SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Original Research Article

Statistical Methods in Medical Research
2023, Vol. 32(2) 229–241

DOI: 10.1177/09622802221133555
journals.sagepub.com/home/smm
SAGE

Unit information prior for incorporating real-world evidence into randomized controlled trials

Hengtao Zhang and Guosheng Yin

Abstract

Randomized controlled trials (RCTs) have been widely recognized as the gold standard to infer the treatment effect in clinical research. Recently, there has been growing interest in enhancing and complementing the result in an RCT by integrating real-world evidence from observational studies. The unit information prior (UIP) is a newly proposed technique that can effectively borrow information from multiple historical datasets. We extend this generic approach to synthesize the non-randomized evidence into a current RCT. Not only does the UIP only require summary statistics published from observational studies for ease of implementation, but it also has clear interpretations and can alleviate the potential bias in the real-world evidence via weighting schemes. Extensive numerical experiments show that the UIP can improve the statistical efficiency in estimating the treatment effect for various types of outcome variables. The practical potential of our UIP approach is further illustrated with a real trial of hydroxychloroquine for treating COVID-19 patients.

Keywords

Clinical trials, evidence synthesis, informative prior, observational studies, summary statistics

1 Introduction

Randomized controlled trials (RCTs) are widely recognized as the gold standard to evaluate the efficacy of a treatment in clinical studies. However, the rigorous experimental settings and inclusion criteria restrict the generalisability of trial results to the routine clinical practice.[1] It is also costly and slow for RCTs to deliver the findings if a large number of subjects are required to be followed for a long period of time. In order to complement and enhance the evidence in an RCT, information synthesis from the ubiquitous real-world evidence or observational studies into therapeutic development has recently received growing attention in the medical literature.[2–4] Motivated by the above concerns, a variety of statistical methods have been proposed to integrate the non-randomized and randomized evidence, most of which are based on the Bayesian framework due to its prominent flexibility in combining multi-source information. When only aggregate data are available for both RCTs and observational studies, the real-world evidence is typically quantified using meta-analysis techniques,[5–8] such as through the prior distribution of the treatment effect parameter in a network meta-analysis model, or through a multi-level Bayesian hierarchical model. The aforementioned methods are typically designed for aggregating the published summary statistics from all studies, including the treatment effect estimates with the corresponding standard errors, and thus they cannot be directly applied to the scenario where the individual patient data are available for a current RCT.

For the cases where investigators have access to patient-level data of both randomized and non-randomized studies, many approaches have been proposed for evidence synthesis following the idea of propensity scores.[9] Lin et al.[10,11] leveraged the observational samples to augment the disproportionate RCT into a 1:1 randomized trial, and then discounted the

Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, China

Corresponding author:
Guosheng Yin, Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, China.
Email: gyin@hku.hk

borrowed samples via propensity score-based priors. Wang et al.[12] proposed a Bayesian nonparametric approach to jointly modeling the outcomes and the corresponding propensity scores from different types of studies. Liu et al.[13] introduced a stratum-specific meta-predictive-analytic prior to including the observational evidence following propensity scores. A three-stage framework was proposed to validate observational data with the RCT as an anchor and integrate the statistical evidence from both sources.[14] Although the individual patient data could allow flexible and effective synthesis of information, they are difficult to obtain for all observational studies in practice while typically only summary statistics are available in the published literature.

We hence focus on the situation where the investigator needs to infer the treatment effect from the patient-level data of a current RCT, while additional summary statistics can be obtained from publications of related observational studies to enhance the RCT results. We assume that the causal estimands are the same between the RCT and observational studies, such as the hazard ratio or odds ratio. This scenario is particularly common under the current COVID-19 pandemic, because the urgency of health care has promoted numerous observational analyses prior to the report released from a well-conducted RCT.[15,16] Some attempts have been made in the literature to deal with the synthesis for mixed data that consist of individual patient data and aggregate information, such as the simulated treatment comparison,[17] matching-adjusted indirect comparisons,[18] Bayesian aggregation[19] and Bayesian meta-analysis with individual data.[20,21] However, those approaches are either only applicable to some specific types of outcomes,[20,21] not pertinent to our scenario,[17,16,19] or unable to account for the potential bias of observational evidence caused by heterogeneity and confounding effects.[20] Recently, Jin and Yin[22] proposed the unit information prior (UIP) to borrow evidence from historical studies. This approach has several salient advantages to address the information synthesis in our scenario. First, the UIP only requires summary statistics of the historical trials, which is common in the published literature. Second, the UIP can account for the heterogeneity between the current and historical datasets through an appropriate weighting scheme, which helps to discount the biased information.[5,6] Furthermore, it automatically determines the amount of information for integration from different historical datasets. Finally, this method is generic and can be readily modified to accommodate various types of outcomes. However, Jin and Yin[22] focused on the development of UIP for binary and continuous data from historical RCTs rather than observational studies. It is thus of great interest to explore the application of this method to incorporate the historical observational evidence into a current RCT with the availability of individual patient data involving various covariates and different types of outcomes. Particularly, we construct the UIP for the causal parameter of treatment effect using the point estimators from the observational studies with their standard errors, which are invariant to the types of outcome variables. We further evaluate the performance of UIP on continuous, binary, and time-to-event outcomes under different configurations.

The remainder of this article is organized as follows. Section 2 introduces how the real-world evidence can be incorporated using the UIP in the estimation of the parameter of interest in an RCT. We present the settings and results of comprehensive numerical experiments in Section 3, where we find that UIP can effectively improve the statistical efficiency of estimation. In Section 4, we re-analyze a randomized trial of hydroxychloroquine for treating COVID-19 by synthesizing the evidence from several observational studies. We conclude with a brief discussion in Section 5.

2 Methodology

2.1 Problem formulation

Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top \in \mathbb{R}^{n\times d}$ denote the covariate matrix for n enrolled patients in a current RCT. We assume that the RCT only involves one treatment and one control arm, and accordingly let $Z = (Z_1, \ldots, Z_n)^\top$ be the binary allocation indicators with $Z_i \in \{0, 1\}$ for $i = 1, \ldots, n$. The ith patient is assigned to the treatment arm if $Z_i = 1$ and to the control arm otherwise. The allocation variable Z_i is generated from randomization. The outcome variables $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ are then collected based on the treatment assignment. Our methodology is general enough to consider different types of outcomes, including continuous, binary, and survival data. For the survival outcome, the variable Y_i refers to the observed follow-up time under an independent right-censoring assumption. We additionally observe the event indicators $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^\top$ with $\delta_i \in \{0, 1\}$ for the survival outcome, where we have $\delta_i = 1$ if the event occurs otherwise $\delta_i = 0$ indicating that Y_i is a censored observation. We thus have the individual patient data $\mathcal{D} = \{\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \delta\}$ for the current RCT data, where δ is fixed as a vector of 1s denoted by $\mathbf{1}$ (i.e. no censoring) under continuous and binary outcomes for convenience.

Given the observed RCT data $\mathcal{D}$, we consider a general modeling structure that can cover different types of outcomes

$$P(Y_i|X_i, Z_i, \delta_i) = f(Y_i, \delta_i, \beta_0 + \theta Z_i + \boldsymbol{\beta}^\top X_i), \tag{1}$$

which denotes the probabilistic model for treatment inference with β_0, θ, and β as parameters for the linear predictor.

The function $f(\cdot)$ is specified accordingly depending on the type of outcome Y. For the continuous outcome, we adopt a linear model

$$Y_i = \beta_0 + \theta Z_i + \beta^{\mathrm{T}} \mathbf{X}_i + \varepsilon_i,$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ and σ_ε^2 is the unknown variance. For the binary outcome, the logistic regression is considered with

$$\operatorname{logit}\{P(Y_i = 1|\mathbf{X}_i)\} = \beta_0 + \theta Z_i + \beta^{\mathrm{T}} \mathbf{X}_i.$$

As for the survival outcome, we use the Cox proportional hazards model with the hazard function $h(t|Z_i, \mathbf{X}_i)$ defined as

$$h(t|Z_i, \mathbf{X}_i) = h_0(t) \exp(\theta Z_i + \beta^{\mathrm{T}} \mathbf{X}_i),$$

where $h_0(t)$ is the unknown baseline hazard function. The parameter θ or its transformation is related to the treatment effect of interest, which represents the additive treatment effect for the continuous outcome, the odds ratio (OR) for the binary outcome, and the hazard ratio (HR) for the survival outcome. The parameter β_0 corresponds to the intercept under the linear and logistic models and β represents the coefficient vector for covariate effects. Our major goal is to make a comprehensive estimation of θ by synthesizing $\mathcal{D}$ and aggregate evidence from observational studies. Before the formal integration analysis is carried out, we should collect all the necessary aggregate data from relevant observational studies.

2.2 Aggregate data from observational studies

Suppose that we have K historical observational studies that are closely related to the current RCT. There are several fundamental criteria for selecting appropriate observational studies to be included in our synthesis analysis. First, all observational studies should conceptually consider the same clinical objective and intervention as the current RCT, and numerically adopt the same parameter to measure the treatment effect. For example, both observational studies and the current RCT use the HR for survival data to measure the treatment effect. Second, the study design of observational data should be similar to the RCT for controlling the heterogeneity among studies, for example, the inclusion criteria are similar so that participants should have a comparable range of ages and symptoms. Finally, we require that all covariates should be balanced for the estimation of treatment effect in these observational studies to reduce the bias caused by confounding effects. Typically, this refers to analysis based on propensity scores, such as through matching[23] or weighting[24] schemes, which has been widely adopted for analyzing observational data in the medical research.[25,26] All information for checking the criteria can be readily found in the abstract of medical publications.

Let θ_k denote the parameter of treatment effect in the kth observational study for $k = 1, \ldots, K$, which is the counterpart of θ in the current RCT. We only require three summary statistics for θ_k from each observational study. First, we should obtain the maximum likelihood estimator (MLE) $\hat{\theta}_k$ for θ_k and the corresponding standard error $SE(\hat{\theta}_k)$. For binary and survival outcomes, $\hat{\theta}_k$ is obtained by conducting the logarithmic transformation on OR and HR, respectively. The standard error can be derived from the confidence interval (CI) if it is not directly provided in the published literature. Assume that the 95% CI is $[l_k, u_k]$ for the kth study, and we can calculate the standard error by $SE(\hat{\theta}_k) = \{g(l_k) - g(u_k)\}/(2 \times 1.96)$, where $g(\cdot)$ is a transformation function based on the type of outcomes. For example, we have $g(\cdot) = \log(\cdot)$ if $\exp(\theta_k)$ corresponds to OR or HR. Sometimes, the published medical papers may provide multiple estimators for θ_k based on different covariate-balancing approaches,[27] while we only adopt the results from the primary analysis for integration. Second, we need the corresponding sample size n_k used in the primary analysis, which should be distinguished from the original sample size in the observational study. For instance, we should specify n_k as the sample size of matched data if the primary analysis uses a matching method to balance covariates. We use $\mathcal{S}_k = \{\hat{\theta}_k, SE(\hat{\theta}_k), n_k\}$ to denote the summary information in the primary analysis of the kth observational study. After collecting $\mathcal{S}_1, \ldots, \mathcal{S}_K$, we can construct the UIP and then conduct Bayesian inference for θ in the current RCT.

2.3 Unit information prior

Let $\mathcal{D}_k$ denote the underlying individual patient data corresponding to $\mathcal{S}_k$, and $L^{(k)}(\theta_k|\mathcal{D}_k)$ refers to the likelihood function for θ_k. According to Jin and Yin,[22] we first introduce the unit information $I_U(\theta_k)$ for θ_k,

$$I_U(\theta_k) = -\frac{1}{n_k} \frac{\partial^2 \log L^{(k)}(\theta_k|\mathcal{D}_k)}{\partial \theta_k^2},$$

which represents the observed Fisher information at a unit sample level. The UIP to θ in the current RCT is then defined as follows:

$$\theta \big| (M, w_1, \ldots, w_K, \mathcal{S}_1, \ldots, \mathcal{S}_K) \sim N\left(\sum_{k=1}^{K} w_k \hat{\theta}_k, \left\{ M \sum_{k=1}^{K} w_k I_U(\hat{\theta}_k) \right\}^{-1} \right),$$

$$(w_1, \ldots, w_K) \sim \text{Dirichlet}(\gamma_1, \ldots, \gamma_K),$$

$$M \sim \text{Uniform}\left(0, \min\left\{ n, \sum_{k=1}^{K} n_k \right\} \right),$$

where $I_U(\hat{\theta}_k)$ refers to the unit information evaluated at $\hat{\theta}_k$, and $(\gamma_1, \ldots, \gamma_K)$ are pre-specified hyper-parameters with $\gamma_k = \min(1, n_k/n)$ so that a study with a larger sample size would be assigned a higher prior weight. We tend to assign a smaller weight in the Dirichlet prior for the dataset whose sample size is smaller than that of the RCT, and otherwise, the prior weight is bounded by 1. Intuitively, the UIP first incorporates the unit-level information from each observational study through the weight parameters w_k, and then figures out the total amount of information to be synthesized through the parameter M. Closely related to the effective sample size in the prior,[28] M is bounded by the sample size of the RCT (n) as well as the total sample size of matched data in the observational studies ($\sum_{k=1}^{K} n_k$), because the UIP should not dominate the current RCT. Both M and w_k's are adaptively determined by the commensurability between the RCT and observational studies.

The only concern now is to specify $I_U(\hat{\theta}_k)$ for the kth observational study. Jin and Yin[22] suggested performing a case-by-case derivation to obtain the explicit form of $I_U(\hat{\theta}_k)$ based on the type of outcome and the corresponding model. By contrast, we provide a more general alternative to determine the unit information. Following the asymptotic normality of the MLE, we have

$$\sqrt{n_k}(\hat{\theta}_k - \theta_k) \rightarrow N(0, 1/\mathbb{E}\{I_U(\theta_k)\}),$$

where the expectation is taken over the distribution that generates independent individual samples in $\mathcal{D}_k$. The above property shows that the asymptotic variance of $\hat{\theta}_k$ is $1/[n_k \mathbb{E}\{I_U(\theta_k)\}]$, and it thus motivates us to define $I_U(\hat{\theta}_k)$ using the empirical standard error of $\hat{\theta}_k$

$$I_U(\hat{\theta}_k) = \frac{1}{n_k SE^2(\hat{\theta}_k)}.$$

Finally, we adopt non-informative prior (NIP) for the other parameters specified in model (1) and then perform Bayesian inference.

There are several advantages of incorporating real-world evidence using UIP. First, we only need the summary statistics $\mathcal{S}_k$ to construct the prior without the need for access to the individual data $\mathcal{D}_k$. Second, the specification of UIP is independent of the outcome model, which greatly facilitates its broad applicability. Most importantly, the UIP accounts for the potential bias in each observational study by a weighting scheme. Different from RCTs, the estimator obtained from an observational study cannot avoid the bias that originates from the non-randomized treatment allocation, although the bias can be reduced by some post-hoc balancing techniques. Furthermore, the bias may also originate from the heterogeneity between observational studies and the current RCT. Therefore, properly discounting the observational evidence plays an essential role in the synthesis to adjust the possible over-estimation or under-estimation of the treatment effect.[5–7] In the UIP, we have

$$w_k I_U(\hat{\theta}_k) = \frac{1}{n_k \left\{ SE^2(\hat{\theta}_k)/w_k \right\}}$$

for each study, where $w_k \in (0, 1)$ inflates the standard error $SE^2(\hat{\theta}_k)$ to alleviate the underlying unamendable bias. A larger weight w_k leads to a smaller standard error $SE^2(\hat{\theta}_k)/w_k$, which leads to more information borrowing through the UIP.

3 Simulation

In the simulation, we evaluate the performance of the aforementioned UIP in estimating the treatment effect for the current RCT with respect to different types of outcomes. Individual patient data are first simulated for both the RCT and observational studies, and the propensity score analysis is performed on the observational data to generate the summary statistics for further synthesis.

3.1 Continuous outcome

For continuous outcomes, we fix the sample size of the current RCT at $n = 200$, and generate the covariates $\mathbf{X}_i$ and allocation indicator Z_i for the ith sample ($i = 1, \ldots, n$) as follows:

$$\mathbf{X}_i \sim N(\mathbf{0}, (1-\rho)\mathbf{I}_d + \rho\mathbf{1}_d\mathbf{1}_d^\top),$$
$$Z_i \sim \text{Bernoulli}(0.5),$$

where we set the correlation coefficient $\rho = 0.1$ and the covariate dimension $d = 6$. The first two covariates ($X_{ij}, j = 1, 2$) are then dichotomized to 1 if it is non-negative and to 0 otherwise. Given the covariates $\mathbf{X}_i$ and the assignment indicators Z_i, we generate the outcome variables Y_i of RCT through a linear model,

$$Y_i = 1 + \theta Z_i + \beta^\top \mathbf{X}_i + \varepsilon_i,$$

where $\theta = 1$, $\beta = \mathbf{1}_d$, and $\varepsilon_i \sim N(0, 0.5^2)$.

For the evidence from observational studies, we consider $K = 3$ historical observational studies and fix the sample size $N_k = 500$ for the kth study with $k = 1, 2, 3$. It is worth emphasizing that N_k is the original sample size of the observational studies and n_k in the UIP is the sample size after matching the propensity scores to balance the covariates, so $n_k \leq N_k$. Similarly, we first generate the covariates from a normal distribution with $d_k = 6$,

$$\mathbf{X}_i^{(k)} \sim N(\mu_k\mathbf{1}_{d_k}, \sigma_k^2(1-\rho_k)\mathbf{I}_{d_k} + \sigma_k^2\rho_k\mathbf{1}_{d_k}\mathbf{1}_{d_k}^\top),$$

where μ_k and σ_k^2 are the covariate mean and variance, respectively, and ρ_k denotes the correlation coefficient. The first two covariates ($X_{ij}^{(k)}, j = 1, 2$) are then dichotomized to 1 if its value is larger than μ_k and to 0 otherwise. For the observational studies, the treatment indicator variables depend on covariates or confounders, which are generated from the following model:

$$Z_i^{(k)} \sim \text{Bernoulli}\left\{P\left(Z_i^{(k)} = 1|\mathbf{X}_i^{(k)}\right)\right\},$$
$$\text{logit}\left\{P\left(Z_i^{(k)} = 1|\mathbf{X}_i^{(k)}\right)\right\} = \boldsymbol{\eta}^{(k)\top}\mathbf{X}_i^{(k)} = \sum_{j=1}^{d_k} \eta_j^{(k)} X_{ij}^{(k)},$$

where $P(Z_i^{(k)} = 1|\mathbf{X}_i^{(k)})$ is the propensity score with the coefficient vector $\eta^{(k)} = (\eta_1^{(k)}, \ldots, \eta_{d_k}^{(k)})^\top$. After collecting the covariate matrix and the allocation indicators, we generate the continuous outcome for the kth observational study via

$$Y_i^{(k)} = 1 + \theta_k Z_i^{(k)} + \beta^{(k)\top}\mathbf{X}_i^{(k)} + \varepsilon_i,$$

where $\varepsilon_i \sim N(0, 0.5^2)$, θ_k and $\beta^{(k)}$ are the parameters of the treatment effect and covariate coefficients, respectively.

3.2 Binary outcome

Regarding binary outcomes, the covariates and treatment indicators of the RCT and three observational studies are generated following the same procedure as that for continuous outcomes. Under a logistic model

$$\text{logit}\left\{P(Y_i = 1|\mathbf{X}_i)\right\} = \theta Z_i + \beta^\top \mathbf{X}_i,$$

with the coefficients $\beta = 0.1 \times \mathbf{1}_d$, the outcome of RCT is then simulated as

$$Y_i \sim \text{Bernoulli}\left\{P(Y_i = 1|\mathbf{X}_i)\right\}.$$

For the three observational studies, we simulate $Y_i^{(k)}$ for $k = 1, 2, 3$ as follows:

$$Y_i^{(k)} \sim \text{Bernoulli}\left\{P(Y_i^{(k)} = 1|\mathbf{X}_i^{(k)})\right\},$$
$$\text{logit}\left\{P(Y_i^{(k)} = 1|\mathbf{X}_i^{(k)})\right\} = \theta_k Z_i^{(k)} + \beta^{(k)\top}\mathbf{X}_i^{(k)},$$

where the treatment indicator $Z_i^{(k)}$ is confounded with covariates $\mathbf{X}_i^{(k)}$ due to non-randomization.

3.3 Survival outcome

For survival data, we generate the covariates and treatment indicators of the RCT and historical data in the same manner. Following the Cox proportional hazards model, we assume a Weibull baseline hazard function $h_0(t) = st^{s-1}/2^s$, where the values of the shape parameter $s = 0.5, 1, 2$ correspond to decreasing, constant and increasing hazards. We generate the underlying event time $T_{E,i}$ for the RCT as follows:

$$T_{E,i} \sim F(t) = 1 - \exp\{-(t^s/2^s)\exp(\theta Z_i + \beta^{\mathrm{T}}\mathbf{X}_i)\},$$

where the coefficient vector β is kept as $0.1 \times \mathbf{1}_d$. The censoring time $T_{C,i}$ is generated as $T_{C,i} \sim \text{Uniform}(0, C)$, where the constant parameter C is determined to achieve a censoring rate of 10%. We set the observed time $Y_i = \min\{T_{E,i}, T_{C,i}\}$ and the censoring indicator $\delta_i = I(T_{E,i} \le T_{C,i})$, where $I(\cdot)$ denotes the indicator function.

The survival outcome in the kth observational study is similarly simulated from the proportional hazards model

$$T_{E,i}^{(k)} \sim F(t) = 1 - \exp\left\{-(t^{s_k}/2^{s_k})\exp(\theta_k Z_i^{(k)} + \beta^{(k)\mathrm{T}}\mathbf{X}_i^{(k)})\right\},$$

where the parameter s_k determines the shape of the baseline hazard function. The censoring time is generated as $T_{C,i}^{(k)} \sim \text{Uniform}(0, C^{(k)})$, where $C^{(k)}$ is chosen to determine the censoring rate. The observed time is $Y_i^{(k)} = \min\{T_{E,i}^{(k)}, T_{C,i}^{(k)}\}$, and the censoring indicator is $\delta_i^{(k)} = I(T_{E,i}^{(k)} \le T_{C,i}^{(k)})$. For simplicity, we set $s_1 = s_2 = s_3 = s$ and specify $C^{(k)}$ to ensure the censoring rate to be 10% for $k = 1, 2, 3$.

3.4 Parameter settings for observational data generation

We consider five different scenarios for the parameters μ_k, σ_k^2, θ_k and $\beta^{(k)}$ in the observational studies as follows:

Scenario 1: Let $(\mu_1, \mu_2, \mu_3) = (0, 0, 0)$, $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1, 1, 1)$, $(\rho_1, \rho_2, \rho_3) = (0.1, 0.1, 0.1)$, and $(\theta_1, \theta_2, \theta_3) = (1, 1, 1)$. Set $\beta^{(k)} = \mathbf{1}_6$ for continuous outcomes and $\beta^{(k)} = 0.1 \times \mathbf{1}_6$ for both the binary and survival outcomes with $k = 1, 2, 3$.
Scenario 2: Let $(\mu_1, \mu_2, \mu_3) = (0, 0, 0)$, $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1, 1, 1)$, $(\rho_1, \rho_2, \rho_3) = (0.1, 0.1, 0.1)$, and $(\theta_1, \theta_2, \theta_3) = (0.8, 1.1, 1.3)$. Generate $\beta^{(k)} \sim N(\mathbf{1}_6, \mathbf{I}_6)$ for continuous outcomes and $\beta^{(k)} \sim N(0.1 \times \mathbf{1}_6, 0.1^2 \times \mathbf{I}_6)$ for both the binary and survival outcomes with $k = 1, 2, 3$.
Scenario 3: Let $(\mu_1, \mu_2, \mu_3) = (0.5, 1, 1.5)$, $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (0.5, 1.5, 2)$, $(\rho_1, \rho_2, \rho_3) = (0.141, 0.071, 0.082)$, and $(\theta_1, \theta_2, \theta_3) = (0.8, 1.1, 1.3)$. Generate $\beta^{(k)} \sim N(\mathbf{1}_6, \mathbf{I}_6)$ for continuous outcomes and $\beta^{(k)} \sim N(0.1 \times \mathbf{1}_6, 0.1^2 \times \mathbf{I}_6)$ for both the binary and survival outcomes with $k = 1, 2, 3$.

Scenario 1 considers the ideal situation when the observational data share the same treatment effect and patient population as the RCT data except that the participants are not randomized during the allocation in the observational studies. Scenario 2 makes the treatment effects in the observational studies slightly different from that in the RCT. In Scenario 3, we simulate more realistic cases where the observational data are collected from related studies with similar treatment effects but slightly different patient populations in terms of covariate means and variances. Furthermore, we consider two scenarios for continuous and binary outcomes to demonstrate how our method performs in some extreme situations.

Scenario 4: Let $(\theta_1, \theta_2, \theta_3) = (0.1, 0.5, 0.9)$ for continuous outcomes and $(\theta_1, \theta_2, \theta_3) = (0.5, 2.5, 4.5)$ for binary outcomes. Values of other parameters are kept the same as those in Scenario 3.
Scenario 5: Let $(\theta_1, \theta_2, \theta_3) = (2, 3, 4)$ for continuous outcomes and $(\theta_1, \theta_2, \theta_3) = (5, 7, 9)$ for binary outcomes. Values of other parameters are kept the same as those in Scenario 3.

Scenario 4 considers the case where the observational data are not strictly selected, because the treatments adopted in the observational studies are not identical and θ_k's are different from each other as well as deviating from the RCT treatment effect $\theta = 1$. Scenario 5 mimics the situation where the drugs investigated in the observational studies may not even be related to the one in the RCT because treatment effects are substantially off the target. Finally, we fix the coefficient vectors $\eta^{(k)}$'s for the propensity scores for each observational study as follows:

$$\text{Observational study 1}: \quad \boldsymbol{\eta}^{(1)} = (0,\ 0,\ 0.2,\ 0.2,\ -0.2,\ -0.2)^{\top}.$$

$$\text{Observational study 2}: \quad \boldsymbol{\eta}^{(2)} = (1,\ 1,\ 1,\ 1,\ -1,\ -1)^{\top}.$$

$$\text{Observational study 3}: \quad \boldsymbol{\eta}^{(3)} = (2,\ 2,\ 2,\ 2,\ -2,\ -2)^{\top}.$$

Figure 1 shows the distributions of propensity scores for three observational datasets under different generation scenarios. A wider distributional gap or separation between the treatment and control groups indicates a stronger confounding effect, which means the covariates in the corresponding dataset are more difficult to balance.

3.5 Aggregate data generation

Through the above data generation procedure, we can obtain the patient-level datasets $\mathcal{D}_k$ for $k = 1, 2, 3$. We then obtain the aggregate data $\mathcal{S}_k$ summarized from $\mathcal{D}_k$ by (i) fitting a logistic model to estimate the propensity scores; (ii) using the nearest neighbor matching (Match)[23] or inverse probability weighting (IPW)[24] to balance the observational data; and (iii) conducting inference over the balanced dataset. In the second step, we implement Match and IPW using `MatchIt`[29] and `WeightIt`[30] packages, respectively. For the kth aggregate data $\mathcal{S}_k = \{\hat{\theta}_k, SE(\hat{\theta}_k), n_k\}$, the adjusted sample size n_k is the number of samples after the propensity score matching whereas we take $n_k = N_k$ for IPW. Furthermore, the MLEs $\hat{\theta}_k$'s for different types of outcomes are accordingly obtained using the functions `lm()`, `glm()` and `coxph()` in the R language.[31] Finally, we adopt the robust sandwich standard error to calculate $SE(\hat{\theta}_k)$ following the instructions in `MatchIt` and `WeightIt`.

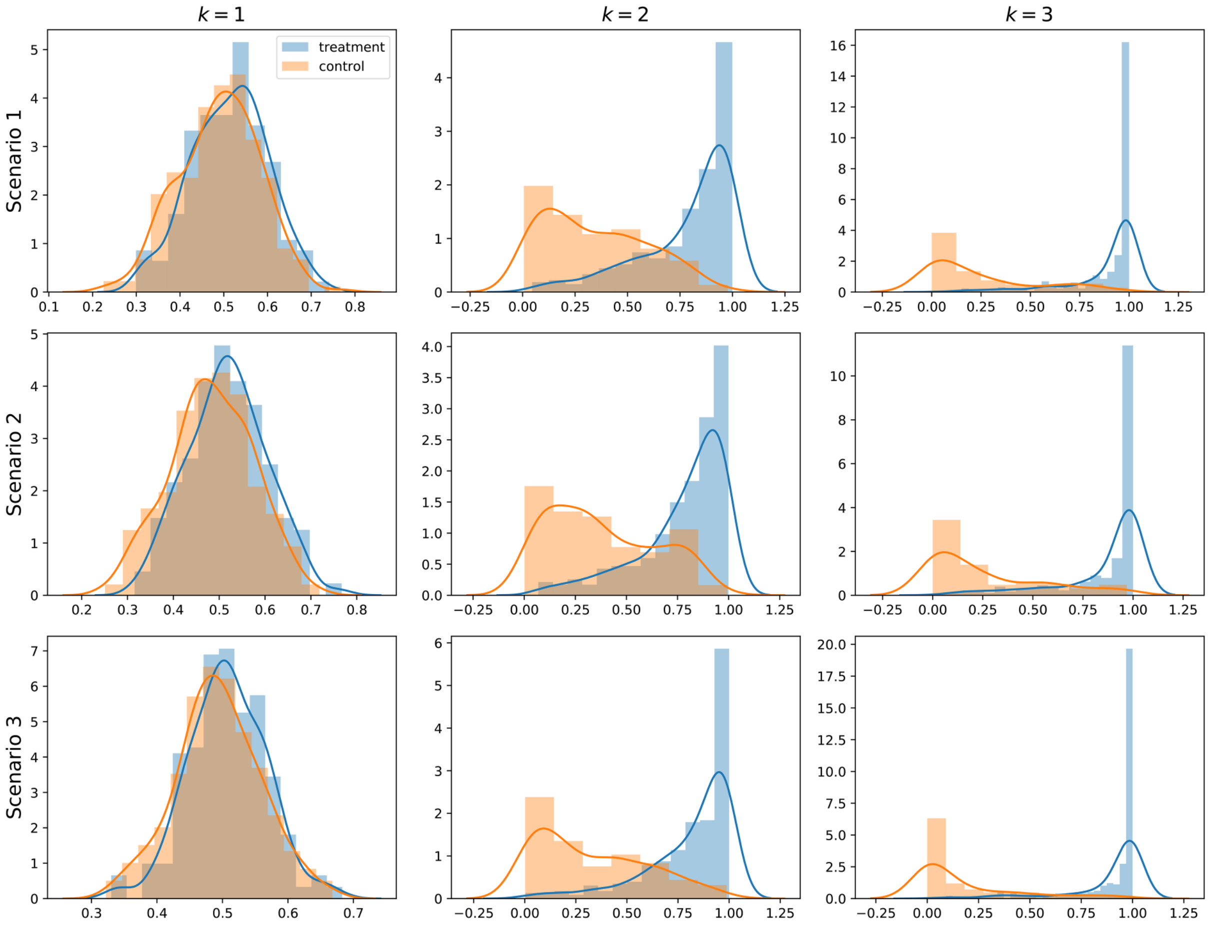

Figure 1. Distributions of propensity scores under different scenarios for each observational dataset.

3.6 Bayesian inference

After collecting the RCT data and aggregrating information from observational data, we conduct the Bayesian inference using the models in Section 2.1. Bayesian linear regression and logistic regression models are fitted to the continuous and binary outcomes, respectively. As for the survival outcome, we use the piecewise exponential proportional hazards model, where the baseline hazard function $h_0(t)$ is modeled as the piecewise constant function with the partition points of the time axis as $0 < \tau_1 < \ldots < \tau_J$. We set the number of intervals to be $J = 10$, and $\{\tau_1, \ldots, \tau_J\}$ are chosen based on the corresponding $100(j/J)$ quantiles ($j = 1, \ldots, J$) of observed event times. Within each interval $(\tau_{j-1}, \tau_j]$, we assume a constant hazard λ_j, which is assigned an NIP Gamma(0.01, 0.01). Under the linear model, the priors for β_0, β, and σ_ε^2 are set to be $\beta_0 \sim N(0, 100^2)$, $\beta \sim N(0, 100^2 \times \mathbf{I}_d)$, and $\sigma_\varepsilon^2 \sim$ Inv-Gamma(0.01, 0.01), and we adopt the same priors for the corresponding parameters in the logistic and Cox models. For the prior treatment effect parameter θ, we compare UIP with the NIP, which is simply defined as $\theta \sim N(0, 100^2)$.

All Bayesian models are implemented by PyMC3,[32] and 5,000 posterior samples are drawn for inference using the Markov chain Monte Carlo with 1,000 burn-in iterations. We simulate 200 repetitions for each configuration of data generation. Based on the repetitions, we evaluate UIP and NIP in terms of the empirical bias, root mean squared error (RMSE), and the width and coverage of the 95% credible interval (CrI) for the parameter θ. The 95% CrI is constructed based on 2.5% and 97.5% quantiles from the posterior samples.

3.7 Results

Table 1 presents the simulation results for the continuous outcome. Because NIP does not require observational data, its results are the same across different scenarios. Considering the UIP, we observe that it can effectively improve the precision of estimating θ in terms of RMSE under various scenarios and both balancing approaches. In addition, UIP reduces the bias under Scenario 1 when the observational data share the same treatment effect with RCT, and only incurs a minor increase in bias under Scenarios 2 and 3 when there exist heterogeneous treatment effects and covariates between the RCT and observational studies. For the CrI, UIP yields shorter widths and slightly larger values of coverage probability compared with NIP for the first three scenarios, which is due to the precise estimation of the treatment effect after synthesizing the observational evidence. Furthermore, UIP tends to borrow more information from $\mathcal{D}_1$ and $\mathcal{D}_2$ as demonstrated by larger values of w_1 and w_2 in Scenarios 2 and 3. It is due to the empirical fact that the first two datasets can more accurately estimate θ_1 and θ_2, which are closer to θ of the RCT, after balancing the relative less confounded covariates (e.g. $n_1 = 482$, $n_2 = 369$, and $n_3 = 366$ under Scenario 3). The values of w_k's are more distinctive when the deviations of θ_k's from θ vary substantially under Scenario 4. The total amount of borrowed information M is slightly above half of the RCT sample size ($n = 200$) for the UIP approaches when θ_k's are close to θ under Scenarios 1 to 3. A drastic decrease can be observed for M under Scenarios 4 and 5, where the treatment effects in the observational studies are more extreme relative to θ in the RCT. In Scenario 5, almost no information is borrowed from observational studies due to such a small value of $M = 2$, and thus UIP delivers performance close to NIP in terms of all evaluation metrics.

Table 1. The simulation results of different prior methods for continuous outcomes (linear regression) under three scenarios.

Scenario	Method	$\text{Bias}_{\times 100}$	$\text{RMSE}_{\times 100}$	Width	Coverage	w_1	w_2	w_3	M
	NIP	1.15	7.27	0.283	0.939				
1	UIP + Match	0.73	5.91	0.246	0.944	0.352	0.336	0.312	111
	UIP + IPW	0.61	5.98	0.246	0.949	0.353	0.340	0.307	111
2	UIP + Match	1.45	6.44	0.256	0.944	0.358	0.349	0.294	109
	UIP + IPW	1.34	6.43	0.256	0.949	0.356	0.351	0.293	109
3	UIP + Match	1.37	6.32	0.255	0.939	0.355	0.348	0.296	109
	UIP + IPW	1.37	6.29	0.256	0.939	0.361	0.349	0.290	108
4	UIP + Match	−2.19	7.44	0.273	0.924	0.140	0.293	0.567	72
	UIP + IPW	−2.03	7.41	0.274	0.939	0.144	0.304	0.552	73
5	UIP + Match	1.70	7.45	0.275	0.914	0.398	0.349	0.253	2
	UIP + IPW	1.66	7.41	0.275	0.929	0.395	0.347	0.258	2

CrI: credible interval; IPW: inverse probability weighting; Match: nearest neighbor matching; NIP: non-informative prior; RMSE: root mean squared error; UIP: unit information prior.

Note: We evaluate the performance of estimating the treatment effect θ using the bias, RMSE, the width and coverage of the 95% CrI. The weights of UIP for three observational studies are denoted by w_1, w_2, w_3, and M corresponds to the amount of information borrowed from observational data measured by the number of units.

The results for the binary outcome are shown in Table 2. Clearly, the UIP can reduce the RMSE under Scenarios 1 to 3, in contrast, to NIP with a minor sacrifice in bias. Narrower CrIs can be obtained under UIP with coverage probabilities closer to the nominal level. Under Scenarios 4 and 5, the first two datasets contribute more to the current RCT as evidenced by relatively larger values of w_1 and w_2 in comparison with w_3. The UIP borrows much less information in Scenario 5 as indicated by small values of M, because of substantial differences between the treatment effects of the observational studies and RCT. The UIP also improves the inference of θ for the survival outcome as shown by Table 3. The UIP can reduce both the bias and RMSE compared with NIP for decreasing ($s = 0.5$), constant ($s = 1$), and increasing ($s = 2$) baseline hazards, which shows the robustness of UIP with respect to different types of hazard functions. It also yields tighter CrIs for θ with coverage probabilities closer to the nominal level. In summary, UIP can enhance the inference of the RCT via integrating real-world evidence, and the improvement is more evident when the RCT and observational studies are similar.

Based on Tables 1 to 3, we can also gain some insight into the factors that affect w_k's and M. In general, the weights are determined by both treatment effects and the covariate confounding effects in the observational data relative to the RCT, and the impact of the treatment effect is more evident. In Scenario 1, all observational data share the same treatment effect with the RCT, that is, $\theta_1 = \theta_2 = \theta_3 = \theta$, but they have different degrees of covariate confounding effect. Since a greater confounding effect would lead to a larger standard error or a more biased treatment effect estimator, the first two datasets, $\mathcal{D}_1$ and $\mathcal{D}_2$, tend to be more informative and we thus observe $w_1 > w_2 > w_3$. In Scenario 2, although θ_2 is closer to θ than θ_1, we observe $w_2 < w_1$ in most cases, because the covariate confounding effect has a greater impact on the weights than the treatment effects when all θ_k's are close to θ. Scenarios 4 and 5 show that the observational treatment effects are more influential than the confounding effect when θ_k's deviate far from θ. Furthermore, by comparing Scenario 2 and Scenario 3, the patient population does not have a strong influence on the order of w_k's.

The parameter M is affected by the level of consistency between the observational studies and the current RCT,[22] for which the more inconsistency the smaller value of M. Scenarios 1 to 3 show that the patient population does not impose a strong influence on the values of M. Scenarios 4 and 5 show that M is sensitive to the dissimilarity between θ_k's and θ.

4 Real application

In the COVID-19 pandemic era, numerous studies have been conducted to find effective treatments. We illustrate the application of UIP by re-analyzing a randomized trial,[33] which was launched to evaluate the efficacy of hydroxychloroquine (HCQ) in reducing the in-hospital mortality of COVID-19 compared with the placebo. The parameter of interest is the HR between the HCQ and placebo groups. Due to inaccessibility to the original survival data, we use the R package `IPDfromKM`[34,35] to reconstruct the patient-level data from the Kaplan–Meier curves of the HCQ and placebo groups in Figure 2 of the original paper.[33] The reconstructed dataset contains 208 samples with 104 samples from the HCQ group and 104 from the control group. We obtain the outcome variable Y_i, treatment indicator Z_i, and the event indicator δ_i for the ith reconstructed observation.

Table 2. The simulation results of different prior methods for binary outcomes (logistic regression) under three scenarios.

Scenario	Method	Bias	RMSE	Width	Coverage	w_1	w_2	w_3	M
	NIP	0.061	0.346	1.238	0.929				
1	UIP + Match	0.075	0.285	1.153	0.965	0.348	0.327	0.325	112
	UIP + IPW	0.089	0.293	1.175	0.960	0.360	0.328	0.313	110
2	UIP + Match	0.092	0.291	1.162	0.975	0.349	0.331	0.321	111
	UIP + IPW	0.101	0.307	1.183	0.970	0.361	0.335	0.304	111
3	UIP + Match	0.081	0.289	1.163	0.975	0.347	0.327	0.325	112
	UIP + IPW	0.090	0.306	1.173	0.955	0.363	0.330	0.307	111
4	UIP + Match	0.207	0.376	1.233	0.894	0.527	0.304	0.169	96
	UIP + IPW	0.208	0.378	1.236	0.879	0.544	0.299	0.156	94
5	UIP + Match	0.136	0.375	1.290	0.929	0.548	0.261	0.191	12
	UIP + IPW	0.131	0.374	1.288	0.939	0.466	0.312	0.222	14

CrI: credible interval; IPW: inverse probability weighting; Match: nearest neighbor matching; NIP: non-informative prior; RMSE: root mean squared error; UIP: unit information prior.

Note: We evaluate the performance of estimating the treatment parameter θ using the bias, RMSE, the width and coverage of the 95% CrI. The weights of UIP for three observational studies are denoted by w_1, w_2, w_3, and M corresponds to the amount of information borrowed from observational data measured by the number of units.

Table 3. The simulation results of different prior methods for survival outcomes (Cox regression) under various cases, where s determines the shape of the baseline hazard function.

s	Scenario	Method	Bias	RMSE	Width	Coverage	w_1	w_2	w_3	M
0.5		NIP	0.062	0.191	0.665	0.919				
	1	UIP + Match	0.046	0.152	0.568	0.944	0.359	0.333	0.308	113
		UIP + IPW	0.044	0.156	0.577	0.949	0.370	0.330	0.300	113
	2	UIP + Match	0.035	0.154	0.576	0.944	0.337	0.343	0.320	112
		UIP + IPW	0.035	0.159	0.581	0.939	0.344	0.342	0.314	111
	3	UIP + Match	0.034	0.154	0.573	0.939	0.342	0.341	0.317	112
		UIP + IPW	0.031	0.157	0.582	0.955	0.349	0.336	0.315	111
1		NIP	0.053	0.184	0.664	0.924				
	1	UIP + Match	0.037	0.149	0.576	0.955	0.356	0.332	0.312	114
		UIP + IPW	0.035	0.156	0.592	0.949	0.369	0.332	0.299	114
	2	UIP + Match	0.048	0.153	0.580	0.955	0.356	0.335	0.308	113
		UIP + IPW	0.044	0.154	0.588	0.955	0.365	0.335	0.300	113
	3	UIP + Match	0.047	0.155	0.578	0.955	0.358	0.334	0.309	113
		UIP + IPW	0.039	0.153	0.586	0.965	0.361	0.334	0.305	113
2		NIP	0.069	0.184	0.678	0.939				
	1	UIP + Match	0.054	0.148	0.584	0.975	0.357	0.332	0.311	114
		UIP + IPW	0.054	0.153	0.592	0.949	0.368	0.330	0.302	114
	2	UIP + Match	0.060	0.153	0.585	0.955	0.348	0.338	0.314	114
		UIP + IPW	0.055	0.153	0.589	0.955	0.356	0.340	0.304	113
	3	UIP + Match	0.063	0.155	0.589	0.944	0.352	0.338	0.310	114
		UIP + IPW	0.059	0.155	0.597	0.960	0.360	0.338	0.301	113

CrI: credible interval; IPW: inverse probability weighting; Match: nearest neighbor matching; NIP: non-informative prior; RMSE: root mean squared error; UIP: unit information prior.

Note: We evaluate the performance of estimating the treatment parameter θ using the bias, RMSE, the width and coverage of the 95% CrI. The weights of UIP for three observational studies are denoted by w_1, w_2, w_3, and M corresponds to the amount of information borrowed from observational data measured by the number of units.

Based on the reconstructed data, we re-estimate the HR of receiving HCQ, respectively, using NIP and UIP under the piecewise exponential model without covariates, i.e.that is, $\beta = 0$. In addition to the Bayesian approaches, we also adopt the frequentist Cox proportional hazards model via the R function `coxph()` to estimate the HR. For the UIP, we collect the aggregate information from four different observational studies denoted by their respective published journals: *PLoS One* (*PLoS*),[36]*New England Journal of Medicine* (*NEJM*),[27]*American Journal of Epidemiology* (*AJE*),[37] and *Clinical Infectious Diseases* (*CID*).[38] The four observational studies are selected following the criteria detailed in Section 2.2 with the aforementioned RCT as the template. First, those studies all considered the effect of HCQ therapy in reducing the mortality of COVID-19 and used the HR to measure the treatment efficacy. Second, similar to the RCT, the enrolled samples in those studies were all non-pregnant hospitalized patients who had been diagnosed with COVID-19 and did not participate in other trials, and the majority of them were middle-aged or older people. Finally, those studies all fitted the Cox proportional hazards model with the propensity score adjustment to estimate the HR, where the propensity score weighting was used for *NEJM*, *AJE*, and *CID*, while *PLoS* leveraged the propensity score stratification. The summary information provided in the published papers includes sample size, the estimated HR $\exp(\hat{\theta}_k)$, and its 95% CI. We perform the logarithmic transformation to obtain $\mathcal{S}_k = \{\hat{\theta}_k, \mathrm{SE}(\hat{\theta}_k), n_k\}$, $k = 1, \ldots, 4$. The UIP and NIP are then, respectively, applied to the Bayesian model following the same configurations used in the simulation.

We report the estimated HR for the reconstructed RCT data and the summary information of the historical studies in Table 4. The estimated HR under UIP is 0.96 (95% CrI, 0.62–1.41), which has a narrower CrI than that under NIP, 0.94 (95% CrI, 0.58–1.44), and the frequentist counterpart, 0.93 (95% CI, 0.59–1.45), thus demonstrating the effectiveness of historical information borrowing. The parameter M is estimated to be 122 for UIP, indicating that UIP synthesizes an intermediate amount of observational information in contrast with the sample size of the randomized trial. The UIP weights for (*PLoS*, *NEJM*, *AJE*, *CID*) are $(w_1, \ldots, w_4) = (0.295, 0.307, 0.284, 0.114)$, suggesting that the least amount of information is borrowed from *CID* and more information is incorporated from *NEJM* and *PLoS*. The sample sizes of the two studies in *NEJM* and *PLoS* are both large and their estimated HRs have narrower CIs and, as a result, UIP tends to borrow more evidence from more reliable studies. In summary, the re-analyzed HR based on UIP concludes that there is no

Table 4. The estimated HR with the CI/CrI for the reconstructed RCT and four historical observational studies using their respective journal names.

	Samples	HR	95% CI	CI Width	UIP Weights
PLoS	2512	1.02	(0.83, 1.27)	0.44	0.295
NEJM	1376	1.04	(0.82, 1.32)	0.50	0.307
AJE	998	1.21	(0.82, 1.76)	0.94	0.284
CID	84	0.89	(0.23, 3.47)	3.24	0.114
RCT	208	0.93	(0.59,1.45)	0.85	
			95% CrI	CrI Width	
RCT + NIP	208	0.94	(0.58, 1.44)	0.86	
RCT + UIP	208	0.96	(0.62, 1.41)	0.79	($M = 122$)

HR: hazard ratio; CrI: credible interval; CI: confidence interval; RCT: randomized controlled trial; NIP: non-informative prior; UIP: unit information prior. *Note*: Journal names: *PLoS*: *PLoS One*; *NEJM*: *New England Journal of Medicine*; *AJE*: *American Journal of Epidemiology*; *CID*: *Clinical Infectious Diseases*. The parameter M corresponds to the amount of information borrowed from observational data measured by the number of units.

significant improvement in the post-hospitalization mortality caused by COVID-19 for patients receiving HCQ, which is consistent with the claim in the original article but the evidence via incorporating the four observational studies is clearly more convincing.

5 Discussion

For a more comprehensive analysis of a current RCT, we develop the UIP methodology for synthesizing the evidence from historical observational studies into the analysis of the RCT data. The UIP is readily applicable to various common types of outcomes considered in clinical trials, and it only requires published summary information. The prior automatically quantifies the amount of information borrowed from each study and simultaneously considers the unamendable bias in the real-world evidence. Extensive numerical experiments show that the UIP can improve the statistical efficiency of inference for the RCT. There are several promising directions to further extend the idea of UIP for information borrowing. For example, it is of interest to leverage the prior to incorporate the patient-level observational data into the single-arm trials as discussed by Liu et al.[13] Furthermore, one can investigate how UIP affects the type I error and power, and its possible application in determining the sample size of an RCT at the design stage. The UIP also sheds light on a meta-analysis from the information borrowing aspect, which warrants further investigation. Finally, one may consider the synthesis when individual patient-level data are available for both the RCT and observational studies, for which other existing methods, such as the power prior or commensurate prior, can also be applicable. The codes for the numerical studies are available at https://github.com/BobZhangHT/UIP4RCT.

Acknowledgements

We thank the two referees for their careful reviews of our manuscript and many insightful comments that have significantly improved the paper.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This research was partially supported by funding from the Research Grants Council of Hong Kong (17308321).

ORCID iDs

Hengtao Zhang https://orcid.org/0000-0003-0284-9748
Guosheng Yin https://orcid.org/0000-0003-3276-1392

Supplemental material

Supplementary material for this article is available online.

References

1. Rothwell PM. External validity of randomised controlled trials: "To whom do the results of this trial apply?". *Lancet* 2005; **365**: 82–93.
2. Sherman RE, Anderson SA, Dal Pan GJ et al. Real-world evidence—what is it and what can it tell us? *N Engl J Med* 2016; **375**: 2293–2297.
3. Katkade VB, Sanders KN and Zou KH. Real world data: An opportunity to supplement existing evidence for the use of long-established medicines in health care decision making. *J Multidiscip Healthc* 2018; **11**: 295–304.
4. Schmidli H, Häring DA, Thomas M et al. Beyond randomized clinical trials: Use of external controls. *Clin Pharmacol Ther* 2020; **107**: 806–816.
5. Schmitz S, Adams R and Walsh C. Incorporating data from various trial designs into a mixed treatment comparison model. *Stat Med* 2013; **32**: 2935–2949.
6. Efthimiou O, Mavridis D, Debray TP et al. Combining randomized and non-randomized evidence in network meta-analysis. *Stat Med* 2017; **36**: 1210–1226.
7. Verde PE. A bias-corrected meta-analysis model for combining, studies of different types and quality. *Biom J* 2021; **63**: 406–422.
8. Jenkins D, Hussein H, Martina R et al. Methods for the inclusion of real world evidence in network meta-analysis. *arXiv preprint arXiv:180506839*, 2021.
9. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.
10. Lin J, Gamalo-Siebers M and Tiwari R. Propensity score matched augmented controls in randomized clinical trials: A case study. *Pharm Stat* 2018; **17**: 629–647.
11. Lin J, Gamalo-Siebers M and Tiwari R. Propensity-score-based priors for Bayesian augmented control design. *Pharm Stat* 2019; **18**: 223–238.
12. Wang C and Rosner GL. A Bayesian nonparametric causal inference model for synthesizing randomized clinical trial and real-world evidence. *Stat Med* 2019; **38**: 2573–2588.
13. Liu M, Bunn V, Hupf B et al. Propensity-score-based meta-analytic predictive prior for incorporating real-world and historical data. *Stat Med* 2021. In Press.
14. Zhang Y, Lin LA, Starkopf L et al. Estimation of causal effect in integrating randomized clinical trial and observational data–an example application to cardiovascular outcome trial. *Contemp Clin Trials* 2021; **107**: 106492.
15. Rubin EJ, Harrington DP, Hogan JW et al. The urgency of care during the covid-19 pandemic—learning as we go. *N Engl J Med* 2020; **382**: 2461–2462.
16. Cohen MS. Hydroxychloroquine for the prevention of covid-19—searching for evidence. *N Engl J Med* 2020; **383**: 585–586.
17. Caro JJ and Ishak KJ. No head-to-head trial? Simulate the missing arms. *Pharmacoeconomics* 2010; **28**: 957–967.
18. Signorovitch JE, Wu EQ, Andrew PY et al. Comparative effectiveness without head-to-head trials. *Pharmacoeconomics* 2010; **28**: 935–945.
19. Weber S, Gelman A, Lee D et al. Bayesian aggregation of average data: An application in drug development. *Ann Appl Stat* 2018; **12**: 1583–1604.
20. Sutton AJ, Kendrick D and Coupland CA. Meta-analysis of individual-and aggregate-level data. *Stat Med* 2008; **27**: 651–669.
21. Riley RD, Lambert PC, Staessen JA et al. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Stat Med* 2008; **27**: 1870–1893.
22. Jin H and Yin G. Unit information prior for adaptive information borrowing from multiple historical datasets. *Stat Med* 2021; **40**: 5657–5672.
23. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res* 2011; **46**: 399–424.
24. Robins JM, Hernán MÁ and Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**: 550–560.
25. Yao XI, Wang X, Speicher PJ et al. Reporting and guidelines in propensity score analysis: A systematic review of cancer and cancer surgical studies. *JNCI: J Nat Cancer Inst* 2017; **109**: djw323.
26. Granger E, Watkins T, Sergeant JC et al. A review of the use of propensity score diagnostics in papers published in high-ranking medical journals. *BMC Med Res Methodol* 2020; **20**: 1–9.
27. Geleris J, Sun Y, Platt J et al. Observational study of hydroxychloroquine in hospitalized patients with COVID-19. *N Engl J Med* 2020; **382**: 2411–2418.
28. Morita S, Thall PF and Müller P. Determining the effective sample size of a parametric prior. *Biometrics* 2008; **64**: 595–602.
29. Ho DE, Imai K, King G et al. MatchIt: Nonparametric preprocessing for parametric causal inference. *J Stat Softw* 2011; **42**: 1–28. https://www.jstatsoft.org/v42/i08/.
30. Greifer N. *WeightIt: Weighting for Covariate Balance in Observational Studies*, 2021. https://CRAN.R-project.org/package=WeightIt. R package version 0.12.0.
31. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. https://www.R-project.org/.
32. Salvatier J, Wiecki TV and Fonnesbeck C. Probabilistic programming in Python using PyMC3. *PeerJ Comput Sci* 2016; **2**: e55.

33. Hernandez-Cardenas C, Thirion-Romero I, Rodríguez-Llamazares S et al. Hydroxychloroquine for the treatment of severe respiratory infection by covid-19: A randomized controlled trial. *PLoS ONE* 2021; **16**: e0257238.
34. Liu N, Zhou Y and Lee JJ. IPDfromKM: Reconstruct individual patient data from published Kaplan-Meier survival curves. *BMC Med Res Methodol* 2021; **21**: 1–22.
35. Guyot P, Ades A, Ouwens MJ et al. Enhanced secondary analysis of survival data: Reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol* 2012; **12**: 1–13.
36. Ip A, Berry DA, Hansen E et al. Hydroxychloroquine and tocilizumab therapy in COVID-19 patients—an observational study. *PLoS ONE* 2020; **15**: e0237693.
37. Gerlovin H, Posner DC, Ho YL et al. Pharmacoepidemiology, machine learning and COVID-19: An intent-to-treat analysis of hydroxychloroquine, with or without azithromycin, and COVID-19 outcomes amongst hospitalized US veterans. *Am J Epidemiol* 2021.
38. Paccoud O, Tubach F, Baptiste A et al. Compassionate use of hydroxychloroquine in clinical practice for patients with mild to severe COVID-19 in a French university hospital. *Clin Infect Dis* 2020; **323**: 2493–2502.