Computing True Parameter Values in Simulation Studies Using Monte Carlo Integration

Ashley I. Naimi,[a] David Benkeser,[b] and Jacqueline E. Rudolph[c]

Abstract: Simulation studies are used to evaluate and compare the properties of statistical methods in controlled experimental settings. In most cases, performing a simulation study requires knowledge of the true value of the parameter, or estimand, of interest. However, in many simulation designs, the true value of the estimand is difficult to compute analytically. Here, we illustrate the use of Monte Carlo integration to compute true estimand values in simple and more complex simulation designs. We provide general pseudocode that can be replicated in any software program of choice to demonstrate key principles in using Monte Carlo integration in two scenarios: a simple three-variable simulation where interest lies in the marginally adjusted odds ratio and a more complex causal mediation analysis where interest lies in the controlled direct effect in the presence of mediator-outcome confounders affected by the exposure. We discuss general strategies that can be used to minimize Monte Carlo error and to serve as checks on the simulation program to avoid coding errors. R programming code is provided illustrating the application of our pseudocode in these settings.

Keywords: Causal inference; Epidemiologic methods; Monte Carlo integration; Monte Carlo simulation; Numeric integration; Statistics

(Epidemiology 2025;36: 690–693)

Simulation studies are often employed to evaluate and compare the properties of statistical methods in controlled experimental settings. Simulation studies can be used for a wide range of goals, from basic conceptual clarification,[1] to formal research questions about the performance of different analytic methods.[2] Monte Carlo simulations relate to a general category of techniques known as Monte Carlo methods,[3] a class of methods that rely on pseudo-random number generation to solve problems.[4]

The design of Monte Carlo simulation studies involves several steps. These include clearly articulating the aims of the simulation study, the data-generating mechanisms that will be used, the estimand of interest, the methods that will be used to estimate the estimand, and the performance measures used to evaluate the properties of the methods under study.[2] Simulation studies proceed by repeatedly simulating datasets from a data-generating mechanism. Each simulated dataset is analyzed, and the performance of the methods is compared by aggregating results across multiple simulated datasets. For example, to approximate the bias of an estimator, we can compute the empirical average (across all simulated datasets) difference between the point estimates and the truth. To approximate the coverage of a method for building a confidence interval, we can compute the empirical proportion (out of all simulated datasets) of intervals that contain the true parameter value. This makes clear the importance for the researcher to know the true value of the estimand that is implied by the selected data-generating process.

Ideally, the true estimand value will be taken directly from the parameter values used in the data-generating mechanism. For example, if the data-generating mechanism involves simulating from a logistic regression $P(Y = 1|A, C) = \text{expit}\ \{\beta_0 + \beta_1 A + \beta_2 C\}$, with $\text{expit}\ \{\bullet\} = \frac{1}{(1+\exp(-\bullet))}$, and interest lies in estimating the conditionally adjusted odds ratio for A, the true estimand value is easily computed as $\exp\ (\beta_1)$, where β_1 is the researcher-selected value used to generate the data.

It is sometimes possible to read off true estimand values directly from the data-generating process. However, often the true estimand value may not be easily computed as a single parameter, or simple combination of parameters. Here, we outline the use of Monte Carlo integration, a process that utilizes pseudo-random sampling to approximate complex analytic expressions, to solve for true estimand values. We illustrate the approach using two example scenarios: a data-generating mechanism where the estimand of interest is the marginally adjusted odds ratio; and a causal mediation setting where interest lies in the controlled direct effect (CDE). We provide pseudocode for each example and provide R code in an associated GitHub repository.

Submitted July 3, 2024; accepted May 20, 2025
From the [a]Department of Epidemiology, Emory University, Atlanta, GA; [b]Department of Biostatistics, Emory University, Atlanta, GA; and [c]Department of Epidemiology, Johns Hopkins University, Baltimore, MD.A.I.N. was supported by NIH grant number R01 HD102313.
Disclosure: The authors report no conflicts of interest.
Correspondence: Ashley I. Naimi, Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Road. Atlanta, GA 30322. E-mail: ashley.naimi@emory.edu.

ISSN: 1044-3983/25/365-690693
DOI: 10.1097/EDE.0000000000001873

EXAMPLE 1: MARGINALLY ADJUSTED ODDS RATIO AS THE ESTIMAND

The first example involves estimating the marginally adjusted odds ratio using methods such as inverse probability weighting or marginal standardization.[5,6] The causal diagram is a simple triangle structure with C as the confounder, A as the exposure, and Y as the outcome. A parametric approach to simulating such data might be to first simulate C from a Normal distribution with mean μ and variance σ^2. Then based on the simulated values of C, simulate the exposure A from a logistic regression model with $P(A = 1|C) = \text{expit}\ (\alpha_0 + \alpha_1 \text{C})$, for some researcher-selected values α_0, α_1. Finally, the outcome is simulated from a logistic regression model

$$P(Y = 1|X, C) = \text{expit}\left\{\beta_0 + \beta_1 \text{A} + \beta_2 \text{C}\right\}. \quad (1)$$

As noted above, if the true parameter of interest was the conditionally adjusted odds ratio, its value could easily be obtained as $\exp(\beta_1)$. However, there are well-known and important differences between marginally versus conditionally adjusted odds ratios,[7–9] due to the fact that the odds ratio is a noncollapsible measure of effect.[10] Noncollapsibility is distinct from confounding,[11] which means that for a given logistic regression model with the same set of correctly chosen confounders, the marginally versus conditionally adjusted odds ratio may be numerically different. Thus, in the context of equation 1, while exponentiating the value of the β_1 coefficient provides the conditionally adjusted odds ratio, the marginally adjusted odds ratio may be a different value.

Under ideal conditions, one would rely on algebraic derivations to obtain an analytic solution to the marginally adjusted odds ratio. Doing this for equation 1 could be accomplished via integration, which would require solving:

$$\mu(a) = \int_c \text{expit}\left\{\beta_0 + \beta_1 \text{a} + \beta_2 \text{c}\right\} \frac{1}{(2\pi)^{1/2}\sigma} \exp\left\{\frac{(\text{c} - \mu)^2}{2\sigma^2}\right\} \text{dc},$$

for $a = 0, 1$. The solution to this integral yields a marginal probability of the outcome that would be observed if $A = a$. One can then construct a marginally adjusted odds ratio as:

$$\psi = \frac{\mu(1)}{1 - \mu(1)} \Big/ \frac{\mu(0)}{1 - \mu(0)}$$

However, even in this simple setting with a single variable C, analytically solving for this integral is challenging. One could employ numeric integration [see GitHub repository]. However, this approach would not scale well to settings with more than one C variable.

Instead, we can use Monte Carlo integration to solve for ψ, as follows:

BOX 1 PSEUDOCODE FOR IMPLEMENTING MONTE CARLO INTEGRATION TO SOLVE FOR THE TRUE MARGINALLY ADJUSTED ODDS RATIO IN A SIMPLE THREE-NODE CAUSAL DIAGRAM

1: set random number generator seed value
2: set large sample size N
3: simulate $i \in 1 \ldots N$ observations C_i from a Normal distribution with mean μ and variance σ^2
4: compute $\mu_i(a) = \text{expit}\ (\beta_0 + \beta_1 \text{a} + \beta_2 \text{C}_i)$ for both $a = 0$ and $a = 1$. Thus, $\mu_i(0) = \text{expit}\,(\beta_0 + \beta_2 \text{C}_i)$ and $\mu_i(1) = \text{expit}\,(\beta_0 + \beta_1 + \beta_2 \text{C}_i)$.
5: the approximated value of $\mu(x)$ is given by the mean of $\mu_i(x)$ over all N simulated observations, $\mu(x) = \frac{1}{N}\sum_{i=1}^{N} \mu_i(x)$. The approximated value of ψ is given by

$$\tilde{\psi} = \frac{[\hat{\mu}(1) / (1 - \hat{\mu}(1))]}{[\hat{\mu}(0) / (1 - \hat{\mu}(0))]}$$

The approximated value of ψ obtained from the pseudocode above can be used as the true estimand value for a simulation study evaluating the properties of an estimator seeking to quantify the marginally adjusted odds ratio. However, it is important to note that this "true" estimand value depends on (i) the specifications for the distribution of C and (ii) the parameter values for the logistic regression model generating Y in this case, $\beta_0, \beta_1, \beta_2$. If the distribution of C changes, or if the regression model coefficients change, this pseudoalgorithm should be run again under these new settings. Additionally, the true value of ψ will be subject to error that depends on the sample size used in the Monte Carlo integration. This can have important consequences for simulation results. To mitigate the impact of Monte Carlo error, a number of strategies could be considered. First, the sample size N used to compute the true value should be as large as possible under available computing resources to minimize Monte Carlo error. Second, one could more comprehensively assess the variance of the Monte Carlo error of the true value under different N to ensure that it is acceptably small in the region of N being used to compute the truth. Finally, several iterations of the pseudocode under different seed values could be run to confirm that the sample is large enough to lead to only negligible changes in the approximated value.

Of note, in step 5 of the above pseudocode, one should compute $\hat{\mu}_i(a = 0)$ and $\hat{\mu}_i(a = 1)$ directly from the logistic model for the outcome, and not to convert these probabilities to binary outcomes with a draw from the Bernoulli distribution. Doing so prevents the introduction of additional Monte

Carlo error, and reduces the variability of the Monte Carlo integration method.

EXAMPLE 2: CONTROLLED DIRECT EFFECT AS THE ESTIMAND

The second example computes the true CDE value. Here, the CDE of interest is on the difference scale:

$$\psi = E\left(Y^{a,m} - Y^{a^\star,m}\right)$$

where $Y^{a,m}$ is the potential outcome that would be observed if the exposure A were set to some value a and the mediator M were set to some value m, while $Y^{a^\star,m}$ is the corresponding value that would be observed under some referent $a^\star$.

Simulating data for this estimand can be done by referring to a causal diagram,[12] with a data-generating mechanism as displayed in the Figure. In this example, the true value of the CDE is difficult to compute analytically. It consists of a function of the magnitudes of the following paths in the Figure:

$$A \rightarrow Y$$

$$A \rightarrow M \rightarrow Y$$

$$A \rightarrow L \rightarrow Y$$

In many such settings, it simply may not be possible to analytically solve for the true CDE. However, we can use Monte Carlo integration as follows:

BOX 2 PSEUDOCODE FOR IMPLEMENTING MONTE CARLO INTEGRATION TO SOLVE FOR THE TRUE CONTROLLED DIRECT EFFECT IN A CAUSAL MEDIATION ANALYSIS DIRECTED ACYCLIC GRAPH WITH A MEDIATOR-OUTCOME CONFOUNDER AFFECTED BY THE EXPOSURE.

1: set random number generator seed value.
2: set large sample size N.
3: simulate $i \in 1 \ldots N$ observations C_i and U_i from a distribution of choice.
4: construct a_1 and a_0 for all N observations corresponding to exposed (e.g., $a_1 = 1$) and referent (e.g., $a_0 = 0$) states.
5: simulate two L_i variables, one corresponding to a_1 ($L_i^{a_1}$) and one corresponding to a_0 ($L_i^{a_0}$), for all N observations. This would be from a regression model that includes the simulated U_i and the specified a_x value, along with relevant parameters.
6: construct m for all N observations corresponding to the specific mediator value of interest in the controlled direct effect contrast (e.g., $m = 0$ for all N).
7: simulate $Q_i^{a_x,m}$ for all $i \in 1 \ldots N$ observations from the model for the outcome:
$$Q_i^{a_x,m} = f\left(\beta_0 + \beta_1 a_x + \beta_2 C_i + \beta_3 m + \beta_4 L_i + \beta_5 U_i\right),$$
where $f(\bullet)$ represents a specified link function of interest.
8: take the mean of $Q_i^{a_x,m}$ over all N to obtain $\mu_{a_x,m} = \frac{1}{N}\sum_{i=1}^{N} Q_i^{a_x,m}$, and construct a difference measure as:
$$\tilde{\psi} = \mu_{a_1,m} - \mu_{a_0,m}$$

Again, the value of $\tilde{\psi}$ obtained from the psuedocode above can be used as the true estimand value for the CDE under the specified simulation settings. The magnitude of this true value will again depend on the specific parameter values, distributions, and functional forms selected to generate all variables used in the outcome model. If any of these are changed, a separate $\tilde{\psi}$ should be quantified.

DISCUSSION

Monte Carlo simulation studies are commonly used to evaluate estimators in a range of scientific settings. However, estimator properties such as bias, mean squared error, and confidence interval coverage require a numerical value for the true estimand of interest. In many settings, this true value cannot be obtained analytically. Several simulation studies using Monte Carlo integration to compute the true parameter values have been conducted.[13,14] However, to our knowledge, there is no practical description of the procedure, or the setbacks that should be considered when deploying the approach, that can serve as a reference for researchers unfamiliar with the technique.

There are several important issues to consider when designing a program to implement Monte Carlo integration to compute estimand values. First, general recommendations suggest that the selected sample size used to compute the true value should be as large as is computationally feasible

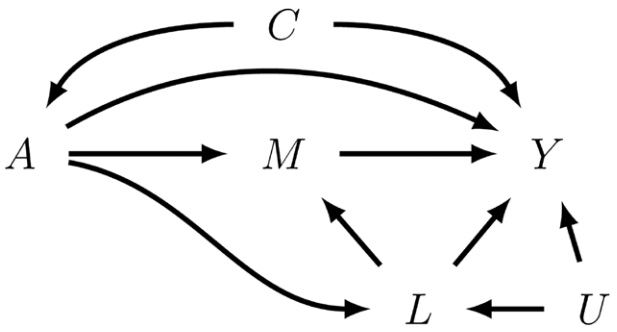

Figure. Causal diagram depicting the relationships among an exposure A, a mediator M, a confounder C, a mediator-outcome confounder affected by the exposure L, and an outcome Y.

to minimize Monte Carlo error to the smallest value possible. However, in practice a degree of Monte Carlo error could be tolerated. For example, if increasing the Monte Carlo sample size past a certain boundary κ only changes the estimand value at the 5th or greater decimal place, it might be advisable to set the Monte Carlo sample size to κ to avoid expending resources for little meaningful gain. Second, it is important to note that the source of Monte Carlo error in any simulation study is derived from programming calls to the system's random number generator. For example, in the R programming language, functions used to generate observations from a distribution of interest [e.g., rnorm(), runif(), or rexp()], or functions used to sample from a set of observations, will contribute to Monte Carlo error. For this reason, one strategy that should be used to minimize Monte Carlo error in the estimand value is to minimize these calls where possible (e.g., Step 5 in Algorithm 1, or Step 7 in Algorithm 2).

Several classes of simulation study design exist. In fully parametric simulation studies, each component of the data-generating mechanism is specified parametrically using researcher-defined functions. For example, confounders, exposures, and outcomes will all be drawn from known distributions with prespecified relationships between each. Alternatively, plasmode simulations can be conducted in which a portion of the data-generating mechanism (e.g., confounders) can be constructed using actual data sampled from a population of interest.[15] Doing so preserves the empirical associations between variables drawn from the sample, thus providing a closer approximation between simulated and real data. To compute the true value using Monte Carlo integration with a plasmode design using data from a fixed sample size, one could resample the original data with replacement to obtain a sufficiently large sample that will minimize Monte Carlo error. Finally, synthetic simulations have recently been developed that rely on machine learning to simulate data whose variables and joint distributions between them closely approximate an index dataset sampled from a real population of interest, but where the effect of the exposure of interest is specified *a priori*.[16,17] In principle, Monte Carlo integration can be used in all three settings to derive the true estimand of interest or to serve as a verification step in the process of simulating complex data. Indeed, the Monte Carlo method is a general methodological strategy that has been used across the sciences to solve simple and complex problems. With increasingly complex simulation analyses being carried out in epidemiology, Monte Carlo integration will become increasingly useful for gaining insights into complex questions.

ACKNOWLEDGMENTS

We thank Dr. Tim Morris for helpful comments on a previous version of this manuscript.

REFERENCES

1. Rudolph JE, Fox MP, Naimi AI. Simulation as a tool for teaching and learning epidemiologic methods. *Am J Epidemiol*. 2021;190:900–907.
2. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38:2074–2102.
3. Metropolis N, Ulam S. The Monte Carlo method. *J Am Stat Assoc*. 1949;44:335–341.
4. McCracken DD. The Monte Carlo method. *Sci Am*. 1955;192:90–96.
5. Naimi AI, Cole SR, Kennedy EH. Introduction to G methods. *Int J Epidemiol*. 2017;46:756–762.
6. Naimi AI, Whitcomb BW. Estimating risk ratios and risk differences using regression. *Am J Epidemiol*. 2020;189:508–510.
7. Neuhaus JM, Jewell NP. A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*. 1993;80:807–815.
8. Daniel R, Zhang J, Farewell D. Making apples from oranges: comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biom J*. 2021;63:528–557.
9. Pang M, Kaufman JS, Platt RW. Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. *Stat Methods Med Res*. 2016;25:1925–1937.
10. Greenland S. Collapsibility. In: Gail MH and Bénichou J (Eds.) *Encyclopedia of Epidemiologic Methods*. John Wiley & Sons, Ltd; 2005.
11. Pang M, Kaufman JS, Platt RW. Mixing of confounding and noncollapsibility: a notable deficiency of the odds ratio. *Am J Cardiol*. 2013;111:302–303.
12. Fox MP, Nianogo R, Rudolph JE, Howe CJ. Illustrating how to simulate data from directed acyclic graphs to understand epidemiologic concepts. *Am J Epidemiol*. 2022;191:1300–1306.
13. Franklin JM, Eddings W, Glynn RJ, Schneeweiss S. Regularized regression versus the high-dimensional propensity score for confounding adjustment in secondary database analyses. *Am J Epidemiol*. 2015;182:651–659.
14. Naimi AI, Moodie EMM, Auger N, Kaufman JS. Constructing inverse probability weights for continuous exposures: a comparison of methods. *Epidemiology*. 2014;25:292–299.
15. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal*. 2014;72:219–226.
16. Athey S, Imbens GW, Metzger J, Munro E. Using Wasserstein generative adversarial networks for the design of Monte Carlo simulations. *J Econometrics*. 2024;240:105076.
17. Parikh H, Varjao C, Xu L, Tchetgen ET. Validating causal inference methods. *International Conference on Machine Learning*. PMLR; 2022.