

# Logistic regression in cancer research: A narrative review of the concept, analysis, and interpretation

## ABSTRACT

Logistic regression is a fundamental statistical technique employed in predictive modeling. It transforms a linear combination of input variables into a probability value, allowing the available data to predict the likelihood of an event occurring. Interpretation involves understanding the coefficients of the model, odds ratios, and the impact of predictor variables on the outcome. Various performance metrics, such as the receiver operating characteristic curve, the area under the curve, and R-squared (measure of the percentage of total variation in the dependent variable that is accounted for by the independent variable), aid in assessing the model accuracy. We conducted an extensive search in the PubMed database for relevant articles published in English between January 2013 and August 2023 using the keywords, “logistic regression,” “binary logistic regression,” “logistic regression in cancer research,” “logistic regression analysis,” and “logistic regression result interpretation.” Of the 118 articles retrieved by the original search, we excluded 103 and included 15 in the review; we manually added six more articles considered classic examples of logistic regression and regression statistics. The review encompasses a wide spectrum of cancer research applications, from tumor classification and prognosis to risk assessment and response prediction. The article takes a step-by-step approach, guiding readers through the data preparation, model construction, and interpretation processes in the context of logistic regression.

**Keywords:** Binary outcome, cancer research, receiver operating characteristic curve, odds ratio, AUC, R-squared

## INTRODUCTION

In recent decades, logistic regression has emerged as a powerful and versatile statistical tool in the field of cancer research.<sup>[1]</sup> With its ability to model the probability of binary outcomes and to elucidate complex relationships between various risk factors and the occurrence of cancer, logistic regression has become an invaluable asset in advancing our understanding of this multifaceted disease.<sup>[2]</sup> As the landscape of cancer research continues to evolve, a comprehensive review of the conceptual foundations, analytical methodologies, and interpretation techniques associated with logistic regression is essential.

This narrative literature review aims to bridge the gap between emerging research findings and the application of logistic regression models in cancer research. By synthesizing and critically examining a wide range of studies and methodologies, we delve into the nuanced intricacies of logistic regression, shedding light on its relevance, versatility,

and limitations within the oncological context. We explore its applications in the assessment of cancer risk, prediction of cancer outcomes, and identification of potential prognostic and diagnostic markers. This manuscript is intended to serve as a useful resource for researchers, clinicians, and statisticians engaged in cancer research, providing

## SHARATH KUMAR, VIKRAM GOTA<sup>1,2</sup>

Lyv Clinical Operations, Product Development, Pumas-AI, Baltimore, Maryland, USA, <sup>1</sup>Department of Clinical Pharmacology, ACTREC, Tata Memorial Center, Navi Mumbai, Maharashtra, India, <sup>2</sup>Homi Bhabha National Institute, Anushakti Nagar, Mumbai, Maharashtra, India

**Address for correspondence:** Dr. Vikram Gota, Department of Clinical Pharmacology, Advanced Center for Treatment, Research and Education in Cancer, Tata Memorial Center, Navi Mumbai – 410 210, Maharashtra, India. E-mail: vgota76@gmail.com

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

For reprints contact: WKHLRPMedknow\_reprints@wolterskluwer.com


**How to cite this article:** Kumar S, Gota V. Logistic regression in cancer research: A narrative review of the concept, analysis, and interpretation. *Cancer Res Stat Treat* 2023;6:573-8.

**Submitted:** 24-Aug-2023

**Revised:** 01-Nov-2023

**Accepted:** 01-Nov-2023

**Published:** 23-Dec-2023

Access this article online	
<b>Website:</b> <a href="https://journals.lww.com/crst">https://journals.lww.com/crst</a>	<b>Quick Response Code</b> 
<b>DOI:</b> 10.4103/crst.crst_293_23	

a comprehensive overview of logistic regression and its applications. In this review, we will explore the concept, analysis, and interpretation of logistic regression with the help of real-world examples.

## METHODS

We conducted a literature search to identify relevant articles, book chapters, and other scholarly sources related to logistic regression. The search was performed in the PubMed database, using a combination of keywords and phrases, including “logistic regression,” “binary logistic regression,” “logistic regression in cancer research,” “logistic regression analysis,” and “logistic regression result interpretation.” The time frame considered for the review was from January 2013 to August 2023. This decade-long period was chosen to strike a balance between capturing recent developments and allowing for a comprehensive examination of the topic. Studies were included for the review if they were relevant, peer-reviewed, and written in English. Sources with significant methodological flaws or limitations that could compromise the reliability and validity of the findings were excluded. In the initial search, we obtained a total of 118 articles. We excluded 103 articles because of various exclusion criteria, leaving 15 articles included in this review [Figure 1]. Six articles were manually chosen (from Google) as they were classic examples of logistic regression analysis and regression statistics conducted in cancer research. Simulated datasets

and plots in this article were generated using the Pumas Integrated Scientific Modeling, Simulation, and artificial intelligence/machine learning (AI/ML) tool (<https://www.pumas.ai/>).<sup>[3]</sup>

## CONCEPT OF LOGISTIC REGRESSION

In this article, we will focus mostly on binary<sup>[4]</sup> (two discrete outcomes, e.g., Yes/No) and multinomial<sup>[5]</sup> (two or more discrete outcomes) logistic regression. The third type, ordinal logistic regression,<sup>[6]</sup> deals with variables consisting of three or more classes in a predetermined order.

### Imagine predicting “yes” or “no”

Logistic regression is like a tool we use to predict things that have only two possible outcomes, like “yes” or “no,” “true” or “false,” or “1” or “0”. It is used when we are interested in understanding how different factors might influence the chance of something happening or not happening.

### S-shaped curve - The sigmoid function

In logistic regression, we use a special mathematical curve called the “sigmoid curve.” This curve looks like an “S” and can take any number you give it and turn it into a value between 0 and 1 (representing probability).<sup>[7]</sup> Let us consider an example, where we model the probability of an event (e.g., cancer) that takes either the value 0 or 1 as a function of an independent variable (e.g., age, any other

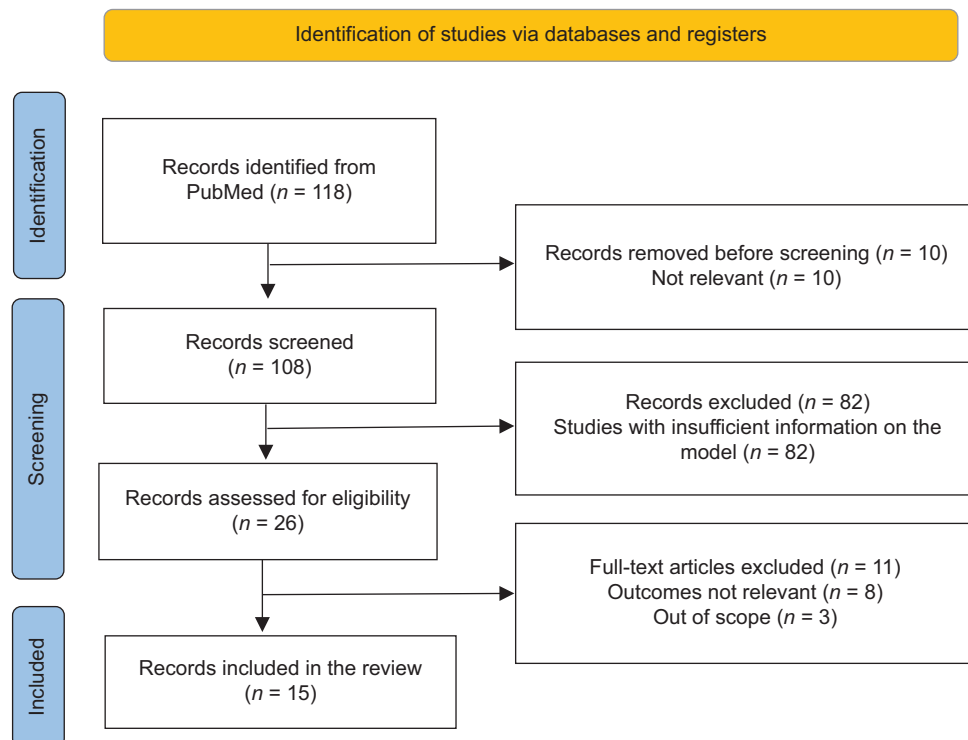


Figure 1: Flow diagram representing the methodology of identification, screening, and inclusion of papers in the review article on logistic regression

risk factor) using both logistic and linear fit represented by the S-curve and straight line, respectively. The plot of this simulated data is shown in Figure 2. In logistic fit, we can observe that the estimated probabilities are bound between 0 and 1, whereas in linear fit, estimates can be less than 0 and more than 1, which makes no sense. This shows the importance of using logistic regression to fit a binary dataset.

## USE OF LOGISTIC REGRESSION IN CANCER RESEARCH

Cancer research has greatly benefited from statistical methods like logistic regression, which play a pivotal role in understanding and predicting various aspects of cancer development, progression, and treatment outcomes. Specific uses of logistic regression in cancer research with real-world examples are described in Table 1.

### Analysis of data using logistic regression

#### Data collection and preparation

To conduct logistic regression in medical research, accurate data collection and pre-processing are crucial. Handling missing data and outliers appropriately is essential to ensure the reliability of the model.<sup>[13,14]</sup>

Arranging a dataset for logistic regression analysis involves several important steps to ensure the data are properly formatted and prepared for the analysis. Here is a step-by-step guide:

- **Data collection and extraction** - Collect the relevant medical data and ensure that the data are accurately extracted and documented.
- **Variable selection** - Identify variables (factors) that you believe might be related to the binary outcome you are studying. These could be patient demographics, medical

history, test results, and other clinical or molecular features.

- **Data cleaning** - Clean the dataset to remove any errors, missing values, or outliers. Missing data can be handled by imputation or removing rows with missing values, depending on the extent of missingness and the nature of the data.
- **Encoding categorical variables** - Convert categorical variables (e.g., sex, type of cancer) into numeric format.
- **Variable scaling** - Scale numerical variables to have similar ranges using standardization or normalization (scaling to a specified range).
- **Creating the outcome variable** - Create a binary outcome variable (dependent variable) that represents the event you are studying (e.g., the presence or absence of a disease). Assign binary values (0 or 1) to represent the two categories.

#### Example: Cancer diagnosis

Here is an example of a simulated cancer study dataset that could be used for logistic regression analysis. The objective is to predict whether a patient has a certain type of cancer, based on their age, tumor size, and genetic marker levels [Table 2].

You, therefore, need to build a logistic regression model to predict the probability of a patient having a specific type of cancer, based on their age, tumor size, and genetic marker levels.

- **Data splitting** - Divide the data into two sets—training and validation. This is important, as the training set is used to build the logistic regression model, and the testing/validation set will be used to assess its performance. Logistic regression models can become overly complex and fit noise in the training data, leading to poor generalization. Validation data help detect and prevent overfitting, ensuring that the model captures relevant patterns rather than noise. The basis for splitting data into training and validation sets can depend on various factors, including the size and nature of the dataset and the specific problem being addressed. Common approaches include Hold-Out validation, k-Fold cross-validation, Leave-One-Out cross-validation, and time-based splitting.<sup>[15]</sup> The following is an example of Hold-Out validation, in which a portion of the available dataset is set aside as the validation set, while the remaining data are used for training. Common split ratios between training and validation are 70-30, 80-20, or 90-10. If two separate datasets are available, they can be combined and partitioned as training and validation sets, and a k-Fold cross-validation can be performed.
- **Model building** - Model building refers to the process of developing a predictive model that can be used

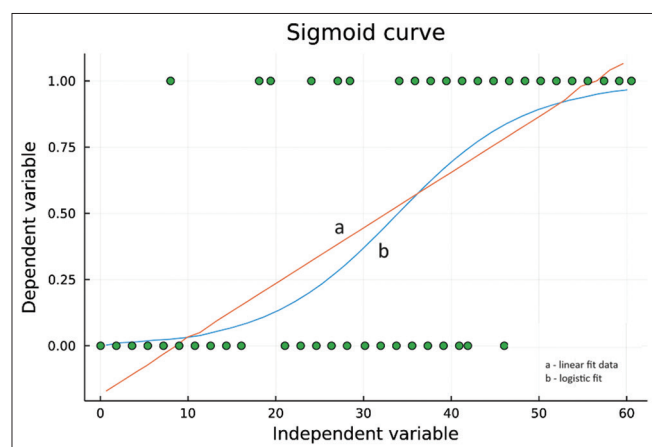


Figure 2: Plots represent the probability of event (dependent variable) which can take either the value 0 or 1 as a function of the independent variable using both logistic and linear fit represented by the S-curve and straight line, respectively

**Table 1: Uses of logistic regression in cancer research with real-world examples**

Use	Outcome	Objective	Data used for building the model	Study
Cancer diagnosis (diagnostic model)	Aids in accurate cancer diagnosis	To ascertain the multi-target stool DNA test's performance characteristics in the detection of colorectal cancer	<i>KRAS</i> mutations, aberrant <i>NDRG4</i> and <i>BMP3</i> methylation, $\beta$ -actin, and a hemoglobin immunoassay	Imperiale <i>et al.</i> , 2014 <sup>[8]</sup>
Risk assessment (risk prediction models)	Enables targeted screening and prevention strategies	To estimate how likely is it for a woman of a certain age and risk factors to develop breast cancer over a specified interval	Age at first live birth, age at menarche, number of first-degree relatives with breast cancer, and number of previous biopsies	Gail <i>et al.</i> , 1989 <sup>[9]</sup>
Prognosis (prognostic models)	To make informed decisions about treatment plans and patient care	To predict the risk of metastasis in patients with uveal melanoma	Chromosomal and clinical features	Vaquero-Garcia <i>et al.</i> , 2017 <sup>[10]</sup>
Treatment response (response prediction models)	To tailor treatments to individual patients, optimizing outcomes and minimizing adverse effects	To predict the response to immunotherapy in patients with advanced melanoma	Patient clinical and demographic characteristics, histology slides	Johannet <i>et al.</i> , 2021 <sup>[11]</sup>
Biomarker identification (biomarker-based predictive models)	Biomarkers can serve as targets for further research and potential therapeutic interventions	To determine if a panel of 10 serum biochemical markers may be used to predict response to chemotherapy, progression, and survival of patients with ovarian cancer	CA-125; kallikreins 5, 6, 7, 8, 10, and 11; B7-H4; regenerating protein IV; and Spondin-2	Oikonomopoulou <i>et al.</i> , 2008 <sup>[12]</sup>

*KRAS*= Kirsten rat sarcoma virus; *NDRG4*= N-MYC Downstream regulated gene 4; *BMP3*= Bone morphogenetic protein 3; CA-125= Cancer antigen-125

**Table 2: Simulated dataset for logistic regression analysis**

Patient number	Age (years)	Tumor size (cm)	Level of genetic marker A	Level of genetic marker B	Cancer type (1 for "yes" and 0 for "no")
1	45	2.5	0.8	1.2	1
2	32	1.8	1.1	0.9	0
3	50	3.2	0.7	1.5	1
4	28	1.4	1.3	1.1	0
5	60	4.0	0.5	1.8	1
6	28	1.0	1.0	1.0	0
* ...	...	...	...	...	...

\*The first six rows of the dataset are shown here. Additional rows are masked for display purposes

to analyze the relationship between one or more independent variables (predictors) and a binary or categorical dependent variable.<sup>[16]</sup> The first step is to apply logistic regression to the training set. You will use the predictor variables to predict the binary outcome. Most statistical software packages, such as Statistical Package for the Social Sciences (SPSS)<sup>[17]</sup> and GraphPad Prism<sup>[18]</sup> provide tools for logistic regression modeling.

- **Model evaluation** – Using the validation set, the performance of the logistic regression model is evaluated using metrics such as Receiver Operating Characteristic curve, confusion matrix, accuracy, precision, etc.

## INTERPRETATION OF LOGISTIC REGRESSION RESULTS

Statistical measures such as logistic coefficients and odds ratios are commonly used in logistic regression analysis.<sup>[19]</sup>

- **Understanding the coefficients** - A positive coefficient implies an increased odds of the binary outcome (e.g., presence of a disease) for each unit increase in the predictor variable. Conversely, a negative coefficient suggests decreased odds.
- **Odds ratios and clinical significance** - The odds ratio represents the factor by which the odds of the event

increase or decrease for a one-unit change in the predictor variable. An odds ratio  $> 1$  indicates higher odds of the outcome, while an odds ratio  $< 1$  indicates lower odds of the outcome.

**Note:** Larger coefficients (beta) and odds ratios both suggest a stronger association between the predictor variable and the outcome. The key difference between coefficients and odds ratios is that the coefficients are useful to assess the direction and strength of the relationship between variables, while the odds ratio provides a more intuitive interpretation. The odds ratio is calculated by taking the exponent of the logistic coefficient for the independent variable. Mathematically, it is represented as  $\exp(\text{Beta})$ .

- **Statistical significance** - A probability value ( $P$  value) less than a chosen significance level (e.g.,  $\alpha = 0.05$ ) indicates that the coefficient is statistically significant. This suggests that there is a relationship between the predictor variable and the outcome. However, statistical significance does not necessarily imply practical or clinical significance.<sup>[20]</sup>
- **Confidence intervals** - Examine the confidence intervals (range of values) associated with the coefficients

and odds ratios. These values indicate the range within which the true population parameter is expected to fall. Narrow confidence intervals indicate more precise estimates.<sup>[21]</sup>

Once the coefficients are estimated, it is important to evaluate how well the model captures the underlying patterns in the data. Some of these methods are shown in Table 3.

### Continuing from the example dataset shown above

Once you have arranged and prepared the dataset as shown in Table 2 (cleaned, encoded, and split into training and testing sets), you can apply logistic regression. Findings from the logistic regression analysis of the simulated dataset are shown in Table 4.

This simulated dataset and its results illustrate how logistic regression can be used to model the relationship between predictor variables and the binary outcome in the context of a cancer study.

- Validation and robustness: It is important to validate the results obtained using different datasets, if possible, to assess the robustness of the model. This ensures the reliability of your findings.
  - Confounding variables: These are the variables that correlate either positively or negatively with both the independent and dependent variables.<sup>[23]</sup> You must look to control the confounding variables during logistic regression. Typically, the “10% rule” applies to controlling the confounders.<sup>[24]</sup> If the odds ratio of the exposure-outcome relationship does not change by

10% or more after adding a confounding variable into the model, the variable does not have to be retained in the model. Let us use an example to illustrate the 10% rule. Consider a study investigating the relationship between smoking and lung cancer. We want to determine if a potential confounder, that is, age, affects the relationship between smoking and lung cancer. Let us fit two logistic regression models, one with only the smoking variable to estimate the effect of smoking on lung cancer and the other model that includes age as a potential confounder. We calculate the percentage change in the coefficient/odds ratio for smoking between the two models using the 10% rule formula. If the percentage change is more than or equal to 10%, we conclude that age is a confounder.

### Reporting of results

- Present your findings in a clear and organized manner. Include explanations of the variables used to build the model, the model's performance (final model's  $R^2$ , -2Loglikelihood or Receiver Operating Characteristic analysis results), and the implications of the results for clinical decision-making.

### CHALLENGES AND LIMITATIONS

- Logistic regression assumes that the relationship between the independent variables and the log odds of the dependent variable is linear, which may not always hold true in complex real-world scenarios. Non-linearity in the data can lead to misinterpretation of the model's predictions.

**Table 3: Methods to assess the fit of a model**

Test	Purpose	Interpretation
Likelihood ratio test	To compare the likelihood of obtaining the outcome when the predictor variable is present with the likelihood of obtaining the outcome when the predictor variable is absent	Likelihood ratio test statistic and $P$ value
Goodness of fit	To measure how best the model describes the response variable, that is, how close the model predicted values are to the observed values	Chi-squared ( $\chi^2$ ) test statistic and $P$ value
Hosmer–Lemeshow test	To assess the goodness of fit of a model	Chi-squared ( $\chi^2$ ) test statistic and $P$ value
$R^2$ (coefficient of determination)	To determine the proportion of variance in the dependent variable that can be explained by the independent variable	Cox and Snell $R^2$ and the Nagelkerke $R^2$ - Higher values denote better prediction <sup>[22]</sup>
Discrimination	How well the model distinguishes patients who achieve the outcome versus those who do not	AUC of ROC (0 to 1) – Higher values denote better discrimination

AUC= Area under the curve; ROC= Receiver Operating Characteristic

**Table 4: Logistic regression analysis results of the simulated dataset**

	Variable	Coefficient		Odds ratio	$P$
For each one-unit increase in...	Intercept*	-3.126	The odds of having the specific cancer increase by a factor of	0.043	<0.001
	Age	0.032		1.033	0.021
	Tumor size	0.732		2.077	<0.001
	Genetic marker A	1.348		3.849	<0.001
	Genetic marker B	0.898		2.456	0.008

Note: The intercept represents the log odds of having the specific cancer when all other variables are zero, \*In this analysis, a  $p$ -value <0.001 can be considered significant



- Logistic regression assumes that the observations are independent of each other, and violations of this assumption can result in biased parameter estimates.
- The model can be sensitive to outliers, which can disproportionately influence the results.
- Logistic regression is not well-suited for situations with high-dimensional data or when there are a large number of predictors relative to the number of observations, as this can lead to overfitting.
- It is essential to consider that logistic regression can only model binary or categorical outcomes and may not be suitable for problems with more complex, multiclass targets.

These challenges and limitations underscore the importance of careful model evaluation, validation, and a nuanced interpretation of the results of the logistic regression.

## CONCLUSION

Logistic regression is a powerful tool in medical research, enabling the prediction of binary outcomes and understanding the influence of predictor variables on patient health. By analyzing coefficients and odds ratios, clinicians can make informed decisions, personalize treatments, and advance medical knowledge. The application of logistic regression in medicine showcases its capacity to facilitate evidence-based practice and revolutionize patient care.

## Financial support and sponsorship

Nil.

## Conflicts of interest

There are no conflicts of interest.

## REFERENCES

1. Pal A. Logistic regression: A simple primer. *Cancer Res Stat Treat* 2021;4:551-4.
2. Sevvanthi K, Ganapathy S, Penumadu P, Harichandrakumar KT. Comparing the predictive performance of a decision tree with logistic regression for oral cavity cancer mortality: A retrospective study. *Cancer Res Stat Treat* 2023;6:103-10.
3. Available from: <https://www.pumas.ai/>. [Last accessed on 2023 Oct 31].
4. Harris JK. Primer on binary logistic regression. *Fam Med Community Health* 2021;9(Suppl 1):e001290.
5. Kwak C, Clayton-Matthews A. Multinomial logistic regression. *Nurs Res* 2002;51:404-10.
6. Bender R, Grouven U. Ordinal logistic regression in medical research. *J R Coll Physicians Lond* 1997;31:546-51.
7. Schober P, Vetter TR. Logistic regression in medical research. *Anesth Analg* 2021;132:365-6.
8. Imperiale TF, Ransohoff DF, Itzkowitz SH, Levin TR, Lavin P, Lidgard GP, *et al.* Multitarget stool DNA testing for colorectal-cancer screening. *N Engl J Med* 2014;370:1287-97.
9. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, *et al.* Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* 1989;81:1879-86.
10. Vaquero-Garcia J, Lalonde E, Ewens KG, Ebrahimzadeh J, Richard-Yutz J, Shields CL, *et al.* PRiMeUM: A model for predicting risk of metastasis in uveal melanoma. *Invest Ophthalmol Vis Sci* 2017;58:4096-105.
11. Johannet P, Coudray N, Donnelly DM, Jour G, Bochaca II, Xia Y, *et al.* Using machine learning algorithms to predict immunotherapy response in patients with advanced melanoma. *Clin Cancer Res* 2021;27:131-40.
12. Oikonomopoulou K, Li L, Zheng Y, Simon I, Wolfert RL, Valik D, *et al.* Prediction of ovarian cancer prognosis and response to chemotherapy by a serum-based multiparametric biomarker panel. *Br J Cancer* 2008;99:1103-13.
13. Darling HS. Basics of statistics-1. *Cancer Res Stat Treat* 2019;2:163-8.
14. Kwak SK, Kim JH. Statistical data preparation: Management of missing values and outliers. *Korean J Anesthesiol* 2017;70:407-11.
15. Bradshaw TJ, Huemann Z, Hu J, Rahmim A. A Guide to cross-validation for artificial intelligence in medical imaging. *Radiol Artif Intell* 2023;5:e220232.
16. Zhang Z. Model building strategy for logistic regression: Purposeful selection. *Ann Transl Med* 2016;4:111.
17. How to perform a binomial logistic regression in SPSS Statistics | Laerd Statistics. Available from: <https://statistics.laerd.com/spss-tutorials/binomial-logistic-regression-using-spss-statistics.php>. [Last accessed on 2023 Oct 30].
18. GraphPad Prism 10 Curve Fitting Guide-Example: Simple logistic regression. Available from: [https://www.graphpad.com/guides/prism/latest/curve-fitting/reg\\_simple\\_logistic\\_example.htm](https://www.graphpad.com/guides/prism/latest/curve-fitting/reg_simple_logistic_example.htm). [Last accessed on 2023 Oct 30].
19. Sperandei S. Understanding logistic regression analysis. *Biochem Med (Zagreb)*. 2014;24:12-8.
20. Darling HS. To "P" or not to "P", that is the question: A narrative review on: P: value. *Cancer Res Stat Treat* 2021;4:756-62.
21. Darling HS. Are you confident about your confidence in confidence intervals? *Cancer Res Stat Treat* 2022;5:139-44.
22. Bewick V, Cheek L, Ball J. Statistics review 14: Logistic regression. *Crit Care* 2005;9:112-8.
23. West R. Causal relationships in medicine. A practical system for critical appraisal. J. Mark Elwood, Oxford University Press, 1988. No. of pages: xi+332. Price: £30. *Stat Med* 1990;9:1543.
24. Budtz-Jørgensen E, Keiding N, Grandjean P, Weihe P. Confounder selection in environmental epidemiology: assessment of health effects of prenatal mercury exposure. *Ann Epidemiol* 2007;17:27-35.