


Article

Analyzing Treatment Effect by Integrating Existing Propensity Score and Outcome Regressions with Heterogeneous Covariate Sets

Yi-Hau Chen ^{1,*} , Szu-Yuan Hsu ² , Jie-Huei Wang ^{3,*}  and Chien-Chou Su ⁴¹ Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan² The Third Research Division, Chung-Hua Institution for Economic Research, Taipei 10672, Taiwan; sheldon770815@gmail.com³ Department of Mathematics, National Chung Cheng University, Chiayi 62102, Taiwan⁴ Clinical Innovation and Research Center, National Cheng Kung University Hospital, Tainan 70403, Taiwan; chienchou.su@gmail.com

* Correspondence: yhchen@stat.sinica.edu.tw (Y.-H.C.); jhwang@ccu.edu.tw (J.-H.W.); Tel.: +886-2-2783-3727 (Y.-H.C.); +886-5-2720411 (ext. 66113) (J.-H.W.)

Abstract: Analyzing treatment or exposure effect is a major research theme in scientific studies. In the current big-data era where multiple sources of data are available, it is of interest to perform a synthesized analysis of treatment effects by integrating information from different data sources or studies. However, studies may contain heterogeneous and incomplete covariate sets, and individual data therein may not be accessible. We apply and extend the generalized meta-analysis method to integrate summary results (e.g., regression coefficients) of outcome and treatment (propensity score, PS) regression analyses across different datasets that may contain heterogeneous covariate sets. The proposed integrated analysis utilizes a reference dataset, which contains data on the complete set of covariates. The asymptotic distribution for the proposed integrated estimator is established. Simulations reveal that the proposed estimator performs well. We apply the proposed method to obtain the causal effect of waist circumference on hypertension by integrating two existing outcomes and PS regression analyses with different sets of covariates.

Keywords: data integration; multi-center study; missing covariate; treatment effect**MSC:** 62P10; 62J12

Citation: Chen, Y.-H.; Hsu, S.-Y.; Wang, J.-H.; Su, C.-C. Analyzing Treatment Effect by Integrating Existing Propensity Score and Outcome Regressions with Heterogeneous Covariate Sets. *Mathematics* **2024**, *12*, 2265. <https://doi.org/10.3390/math12142265>

Academic Editor: Li-pang Chen

Received: 3 June 2024

Revised: 15 July 2024

Accepted: 16 July 2024

Published: 19 July 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The impact of a treatment or exposure on an outcome is a significant focus in many scientific fields. These include clinical trials [1], education [2], economics [3], and medical [4] fields. In observational studies, treatment assignment often correlates with subject characteristics, potentially leading to systematic differences between treated and control groups. This can result in biased conclusions when directly comparing these groups.

To perform causal inferences in observational studies, Rubin [5] proposed the counterfactual or potential outcome framework. Suppose the treatment assignment Z is binary with $Z = 1$ denoting the treated group and $Z = 0$ the control group, and Y_1 and Y_0 are the potential outcome values a subject would have if he/she had a treatment assignment $Z = 1$ and $Z = 0$, respectively. The outcome Y observed for a subject with treatment Z is then given by

$$Y = ZY_1 + (1 - Z)Y_0.$$

A causal treatment effect for a subject is obtained as the difference of two potential outcomes, $Y_1 - Y_0$. Usually, only one potential outcome can be observed for a subject,

and the causal effect is unobserved. However, under suitable assumptions, the average treatment effect (ATE)

$$E(Y_1 - Y_0) = E(Y_1) - E(Y_0)$$

is estimable from observational studies and has been a popular target of causal inferences. For a binary outcome, the ATE amounts to the marginal risk difference $P(Y_1 = 1) - P(Y_0 = 1)$. A key assumption for estimating ATE with data from observational studies is that the treatment assignment is strongly ignorable [6]:

$$(Y_1, Y_0) \perp Z \mid \mathbf{X},$$

where “ \perp ” stands for statistical independence; that is, the potential outcomes (Y_1, Y_0) , and the treatment assignment Z are independent conditioning on the covariate set \mathbf{X} . In other words, the treatment assignment Z is irrelevant to the values of potential outcomes Y_1 and Y_0 once the covariates \mathbf{X} are controlled. Another key assumption is the stable unit-treatment value assumption (SUTVA), which ensures that $Y_1(Y_0)$ is the unique outcome associated with $Z = 1(Z = 0)$ [6]. Under (1) and SUTVA, consistent estimation of the ATE can then be achieved by the technique of matching, stratification, covariate adjustment (CA), or inverse probability of treatment weighting (IPTW) [6–10].

Rosenbaum and Rubin [6] further proposed using propensity score (PS) for conducting causal inferences. A PS is the probability $P(Z = 1 \mid \mathbf{X})$ of a subject being assigned to the treated group conditioning on his/her observed covariates and is the coarsest balancing score such that the covariate distributions are the same between treatment groups once the score is matched. Accordingly, a PS can replace the full covariate set \mathbf{X} used for matching, stratification, CA, or IPTW to conduct causal inferences on treatment effect when the assumption (1) holds Rosenbaum and Rubin [6]. Since the PS is a univariate minimal sufficient statistic, the PS-based method can also be viewed as an effective dimension-reduction technique, since matching, stratification, CA, or IPTW can be simply performed on a scalar PS rather than on a multi-dimensional covariate vector [6–10]. Since its invention, the PS method has gained tremendous popularity in observational studies; see Figure 1 of Simoneau et al. [11].

In the big data era, there is growing interest in conducting integrated analyses of treatment effects by integrating information from databases across various observational studies [12,13]. However, combining data from different databases presents challenges due to variations in covariate variables, even if they share common outcome and treatment variables, and discrepancies in baseline data distributions. Moreover, stringent data privacy regulations such as the European Union’s General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) pose barriers to directly sharing, exchanging, and synthesizing individual-level data from disparate sources.

Analyzing treatment or exposure effects is a major research theme in scientific studies. In the current big-data era where multiple data sources are available, it is interesting to perform a synthesized analysis of treatment effects by integrating information from different data sources or studies. However, studies may contain heterogeneous and incomplete covariate sets, and individual data may not be accessible. This motivates us to consider a framework for integrated regression analysis that can utilize summary results (e.g., regression coefficients) of outcome and treatment (PS) regression analyses across different datasets that may contain heterogeneous covariate sets.

This study is specifically motivated by two recent investigations into the relationship between waist circumference (WC) and hypertension (HT): Ren et al. [14] and Hu et al. [15]. Each study explored how WC influences the risk of HT while accounting for different sets of covariates (detailed in Section 4). The aim here is to examine the impact of WC on HT while controlling for the comprehensive set of covariates (combining the respective sets from both studies). Specifically, this study integrates the regression analysis results from two existing studies on the relationship between WC and HT, considering their respective sets of covariates. Additionally, it incorporates the findings from propensity score analyses

of WC as a treatment variable in both studies. This integrated approach aims to conduct a synthesized analysis of the causal effect of WC on HT, utilizing a comprehensive covariate dataset sourced from the Taiwan BioBank database. Specifically, we apply and extend the generalized meta-analysis method developed by Kundu et al. [16] to integrate the existing PS and outcome regression analyses with heterogeneous covariate variables using a dataset on the complete covariate set, which provides a reference to the covariate distribution and is termed “reference dataset” by Kundu et al. [16]. We obtain consistent estimates for the parameters of the full PS and outcome regression models and for the ATE. Asymptotic theory for the estimators of the regression parameters and the ATE is also established. The proposed integrated analysis allows researchers to integrate summary results from existing outcome and PS regression analyses with heterogeneous covariates sets, and hence to obtain enhanced statistical efficiency and power from existing studies.

In Section 2, we describe the proposed estimation methods for the full PS and outcome regression models and the ATE. In Section 3, we report simulation results to show the satisfactory performance of the proposed method. Section 4 presents the empirical study on the effect of waist circumference on hypertension based on two existing studies and a reference sample of covariate data. Finally, concluding discussions are provided in Section 5.

2. The Proposed Method

In this section, we present an inferential framework that integrates results from existing studies and results from treatment (i.e., propensity score, PS) regression analyses, in which different and incomplete covariates may be present. We use the reference set on the full set of covariates to convert existing outcome and PS regression analyses into consistent estimates of the full outcome and PS regression model based on the full set of covariates, which in turn generate causal inferences about treatment effects.

2.1. The Existing Outcome and PS Regression Analyses

Suppose that there exist K independent studies, where the k th study contains independent data on $\{Y, Z, \mathbf{X}_{(k)}\}$ of size $n_k, k = 1, 2, \dots, K$. Here, Y and Z are, respectively, the common outcome and binary treatment indicator of interest in the K studies, and $\mathbf{X}_{(k)}$ is the covariate set considered in study k and is some subset of \mathbf{X} , the complete covariate set satisfying the strong ignorability condition (1). Suppose that $g_k(Y | Z, \mathbf{X}_{(k)}; \gamma_k)$ and $h_k(Z | \mathbf{X}_{(k)}; \theta_k)$ are, respectively, the “reduced” outcome and PS regression models used in the k th study, where γ_k and θ_k are the respective vectors of regression coefficients in the reduced outcome and the PS models based on the reduced covariate set $\mathbf{X}_{(k)}$. Note that since the reduced PS and outcome models are from existing studies based on reduced sets of the covariate variables, to make the proposed method applicable even when models used in such studies are wrongly specified, we consider the general case where these models can be misspecified. That is, $g_k(Y | Z, \mathbf{X}_{(k)})$ and $h_k(Z | \mathbf{X}_{(k)})$ may not equal to the true conditional distributions $P(Y | Z, \mathbf{X}_{(k)})$ and $P(Z | \mathbf{X}_{(k)})$, respectively. Assume that the estimates of γ_k and θ_k are available for study $k, k = 1, \dots, K$. Individual data in these studies are not required in the proposed method.

Further, assume that the true distribution of Z given the complete covariate set \mathbf{X} is given by the full PS model $f_Z(Z | \mathbf{X}; \alpha)$, where α is the vector of regression coefficients of \mathbf{X} , and hence the full PS is given by $e(\mathbf{X}) = f_Z(Z = 1 | \mathbf{X}; \alpha)$. Also, assume that the true distribution of Y given Z and \mathbf{X} is given by the full outcome regression model $f_Y(Y | Z, \mathbf{X}; \beta)$, where β is the vector of regression coefficients of (Z, \mathbf{X}) . To fix ideas, in the following we consider the underlying distributions $f_Z(Z | \mathbf{X}; \alpha)$ and $f_Y(Y | Z, \mathbf{X}; \beta)$ that follow the generalized linear models (GLMs) of Nelder and Wedderburn [17], although the ideas may simply extend to more general models. We assume that the underlying treatment and outcome distributions $f_Z(Z | \mathbf{X}; \alpha)$ and $f_Y(Y | Z, \mathbf{X}; \beta)$ are essentially the same across different studies, except that their baselines, namely their intercept parameters in α and β

are allowed to vary across different studies to accommodate possible differences in baseline data distributions among existing studies; however, we do not make this explicit in the notation for the regression parameters to keep the notation simple.

2.2. Reference Data

In addition to the existing regression results, the proposed method needs a reference dataset with data on the complete covariate set.

Specifically, let $\{\mathbf{X}_i\}_{i=1}^n$ be the reference sample of n independent observations of \mathbf{X} . Since the intercept parameters in the treatment and outcome distributions $f_Z(Z | \mathbf{X}; \alpha)$ and $f_Y(Y | Z, \mathbf{X}; \beta)$ are already allowed to be different among the existing studies, the covariate distributions among the existing studies and the reference sample are assumed to be the same up to location shifts, namely these covariate distributions can have different means but become the same after mean removal. In fact, under the above assumptions, it is possible just to assume the covariate distributions among the studies and the reference dataset are the same, and to adjust the intercept parameters of the underlying treatment and outcome distributions $f_Z(Z | \mathbf{X}; \alpha)$ and $f_Y(Y | Z, \mathbf{X}; \beta)$ in each study, such that $f_Y(Y | Z, \mathbf{X}; \beta) \times f_Z(Z | \mathbf{X}; \alpha) \times f_X(\mathbf{X})$ equals the joint distribution of (Y, Z, \mathbf{X}) in each study population. It is such adjusted full models $f_Z(Z | \mathbf{X}; \alpha)$ and $f_Y(Y | Z, \mathbf{X}; \beta)$ that are our estimation targets.

2.3. Estimation of the Full Propensity and Outcome Regression Models

The generalized meta-analysis method of Kundu et al. [16] is a data integration method for combining information on parameters of various outcome regression models with disparate covariate sets. Assuming a common underlying data distribution, this method integrates the estimating equations for parameter estimates of various outcome regression models using the generalized method of moments approach [18] and a reference dataset on the covariates to yield consistent estimation for the full outcome regression model.

We apply and extend the generalized meta-analysis method to estimate the parameters in both the full PS and outcome models $f_Z(Z | \mathbf{X}; \alpha)$ and $f_Y(Y | Z, \mathbf{X}; \beta)$ using the available estimates $\hat{\theta}_k$ and $\hat{\gamma}_k, k = 1, \dots, K$, from the existing K studies, as well as the reference dataset $\{\mathbf{X}_i\}_{i=1}^n$. Following the usual practice, for $k = 1, \dots, K$, we assume the estimates $\hat{\theta}_k$ and $\hat{\gamma}_k$ are obtained by the maximum likelihood estimation (of the GLMs) based on the (reduced) models $h_k(Z | \mathbf{X}_{(k)}; \theta_k)$ and $g_k(Y | Z, \mathbf{X}_{(k)}; \gamma_k), k = 1, \dots, K$, respectively. Let $t_k(Z | \mathbf{X}_{(k)}; \theta_k) = \frac{\partial}{\partial \theta_k} \log(h_k(Z | \mathbf{X}_{(k)}; \theta_k))$ and $s_k(Y | Z, \mathbf{X}_{(k)}; \gamma_k) = \frac{\partial}{\partial \gamma_k} \log(g_k(Y | Z, \mathbf{X}_{(k)}; \gamma_k))$ be the score functions of the k th reduced PS and outcome models, respectively, and consider the expected scores

$$t_k(\mathbf{X}; \alpha, \theta_k) = E_{Z|\mathbf{X}}\{t_k(Z | \mathbf{X}_{(k)}; \theta_k)\} = \sum_{Z=0,1} t_k(Z | \mathbf{X}_{(k)}; \theta_k) f_Z(Z | \mathbf{X}; \alpha), \text{ and}$$

$$s(\mathbf{X}; \alpha, \beta, \theta_k, \gamma_k) = E_{(Y,Z)|\mathbf{X}}\{s_k(Y | Z, \mathbf{X}_{(k)}; \theta_k, \gamma_k)\}$$

$$= \sum_{Z=0,1} \int s_k(Y | Z, \mathbf{X}_{(k)}; \theta_k, \gamma_k) f_Y(Y | Z, \mathbf{X}; \beta) f_Z(Z | \mathbf{X}; \alpha) dY.$$

Let

$$T(\mathbf{X}; \alpha) = \begin{pmatrix} t_1(\mathbf{X}; \alpha, \hat{\theta}_1) \\ \vdots \\ t_K(\mathbf{X}; \alpha, \hat{\theta}_K) \end{pmatrix}, S(\mathbf{X}; \alpha, \beta) = \begin{pmatrix} s_1(\mathbf{X}; \alpha, \beta, \hat{\theta}_1, \hat{\gamma}_1) \\ \vdots \\ s_K(\mathbf{X}; \alpha, \beta, \hat{\theta}_K, \hat{\gamma}_K) \end{pmatrix}, \text{ and}$$

$$U(\mathbf{X}; \alpha, \beta) = \begin{pmatrix} T(\mathbf{X}; \alpha) \\ S(\mathbf{X}; \alpha, \beta) \end{pmatrix}.$$

The estimator $(\hat{\alpha}, \hat{\beta})$ for (α, β) is obtained by minimizing the objective function $U_n(\alpha, \beta)^T C U_n(\alpha, \beta)$, or equivalently solving the estimating equation $\dot{U}_n(\alpha, \beta)^T C U_n(\alpha, \beta) = 0$,

where $U_n(\alpha, \beta) = \sum_{i=1}^n U(\mathbf{X}_i; \alpha, \beta)$, $\dot{U}_n(\alpha, \beta) = \partial U_n(\alpha, \beta) / \partial(\alpha, \beta)$. The matrix \mathbf{C} is an arbitrary positive semi-definite weighting matrix and the optimal (minimal asymptotic covariance of the resulting estimator) choice of \mathbf{C} is $\mathbf{C} = (\Delta + \Lambda)^{-1}$ [16], where Δ is the covariance matrix of $U(\mathbf{X}; \alpha, \beta)$, and Λ arises from the covariances of $\sqrt{n_k}(\hat{\theta}_k, \hat{\gamma}_k)$, $k = 1, \dots, K$, which may be available from the existing studies or estimated using the reference dataset. The way of estimation of the matrices Δ and Λ is given in Appendix A.

Following similar arguments in Kundu et al. [16], $\sqrt{n}\{(\hat{\alpha}, \hat{\beta}) - (\alpha, \beta)\}$ has the asymptotic normal distribution with zero mean and the covariance matrix given by $\{\mathbf{\Gamma}^T(\Delta + \Lambda)^{-1}\mathbf{\Gamma}\}^{-1}$, where $\mathbf{\Gamma} = E\{\partial U(\mathbf{X}; \alpha, \beta) / \partial(\alpha, \beta)\}$ whose estimation can be based on its sample analog in the reference dataset, and the matrices Δ and Λ are estimated in the way mentioned in Appendix A using the final estimator $(\hat{\alpha}, \hat{\beta})$ for (α, β) .

We conclude this section by noting that, the covariance matrices of the regression parameter estimates from the existing studies are not necessary for implementing the proposed method. However, when the covariance matrix for the regression parameter estimates is unavailable and the outcome regression model in an existing study contains a dispersion parameter not fixed to 1 (e.g., the normal regression model), the estimate for that dispersion parameter is required for the proposed method to implement estimation of the covariance matrix of the regression parameters (see Appendix A for detail).

2.4. Estimation of the Average Treatment Effect (ATE)

Let $Q(\mathbf{X}) = E(Y | Z = 1, \mathbf{X}) - E(Y | Z = 0, \mathbf{X})$. By the SUTVA and the strong ignorability assumption (1), $Q(\mathbf{X}) = E(Y_1 | Z = 1, \mathbf{X}) - E(Y_0 | Z = 0, \mathbf{X}) = E(Y_1 | \mathbf{X}) - E(Y_0 | \mathbf{X})$ corresponds to the conditional average treatment effect (CATE) at \mathbf{X} , and $E_{\mathbf{X}}\{Q(\mathbf{X})\} = E(Y_1) - E(Y_0) \equiv \mu$ corresponds to the ATE.

Let $Q(\mathbf{X}; \beta) = \int Y\{f_Y(Y | Z = 1, \mathbf{X}; \beta) - f_Y(Y | Z = 0, \mathbf{X}; \beta)\}dY$. Given the estimate $\hat{\beta}$ for the full outcome regression parameter obtained in Section 2.3, we can estimate $E(Y | Z, \mathbf{X})$ by $\int Y f_Y(Y | Z, \mathbf{X}; \hat{\beta})dY$, the CATE $Q(\mathbf{X})$ at \mathbf{X} by $Q(\mathbf{X}; \hat{\beta})$, and the ATE by the reference sample data $\{\mathbf{X}_i\}_{i=1}^n$ via

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n Q(\mathbf{X}_i; \hat{\beta}).$$

The consistency of the proposed ATE estimator $\hat{\mu}$ follows directly from the consistency of $\hat{\beta}$. Also, using the delta method, we can obtain the asymptotic normal distribution of $\hat{\mu}$; details are provided in Appendix B.

3. The Simulation Studies

In this section, we conduct simulations to assess the performance of the proposed estimators for the regression coefficients of the underlying outcome and treatment distributions, as well as for the ATE. We specifically report results for bias, standard deviation (SD), mean of estimated standard errors (ESE), and coverage probability 1 and 2 (CP1 and CP2) of the 95% and 90% Wald-type confidence intervals, respectively, across 2000 simulation replications.

3.1. The Simulation Setting

We consider a simple simulation setting where there exist independent datasets D_1 on $\{Y, Z, X_1\}$, D_2 on $\{Y, Z, X_2\}$, and D_3 on $\{X_1, X_2\}$, where Y is the common outcome of interest, Z is the common treatment indicator variable, X_1 is the covariate observed in dataset D_1 , X_2 is the covariate observed in dataset D_2 , and (X_1, X_2) is the set of covariates in the underlying full regression models for the outcome and the treatment (i.e., the PS). We set the sample sizes of the datasets D_1 , D_2 , and D_3 to $n_1 = 500$, $n_2 = 500$, $n = 50$, 100 or 200, respectively.

The covariates (X_1, X_2) in D_1 , D_2 , and D_3 are generated by independent standard normal random variables, and given (X_1, X_2) , the treatment assignment Z in both D_1 and D_2 is gener-

ated using the linear logistic regression model: $P(Z = 1 | X_1, X_2) = \text{expit}(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2)$ with $\text{expit}(x) = \exp(x)/(1 + \exp(x))$. The outcome Y in both D_1 and D_2 is binary and generated by the linear logistic regression model for given treatment assignment Z and covariates (X_1, X_2) : $P(Y = 1 | Z, X_1, X_2) = \text{expit}(\beta_0 + \beta_z Z + \beta_1 X_1 + \beta_2 X_2)$. The coefficients for the covariates, (α_1, α_2) and (β_1, β_2) , in the PS and outcome models are both fixed at $(\log(1.2), \log(1.5))$, and the coefficient β_z for the treatment indicator is set to 0, $\log(1.2)$, $\log(1.5)$, or $\log(2)$ when generating both datasets D_1 and D_2 . On the other hand, the intercept parameters (α_0, β_0) are set to $(-1, -1)$ and $(0, -1.5)$ when generating datasets D_1 and D_2 , respectively, to reflect different baselines in the underlying distributions of the two datasets. The reduced PS and outcome regression coefficient estimates are obtained from both D_1 and D_2 ; in D_1 , the working PS and outcome regression models are, respectively, the linear logistic models for $(Z | X_1)$ and $(Y | Z, X_1)$, while in D_2 they are, respectively, the linear logistic models for $(Z | X_2)$ and $(Y | Z, X_2)$. Such reduced-model regression coefficient estimates are used in the later estimation procedure while individual data in D_1 and D_2 are not. Individual data in D_3 on the complete covariate set (X_1, X_2) are used as the reference data for the proposed estimation.

The method in Section 2.3 is applied to the regression coefficient estimates from D_1 and D_2 and the individual data from D_3 to obtain the estimates of the parameters in the full PS and outcome regression models, which are correctly specified in the estimation as the data generating models mentioned above. Further, the method in Section 2.4 is applied to estimate the ATE of Z on Y .

We also perform an extended simulation with $K = 4$ existing datasets over 1000 simulation replications, and the distribution of the outcome Y is binomial, normal, or Poisson. The complete covariate set is $\mathbf{X} = (X_1, X_2, X_3, X_4)$, where (X_1, X_2) are generated from standard normal with a correlation coefficient of 0.5, and (X_3, X_4) are Bernoulli random variables with success probabilities $\text{expit}(X_1 + E_1)$ and $\text{expit}(X_2 + E_2)$, where X_1, X_2 are the covariates mentioned above, E_1, E_2 are independent standard normal, and $\text{expit}(x) = \exp(x)/\{1 + \exp(x)\}$. Accordingly, X_3 is correlated with X_1 , and X_4 is correlated with X_2 and hence also X_1 since X_1 and X_2 are correlated. The four existing datasets contain data on the common outcome variable Y and treatment variable Z , and data on the different covariate sets (X_1, X_2) , (X_2, X_3) , (X_3, X_4) , and (X_1, X_4) , respectively. The treatment variable Z in the k th dataset is generated using the linear logistic regression model: $P_k(Z = 1 | X_1, X_2, X_3, X_4) = \text{expit}(\alpha_{0k} + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4)$, $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (\log(1.2), \log(1.5), -\log(1.2), -\log(1.5))$, and $\alpha_{0k} = -1, -1.5, -2, -2.5$ for $k = 1, 2, 3, 4$. The outcome Y in the k th dataset is generated using the linear logistic regression model: $P_k(Y = 1 | Z, X_1, X_2, X_3, X_4) = g(\beta_{0k} + \beta_z Z + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4)$ with g the expit, identity, or exponential function when the distribution of Y is binomial, normal, or Poisson (i.e., the inverse canonical link function), respectively. The true parameter values $(\beta_1, \beta_2, \beta_3, \beta_4) = (\log(1.2), \log(1.5), -\log(1.5), -\log(1.2))$, $\beta_{0k} = -1, -1.5, -2, -2.5$ for $k = 1, 2, 3, 4$, and $\beta_z = 0$ or $\log(1.5)$. The reduced propensity (PS) and outcome regression coefficient estimates are obtained from the four existing datasets, using the models having the same link functions as in the data-generating models but the reduced covariate sets observed in these datasets, as mentioned above. Such reduced regression coefficient estimates are used in the estimation procedure while individual data in the existing datasets are not. Individual data in the reference sample, which contain data on the complete covariate set $\mathbf{X} = (X_1, X_2, X_3, X_4)$, are also used in the estimation procedure as proposed in Section 2.3. The full PS and outcome regression models in the estimation are correctly specified as the data-generating models mentioned above. Further, the method in Section 2.4 is applied to estimate the average treatment effect (ATE) of Z on Y in the population of each dataset. All four existing studies and the reference sample have the same sample size, set to 200, 500, or 1000.

3.2. The Simulation Results

Table 1 presents the simulation results for the proposed estimation of the regression coefficient for the treatment variable in the full outcome model and the ATEs for the populations

of the two existing datasets. The findings in Table 1 suggest that the proposed estimation for the full outcome model parameter and the ATE is essentially unbiased, with the absolute bias of the proposed estimates for both parameters being less than 1%. Also, the estimated standard error (ESE) based on the asymptotic theory is close to the simulation standard deviation of the estimator with the absolute difference less than 1%, and the coverage probability of the 95% and 90% Wald-type confidence intervals based on the asymptotic normality of the estimator is close to the nominal levels 0.95 and 0.90, respectively. These results reveal that the proposed estimators perform well in finite samples.

Table 2 presents the simulation results for the proposed estimation of the regression coefficients in the full PS model under the same simulation settings as in Table 1. The findings demonstrate that our proposed method performs well in the estimation of the full PS model.

Table 1. Simulation results (multiplied by 100) for the estimates of the coefficient β_Z of the treatment variable in the full outcome model and the ATEs with true parameter values $\beta_0 = -1$ in Study 1, and -1.5 in study 2, $(\beta_{X_1}, \beta_{X_2}) = (\log(1.2), \log(1.5))$, and different β_Z and ATE values (in parenthesis), $\alpha = (\alpha_0, \alpha_1, \alpha_2) = (-1, \log(1.2), \log(1.5))$ in Study 1, and $(0, \log(1.2), \log(1.5))$ in Study 2.

	Bias	SD	$n = 50$			Bias	SD	$n = 100$			Bias	SD	$n = 200$		
			ESE	CP1	CP2			ESE	CP1	CP2			ESE	CP1	CP2
$\beta_Z(0)$	-0.19	17.0	17.7	96.5	91.6	-0.32	17.7	17.5	95.1	89.8	-0.94	17.3	17.4	95.1	89.7
ATE1 (0)	0.03	3.23	3.37	96.4	91.5	0.01	3.37	3.34	95.0	89.8	-0.11	3.29	3.32	94.8	89.5
ATE2 (0)	-0.01	2.53	2.64	96.5	91.5	-0.04	2.62	2.60	95.1	89.8	-0.14	2.58	2.59	95.0	89.6
$\beta_Z(\log(1.2))$	-0.08	16.5	17.2	96.2	91.7	-0.21	17.2	17.0	95.1	89.2	-0.53	17.0	16.9	94.7	89.8
ATE1 (0.04)	-0.00	3.33	3.47	96.3	91.4	-0.00	3.48	3.44	94.9	89.4	-0.07	3.41	3.42	94.6	89.6
ATE2 (0.03)	-0.02	2.61	2.72	95.9	91.8	-0.05	2.70	2.68	95.3	89.3	-0.10	2.67	2.67	94.8	89.9
$\beta_Z(\log(1.5))$	-0.13	16.2	16.8	96.1	91.5	-0.14	16.4	16.6	95.4	90.2	-0.35	16.4	16.5	94.8	90.0
ATE1 (0.08)	-0.07	3.46	3.59	96.4	91.4	-0.04	3.52	3.55	95.5	90.2	-0.09	3.49	3.52	94.5	89.8
ATE2 (0.07)	-0.06	2.74	2.82	96.1	90.9	-0.08	2.75	2.78	95.6	91.0	-0.10	2.75	2.77	95.3	89.6
$\beta_Z(\log(2))$	0.04	15.8	16.3	96.2	91.5	0.15	15.9	16.2	95.5	90.1	-0.17	15.9	16.1	95.6	90.3
ATE1 (0.15)	-0.12	3.56	3.69	96.2	91.2	-0.03	3.59	3.65	95.0	90.0	-0.12	3.59	3.62	95.1	90.0
ATE2 (0.12)	-0.09	2.87	2.96	95.9	90.6	-0.08	2.85	2.91	95.4	90.2	-0.12	2.85	2.89	95.6	90.1

n , size of reference data; SD, standard deviation; ESE, estimated standard error; CP1 and CP2, coverage probability of 95% and 90% confidence intervals, respectively.

Table 2. Simulation results (multiplied by 100) for the estimates of the coefficients $(\alpha_{X_1}, \alpha_{X_2})$ of the covariate variables in the full propensity score model with true parameter values $\beta_0 = -1$ in Study 1, and -1.5 in study 2, $(\beta_{X_1}, \beta_{X_2}) = (\log(1.2), \log(1.5))$, and different β_Z values, $\alpha = (\alpha_0, \alpha_{X_1}, \alpha_{X_2}) = (-1, \log(1.2), \log(1.5))$ in Study 1, and $(0, \log(1.2), \log(1.5))$ in Study 2.

	Bias	SD	$n = 50$			Bias	SD	$n = 100$			Bias	SD	$n = 200$		
			ESE	CP1	CP2			ESE	CP1	CP2			ESE	CP1	CP2
$\beta_Z = 0$															
α_{X_1}	0.99	13.0	12.8	96.0	90.3	1.19	12.0	11.7	95.7	89.5	0.32	11.0	11.1	95.2	90.2
α_{X_2}	1.29	10.8	10.9	96.4	91.6	0.91	10.1	10.2	95.6	90.9	0.44	9.82	9.84	95.5	90.3
$\beta_Z = \log(1.2)$															
α_{X_1}	0.99	13.0	12.8	96.2	90.4	1.20	12.0	11.7	95.6	89.6	0.33	11.0	11.1	95.2	90.3
α_{X_2}	1.29	10.8	10.9	96.3	91.7	0.91	10.1	10.2	95.6	90.9	0.44	9.82	9.84	95.6	90.5
$\beta_Z = \log(1.5)$															
α_{X_1}	1.00	13.0	12.8	96.2	90.2	1.20	12.0	11.7	95.5	89.5	0.33	11.0	11.1	95.3	90.3
α_{X_2}	1.30	10.8	10.9	96.3	91.9	0.93	10.1	10.2	95.9	91.1	0.44	9.82	9.84	95.4	90.3
$\beta_Z = \log(2)$															
α_{X_1}	1.00	13.0	12.8	96.0	90.4	1.20	12.0	11.7	95.7	89.5	0.33	11.0	11.1	95.3	90.3
α_{X_2}	1.32	10.8	10.9	96.1	91.7	0.94	10.1	10.2	95.9	91.1	0.46	9.82	9.83	95.3	90.3

n , size of reference data; SD, standard deviation; ESE, estimated standard error; CP1 and CP2, coverage probability of 95% and 90% confidence intervals, respectively.

The extended simulations, performed under the settings with $K = 4$ studies, correlated covariates, and binomial, normal, and Poisson distributed outcome variables (see Section 3.1 for detail), still reveal satisfactory performances of the proposed estimation method; the results are shown in the Appendix C.

The performance of the proposed method in terms of computation time is summarized as follows. When $K = 2$, the computation time for a simulation case with $n = 200$ is 0.5 s in a desktop computer with i7-9700 CPU, and the time increases to 1.2 s when $K = 4$ and $n = 200$. Essentially, the computation time increases linearly with K . Also, when $K = 4$, the computation time for running a case in the simulation study is 5.0 s when the size of the reference sample increases to $n = 1000$.

4. A Real Data Application

In this section, we apply the proposed method to analyze the impact of waist circumference (WC) on hypertension (HT) risk, while controlling for covariates such as age, sex, body mass index (BMI), smoking status (SMK), drinking status (DRK), body fat percentage (BFP), heart rate (HR), and hip circumference (HC) among working-age individuals. The analysis leverages regression analyses from two existing studies on the relationship between WC and HT and a reference dataset encompassing the complete set of covariates.

4.1. Two Existing Studies on the Effect of Waist Circumference on Hypertension

The WC reflects the size of the visceral fat depot and is an effective clinical tool for assessing the risk of diabetes and cardiovascular diseases [19]. Guagnano et al. [20] indicated that WC seems to have a strong association with the risk of hypertension. In recent years, Ren et al. [14] investigated the cut-off values for the obesity indices that represent the elevated incidence of hypertension in Chinese adults aged between 18 and 65. Hu et al. [15] indicated that a combination of WC and BMI was superior to individual indices for identifying hypertension. Data from Ren et al. [14] and Hu et al. [15] were publicly provided (<https://doi.org/10.6084/m9.figshare.2151271.v1> (accessed on 15 February 2016), <https://doi.org/10.1371/journal.pone.0170238.s001> (accessed on 5 January 2017)). The study of Hu et al. [15], termed Study 1, contains data on the covariate set $X_{(1)}$ including age, sex, BMI, SMK, DRK, BFP, and HR, while the study of Ren et al. [14], termed Study 2, contains data on the covariate set $X_{(2)}$ including age, sex, BMI, SMK, DRK, and HC. We focus on working-age (20–65 years old) people and the subsamples from the two studies meeting this criterion are of sizes $n_1 = 9926$ and $n_2 = 2970$, respectively, (after removing missing observations).

4.2. The Reference Dataset with Complete Covariates

The Taiwan BioBank (TWB) database, created by Academia Sinica, comprises a community-based cohort of over 200,000 study participants. It includes comprehensive data on demographics, health behaviors, environmental factors, and biomarkers collected through meticulously conducted questionnaires and thorough examinations. Details about the TWB data can be found at <https://www.twbiobank.org.tw/> (accessed on 1 June 2024). The reference dataset we employ in the current analysis is based on the released subsample of the TWB cohort consisting of 4575 randomly sampled study subjects aged 20–65 years. The reference dataset contains data on the complete covariate set X including age, sex, BMI, SMK, DRK, BFP, HR, and HC, but contains no data on either the treatment (WC) or the outcome (HT).

4.3. The Proposed Analysis

In the following analysis, both WC and HT are defined as binary variables, classified according to $WC = I(\text{waistcircumference} > 80 \text{ cm})$ and $HT = I(\text{SBP} \geq 140 \text{ mmHg or DBP} \geq 90 \text{ mmHg})$, where $I(\cdot)$ is the indicator function and SBP and DBP denote systolic and diastolic blood pressures, respectively; the classification rules follow those in Lean et al. (1998) [19]. The covariates age (years), BMI (kg/m^2), BFP (%), HR (beats/min), and HC (cm) are continuous variables, while the covariates sex (female vs. male), SMK (yes vs. no), and DRK (yes vs. no) are binary.

We apply the proposed methods in Sections 2.3 and 2.4 to assess the treatment effect of WC on the risk of HT controlling for the covariates age, sex, BMI, SMK, DRK, BFP, HR,

and HC, which are regarded as the complete covariate set among working-age people. Specifically, the proposed analysis uses the results of the regression analyses from Study 1 (Hu et al. [15]) and Study 2 (Ren et al. [14]), as well as the reference dataset from the TWB database. Both the analyses in Studies 1 and 2 employ linear logistic regressions to examine the association between WC and HT by adjusting the covariate sets $\mathbf{X}_{(1)}$ and $\mathbf{X}_{(2)}$, respectively; see Section 4.1 for detail. Also, both PS analyses in the two studies are based on the linear logistic regressions for WC with the covariate sets $\mathbf{X}_{(1)}$ and $\mathbf{X}_{(2)}$, respectively.

In the proposed analysis, only the outcome and regression parameter estimates from Studies 1 and 2 are employed, while individual data are not. The full outcome (HT) model is specified by the linear logistic regression model for HT given the treatment (WC) and the complete covariate set \mathbf{X} , and the full PS model is specified as the linear logistic regression model for WC given \mathbf{X} ; in these logistic regression models only the main effects of the treatment and the covariate variables are considered. To account for possible differences between the baselines of Studies 1 and 2, the intercept parameters of the full models, including the outcome and the PS models, across the studies are treated as different.

4.4. The Analysis Results

The results for the proposed estimation of the logistic regression models for the PS (treatment, WC) and the outcome (HT) adjusting for the complete covariate set are provided in Tables 3 and 4, respectively. We can see from Table 3 that, older age, male, larger BMI, higher heart rate, and larger hip circumference tend to have a waist circumference greater than 80 cm (treatment group), and the estimation result seems to nicely summarize, synthesize, and complement the results from the two existing studies. Also, we see from Table 4 that, after adjusting for the covariates age, sex, body mass index (BMI), smoking (SMK), drinking (DRK), body fat percentage (BFP), heart rate (HR), and hip circumference (HC), the effect of waist circumference (WC) on the risk of hypertension (HT) is strongly significant; the odds for hypertension in working-age people with waist circumference greater than 80 cm is 1.4 ($\approx \exp(0.323)$) times as high as those with waist circumference no greater than 80 cm (p -value < 0.001). In contrast, the effect of WC on the risk of HT obtained by adjusting an incomplete covariate set can be somewhat higher (when adjusting only for age, sex, BMI, SMK, DRK, BFP, and HR in Study 1) or lower (when adjusting only for age, sex, BMI, SMK, DRK, and HC in Study 2). Since HR has its own significant effects on both WC and HT, the lower effect of WC on HT without adjusting for HR obtained in Study 2 is likely to be biased.

The average treatment effect of WC on HT, averaged over the covariate distribution, is obtained as 0.044 with SE (standard error) = 0.009, p -value < 0.001 , in Study 1 population, and as 0.042 with SE = 0.009, p -value < 0.001 , in Study 2 population. That is overall working-age people with waist circumference larger than 80 cm can have 44 (42) additional cases of HT per 1000 people in the Study 1 (2) population, 95% confidence interval 26–62 (24–60), compared to those who with waist circumference no larger than 80 cm.

From the results mentioned above and shown in Table 3, we conclude that the proposed integrated analysis, using information from both studies and the reference data with complete covariates, can lead to less biased and possibly more efficient analysis results than those from the original individual studies.

Table 3. Results of the real-data analysis. The propensity score (PS) for waist circumference (WC) with the covariates age, sex, body mass index (BMI), smoking status (SMK), drinking status (DRK), body fat percentage (BFP), heart rate (HR), and hip circumference (HC) based on results from two studies and a reference sample.

	PS Model		
	Study 1 $n_1 = 9926$ Est (SE)	Study 2 $n_1 = 2970$ Est (SE)	Proposed $n = 4575$ Est (SE)
Age	0.025 (0.002) *	0.041 (0.006) *	0.046 (0.003) *
Sex	−1.234 (0.074) *	−0.864 (0.150) *	−1.237 (0.157) *
BMI	0.597 (0.013) *	0.301 (0.021) *	0.310 (0.028) *
SMK	0.009 (0.073)	0.055 (0.150)	0.052 (0.073)
DRK	0.065 (0.067)	0.229 (0.129)	0.118 (0.097)
BFP	0.035 (0.004) *	-	0.025 (0.013)
HR	0.010 (0.002) *	-	0.013 (0.004) *
HC	-	0.217 (0.011) *	0.217 (0.017) *

Est: parameter estimate; SE: standard error; *: p -value < 0.05.

Table 4. Results of the real-data analysis. The risk of hypertension (HT) with treatment of waist circumference (WC) adjusting for age, sex, body mass index (BMI), smoking status (SMK), drinking status (DRK), body fat percentage (BFP), heart rate (HR), and hip circumference (HC) based on results from two studies and a reference sample.

	Outcome Model		
	Study 1 $n_1 = 9926$ Est (SE)	Study 2 $n_1 = 2970$ Est (SE)	Proposed $n = 4575$ Est (SE)
WC	0.418 (0.067) *	0.047 (0.131)	0.323 (0.067) *
Age	0.085 (0.003) *	0.061 (0.006) *	0.080 (0.003) *
Sex	−0.196 (0.074) *	−0.382 (0.151) *	−0.251 (0.126) *
BMI	0.101 (0.010) *	0.163 (0.020) *	0.109 (0.024) *
SMK	0.027 (0.071)	0.115 (0.146)	0.044 (0.064)
DRK	−0.060 (0.066)	−0.035 (0.131)	−0.053 (0.084)
BFP	0.005 (0.004)	-	0.005 (0.010)
HR	0.021 (0.002) *	-	0.021 (0.003) *
HC	-	0.002 (0.010)	0.003 (0.013)

Est: parameter estimate; SE: standard error; *: p -value < 0.05.

5. Discussion and Conclusions

In this study, we propose a new inference framework that integrates the results of the outcome and the treatment (i.e., the PS) regression analyses across multiple existing databases. These databases may vary in their coverage of covariate variables and may contain incomplete data, potentially introducing bias in individual database analyses. Moreover, access to individual-level data from these databases may be restricted. Our proposal integrates the existing PS and outcome regression analyses through a reference sample, which contains only data on the complete covariate set. We obtain consistent estimates for the parameters of the full PS and outcome regression models and for the ATE. The new proposal extends the original generalized meta-analysis method of Kundu et al. [16] by further considering the treatment (propensity score) regression in addition to the outcome regression. Also, the new proposal can apply with a general outcome variable, such as one following a generalized linear model, and hence is more flexible than the work of Li et al. [21], which considers the setting similar to ours but is restricted to normality outcome and linear regression model.

Our approach necessitates a dataset with comprehensive covariate information, which acts as a benchmark for the underlying covariate distribution [16]. Such a reference dataset could be sourced from a large-scale database like the Taiwan Biobank, as outlined in Section 4 of our application. Alternatively, a reference sample might be gathered through a smaller validation study, a method commonly discussed in the epidemiological literature (e.g., Stümer et al. [22]).

As in the existing methods, such as Kundu et al. [16], for integrating common information from different studies, we require the underlying treatment and outcome distributions

to be the same across various studies. When this assumption is not satisfied, we should interpret the parameter estimates from the proposed method with caution, since they no longer represent consistent estimates for some common parameters, but instead represent the estimates for some “average effects” over different studies.

In summary, our proposal is the best applicable in the following two scenarios: (1) A multi-center study where individual data from each of the centers are not accessible except for the derived summary statistics (e.g., regression coefficient estimates), and an independent reference sample of complete covariate data is available. (2) A meta-analysis where results for both the outcome and the treatment (propensity score) regression analyses are available for various studies, together with a reference sample of complete covariate data. In both scenarios, our approach optimally integrates analysis results from diverse data sources to yield valid inferences on treatment effects using summarized and synthesized information.

Author Contributions: Conceptualization, Y.-H.C.; methodology, Y.-H.C. and S.-Y.H.; software, S.-Y.H., J.-H.W. and C.-C.S.; validation, S.-Y.H., J.-H.W. and C.-C.S.; formal analysis, S.-Y.H., J.-H.W. and C.-C.S.; investigation, S.-Y.H., J.-H.W. and C.-C.S.; resources, Y.-H.C. and J.-H.W.; data curation, S.-Y.H., J.-H.W. and C.-C.S.; writing—original draft preparation, Y.-H.C.; writing—review and editing, Y.-H.C. and J.-H.W.; visualization, Y.-H.C. and J.-H.W.; supervision, Y.-H.C.; project administration, Y.-H.C.; funding acquisition, Y.-H.C. and J.-H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the grant NSTC 112-2118-M-194-003-MY2 from the National Science and Technology Council of the Republic of China (Taiwan).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ATE	Average treatment effect
SUTVA	Stable unit-treatment value assumption
CA	Covariate adjustment
IPTW	Inverse probability of treatment weighting
PS	Propensity score
GDPR	General Data Protection Regulation
HIPAA	Health insurance portability and accountability act
WC	Waist circumference
HT	Hypertension
CATE	Conditional average treatment effect
SD	Standard deviation
ESE	Estimated standard errors
CP	Coverage probability
BMI	Body mass index
SMK	Smoking status
DRK	Drinking status
BFP	Body fat percentage
HR	Heart rate
HC	Hip circumference
TWB	Taiwan BioBank
SBP	Systolic blood pressures
DBP	Diastolic blood pressures

Appendix A. Estimation of the Matrices Δ and Λ

The matrices Δ and Λ are defined as $\Delta = E\{U(\mathbf{X}; \alpha, \beta)U^T(\mathbf{X}; \alpha, \beta)\}$, and

$$\Lambda = \begin{pmatrix} \Lambda^{(1)} & \Lambda^{(1,2)} \\ \Lambda^{(1,2)T} & \Lambda^{(2)} \end{pmatrix}$$

with $\Lambda^{(j)} = \text{diag}(\Lambda_1^{(j)}, \dots, \Lambda_K^{(j)})$, $j = 1, 2$, $\Lambda^{(1,2)} = \text{diag}(\Lambda_1^{(1,2)}, \dots, \Lambda_K^{(1,2)})$, where

$$\begin{pmatrix} \Lambda_k^{(1)} & \Lambda_k^{(1,2)} \\ \Lambda_k^{(1,2)T} & \Lambda_k^{(2)} \end{pmatrix} = \frac{1}{c_k} E \left\{ \begin{pmatrix} t_k(Z | \mathbf{X}_{(k)}; \alpha, \theta_k) \\ s_k(Y | Z, \mathbf{X}_{(k)}; \alpha, \beta, \theta_k, \gamma_k) \end{pmatrix} \begin{pmatrix} t_k(Z | \mathbf{X}_{(k)}; \alpha, \theta_k) \\ s_k(Y | Z, \mathbf{X}_{(k)}; \alpha, \beta, \theta_k, \gamma_k) \end{pmatrix}^T \right\} \\ \equiv \Lambda_k^*, \text{ for } k = 1, \dots, K,$$

$c_k = n_k/n$. In practice, we can first use the identity matrix for the weight \mathbf{C} , namely minimize the objective function $U_n(\alpha, \beta)^T U_n(\alpha, \beta)$ to obtain the initial estimator $(\tilde{\alpha}, \tilde{\beta})$ for (α, β) , and use the initial estimator to estimate the matrix Δ by

$$\hat{\Delta} = \frac{1}{n} \sum_{i=1}^n U(\mathbf{X}_i; \tilde{\alpha}, \tilde{\beta}) U^T(\mathbf{X}_i; \tilde{\alpha}, \tilde{\beta})$$

and the matrix Λ_k^* ($k = 1, \dots, K$) by

$$\hat{\Lambda}_k^* = \left(\frac{1}{c_k} \right) P_n \left[E_{(Y,Z)|\mathbf{X}} \left\{ \begin{pmatrix} t_k(Z | \mathbf{X}_{(k)}; \hat{\theta}_k) \\ s_k(Y | Z, \mathbf{X}_{(k)}; \hat{\theta}_k, \hat{\gamma}_k) \end{pmatrix} \begin{pmatrix} t_k(Z | \mathbf{X}_{(k)}; \hat{\theta}_k) \\ s_k(Y | Z, \mathbf{X}_{(k)}; \hat{\theta}_k, \hat{\gamma}_k) \end{pmatrix}^T \right\} \right]$$

where $P_n\{f(\mathbf{X})\} = 1/n \sum_{i=1}^n f(\mathbf{X}_i)$ is the empirical measure with respect to the reference sample and is evaluated at $(\alpha, \beta) = (\tilde{\alpha}, \tilde{\beta})$. The optimal weight \mathbf{C} is then estimated by $\hat{\Delta} + \hat{\Lambda}$.

Note that, when the outcome regression in study k is specified by a GLM with a dispersion parameter not fixed to 1 and estimated from study data, the calculation of $\hat{\Lambda}_k^*$ involves the estimated value of the dispersion parameter.

Appendix B. Large Sample Theory for the ATE Estimator $\hat{\mu}$

By the consistency of the estimator $\hat{\beta}$ and the continuous mapping theorem, in large samples, the ATE estimator $\hat{\mu}$ converges to $\mu = E(Y_1) - E(Y_0)$ with probability one. Also, by the delta method, $\sqrt{n}(\hat{\mu} - \mu)$ converges in distribution to a zero-mean normal distribution with the variance given by $\omega + \Phi \Omega_\beta \Phi^T - 2\mathbf{A} \mathbf{C}^T \Gamma^T \Phi^T$ with $\omega = \text{var}\{Q(\mathbf{X}; \beta)\}$, $\Phi = E\{\partial Q(\mathbf{X}; \beta) / \partial \beta\}$, Ω_β being the lower right $p \times p$ ($p = \text{dim}(\beta)$) submatrix of $\Omega = \{\Gamma^T(\Delta + \Lambda)^{-1}\Gamma\}^{-1}$, i.e., the asymptotic covariance of $\sqrt{n}(\hat{\beta} - \beta)$, the matrix $\mathbf{A} = E\{Q(\mathbf{X}; \alpha, \beta)U(\mathbf{X}; \alpha, \beta)^T\}$, \mathbf{H} being the matrix formed by the lower p rows of Ω , and Γ, Δ , and Λ are given in Section 2.3 and Appendix A. The estimation of the asymptotic variance of $\hat{\mu}$ can be performed similarly to that of the asymptotic covariance of $(\hat{\alpha}, \hat{\beta})$ as mentioned in Appendix A.

Appendix C. Extended Simulations

The extended simulations, performed under the settings with $K = 4$ studies, correlated covariates, and binomial, normal, and Poisson distributed outcome variables (see Section 3.1 of the main text for details about the simulation settings), are reported in the following supplementary tables, including results for both the small sample (with sample size of 200 or 500) and the large sample (with samples size of 1000), and both the outcome and the propensity score regressions. Tables A1–A4 are for the setting with $\beta_z = 0$,

while Tables A5–A8 are for the setting with $\beta_z = \log(1.5)$. As seen from these tables, the simulation results still reveal satisfactory performances of the proposed estimation method.

Table A1. Simulation results (multiplied by 100) for the estimates of the coefficient β_z of the treatment variable in the full outcome model and the ATEs for the populations of the existing datasets under the binomial, normal, or Poisson outcome distribution; true value of $\beta_z = 0$ (small sample case with sample size of 200 or 500).

	Bias	$n = 200$			Bias	$n = 500$		
		SD	ESE	CP		SD	ESE	CP
Binomial								
$\beta_z(0)$	−2.06	31.0	31.9	96.6	−1.33	20.0	19.4	94.6
ATE1(0)	−0.13	4.98	5.17	95.5	−0.11	3.29	3.20	94.2
ATE2(0)	0.04	3.73	3.88	95.6	−0.05	2.45	2.40	94.4
ATE3(0)	0.07	2.67	2.76	95.7	0.00	1.72	1.69	94.4
ATE4(0)	0.10	1.77	1.86	97.1	0.00	1.14	1.15	94.5
Normal								
$\beta_z(0)$	0.66	22.7	22.7	95.1	−0.11	14.3	14.1	95.2
ATE1(0)	0.66	22.7	22.7	95.1	−0.11	14.3	14.1	95.2
ATE2(0)	0.66	22.7	22.7	95.1	−0.11	14.3	14.1	95.2
ATE3(0)	0.66	22.7	22.7	95.1	−0.11	14.3	14.1	95.2
ATE4(0)	0.66	22.7	22.7	95.1	−0.11	14.3	14.1	95.2
Poisson								
$\beta_z(0)$	−0.53	9.94	10.1	95.6	−0.20	6.12	6.17	94.7
ATE1(0)	−0.69	22.3	22.7	95.6	−0.25	13.8	13.9	95.1
ATE2(0)	−0.28	13.5	13.9	95.6	−0.10	8.37	8.43	95.0
ATE3(0)	−0.14	8.29	8.42	95.6	−0.05	5.08	5.11	95.0
ATE4(0)	−0.04	5.06	5.13	95.9	−0.02	3.09	3.11	95.1

n , size of reference data; SD, standard deviation; ESE, estimated standard error; CP, coverage probability of 95% confidence intervals.

Table A2. Simulation results (multiplied by 100) for the estimates of the coefficients $(\alpha_{X_1}, \alpha_{X_2}, \alpha_{X_3}, \alpha_{X_4})$ of the covariate variables in the full propensity score model with true parameter values $(\alpha_{X_1}, \alpha_{X_2}, \alpha_{X_3}, \alpha_{X_4}) = (\log(1.2), \log(1.5)), -\log(1.2), -\log(1.5))$, under the binomial, normal, or Poisson outcome distribution; true value of $\beta_z = 0$ (small sample case with sample size of 200 or 500).

	Bias	$n = 200$			Bias	$n = 500$		
		SD	ESE	CP		SD	ESE	CP
Binomial								
α_{X_1}	0.61	23.1	23.3	96.1	−0.26	14.5	14.2	96.2
α_{X_2}	1.40	21.8	22.6	96.1	1.12	13.3	13.6	95.6
α_{X_3}	−0.76	41.3	39.7	95.3	0.53	23.8	24.2	95.5
α_{X_4}	−5.39	97.0	47.8	95.9	−1.00	29.0	29.0	95.0
Normal								
α_{X_1}	0.57	23.5	23.4	95.7	−0.12	14.6	14.2	95.4
α_{X_2}	1.35	23.3	22.5	95.2	0.28	14.1	13.6	94.3
α_{X_3}	−2.58	40.7	39.8	95.6	−0.54	24.1	24.2	95.5
α_{X_4}	−3.00	50.3	47.9	95.5	0.17	29.2	29.0	95.2
Poisson								
α_{X_1}	0.58	23.1	23.3	96.2	−0.29	14.5	14.2	96.2
α_{X_2}	1.72	21.6	22.4	96.3	1.12	13.1	13.4	96.0
α_{X_3}	−0.85	41.1	39.5	95.4	0.51	23.6	24.1	95.5
α_{X_4}	−5.92	97.0	47.6	96.3	−1.02	29.0	28.9	95.0

n , size of reference data; SD, standard deviation; ESE, estimated standard error; CP, coverage probability of 95% confidence intervals.

Table A3. Simulation results (multiplied by 100) for the estimates of the coefficient β_Z of the treatment variable in the full outcome model and the ATEs for the populations of the existing datasets under the binomial, normal, or Poisson outcome distribution; true value of $\beta_z = 0$ (large sample case with sample size of 1000).

	Bias	$n = 1000$ SD	ESE	CP
Binomial				
$\beta_Z (0)$	-0.22	13.3	13.5	94.5
ATE1 (0)	0.01	2.22	2.26	94.4
ATE2 (0)	0.02	1.65	1.68	94.2
ATE3 (0)	0.02	1.15	1.18	94.4
ATE4 (0)	0.03	0.78	0.79	94.7
Normal				
$\beta_Z (0)$	-0.08	10.14	9.95	94.1
ATE1 (0)	-0.08	10.14	9.95	94.1
ATE2 (0)	-0.08	10.14	9.95	94.1
ATE3 (0)	-0.08	10.14	9.95	94.1
ATE4 (0)	-0.08	10.14	9.95	94.1
Poisson				
$\beta_Z (0)$	-0.11	4.22	4.32	95.5
ATE1 (0)	-0.17	9.47	9.71	95.6
ATE2 (0)	-0.08	5.76	5.90	95.8
ATE3 (0)	-0.05	3.50	3.58	95.7
ATE4 (0)	-0.02	2.11	2.17	95.7

n , size of reference data; SD, standard deviation; ESE, estimated standard error; CP, coverage probability of 95% confidence intervals.

Table A4. Simulation results (multiplied by 100) for the estimates of the coefficients ($\alpha_{X_1}, \alpha_{X_2}, \alpha_{X_3}, \alpha_{X_4}$) of the covariate variables in the full propensity score model with true parameter values ($\alpha_{X_1}, \alpha_{X_2}, \alpha_{X_3}, \alpha_{X_4} = (\log(1.2), \log(1.5)), -\log(1.2), -\log(1.5)$), under the binomial, normal, or Poisson outcome distribution; true value of $\beta_z = 0$ (large sample case with sample size of 1000).

	Bias	$n = 1000$ SD	ESE	CP
Binomial				
α_{X_1}	-0.33	9.79	9.90	96.2
α_{X_2}	0.87	9.48	9.51	94.9
α_{X_3}	0.38	17.2	16.9	94.9
α_{X_4}	-1.27	20.3	20.3	95.3
Normal				
α_{X_1}	-0.38	9.82	9.86	94.8
α_{X_2}	-0.02	9.14	9.45	96.1
α_{X_3}	0.09	16.6	16.9	95.6
α_{X_4}	0.38	20.3	20.2	94.9
Poisson				
α_{X_1}	-0.32	9.78	9.88	96.3
α_{X_2}	0.85	9.42	9.39	95.4
α_{X_3}	0.43	17.1	16.8	94.9
α_{X_4}	-1.25	20.3	20.2	95.4

n , size of reference data; SD, standard deviation; ESE, estimated standard error; CP, coverage probability of 95% confidence intervals.

Table A5. Simulation results (multiplied by 100) for the estimates of the coefficient β_Z of the treatment variable in the full outcome model and the ATEs for the populations of the existing datasets under the binomial, normal, or Poisson outcome distribution; true value of $\beta_z = \log(1.5) = 0.41$ (small sample case with sample size of 200 or 500).

	$n = 200$				$n = 500$			
	Bias	SD	ESE	CP	Bias	SD	ESE	CP
Binomial								
$\beta_Z(\log(1.5) = 0.41)$	-0.40	28.4	29.4	96.7	-0.94	17.8	17.8	95.3
ATE1 (0.08)	-0.15	5.30	5.53	95.5	-0.20	3.40	3.42	94.4
ATE2 (0.06)	0.04	4.31	4.44	95.3	-0.10	2.72	2.74	95.0
ATE3 (0.04)	0.11	3.30	3.36	95.2	-0.02	2.05	2.06	94.0
ATE4 (0.03)	0.14	2.37	2.41	93.4	-0.01	1.44	1.47	93.5
Normal								
$\beta_Z(\log(1.5) = 0.41)$	0.66	22.7	22.7	95.1	-0.11	14.3	14.1	95.2
ATE1 (0.41)	0.66	22.7	22.7	95.1	-0.11	14.3	14.1	95.2
ATE2 (0.41)	0.66	22.7	22.7	95.1	-0.11	14.3	14.1	95.2
ATE3 (0.41)	0.66	22.7	22.7	95.1	-0.11	14.3	14.1	95.2
ATE4 (0.41)	0.66	22.7	22.7	95.1	-0.11	14.3	14.1	95.2
Poisson								
$\beta_Z(\log(1.5) = 0.41)$	-0.63	9.04	9.08	95.9	-0.18	5.56	5.58	95.0
ATE1 (1.13)	-1.55	28.1	28.2	95.2	-0.52	17.6	17.3	94.4
ATE2 (0.68)	-0.67	17.9	17.8	94.8	-0.19	11.0	10.9	94.2
ATE3 (0.41)	-0.31	11.3	11.3	94.6	-0.03	6.85	6.96	95.6
ATE4 (0.25)	-0.17	7.04	7.15	95.4	-0.03	4.38	4.39	94.8

n , size of reference data; SD, standard deviation; ESE, estimated standard error; CP, coverage probability of 95% confidence intervals.

Table A6. Simulation results (multiplied by 100) for the estimates of the coefficients $(\alpha_{X_1}, \alpha_{X_2}, \alpha_{X_3}, \alpha_{X_4})$ of the covariate variables in the full propensity score model with true parameter values $(\alpha_{X_1}, \alpha_{X_2}, \alpha_{X_3}, \alpha_{X_4}) = (\log(1.2), \log(1.5)), -\log(1.2), -\log(1.5))$, under the binomial, normal, or Poisson outcome distribution; true value of $\beta_z = \log(1.5) = 0.41$ (small sample case with sample size of 200 or 500).

	$n = 200$				$n = 500$			
	Bias	SD	ESE	CP	Bias	SD	ESE	CP
Binomial								
α_{X_1}	0.62	23.1	23.3	96.1	-0.25	14.5	14.2	96.1
α_{X_2}	1.37	21.8	22.6	96.2	1.13	13.3	13.6	95.6
α_{X_3}	-0.76	41.3	39.7	95.5	0.51	23.9	24.2	95.6
α_{X_4}	-5.38	97.0	47.7	96.0	-1.02	29.0	29.0	94.9
Normal								
α_{X_1}	0.57	23.5	23.4	95.7	-0.12	14.6	14.2	95.4
α_{X_2}	1.35	23.3	22.5	95.2	0.28	14.1	13.6	94.3
α_{X_3}	-2.58	40.7	39.8	95.6	-0.54	24.1	24.2	95.5
α_{X_4}	-2.98	50.3	47.9	95.5	0.17	29.2	29.0	95.2
Poisson								
α_{X_1}	-0.51	23.2	23.2	96.7	-0.12	14.0	14.1	95.9
α_{X_2}	1.38	22.3	22.1	95.5	0.66	12.9	13.4	96.0
α_{X_3}	0.28	38.7	39.3	95.6	-0.72	23.7	24.1	95.7
α_{X_4}	-3.51	45.5	47.2	97.3	-1.03	29.0	28.9	95.4

n , size of reference data; SD, standard deviation; ESE, estimated standard error; CP, coverage probability of 95% confidence intervals.

Table A7. Simulation results (multiplied by 100) for the estimates of the coefficient β_Z of the treatment variable in the full outcome model and the ATEs for the populations of the existing datasets under the binomial, normal, or Poisson outcome distribution; true value of $\beta_z = \log(1.5) = 0.41$ (large sample case with sample size of 1000).

	Bias	$n = 1000$ SD	ESE	CP
Binomial				
$\beta_Z(\log(1.5) = 0.41)$	0.22	12.3	12.4	95.5
ATE1 (0.08)	0.04	2.39	2.41	94.9
ATE2 (0.06)	0.05	1.90	1.93	94.5
ATE3 (0.04)	0.06	1.42	1.45	94.6
ATE4 (0.03)	0.06	1.03	1.04	95.3
Normal				
$\beta_Z(\log(1.5) = 0.41)$	-0.08	10.1	9.95	94.1
ATE1 (0.41)	-0.08	10.1	9.95	94.1
ATE2 (0.41)	-0.08	10.1	9.95	94.1
ATE3 (0.41)	-0.08	10.1	9.95	94.1
ATE4 (0.41)	-0.08	10.1	9.95	94.1
Poisson				
$\beta_Z(\log(1.5) = 0.41)$	0.19	3.80	3.86	95.1
ATE1 (1.13)	0.42	11.7	12.0	96.2
ATE2 (0.68)	0.41	7.42	7.61	95.3
ATE3 (0.41)	0.35	4.76	4.85	95.5
ATE4 (0.25)	0.16	3.02	3.06	95.9

n , size of reference data; SD, standard deviation; ESE, estimated standard error; CP, coverage probability of 95% confidence intervals.

Table A8. Simulation results (multiplied by 100) for the estimates of the coefficients ($\alpha_{X_1}, \alpha_{X_2}, \alpha_{X_3}, \alpha_{X_4}$) of the covariate variables in the full propensity score model with true parameter values ($\alpha_{X_1}, \alpha_{X_2}, \alpha_{X_3}, \alpha_{X_4} = (\log(1.2), \log(1.5)), -\log(1.2), -\log(1.5)$), under the binomial, normal, or Poisson outcome distribution; true value of $\beta_z = \log(1.5) = 0.41$ (large sample case with sample size of 1000).

	Bias	$n = 1000$ SD	ESE	CP
Binomial				
α_{X_1}	-0.33	9.80	9.90	95.8
α_{X_2}	0.90	9.46	9.51	95.2
α_{X_3}	0.37	17.17	16.89	94.9
α_{X_4}	-1.31	20.23	20.27	95.4
Normal				
α_{X_1}	-0.38	9.82	9.86	94.8
α_{X_2}	-0.02	9.14	9.45	96.1
α_{X_3}	0.09	16.61	16.86	95.6
α_{X_4}	0.38	20.27	20.18	94.9
Poisson				
α_{X_1}	0.04	9.53	9.84	95.6
α_{X_2}	-0.11	9.08	9.34	95.6
α_{X_3}	0.06	16.6	16.8	95.6
α_{X_4}	0.32	20.3	20.2	94.3

n , size of reference data; SD, standard deviation; ESE, estimated standard error; CP, coverage probability of 95% confidence intervals.

References

1. Wang, D.; Zheng, S.; Cui, Y.; He, N.; Chen, T.; Huang, B. Adjusted win ratio using the inverse probability of treatment weighting. *J. Biopharm. Stat.* **2023**, *10*, 1–16. [[CrossRef](#)] [[PubMed](#)]
2. Liang, J.; Liu, J. Evaluation of educational interventions based on average treatment effect: A case study. *Mathematics* **2022**, *10*, 4333. [[CrossRef](#)]
3. Hsu, Y.C.; Lai, T.C.; Lieli, R.P. Estimation and inference for distribution and quantile functions in endogenous treatment effect models. *Econom. Rev.* **2020**, *41*, 22–50. [[CrossRef](#)]
4. Yang, S.; Ding, P. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika* **2018**, *105*, 487–493. [[CrossRef](#)]
5. Rubin, D.B. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **1974**, *66*, 688–701. [[CrossRef](#)]
6. Rosenbaum, P.R.; Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **1983**, *70*, 41–55. [[CrossRef](#)]
7. Austin, P.C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar. Behav. Res.* **2011**, *46*, 399–422. [[CrossRef](#)] [[PubMed](#)]
8. D’Agostino, R.B. Tutorial in biostatistics propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat. Med.* **1998**, *17*, 2265–2281. [[CrossRef](#)]
9. Lunceford, J.K.; Davidian, M. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Stat. Med.* **2004**, *23*, 2937–2960. [[CrossRef](#)] [[PubMed](#)]
10. Athey, S.; Imbens, G.; Pham, T.; Wager, S. Estimating average treatment effects: Supplementary analyses and remaining challenges. *Am. Econ. Rev.* **2017**, *107*, 278–281. [[CrossRef](#)]
11. Simoneau, G.; Pellegrini, F.; Debray, T.P.; Rouette, J.; Muñoz, J.; Platt, R.W.; Petkau, J.; Bohn, J.; Shen, C.; de Moor, C.; et al. Recommendations for the use of propensity score methods in multiple sclerosis research. *Mult. Scler.* **2022**, *28*, 1467–1480. [[CrossRef](#)] [[PubMed](#)]
12. Taylor, S.A.; Phillips, K.J.; Gertzog, M.G. Use of synthesized analysis and informed treatment to promote school reintegration. *Behav. Interv.* **2018**, *33*, 364–379. [[CrossRef](#)]
13. Hamada, A. Using meta-analysis and propensity score methods to assess treatment effects toward evidence-based practice in extensive reading. *Front. Psychol.* **2020**, *11*, 617. [[CrossRef](#)] [[PubMed](#)]
14. Ren, Q.; Su, C.; Wang, H.; Wang, Z.; Du, W.; Zhang, B. Prospective study of optimal obesity index cut-off values for predicting incidence of hypertension in 18–65-year-old Chinese adults. *PLoS ONE* **2016**, *11*, e0148140. [[CrossRef](#)] [[PubMed](#)]
15. Hu, L.; Huang, X.; You, C.; Li, J.; Hong, K.; Li, P.; Wu, Y.; Wu, Q.; Bao, H.; Cheng, X. Prevalence and risk factors of prehypertension and hypertension in southern China. *PLoS ONE* **2017**, *12*, e0170238. [[CrossRef](#)] [[PubMed](#)]
16. Kundu, P.; Tang, R.; Chatterjee, N. Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika* **2019**, *106*, 567–585. [[CrossRef](#)] [[PubMed](#)]
17. Nelder, J.A.; Wedderburn, R.W.M. Generalized linear models. *J. R. Stat. Soc. Ser. A Stat. Soc.* **1972**, *135*, 370–384. [[CrossRef](#)]
18. Hansen, L.P. Large sample properties of generalized method of moments estimators. *Econometrica* **1982**, *50*, 1029–1054. [[CrossRef](#)]
19. Lean, J.; Han, T.S.; Seidell, J.C. Impairment of health and quality of life in people with large waist circumference. *Lancet* **1998**, *351*, 853–856. [[CrossRef](#)] [[PubMed](#)]
20. Guagnano, M.T.; Ballon, E.; Colagrande, V.; Della Vecchia, R.; Manigrasso, M.R.; Merlitti, D.; Riccioni, G.; Sensi, S. Large waist circumference and risk of hypertension. *Int. J. Obes.* **2001**, *25*, 1360–1364. [[CrossRef](#)]
21. Li, H.; Miao, W.; Cai, Z.; Liu, X.; Zhang, T.; Xue, F.; Geng, Z. Causal data fusion methods using summary-level statistics for a continuous outcome. *Stat. Med.* **2020**, *39*, 1054–1067. [[CrossRef](#)] [[PubMed](#)]
22. Stümer, T.; Schneeweiss, S.; Avorn, J.; Glynn, R.J. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *Am. J. Epidemiol.* **2005**, *162*, 279–289. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.