# Role of Covariates in Case Control Studies with Skewed Exposure: Evidence from Monte Carlo Simulations

*Yashaswini K*[(1)], *Anna Maria Pinto*[(1)]

*(1) Department of Statistics, Yenepoya (Deemed to be University), Mangalore, Karnataka, India.*

CORRESPONDING AUTHOR: Yashaswini K, Department of Statistics, Yenepoya (Deemed to be University), Mangalore, Karnataka, India. e-mail: yashaswinik33@gmail.com

## SUMMARY

Case-control studies, a widely used observational study design, are essential for investigating the association between exposure and outcomes. In such studies, logistic regression is commonly employed to analyse the relationship between binary outcome and exposure, accounting for covariates, confounders, and effect modifiers. However, skewed exposure distributions, where the exposure is disproportionately distributed among cases and controls, pose significant challenges. In this case, the parameter estimates may be biased, leading to an over- or underestimation of the true effect size, and this can affect the interpretability and reliability of the estimated coefficients.

This study aims to address these challenges by conducting a series of Monte Carlo simulation experiments to assess the impact of skewed exposure on the power of the Wald test and the bias in estimated logistic regression coefficients. The simulations focus on the role of continuous covariates in producing reliable estimates of exposure effects. The study highlights the importance of preliminary knowledge of exposure and covariate effects, as these factors play a crucial role in selecting an appropriate sample size. These simulations, which required significant computational time, highlight the robustness of the estimates with larger sample sizes and a greater number of covariates, eliminating the potential bias introduced by skewed exposure.

*Keywords: Case control study; logistic regression; skewed exposure; odds ratio; sample size.*

## INTRODUCTION

Case-Control study design, is a common observational study that involves researchers observing and measuring both exposure and outcome among participants to examine their association [1]. In this design, individuals with a particular outcome (cases) are compared to those without the outcome (controls), assessing the presence or absence of exposure in both groups to identify potential risk factors. Due to its speed and efficiency, the case-control study is frequently the preferred design for research on the causes of disease [2].

In case-control studies, the odds ratio (OR) is often used as a measure of association between exposure and a binary outcome. Logistic regression is a widely used statistical technique for analyzing case-control data, as it allows researchers to account for the effects of covariates, confounders, and potential effect modifiers [3].

The logistic regression model to study the relationship between the binary outcome ($Y$) and exposure ($X$), in the presence of $p$ covariates/confounders/effect modifiers $C_1, C_2, C_3 \ldots C_p$ is given by,

$$\log \frac{p}{1-p} = \alpha + \beta X + \beta_1 C_1 + \beta_2 C_2 + \ldots + \beta_p C_p$$

Where $p = E\left(Y \mid X, C_1, C_2, ..C_p\right)$

By including covariates/confounders/effect modifiers in the logistic regression model, their effects are accounted for, but the researcher is particularly interested in coefficient β. A positive coefficient indicates a positive association between the exposure and the outcome, while a negative coefficient indicates a negative association. The magnitude of the coefficient represents the strength of the association.

DOI: 10.54103/2282-0930/27302
Accepted: 15th January 2025

A skewed exposure in case control studies is defined as an exposure which is disproportionately distributed among cases and controls. A binary skewed exposure is defined as a categorical variable with despaired marginal distribution.

Foxman et al. [4] conducted a case-control study of contraceptives and Urinary Tract Infection among college women and the cross tabulation reported clearly indicated that Diaphragm usage as a skewed exposure with majority of the women not using it.

*Table 1. Cross tabulation of Diaphragm usage and Urinary Tract Infection, Foxman et al [4]*

| Diaphragm usage | Urinary Tract Infection | |
|---|---|---|
| | Yes | No |
| Yes | 7 | 0 |
| No | 140 | 290 |

A skewed exposure may lead to one of two potential issues:

(i) Problem of Separation: This occurs when the maximum likelihood estimates do not converge, leading to what is known as complete or quasi-complete separation. In such cases, standard logistic regression fails to provide finite estimates [5,6,7].

(ii) Biased Estimates with Wide Confidence Intervals: Even if the maximum likelihood estimates do converge, they may be biased, resulting in wide confidence intervals. This leads to underpowered inference regarding the exposure, making it difficult to draw reliable conclusions [8].

The exact logistic regression and Firth's approach are two very well-known methods to handle the problem of separation. The exact logistic regression is regression technique that provides precise parameter estimates by conditioning on sufficient statistics, making it especially suitable for small or sparse datasets. However, it is often criticized for its computational cost and inability to handle large number of covariates and large sample size [9,10,11]. Firth's approach is a bias-reduction method for logistic regression that applies a penalized likelihood function to improve parameter estimation. Also, there are other methods to handle the problem of separation and are discussed in detail by Mansournia et al [12].

But there are instances when the model does converge, the results obtained may be unreliable due to the skewness in the exposure distribution. This aspect is investigated by Alkhalaf and Zumbo (2017) [8] by considering only one covariate and they concluded that estimates are not reliable. However, the specific effects of skewed exposure on case-control studies, considering factors such as the degree of skewness, sample size, and covariate effect size, and number of covariates remain unexplored.

Thus, this paper aimed to conduct a series of Monte Carlo simulations to assess the power of the Wald test and the bias in estimated logistic regression coefficients, demonstrated the role of continuous covariates in producing the reliable estimates of exposure effect.

## Simulation Experiments

Six simulation experiments were conducted to evaluate the Wald test in logistic regression. Experiment I focused on the Type I error rate of the Wald test, varying sample sizes and skewness probabilities. Experiment II assessed the power of the Wald test, bias in exposure coefficient estimates, and confidence interval widths with a single covariate, considering different sample sizes, skewness probabilities, and effect sizes. Experiments III through VI extended this by examining the effects of increasing numbers of covariates (two to five) on the Wald test's performance, following similar procedures to those in Experiment II. These simulation experiments help understand how the inclusion of multiple covariates affects the performance of logistic regression models in case of skewed exposure. In R (Version 4.4.0), user defined functions were specialy developed to run different simulation experiments.

## Simulation Experiment I

This is a $22\times10$ experimental design where Monte Carlo simulations were conducted, considering a response variable ($Y$), a skewed binary exposure ($X$), and a covariate $C_1$.

The simulations considered two experimental factors: the sample size, denoted as $n$, the probability of skewness (or prevalence) of exposure, denoted as $p_s$, and there were 22 values for $n$ ranging from 20 to 500, 10 values for $p_s$, ranging from 0.05 to 0.5. The simulation is executed for 1000 times for each cell of $22\times10$ experimental design. The $(i,j)^{th}$ cell of this experimental design represents the simulation corresponding to $i^{th}$ sample size and $j^{th}$ probability of skewness.

The logistic regression model considered is

$$\log\frac{p}{1-p}=\alpha+\beta X+\beta_1 C_1$$

This simulation experiment is carried out to estimate the type I error rate of Wald test used in logistic regression model to detect the impact of the exposure. The null and alternative hypothesis of interest are $H_0:\beta=0\ Vs H_1:\beta\neq0$.

The type I error rate is estimated only for skewed exposure as it is of primary interest in most of the epidemiological context.

The simulation procedure for $(i,j)^{th}$ cell of the experimental design is as follows:

- The covariates are generated from a prespecified probability distributions. $X\sim Bernoulli(p_s)$ and $C_1\sim Normal(22,5)$

- The regression coefficients are set to zero and the probability *p* is generated using logistic regression model. The predicted probability serves as the expected value for the Bernoulli distribution from which the data on outcome variable is drawn.
- The logistic regression model is fitted to the simulated data, and the occurrence of type I error is noted down.
- The procedure is repeated for 1000 simulations.

The empirical type I error rate for $(i,j)^{th}$ cell of the experimental design is computed as the number of times the true null hypothesis is rejected at the level of 5% divided by 1000.

The Bradley's rule [13] was used to decide whether the type I error rate meets the criteria. The type I error rate between 0.025 and 0.075 was considered as the meeting the liberal criteria. In simulations if generated datasets fail to yield convergence with the maximum likelihood estimation method, such datasets are excluded.

## Simulation Experiment II

This is a $22 \times 10 \times 3 \times 3$ experimental design with four design factors: the sample size, denoted as *n*, the probability of skewness of exposure, denoted as $p_s$, and the regression coefficient of *X* denoted as $\beta_X$, the regression coefficient of covariate is $\beta_C$. There were 22 values for *n* ranging from 20 to 500, 10 values for $p_s$, ranging from 0.05 to 0.5, three distinct values for $\beta_X$ were considered, namely 0.683 (odds ratio=1.98), 1.1(odds ratio=3), and 1.38 (odds ratio=3.97), corresponding to small, moderate, and large effects of exposure, respectively [14]. There were 3 values for $\beta_C$ namely 0.683, 1.1 and 1.38 corresponding to small, moderate, and large effects of covariate.

In the $(i,j,k,l)^{th}$ simulation of this experiment, 1000 datasets were generated according to the previously outlined procedure (as in simulation experiment I), considering $i^{th}$ sample size, $j^{th}$ probability of skewness, $k^{th}$ value of $\beta_X$ and $l^{th}$ value of $\beta_C$.

The power of Wald test to detect the effect of exposure is defined as the number of times the false null hypothesis is rejected during $(i,j,k,l)^{th}$ simulation divided by 1000. The bias of $(i,j,k,l)^{th}$ simulation is the difference between estimated regression coefficient and the actual value. The mean bias is the mean of all these differences. The mean squared error for $(i,j,k,l)^{th}$ simulation is the average squared deviation of these differences.

$$Bias = \bar{\beta}_{ijkl} - \beta$$

$$Mean\ Bias = \frac{\sum(\bar{\beta}_{ijkl} - \beta)}{n}$$

$$Mean\ Squared\ Error = \frac{\sum\left(\bar{\beta}_{ijkl} - \beta\right)^2}{n}$$

This simulation experiment was run for 9 scenarios depending on the strength of effects of exposure and covariates. These 9 scenarios are tabulated below:

*Table 2. Combinations of exposure and covariate effects*

| | | **Covariate effect** | | |
|---|---|---|---|---|
| | | **Small** | **Moderate** | **Large** |
| **Exposure effect** | **Small** | Scenario 1 | Scenario 2 | Scenario 3 |
| | **Moderate** | Scenario 4 | Scenario 5 | Scenario 6 |
| | **Large** | Scenario 7 | Scenario 8 | Scenario 9 |

For each of these nine scenarios, the mean bias and mean squared error of the regression coefficients are computed for various values of *n* and $p_s$. The percentage bias was calculated as $\%bias = \frac{(\bar{\beta}_{ijkl} - \beta)}{\beta} \times 100$ where $\bar{\beta}_{ijkl}$ is the average estimate of β for $(i,j,k,l)^{th}$ simulation.

The estimator is unbiased if the percentage bias is 0%. The percentage bias within ±5% was considered acceptable [15]. The 95% confidence interval for β during $(i,j,k,l)^{th}$ simulation was computed as $[\bar{\beta}_{ijkl} - Z_{1-\frac{\alpha}{2}} SE(\bar{\beta}_{ijkl})\ \bar{\beta}_{ijkl} + Z_{1-\frac{\alpha}{2}} SE(\bar{\beta}_{ijkl})]$ where $SE\left(\bar{\beta}_{ijkl}\right)$ is the square root of mean squared error.

The width of the confidence interval in each simulation was calculated as

$$Width = 2 \times Z_{1-\frac{\alpha}{2}} \times SE(\bar{\beta}_{ijkl})$$

Plots of sample size versus percentage bias for different probabilities of skewness, and plots of sample size versus width of the confidence interval for various probabilities of skewness, are generated. These plots help us understand how percentage bias decreases and the width of the confidence interval narrows as sample size increases. The optimal sample size is identified as the point where the percentage bias is within ±5%, and after which the width of the confidence interval reaches a minimum and then saturates.

## Simulation Experiment III-VI

The simulation experiments III, IV, V and VI work similar to experiment II, with number of covariates increased. In all experiments, the values of $\beta_X$ are fixed at 0.683, 1.1, and 1.38, corresponding to small, moderate, and large effect sizes of the exposure, respectively. The values of $\beta_C$ correspond to the effect sizes of these covariates on the outcome, with different sets of values used to simulate small, moderate, and large effects. As the number of covariates increases across experiments III to VI, the values of $\beta_C$ correspond to combinations of effect sizes for each covariate, reflecting various strength levels of their association with the outcome.

*Table 3. The Details of Simulation Experiments*

| Experiment | III | IV | V | VI |
|---|---|---|---|---|
| Design | $22 \times 10 \times 3 \times 3$ | $22 \times 10 \times 3 \times 3$ | $22 \times 10 \times 3 \times 3$ | $22 \times 10 \times 3 \times 3$ |
| Values of $\beta_X$ | $\begin{bmatrix} 0.683 \\ 1.1 \\ 1.38 \end{bmatrix}$ | $\begin{bmatrix} 0.683 \\ 1.1 \\ 1.38 \end{bmatrix}$ | $\begin{bmatrix} 0.683 \\ 1.1 \\ 1.38 \end{bmatrix}$ | $\begin{bmatrix} 0.683 \\ 1.1 \\ 1.38 \end{bmatrix}$ |
| $\beta_C$ | $\begin{bmatrix} 0.683 & 0.683 \\ 1.1 & 1.1 \\ 1.38 & 1.38 \end{bmatrix}$ | $\begin{bmatrix} 0.683 & 0.683 & 0.683 \\ 1.1 & 1.1 & 1.1 \\ 1.38 & 1.38 & 1.38 \end{bmatrix}$ | $\begin{bmatrix} 0.683 & 1.1 & 1.38 \\ 0.683 & 1.1 & 1.38 \\ 0.683 & 1.1 & 1.38 \\ 0.683 & 1.1 & 1.38 \end{bmatrix}^T$ | $\begin{bmatrix} 0.683 & 1.1 & 1.38 \\ 0.683 & 1.1 & 1.38 \\ 0.683 & 1.1 & 1.38 \\ 0.683 & 1.1 & 1.38 \\ 0.683 & 1.1 & 1.38 \end{bmatrix}^T$ |
| Covariates | $C_1 \sim Normal(22,5)$<br>$C_2 \sim Normal(35,15)$ | $C_1 \sim Normal(22,5)$<br>$C_2 \sim Normal(35,15)$<br>$C_3 \sim Gamma(80,15)$ | $C_1 \sim Normal(22,5)$<br>$C_2 \sim Normal(35,15)$<br>$C_3 \sim Gamma(80,15)$<br>$C_4 \sim Beta(3,4)$ | $C_1 \sim Normal(22,5)$<br>$C_2 \sim Normal(35,15)$<br>$C_3 \sim Gamma(80,15)$<br>$C_4 \sim Beta(3,4)$<br>$C_5 \sim Normal(5,1)$ |

In the $(i,j,k,l)^{th}$ simulation of this experiment, 1000 datasets were generated according to the previously outlined procedure (as in simulation experiment I and II), considering $i^{th}$ sample size, $j^{th}$ probability of skewness, $k^{th}$ value of $\beta_X$ and $l^{th}$ row of $\beta_C$. For each scenario the plots of percentage bias and width of confidence interval are generated and the optimal sample size is decided.

## RESULTS

The Monte Carlo simulations of this study took approximately 1874 minutes i.e., 31.23 hours of execution time. Even though these experiments were time-consuming, a pattern in percentage bias and the width of the confidence interval emerged, providing the sample size guidelines. The results of this experiment may not be particularly pleasing to researchers who wish to conduct case-control studies with limited small sample sizes. However, preliminary knowledge about exposure and covariate effect would help researchers choose the optimal sample size.

The results of simulation experiment I is presented in Table 4. The aim of this experiment was to estimate the type I error rate of Wald test used in logistic regression in the presence of skewed exposure. It is observed that as the sample size increases, the type I error rates tend to approach Bradley's criteria more closely. However,

*Table 4. Type I Error Rates by Sample Size and Probability of Skewness*

| | | **Probability of Skewness** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0.05** | **0.1** | **0.15** | **0.2** | **0.25** | **0.3** | **0.35** | **0.4** | **0.45** | **0.5** |
| **Sample size** | 20 | 0.006 | 0.004 | 0.003 | 0.011 | 0.012 | 0.018 | 0.016 | 0.032 | 0.03 | 0.021 |
| | 50 | 0.005 | 0.006 | 0.015 | 0.032 | 0.029 | 0.052 | 0.044 | 0.051 | 0.047 | 0.046 |
| | 100 | 0.006 | 0.034 | 0.046 | 0.048 | 0.048 | 0.044 | 0.05 | 0.054 | 0.051 | 0.062 |
| | 150 | 0.006 | 0.054 | 0.046 | 0.034 | 0.058 | 0.051 | 0.052 | 0.044 | 0.045 | 0.044 |
| | 200 | 0.018 | 0.048 | 0.038 | 0.057 | 0.041 | 0.059 | 0.054 | 0.047 | 0.047 | 0.041 |
| | 250 | 0.034 | 0.045 | 0.046 | 0.042 | 0.052 | 0.054 | 0.058 | 0.038 | 0.053 | 0.05 |
| | 300 | 0.033 | 0.057 | 0.054 | 0.047 | 0.032 | 0.049 | 0.057 | 0.043 | 0.07 | 0.04 |
| | 350 | 0.037 | 0.048 | 0.054 | 0.053 | 0.054 | 0.046 | 0.05 | 0.039 | 0.047 | 0.066 |
| | 400 | 0.035 | 0.043 | 0.058 | 0.047 | 0.051 | 0.049 | 0.052 | 0.042 | 0.048 | 0.055 |
| | 450 | 0.048 | 0.041 | 0.038 | 0.05 | 0.043 | 0.063 | 0.046 | 0.043 | 0.049 | 0.055 |
| | 500 | 0.035 | 0.046 | 0.04 | 0.052 | 0.044 | 0.045 | 0.044 | 0.049 | 0.043 | 0.049 |
| | 550 | 0.049 | 0.04 | 0.052 | 0.054 | 0.047 | 0.047 | 0.046 | 0.062 | 0.039 | 0.043 |
| | 600 | 0.035 | 0.045 | 0.054 | 0.05 | 0.05 | 0.055 | 0.043 | 0.059 | 0.048 | 0.047 |

(continua)

*Table 4. Type I Error Rates by Sample Size and Probability of Skewness (continua)*

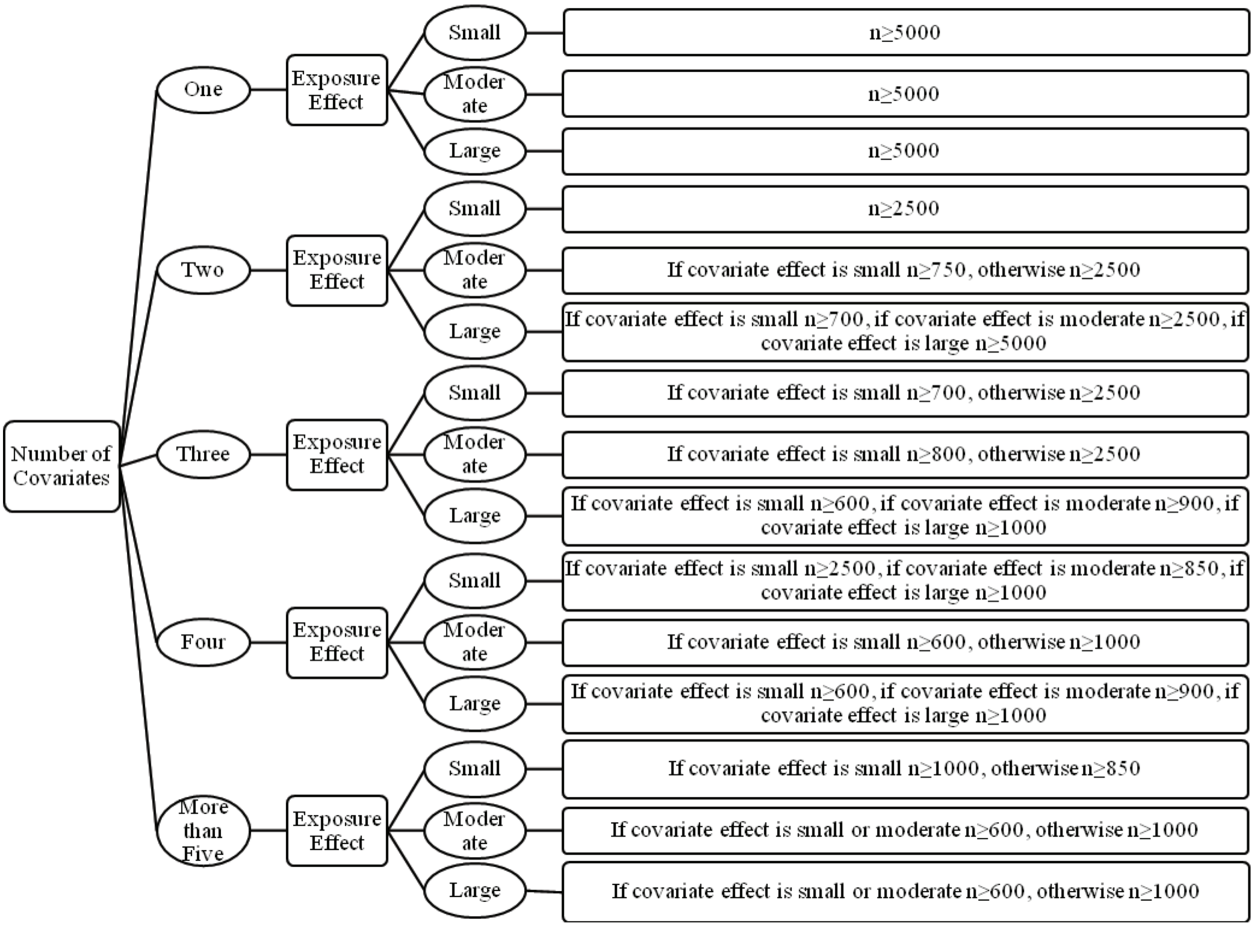
| | Probability of Skewness | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **0.05** | **0.1** | **0.15** | **0.2** | **0.25** | **0.3** | **0.35** | **0.4** | **0.45** | **0.5** |
| 650 | 0.045 | 0.049 | 0.05 | 0.037 | 0.054 | 0.054 | 0.049 | 0.042 | 0.043 | 0.039 |
| 700 | 0.035 | 0.05 | 0.046 | 0.041 | 0.05 | 0.044 | 0.057 | 0.051 | 0.07 | 0.045 |
| 750 | 0.037 | 0.041 | 0.052 | 0.039 | 0.043 | 0.072 | 0.053 | 0.052 | 0.044 | 0.045 |
| 800 | 0.044 | 0.055 | 0.045 | 0.042 | 0.04 | 0.042 | 0.062 | 0.046 | 0.047 | 0.052 |
| 850 | 0.05 | 0.046 | 0.051 | 0.043 | 0.051 | 0.053 | 0.064 | 0.051 | 0.055 | 0.059 |
| 900 | 0.045 | 0.049 | 0.039 | 0.049 | 0.039 | 0.052 | 0.041 | 0.059 | 0.056 | 0.058 |
| 1000 | 0.037 | 0.048 | 0.045 | 0.05 | 0.06 | 0.075 | 0.052 | 0.047 | 0.062 | 0.051 |
| 2500 | 0.063 | 0.055 | 0.062 | 0.056 | 0.039 | 0.044 | 0.046 | 0.05 | 0.047 | 0.058 |
| 5000 | 0.063 | 0.053 | 0.058 | 0.054 | 0.048 | 0.055 | 0.05 | 0.045 | 0.046 | 0.048 |

with smaller sample sizes, particularly when $n = 20$, there is a higher likelihood of deflated type I error rates.

The plots of sample size versus percentage bias and plots of sample size versus width of confidence interval for all nine scenarios of simulation experiments II to VI are provided in the supplementary material.

The figure presents general guidelines for determining sample size, based on percentage bias and the width of the confidence interval. If the percentage bias is within ±5% and the confidence interval is also narrower, that sample size is reported as the optimal one. However, several inputs, such as exposure effect size and covariate effect size, are required. This may seem challenging to researchers, as using these guidelines necessitates a significant amount of prior knowledge. However, this paper offers additional insights. In case-control studies, where numerous covariates are often present, it is safe to assume that there are at least five covariates. If this is the case, and the exposure effect is greater than the covariate effect, a sample size of 600 is sufficient. However, if the sample size is around 1,000, a skewed exposure does not significantly impact the estimates of the odds ratio, the power of the Wald test, or the width of the confidence interval if the number of covariates is at least five. Hence a larger sample size and an increased number of covariates contribute to more stable and reliable estimates, reducing the potential bias that might be introduced by a skewed exposure.

*Figure 1. Flowchart of Optimal Sample Size Guidelines*


Number of Covariates
One
Exposure Effect
Small
n≥5000
Moder ate
n≥5000
Large
n≥5000
Two
Exposure Effect
Small
n≥2500
Moder ate
If covariate effect is small n≥750, otherwise n≥2500
Large
If covariate effect is small n≥700, if covariate effect is moderate n≥2500, if covariate effect is large n≥5000
Three
Exposure Effect
Small
If covariate effect is small n≥700, otherwise n≥2500
Moder ate
If covariate effect is small n≥800, otherwise n≥2500
Large
If covariate effect is small n≥600, if covariate effect is moderate n≥900, if covariate effect is large n≥1000
Four
Exposure Effect
Small
If covariate effect is small n≥2500, if covariate effect is moderate n≥850, if covariate effect is large n≥1000
Moder ate
If covariate effect is small n≥600, otherwise n≥1000
Large
If covariate effect is small n≥600, if covariate effect is moderate n≥900, if covariate effect is large n≥1000
More than Five
Exposure Effect
Small
If covariate effect is small n≥1000, otherwise n≥850
Moder ate
If covariate effect is small or moderate n≥600, otherwise n≥1000
Large
If covariate effect is small or moderate n≥600, otherwise n≥1000

## DISCUSSION

There are various approaches in the literature addressing case-control studies where logistic regression fails to provide finite estimates for the odds ratio of exposure. This paper investigates the reliability of these estimates when they are finite.

The patterns in both percentage bias and the width of the confidence interval, lead to the development of sample size guidelines. These guidelines are crucial for researchers, particularly those engaged in case-control studies. The results also emphasize the importance of preliminary knowledge regarding exposure and covariate effects. Such knowledge enables researchers to select an appropriate sample size from a range of options.

The covariates considered in all the simulation experiments are continuous in nature, and hence simulated using continuous probability distributions. Further simulation can be conducted to study the impact of categorical covariates.

As the number of covariates increases, the impact of skewed exposure on the estimates of the exposure effect diminishes, particularly when the sample size is larger. This finding is significant because it suggests that, with sufficient sample size, the skewness of exposure does not substantially affect the exposure effect estimates, in studies with numerous covariates. The larger sample sizes and a greater number of covariates not only enhance the robustness of the estimates but also mitigate the potential bias introduced by skewed exposure.

## CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

## FUNDING

No funding was received for conducting this study.

## FINANCIAL INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

## REFERENCES

1. Setia MS. Methodology Series Module 2: Case-control Studies. *Indian Journal of Dermatology*. 2016; 61(2):146–51.
2. Rosendaal FR. Bridging Case-control Studies and Randomized Trials. *Current Controlled Trials in Cardiovascular Medicine*. 2001; 2(3):109–10.
3. Suárez E, Pérez CM, Rivera R, Martínez MN, Applications of Regression Models in Epidemiology. 2017.
4. Foxman B, Marsh J, Gillespie B, Rubin N, Koopman JS, Spear S. Condom Use and First-Time Urinary Tract Infection. *Epidemiology*. 1997; 8(6):637–41.
5. Day NE, Kerridge DF. A General Maximum Likelihood Discriminant. *Biometrics*. 1967; 23(2):313–23.
6. Albert A, Anderson JA. On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*. 1984; 71(1):1–10.
7. Anderson JA. Separate Sample Logistic Discrimination. *Biometrika*. 1972 Apr; 59(1):19–35.
8. Alkhalaf A, Zumbo BD. The Impact of Predictor Variable(s) with Skewed Cell Probabilities on Wald Tests in Binary Logistic Regression. *Journal of Modern Applied Statistical Methods*. 2017; 16(2):40–80.
9. Agresti A. Categorical Data Analysis. Wiley Series in Probability and Statistics. 2002 Jul 3;
10. Heinze G, Schemper M. A Solution to the Problem of Separation in Logistic Regression. *Statistics in Medicine*. 2002; 21(16):2409–19.
11. Heinze G, Puhr R. Bias-reduced and Separation-proof Conditional Logistic Regression with Small or Sparse Data Sets. *Statistics in Medicine*. 2010; 29(7-8):770–7.
12. Mansournia MA, Geroldinger A, Greenland S, Heinze G. Separation in Logistic Regression: Causes, Consequences, and Control. *American Journal of Epidemiology*. 2017; 187(4):864–70.
13. Bradley, J. V. Robustness? *British Journal of Mathematical and Statistical Psychology*.1978; 31(2):144–52.
14. Ferguson CJ. An Effect Size Primer: a Guide for Clinicians and Researchers. *Professional Psychology: Research and Practice*. 2009; 40(5):532–8.
15. Kuo CL, Duan Y, Grady J. Unconditional or Conditional Logistic Regression Model for Age-Matched Case–Control Data? *Frontiers in Public Health*. 2018; 6(57).