BMC Medical Research Methodology

RESEARCH

Open Access

The proper application of logistic regression model in complex survey data: a systematic review

Devjit Dey[1], Md. Samio Haque[1], Md. Mojahedul Islam[1], Umme Iffat Aishi[1], Sajida Sultana Shammy[1], Md. Sabbir Ahmed Mayen[1], Syed Toukir Ahmed Noor[1,2] and Md. Jamal Uddin[1,3*]

Abstract

Background Logistic regression is a useful statistical technique commonly used in many fields like healthcare, marketing, or finance to generate insights from binary outcomes (e.g., sick vs. not sick). However, when applying logistic regression to complex survey data, which includes complex sampling designs, specific methodological issues are often overlooked.

Methods The systematic review extensively searched the PubMed and ScienceDirect databases from January 2015 to December 2021, following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines, focusing primarily on the Demographic and Health Surveys (DHS) and Multiple Indicator Cluster Surveys (MICS). 810 articles met the inclusion criteria and were included in the analysis. When discussing logistic regression, the review considered multiple methodological problems such as the model adequacy assessment, handling dependence of observations, utilization of complex survey design, dealing with missing values, outliers, and more.

Results Among the selected articles, the DHS database was used the most (96%), with MICS accounting for only 3%, and both DHS and MICS accounting for 1%. Of these, it was found that only 19.7% of the studies employed multilevel mixed-effects logistic regression to account for data dependencies. Model validation techniques were not reported in 94.8% of the studies with limited uses of the bootstrap, jackknife, and other resampling methods. Moreover, sample weights, PSUs, and strata variables were used together in 40.4% of the articles, and 41.7% of the studies did not use any of these variables, which could have produced biased results. Goodness-of-fit assessments were not mentioned in 75.3% of the articles, and the Hosmer–Lemeshow and likelihood ratio test were the most common among those reported. Furthermore, 95.8% of studies did not mention outliers, and only 41.0% of studies corrected for missing information, while only 2.7% applied imputation techniques.

Conclusions This systematic review highlights important gaps in the use of logistic regression with complex survey data, such as overlooking data dependencies, survey design, and proper validation techniques, along with neglecting outliers, missing data, and goodness-of-fit assessments, all of which point to the need for clearer methodological standards and more thorough reporting to improve the reliability of results. Future research should focus on consistently following these standards to ensure stronger and more dependable findings.

Keywords Logistic regression, Complex survey data, Methodological challenges, Model selection, DHS, MICS

*Correspondence:
Md. Jamal Uddin
jamal-sta@sust.edu
Full list of author information is available at the end of the article

Introduction

The logistic regression (LR) model, despite its origins in the nineteenth century, has seen increased use in the past two to three decades [1]. This model, used the categorical outcome variables most especially in binary classification where the response variable is limited to two values and represented as either 0 or 1. For instance, if we have two events, 0 might be symbolic of no occurrence or failure, and 1 might symbolize the occurrence or success of the event. LR models the probability of the response variable being 1, enabling the prediction and understanding of binary outcomes based on the provided data [2]. Higher orders of this model are ordinal or multinomial logistic regression, which is useful in handling outcomes with more than two categories [1, 3]. Furthermore, the LR model is frequently utilized in machine learning for binary classification, employing the logistic function to transform real numbers into probabilities between 0 and 1 [4]. The number of published research papers that used LR models by year between 1971 and 2021 was presented in Fig. 1.

In health-related research, such as public health, epidemiology, obstetrics, and gynecological research [5, 6], for several purposes such as predicting the outcome of an intervention (or risk), assessing if certain diseases or conditions present or absent to some degree which helps early diagnosis; exploring relationships among predictor variables with more accuracy than inferential statistics alone provide; identifying important/predictors that can be associated predicting the occurrence of morbidity/mortality rate unlike clinical specialties where it difficult access outside laboratory control tests parameters; evaluating performance (new predictors) either positively affect model accuracy [7, 8].

Complex health survey data are frequently used in these research areas, such as Demographic and Health Surveys (DHS) and Multiple Indicator Cluster Surveys (MICS). These comprehensive health surveys are important for identifying disease patterns, behaviors, and healthcare costs. These surveys enable subtle linkage studies of health factors, disease studies, or even examination of groups with conditions. The results of these surveys are influential for both medical care practices and policymaking [9, 10]. These surveys extensively collect data in developing countries, providing information on maternal and child health, family planning, child mortality, education, and sanitation [11, 12]. These are critical data sets for monitoring progress toward international development goals, including the Sustainable Development Goals (SDGs) [13].

For example, many studies that analyze complex health survey data focus on binary indicators, such as whether a person has a disease or complication based on certain measurable conditions. The LR models are an important tool, helping guide our health insights and shaping organizational decision-support regarding healthcare [14–16].

Like all statistical models, LR assumes certain conditions, including independence of observation, linear relationship between each predictor variable and logit of the outcome, no significant multicollinearity, absence of strongly influential outliers, and adequate sample size [17]. Meeting these conditions helps ensure that the LR model provides reliable and valid results. Using logistic regression on complex health survey data can be tricky because of the complex sampling methods and many factors to consider, so it's important to use clear and reliable methods [18, 19]. Besides, it is necessary to pay more attention to the selection of the variables, assessment of collinearity, and the evaluation of the model fit and

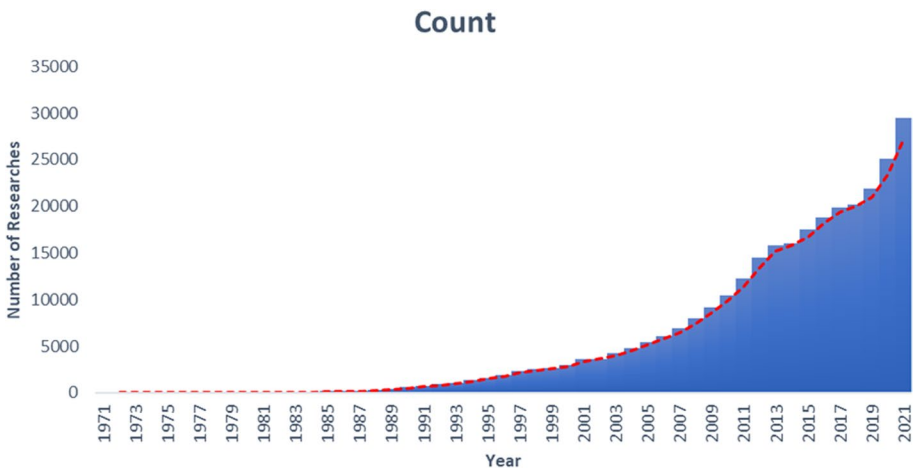

Fig. 1 Trends in research papers using logistic regression from 1971 to 2021

performance. The researcher has also to manage outliers and recognize interaction effects, account for survey weights, clustering, stratification, etc [20–22]..

It is essential for researchers to strictly validate logistic regression (LR) assumptions, follow sound scientific practices, and prioritize methodological aspects in their work. Failure to evaluate assumptions can threaten model projections, confusing variable coding can lead to ambiguous results, and mishandling missing data can impact results [23–28]. Considering these challenges, our study aims to bridge the gap in available guidance on using LR models properly in research, summarizing problems from previous studies and providing clear instructions for researchers to ensure the correct application of the LR [1] model.

The purpose of this systematic review is to provide a clear methodological insight for constructing LR models in health-related research. This review aims to provide trustworthy guidance for researchers through the analysis of the available literature so that LR models can be appropriately applied in health studies. This review aims to evaluate the state of LR models concerning health research by reviewing trends and challenges. It seeks to contribute to the field by presenting emerging methodologies and addressing gaps in existing literature. The ultimate objective is to enhance methodological standards in health-focused LR studies. Ultimately, this study should aim to improve methodological quality across health-focused LR studies. This review therefore intends to advise researchers on what is the current state of the art and possible pitfalls concerning the application of the LR model. Overall, the objective of this review is to provide useful information for researchers to perform proper and effective health-related research using the LR model.

Methodology

This systematic review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines, excluding the meta-analysis components, which are not applicable [29]. The PRISMA guidelines offer a structure for presenting systematic reviews and meta-analyses, promoting clear and thorough scientific reporting. The guidelines consist of both a checklist and a flow diagram to assist authors in conveying clear and thorough results. The study protocol was registered with the International Prospective Register of Systematic Reviews (PROSPERO) (registration number ID: CRD42020182867).

Logistic regression model

The LR shall provide each predictor with an independent coefficient for its contribution to the change in a dependent variable. The dependent variable Y will take a value of 1 if the response is "yes", and 0 if it's "no". The predicted probabilities model is formulated using the natural logarithm of the odds ratio (OR):

$$\ln\left[\frac{P(Y=1|X_{ij})}{1-P(Y=1|X_{ij})}\right] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} \quad (1)$$

where the expression $\ln[\frac{P(Y=1|X_{ij})}{1-P(Y=1|X_{ij})}]$ presents the natural logarithm of the odds of the outcome Y, where Y is a binary outcome. By exponentiating both sides, we convert the log odds into the odds, turning the logarithmic expression into an exponential one. Then, we solve probability by rearranging the equation to isolate the probability term. This process leads us to the logistic function (Eq. 2), which expresses the probability as a function of the exponentiated combination of the predictors.

$$P\left(Y=1|X_{ij}\right) = \pi(X_i) = \frac{e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots+\beta_p X_{ip}}}{1+e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots+\beta_p X_{ip}}} \quad (2)$$

Here, vector $X_i=(X_{i1}, X_{i2}, \ldots, X_{ip})$ are the independent/predictor variables in the model, and $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the coefficients of these predictors, with β_0 being the intercept term and $\pi(X_i)$ is the probability of the binary outcome (Y=1) for observation i, given its associated predictor variables X_{ij}.

In logistic regression, predictor variables influence the log odds of the outcome through the logit link function. The linear combination of predictors gives the log odds, which are then transformed back into a probability using the logistic function. Therefore, the relationship between predictors and the outcome is nonlinear in terms of probability but linear in terms of log odds. The goal of the LR model is to estimate the parameters using maximum likelihood estimation (MLE), which finds the set of parameters that maximizes the probability of the observed data. The regression coefficients reflect the strength and direction of the relationship between each independent variable and the outcome. Specifically, they indicate how much the odds of the outcome change with a one-unit change in the predictor variable. The LR model predicts the category of outcomes by calculating the odds of success versus failure, often presented as an OR. An OR measures the likelihood of an outcome given a particular exposure compared to the likelihood of the outcome without that exposure. This concept is widely used in case–control studies and applicable in cross-sectional and cohort studies.

Methodological issues

As mentioned, this systematic review comprehensively examined various methodological issues within the domain of the LR model applied to complex survey data, particularly DHS and MICS surveys (Fig. 2). Several

Fig. 2 Methodological issues of logistic regression

methodological issues that must be considered when performing a LR model are given here.

Model adequacy assessment

The goodness-of-fit statistic shows how well the model applied to the data reflects observed outcomes. Various methods exist for evaluating goodness of fit such as the Hosmer–Lemeshow statistic, which resembles R-squared in linear regression, as well as Pearson's test, deviation, and likelihood ratio among other techniques [30, 31]. The evaluation of classification model performance, particularly for binary tasks is done using AUC (Area Under the Curve). This represents the area under the ROC curve that examines true positive rates against false positive rates for different thresholds. The smallest value that can be obtained from AUC ranges from 0.5, meaning no ability to discriminate between groups, while a value of 1 indicates perfect discrimination ability; thus higher values indicate better abilities to distinguish between classes.

Furthermore, model validation is a key part of logistic regression because it establishes the credibility of models to generalize for any sample other than the one used in development. A model needs to be able to reflect genuine correlation in the research domain rather than just being a chance association. This involves testing its performance on different data from that utilized during construction. In contrast, external validation involves using independent data from either similar or different populations to evaluate the model's generalizability. Thus, if a well-fitting model aligns with new data, then its usefulness might be evident but otherwise, it could indicate either context differences or a pure lack of model fit [32].

Handling dependence of observations in binary response data due to multi-stage cluster sampling designs

In this regard, the parameters of the logistic regression model are estimated through the pseudo maximum likelihood method or weighted maximum likelihood method. This approach estimates considering both the sampling design and weights in the process. The pseudo-log-likelihood function is approximately the likelihood function

of a finite population based on observed samples and known sampling weights.

For a binary logistic regression model, the likelihood function is given by:

$$L(\beta) = \prod_{i=1}^{n} (\pi_i)^{y_i}(1 - (\pi_i))^{1-y_i} \quad (3)$$

where $\pi_i = \frac{e^{X_i\beta}}{1+e^{X_i\beta}}$ is the probability of the i-th observation having a positive outcome, given the predictors x_i, and y_i is the binary outcome.

The pseudo maximum likelihood function can be expressed as:

$$L_{pseudo}(\beta) = \prod_{i=1}^{n} \prod_{j=1}^{m_i} (\pi_{ij})^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}} \quad (4)$$

where π_{ij} is the probability for the i-th observation in the j-th cluster, and m_i represents the number of observations within cluster i.

Due to the dependent observation problem from the multi-stage cluster sampling designs, where ML estimation for binary response data is invalid due to a violation of the independence assumption, the use of advanced statistical methods often allows solving the problem associated with accounting for the correlation within groups. In general, the steps below describe a common way to address the described issue.

Generalized estimating equations (GEE)

Among clustered data, generalized estimating equations (GEE) remain one of the most applied alternatives to address the dependence of observations due to the violation of the independence assumption in a broad range of applications. These are generalizations of linear models to account for multiple related observations, and in cases of binary data, the robustness is provided through the change in the standard errors of the respective estimates. With a guarantee of robust standard errors even in the cases of errors in correlation structure definition, GEE is the most popular method for handling clustered binary data analysis [33].

$$\widehat{\beta}_{\text{GEE}} = \left(\sum_{i=1}^{n} (X_i^T V_i^{-1} X_i)\right)^{-1} \left(\sum_{i=1}^{n} X_i^T V_i^{-1} Y_i\right) \quad (5)$$

where V_i is the working covariance matrix that captures the within-cluster correlation structure, and X_i, and Y_i are the design matrix and outcome vector for the i-th cluster, respectively.

Multilevel mixed-effects logistic regression

Mixed-effects (or multilevel) logistic regression models use random effects to incorporate within-cluster correlations, thereby permitting hierarchical data structures (for instance, data grouped by regions, schools, or families). This model does not consider that all the clusters are equal, instead, it considers that changes across clusters can be treated as random effects, which enables the consideration of within-cluster variation in the presence of fixed effects. This is especially useful in situations where the intervention to be researched has an individual effect as well as a cluster effect because the relationships in clusters and between clusters can be estimated more accurately [34].

In Eq. 1 and Eq. 2, we consider a collection of p independent variables, denoted by the vector $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$. The corresponding multivariable logistic regression model (Eq. 1) can then be written as:

$$\ln\left[\frac{P(Y=1|X_{ij})}{1-P(Y=1|X_{ij})}\right] = logit(\pi_i) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} \quad (6)$$

where π_i is the probability of the outcome $Y=1$ for the i-th observation, X_{ij} represents the j-th independent variable for the i-th observation, β_0 is the intercept, and β_j are the coefficients of the predictors.

Before performing the multilevel analysis, the intra-class correlation coefficient (ICC) should be computed utilizing the formula,

$$ICC = \frac{var(u_{0j})}{var(u_{0j}) + \left(\frac{\pi^2}{3}\right)} \quad (7)$$

where $var(u_{0j})$ is the random intercept variance, i.e., the level-2 variance. The ICC value varies between 0 and 1. A multilevel logistic regression model can only be utilized when the ICC exceeds 0 [35].

So, the formula for Eq. 6 in the multilevel model can be expressed as follows:

$$logit(\pi_{ij}) = \beta_0 + \sum_{k=1}^{p} \beta_k X_{ijk} + u_j \quad (8)$$

where i refers to level 1 units (e.g., individuals), and j refers to level 2 units (e.g., clusters or groups). In this model, X_{ijk} represents the k-th independent variable for an individual i in group j. β_0 represents the overall fixed intercept, while β_k denotes the fixed effect coefficients for the covariates X_{ij}. The term u_j is the random intercept specific to the j-th group, accounting for variability between groups. The random intercept u_j is assumed to follow a normal distribution, i.e. $u_j \sim (0, \sigma^2)$, where σ^2 is the variance of the random effects. This variance determines the extent to which group-specific intercepts deviate from the overall fixed intercept β_0.

Handling survey design effects (sampling weights, primary sampling unit (PSU) or cluster, and strata variables) in estimation of binary outcomes or regression coefficients

Utilization of sampling weights, primary sampling unit (PSU) or cluster, and strata variables

In dealing with complex sample survey data analysis, it is necessary to consider the sampling design

characteristics. The lack of consideration for these factors may bring about wrong estimates of standard errors and increase false positive findings [36]. Three main aspects are involved when dealing with complex survey data: sampling weights, stratification, and cluster sampling. Sampling weights or probability weights are used to correct selection behavior among participants with changing probabilities so that the sample reflects the true population accurately. Stratification involves dividing a population into strata or subgroups that are as similar as possible based on specific attributes. Each stratum is sampled independently to reduce sampling errors, which are influenced by the variance within the strata rather than between them. Typically, samples are stratified by geographic region and urban/rural areas within each region. Within each stratum, the sample design determines the number of households to be selected. Most complex surveys use a consistent number of households per cluster, usually around 25–30 households, to decide the number of clusters needed. In the first stage, primary sampling units (PSUs), often census enumeration areas (EAs), are chosen with a probability proportional to their size within each stratum. These PSUs form the survey clusters. In the second stage, a complete list of households within each selected cluster is compiled [37]. From this list, a fixed number of households is selected using equal probability systematic sampling. This multi-stage sampling process ensures the sample accurately mirrors the population structure while minimizing sampling errors.

Considered is the use of sampling weights when fitting models to complex survey data. Ignoring the sample selection scheme in the inference process can lead to misleading results even when conditioning on all available design information if the sample is chosen with unequal selection probabilities related to the response variables. Weighting the sample observations by probability results in reliable estimations of the model parameters and guards against model misrepresentation, although to a certain extent. Different ways of integrating the sampling weights into the inference process are examined and contrasted with the application of probability weighting.

Weighted and survey-weighted logistic regression

Weighted logistic regression is an advanced type of logistic regression that uses survey weights in the estimation procedure, and it accommodates factors such as different selection probabilities. This is important in dealing with data from complex surveys because it ensures that the estimates are representative of the overall population without bias arising from sampling design [38].

As described in Eq. (2), the logistic model includes the term $P\left(Y=1|X_{ij}\right)=\pi\left(X_i\right)$, which represents the conditional probability that Y is equal to 1 given X_i. In weighted logistic regression, survey weights w_i are incorporated into the likelihood function to adjust for the unequal probabilities of selection. For calculating a Bernoulli distribution's parameter, the weights in logistic regression are determined by the covariates. The likelihood function for weighted logistic regression can be expressed as:

$$l(\mathrm{X},\ \mathrm{W},\beta)=\prod_{i=1}^{n}(\pi_i)^{w_i y_i}(1-\pi_i)^{w_i(1-y_i)} \tag{9}$$

or equivalently, the weighted log-likelihood function of covariates is:

$$L(X,W,\beta)=ln(l(X,W,\beta))=\sum_{i=1}^{n}w_i\left[y_i\log(\pi_i)+\left(1-y_i\right)\log(1-\pi_i)\right] \tag{10}$$

The value of β that maximizes the log-likelihood function is called the maximum likelihood estimator (MLE), denoted by $\widehat{\boldsymbol{\beta}}=(\widehat{\boldsymbol{\beta}}_0,\widehat{\boldsymbol{\beta}}_1,\ \ldots,\widehat{\boldsymbol{\beta}}_p)^{\mathrm{T}}$. It is easy to see that unweighted logistic regression is a special case of weighted regression with all weights equal to 1.

Survey-weighted logistic regression is often used to address survey design and cluster correlation in complex surveys. The survey weights are included in the logistic regression model along with adjustments of standard errors with methods like Taylor series linearization or resampling techniques (e.g., jackknife or bootstrap) to enable valid statistical inference [39]. The survey-weighted logistic regression links the parameters within the pseudo maximum likelihood (PML) framework, which is an extension of the standard maximum likelihood estimation (MLE) to the case of survey sampling due, for instance, to stratification and selection probabilities that are not equal. In the PML framework, survey weights, w_i, which adjust for the probability of inclusion of an observation, are included in the adjustment of the log-likelihood. This adjustment ensures that the parameter estimates are consistent and unbiased, better representing the target population compared to standard logistic regression. Variance estimation techniques, such as the Huber-White sandwich estimator, are often used to account for design effects in variance estimation, enhancing inference reliability [40].

Handling error of estimation and outliers in binary outcome data from complex surveys

When handling errors in estimation and outliers in binary outcome data from complex surveys, there are several important approaches. To ensure that estimates are representative, design effects and weights are used to account for sampling biases and inefficiencies.

Design effects and variance are adjusted using standard errors and special software (e.g., Stata, R packages). For example, logistic regression and the Huber-White Sandwich Estimator are among the regression techniques that mitigate their effects [41].

The variance of the estimator $\widehat{\beta}$ is adjusted by:

$$Var\left(\widehat{\beta}\right) = DesignEffect \times \text{Standard Error}^2 \tag{11}$$

For regression techniques, logistic regression estimates $\widehat{\beta}$ can be adjusted using the Huber-White Sandwich Estimator to account for heteroscedasticity and clustering:

$$\text{Var}\left(\widehat{\beta}\right) = \left(X^T WX\right)^{-1}\left(X^T W\Omega WX\right)\left(X^T WX\right)^{-1} \tag{12}$$

Outlier variables are characterized by significantly larger residuals (the difference between observed and predicted values) compared to other variables. Outliers can affect the coefficients significantly and if not addressed may result in biases that can compromise the LR model accuracy. Employing outlier-checking techniques is essential to ensure the accuracy and reliability of the LR analysis. There are effective methods to identify outliers, including residual assessment (using, for example, Pearson residuals and deviance residuals) and methods to calculate the effect of outliers on the regression model, including DFBETA(s) and other methods [42]. If an outlier's checking technique was given, we used that information to categorize the articles.

Outliers in complex survey data can be specifically challenging owing to the design effects, which include stratification, clustering, and unequal sampling weights. Because traditional outlier detection methods do not consider these aspects, they may not be appropriate. The following are strategies that are suited for outlier detection in complex survey data:

- *Z-square residual:* The z-square residual can be used to detect outliers in LR for observation i; the z-square residual can be calculated as:

$$Z_i^2 = \left(\frac{y_i - \widehat{\pi}_i}{\sqrt{\widehat{\pi}_i(1-\widehat{\pi}_i)}}\right)^2 \tag{13}$$

  where y_i is the outcome of the i-th case, $\widehat{\pi}_i$ is the predicted probability for the i-th case (from the logistic regression model) and $\widehat{\pi}_i\left(1-\widehat{\pi}_i\right)$ is the variance of the predicted probability.
- *Pearson residual:* In logistic regression, the Pearson residual measures the difference between observed and predicted counts, which is then standardized by variance. The calculation for observation i is as follows:

$$r_i = \frac{y_i - \widehat{\pi}_i}{\sqrt{\widehat{\pi}_i(1-\widehat{\pi}_i)}} \tag{14}$$

  where y_i is the observed value (0 or 1), $\widehat{\pi}_i$ is the predicted probability for the i-th case (from the logistic regression model), and $\widehat{\pi}_i\left(1-\widehat{\pi}_i\right)$ is the variance of the predicted probability.
- *Deviance residual:* The deviance residual indicates how much each observation adds to the overall model deviation. The deviation residual for observation i is provided as:

$$d_i = sign\left(y_i - \widehat{\pi}_i\right)\sqrt{2[y_i \log\left(\frac{y_i}{\widehat{\pi}_i}\right) + \left(1-y_i\right)\log\left(\frac{1-y_i}{1-\widehat{\pi}_i}\right)]} \tag{15}$$

  where y_i is the observed value (0 or 1), $\widehat{\pi}_i$ is the predicted probability for the i-th observation (from the logistic regression model) and the function *sign*$\left(y_i - \widehat{\pi}_i\right)$ provides the direction (positive or negative) of the residual.

Measurement error in binary outcomes is addressed using modified error models and sensitivity analyses that help evaluate how these measurement errors impact results [43, 44]. All these methods play a significant role in ensuring rigorous and dependable statistical analyses within complex survey designs.

Handling missing values in surveys with complex study design

The absence or partial availability of specified information within a data set is called missing data. Some variables may have missing values for some observers because of missing data in logistic regression. It's always important to properly manage missing data, as this will avoid incorrect conclusions that might be drawn if not managed well. Analyzing without taking into account the existing missing data would lead to a reduced sample size together with a loss of vital information, consequently affecting the validity as well as reliability of the statistical analysis. Researchers handle missing data in logistic regression using a variety of strategies, including elimination of instances with missing data, imputation (estimating missing values), and improved statistical methods that account for missingness. The proper management of missing data guarantees that the analysis is based on a complete and representative dataset, leading to more accurate and reliable results [45].

Here, we also explored several important statistical consequences that occur with the usage of complex survey design, as follows:

Appropriate treatment for missing outcome data gathered through a complex survey design is a difficult task—particularly when some variables used to estimate nonresponses might also have missing values. For tackling surveys with complex samples that collect missing-at-random outcomes, inverse probability weighting (IPW) is the most popular solution. IPW accounts for selection bias by weighting observations in inverse proportion to their chances of inclusion in the sample. If $\widehat{\pi}_i$ is the probability of selection for observation i, then the IPW weight w_i is given below:

$$w_i = \frac{1}{\pi_i} \tag{16}$$

The weighted estimate for a parameter β is then:

$$\widehat{\beta}_{IPW} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \tag{17}$$

where y_i is the outcome for observation i, and n is the total number of observations.

Nevertheless, most IPW procedures depend on the existence of fully observed predictor variables for every sampling unit which makes it uncertain how these methods should be implemented in cases where one or more independent predictors contain missing values. Multiple imputation (MI), which is another popular technique for handling missing data can address various types of missingness patterns although they are less applicable to intricate sampling frameworks [46]. For an observed dataset with missing values y_i, MI involves:

$$\widehat{\beta}_{MI} = \frac{1}{m} \sum_{j=1}^{m} \widehat{\beta}_j \tag{18}$$

In this formula, m represents the number of imputed datasets. each $\widehat{\beta}_j$ refers to the estimate of the parameter β obtained from the j-th imputed dataset. The final pooled estimate of β, denoted by $\widehat{\beta}_{MI}$, is calculated by averaging the estimates across all m datasets.

The variance-combining formula for this method is:

$$Var(\widehat{\beta}_{MI}) = \frac{1}{m} \sum_{l=1}^{m} Var(\widehat{\beta}_j) + \left(1 + \frac{1}{m}\right) Var_{between} \tag{19}$$

where $Var_{between}$ is the variance between the imputed dataset estimates.

Handling the standard errors of coefficients' estimation in complex survey data

Standard errors of regression coefficients in binary outcome data obtained through complex surveys require special techniques for their estimation. These methods account for survey designs such as clustering, stratification, and weighting that are not accommodated by traditional standard error estimation methods. The following are some of the most commonly used methods for estimating standard errors when dealing with this scenario [47].

Jackknife repeated replication (JRR)

Jackknife repeated replication is a type of resampling method used to estimate the standard errors of coefficients in complex survey data. In this method, one primary sampling unit or PSU is systematically left out of the estimate based on the sample. The procedure then repeats with each subsequent data extraction containing the PSU that was previously omitted. At the end of the extraction stage, the differences in estimates using different excerpts of the data are recorded and their mean is used as the standard error.

The standard error is estimated from the variability of these replicate coefficients, as follows:

$$Variance\left(\widehat{\beta}\right) = \frac{m-1}{m} \sum_{i=1}^{m} \left(\widehat{\beta}_{(i)} - \widehat{\beta}\right)^2 \tag{20}$$

where $\widehat{\beta}_{(i)}$ is the estimate with the i-th PSU omitted, and $\widehat{\beta}$ is the average of the estimates across all replicates.

Balanced repeated replication (BRR)

Balanced repeated replication, or BRR, is another resampling method that can be used with complex survey data. This method is used with stratified, multistage survey data. Replicates of the original weight are created by systematic perturbation, and the statistic of the interests is calculated for each set of replicate weights. In this method, a stratification of data is conducted, and replicate weights are produced, recalculating the desired sample estimate for each set of weights of the replicates, then using variability across replicates to estimate the sampling variance of the sample estimate. In balanced repeated replication (BRR), the standard error of a coefficient $\widehat{\beta}$ is approximated as:

$$SE\left(\widehat{\beta}\right) \approx \sqrt{\frac{1}{B} \sum_{b-1}^{B} \left(\widehat{\beta}_b - \overline{\widehat{\beta}}\right)^2} \tag{21}$$

where $\widehat{\beta}_b$ is the coefficient estimate from the b-th replicate, $\overline{\widehat{\beta}}$ is the average of all replicate estimates, and B is the number of replicates.

Bootstrap method

In complex survey data, a very useful resampling method for estimating the standard errors of coefficients is the bootstrap method. In this case, each should perform a repeated large number of samples called bootstrap samples from the original data set with

repetition. Mathematically, if $\widehat{\beta}_b^*$ denotes the coefficient estimates from the b-th bootstrap sample, then the standard error is given by:

$$SE\left(\widehat{\beta}\right) \approx \sqrt{\frac{1}{1-B}\sum_{b=1}^{B}\left(\widehat{\beta}_b^* - \overline{\widehat{\beta}}^*\right)^2} \quad (22)$$

Where B is the number of bootstrap samples and $\overline{\widehat{\beta}}^*$ is the mean of the bootstrap estimates.

Every bootstrap sample is utilized to re-establish the regression model, leading to the formation of a distribution of coefficient estimates. The standard errors are thus calculated based on how they vary across all bootstrapped samples. This technique works especially well in complex surveys as it respects the original survey design by maintaining an elaborate layout that includes clustering and stratification features within its resampling process. By so doing, therefore, the bootstrap method produces strong and dependable size estimates that make statistical inference valid [48].

There are also several methodological concerns with the LR model, including multicollinearity, model selection, coding of variables, interactions, statistical software, and conformity of the linear gradient. However, these issues are not specific to or negligible in the context of complex survey designs.

Search strategy

The searches were conducted in two electronic databases: PubMed and ScienceDirect. To identify additional studies that were not included in this search, we checked all references of included studies and consulted experts for missing studies that met the inclusion criteria. The searches were performed on 12 December 2022 by one author who was experienced in searching electronic databases mentioned above. The period was January 1, 2015, to December 31, 2021. The search terms were: ("Complex Survey Analysis" OR "Complex Survey" OR "Complex Survey Design" OR "Complex Survey Data" OR "Survey Analysis" OR "Survey" OR "Survey Design" OR "Survey Data" OR "Demographic and Health Survey" OR "DHS" OR "Multiple Cluster Indicator Survey" OR "MICS") AND ("Logistic Regression" OR "Generalized Linear Model" OR "Multi-Level Logistic Regression" OR "Logit Regression" OR "Binary model"). Table 1 displays all search terms:

Inclusion and exclusion criteria

We chose articles that reported on the LR model and utilized complex survey designs such as DHS and MICS surveys that were published in English in any country in the world. We also chose full-text journal articles and articles on the human population from January 2015 to December 2021. We excluded reviews, editorials, case reports, letters to the editor, guidelines, dissertations and theses, book chapters, full text, not available journals, animals, objectives and materials, meta-analyses, and languages other than English. Table 2 shows the detailed inclusion and exclusion criteria that were used to select articles.

Table 1 Search terms

#1	Complex Survey Analysis Complex Survey Complex Survey Design Complex Survey Data Survey Analysis Survey Survey Design Survey Data
#2	Demographic and Health Surveys DHS
#3	Multiple cluster indicator surveys MICS
#4	Logistic Regression Generalized Linear Model Multi-Level Logistic Regression Logit Regression Binary Model

Combining them with Boolean operators

(#1 AND #4) AND (#2 OR #3)

Data selection and extraction

The titles and abstracts were screened to identify articles that substantially met our inclusion criteria. To manage duplicate publications, we employed Endnote [49] and Rayyan [50]. The preliminary selection of abstracts and titles was carried out using Rayyan [50]. After identifying potentially eligible studies, we reviewed the full-text articles to extract the necessary data. All articles were reviewed independently by the co-authors. Any disagreements among co-authors were resolved collaboratively, following the supervisor's directives.

The country of study was determined based on the first author's affiliation, and the reviewers were not provided with information regarding the journal or the authors' names during the extraction of data from the articles. Furthermore, the publication year for each article was determined to assess the time trends in the use of LR. Overall, this study included 810 research publications, and the details extracted from each article are provided in Appendix A. The data related to the methodological issues discussed in "Methodological issues" section were extracted and analyzed for this review.

Table 2 Inclusion and exclusion criteria

	Inclusion Criteria	Exclusion Criteria
Study Design	• Cross-Sectional Studies	• Observational Cohort Studies • Observational Case–Control Studies • Experimental Studies (Randomized Control Trials, Clinical Trials, Interventional Studies)
Sampling Design	• Complex Survey Design	• All other studies don't utilize a complex survey design
Method	• Logistic Regression	• All other studies that don't perform a logistic regression analysis to compute estimates
Participant	• Human Population	• Animals, Objectives, and Materials
Article	• Full-Text Available Journal Articles	• Editorials, letters, and commentaries • Case studies, reports, or case series • Literature reviews • Guidelines • Dissertations and Thesis • Book Chapters • Full text not available journal • Systematic Review/Meta-Analysis
Survey Database	• Demographic Health Surveys (DHS) • Multiple Indicator Cluster Surveys (MICS)	• All other survey databases
Publication	• Jan 2015 to Dec 2021	
Language	• English	• Languages except English
Location	• All Countries	

Software

For data analysis, we used SPSS (v26) and Excel 2019.

Results

The search terms yielded 1,783 articles in PubMed and 1,842 articles in ScienceDirect, resulting in a total of 3,625 articles. After removing duplicates, our search yielded 2940 articles. Following the selection of titles and abstracts, 1322 articles were included for full-text selection. In total, 810 studies met our inclusion criteria and were eligible for inclusion in the review (Fig. 3).

The publication timeline of the review provides evidence of an increasing number of qualitative studies in recent years. Papers from 2020 and 2021 collectively account for 56.7% of the total, demonstrating that over half of the study's sources are concentrated in these two years alone. In addition, publications from 2019 contributed 12.7%, while earlier years from 2015 to 2018, each contributed between 7.0% and 9.1% Fig. 4(a). Most authors (41%) were from the Department of Public Health, followed subsequently by the Department of Statistics (7.5%), Epidemiology (4.6%), and other affiliations (32%). The rest of them are combined affiliations mentioned in Fig. 4(b). In terms of the database types, the DHS database was the most widely used (96%), with MICS being used by a lower proportion of users (3%) and both DHS & MICS at 1% Fig. 4(c). Furthermore, analytical software choices were also analyzed. The findings highlight a diverse picture of software adoption. Notably, the most preferred package out of all was STATA (57.8%), followed by SPSS (23.0%), SAS (6.8%), and R (4.4%). A smaller proportion of papers (1.2%) reported on the use of additional software packages, while a great number (6.8%) did not specify their software Fig. 4(d).

The complex surveys of DHS & MICS covered a comprehensive set of countries' data with significant representation from Ethiopia (20.7%), Bangladesh (12.1%), Ghana (7.0%), Nepal (6.4%), and Nigeria (5.4%). The world map (Fig. 5) showing the use of logistic regression in DHS and MICS data reveals that research is concentrated in regions like South Asia, Sub-Saharan Africa, and parts of Southeast Asia. This suggests common research interests and health priorities in these areas. In contrast, limited use in other parts of the world indicates less research or data available in those regions.

We obtained the following results based on the key methodological issues encountered when applying the logistic regression model to complex survey data, as discussed in "Methodological issues" section. The results are presented in Table 3.

Model adequacy assessment

The review found that 24.7% addressed the goodness of fit (GOF) of the model. In those cases, 32.8% utilized the Hosmer–Lemeshow test, 35.9% used the likelihood ratio test, 20.7% relied on the deviance measure, 6.6% applied the Pearson test and 1.0% employed c-statistic. In about 3.0% of the studies, no special method was mentioned. This indicates that in research on logistic regression

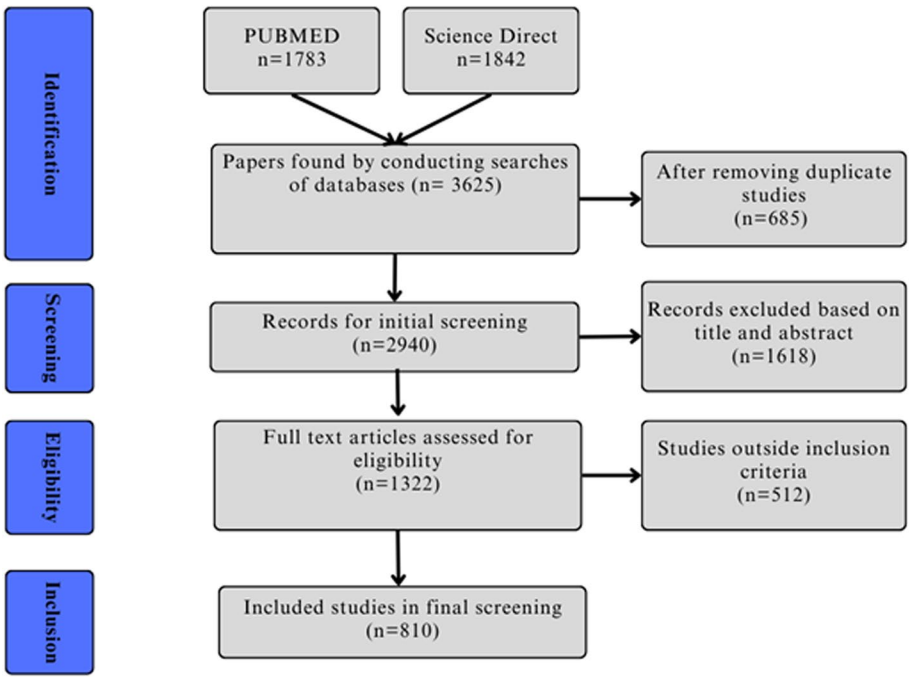

Fig. 3 Flow chart for the studies included in the systematic review

in this field, there must be a continued assessment of GOF and truthful reporting to uphold methodological accuracy.

Model validation discussions were noticeably lacking in most of the evaluated publications (94.8%) performing logistic regression analyses, indicating the absence of attention to this critical modeling characteristic. 5.2% of the articles discussed the model validation techniques where various strategies were used by people who debated it: cross-validation (23.8%) demonstrated an understanding of generalizability testing, while split methods (21.4%) demonstrated the importance of validation on diverse datasets.

Model test methods for handling dependence of observations in binary data

With regards to techniques for testing models, maximum partial likelihood estimation was widely used in 46.3% of the cases with a very low percentage (1.7%) of implementation of pseudo maximum likelihood. Furthermore, out of the 810 studies assessed, 19.7% utilized multilevel logistic regression analysis to address data dependencies, whereas the overwhelming majority (80.2%) failed to account for the hierarchical structure of the data.

Handling survey design effects (sampling weights, primary sampling unit (PSU) or cluster, and strata variables)

Data collection based on sampling weights is an issue in (58.3%) making the matter of taking into account complex survey designs to obtain precise and representative results very important. However, this implies that 41.7% of the papers did not take this into account which indicates that more research is required on its effect on validity. Concerning dealing with clustered data, 73% of the studies were carried out using primary sampling unit (PSU) characteristics to realize reliable analysis. On the other side, 26.4% of the studies did not report using PSUs. Utilization of strata variable was used in 64.9% of the research, highlighting the need to account for stratified sampling designs. Surprisingly, 40.4% of the study used all three variables. Sampling weight, PSU, or cluster and strata variables improves the methodological difficulty and reliability of the results. The command "svy" (typically used after survey data has been imported into Stata and survey design information has been specified) was used to account for complex survey designs in 29.6% of the research, showing the use of advanced statistical methods to ensure accurate interpretation of findings and appropriately address the inherent complexities of survey data.

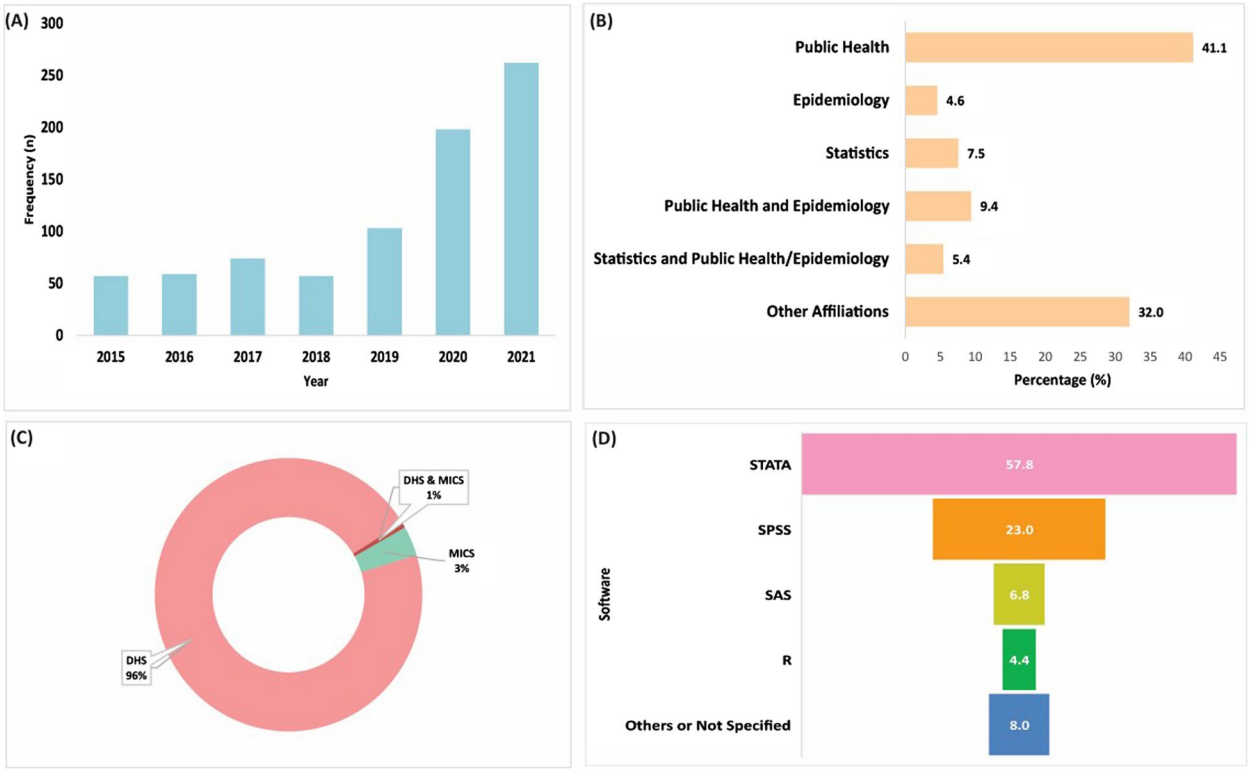

Fig. 4 Screening of articles by (**a**) publication year, (**b**) author affiliation, (**c**) survey database type, and (**d**) software for data analysis ($n = 810$)

Handling error of estimation and outliers

According to the review, surprisingly, considering the significance of outliers in data quality assessment, a large proportion of research studies (95.8%) did not recognize their presence. Various strategies were used in the subset of research that addressed outliers (4.2%). In particular, the approach most used was z square residual (40%), followed by Pearson residual (8.6%), deviance residual (11.4%), scatter diagram (2.9%) and 37.1% did not mention the specific method. These findings emphasize the importance of outlier detection and its consequences for logistic regression research in this sector.

Handling standard errors

To handle the standard errors in complex survey data resampling approaches such as bootstrapping (7.2%) and other resampling methods (14.3%) examined model stability, while a subset investigated reliability (2.4%) for validation. However, transparency might be at risk because a subgroup (30.9%) didn't clearly explain the validation process, making it harder to reproduce and understand the findings.

Handling missing data

Data analysis carried out ushered in missing data revealing that 41.0% of the studies dealt with missing data, thus showing that data strategy is crucial for doing complete studies. Out of all these investigations, 88.3% opted to leave out missing values while only 2.7% adopted different imputations such as mean/mode imputation, multiple imputations, and k-nearest neighbors (KNN) imputation techniques respectively. In addition to this, 9% of them did not indicate their way (IPW or multiple imputation) of handling missing values whereby they were either excluded or included sometimes.

Discussion

In this systematic review, it has been revealed that the significance of clearly defining how the parameters of the included variables are estimated properly and unbiasedly in the context of complex survey design is crucial for enhancing accuracy and reproducibility [51]. Some studies provide thorough explanations while others fail to do so on account of poor transparency, thus suggesting that there is always room for improvement in reporting standards. The systematic analysis places great emphasis on the critical nature of trustworthy reporting practices in studies that use logistic regression. The clarity of variable selection and carefully chosen methods for significance testing are vital to research integrity and the improvement of the quality of empirical research across different fields [52]. By advocating for these transparent practices,

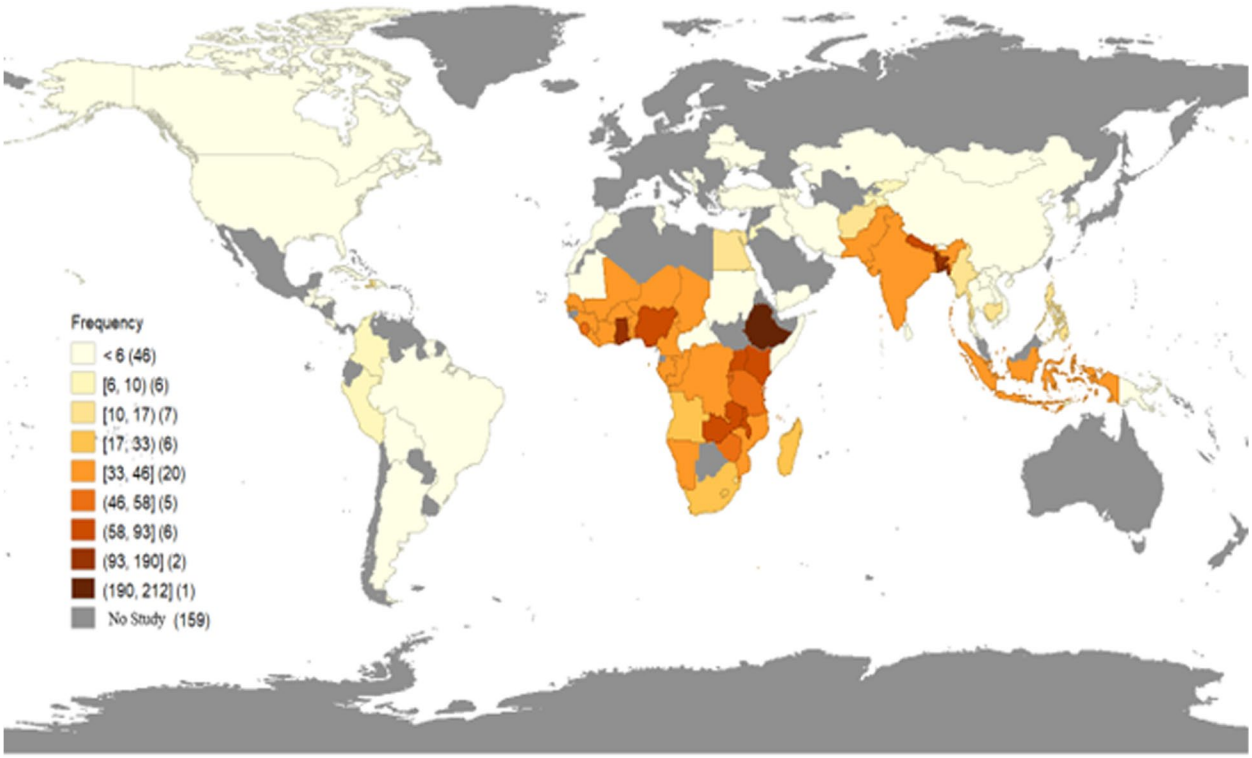

Fig. 5 Region-wise frequencies of uses on logistic regression in complex survey databases

researchers and publishers can collectively raise the standards of reporting and methodological accuracy in logistic regression analyses.

Surprisingly, the systematic review brings to attention that maximum partial likelihood estimation is widely preferred as a method for testing models, which is in accordance with its efficiency in estimating categorical predictors under certain circumstances. However, the minimal application of pseudo maximum likelihood (1.7%) suggests a gap in awareness of its potential benefits [53]. Similarly, the use of the wald chi-square test is underscored on account of its importance as far as the evaluation of individual predictor variables' contribution to a model is concerned. However, the presence of a significant number of studies that did not specify the procedure to fit or test the model underscores the need for more consistent and thorough reporting standards. This, in turn, strengthens transparency and the potential to reproduce results in logistic regression analyses [54].

The use of multilevel models is particularly important in the sense that they make it possible to analyze clustered data: they deal with intra-cluster correlation and give more trustworthy estimates of parameters [55]. Previous research demonstrated that not accounting for these dependencies can produce biased point estimates with low standard errors [56]. In consideration of the previous remark, it's suggested that fields featuring underlying structures with clustered data should embrace the use of multilevel modeling.

Model validation is critical in determining if the logistic regression model can generalize beyond the sample data. Cross-validation and other techniques, which divide data into training and testing sets, are generally suggested. Zhang discusses purposeful selection as a strategy for model building and validation, emphasizing the importance of validating models on different datasets to ensure they perform well in different contexts [54]. This systematic review's findings are quite revealing because there is an overwhelming number of studies that didn't even mention anything about model validation despite it being an important aspect of evaluation and reporting transparency. However, among the few studies that were carried out to validate, we note a thorough knowledge of model generalizability and reliability in the use of cross-validation and various sampling methods (such as holdout validation, k-fold cross-validation, leave-one-out cross-validation (LOOCV), stratified k-fold cross-validation, bootstrap validation). Similarly, split-sample methods fall into this category since they allow the division of data into parts for both training purposes as well as testing making it easier for the indivisibility of internal validation procedures as pointed out in reference [32].

Table 3 Descriptive statistics of all collected information

ASPECTS	CATEGORIES	FREQUENCY	PERCENTAGE
Explanation of selection variable	No	132	16.3
	Yes	678	83.7
Model test methods of the fitting procedure	Conditional Parameter Estimation	30	24.8
	Maximum Partial Likelihood Estimation	56	46.2
	Pseudo Maximum Likelihood Estimation	2	1.7
	Wald Chi-Square Test	33	27.3
Multilevel	No	650	80.25
	Yes	160	19.75
Model Validation Discussed	No	768	94.8
	Yes	42	5.2
Model Validation Method	Cross Validation	10	23.8
	Bootstrapping	3	7.2
	Other Resampling Methods	6	14.3
	Robustness	1	2.4
	Split Sample Methods	9	21.4
	Not Specified	13	30.9
Method of Goodness of Fit Test	Hosmer-Lemeshow Test	66	32.8
	Likelihood Ratio Test	72	35.9
	Deviance	42	20.7
	Pearson Test	14	6.4
	C-Statistic	2	1.0
	Others	7	3.0
Checking for Outliers	No	775	95.8
	Yes	35	4.2
Outliers Discussion Method	Z Square	14	40
	Pearson Residual	3	8.6
	Deviance Residual	4	11.4
	Scatter Diagram	1	2.9
	Not Specified	13	37.1
The use of survey weight in analytical models	No	338	41.7
	Yes	472	58.3
The use of primary sampling unit (PSU) in analytical models	No	214	26.4
	Yes	569	73.6
The use of Strata in analytical models	No	284	35.1
	Yes	526	64.9
The use of all three variables in analytical models(Sampling weight, PSU or cluster and strata variables)	No	493	59.6
	Yes	317	40.4
The use of Survey Setup for complex survey design in analytical models	No	570	70.4
	Yes	240	29.6
Missing Data Discussed	No	478	59
	Yes	332	41
Explanation of Missing Data was discussed	Excluded	293	88.3
	Adjusted	9	2.7
	Not Specified	30	9

The systematic review underscores a significant want for emphasis on evaluating how well logistic regression models fit the data. Several research works have neglected this issue, hence more efforts must be put in place to raise awareness on model evaluation and performance. At the same time, some studies used various

methods such as the Hosmer–Lemeshow test, likelihood ratio test, deviance, Pearson test, and c-statistic when addressing goodness of fit. These methods offer various approaches to assess how well the model fits [57, 58].

The systematic review points out a notable deficiency in handling outliers in logistic regression analyses. Outliers, which are observations that significantly differ from the rest of the data, can greatly affect the outcome of a logistic regression model by skewing the estimation of coefficients and resulting in incorrect conclusions. The evaluation revealed that numerous studies did not acknowledge or deal with outliers properly, causing doubts about the credibility of the results presented in these papers. In the review, estimations of standard errors in complex survey data are analyzed with a special emphasis on their robustness. It was used in just 7.2% of the research. Resampling techniques such as bootstrapping, jackknife repeated replication (JRR), and balanced repeated replication (BRR) are necessary for dealing with complex survey designs. Still, very few researchers employ them and some are ambiguous on how they arrived at their figures. It is important that researchers use better methods when estimating standard errors to increase the trustworthiness and credibility of findings in public health research. This builds an argument for people engaging in research activity to adopt advanced methods that are used for determining standard errors to come up with dependable results whose credibility can be validated in research [59].

On the other hand, properly conducted studies understand the significance of identifying and handling outliers to maintain the strength of their models. For example, Hosmer et al. (2013) highlight the potential impact of outliers on logistic regression outcomes and suggest regular checks like analyzing standardized residuals, leverage, and Cook's distance to detect influential data points [31]. These techniques offer a thorough evaluation of the impact of individual data points on the overall model, which is essential for ensuring the accuracy of the analysis.

Moreover, research by Nurunnabi & West explains how various methods can be used to identify outliers depending on the specific characteristics of a data set, for example, z-square, Pearson residuals, and deviance residuals have all been identified as newer innovative approaches [60]. Their findings indicate that while z-square is suitable for identifying extreme values in general, Pearson and deviance residuals are more suitable for identifying some influential data points that are not easily detected using other methods.

Similarly, the systematic review highlights that a large majority of research studies commonly considered sample weight adjustments. This underscores the importance of accounting for complex survey designs to obtain accurate and representative results. However, some publications within this subset did not take sample weights into account, emphasizing the need for further investigation into how this choice might affect the trustworthiness and dependability of the study findings. The review also draws attention to the various ways in which PSU and strata variables were handled in logistic regression analyses. The consideration of the PSU variable in a significant portion of research indicates an understanding of how to effectively manage clustered data for reliable analysis [61]. Considering the vital importance of stratified sampling designs is emphatic in having many strata variables, while the application of cluster sampling shows that observational relationships can be recognized. A subgroup of all three variables- sample weights, PSUs, and strata- was used by some studies; therefore their results were methodologically more precise and trustworthy [62]. Moreover, this review draws attention to the "svy" command usage for complex survey designs on logistic regression analyses. The use of such an advanced statistical method in almost one-third of the studies displays a commitment towards accurate findings interpretation thereby effectively addressing inherent challenges associated with survey data complexities. In this sense, these findings stress how imperative it is to have sound approaches that yield dependable results in logistic regression research.

Above all, missing data in logistic regression modeling is an important issue discussed in this systematic review. Since over 40% of studies acknowledge missing values, it goes without saying that detailed investigations are needed. Although most studies chose to leave out missing data, a smaller portion uses adjusted analyses. Researchers need to acknowledge and address missing data by using proper techniques and reporting their methods. This could involve utilizing advanced imputation techniques, conducting sensitivity analyses to evaluate the influence of missing data, and providing detailed documentation of the selected method. Schafer and Graham endorse utilizing sophisticated imputation methods, like multiple imputation, to handle missing data because it maintains dataset integrity and minimizes bias [63]. In the context of complex survey data, the lack of mention of the inverse probability weighting (IPW) technique, is an effective approach when working with missing information because it can take into consideration what the chances are that any particular observation is included to adjust for such cases [64]. Weighting observations by their probabilities of missingness can minimize biases while giving greater precision in estimates. Doing so highlights the importance of transparent reporting and careful management of missing data to improve the

trustworthiness and reliability of logistic regression research in this context [45, 65].

Recommendation

After exploring the existing literature on logistic regression analysis with complex survey data, we identified many key methodological issues that are missing from the included studies and which researchers should incorporate into their work. To start with, staying up to date on current trends and standard procedures in logistic regression analysis is vital to assuring research rigor and relevance. Researchers should be given attention to model validation and cross-validation or split-sample methods should be used to establish whether results can be generalized; validation outcomes should be reported to mitigate credibility debt. Checks for outliers should be systematic and employ data points such as z-square, Pearson, and standardized residuals that might have a disproportionately high effect on the outcomes. Simple resampling techniques like bootstrap, jackknife, and other methods can be employed to overcome the standard error problem in the case of complex survey data. It is strongly advised to include survey weights, PSU, clusters, and strata in the data set to improve the performance of the logistic model. In addition, it is also suggested to practice sophisticated replacement approaches like multiple imputation and inverse probability weighting that are keen on tackling the issue of missing data and through this, a solution for minimizing bias is also uncovered. To address clustered data effectively, researchers should consider using multilevel mixed-effects logistic regression models. These models are particularly recommended as they account for intracluster correlation, which can otherwise bias estimates and reduce the accuracy of standard errors. Incorporating multilevel modeling ensures more reliable parameter estimation by accommodating the hierarchical structure often inherent in complex survey data. Applying these methodological aspects to the study can significantly improve the model's performance, hence reliable interpretation, and prediction.

Strengths

The strength of this systematic review is founded on a methodological approach that thoroughly discusses different elements related to logistic regression analyses, giving an exhaustive and clear presentation of the overall research scenario. The necessity for enhancing reporting norms, taking into account model validation to a larger extent, as well as considering the interaction effects and outliers better are some findings that were very powerful and discussions that were very enlightening. This review highlights areas needing improvement concerning transparency and rigor, making it a useful tool for researchers, thus contributing to raising standards for methodology and reproducibility of logistic regression research in various fields.

Limitations

The systematic review can be said to give a good idea of how logistic regression studies are done, their results, and the implications that come with them, but it is not without limitations. There may be publication bias because studies with significant results are more likely to be published, statistical practices keep changing over time since there was no knowledge update after December 31, 2021, and thirdly there is a wide range of fields and research areas covered by this review making the methodologies and reporting criteria also differ. Furthermore, this review assumes that the data reported in the examined studies was accurate and complete which might change; therefore their overall findings could have been less comprehensive. Hence, even though this review was able to present important findings regarding logistic regression research improvement other than these limitations should be emphasized when making sense out of its conclusions as well as implications.

Conclusion

This study aims to systematically investigate the methodological aspects of the existing literature that has used a binary logistic model in complex survey data. The results highlight the need for more effective measures to ensure proper validation, clearly defined outcomes, handling dependence of observations, handling of outliers and missing values to produce valid and reliable results. The study also suggests incorporating complex survey design variables such as survey weights, primary sampling units, and stratification, which help to reduce bias and achieve the actual population structure.

The article notes that most studies present some good practices, but most of them do not meet significant validity criteria such as model verification and outlier treatment, which puts into question the entire findings. Overcoming such limitations is critical for progress in this area because advanced and sometimes even simple approaches become systematic over time leading to increased validity and expanded applicability of findings based on logistic regression modeling. Researchers can significantly improve the performance and reliability of logistic regression models by systematically corresponding to these key assumptions and including the recommended methodological aspects.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02454-5.

Supplementary Material 1.

Acknowledgements
We extend our sincere gratitude to the Demographic and Health Surveys (DHS) and Multiple Indicator Cluster Surveys (MICS) teams for their invaluable contributions to the collection and dissemination of high-quality data. Their dedication to providing open-access datasets has been instrumental in advancing research and informing evidence-based policies worldwide.

Authors' contributions
M.J.U proposed and designed the study, made substantial contributions to the interpretation, and oversaw the study project. D.D, M.S.H, M.M.I, U.M.A and S.S.S reviewed the literature and conducted a systematic review after extracting the data. M.S.H and D.D contributed significantly to the data analysis and drafted the first copy of the manuscript. M.M.I, U.M.A and S.S.S assisted in the writing of the manuscript. M.S.A.M and S.T.A.N provided instructive assistance throughout the process, helped write the first draft of the manuscript, reviewed and edited the manuscript. All the contributors have read and accepted the final version of the manuscript.

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability
The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Statistics, Shahjalal University of Science and Technology, Sylhet 3114, Bangladesh. [2]Maternal and Child Health Division, International Centre for Diarrhoeal Disease Research, Bangladesh (Icddr,B), Dhaka, Bangladesh. [3]Faculty of Graduate Studies, Daffodil International University, Dhaka, Bangladesh.

Received: 6 May 2024 Accepted: 27 December 2024
Published online: 22 January 2025

References

1. Boateng EY, Abaye DA. A review of the logistic regression model with emphasis on medical research. J Data Anal Inform Processing. 2019;7(4):190–207. https://doi.org/10.4236/jdaip.2019.74012.
2. Bewick V, Cheek L, Ball J. No title found. Crit Care. 2005;9(1):112. https://doi.org/10.1186/cc3045.
3. López L, Sánchez JL. Discriminant methods for radar detection of hail. Atmos Res. 2009;93(1–3):358–68. https://doi.org/10.1016/j.atmosres.2008.09.028.
4. Zaidi A, Al Luhayb ASM. Two statistical approaches to justify the use of the logistic function in binary logistic regression. Math Probl Eng. 2023;2023(1):5525675. https://doi.org/10.1155/2023/5525675.
5. Chien K, Cai T, Hsu H, et al. A prediction model for type 2 diabetes risk among Chinese people. Diabetologia. 2009;52(3):443–50. https://doi.org/10.1007/s00125-008-1232-4.
6. Khan KS, Chien PF, Dwarakanath LS. Logistic regression models in obstetrics and gynecology literature. Obstet Gynecol. 1999;93(6):1014–20. https://doi.org/10.1016/s0029-7844(98)00537-7.
7. Kim Y, Kwon S, Heun SS. Multiclass sparse logistic regression for classification of multiple cancer types using gene expression data. Comput Stat Data Anal. 2006;51(3):1643–55. https://doi.org/10.1016/j.csda.2006.06.007.
8. Howell P, Davis S. Predicting persistence of and recovery from stuttering by the teenage years based on information gathered at age 8 years. J Dev Behav Pediatr. 2011;32(3):196–205. https://doi.org/10.1097/DBP.0b013e31820fd4a9.
9. Kindratt TB. Complex survey design features. Published online August 1, 2022. https://uta.pressbooks.pub/bigdataforepidemiology/chapter/chapter5-complexsurvey/. Accessed 15 Mar 2024.
10. Use of Design Effects and Sample Weights in Complex Health Survey Data: A Review of Published Articles Using Data From 3 Commonly Used Adolescent Health Surveys - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3477989/. Accessed 15 Mar 2024.
11. Boerma JT, Sommerfelt AE. Demographic and health surveys (DHS): contributions and limitations. World Health Stat Q. 1993;46(4):222–6.
12. The DHS Program - Demographic and Health Survey (DHS). https://dhsprogram.com/Methodology/Survey-Types/DHS.cfm. Accessed 15 Mar 2024.
13. THE 17 GOALS | Sustainable Development. https://sdgs.un.org/goals. Accessed 15 Mar 2024.
14. Logistic Regression Modelling for Complex Survey Data with an Application for Bed Net Use in Mozambique. https://www.scirp.org/journal/paperinformation?paperid=71436. Accessed 15 Mar 2024.
15. Determinants of Stunting among under-five Years Children in Ethiopia from the 2016 Ethiopia Demographic and Health Survey: Application of Ordinal Logistic Regression Model using Complex Sampling Designs. https://www.researchgate.net/publication/336096037_Determinants_of_Stunting_among_under-five_Years_Children_in_Ethiopia_from_the_2016_Ethiopia_Demographic_and_Health_Survey_Application_of_Ordinal_Logistic_Regression_Model_using_Complex_Sampling_Design. Accessed 15 2024.
16. Multivariate Logistic Regression Analysis of Complex Survey Data with Application to BRFSS Data. https://www.researchgate.net/publication/267941983_Multivariate_Logistic_Regression_Analysis_of_Complex_Survey_Data_with_Application_to_BRFSS_Data. Accessed 15 Mar 2024.
17. Zach. The 6 Assumptions of Logistic Regression (With Examples). Statology. October 13, 2020. https://www.statology.org/assumptions-of-logistic-regression/. Accessed 8 Sept 2023.
18. Amoakoh-Coleman M, Ansah EK, Agyepong IA, Grobbee DE, Kayode GA, Klipstein-Grobusch K. Predictors of skilled attendance at delivery among antenatal clinic attendants in Ghana: a cross-sectional study of population data. BMJ Open. 2015;5(5):e007810. https://doi.org/10.1136/bmjopen-2015-007810.
19. Atteraya M, Kimm H, Song IH. Caste- and ethnicity-based inequalities in HIV/AIDS-related knowledge gap: a case of Nepal. Health Soc Work. 2015;40(2):100–7. https://doi.org/10.1093/hsw/hlv010.
20. Lakew Y, Haile D. Tobacco use and associated factors among adults in Ethiopia: further analysis of the 2011 Ethiopian Demographic and Health Survey. BMC Public Health. 2015;15(1):487. https://doi.org/10.1186/s12889-015-1820-4.
21. Rahman MM, Gilmour S, Akter S, Abe SK, Saito E, Shibuya K. Prevalence and control of hypertension in Bangladesh: a multilevel analysis of a nationwide population-based survey. J Hypertens. 2015;33(3):465–72. https://doi.org/10.1097/HJH.0000000000000421.
22. Application of logistic regression to explain internet use among older adults: a review of the empirical literature. https://ouci.dntb.gov.ua/en/works/4O2bwxb7/. Accessed 15 Mar 2024.
23. Austin PC, van Buuren S. The effect of high prevalence of missing data on estimation of the coefficients of a logistic regression model when using multiple imputation. BMC Med Res Methodol. 2022;22(1):196. https://doi.org/10.1186/s12874-022-01671-0.
24. Senaviratna NAMR, A. Cooray TMJ. Diagnosing Multicollinearity of Logistic Regression Model. AJPAS. Published online October 1, 2019:1–9. https://doi.org/10.9734/ajpas/2019/v5i230132.

25. Regression Analysis - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/psychology/regression-analysis. Accessed 15 Mar 2024.
26. Pawakapan P. Logistic Regression is sensitive to outliers? Using on synthetic 2D dataset. Stack Overflow. April 1, 2018. https://stackoverflow.com/q/49603548. Accessed 15 Mar 2024.
27. jan. Answer to "Logistic Regression is sensitive to outliers? Using on synthetic 2D dataset." Stack Overflow. April 2, 2018. https://stackoverflow.com/a/49603934. Accessed 15 Mar 2024.
28. appletree. Answer to "Logistic Regression is sensitive to outliers? Using on synthetic 2D dataset." Stack Overflow. September 6, 2018. https://stackoverflow.com/a/52212988. Accessed 15 Mar 2024.
29. PRISMA statement. PRISMA statement. https://www.prisma-statement.org. Accessed 2 Aug 2024.
30. Hosmer Jr. DW, Lemeshow S, Sturdivant RX. Assessing the Fit of the Model. In: Applied Logistic Regression. 2013th ed. ; 2013:153–225. https://doi.org/10.1002/9781118548387.ch5.
31. Hosmer Jr. DW, Lemeshow S, Sturdivant RX. Model-Building Strategies and Methods for Logistic Regression. In: Applied Logistic Regression. 2013th ed. ; 2013:89–151. https://doi.org/10.1002/9781118548387.ch4.
32. Anderson WN. Statistical techniques for validating logistic regression models. Ann Thorac Surg. 2005;80(4):1169. https://doi.org/10.1016/j.athoracsur.2005.06.049.
33. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73(1):13–22. https://doi.org/10.1093/biomet/73.1.13.
34. Vermunt JK. Mixed-effects logistic regression models for indirectly observed discrete outcome variables. Multivar Behav Res. 2005;40(3):281–301. https://doi.org/10.1207/s15327906mbr4003_1.
35. Roberts JK. An Introductory Primer on Multilevel and Hierarchical Linear Modeling. Learning Disabilities: Published online; 2004.
36. Chantala K. Using STATA to Analyze Data from a Sample Survey. Published online January 1, 2001.
37. Guide to DHS Statistics (English). https://dhsprogram.com/publications/publication-dhsg1-dhs-questionnaires-and-manuals.cfm. Accessed 9 Aug 2024.
38. Skinner C, Mason B. Weighting in the regression analysis of survey data with a cross-national application. Canadian J Stat / La Revue Canadienne de Statistique. 2012;40(4):697–711.
39. Sakaya K. Incorporating survey weights into binary and multinomial logistic regression models. Sci J Appl Math Stat. 2015;3:243. https://doi.org/10.11648/j.sjams.20150306.13.
40. Heeringa SG, West BT, Heeringa SG, Berglund PA, Berglund PA. Applied Survey Data Analysis. 2nd ed. Chapman and Hall/CRC; 2017. https://doi.org/10.1201/9781315153278.
41. Freedman DA. On The So-Called "Huber Sandwich Estimator" and "Robust Standard Errors." Am Stat. 2006;60(4):299–302. https://doi.org/10.1198/000313006X152207.
42. Zhang Y ying, Zhou X bin, Wang Q zhen, Zhu X yan. Quality of reporting of multivariable logistic regression models in Chinese clinical medical journals. Medicine. 2017;96(21):e6972. https://doi.org/10.1097/MD.0000000000006972.
43. Surupa R, Tathagata B. Analysis of Mixed Outcomes: Misclassified Binary Responses and Measurement Error in Covariates. J Stat Comp Simul. 2007;80. https://doi.org/10.1080/00949650903008593.
44. Proschan MA, McMahon RP, Shih JH, et al. Sensitivity analysis using an imputation method for missing binary data in clinical trials. Journal of Statistical Planning and Inference. 2001;96(1):155–65. https://doi.org/10.1016/S0378-3758(00)00332-3.
45. School of Transportation Engineering, Suranaree University of Technology, 111 University Ave., Muang, Nakhon Ratchasima, 30000, Thailand, Meeyai S. Logistic Regression with Missing Data: A Comparison of Handling Methods, and Effects of Percent Missing Values. *JTLE*. Published online 2016. https://doi.org/10.18178/jtle.4.2.128-134.
46. Kalpourtzi N, Carpenter JR, Touloumi G. Handling missing values in surveys with complex study design: a simulation study. J Surv Stat Methodol. 2024;12(1):105–29. https://doi.org/10.1093/jssam/smac039.
47. Rao JNK, Wu CFJ. Resampling inference with complex survey data. J Am Stat Assoc. 1988;83(401):231–41. https://doi.org/10.2307/2288945.
48. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Springer US; 1993. https://doi.org/10.1007/978-1-4899-4541-9.
49. EndNote | The Best Citation & Reference Management Tool. EndNote. https://endnote.com/. Accessed 28 Sept 2023.
50. Rayyan - AI Powered Tool for Systematic Literature Reviews. November 8, 2021. https://www.rayyan.ai/. Accessed 28 Sept 2023.
51. Variable Selection for Logistic Regression Model Using Adjusted Coefficients of Determination. Korean J Appl Stat. 2005;18(2):435–443. https://doi.org/10.5351/KJAS.2005.18.2.435.
52. Heinze G, Wallisch C, Dunkler D. Variable selection – A review and recommendations for the practicing statistician. Biom J. 2018;60(3):431–49. https://doi.org/10.1002/bimj.201700067.
53. Solomon G, Weissfeld L. Pseudo maximum likelihood approach for the analysis of multivariate left-censored longitudinal data. Stat Med. 2017;36(1):81–91. https://doi.org/10.1002/sim.7080.
54. Zhang Z. Model building strategy for logistic regression: purposeful selection. Ann Transl Med. 2016;4(6):111–111. https://doi.org/10.21037/atm.2016.02.15.
55. (PDF) Multilevel Analysis: An introduction to basic and advanced multilevel modeling. https://www.researchgate.net/publication/44827177_Multilevel_Analysis_An_Introduction_to_Basic_and_Advanced_Multilevel_Modeling. Accessed 3 Nov 2024.
56. Hierarchical Linear Models. SAGE Publications Inc. October 2, 2024. https://us.sagepub.com/en-us/nam/hierarchical-linear-models/book9230. Accessed 3 Nov 2024.
57. Hosmer DW, Taber S, Lemeshow S. The importance of assessing the fit of logistic regression models: a case study. Am J Public Health. 1991;81(12):1630–5. https://doi.org/10.2105/AJPH.81.12.1630.
58. Hosmer DW, Lemeshow S. Applied Logistic Regression. 1st ed. Wiley; 2000. https://doi.org/10.1002/0471722146.
59. Ahmad T. A resampling technique in complex survey data. J Ind Soc Agri Stat. 1997;50:364–79.
60. Nurunnabi A, West G. Outlier Detection in Logistic Regression: A Quest for Reliable Knowledge from Predictive Modeling and Classification. In: 2012 IEEE 12th International Conference on Data Mining Workshops. IEEE; 2012:643–652. https://doi.org/10.1109/ICDMW.2012.107.
61. Galbraith S, Daniel JA, Vissel B. A study of clustered data and approaches to its analysis. J Neurosci. 2010;30(32):10601–8. https://doi.org/10.1523/JNEUROSCI.0362-10.2010.
62. Lumley T. Analysis of complex survey samples. J Stat Softw. 2004;9:1–19. https://doi.org/10.18637/jss.v009.i08.
63. Schafer J, Graham J. Missing data: our view of the state of the art. Psychol Methods. 2002;7:147–77. https://doi.org/10.1037/1082-989X.7.2.147.
64. Chesnaye NC, Stel VS, Tripepi G, et al. An introduction to inverse probability of treatment weighting in observational research. Clin Kidney J. 2021;15(1):14–20. https://doi.org/10.1093/ckj/sfab158.
65. Mahdy S, Abonazel M, Ghallab M. A review of ten imputation methods for handling missing values in logistic regression: a medical application. JPAS. 2021;21(3):434. https://doi.org/10.5455/sf.117832.

Publisher's Note