Universal Journal of Applied Mathematics 11(1): 8-15, 2023
DOI: 10.13189/ujam.2023.110102 http://www.hrpub.org

Confounding and Effect Modification in Biostatistics: An Overview

Mário Basto[1,*], **Teresa Abreu**[1], **Ricardo Gonçalves**[1], **José M. Pereira**[2]

[1]Higher School of Technology, Polytechnic Institute of C´avado and Ave, Barcelos, Portugal

[2]CICF - Research Center on Accounting and Taxation, Polytechnic Institute of C´avado and Ave, Barcelos, Portugal

Received December 24, 2022; Revised April 15, 2023; Accepted April 25, 2023

Cite This Paper in the following Citation Styles

***(a)**: [1] Mário Basto, Teresa Abreu, Ricardo Gonçalves, José M. Pereira, "Confounding and Effect Modification in Biostatistics: An Overview," Universal Journal of Applied Mathematics, Vol.11, No.1, pp. 8-15, 2023. DOI: 10.13189/ujam.2023.110102*

***(b)**: Mário Basto, Teresa Abreu, Ricardo Gonçalves, José M. Pereira, (2023). Confounding and Effect Modification in Biostatistics: An Overview. Universal Journal of Applied Mathematics, 11(1), 8-15. DOI: 10.13189/ujam.2023.110102*

Copyright ©2023 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract Understanding concepts like confounding and effect modification is essential to biostatistics due to their potential influence on the interpretation of statistical results. The ability to appropriately identify and comprehend the links between research variables relies on having a firm grasp of these ideas. Statistical methods, such as stratification-related procedures and logistic regression, can be used to account for potential confounding factors. This helps to determine the real link between two variables of interest by controlling for the potentially skewing effects of the confounding variables. The ability to identify subsets of the population that may be more or less receptive to an intervention necessitates an understanding of effect modification. If, for instance, one learns that regular exercise is particularly useful in warding off heart disease in younger people, one could direct preventative efforts toward them. This paper aims to highlight the existence of associations between two binary variables that may be misleading or distorted due to the existence of confounding or effect modifier variables, that must be accounted for. The Mantel-Haenszel analysis and logistic regression are two techniques addressed in this study that help to statistically adjust the association between variables. Typically, only one of these methods is used in a study. This paper contrasts and illustrates the application of both techniques. To do this, four hypothetical situations are examined in order to provide the researcher with a comparative analysis of the two procedures and how to interpret the outcomes from each. Although the data may need to be adapted for each of the analyses, the outcomes are usually the same. The results indicate that both approaches are useful as long as they are used correctly. Nevertheless, depending on the situation, one or the other may be advisable.

Keywords Odds Ratio, Confounding, Interaction, Effect Modification, Mantel-Haenszel Analysis, Logistic Regression

1 Introduction

Many problems in biostatistics revolve around determining the link between a potentially damaging factor and a concrete outcome, like an illness. In this situation, one wants to look into the connection between two variables. However, it is not sufficient to explore the potential relationship between these two factors without additionally taking into account other variables that may interfere with its association [1], [2]. In the majority of cases, in order to conduct a proper statistical analysis, one is required to make adjustments for other variables that could potentially influence the results of the analysis. Some of these additional variables are known as confounders. A confounder is a variable that is connected to both the disease (result or outcome) and the underlying factor (exposure) [2], [3], [4], [5], [6], [7], [8]. To put it another way, confounding is a circumstance in which the effect or association between an exposure and a result is distorted by the presence of a confounding variable or variables unequally distributed throughout the different exposure levels. It can be either positive or negative, depending on whether the observed link or correlation is biased positively (away from the null) or negatively (toward the null). Each set of individuals defined by the confounders experiences the effect to varying degrees, but the impact of the exposure remains unchanged.

For instance, a case-control study could be carried out to look into the connection between living close to high voltage lines and Alzheimer's disease. In epidemiology, case-control studies are among the most common types of studies con-

ducted. In contrast to cohort studies, case-control research can often be conducted in a significantly shorter amount of time and at a lower financial outlay. In this case, in order to perform a proper statistical analysis, it is necessary to account for a number of confounders that have already been recognized as risk factors, including age (Alzheimer's disease typically affects people over the age of 65), family history, sex (women are more prone than men to get the condition, owing in part to their greater life expectancy), mild cognitive impairment, and lifestyle (high blood pressure, high cholesterol, and poorly controlled diabetes are all known risk factors that may potentially increase the likelihood of developing Alzheimer's disease). It is essential to draw accurate conclusions from the data, and this can only be done by minimizing the impact of potential confounding factors [8], [9], [10], [11], [12].

Randomization and stratification are both potential strategies that can be considered during the study's design phase [11], [12]. The advantage of incorporating stratification during the design stage is that it allows one to prevent errors that would otherwise be caused by having some strata in which there is a significant imbalance between the number of cases and controls [2], [5], [6]. When cases are paired with controls that have comparable features, that is just an example of stratified design. There are a number of benefits to these matching configurations, such as allowing for confounding variables to be accounted for and adjusted for in the study that aren't easy to get a handle on in the analysis phase since they are hard to quantify directly. Also, given that matching does not need any particular assumptions as to functional form, which may be necessary in regression models, it is able to give more accurate control of confounding than can be adjusted in analysis using regression. But challenges might emerge. Unique individuals have a difficult time finding compatible matches. Another drawback of matching cases and controls on a particular feature is that it makes it impossible to determine the impact of that factor on a disease. Furthermore, a collection of matched cases and controls is seldom typical of any particular population, which limits the capacity to generalize study results [3], [13], [14].

The existence of effect modifier variables is a further factor to consider. An effect modifier alters the observed impact of a risk factor on an outcome. When effect modification is occurring, the risk estimates for various groups of individuals differ. Interaction is a concept that is frequently used to denote a similar data phenomenon. Generally, interaction refers to the circumstance in which the effect of one variable on an outcome is contingent upon the value of another variable [2], [3], [12], [14], [15], [16], [17]. The words interaction and effect modification are frequently used synonymously. This is illustrated in regression studies by the inclusion of product terms. However, they are two distinct but related statistical and epidemiological concepts.

This paper aims to examine a scenario with a binary exposure and a binary outcome, and illustrates how to conduct a Mantel-Haenszel analysis [18], [19], [20] and a logistic regression [21], [22] at the analysis stage, while adjusting for potential confounding variables. The similarities between these two statistical approaches are highlighted. This paper is organized as follows:

In section 2, the Mantel-Haenszel and logistic regression analyses are described briefly. Four examples contrasting Mantel-Haenszel analysis and logistic regression are included and discussed in Section 3. Section 4 provides concluding remarks.

2 Methods

2.1 Mantel-Haenszel analysis

The Mantel-Haenszel analysis is a stratified analysis that adjusts for potential confounding variables [18]. It permits the gathering of existing data in several contingency tables depending on distinct subgroups or strata of a larger population. Each table is organized as shown in Table 1, where the letters a, b, c, and d represent the cell frequencies.

The odds of an event are the ratio of the probability that the event will happen to the probability that it won't happen. The odds ratio (OR) compares the odds of an outcome occurring with and without a particular exposure [23]. The OR provides an estimate of the strength of the association between the exposure and outcome variables. Referring to Table 1, the OR for experiencing the outcome while exposed versus not exposed can be computed as equation (1).

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc} \tag{1}$$

These strata are distinguished from one another based on different levels of a potential confounding variable. If the data in the various strata are actually homogeneous, the data in the various tables can be integrated. This is accomplished by using an OR homogeneity test. A non-significant result suggests that the data support the hypothesis that the putative confounding variable is not an effect modifier and that the different strata are homogenous. In this case, this approach produces a summary statistic known as the common OR, which measures the relationship between the exposure and the outcome. The OR is calculated independently for each stratum, and then the overall common OR is derived by averaging the individual odds ratios using a weighted mean. This enables the results to be merged in a manner that takes sample size and precision into account [24].

The robustness of the Mantel-Haenszel approach in the presence of confounding variables is one of its advantages. The Mantel-Haenszel method can account for their impact on the association between two binary variables by stratifying the data according to the levels of the confounders. This can lessen the likelihood of drawing erroneous inferences on the link between the two variables [18]. It is widely used in medical research to test the efficacy of a treatment or intervention or to estimate the risk of exposure to a given factor, by comparing the result of interest between two groups of participants [2].

However, the Mantel-Haenszel method has limitations. It is not applicable when either variable is continuous or ordinal [25]. Potential confounding variables must be handled as discrete or aggregated variables. Also, the Mantel-Haenszel

Table 1. Contingency table for each strata

	Outcome	
Exposure	yes	no
yes	a	b
no	c	d

approach presupposes that the data be randomly picked from the target population. In addition, the approach relies on large sample sizes to get accurate estimates, hence it may not be suitable for small or sparse data sets.

Procedure:

Break down the study population into subgroups according to the levels of potential confounding variables, that is, create m tables 2×2, exposure versus outcome, each representing a different level of the confounding variable.

Calculate the ORs for each of the m contingency tables and run a homogeneity test on the m ORs.

In case the homogeneity test does reject the null hypothesis (p-value less than 0.05), the data cannot be merged since the m OR are not homogeneous and effect modification may be present, and separate bivariate analyses should be conducted.

Otherwise, if the homogeneity test does not reject the null hypothesis (p-value greater than 0.05), one can assume that homogeneity (lack of effect modification) exists, and the data from the m tables can be combined on a statistic called the common OR [18]. Then, the Mantel-Haenszel conditional independence test is used to assess the independence of the m tables across the various levels of the potential confounder. The test compares the null hypothesis, which asserts that there is no partial link between exposure and outcome in any stratum, to the alternative, which asserts that the partial link, as assessed by the common OR, is not equal to one. It is an extension of the Chi-square test. This is an overall test of association. The test provides a p-value indicating if conditional independence is met.

2.2 Logistic regression

In this type of regression, the logit function is used as a link function to establish a linear relationship between p independent variables and the probability that the response variable y, would be equal to one, according to equation (2) [21], [22].

$$\begin{aligned} logit\,(P(y=1)) &= \log(odds(P(y=1))) = \\ &= \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \end{aligned} \tag{2}$$

The purpose of logistic regression is to model the association between a binary dependent variable and a set of independent variables. The dependent variable is assumed to follow a logistic distribution. This distribution is commonly used to model discrete events with just two possible outcomes. The objective of logistic regression is to determine the model that best predicts a binary result, such as the probability of an event happening. This makes it useful for tasks like assessing the likelihood that a consumer will make a purchase based on their demographics, or forecasting whether an individual will develop a given disease based on their risk factors [26], [27].

In order to account for confounding, confounder variables must be included in the collection of independent variables.

For a binary variable x_j coded as zero and one, the OR is computed as the exponencial of the coefficient β_j estimated from the logistic regression model, that is e^{β_j}.

By incorporating cross terms into the model, logistic regression accounts for effect modification. If the binary variable x_j coded as zero and one is part of an interaction cross term $x_j x_k$, the OR is given by $e^{\beta_j + \beta_r x_k}$, being β_j and β_r the coefficients of x_j and $x_j x_k$ respectively. Hence, the effect of x_j on the outcome is dependent on the value of x_k.

Logistic regression has various advantages. It is relatively quick to train and is capable of processing massive datasets. Furthermore, it is robust to noise and can manage scenarios in which the data contains a significant number of irrelevant or overlapping characteristics [28], [29].

Logistic regression, on the other hand, has some drawbacks. It is a binary classifier. So it can only be used to make predictions between two categories. It is also quite sensitive to datasets in which one category is far more common than the other. It is clear that if the model favors the more common category, it will underperform on the other category [30].

The so-called crude OR is estimated modelling the relation between the outcome variable and the exposure variable without accounting for other variables. It is an estimate of the OR for the association between two variables: exposure and outcome. The crude OR is the ratio of the odds or chances that the outcome will happen in the exposed group to the odds or chances that it will happen in the control group. This figure can be accomplished using logistic regression. The exponencial of the coefficient estimated from the logistic regression model that includes only the exposure and outcome variables, gives the crude OR of the data. Alternatively, it can be computed using a single table including all the data. However, the crude OR might not adequately represent the actual association between the exposure and outcome variables. In contrary, a common OR is a more precise assessment of the link between the exposure and outcome variables since it accounts for the impact of potential confounding variables [23], [31]. It is generally computed with the Mantel-Haenszel technique, which adjusts the crude OR for confounding variables. Yet, the common OR from Mantel-Haenszel analysis, can be computed using logistic regression. In order to do this it is sufficient to analyze the association between the outcome and the exposure variables while controlling for the relevant confounders in logistic regression. The common ORs is then interpreted by taking into account its magnitude and statistical significance [23], [31].

This way, Mantel-Haenszel analysis can be replaced by logistic regression. Indeed, logistic regression is a generalization of the Mantel-Haenszel method, which is used to investigate the association between two binary variables. Logistic

regression, on the other hand, can be used to investigate the connection between a binary outcome and one or more predictor variables. It is also capable of dealing with more complex data structures, such as interactions between numerous predictor variables [24]. Nevertheless, the precise statistical method to employ will be determined by the data and study issue being investigated.

In short, the association between two variables can be examined using either the Mantel-Haenszel technique or logistic regression. Confounding variables can be accounted for and the strength of the association between the variables can be assessed using both approaches. Mantel-Haenszel analysis uses stratification by levels of confounders to eliminate the effects of these variables. By incorporating them as new independent variables in the regression model, confounding factors can be accounted for and removed in logistic regression.

To assess the significance of the various models in logistic regression, the likelihood ratio test is used. The likelihood ratio test for a group of variables compares the likelihood L of the full model, with the likelihood of the reduced model without those variables. The statistic used is $-2log(L) = -2LL$, which takes only non-negative values, with zero representing the perfect fit.

3 Examples

First study (Table 2): Consider a case-control study that was conducted to look at the connection between working in a poorly ventilated auto shop and developing lung cancer. In this instance, smoking is a potential confounding variable because it is known to be associated with lung cancer and because mechanics that work at poorly ventilated auto shops are likely smokers.

The main issue is whether there is a causal relationship between working in a poorly ventilated auto shop and having lung cancer. The exposure variable, working in a poorly ventilated auto shop, is being investigated as the primary risk factor. Nevertheless, if the confounding variable, smoking, is not an effect modifier (smoking does not modify the association between working in a poorly ventilated auto shop and lung cancer), one does not want to draw different results for each level of smoking. In this case, it is desirable to combine the data in order to draw a conclusion. The Mantel-Haenszel method is frequently employed to achieve this goal when both the exposure and the disease are binary.

Here is an overview of the procedure: four strata or levels of smokers were considered (non-smokers, casual smokers, moderate smokers and heavy smokers), hence the data are presented in four tables 2×2, one table for each level of smokers. Three examples are considered. The data is fictitious and is shown in Table 2. This data is used for examples 3.1 through 3.3.

Second study (Table 3): Consider a retrospective observational research in which it is intended to determine whether the cure rate for a given condition using a significantly more expensive new drug is greater than the cure rate using a drug that has been on the market for many years. Considering that hypertension worsens the condition in concern and that individuals may not be equally distributed across the two therapies, hypertension was considered as a possible confounding variable. The data is shown in Table 3. This data is used for Example 3.4.

3.1 Example 1

3.1.1 Mantel-Haenzel analysis

The crude OR obtained for the entire data table is 1.970. This statistics compares the odds of developing lung cancer when working in a poorly ventilated auto shop to the odds of having lung cancer while not doing so.

The homogeneity test of the ORs yields a p-value = 0.746. The result is non-significant, therefore the null hypothesis is not rejected, and it is reasonable to believe that the ORs for the four populations characterized by the four levels of daily cigarette consumption are identical.

The Mantel-Haenszel common OR is then calculated and yields a value of 1.972.

The common OR and the crude OR are quite close, suggesting that there is no confounding between the exposure (working in an auto shop with poor ventilation) and the outcome (lung cancer) association. When confounding is present the crude OR differs by 10 to 20% from the common OR.

Nevertheless a chi-square test can be run between the potential confounding variable (smoking habits) and the exposure (working in an auto shop with poor ventilation) to uncover potential imbalances in the distribution of smoking habits among the two exposure levels. As there is a documented relationship between smoking and lung cancer, an imbalance in the distribution may trigger confounding in the association between working in an auto shop with poor ventilation and lung cancer. The Pearson chi-square p-value yields the value 0.972, which is not statistically significant and indicates that confusion is not an issue.

The test for conditional independence of Mantel-Haenszel yields a p-value < 0.001, indicating a significant association between the two variables (working in a poorly ventilated auto shop and lung cancer).

3.1.2 Logistic regression

Table 4 displays the obtained outcomes.

Model 1 is significant and describes the association between the exposure (working in an auto shop with poor ventilation) and the outcome (lung cancer). The exponencial of the variable exposure's coefficient yields the crude OR, identical to the one derived from the Mantel-Haenszel analysis, $e^{0.678} = 1.970$.

Model 2 adds smoking habits as a potential confounder to Model 1. This addition is non-significant as revealed by the p-value = 0.408, indicating that the confounding variable is unnecessary because confounding may not be present. The exponencial of the variable exposure's coefficient yields the common OR identical to the one derived from the Mantel-Haenszel analysis, $e^{0.679} = 1.972$. Both the crude and the common OR are similar since confounding is not an issue.

The assessment of homogeneity across strata of the confounding variable is based on the likelihood ratio test from

Table 2. Data related to the first three examples

Exposure	Outcome	Confounder	Frequency Example 1	Frequency Example 2	Frequency Example 3
yes	case	non-smoker	57	60	30
yes	control	non-smoker	78	20	65
no	case	non-smoker	25	56	50
no	control	non-smoker	75	67	145
yes	case	casual smoker	120	40	55
yes	control	casual smoker	156	15	91
no	case	casual smoker	54	57	105
no	control	casual smoker	150	43	230
yes	case	moderate smoker	90	150	30
yes	control	moderate smoker	117	51	20
no	case	moderate smoker	40	46	83
no	control	moderate smoker	112	45	210
yes	case	heavy smoker	138	200	45
yes	control	heavy smoker	183	40	21
no	case	heavy smoker	77	61	61
no	control	heavy smoker	172	40	250

Table 3. Data applicable to the fourth example

Treatment	Cure	Confounder	Frequency
New drug	no	non-hypertensive	5
Old drug	no	non-hypertensive	37
New drug	yes	non-hypertensive	84
Old drug	yes	non-hypertensive	230
New drug	no	hypertensive	67
Old drug	no	hypertensive	24
New drug	yes	hypertensive	210
Old drug	yes	hypertensive	57

model 2 to model 3 when the crossed term is included. In this case the interaction term is non-significant.

In conclusion, there is likely a link between the exposure (working in an auto shop with poor ventilation) and the outcome (lung cancer).

3.2 Example 2

3.2.1 Mantel-Haenzel analysis

The crude OR for the table containing all the data is 3.166.

The p-value for the homogeneity test of the ORs is 0.645. Because the result is not statistically significant, the null hypothesis is not rejected, and the four ORs can be combined. The ORs are identified to be homogeneous, and there appears to be no effect modification, so the four ORs can be merged. The Mantel-Haenszel common OR produces a value of 2.955 in this instance, a variation of less than 10% in the crude OR.

Nevertheless, the Pearson chi-square p-value between the potential confounding variable (smoking) and the exposure (working in an auto shop with poor ventilation) is less than 0.001, a statistically significant value indicating the possibility of confusion.

Being the ORs homogeneous, the test for conditional independence of Mantel-Haenszel yields a p-value < 0.001, indicating a significant association between the two variables (working in a poorly ventilated auto shop and lung cancer).

3.2.2 Logistic regression

The observed results are reported in Table 5.

Model 1 is significant and describes the association between the exposure (working in an auto shop with poor ventilation) and the outcome (lung cancer). The exponencial of the variable exposure's coefficient yields the crude OR identical to the one derived from the Mantel-Haenszel analysis, $e^{1.152} = 3.166$.

Model 2 adds smoking habits as a possible confounding factor to Model 1. This addition is statistically significant, as indicated by the p-value = 0.007. Confusion is a concern that must be handled. The exponencial of the variable exposure's coefficient provides an OR of $e^{1.050} = 2.858$, which differs slightly from the 2.955 obtained for the common OR via the Mantel-Haenszel analysis. The ordinal nature of the confounding variable in logist regression, which is treated as a numeric variable by the model, is one of the factors contributing to this little difference, besides the fact that the computations for the two methods are clearly distinct. Another alternative that could be explored, is to create three dummy variables to treat the ordinal variable as nominal in logistic regression.

Model 3 includes an interaction term that was found to be non-significant. The conclusion is that it is fair to assume that effect modification is absent and that the strata are homogeneous.

Model 2 is the final model and reveals that there is a significant association between the two variables, working in a poorly ventilated auto shop and lung cancer (p-value < 0.001).

3.3 Example 3

3.3.1 Mantel-Haenzel analysis

The calculated crude OR for the table containing all of the data is 2.268.

Table 4. Successive models for the logistic regression (example 1)

Model	constant	exposure	confounder	cross term	$-2LL$	$-2LL$ reduction	p-value
1	-0.954	0.678			2117.358	41.401	< 0.001
2	-1.026	0.679	0.040		2116.675	0.684	0.408
3	-1.132	0.849	0.099	-0.095	2115.771	0.903	0.342

Table 5. Successive models for the logistic regression (example 2)

Model	constant	exposure	confounder	cross term	$-2LL$	$-2LL$ reduction	p-value
1	0.121	1.152			1178.975	69.259	< 0.001
2	-0.122	1.050	0.173		1171.618	7.357	0.007
3	-0.105	1.005	0.161	0.027	1171.574	0.044	0.834

The p-value for the homogeneity test of the ORs is less than 0.001. Because the result is statistically significant, the null hypothesis is rejected, and the four ORs cannot be combined. There is evidence that different smoking habits can modify the link between working in a poorly ventilated auto shop and lung cancer. The four ORs are not homogeneous. Analyses must be conducted within each strata. As a result, any calculation of the OR of developing lung cancer as a result of working in a poorly ventilated auto shop must take smoking habits into account.

3.3.2 Logistic regression

Table 6 displays the observed outcomes.

Model 1 is significant and describes the association between the exposure (working in an auto shop with poor ventilation) and the outcome (lung cancer). The crude OR is identical to the one derived from the Mantel-Haenszel analysis, $e^{0.819} = 2.268$.

Model 2 adds smoking habits as a possible confounding factor to Model 1. This addition is not statistically significant, as indicated by the p-value = 0.383. The exponencial of the variable exposure's coefficient provides an OR of $e^{0.838} = 2.311$, which slightly deviates from the value 2.277, that would be produced for the common OR using the Mantel-Haenszel method, if the strata were homogenous.

The interaction term in Model 3 was found to be statistically significant. As shown already by the Mantel-Haenszel analysis, there is evidence that effect modification is present. Analyses must be carried out within each strata.

3.4 Example 4

3.4.1 Mantel-Haenzel analysis

With all of the data, a crude OR of 0.868 was obtained. This figure indicates that the odds of getting cured with the new drug is 0.868 the odds of being cured with the old drug. Given that the value is less than one, it seems likely that the older drug is superior.

The p-value for the homogeneity test of the ORs is 0.200. Because the result is non-significant, the null hypothesis is not rejected, and the two ORs are identified to be identical and must be combined in one value. The Mantel-Haenszel common OR is then computed and yields a value of 1.641. The OR increased by nearly 90%. This figure provides a 180-degree turn. With this value, the odds of being cured by the new drug is 1.642 more than the odds of being cured by the old drug. The update conclusion indicates that the new drug is superior to the old one. These results reveal considerable confusion over the relative efficacy of the drugs.

As the ORs are believed to be homogeneous, the test for conditional independence of Mantel-Haenszel produces a p-value of 0.053, indicating that the prescription of the new drug may result in a significant improvement in the cure rate.

3.4.2 Logistic regression

The results observed are shown in Table 7.

Model 1 is non-significant. The exponencial of the coefficient of the variable treatment yields the crude OR identical to the one obtained from the Mantel-Haenszel analysis, $e^{-0.142} = 0.868$.

Model 2 adds the confounder (to be hypertensive or not). This addition is statistically significant, as indicated by the p-value < 0.001. In the previous association, there is confusion. The exponencial of the coefficient of the variable treatment provides a common OR of $e^{0.473} = 1.605$ (similar to the value 1.641 via the Mantel-Haenszel analysis), which differs patently from the crude OR obtained.

Model 3 includes an interaction term that was found to be non-significant. The conclusion is that it is fair to assume that effect modification is absent and that the strata are homogeneous.

Model 2 is the final model and reveals that there is a significant association between treatment and cure rate (p-value = 0.041, not displayed in Table 7) and that the prescription of the new drug may result in a significant increase in the cure rate and not a decrease as suggested by the data before controlling for the confounder (being hypertensive).

4 Conclusion

This paper addresses four case studies that include confounding and/or effect modification, in an effort to help readers become better decision-makers throughout the analysis phase of a research.

The association between exposure and outcome variables can be skewed if confounders aren't accounted for, hence doing so is a crucial part of biostatistics. By removing the influence of any confounding factors, researchers are better able to determine the direction and the strength of any association between

Table 6. Successive models for the logistic regression (example 3)

Model	constant	exposure	confounder	cross term	$-2LL$	$-2LL$ reduction	p-value
1	-1.027	0.819			1799.382	41.607	< 0.001
2	-1.105	0.838	0.047		1798.620	0.762	0.383
3	-0.802	-0.105	-0.141	0.695	1766.860	31.760	< 0.001

Table 7. Successive models for the logistic regression (example 4)

Model	constant	treatment	confounder	cross term	$-2LL$	$-2LL$ reduction	p-value
1	1.549	-0.142			686.006	0.541	0.462
2	1.909	0.473	-1.189		660.050	25.956	< 0.001
3	1.827	0.994	-0.962	-0.717	658.310	1.740	0.187

the exposure and the result of interest. For instance, this can be useful for figuring out what might go wrong with a health problem and what might work to fix it.

Because the impact of treatments or events on results might vary according to the values of other variables, adjusting for effect modification is also a crucial step in biostatistical analyses. For instance, if one wants to investigate the link between smoking and lung cancer, one might find that the association weakens or strengthens with increasing or decreasing age. Thus, age acts as a moderator or effect modifier of the association between smoking and lung cancer. Understanding how the association between exposure and outcome changes is improved by controlling for effect modification. This is useful for determining which segments of the population are more likely to be adversely affected by an event and which efforts could be most successful in protecting those segments. It is possible to better pinpoint potential confounders if one accounts for the possibility of an altered effect. If one find that the association between a treatment and a endpoint changes with age, for instance, that could indicate that age is a confounder in the study.

Two common techniques for doing so are the Mantel-Haenszel analysis and the logistic regression. Mantel-Haenzel analysis demands that the exposure variable be segmented as binary and that the sample be stratified by potential confounding variables. Therefore, the confounding variables must likewise be segmented. The use of continuous variables and multiple interaction terms are not only permitted, but recommended if necessary, in logistic regression. Despite the fact that the data may need to be adjusted for each of the analyses, the results are typically the same. Both approaches have their merits, and how they are put to use will vary from one study to the next. Nevertheless, it is essential to accurately conduct and evaluate the results of these approaches.

REFERENCES

[1] Benasseur, I., Talbot, D., Durand, M., Holbrook, A., Matteau, A., Potter, B.J., Renoux, C., Schnitzer, M.E., Tarride, J.E., Guertin, J.R., "A comparison of confounder selection and adjustment methods for estimating causal effects using large healthcare databases," vol. 31, no. 4, pp. 424-433, 2021. https://doi.org/10.1002/pds.5403

[2] Le C.T., "Descriptive Methods for Categorical Data," in Introdutory Biostatistics, 1st ed, John Wiley & Sons, 2003, pp. 1-56. https://www.wiley.com/en-us/Introductory+Biostatistics-p-9780471308881

[3] Bruce N., Pope D., Stanistreet D., "Quantitative Methods for Health Research. A Practical Interactive Guide to Epidemiology and Statistics," John Wiley & Sons Ltd, 2018. https://doi.org/10.1002/9781118665374

[4] Lazcano G., Papuzinski C., Madrid E., Arancibia M., "General Concepts in Biostatistics and Clinical Epidemiology: Observational Studies with Cohort Design," Medwave, vol. 19, no. 11, e7748, 2019. https://doi.org/10.5867/medwave.2019.11.7748

[5] Cataldo R., Arancibia M., Stojanova J., Papuzinski C., "General Concepts in Biostatistics and Clinical Epidemiology: Observational Studies with Cross-Sectional and Ecological Designs," Medwave, vol. 19, no. 8, e7698, 2019. https://doi.org/10.5867/medwave.2019.08.7698

[6] Martínez D., Papuzinski C., Stojanova J., Arancibia M., "General Concepts in Biostatistics and Clinical Epidemiology: Observational Studies with Case-Control Design," Medwave, vol. 19, no. 10, e7716, 2019. https://doi.org/10.5867/medwave.2019.10.7716

[7] D'Onofrio B.M., Sjölander A., Lahey B.B., Lichtenstein P., Öberg A.S., "Accounting for Confounding in Observational Studies,", Annual Review of Clinical Psychology, vol. 16, 25-48, 2020. https://doi.org/10.1146/annurev-clinpsy-032816-045030

[8] Franklin J.M., Platt R., Dreyer N.A., London A.J., Simon G.E., Watanabe J.H., Horberg M., Hernandez A., Califf R.M., "When Can Nonrandomized Studies Support Valid Inference Regarding Effectiveness or Safety of New Medical Treatments?," vol. 111, no. 1, 108-115, 2022. https://doi.org/10.1002/cpt.2255

[9] Mathur M.B., VanderWeele T.J., "Assessing Uncontrolled Confounding in Associations of Being Overweight With All-Cause Mortality," JAMA Network Open, vol. 5, no. 3, e222614, 2022. https://doi.org/10.1001/jamanetworkopen.2022.2614

[10] Sjölander A., "Estimation of Marginal Causal Effects in the Presence of Confounding by Cluster," Biostatistics, vol. 22, no. 3, pp. 598–612, 2021. https://doi.org/10.1093/biostatistics/kxz054

[11] Zetterqvist J., Vansteelandt S., Pawitan Y., Sjölander A., "Doubly Robust Methods for Handling Confounding by Cluster," Biostatistics, vol. 17, no. 2, pp. 264–276, 2016. https://doi.org/10.1093/biostatistics/kxv041

[12] Vetter T.R., Mascha E.J., "Bias, Confounding, and Interaction: Lions and Tigers, and Bears, Oh My!", Anesthesia & Analgesia, vol. 125, no. 3, pp. 1042-1048, 2017. https://doi.org/10.1213/ANE.0000000000002332

[13] Kashner T.M., Henley S.S., Golden R.M., Zhou X.-H., "Making causal inferences about treatment effect sizes from observational datasets," Biostatistics & Epidemiology, vol. 4, no. 1, pp. 48-83, 2020. https://doi.org/10.1080/24709360.2019.1681211

[14] Chen J., Ho M., Lee K., Song Y., Fang Y., Goldstein B.A., He W., Irony T., Jiang Q., Laan M., Lee H., Lin X., Meng Z., Mishra-Kalyani, P., Rockhold, F., Wang, H., White, R., "The Current Landscape in Biostatistics of Real-World Data and Evidence: Clinical Study Design and Analysis," Statistics in Biopharmaceutical Research, vol. 15, no. 1, pp. 29-42, 2023. https://doi.org/10.1080/19466315.2021.1883474

[15] Andriani H., "Exposure to Parental Smoking and Children being Overweight: Residence as an Effect Modifier," Journal of Public Health (Berl.), vol. 29, pp. 495–502, 2021. https://doi.org/10.1007/s10389-019-01153-6

[16] Bahamyirou A., Schnitzer M.E., Kennedy E.H., Blais L., Yang Y., "Doubly Robust Adaptive LASSO for Effect Modifier Discovery," The International Journal of Biostatistics, vol. 18, no. 2, pp. 307–327, 2020. https://doi.org/10.1515/ijb-2020-0073

[17] Karmakar B., Heller R., Small D.S., "False Discovery Rate Control for Effect Modification in Observational Studies," Electronic Journal of Statistics, vol. 12, pp. 3232–3253, 2018. https://doi.org/10.1214/18-EJS1476

[18] Mantel N., Haenszel W., "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease," Journal of the National Cancer Institute, vol. 22, no. 4, pp. 719-748, 1959. https://doi.org/10.1093/jnci/22.4.719

[19] Kuritz S.J., Landis J.R., Koch G.G, "A General Overview of Mantel-Maenszel Methods: Applications and Recent Developments," Annual Review of Public Health, vol. 9, pp. 123–160, 1988. https://doi.org/10.1146/annurev.pu.09.050188.001011

[20] Vierra A., Razzaq A., Andreadis A., "Categorical Variable Analyses: Chi-square, Fisher Exact, and Mantel–Haenszel," in Translational Surgery (Handbook for Designing and Conducting Clinical and Translational Research), Academic Press, 2023, pp. 171-175. https://doi.org/10.1016/B978-0-323-90300-4.00095-1

[21] Sperandei S., "Understanding logistic regression analysis," Biochemia Medica, vol. 24, no. 1, pp. 12-18, 2014. https://doi.org/10.11613/BM.2014.003

[22] Osborne J:W., "Best Practices in Logistic Regression," SAGE Publications, Inc., 2015. https://doi.org/10.4135/9781483399041

[23] Breslow N.E., Day N.E, "Statistical Methods in Cancer Research Volume I: The Analysis of Case-Control Studies," IARC Scientific Publications, 1980. https://publications.iarc.fr/Book-And-Report-Series/Iarc-Scientific-Publications/Statistical-Methods-In-Cancer-Research-Volume-I-The-Analysis-Of-Case-Control-Studies-1980

[24] Agresti A., "Categorical Data Analysis," 3rd ed, John Wiley & Sons, 2013. https://www.wiley.com/en-us/Categorical+Data+Analysis2C+3rd+Edition-p-9780470463635

[25] Fleiss J.L., Levin B., Paik M.C., "Combining Evidence from Fourfold Tables," in Statistical Methods for Rates and Proportions, 3rd ed, John Wiley & Sons, 2003, pp. 234-283. https://doi.org/10.1002/0471445428.ch10

[26] Christopher M.B., "Linear Models for Classification," in Pattern Recognition and Machine Learning, Springer, 2006, pp. 179-224. https://link.springer.com/book/9780387310732

[27] Hosmer D.W., Lemeshow S., Sturdivant R.X., "Special Topics," in Applied Logistic Regression, 3rd ed, John Wiley & Sons, 2013, pp. 377-458. https://doi.org/10.1002/9781118548387

[28] Hastie T., Tibshirani R., Friedman J., "Linear Methods for Classification," in The Elements of Statistical Learning Data Mining Inference and Prediction, 2nd ed, Springer, 2009, pp. 101-138. https://link.springer.com/book/10.1007/978-0-387-84858-7

[29] Alpaydin E., "Supervised Learning," in Introduction to Machine Learning, 4th ed, MIT Press, 2020, pp. 23-50. https://mitpress.ublish.com/book/introduction-to-machine-learning--4

[30] Chawla N.V., Bower K.W., Hall L.O., Kegelmeyer W.P., "SMOTE: Synthetic Minority Over-Sampling Technique," Journal of Artificial Intelligence Research, vol. 16, no. 1, 321–357, 2002. DOI:10.1613/jair.953

[31] Rothman K.J., Greenland S., Lash T.L., "Modern Epidemiology," 4th ed, Lippincott Williams & Wilkins, 2008. https://www.rti.org/publication/modern-epidemiology-3rd-edition