

$$\operatorname{var}\left(\hat{B}_k\right) = J^{-1}\Gamma_{kk} + \phi, k = 1, \ldots, K, \tag{B9}$$

and because $B_k \perp B_{\mathrm{l}}$,

$$\operatorname{cov}\left(\hat{B}_k, \hat{B}_l\right) = J^{-1}\Gamma_{kl}, k \neq I = 1, \ldots, K\,. \tag{B10}$$

We denote the unconditional covariance matrix for $\hat{B}$ as $\operatorname{cov}\left(\hat{B}\right) = \Psi\left(\phi\right) = J^{-1}\Gamma + \Delta\phi$ where $\Psi_{kk}(\phi)$ is given by (B9), $\Psi_{kl}(\phi)$ is given by (B10), and Δ is a $K \times K$ identity matrix. Given an estimate of ϕ, we estimate this covariance matrix by

$$\widehat{\operatorname{Cov}}\left(\hat{B}\right) = J^{-1}\hat{\Gamma} + \Delta\hat{\phi} = \hat{\Psi}\left(\hat{\phi}\right).$$

REFERENCES

Akkaya Hocagil T, Cook RJ, Jacobson SW, Jacobson JL, & Ryan LM (2021). Propensity score analysis for a semi-continuous exposure variable: A study of gestational alcohol exposure and childhood cognition. Journal of the Royal Statistical Society: Series A (Statistics in Society), 184, 1390–1413. 10.1111/rssa.12716

Axelrad DA, Bellinger DC, Ryan LM, & Woodruff TJ (2007). Dose-response relationship of prenatal mercury exposure and IQ: An integrative analysis of epidemiologic data. Environmental Health Perspectives, 115(4), 609–615. [PubMed: 17450232]

Brown JV, Bakeman R, Coles CD, Sexson WR, & Demi A (1998). Maternal drug use during pregnancy: Are preterm and full-term infants affected differently? Developmental Psychology, 34, 540–554. [PubMed: 9597363]

Cahalan D, & Cisin IH (1968). American drinking practices: Summary of findings from a national probability sample. I. Extent of drinking by population subgroups. Quarterly Journal of Studies on Alcohol, 29(1), 130–151. 10.15288/qjsa.1968.29.130

Carter RC, Jacobson JL, Molteno CD, Dodge NC, Meintjes EM, & Jacobson SW (2016). Fetal alcohol growth restriction and cognitive impairment. Pediatrics, 138(2), e20160775. 10.1542/peds.2016-0775

Cheung MW-L (2013). Multivariate meta-analysis as structural equation models. Structural Equation Modeling, 20(3), 429–454.

Cheung MW-L (2019). A guide to conducting a meta-analysis with non-independent effect sizes. Neuropsychology Review, 29, 387–296. [PubMed: 31446547]

Coles C, Platzman K, Raskind-Hood C, Brown R, Falek A, & Smith I (2006). A comparison of children affected by prenatal alcohol exposure and attention deficit, hyperactivity disorder. Alcoholism: Clinical and Experimental Research, 21, 150–161.

Day N, Sambamoorthi U, Taylor P, Richardson G, Robles N, Jhon Y, Scher M, Stoffer D, Cornelius M, & Jasperse D (1991). Prenatal marijuana use and neonatal outcome. Neurotoxicology and Teratology, 13(3), 329–334. [PubMed: 1886543]

Hedges LV, Tipton E, & Johnson MC (2010). Robust variance estimation in meta-regression with dependent effect size estimates. Research Synthesis Methods, 1(1), 39–65. [PubMed: 26056092]

Hoyme HE, May PA, Kalberg WO, Kodituwakku P, Gossage JP, Trujillo PM, Buckley DG, Miller JH, Aragon AS, Khaole N, Viljoen DL, Jones KL, & Robinson LK (2005). A practical clinical approach to diagnosis of fetal alcohol spectrum disorders: Clarification of the 1996 Institute of Medicine Criteria. Pediatrics, 115(1), 39–47. [PubMed: 15629980]

Jacobson JL, Akkaya-Hocagil T, Ryan LM, Dodge NC, Richardson GA, Olson HC, Coles CD, Day NL, Cook RJ, & Jacobson SW (2021). Effects of prenatal alcohol exposure on cognitive and behavioral development: Findings from a hierarchical meta-analysis of data from six prospective longitudinal U.S. cohorts. Alcoholism: Clinical and Experimental Research, 45, 2040–2058. 10.1111/acer.14686 [PubMed: 34342030]

Jacobson JL, Jacobson SW, Sokol RJ, Martier SS, Ager JW, & Kaplan-Estrin MG (1993). Teratogenic effects of alcohol on infant development. Alcoholism: Clinical and Experimental Research, 17(1), 174–183. 10.1111/j.1530-0277.1993.tb00744.x [PubMed: 8452200]

Jacobson SW, Chiodo LM, Sokol RJ, & Jacobson JL (2002). Validity of maternal report of prenatal alcohol, cocaine, and smoking in relation to neurobehavioral outcome. Pediatrics, 109(5), 815–825. [PubMed: 11986441]

Jacobson SW, Jacobson JL, Sokol RJ, Chiodo LM, & Corobana R (2004). Maternal age, alcohol abuse history, and quality of parenting as moderators of the effects of prenatal alcohol exposure on 7.5-year intellectual function. Alcoholism: Clinical and Experimental Research, 28(11), 1732–1745. 10.1097/01.ALC.0000145691.81233.FA [PubMed: 15547461]

Jacobson SW, Stanton ME, Molteno CD, Burden MJ, Fuller DS, Hoyme HE, Robinson LK, Khaole N, & Jacobson JL (2008). Impaired eye-blink conditioning in children with fetal alcohol syndrome. Alcoholism: Clinical and Experimental Research, 32(2), 365–372. 10.1111/j.1530-0277.2007.00585.x [PubMed: 18162064]

Konstantopoulos S (2011). Fixed effects and variance components estimation in three-level meta-analysis. Research Synthesis Methods, 2(1), 61–76. [PubMed: 26061600]

Kontopantelis E (2018). A comparison of one-stage vs two-stage individual patient data meta-analysis methods: A simulation study. Research Synthesis Methods, 9(3), 417–430. 10.1002/jrsm.1303 [PubMed: 29786975]

Lin DY, & Zeng D (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. Biometrika, 97(2), 321–332. 10.1093/biomet/asq006 [PubMed: 23049122]

Little RJA, & Rubin DB (2019). Statistical analysis with missing data, 3rd edition. Hoboken, NJ: John Wiley & Sons.

Mathew T, & Nordstrom K (1999). On the equivalence of meta-analysis using literature and using individual patient data. Biometrics, 55(4), 1221–1223. [PubMed: 11315071]

Mattson SN, Bernes GA, & Doyle LR (2019). Fetal alcohol spectrum disorders: A review of the neurobehavioral deficits associated with prenatal alcohol exposure. Alcoholism: Clinical and Experimental Research, 43(6), 1046–1062. 10.1111/acer.14040 [PubMed: 30964197]

Olkin I, & Sampson A (1998). Comparison of meta-analysis versus analysis of variance of individual patient data. Biometrics, 54(1), 317–322. [PubMed: 9544524]

Richardson GA, Hamel SC, Goldschmidt L, & Day NL (1999). Maternal drug use during pregnancy: Are preterm and full-term infants affected differently? Pediatrics, 104, 540.

Riley RD, Abrams KR, Sutton AJ, Lambert PC, & Thompson JR (2007). Bivariate random-effects meta-analysis and the estimation of between-study correlation. BMC Medical Research Methodology, 7(1), 3. 10.1186/1471-2288-7-3 [PubMed: 17222330]

Riley RD, Lambert P, & Abo-Zaid GMA (2010). Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ, 340, c221. [PubMed: 20139215]

Riley RD, & Steyerberg EW (2010). Meta-analysis of a binary outcome using individual participant data and aggregate data. Research Synthesis Methods, 1(1), 2–19. 10.1002/jrsm.4 [PubMed: 26056090]

Rosenbaum PR, & Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41–55. 10.1093/biomet/70.1.41

Schuetze P, Eiden RD, & Coles CD (2007). Prenatal cocaine and other substance exposure: Effects on infant autonomic regulation at 7 months of age. Developmental Psychobiology, 49(3), 276–289. 10.1002/dev.20215 [PubMed: 17380506]

Simmonds MC, & Higgins JPT (2007). Covariate heterogeneity in meta-analysis: Criteria for deciding between meta-regression and individual patient data. Statistics in Medicine, 26(15), 2982–2999. 10.1002/sim.2768 [PubMed: 17195960]

Stratton K, Howe C, & Battaglia FC (1996). Fetal alcohol syndrome: Diagnosis, epidemiology, prevention, and treatment. Washington, D.C.: National Academy Press.

Streissguth AP, Martin DC, Martin JC, & Barr HM (1981). The seattle longitudinal prospective study on alcohol and pregnancy. Neurobehav Toxicol Teratol, 2(3), 223–233.

Van den Noortgate W, López - López JA, Marín-Martínez F, & Sánchez-Meca J (2013). Three-level meta-analysis of dependent effect sizes. Behavior Research Methods, 45, 576–594. [PubMed: 23055166]

Van den Noortgate W, López-López J, Marín-Martínez F, & Sanchez-Meca J (2014). Meta-analysis of multiple outcomes: A multilevel approach. Behavior research methods, 47, 1274–1294.

Viechtbauer W (2010). Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36(3), 1–48. https://www.jstatsoft.org/v36/i03/

Whitehead A (2002). Meta-analysis of controlled clinical trials. West Sussex, England: John Wiley & Sons.

TABLE 1

Results of a simulation study assessing the performance of our hierarchical meta-analysis and a full multivariate analysis in estimating the effect size for the exposure in a variety of settings

		Hierarchical meta-analysis				One-stage method			
τ^2	k	**EBIAS**	**ASE**	**ESE**	**ECP (%)**	**EBIAS**	**ASE**	**ESE**	**ECP (%)**
0.10									
	10	0.007	0.16	0.14	96.0	<0.001	0.10	0.14	92.0
	5	0.006	0.20	0.20	96.0	<0.001	0.13	0.14	99.0
	3	0.009	0.23	0.25	95.0	<0.001	0.14	0.14	93.0
0.25									
	10	<0.000	0.11	0.10	95.0	0.020	0.18	0.24	88.0
	5	<0.000	0.13	0.13	95.0	0.030	0.19	0.21	91.0
	3	0.010	0.15	0.15	95.0	0.020	0.21	0.23	94.0
0.50									
	10	0.004	0.16	0.15	95.0	0.040	0.26	0.29	96.0
	5	0.009	0.21	0.20	94.0	0.060	0.27	0.30	97.0
	3	0.010	0.23	0.24	94.0	0.120	0.31	0.34	92.0

TABLE 2

IQ related outcomes assessed at age 7 in the six cohorts

Cohort	Endpoints	Mean (SD)
Detroit (*n* = 336		
	WISC Verbal IQ	87.4 (12.4)
	WISC Performance IQ	83.6 (13.0)
	WISC Freedom from distractibility	93.6 (14.9)
Pittsburgh 1 (*n* = 720)		
	Stanford-Binet Verbal reasoning	100.0 (11.8)
	Stanford-Binet Abstract reasoning	85.0 (13.9)
	Stanford-Binet Quantitative reasoning	98.0 (18.0)
	Stanford-Binet Short-term memory	92.6 (15.2)
Pittsburgh 2 (*n* = 268)		
	Stanford-Binet Verbal reasoning	96.3 (12.2)
	Stanford-Binet Abstract reasoning	88.0 (16.2)
	Stanford-Binet Quantitative reasoning	94.4 (18.4)
	Stanford-Binet Short-term memory	91.4 (15.6)
Atlanta 1 (*n* = 223)		
	Kaufman ABC Simultaneous processing	88.6 (14.1)
	Kaufman ABC Sequential processing	89.1 (14.1)
Atlanta 2 (*n* = 138)		
	DAS Verbal standard score	79.6 (15.2)
	DAS Nonverbal standard score	87.7 (14.9)
	DAS Spatial standard score	81.7 (14.2)
Seattle (*n* = 510)		
	WISC Verbal IQ	106.3 (15.5)
	WISC Performance IQ	107.8 (13.9)

TABLE 3

Estimated outcome-specific IQ-related effects at age 7 from Stage I estimation

Cohort	Response type	Effect size	SE
Detroit	WISC Verbal IQ	−4.2	3.2
Detroit	WISC Performance IQ	−3.7	3.2
Detroit	WISC Freedom from distractibility	−10.3	3.1
Pittsburgh Cohort 1	Stanford Binet Verbal reasoning	−5.8	3.0
Pittsburgh Cohort 1	Stanford Binet Abstract reasoning	−5.0	3.0
Pittsburgh Cohort 1	Stanford Binet Quantitative reasoning	−1.9	3.0
Pittsburgh Cohort 1	Stanford Binet Short term memory	−5.3	3.0
Pittsburgh Cohort 2	Stanford Binet Verbal reasoning	−0.3	3.1
Pittsburgh Cohort 2	Stanford Binet Abstract reasoning	−1.8	3.0
Pittsburgh Cohort 2	Stanford Binet Quantitative reasoning	−1.1	3.1
Pittsburgh Cohort 2	Stanford Binet Short term memory	−3.5	3.1
Atlanta Cohort 1	Kaufman ABC Simultaneous processing	−6.9	2.9
Atlanta Cohort 1	Kaufman ABC Sequential processing	−1.9	2.9
Atlanta Cohort 2	DAS Verbal standard score	−5.9	3.2
Atlanta Cohort 2	DAS Nonverbal standard score	1.7	3.3
Atlanta Cohort 2	DAS Spatial standard score	−0.9	3.3
Seattle	WISC Verbal IQ	−0.5	2.6
Seattle	WISC Performance IQ	−1.9	2.6

TABLE 4

Pooled effect size estimates of prenatal alcohol exposure for each cohort at Stage II

	Hierarchical Approach			Multivariate Approach		
Cohort	**Effect Size**	**SE**	$\hat{\tau}^2$	**Effect Size**	**SE**	$\hat{\tau}^2$
Detroit	−6.1	3.2	7.8	−6.1	3.1	6.2
Pittsburgh Cohort 1	−4.3	2.4	0.0	−4.0	2.6	0.0
Pittsburgh Cohort 2	−1.6	2.5	0.0	−1.6	2.5	0.0
Atlanta Cohort 1	−4.4	3.0	5.1	−4.4	3.4	6.2
Atlanta Cohort 2	−1.9	3.0	7.8	−2.0	3.2	8.2
Seattle	−1.2	2.3	0.0	−1.2	2.3	0.0

TABLE 5

Stage III estimated effect sizes of prenatal alcohol exposure on IQ at age 7

Method	**Global effect size**	**SE**	$\hat{\tau}^2(\mathrm{se})$
Hierarchical meta-analytic approach	−3.2	0.8	1.0 (2.3)
One-stage (full multivariate approach)	−3.1	0.8	0.9 (2.3)
Conventional meta-analysis (ignoring the correlation)	−3.5	0.6	4.7 (2.7)