Special Proceedings: ISBN #: 978-81-950383-2-9
25th Annual Conference, 15-17 February 2023; pp 169-203

Analysis of Large Medical Databases: Addressing the Clinical Relevance of Statistically Significant Findings

Rachana Lele[1a,b], Anand Seth[2], Sameer Patel[1d], Jianmin Pan[1a-c], Marepalli B. Rao[1a], and Shesh N. Rai[1a-c, 3]

[1a] Division of Biostatistics and Bioinformatics, Department of Environmental and Public Health Sciences, College of Medicine, University of Cincinnati, Ohio
[1b]Cancer Data Science Center, Department of Environmental and Public Health Sciences, College of Medicine, University of Cincinnati, Ohio
[1c]Biostatistics and Informatics Shared Resources, University of Cincinnati Cancer Center, Ohio
[1d]Department of Surgery, College of Medicine, University of Cincinnati, Ohio
[2]Research and Development, SK Patent Associates, LLC, Ohio
[3]Division of Environmental Cardiology, School of Medicine, University of Louisville

Received: 01 August 2023; Revised: 13 August 2023; Accepted: 16 August 2023

Abstract

The analysis of studies using large medical databases has gained popularity due to their ability to provide extensive and diverse samples. However, in the currently published literature, the selection of samples in such studies often relies on inclusion criteria based solely on the study's objectives, rather than utilizing formal sample size calculation techniques. Also, inferences are predominantly drawn based on *p*-values, which tend to be highly significant due to large samples but may lack clinical relevance. In this article, we explore the issue of statistically significant *p*-values but with limited clinical relevance when analyzing large databases. We propose the incorporation of effect sizes, a concept well-established in the literature, to supplement *p*-values in assessing the practical significance of research findings. To address the unique challenges of analyzing large samples using logistic regression, we present a novel effect size measure specifically tailored for this context. Moreover, we introduce conventions for interpreting effect sizes when analyzing large databases, thus providing researchers with a standardized approach for evaluating the magnitude of the observed associations. To validate the proposed effect size measure, we employ state-of-the-art machine learning techniques on the same datasets and demonstrate its robustness and utility in large-scale medical studies. To illustrate the statistical challenges and the application of our novel effect size measure, we present a compelling case study utilizing breast cancer data from the National Cancer Database (NCDB). Our findings shed light on the potential pitfalls of relying solely on *p*-values in large database studies and highlight the significance of incorporating effect sizes to better understand the clinical implications of research results. By emphasizing the importance of effect sizes in addition to *p*-values, this study aims to improve the accuracy and clinical relevance of statistical analyses for large medical databases. Implementing our suggested approach can lead to more informative and meaningful insights, thereby contributing to the advancement of evidence-based medicine and patient care.

Corresponding Author: Shesh N. Rai
Email: raise@ucmail.uc.edu

Key words: National cancer database; NCDB; Breast cancer; *p*-values; Effect sizes; Machine learning.

1. Introduction

Most traditional statistical analysis methods, such as linear regression, *t*-test, ANOVA, *etc.*, require the assumption of normality and randomized study sample selection. While conducting clinical research to test the safety and efficacy of new drugs, randomization is an essential component as well. Such studies require careful selection of a representative sample which is achieved using formal randomization techniques.

A relatively new branch of study, often referred as Real-World Data (RWD) and Real World Evidence (RWE), involves analyzing databases maintained by various government and private institutions to discover new insights related to public health which were previously underutilized, Breckenridge *et al.* (2019). Since data collection is lengthy and expensive, readily available databases provide an excellent alternative to conducting research and help save time, cost, and resources and complement evidence obtained from randomized clinical studies. However, large databases (which may also be referred as 'big data') are repositories of majority of the actual observed cases; hence, these are not randomized. Being extremely large, normality assumption is unrealistic and often unmet for most of these databases. Hence, using traditional parametric tests for such databases tend to produce highly significant results which may have no clinical relevance. The traditional statistical tests run using these large databases tend to produce highly significant *p*-values and inferences based on *p*-values alone and could lead to misleading or incorrect conclusions. Several articles such as Sullivan and Feinn (2012) and Solla *et al.* (2018) explore the alternative of using effect sizes and provide criticism for the use of *p*-values alone talk about the use of effect sizes and confidence intervals in addition to using p-values. They note highly significant *p*-values with small effect sizes may be clinically irrelevant as suggested by Ranstam *et al.* (2012) that confidence interval is a better alternative to using *p*-values.

Cohen (1988) and Cohen (1992) introduced the concept of effect sizes and defined it as the discrepancy between the null and the alternative hypotheses. He suggested formulae for effect sizes using normally distributed outcomes as well as proportions which were respectively popularized as Cohen's *d* and Cohen's *h*. A strength of the effect size measure is that it does not directly depend on the sample size and hence, is unaffected by large sample sizes. Consequently, for big data analysis involving large databases, effect size could be a better inferential measure than the traditional *p*-values. However, in the case of a logistic regression in which we are comparing effects of two treatments within two different categories of a variable, the Cohen's *h* effect size cannot be used in the present form.

In this paper, we argue that *p*-values alone can provide misleading results for extremely large sample sizes since *p*-value calculations depend on sample size. We compare the p-values obtained from an overall test and those obtained from individual tests as well as Bonferroni adjusted *p*-values. As an alternative to using *p*-values, we propose a modification/extension of Cohen's *h* effect size estimator for logistic regression. We validate our results obtained using the new Cohen's *h* measure using machine learning techniques such as Association Rule Mining (ARM) and Naïve Bayes classifier.

The organization of the paper is described as follows. In Section 2, we introduce statistical formulation of the issue of obtaining highly significant *p*-values for large sample sizes. We formally introduce the concept of alpha adjustment for multiple comparisons and introduce the theory of Bonferroni method of multiplicity adjustment. Furthermore, we

introduce the theory of logistic regression, and we introduce the concept of effect size and explore the theory of different effect size measures. Lastly, we provide the theory for ARM and Naïve Bayes classifier. In Section 3, we describe our proposed modification/extension of Cohen's *h* measure which can be applied to logistic regression analysis. In Section 4, we describe different statistical issues in the analysis of large databases using National Cancer Database (NCDB) as an example and present a literature review of articles published using NCDB. In Section 5, we present a case study using breast cancer data from NCDB to demonstrate the central issue of p-values addressed in this paper and how our proposed modified Cohen's *h* effect size will lead to 'meaningful statistical significance' as opposed to clinically irrelevant statistical significance. We provide a strong support for our arguments by using ARM and Naïve Bayes classifier methods.

2. Statistical methods

2.1. Wald test

When analyzing data using statistical tests, *p*-values are often used to draw inferences. We will illustrate the effect of large sample size on *p*-values using the Wald test as established by Wald (1974). The traditional *t*-test statistic, binomial test statistic, Poisson test statistic *etc*. are special cases of the Wald test statistic. We will illustrate the dependence of the test statistic on the sample size using the simple case of Wald test for Bernoulli random variable. In the case of Bernoulli test as described by Klotz (1973), we observe independent binary responses, and we wish to draw inferences about the probability of an event in the population.

Suppose we sample *n* individuals from a pre-specified population and the probability of occurrence of an event in this population is the same for an individual, say, *p*.

Let Y_i denote the occurrence of an event for each individual *i*. Here, we define $Y_i = 1$ if an event occurs and $Y_i = 0$, otherwise. Thus, the observed data would be given by $Y_1, Y_2, \ldots, Y_n$.

The maximum likelihood estimate (MLE) of *p* is given by

$$\hat{p} = \frac{\sum_{i=1}^{n} Y_i}{n} \tag{1}$$

Now, suppose we are testing the hypothesis $H_0: p = p_0$ *vs.* $H_1: p \neq p_0$.

The Wald test statistic (*W*) is given by a difference in the MLE estimate of *p* and the hypothesized value, normalized by the MLE estimate of the standard deviation.

Thus, we have

$$W = \frac{(\hat{p}-p_0)^2}{\hat{p}(1-\hat{p})/n} \tag{2}$$

This test can be extended for the case of logistic regression to test the significance of regression coefficients.

For $E(Y_i) = \pi_i$, we have

$logit(\pi_i) = \beta_0 + \beta_0 x_{i1} + \cdots + \beta_p x_{ip} = x_i'\beta$, where $x_i = \left(1, x_{i1}, \ldots, x_{ip}\right)'$ and $\beta = \left(\beta_0, \beta_1, \ldots, \beta_p\right)'$.

To test a single β coefficient value, the Wald test statistic will be given by

$$Z = \frac{\hat{\beta}_j - \beta_{j0}}{\widehat{se}(\hat{\beta})} \sim N(0,1) \tag{3}$$

where $\widehat{se}\left(\hat{\beta}\right)$ is calculated by taking the inverse of the estimated information matrix.

From equation (2), we observe that *W* statistic depends on the sample size, *n*. Note as *n* becomes large in equation (2), *i.e.*, as $n \to \infty$, $W \to \infty$. Similarly, for the case of logistic regression, using equation (3), as $n \to \infty$, $Z \to \infty$.

The *p*-value may be defined as the probability of observing a test statistic as extreme as the one observed if the null hypothesis were true. Alternatively, *p*-value is the observed risk of rejecting H_0. For the Wald test, we have *p*-value, $p' = P(|Z| > |T_{obs}|)$, where T_{obs} is the observed value of the test statistic.

Thus, using (3) and definition of p', as $Z \to \infty$, $p' \to 0$.

As described above, increasing the sample size leads to a significant increase in the value of test statistics, resulting in a very low *p*-value. This is considered highly significant in statistical terms. However, it's important to note that simply increasing the sample size does not guarantee clinical relevance. In fact, using an infinitely large sample can lead to significant results even if there is no real clinical difference, as is often the case with studies that use large databases. Therefore, it is important to reconsider the use of *p*-values when analyzing large databases to ensure that clinical relevance is accurately assessed. Thus, *p*-values alone cannot provide reliable results when sample size becomes extremely large. In addition to the use of the Wald test statistic, multiple comparisons and multiplicity adjustment are discussed in the next section.

2.2. Multiple comparisons and multiplicity adjustment

In exploratory analyses on large datasets, many hypotheses are evaluated. Sometimes when an experiment is conducted to answer a research question, multiple hypotheses may need to be tested, thereby requiring multiple comparisons to be performed. If all comparisons are simultaneously performed with an error rate of 0.05, the actual error rate gets inflated to a quantity equal to 0.05 times the number of hypothesis tests. This would reduce the reliability of the results and hence, we require an appropriate statistical inferential procedure to handle such a situation. Therefore, multiple comparison adjustments have been suggested in the literature which help in maintaining the allowable error rate at 5%. Consider a family of *k* independent null hypotheses being tested at level α. In this case, the family wise error rate (FWER) described by Ranstam *et al.* (2012) would be $1-(1-\alpha)^k$. Some commonly used multiplicity adjustment techniques are Bonferroni test, Tukey test, and Scheffé test as shown by Lee and Lee (2018). The Bonferroni method offers a higher level of rigor compared to the Tukey test, which is more permissive toward Type I errors. It also provides more leniency compared to the highly conservative Scheffé's method as indicated by Lee and Lee (2018). For a detailed description of the Bonferroni method and its application in this study, please refer to the next section.

2.2.1. Bonferroni test

When working with a family of hypotheses and their corresponding *p*-values, the Bonferroni correction can be used to control the FWER. The FWER is the probability of incorrectly rejecting at least one true null hypothesis(H_i). The Bonferroni correction involves rejecting the null hypothesis for each *p*-value that is less than or equal to alpha divided by the total number of hypotheses as described by Lee and Lee (2018). This approach effectively controls the FWER at a level of α. Boole's inequality as shown by Khrennikov (2008) provides proof that this control is achieved as follows:

$$FWER = P\left\{\cup_{i=1}^{m_0}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq \sum_{i=1}^{m_0}\left\{P\left(p_i \leq \frac{\alpha}{m}\right)\right\} = m_0\frac{\alpha}{m} \leq \alpha \quad (3)$$

where *m* = total number of null hypotheses, $H_1, \ldots, H_m$ are a family of hypotheses and $p_1, \ldots, p_m$ are corresponding *p*-values.

This control method is very versatile and flexible, though conservative, as it doesn't rely on any assumptions about the relationships between *p*-values or how many of the null hypotheses are actually true.

2.3. Logistic regression

There are different types of logistic regression, including simple, ordinal, and multiple versions of logistic and ordinal regression as described by McNulty (2021). Simple logistic regression is used when the outcome variable is binary, while ordinal regression is used when the outcome variable has multiple ordered categories. Multiple versions of logistic and ordinal regression are used depending on the complexity of the data and research question. When researchers conduct multiple statistical tests within these regression models, they may encounter multiple comparisons, which can lead to false positives. A logistic regression model, also known as the logit model, estimates the probability of occurrence of an event, such as treatment was beneficial or not, based on certain set of independent variables as described by McNulty (2021). In logistic regression, a logit transformation is performed on the odds, *i.e.*, the probability of success divided by the probability of failure. The logistic function is written as follows

$$logit\ (p_i) = \frac{1}{1+e^{-p_i}}. \quad (4)$$

The logistic regression model is written as follows.

$$ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (5)$$

Here, $logit\ (p_i) = \ln\left(\frac{p_i}{1-p_i}\right)$ is the dependent variable and $\underline{X} = (X_1, X_2, \ldots, X_k)^T$ is the vector of independent variables. Generally, the maximum likelihood estimation (MLE) method is used to estimate the beta coefficients of the logistic model.

2.4. Effect size

Effect sizes represent quantitative measures of the relationships between variables. While the term "effect size" has historically been associated with various specific measures, it is now commonly used to denote any index indicating the relationship between variables.

Effect sizes serve as a means to convey the magnitude of the relationship observed between variables in a scientific study performed by Hedges *et al.* (2008). Effect size helps in quantifying the difference between the comparison groups as described by Grissom and Kim (2005). It gives an idea about the actual difference between groups and does not directly depend on the sample size. We describe some of the common effect size measures in the following subsections.

2.4.1. Cohen's *d*

The effect size using Cohen's *d* presented by Cohen (1988) and Cohen (1992) is calculated as follows:

$$d = \frac{\mu_1 - \mu_2}{s} \tag{6}$$

Here, $\mu_1$and μ_2 are the means of the two comparison groups and *s* is the pooled standard deviation. Cohen's *d* is used for continuous outcomes and follows a general convention that *d* = 0.2 implies small effect, *d* = 0.5 implies medium effect and *d* = 0.8 implies large effect.

2.4.2. Glass's Δ

The effect size using Glass's Δ presented by Rosenthal *et al.* (1994) is calculated as follows.

$$\Delta = \frac{\mu_1 - \mu_2}{s_c} \tag{7}$$

Here, $\mu_1$and μ_2 are the means of the two comparison groups and s_c is the standard deviation of the control group. The same convention is followed for Glass's Δ effect size estimates as that for Cohen's *d* described above with respect to interpretation based on cutoff values.

2.4.3. Cohen's *h*

In case of categorical outcomes, Cohen's *d*, or Glass's Δ cannot be used. In such cases, difference between proportions is tested instead of means presented by Cohen (1988) and Rosenthal *et al.* (1994).

Suppose p_1 and p_2 represent two proportions. Cohen's *h* effect size measure is represented by

$$h = \varphi_1 - \varphi_2 \tag{8}$$

$$\text{where } \varphi_i = 2\arcsin\left(\sqrt{p_i}\right) \tag{9}.$$

The same convention is followed for Cohen's *h* effect size estimates as that for Cohen's *d* described above with respect to interpretation based on cutoff values as shown by Cohen (1988), Cohen (1992), Hedges *et al.* (2008) and Grissom and Kim (2005).

2.4.4. Odds ratio (OR)

OR is used to assess degree of association between binary outcomes and is interpreted as follows as reported by Chinn (2000). *OR* = 1.5 indicates weak association, *OR* = 2 indicates medium association and *OR* = 3 indicates strong association.

Consider the following 2x2 table.

Table 1: 2×2 Contingency table

	Event	
Exposure	**Yes**	**No**
Yes	*a*	*b*
No	*c*	*d*

$$Odds\ ratio\ (OR) = \frac{odds\ of\ the\ event\ in\ the\ exposed\ group}{odds\ of\ the\ event\ in\ the\ non-exposed\ group} \tag{10}$$

$$OR = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc} \tag{11}$$

2.5. Machine learning techniques

Multiple Linear Regression (MLR) is a powerful statistical technique commonly used in large data set analyses. MLR aims to model the relationship between a dependent variable and multiple independent variables by estimating the best-fitting regression equation as described by Ayyadevara (2018). In scenarios where data sets are large and complex, MLR serves as a valuable tool to identify and quantify the effects of multiple predictors on the outcome of interest. By incorporating multiple independent variables simultaneously, MLR allows researchers to understand the collective influence of various factors on the dependent variable, enabling them to uncover complex patterns and associations within the data. Although there are multiple MLR models, the choice of MLR model depends on the nature of the data, the research question, and the assumptions underlying the analysis as described by Ayyadevara (2018). In this work, we use ARM and the Naïve Bayes Classifier within Multiple Linear Regression (MLR) to provide valuable insights and enhance the interpretability and statistical analysis of big datasets.

2.5.1. Association rule mining (ARM)

To discover interesting relations between variables in large databases, ARM can be used as denoted by Ayyadevara (2018). ARM is a rule-based machine learning approach. The main concept in ARM is to discover rules that govern how certain sets of variables relate to each other. To find the degree of these relations, different measures such as lift (*L*), support (*S*) and confidence (*C*) can be used as described below.

In order to distinguish a trivial rule from a non-trivial rule, a measure used in ARM called the lift (*L*) can be calculated as follows as denoted by Geurts *et al.* (2003).

$$L = \frac{s(X \Rightarrow Y)}{s(X) \cdot s(Y)} \tag{12}$$

X is known as the antecedent of the rule and *Y* is known as the consequent. The numerator $s(X \Rightarrow Y)$ measures the observed frequency of the items in *X* and *Y* occurring together

and the denominator $s(X) \bullet s(Y)$ measures the expected frequency of the items in X and Y occurring together under the assumption of conditional independence as denoted by Geurts *et al.* (2003).

If L has a value greater than 1, we conclude that there is positive interdependence between X and Y. If the value of L is less than 1, we conclude that there is negative interdependence between X and Y. Lastly, if $L = 1$, X and Y are said to be conditionally independent. The greater the value of lift L, the stronger is the dependence between X and Y.

Two other important parameters for the ARM are the support (S) and confidence (C) of a rule by means of which the algorithm to produce a set of rules describing the underlying patterns in the data. Support of a rule indicates the frequency with which a rule occurs in a dataset and confidence measures the reliability of an association rule as indicated by Geurts *et al.* (2003). Suppose we are studying the association of different predictor variables with different surgery types for breast cancer.

Table 2: Interpretation of lift values

Outcome	Interpretation of lift (L)
L < 1	Negative interdependence between X and Y
L = 1	Conditional independence between X and Y
L > 1	Positive interdependence between X and Y

Then,

$$S\{X\} = \frac{number\ of\ patients\ receiving\ surgery\ type\ X}{total\ number\ of\ patients} \text{ for a rule } \{X \Rightarrow Y\} \quad (13)$$

$$C\{X \geq Y\} = \frac{number\ of\ patients\ receiving\ surgery\ type\ X\ in\ predictor\ variable\ category\ Y}{total\ number\ of\ patients\ receiving\ surgery\ type\ X} \quad (14)$$

2.5.2. Naïve Bayes classifier

Naïve Bayes classifier is a machine learning algorithm based on Bayes' theorem that follows a probabilistic approach for solving classification problems. In real-world scenarios, variables have some correlations and are not entirely independent. However, the algorithm is called 'Naïve' Bayes classifier because it assumes independence between predictor variables as described by Zhang (2016) and Ayyadevara (2018).

The equation for Bayes' theorem is given as

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)} \quad (15)$$

Here, $P(A \mid B)$: Conditional probability of an event A, given the event B,
$P(A)$: Probability of event A
$P(B)$: Probability of event B
$P(B|A)$: Conditional probability of an event B, given the event A

The equation (15) represents a case with a single predictor. However, in real-world scenarios, there are more than one predictor variables and for a classification problem, there are multiple output classes. Let us represent these classes as $C_1, C_2, \ldots, C_k$ and the predictor variables as $x_1, x_2, \ldots, x_n$.

The objective of a Naïve Bayes algorithm is to estimate the conditional probability that an event with a feature vector $x_1, x_2, \ldots, x_n$ belongs to a particular class C_i.

Given these conditions, the equation for Naïve Bayes' classifier can be written as follows.

$$P(C_i \mid x_1, x_2, \ldots, x_n) \propto \left(\prod_{j=1}^{n} P\left(x_j \mid C_i\right)\right) \cdot P(C_i) \text{ for } 1 < i < k \tag{16}$$

Two statistical measures, namely, misclassification and accuracy, can be calculated for the Naïve Bayes classifier, based on which the model performance can be evaluated. *Misclassification* is the percentage of times a classifier incorrectly classifies an item into a class or category. *Accuracy* is the percentage of times a classifier correctly classifies an item into a class or category.

Other classification techniques include the Random Forest method that combines multiple decision trees to improve accuracy and reduce overfitting as explained by Ayyadevara 2018. The Neural Networks classification approach uses deep learning models with multiple layers of interconnected nodes that could be used for complex classification tasks. Another classification technique is the K-Nearest Neighbors (KNN), a non-parametric method that assigns class labels based on the majority class of the k-nearest data points. While other classification methods like Random Forest, Neural Networks, and K-Nearest Neighbors also offer their respective advantages, we have chosen the Naïve Bayes Classifier for this study as shown by Zhang (2016) and Ayyadevara (2018). The decision to use this method is based on factors such as interpretability, computational efficiency, and the specific characteristics of the NCDB dataset and the research question.

3. Modification/Extension of Cohen's *h* for logistic regression

Consider the following notations for effect size calculation.

Table 3: Variables and notations for effect size calculations

		Outcome variable			
Variable 1	**Total**	**Category 1**	**Category 2**	**…**	**Category *n*'**
Category 1	n_1	n_{11}	n_{12}		$n_{1n'}$
Category 2	n_2	n_{21}	n_{22}		$n_{2n'}$
…					
Category *n*	n_n	n_{n1}	n_{n2}		$n_{nn'}$

Notations

- n_{ij}: Number of patients in category i of variable 1 and category j of *variable 2*; $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, n'$.
- n_i: Total number of patients in i^{th} category of variable 1.
- p_{ij}: Prevalence for category i of variable 1 and category j of *variable 2*. We calculate p_{ij} as
  $p_{ij} = n_{ij}/n_i$

Cohen's h effect size for the i^{th} category will be given by

$$h_i = \varphi_{i1} - \varphi_{i2};\ i = 1, 2 \tag{17}$$

$$\varphi_{ij} = 2\, logit\left(\sqrt{p_{ij}}\right);\ i = 1, 2;\ j = 1, 2 \tag{18}$$

These are defined for comparing two categories within two variables.

Here, instead of the traditional *arcsin* transformation used for Cohen's *h* shown by Catarino *et al.* (2011), we use *logit* transformation described by Collins *et al.* (1992). A *logit* transformation is more appropriate in the case of logistic regression.

For our case we will calculate h_1 and h_2 corresponding to the two groups that we are comparing using equation (17). Then the effect size *h* is given by

$$h = h_1 - h_2 \tag{19}$$

Here, we are comparing the effect sizes for one category of predictor variable with another category of predictor variable based on different categories of outcome variable. Since we are comparing the two categories of a predictor variable, a difference of differences is proposed. This difference, *h* defined using equations (17), (18) and (19) is the novel effect size measure which would help in determining the meaningfully significant differences.

The convention used for the interpretation of the effect sizes is described in the table below. For large sample sizes such as the NCDB database, effects show up quickly due to the large sample. Hence, the convention that we have suggested considers an effect of approximately 93% as a small effect, 99.3% as medium effect and 99.9% as large effect. We suggest using this convention owing to the large sample size and using the guidelines suggested by Cohen for determining small, medium and large effect sizes as detailed by Cohen (1988) and Cohen (1992).

Thus, in the case of modified Cohen's *h* – 1.5: small effect, 2.5: medium effect, 3: large effect (Table 4).

In this paper, using a case study from the National Cancer Database (NCDB), we have presented how the proposed novel effect size measure above can be utilized to help in obtaining meaningfully significant results. In addition, we have also validated the novel effect size measure using machine learning techniques.

Table 4: Convention for modified Cohen's *h*

Relative size	**Effect size**	**Difference between the comparison groups**
	0.0	50%
Small	1.5	93.3%
Medium	2.5	99.3%
Large	3	99.9%
	5.5	100%

The next section describes a brief literature review which is followed by the case study which demonstrates the statistical issues in the analysis of large databases, particularly expanding on *p*-values, and illustrates the use of the novel effect size measure.

4. Literature review

In this section, we provide a short literature review that comprises of 15 research articles that were carefully selected using the flowchart presented in Figure 1. Our main objective was

to gain insights into the prevailing statistical issues surrounding sample selection, missing data imputation techniques, commonly used statistical methods and inference measures used. Our search revealed that a total of 3,331 articles were published between 2004 and 2014 using the NCDB, out of which 257 were focused on female breast cancer. To maintain uniformity, since the case study presented in this paper is focused on the association of surgery types with different demographic predictor variables, we only included articles that dealt with 'mastectomy' and 'lumpectomy' surgeries. This led us to 22 articles, and after removing duplicates, we were left with 15 articles that were used for our literature review.

4.1. Article search protocol

In the current article, we have presented a case study to examine association of surgery types with different demographic predictor variables to demonstrate statistical issues while analyzing large databases using NCDB as an example. The three surgery types for this study included from the NCDB were ‘lumpectomy’, ‘mastectomy without reconstruction’ and ‘mastectomy with reconstruction’. Hence, we designed the literature to identify and demonstrate statistical issues in the analysis of large medical databases. We identified articles published using female breast cancer data from NCDB and we performed keyword search using PubMed, MEDLINE (Web of Science), and Embase databases. We used the following keywords to search relevant articles: ‘NCDB’, ‘National Cancer Database’, ‘Breast Cancer’, ‘surgery’, ‘mastectomy’, ‘lumpectomy’, and ‘female’ and narrowed down to 15 articles that were most relevant.

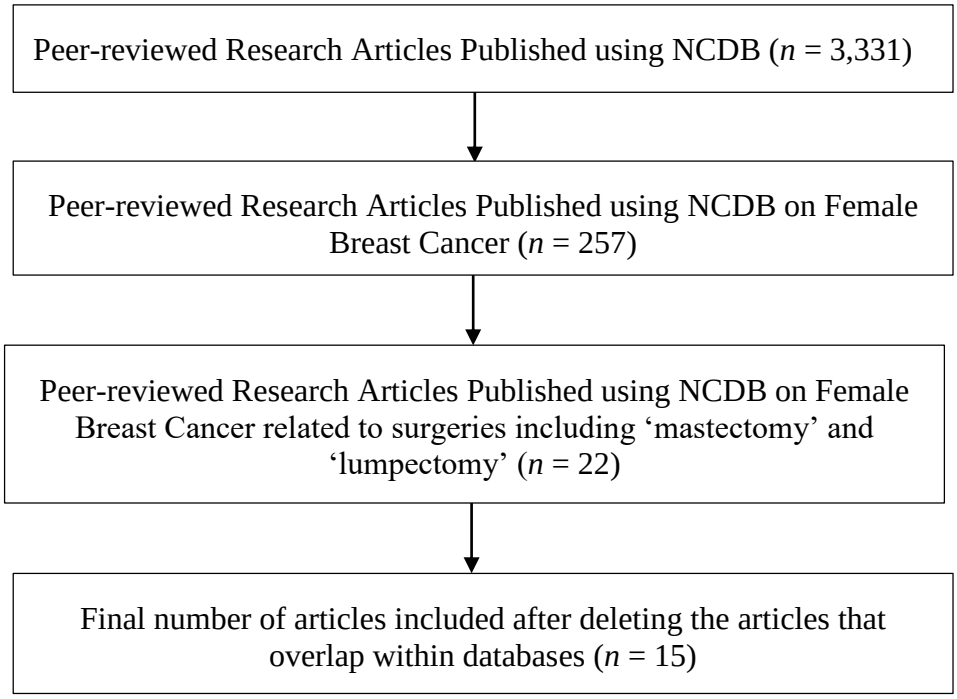

Figure 1: Schematic representation for selection of articles

4.2. Literature review results

Table 5 presents an overview of the research articles with respect to important statistical considerations.

Table 5: General overview of research articles

Article reference (Year)		Details of the article
Hotsinpiller *et al.* (2021)	Objective	Describe rates and predictors of positive margins for invasive breast cancers in the NCDB
	Sample Size	707,798
	Missing/Imputation	None
	Statistical Methods	Two-sided *t*-test; Chi-square test; Multivariable logistic regression
	Inference Measures	Odds ratios with 95% CI; *p*-values
Wrubel *et al.* (2021)	Objective	Compare BCT with mastectomy for treatment of early-stage breast cancer
	Sample Size	202,236
	Missing/Imputation	None
	Statistical Methods	Chi-square test; Kaplan-Meier analysis; Log-rank test
	Inference Measures	Kaplan-Meier survival curves; Overall survival (%); *p*-values
Weiser *et al.* (2021)	Objective	Identify sub-groups of node-positive patients with low to intermediate RS who still benefit from adjuvant chemotherapy
	Sample Size	28,591
	Missing/Imputation	None
	Statistical Methods	*t*-test; Chi-square test; Multivariable logistic regression; Kaplan-Meier method; Log-rank test; Multivariable Cox proportional hazards model
	Inference Measures	Hazard ratios with 95% CI; Odds ratios with 95% CI; Kaplan-Meier survival curves; *p*-values
Lehrberg *et al.* (2021)	Objective	Evaluate the outcomes and predictors for patients receiving BCS treatment outside of the standard NCCN guidelines, compared with patients receiving standard MRM treatment
	Sample Size	10,610
	Missing/Imputation	None
	Statistical Methods	*t*-test; Chi-square test; Cochran-Armitage trend test; Multivariate Cox proportional hazards model
	Inference Measures	Adjusted hazards ratios; *p*-values
Pratt *et al.* (2021)	Objective	Examine the association between the time interval from time of diagnosis to completion of all acute breast cancer treatment modalities (surgery, chemotherapy, and radiation therapy) and survival
	Sample Size	50,720
	Missing/Imputation	None

Article reference (Year)		Details of the article
	Statistical Methods	Univariate and multivariate Cox proportional hazards model; Log-rank test; Kaplan-Meier method; Chi-square test; Fisher's exact test; Two-sample *t*-test
	Inference Measures	Hazard ratios with 95% CI; Kaplan-Meier survival curves; 5-year survival (%); *p*-values
Lewis *et al.* (2019)	Objective	Determine the clinical characteristics, outcomes, and propensity for lymph node metastasis of patients with IMPC of the breast recorded in the NCDB
	Sample Size	2660
	Missing/Imputation	None
	Statistical Methods	Log-rank test; Cox proportional hazards model; Kaplan-Meier method
	Inference Measures	Hazard ratios with 95% CI; Kaplan-Meier survival curves; *p*-values
Mazor *et al.* (2019)	Objective	Assess patterns and outcomes of BCT for T3 tumors
	Sample Size	37,268
	Missing/Imputation	None
	Statistical Methods	Sensitivity analysis; Chi-square test; Wilcoxon rank sum test; Multivariable logistic regression; Cochran-Armitage trend test; Spearman's correlation; Kaplan-Meier method; Cox proportional hazards model
	Inference Measures	Odds ratios with 95% CI; Hazard ratios with 95% CI; Kaplan-Meier survival curves; *p*-values
Zhu *et al.* (2019)	Objectives	Study clinicopathological features, treatment patterns and prognosis of SCC; Investigate whether SCC (*vs.* IEDC) is associated with poor clinicopathological characteristics, different treatment patterns and worse survival; Perform exploratory analysis of the benefits of systematics therapy for SCC patients
	Sample Size	3,430
	Missing/Imputation	None
	Statistical Methods	Chi-square test; Kaplan-Meier analysis; Cox regression model
	Inference Measures	Hazard ratios with 95% CI; Kaplan-Meier survival curves; *p*-values
Landercasper *et al.* (2019)	Objective	Determine if there were differences in the OS of matched breast cancer patients undergoing lumpectomy *vs.* mastectomy in the NCDB
	Sample Size	845,136
	Missing/Imputation	None

Article reference (Year)		Details of the article
	Statistical Methods	Kaplan-Meier method; Propensity score matched analysis; Cox proportional hazards model; Subgroup analysis
	Inference Measures	Hazard ratios with 95% CI; Kaplan-Meier survival curves; *p*-values
McClelland *et al.* (2019)	Objective	To assess trends in patterns of care and clinical outcomes to manage localized breast angiosarcoma
	Sample Size	826
	Missing/Imputation	None
	Statistical Methods	Chi-square test; Cochran-Armitage trend test; Univariate and multivariate logistic regression analysis; Univariate and adjusted Cox models; Survival analysis
	Inference Measures	Hazard ratios with 95% CI; Kaplan-Meier survival curves; Forest plots; *p*-values
Mills *et al.* (2018)	Objective	Utilize data from NCDB to complete investigation of the prognostic importance of histology within TMBC
	Sample Size	89,220
	Missing/Imputation	None
	Statistical Methods	Kaplan-Meier method; Log-rank test; Multivariate Cox proportional hazards model
	Inference Measures	Hazard ratios with 95% CI; Kaplan-Meier survival curves; *p*-values
Chiba *et al.* (2017)	Objective	Evaluate national trends in NET use in relation to conduct of Z1031 trial and impact of NET on the rates of BCS
	Sample Size	77,272
	Missing/Imputation	None
	Statistical Methods	Cochran-Armitage trend test; Chi-square test; Two-sample *t*-test; Multivariable logistic regression
	Inference Measures	Odds ratios with 95% CI; *p*-values
Landercasper *et al.* (2017	Objective	Investigate whether the receipt of NAC is associated with fewer reoperations
	Sample Size	71,627
	Missing/Imputation	None
	Statistical Methods	Cochran-Armitage trend test; Chi-square test; Multivariable logistic regression; Propensity score matching
	Inference Measures	Odds ratios with 95% CI; Forest plots; *p*-values
Rusthoven *et al.* (2016)	Objective	Evaluate the impact of PMRT and RNI for women with clinically node positive breast cancer treated with NAC
	Sample Size	15,315
	Missing/Imputation	None

Article reference (Year)		Details of the article
	Statistical Methods	Kaplan-Meier method; Log-rank test; Multivariate Cox models; Propensity score matched analysis
	Inference Measures	Hazard ratios with 95% CI; Kaplan-Meier survival curves; Forest plot; *p*-values
Chen *et al.* (2015)	Objective	Compare efficacy of BCS with RT and mastectomy using NCDB.
	Sample Size	160,880
	Missing/Imputation	None
	Statistical Methods	Kaplan-Meier method; Cox regression; Propensity score analysis
	Inference Measures	Hazard ratios with 95% CI; Kaplan-Meier survival curves; *p*-values

BCS= breast conserving surgery, RT= radiotherapy, CI= Confidence interval, PMRT= Postmastectomy radiotherapy, RNI= Regional nodal irradiation, NAC = Neoadjuvant chemotherapy, NET = Neoadjuvant endocrine therapy, IMPC= invasive micropapillary carcinoma, BCT= Breast conservation therapy, SCC= Squamous cell carcinoma, IDC= Infiltrating ductal carcinoma, RS= recurrence score assay, OS=overall survival, NCCN= National comprehensive cancer network, MRM= Modified radical mastectomy, TMBC= Triple negative breast cancer

4.3. Observations from the literature review

After conducting a comprehensive literature review, we found that there is a commonality between the sample size selection techniques, missing value imputation methods, statistical methods, and inference measures used in published research studies. Table 5 provides a helpful summary of these various approaches, which we organized according to the study objective and statistical considerations. Our analysis of this information yielded some interesting findings, which we will now discuss in more detail in next sections.

4.3.1. Sample selection techniques

Based on our analysis, appropriate sample size selection is a crucial aspect of statistical research. We have found that most of the studies that we reviewed worked with large sample sizes. In fact, we observed that the largest sample size among the 15 studies was an impressive 707,798 as observed in Hotsinpiller *et al.* (2021) while the smallest sample size was just 826 as observed in McClelland *et al.* (2019).

Interestingly, we also found that none of the studies that we reviewed described any formal sample size calculation techniques used for sample selection. Instead, it appears that samples were selected primarily based on data availability and filtering based on the study objective. This means that convenience/purposive sampling was used, and analysis methods designed for randomized data were employed, which could lead to misleading results and conclusions.

4.3.2. Missing data imputation

It is important to note that missing data can be a common occurrence, especially in large datasets like the NCDB. If missing data is simply deleted without any imputation, it can lead to a biased sample with biased results. Unfortunately, many of the studies such as Landercasper *et al.* (2017), Landercasper *et al.* (2019), Lewis *et al.* (2019), Mazor *et al.* (2019), McClelland *et al.* (2019), Zhu and Chen (2019), Lehrberg *et al.* (2021), Weiser *et al.* (2021), and Wrubel

et al. (2021) included in the literature review did not describe or perform any missing data imputation. Instead, they chose to perform a complete case analysis. However, it is worth noting that some studies, like Rusthoven *et al.* (2016), conducted a sensitivity analysis before and after excluding unknown variable values and still obtained similar results. Others, like Chiba *et al.* (2017), reported missing values in their table for tumor characteristics but did not mention whether these values were imputed or deleted before analysis. Lastly, Landercasper *et al.* (2019) pointed out that all but one of the studies they cited did not include any missing data imputation.

4.3.3. Statistical analysis methods

The choice of analysis methods depends upon the study objective. However, every statistical technique involves certain assumptions and if these assumptions are not satisfied, the analysis may not result in reliable conclusions. This is a common issue with statistical analysis of the NCDB. For example, application of a common Cox proportional hazard model to non-randomized studies (case-control and databases) results in unreliable estimate of hazard ratio (relative risk) due to heterogeneity, time-varying exposure, corelated risk factors, and confounding *etc.* as presented by Moolgavkar *et al.* (2018).

From the studies included in the present literature review, we observed that the most common statistical analysis methods used were univariable as shown in Weiser *et al.* (2021) and multivariable logistic regression as used in Chiba *et al.* (2017), Landercasper *et al.* (2017), Mazor *et al.* (2019), McClelland *et al.* (2019), Hotsinpiller *et al.* (2021), and Weiser *et al.* (2021) and survival analysis as shown in Chen *et al.* (2015), Rusthoven *et al.* (2016), Mills *et al.* 2018, Landercasper *et al.* (2019), Lewis *et al.* (2019), Mazor *et al.* (2019), McClelland *et al.* (2019), Zhu and Chen (2019), Lehrberg *et al.* (2021), Pratt *et al.* (2021), Weiser *et al.* (2021), and Wrubel *et al.* (2021). Baseline characteristics are commonly compared using either a *t*-test (continuous variables) as indicated in Chiba *et al.* (2017), Hotsinpiller *et al.* (2021), Lehrberg *et al.* (2021), Pratt *et al.* (2021), and Weiser *et al.* (2021) or a chi-square test (categorical variables) as indicated in (Chiba *et al.* (2017), Landercasper *et al.* (2017), Mazor *et al.* (2019), McClelland *et al.* (2019), Zhu and Chen 2019, Hotsinpiller *et al.* (2021), Lehrberg *et al.* (2021), Pratt *et* al. (2021), Weiser *et al.* (2021), and Wrubel *et al.* (2021). Propensity score matching methods shown in Chen *et al*. (2015), Rusthoven *et al.* (2016), Landercasper *et al.* (2017), and Landercasper *et al.* (2019) and the Cochran-Armitage trend test as shown in Chiba *et al.* (2017), Landercasper *et al.* (2017), Mazor *et al.* (2019), McClelland *et al.* (2019), and Lehrberg *et al.* (2021) are also popular techniques for analyzing NCDB.

4.3.4. Inference measures

Different statistical analysis methods involve different inference measures based on which we draw conclusions. The most common inference measures in the cancer studies are hazard ratio, odds ratio, Kaplan-Meier survival curve, and the ubiquitous *p*-value which is used for making conclusions in most of the analysis procedures.

In the articles included in the literature review as well, *p*-values were the most common statistical inference measure were used in Chen *et al.* (2015), Rusthoven *et al.* (2016), Chiba *et al.* (2017), Landercasper *et al.* (2017), Mills *et al.* 2018, Landercasper *et al.* (2019), Lewis *et al.* (2019), Mazor *et al.* (2019), McClelland *et al.* (2019), Zhu and Chen (2019), Hotsinpiller *et al.* (2021), Lehrberg *et al.* (2021), Pratt *et al.* (2021), Weiser *et al.* (2021), and Wrubel *et al.* (2021). Odds ratios was used in Chiba et al. (2017), Landercasper *et al.* (2019), Mazor *et al.* (2019), Hotsinpiller *et al.* (2021), and Weiser *et al.* (2021), hazard ratios was used in Chen *et al.* (2015), Rusthoven *et al.* (2016), Mills *et al.* 2018, Landercasper *et al.* (2019), Lewis *et al.*

(2019), Mazor *et al.* (2019), McClelland *et al.* (2019), Zhu and Chen (2019), Lehrberg *et al.* (2021), Pratt *et al.* (2021), and Weiser *et al.* (2021), and their 95% confidence intervals, Kaplan-Meier survival curves was shown in Chen *et al.* (2015), Rusthoven *et al.* (2016), Mills *et al.* 2018, Landercasper *et al.* (2019), Lewis *et al.* (2019), Mazor *et al.* (2019), McClelland *et al.* (2019), Zhu and Chen (2019), Pratt *et al.* (2021), Weiser *et al.* (2021), and Wrubel *et al.* (2021) were also a common choice. Some studies such as Rusthoven *et al.* (2016), Landercasper *et al.* (2017), and McClelland *et al.* (2019) used forest plots.

4.3.5. Statistical issues with large databases – Summary

After conducting the literature review and carefully studying the published literature, we observed the following statistical issues with large databases.

Based on our comprehensive literature review and analysis of large databases, we have identified several design and statistical analysis issues that need to be addressed. Firstly, databases such as NCDB have extremely large sample sizes, which can create challenges in analyzing the data effectively. Secondly, data from large databases are not randomized, and therefore include close to all observed cases, making it difficult to control for bias. There is also a risk of duplicate records if a patient gets treated at multiple facilities and gets included in the databases that many times they were treated at different locations, which could significantly confound the results.

Another issue is that since databases such as NCDB are huge, the assumption of normality fails, which can impact the validity of parametric statistical analysis, most of which is built on the assumption of normality of data. Furthermore, due to extremely large sample size, analysis results in highly significant *p*-values, which may have no clinical relevance and could lead to false conclusions. Lastly, Simpson's paradox is a concern when using the entire database, as we may get highly significant results that might get reversed if we use a smaller sample selected using formal sample size selection techniques as reported by Hernán *et al.* (2011), and Pearl (2022).

5. Case study

5.1. Study design

Using the general study considerations outlined previously and based on the most common statistical analysis methods, we designed a case study using female breast cancer data from the NCDB collected between 2004 and 2014. This case study was designed and presented to demonstrate statistical issues related to analyzing large databases using NCDB as an example and suggesting alternative inference techniques that could describe real-world scenarios better than the current methods. Additionally, using the case study, we demonstrate the effectiveness of the novel modified Cohen's *h* effect size estimator. We also present category-wise comparisons which result in multiple *p*-values and show Bonferroni multiplicity adjustment fails to produce meaningful results as reported by Leon (2004). Note that the Bonferroni method is the most conservative method for adjusting for multiple comparisons.

The objective of the case study was to examine whether there is an association between surgery types and different demographic predictor variables. The dependent and independent variables and their levels/categories is given in Table 6.

The study used participants that satisfied certain inclusion and exclusion criteria based on the type of malignancy, diagnosis year, cancer stage, surgery type, *etc*. This outline has

been explained in detail in Figure 2. After choosing participants, we investigated the data for missingness. In the current design, we deleted the missing observations as the proportion of missing values was low (approximately 1%). Details about the analysis have been discussed in section.

As described in Table 6, the dependent variable used in the case study was surgery type and the independent variables were age, race, insurance status, facility type, stage (of cancer) and great circle distance (distance between the medical facility and patient's residence).

Figure 2 presents the flowchart for selection of study participants. We included nine primary tumor sites as shown in Figure 2. The initial sample size was 2,445,870 for subjects who had tumors detected at the given primary sites. We, then, included patients who had a single malignant primary tumor, invasive or microinvasive breast cancer behavior, were diagnosed between 2004 and 2014, experienced breast cancer stages I, II or III and were female subjects. Later, we excluded the patients who did not undergo any surgery or had surgeries other than lumpectomy, mastectomy without reconstruction or mastectomy with reconstruction. The final sample size was 1,158,387 after applying all the inclusion and exclusion criteria. This is a large sample size that fits the definition of big data and hence, analysis using traditional approaches poses issues that need a robust solution. We will demonstrate the statistical issues and the effectiveness of our novel modified Cohen's *h* effect size estimator in the following sections.

Table 6: Dependent and independent variables for the case study

	Variable	Levels/Categories
Dependent Variable	Surgery Type	1. Lumpectomy 2. Mastectomy without reconstruction 3. Mastectomy with reconstruction
Independent Variables	Age	1. < 40 years 2. 40 – 50 years 3. 50 – 65 years 4. > 65 years
	Race	1. White, 2. Black 3. Others
	Insurance status	1. Medicare + Other Govt. 2. Private 3. Medicaid 4. Not insured
	Facility type	1. Academic 2. Community Cancer Center 3. Comprehensive Community Center 4. Integrated Network
	Stage (of cancer)	1. Stage I 2. Stage II 3. Stage III
	Great Circle Distance (Distance from the medical facility and patient's residence)	1. < 50 miles 2. 50 – 150 miles 3. > 150 miles

5.2. Analysis plan

The study sample was divided into three groups based on the surgery types, and baseline characteristics. Values were tabulated for each of the three groups with respect to the different categories of predictor variables (see Table 7). We compared the baseline characteristics of the study participants using chi-square test, and the results are presented in Table 7.

To examine the association between surgery types and different predictor variables with multiple levels/categories for each, we applied multinomial multivariable logistic regression. We presented results along with odds ratios, their 95% confidence intervals (CI) and *p*-values in Table 9. In addition, we calculated the effect sizes using modified Cohen's *h* and presented the results for comparison in Table 10. A schematic of the study design has been presented in Figure 3.

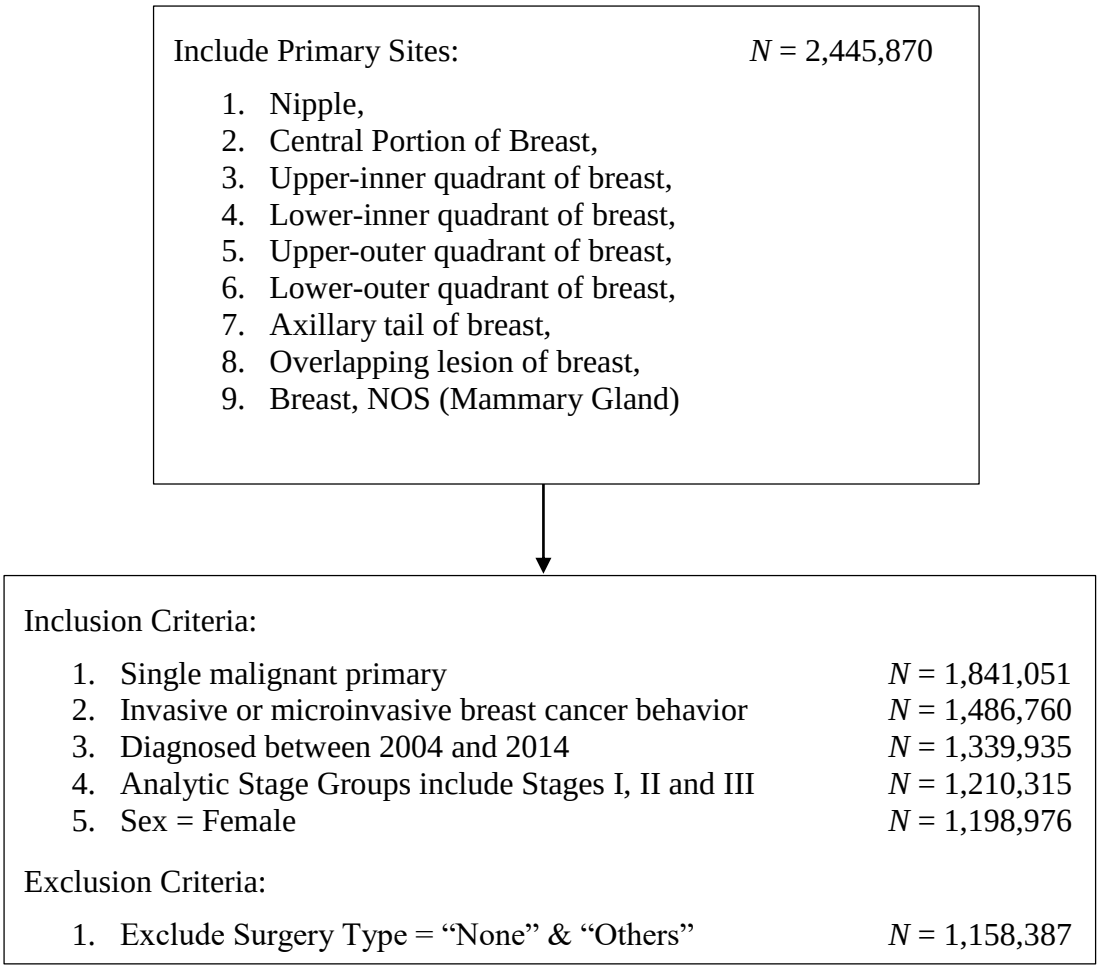

Figure 2: Study flowchart

Furthermore, we used ARM and NBC to understand the underlying associations and identify clinically relevant variables. Both ARM and NBC were used to validate the results obtained using our novel effect size estimator and to identify the important variables. This was especially important given the large and complex datasets we were working with.

5.3. Methods

A detailed protocol stating the study objectives, methods, analysis plan *etc*. was submitted for IRB approval in order to gain access to the data from NCDB for the present

research. The protocol was approved by the University of Louisville and James Graham Brown Cancer Center Cancer Committee which is a CoC Accredited Cancer Center. After receiving the IRB approval (IRB Number: 20.0049, we submitted access request to NCDB, which was also approved and access to NCDB was then granted to the authors of this manuscript. IRB approval request at the University of Cincinnati is pending. This study falls under IRB exemption since it is a retrospective study that looks at and utilizes NCDB data for analyses but does not involve working with human subjects.

5.4. Results

This section presents and describes the results obtained from analyzing our retrospective study. Table 7 shows the frequency distribution of breast cancer patients' three most common surgery types. For analyzing these data, we deleted the missing observations and compared Lumpectomy *vs.* Mastectomy without reconstruction and Mastectomy with reconstruction using the complete cases available. We present an overall *p*-value along with *p*-values for individual comparisons for each category of dependent variable. These *p*-values were then adjusted for multiple comparison using Bonferroni's test and the results are presented in Table 8. Majority of the studies present an overall *p*-value without looking at individual level comparisons. Since logistic regression compares difference between surgery types within categories of predictor variables, it is useful to present individual level *p*-values to help direct comparison of results.

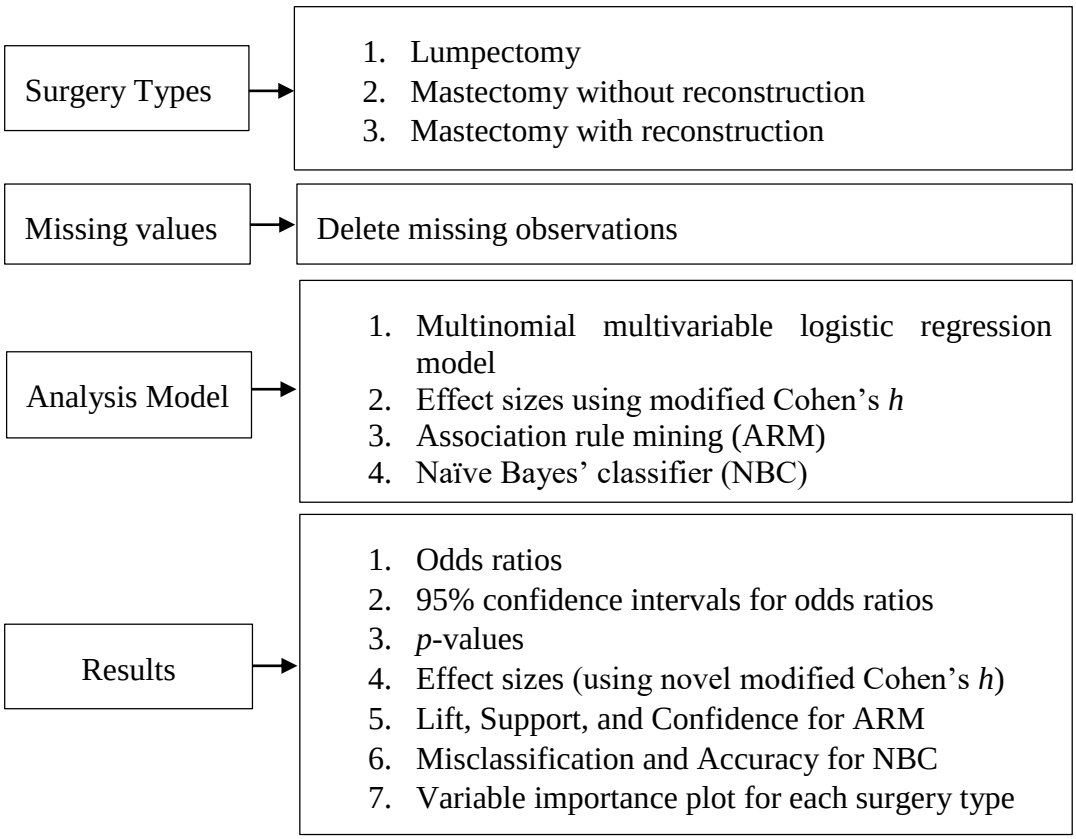

Figure 3: Study design

5.4.1. Baseline characteristics

Table 7 presents the baseline characteristics (lumpectomy (LT) *vs.* mastectomy without reconstruction (MTnR)/mastectomy with reconstruction (MTR)). It can be observed that *p*-values for all the variables shown in Table 7 are $<10^{-5}$ due to the large sample size and such a situation where almost all comparisons appear to be statistically significant, regardless of their practical importance or real-world significance. Therefore, we provide an alternative approach as presented in Table 8.

Table 8 presents the raw *p*-values in addition to Bonferroni-adjusted *p*-values to aid in comparing logistic regression results with baseline characteristics results.

Table 7: Baseline characteristics (lumpectomy (LT) *vs.* mastectomy without reconstruction (MTnR)/mastectomy with reconstruction (MTR))

Variable		**Surgery type**			***p*-value**
	Total (*N* = 1157322)	**LT (*N* = 692564)** (59.84%)	**MTnR (*N* = 325104)** (28.09%)	**MTR (*N* = 139654)** (12.07%)	
	Frequency (%)				
Age					$< 10^{-5}$
< 40 years	65404 (5.65)	23133 (3.34)	21327 (6.56)	20944 (15.00)	--
40 – 50 years	217557 (18.8)	110796 (15.99)	57470 (17.68)	49291 (35.30)	Ref
50 – 65 years	483166 (41.75)	302250 (43.64)	123114 (37.87)	57802 (41.39)	$9.9x10^{-5}$
> 65 years	391195 (33.80)	256385 (37.02)	123193 (37.89)	11617 (8.32)	$9.9x10^{-5}$
Race					$< 10^{-5}$
White	971210 (83.92)	587089 (84.77)	265520 (81.67)	118601 (84.92)	Ref
Black	124883 (10.79)	71636 (10.34)	40003 (12.30)	13244 (9.48)	$9.9x10^{-5}$
Others	48959 (4.23)	26181 (3.78)	16482 (5.07)	6296 (4.51)	$9.9x10^{-5}$
Insurance Status					$< 10^{-5}$
Medicare + Other Govt	399491 (34.52)	256126 (36.98)	127127 (39.10)	16238 (11.63)	Ref
Private	643302 (55.59)	378917 (54.71)	152246 (46.83)	112139 (80.30)	$9.9x10^{-5}$
Medicaid	70271 (6.07)	35047 (5.06)	27374 (8.42)	7850 (5.62)	$9.9x10^{-5}$
Not Insured	23876 (2.06)	11782 (1.70)	10169 (3.13)	1925 (1.38)	$9.9x10^{-5}$
Facility Type					$< 10^{-5}$
Academic	321773 (27.80)	196015 (28.30)	82863 (25.49)	42895 (30.72)	Ref

Variable	Total (*N* = 1157322)	Surgery type			*p*-value
		LT (*N* = 692564) (59.84%)	MTnR (*N* = 325104) (28.09%)	MTR (*N* = 139654) (12.07%)	
Community Cancer	116406 (10.06)	72115 (10.41)	37621 (11.57)	6670 (4.78)	$9.9x10^{-5}$
Comprehensive Community	532529 (46.01)	327628 (47.31)	151580 (46.63)	53321 (38.18)	$9.9x10^{-5}$
Integrated Network	121210 (10.47)	73673 (10.64)	31713 (9.75)	15824 (11.33)	0.004
Analytic Stage					$< 10^{-5}$
I	620662 (53.63)	456319 (65.89)	101909 (31.35)	62434 (44.71)	Ref
II	399934 (34.56)	204739 (29.56)	137980 (42.44)	57215 (40.97)	$9.9x10^{-5}$
III	136726 (11.81)	31506 (4.55)	85215 (26.21)	20005 (14.32)	$9.9x10^{-5}$
Great Circle Distance					$< 10^{-5}$
< 50 miles	1069529 (92.41)	646664 (93.37)	296028 (91.06)	126837 (90.82)	Ref
50 – 150 miles	61414 (5.31)	31426 (4.54)	20605 (6.34)	9383 (6.72)	$9.9x10^{-5}$
> 150 miles	16632 (1.44)	8656 (1.25)	5375 (1.65)	2601 (1.86)	$9.9x10^{-5}$

Table 8: Raw and Bonferroni-adjusted *p*-values for baseline characteristics

Variable	Raw *p*-value	Bonferroni adjusted *p*-value
Age		
40 – 50 years	Ref	
50 – 65 years	$9.9x10^{-5}$	$20x10^{-5}$
> 65 years	$9.9x10^{-5}$	$20x10^{-5}$
Race		
White	Ref	
Black	$9.9x10^{-5}$	$20x10^{-5}$
Others	$9.9x10^{-5}$	$20x10^{-5}$
Insurance Status		
Medicare + Other Govt	Ref	
Private	$9.9x10^{-5}$	$30x10^{-5}$
Medicaid	$9.9x10^{-5}$	$30x10^{-5}$
Not Insured	$9.9x10^{-5}$	$30x10^{-5}$
Facility Type		
Academic	Ref	
Community Cancer	$9.9x10^{-5}$	$30x10^{-5}$
Comprehensive Community	$9.9x10^{-5}$	$30x10^{-5}$
Integrated Network	0.004	0.012
Analytic Stage		

Variable	Raw *p*-value	Bonferroni adjusted *p*-value
I	Ref	
II	$9.9x10^{-5}$	$20x10^{-5}$
III	$9.9x10^{-5}$	$20x10^{-5}$
Great Circle Distance		
< 50 miles	Ref	
50 – 150 miles	$9.9x10^{-5}$	$20x10^{-5}$
> 150 miles	$9.9x10^{-5}$	$20x10^{-5}$

The variables that have three categories and two comparisons (age, race, analytical stage, and great circle distance) will be tested at a significance level of 0.05/2 = 0.025. The variables that have four categories and three comparisons (insurance status and facility type) will be tested at a significance level of 0.05/3 = 0.017.

5.4.2. Logistic regression: lumpectomy *vs.* mastectomy without reconstruction and mastectomy with reconstruction

Table 9 presents the results of multinomial multivariable logistic regression using surgery types as the dependent variable and same predictor variables as those presented in the baseline characteristics table.

Table 9: Multinomial multivariable logistic regression results

	Multiple logistic regression		
Predictors	**Odds ratio**	**95% CI**	***p*-value**
	LT *vs.* MTnR		
Age			
40 – 50 years (Ref)	1	NA	NA
50 – 65 years	0.85	(0.85, 0.87)	$< 10^{-5}$
> 65 years	0.97	(0.95, 0.99)	0.0007
Race			
White (Ref)	1	NA	NA
Black	1.05	(1.03, 1.07)	$< 10^{-5}$
Others	1.41	(1.38, 1.45)	$< 10^{-5}$
Insurance status			
Medicare + Other Govt (Ref)	1	NA	NA
Private	0.75	(0.74, 0.76)	$< 10^{-5}$
Medicaid	1.17	(1.14, 1.19)	$< 10^{-5}$
Not Insured	1.31	(1.27, 1.36)	$< 10^{-5}$
Facility Type			
Academic (Ref)	1	NA	NA
Community Cancer	1.29	(1.27, 1.31)	$< 10^{-5}$
Comprehensive Community	1.19	(1.17, 1.20)	$< 10^{-5}$
Integrated Network	1.09	(1.07, 1.11)	$< 10^{-5}$
Stage			
I (Ref)	1	NA	NA
II	3.05	(3.02, 3.08)	$< 10^{-5}$
III	12.37	(12.18, 12.57)	$< 10^{-5}$
Great Circle Distance			
< 50 miles (Ref)	1	NA	NA

	Multiple logistic regression		
Predictors	**Odds ratio**	**95% CI**	***p*-value**
50 – 150 miles	1.48	(1.46, 1.52)	$< 10^{-5}$
> 150 miles	1.22	(1.18, 1.28)	$< 10^{-5}$
LT *vs.* MTR			
Age			
40 – 50 years (Ref)	1	NA	NA
50 – 65 years	0.45	(0.44, 0.46)	$< 10^{-5}$
> 65 years	0.15	(0.14, 0.152)	$< 10^{-5}$
Race			
White (Ref)	1	NA	NA
Black	0.73	(0.71, 0.75)	$< 10^{-5}$
Others	0.92	(0.89, 0.95)	$< 10^{-5}$
Insurance			
Medicare + Other Govt (Ref)	1	NA	NA
Private	1.49	(1.45, 1.52)	$< 10^{-5}$
Medicaid	0.94	(0.90, 0.97)	0.0005
Not Insured	0.67	(0.63, 0.71)	$< 10^{-5}$
Facility Type			
Academic (Ref)	NA	NA	NA
Community Cancer	0.48	(0.47, 0.49)	$< 10^{-5}$
Comprehensive Community	0.80	(0.78, 0.81)	$< 10^{-5}$
Integrated Network	1.04	(1.02, 1.07)	0.0001
Stage			
I	NA	NA	NA
II	1.83	(1.80, 1.85)	$< 10^{-5}$
III	4.07	(3.98, 4.16)	$< 10^{-5}$
Great Circle Distance			
< 50 miles (Ref)	NA	NA	NA
50 - 150 miles	1.43	(1.39, 1.47)	$< 10^{-5}$
> 150 miles	1.43	(1.36, 1.50)	$< 10^{-5}$

5.4.3. Effect size

Table 10 presents the effect sizes using modified Cohen's *h* estimator. These results will be compared with the *p*-values obtained using logistic regression to identify the important associated variables for surgery types that the two procedures predict.

5.4.4. Association rule mining

Table 11 presented below shows the results obtained using the ARM procedure. Here, we present lift, support, and confidence for each of the associations obtained using this procedure. The ARM procedure was run to check associations between the surgery types and all the predictor variables. Conclusions were mainly drawn using lift values and support and confidence values were presented to demonstrate the strength of the lift values.

Table 10: Effect sizes using modified Cohen's *h*

Variable	Effect size using modified Cohen's *h* procedure	
	LT *vs.* MTnR	**LT *vs.* MTR**
Age		
40 – 50 years	Ref	
50 – 65 years	-0.91	-1.92
> 65 years	-0.69	-4.01
Race		
White	Ref	
Black	0.57	0.01
Others	0.96	0.58
Insurance Status		
Medicare + Other Govt	Ref	
Private	-0.23	2.48
Medicaid	1.53	2.40
Not Insured	1.81	1.97
Facility Type		
Academic	Ref	
Community Cancer	0.41	-1.28
Comprehensive Community	0.16	-0.48
Integrated Network	0.04	-0.02
Analytic Stage		
I	Ref	
II	3.22	2.28
III	7.16	4.32
Great Circle Distance		
< 50 miles	Ref	
50 – 150 miles	1.09	1.06
> 150 miles	0.94	1.04

Table 11: Association rule mining – testing association between surgery types and predictor variables lift (*L*), support (*S*), and confidence (*C*)

Association	*L*	*S*	*C*
Surgery types and Age			
Lumpectomy and age group '> 65 years'	1.09	0.24	0.66
Lumpectomy and age group '50 – 65 years'	1.04	0.24	0.62
Lumpectomy and age group '40 – 50 years'	0.85	0.10	0.51
Surgery types and Race			
Mastectomy with Reconstruction and White	1.01	0.10	0.86
Lumpectomy and White	1.01	0.51	0.86
Mastectomy without Reconstruction and White	0.97	0.23	0.82
Surgery types and Insurance status			
Mastectomy with Reconstruction and Private	1.43	0.10	0.81
Lumpectomy and Medicare + Other Govt.	1.07	0.22	0.64
Lumpectomy and Private	0.98	0.33	0.59
Surgery types and Facility type			
Lumpectomy and Comprehensive Community Cancer Program	1.03	0.28	0.62
Lumpectomy and Academic/Research Program	1.02	0.17	0.61

Association	*L*	*S*	*C*
Lumpectomy and Integrated Network Cancer Program	1.02	0.06	0.10
Lumpectomy and Community Cancer Program	0.88	0.08	0.52
Surgery types and Stage of cancer			
Mastectomy with Reconstruction and Stage III	2.22	0.07	0.62
Lumpectomy with Stage I	1.23	0.39	0.74
Lumpectomy and Stage II	0.86	0.18	0.51
Surgery types and Great circle distance			
Lumpectomy and < 50 miles	1.01	0.56	0.94
Lumpectomy and > 150 miles	0.87	0.01	0.52
Lumpectomy and 50 – 150 miles	0.86	0.03	0.51

5.4.5. Naïve Bayes classifier

To validate the results obtained using modified effect sizes, we used Naïve Bayes classifier. To identify important variables using Naïve Bayes classifier, the measures used were misclassification and accuracy as described in Section 2.5.2. The variable with the least misclassification and highest accuracy was concluded to have highest association with the dependent variable, surgery type. Furthermore, for the Naïve Bayes classifier, we plotted a variable importance plot by surgery type. The variable importance plot is presented in Figure 3. The variables highly associated with each surgery type is presented in the plot in the order of the strength of association. Each variable has an associated bar indicating the magnitude of its importance.

Table 12: Naïve Bayes classifier – misclassification percentages along with accuracy

Association of surgery types with	Misclassification	Accuracy
Age	40.3%	59.7%
Race	40.9%	59.1%
Insurance Status	40.3%	59.7%
Facility Type	40.04%	59.96%
Stage	35.4%	64.6%
Great Circle Distance	41.3%	58.7%

5.5. Interpretation

5.5.1. Baseline characteristics

Our selected study population had approximately 60% of the patients who received lumpectomy, 28% received mastectomy without reconstruction, 12% received mastectomy with reconstruction, and for around 1% of the subjects' surgery information was missing.

Table 7 shows that the age group of 50 – 65 years receives the maximum proportion of breast cancer surgery. Among this group, lumpectomy is the most common surgery type, followed by mastectomy with reconstruction. The study also revealed that 84% of the patients were white, and 11% were black. The most common surgery type for white patients was lumpectomy and mastectomy without reconstruction, and for black patients, it was also lumpectomy and mastectomy without reconstruction. Private insurance was the preferred choice of the majority of the patients, and within this group, mastectomy with reconstruction was the most common surgery type. The study also found that about 46% of the patients

received treatment at a comprehensive community center, and the maximum proportion of patients received lumpectomy. Among the patients with Stage I and Stage II breast cancer, 74% received lumpectomy and 51%, respectively, while most of the patients with Stage III received mastectomy without reconstruction (62%). The study also revealed that 92% of the patients lived within 50 miles of the facility where they received treatment, and within this group, 60% received lumpectomy followed by mastectomy without reconstruction (28%).

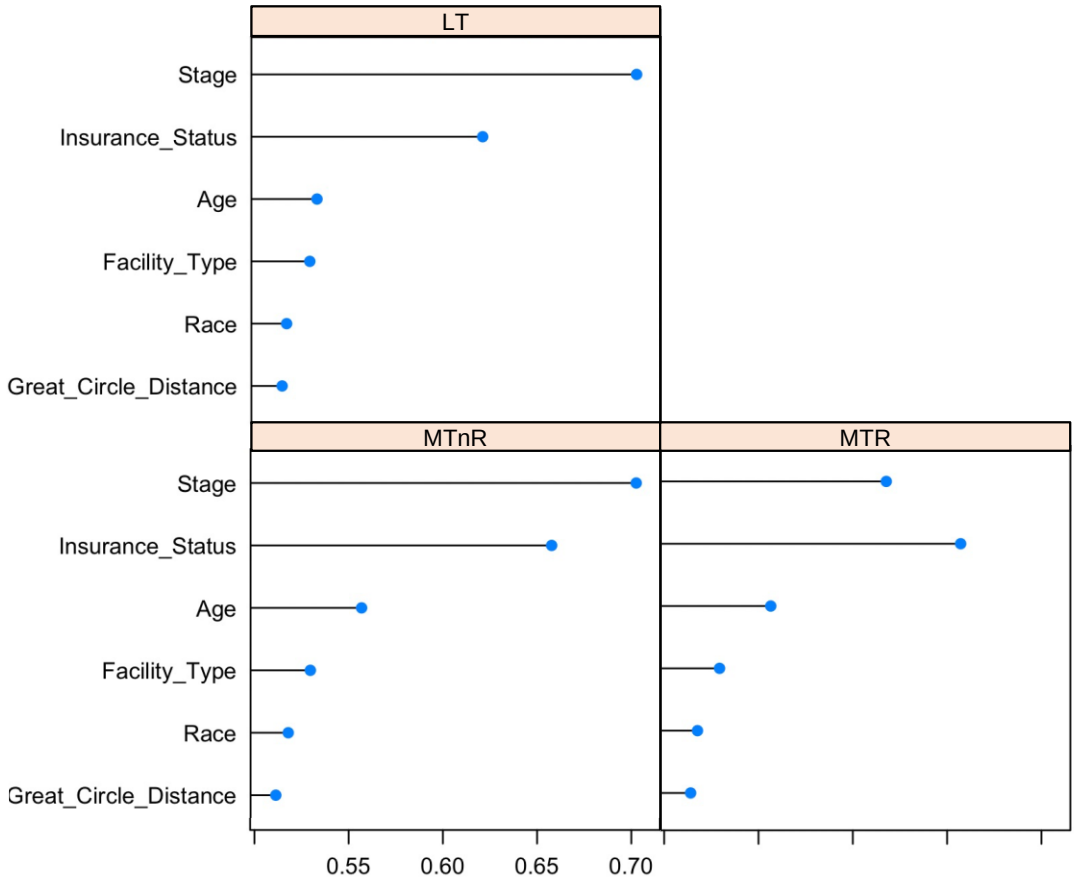

Figure 4: Variable importance plot

The chi-square tests run on the baseline characteristics reveal a highly significant association between the surgery types and all predictors, i.e., age, race, insurance status, facility type, analytic stage, and great circle distance. When we conducted pairwise comparisons using one of the categories as the reference category for each of the predictor variables, we obtained highly significant *p*-values for all the comparisons. Table 8 presents Bonferroni adjusted *p*-values in addition to the raw *p*-values. The Bonferroni adjusted *p*-values are highly significant for all variables as well. From tables 7 and 8, we conclude that surgery type has a highly significant association with all the predictor variables included in the study.

5.5.2. Logistic regression – *p*-values and odds ratios

It is interesting to note that Table 9 displays highly significant *p*-values ($< 10^{-5}$) for all the variables. However, the odds ratios tell a different story. According to Sullivan and Feinn (2012) an odds ratio around 1.5 indicates a small difference, around 2 indicates a medium difference, and around 3 indicates a significant difference. Except for the stage of cancer, none

of the other variables have an odds ratio of >1.5, and thus, based on *p*-values, it can be concluded that stage is the only variable with a significant difference.

It should be noted that the statistical fallacy in this case is that the p-values are highly significant, with an odds ratio close to 1. This is demonstrated by the odds ratio of 0.97 and a *p*-value of 0.0007 for age group > 65 years *vs*. 40 – 50 years when comparing LT *vs.* MTnR. From the odds ratio, which is very close to 1, we would conclude that the age groups > 65 years and 40 – 50 years does not significantly differ from each other in terms of the surgery types. The 95% CI for this case is (0.95, 0.99) which almost cuts through the value of 1, thus hinting non-significant difference. This is a classic example of how *p*-values can be misleading, and concluding significant differences based on *p*-values alone would lead to misleading conclusions.

5.5.3. Effect sizes

In Table 10, the effect sizes calculated using modified Cohen's *h* formula are presented. To interpret the effect sizes, we use the cutoffs presented in Table 4 suggested by the authors of this paper. If the effect size is close to 1.5, it indicates a small effect, 2.5 indicates a medium effect, and 3 indicates a large effect. For a significant difference, we require a large effect size as it indicates a large underlying difference.

From Table 10, we can see that for LT *vs*. MTnR, we obtained an effect size of 3.22 for Stage II *vs*. Stage I and an effect size of 7.16 for Stage III *vs*. Stage I using the novel modified Cohen's *h* effect size. This means that patients with Stage II breast cancer have a significantly higher chance of receiving MTnR than LT when compared with patients who have Stage I breast cancer. A similar interpretation applies to Stage III *vs*. Stage I. For all other comparisons, we do not obtain a high effect size and thus, we may conclude that except stage of cancer, the other predictors do not have a significant association with surgery type.

When comparing LT *vs*. MTR, a large effect size of -4.01 was obtained for age group > 65 years *vs*. 40 – 50 years. This indicates that the age group > 65 years has a significantly lower chance of receiving LT *vs*. MTR when compared with patients in the age group 40 – 50 years.

Similarly, we obtained a large effect size of 4.32 for comparing stage III *vs.* stage I breast cancer for LT *vs.* MTR. This also indicates that except age group and stage of cancer, none of the other predictors show a highly significant association with surgery types.

Based on odds ratios and effect sizes, we find only one significant association, *i.e*., between surgery types and stage of cancer. Even though the *p*-values are highly significant for all comparisons, odds ratios and effect sizes do not support this result. Thus, using effect sizes in addition to *p*-values when analyzing large datasets may be a more statistically sound approach.

5.5.4. Association rule mining

We observe from Table 11 that the values for lift are either very close to 1 or just below 1 for most associations except for the variable 'stage'. Thus, we may say that the predictor variables age, race, insurance status, facility type, and great circle distance have a weak correlation with surgery types. There is, however, a strong association between stage of cancer and surgery types. From the lift value of 2.22, we may conclude that patients suffering from stage III breast cancer have a strong possibility of receiving mastectomy with reconstruction.

Also, patients suffering from stage I breast cancer have a slightly positive correlation with receiving lumpectomy which agrees with our conclusions from Table 7. The lift value for Mastectomy with reconstruction and private is 1.43 which is higher than 1. This implies a positive correlation between the two values with a confidence of 0.81.

5.5.5. Naïve Bayes classifier

Table 12 presents the results obtained using the Naïve Bayes classifier. The misclassification proportion is 35% for stage of cancer which is the lowest and the corresponding accuracy is 65% which is the highest. This indicates and supports all the previous results and arguments that the only significantly associated variable with surgery types is the stage of cancer.

From the variable importance plot, we see that stage is the most important variable associated with the surgery types LT and MTnR and the second most important variable associated for surgery type MTR.

From all the above results, we see that *p*-values indicate highly significant associations for surgery types with all the predictors. However, this statistical significance is not clinically relevant as indicated by the odds ratios that are close to 1. The modified effect sizes indicate a highly significant association between stage and surgery types and one of the age groups for LT *vs.* MTR comparison. The significant association between stage of cancer and surgery type is supported by ARM and NBC results as well.

6. Conclusion

From our analysis, we observed that *p*-values alone can lead to misleading conclusions since they are very sensitive to sample sizes. As sample size increases, *p*-values tend to decrease and produce highly significant but clinically irrelevant results. Thus, an alternative to *p*-values when analyzing extremely large datasets is crucial. For this purpose, we explored effect sizes. However, to the best of our knowledge, effect size measures have not been suggested for the case of a logistic regression when we are comparing effects of two treatments within two different categories of a variable. To handle such a situation, we suggested an extension to Cohen's *h* effect size measure and demonstrated its use with the help of a case study using NCDB as an example. We proved its utility using machine learning tools such as ARM and NBC. We suggest using this modified version of Cohen's *h* for large databases when using logistic regression and comparing multiple treatments across different categories of predictor variables.

7. Discussion

The study aimed to address critical gaps in the existing literature related to the analysis of large electronic health record (EHR) databases, sample selection methods for such databases, and the over-reliance on *p*-values for drawing clinical inferences. By doing so, it sought to provide valuable insights into the limitations of *p*-values in the context of large sample sizes and propose a novel effect size measure tailored for logistic regression. Furthermore, the study aimed to validate the effectiveness of the proposed effect size measure using machine learning techniques.

The first research question focused on the clinical relevance of *p*-values when dealing with sampling from large databases. It is well-known that large sample sizes can lead to statistically significant *p*-values, even when the observed effect sizes are trivial or lack clinical importance. The study's objective was to demonstrate this phenomenon when analyzing large medical databases, which is critical in guiding researchers to avoid misinterpreting significant *p*-values as clinically meaningful results. By highlighting the limitations of relying solely on *p*-values, the study encourages researchers to adopt a more comprehensive approach that includes effect sizes for a more nuanced interpretation of results. For large databases, in general, even if the large number of comparisons are subjected to the most stringent procedure of controlling Type I error *i.e*., Bonferroni adjustment, let alone less stringent procedures like Holm, Hochberg, Hommel, and Benjamini-Hochberg, will result in highly significant *p*-values.

The second research question sought to explore the clinical relevance of effect sizes compared to *p*-values. Effect sizes provide a quantitative measure of the magnitude of an observed effect, indicating the practical importance of a finding. The study recognized the value of effect sizes in determining clinical significance, especially when dealing with large samples. By proposing a novel effect size measure specifically designed for logistic regression, the study aimed to overcome the limitations of *p*-values and offer a more meaningful and informative measure for interpreting results.

The study's primary objective was to propose and validate a novel effect size measure for logistic regression using machine learning techniques. Machine learning methods, specifically, association rule mining and Naïve Bayes classifier served as complementary tools to corroborate the findings obtained from the effect size measure. The validation process aimed to strengthen the credibility of the proposed measure and ensure its applicability in real-world scenarios.

The study's findings shed light on the importance of considering both *p*-values and effect sizes in data analysis. It emphasized that large sample sizes can lead to significant *p*-values without necessarily indicating clinical relevance. The proposed novel effect size measure offered a valuable alternative for assessing practical significance, particularly in logistic regression models. The validation through machine learning techniques provided additional support for the effectiveness and reliability of the novel effect size measure.

7.1. Baseline characteristics

The baseline characteristics of the selected study population provided valuable insights into the distribution of breast cancer surgery types and the patient demographics. The majority of patients (approximately 60%) underwent lumpectomy, followed by 28% who received mastectomy without reconstruction and 12% who underwent mastectomy with reconstruction. While the proportion of missing surgery information was minimal (around 1%), it is essential for future studies to address and minimize missing data to ensure the completeness and accuracy of the analysis.

One of the key findings of the study was the prominence of the age group between 50 and 65 years, as it received the highest proportion of breast cancer surgeries. Within this age group, lumpectomy was the most common surgery type, followed by mastectomy with reconstruction. This observation aligns with the current clinical guidelines, which often recommend lumpectomy as a preferred option for early-stage breast cancer in older patients due to its less invasive nature and potential for better cosmetic outcomes Pusic *et al.* (1999).

It is essential to recognize the limitations of the study, including potential selection bias and generalizability. The study population may not be fully representative of the broader breast cancer patient population, particularly in settings with different healthcare systems or demographics. Researchers should consider such factors when interpreting and applying the study's findings to diverse patient populations.

7.2. Logistics regression – odds ratio and *p*-value

The findings from the logistic regression analysis revealed highly significant *p*-values ($< 10^{-5}$) for all the variables under investigation. Such high significance levels might lead one to believe that all predictor variables have a strong impact on the outcome (surgery types). However, a closer examination of the odds ratios indicated that, except for the stage of cancer, none of the other variables demonstrated odds ratios greater than 1.5. This observation suggests that most of the predictor variables might not have a substantial effect on the choice of surgery types, except for the stage variable, which appears to be significantly associated with surgery types.

One notable concern arising from the analysis is the occurrence of a statistical fallacy when highly significant *p*-values are accompanied by odds ratios close to 1. This implies that, despite the statistical significance, the observed effect sizes might be minimal or practically negligible. In the context of the study, this phenomenon is particularly evident in the comparison between age groups and surgery types received. Although the *p*-values suggest a significant association, the odds ratios close to 1 indicate that age groups may not play a substantial role in determining the choice of surgery.

Relying solely on *p*-values to draw conclusions can be misleading, as highlighted by the findings. Focusing solely on the significance levels without considering the effect sizes might lead to incorrect interpretations of significant differences. It is essential to consider both the statistical significance and the practical significance (effect sizes) of the predictor variables to gain a comprehensive understanding of their impact on the outcome.

To avoid this statistical fallacy and ensure a more meaningful interpretation of the results, researchers should adopt a more holistic approach that considers both *p*-values and effect sizes. By considering the magnitude and direction of the effect sizes, researchers can better understand the clinical relevance of the predictor variables in relation to the outcome of interest.

Furthermore, the study underscores the importance of interpreting logistic regression results in the context of the research question and the clinical significance of the variables under investigation. While highly significant *p*-values are essential in identifying potential associations, they should not be the sole basis for decision-making or drawing conclusions. Instead, researchers should use them as a starting point to explore effect sizes and consider the practical implications of the findings.

7.3. Effect sizes

The analysis of effect sizes provided valuable insights into the magnitude and clinical relevance of the associations between predictor variables and surgery types in breast cancer patients. The effect sizes demonstrated that patients with Stage II and Stage III breast cancer had significantly higher chances of receiving mastectomy without reconstruction (MTnR)

compared to lumpectomy (LT) when compared to patients with Stage I breast cancer. The large effect sizes of 3.22 and 7.16 for Stage II and Stage III, respectively, indicated a substantial impact of cancer stage on the choice of surgery type. These findings align with clinical practice, as more advanced stages of cancer often require more extensive surgical interventions.

However, apart from the stage of cancer, the analysis of effect sizes revealed that other predictor variables did not show a significant association with surgery types. Effect sizes not being high for these comparisons indicated that variables such as age group, race, insurance status, facility type, analytic stage, and great circle distance might not have a considerable impact on the choice of surgery type. It is crucial to consider these results when making treatment decisions and designing interventions, as they highlight the relative importance of different predictors in guiding surgical decisions for breast cancer patients.

An interesting finding emerged when comparing lumpectomy (LT) *vs.* mastectomy with reconstruction (MTR) within different age groups. The effect size of -4.01 for the age group > 65 years indicated a significantly lower chance of receiving lumpectomy compared to patients in the age group 40-50 years. This observation suggests that age plays a critical role in determining the choice of surgical treatment, and older patients are more likely to undergo mastectomy with reconstruction. These insights can help inform patient counseling and shared decision-making, enabling healthcare providers to better tailor treatment plans based on age-related preferences and concerns.

Moreover, the discussion emphasizes the added value of incorporating effect sizes alongside *p*-values in analyzing large datasets. While *p*-values indicate statistical significance, they might not fully convey the practical relevance of the findings. In contrast, effect sizes provide a quantitative measure of the strength of the associations, allowing researchers to assess the clinical significance of the predictor variables. The finding that only the stage of cancer showed significant associations based on both odds ratios and effect sizes suggests that effect sizes serve as a more robust tool for identifying clinically relevant relationships.

Using effect sizes in conjunction with *p*-values in the analysis of large datasets is suggested as a more statistically sound approach. By combining these measures, researchers can gain a more comprehensive understanding of the study results, identify meaningful associations, and avoid drawing conclusions based solely on statistical significance. This approach ensures that the reported findings have practical implications in clinical decision-making and can guide evidence-based practices.

The effect size measure proposed in this study was developed based on breast cancer data from the National Cancer Database (NCDB). To ensure its broader applicability, the robustness of this measure, along with the recommended conventions for interpretation, needs to be thoroughly investigated across various types of cancer and in diverse medical databases. Only through such comprehensive validation can the suggested novel effect size measure be established as a globally applicable and reliable metric for analyzing large medical datasets.

7.4. Naïve Bayes classifier (NBC)

The NBC results provide valuable insights into the relationship between predictor variables and surgery types in breast cancer patients. The findings show that the misclassification proportion for the stage of cancer is the lowest at 35%, and the corresponding accuracy is the highest at 65%. These results align with previous arguments that the stage of

cancer is the most influential variable associated with surgery types, indicating that patients' cancer stage significantly impacts the choice of surgical treatment.

The variable importance plot further supports the importance of the stage of cancer in determining surgery types. The plot reveals that the stage variable is the most critical factor associated with lumpectomy (LT) and mastectomy without reconstruction (MTnR). Additionally, the stage variable is the second most important variable linked to mastectomy with reconstruction (MTR). These results emphasize the significance of cancer staging in guiding surgical decisions for breast cancer patients.

Interestingly, the *p*-values indicate highly significant associations for surgery types with all predictor variables. However, the odds ratios are close to 1, suggesting that the observed statistical significance might not translate into substantial clinical relevance. This discrepancy between statistical significance and practical significance can lead to the statistical fallacy discussed earlier, where highly significant *p*-values might not provide meaningful insights into the impact of predictor variables on surgery types. This highlights the importance of using effect sizes, as demonstrated in the modified effect sizes, to assess the clinical relevance of the associations.

The modified effect sizes demonstrate a highly significant association between the stage of cancer and surgery types, as well as one of the age groups in the comparison between LT and MTR. These effect sizes provide a more accurate and clinically meaningful measure of the associations, helping researchers understand the practical implications of the predictor variables in determining surgical choices for breast cancer patients.

The consistency of the significant association between the stage of cancer and surgery types across the Naïve Bayes Classifier, ARM, and other methods reinforces the robustness of the findings. These complementary techniques lend additional support to the conclusion that the stage of cancer is the most critical predictor variable influencing surgery types.

Overall, the study contributed to the growing body of literature on statistical analysis methods, offering insights into how to avoid misinterpretations and ensure more robust and clinically meaningful inferences. By addressing the gaps in the existing literature and proposing a novel effect size measure, this research provides valuable guidance to researchers, helping them make informed decisions and draw more accurate conclusions from large EHR databases.

References

Ayyadevara, V. K. (2018). *Pro Machine Learning Algorithms: A Hands-On Approach To Implementing Algorithms In Python And R*. Apress.

Breckenridge, A. M., Breckenridge, R. A., and Peck, C. C. (2019). Report on the current status of the use of real-world data (RWD) and real-world evidence (RWE) in drug development and regulation. *British Journal of Clinical Pharmacology*, **85**, 1874-1877.

Catarino, A., Churches, O., Baron-Cohen, S., Andrade, A., and Ring, H. (2011). Atypical EEG complexity in autism spectrum conditions: a multiscale entropy analysis. *Clinical Neurophysiology*, **122**, 2375-2383.

Chen, K., Liu, J., Zhu, L., Su, F., Song, E., and Jacobs, L. K. (2015). Comparative effectiveness study of breast-conserving surgery and mastectomy in the general population: a NCDB analysis. *Oncotarget*, **6**, 40127.

Chiba, A., Hoskin, T. L., Heins, C. N., Hunt, K. K., Habermann, E. B., and Boughey, J. C. (2017). Trends in neoadjuvant endocrine therapy use and impact on rates of breast conservation in hormone receptor-positive breast cancer: a national cancer data base study. *Annals of Surgical Oncology*, **24**, 418-424.

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, **19**, 3127-3131.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates Publishers.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, **1**, 98-101.

Collins, J., Fanciulli, M., Hohlfeld, R., Finch, D., Sandri, G. V. H., and Shtatland, E. (1992). A random number generator based on the logit transform of the logistic variable. *Computers in Physics*, **6**, 630-632.

Geurts, K., Wets, G., Brijs, T., and Vanhoof , K. (2003). Profiling of high-frequency accident locations by use of association rules. *Transportation Research Record*, **1840**, 123-130.

Grissom, R. J. and Kim, J. J. (2005). *Effect Sizes for Research: A Broad Practical Approach*. Lawrence Erlbaum Associates Publishers.

Hedges, S. B., Duellman, W. E., and Heinicke, M. P. (2008). New World direct-developing frogs (Anura: Terrarana): molecular phylogeny, classification, biogeography, and conservation. *Zootaxa*, **1737**, 181–182.

Hernán, M. A., Clayton, D., and Keiding, N. (2011). The Simpson's paradox unraveled. *International Journal of Epidemiology*, **40**, 780-785.

Hotsinpiller, W., Everett, A., Richman, J., Parker, C., and Boggs, D. (2021). Rates of margin positive resection with breast conservation for invasive breast cancer using the NCDB. *The Breast*, **60**, 86-89.

Khrennikov, A. (2008). Bell-Boole inequality: nonlocality or probabilistic incompatibility of random variables?. *Entropy*, **10**, 19-32.

Klotz, J. (1973). Statistical inference in Bernoulli trials with dependence. *The Annals of Statistics*, **1**, 373-379.

Landercasper, J., Bennie, B., Parsons, B. M., Dietrich, L. L., Greenberg, C. C., Wilke, L. G., and Linebarger, J. H. (2017). Fewer reoperations after lumpectomy for breast cancer with neoadjuvant rather than adjuvant chemotherapy: a report from the national cancer database. *Annals of Surgical Oncology*, **24**, 1507-1515.

Landercasper, J., Ramirez, L. D., Borgert, A. J., Ahmad, H. F., Parsons, B. M., Dietrich, L. L., and Linebarger, J. H. (2019). A reappraisal of the comparative effectiveness of lumpectomy versus mastectomy on breast cancer survival: a propensity score–matched update from the National Cancer Data Base (NCDB). *Clinical Breast Cancer*, **19**, e481-e493.

Lee, S. and Lee, D. K. (2018). What is the proper way to apply the multiple comparison test?. *Korean Journal of Anesthesiology*, **71**, 353-360.

Lehrberg, A., Sebai, M., Finn, D., Lee, D., Karabon, P., Kiran, S., and Dekhne, N. (2021). Trends, survival outcomes, and predictors of nonadherence to mastectomy guidelines for nonmetastatic inflammatory breast cancer. *The Breast Journal*, **27**, 753-760.

Leon, A. C. (2004). Multiplicity-adjusted sample size requirements: a strategy to maintain statistical power with Bonferroni adjustments. *Journal of Clinical Psychiatry*, **65**, 1511-1514.

Lewis, G. D., Xing, Y., Haque, W., Patel, T., Schwartz, M., Chen, A., Farach, A., Hatch, S. S., Butler, E. B., and Chang, J. (2019). Prognosis of lymphotropic invasive micropapillary breast carcinoma analyzed by using data from the National Cancer Database. *Cancer Communications*, **39**, 1-9.

Mazor, A. M., Mateo, A. M., Demora, L., Sigurdson, E. R., Handorf, E., Daly, J. M., A. A., Aggon, Anderson, P. R., Weiss, S. E., and Bleicher, R. J. (2019). Breast conservation versus mastectomy in patients with T3 breast cancers (> 5 cm): An analysis of 37,268 patients from the National Cancer Database. *Breast Cancer Research and Treatment*, **173**, 301-311.

McClelland, S., Hatfield, J., Degnin, C., Chen, Y., and Mitin, T. (2019). Extent of resection and role of adjuvant treatment in resected localized breast angiosarcoma. *Breast Cancer Research and Treatment*, **175**, 409-418.

McNulty, K. (2021). *Handbook of Regression Modeling in People Analytics: With Examples in R and Python*. CRC Press.

Mills, M. N., Yang, G. Q., Oliver, D. E., Liveringhouse, C. L., Ahmed, K. A., Orman, A. G., Laronga, C., Hoover, S. J., Khakpour, N., and Costa, R. L. (2018). Histologic heterogeneity of triple negative breast cancer: A National Cancer Centre Database analysis. *European Journal of Cancer*, **98**, 48-58.

Moolgavkar, S. H., Chang, E. T., Watson, H. N., and Lau, E. C. (2018). An Assessment of the Cox Proportional Hazards Regression Model for Epidemiologic Studies. *Risk Analysis*, **38**, 777-794.

Pearl, J. (2022). Comment: understanding Simpson's paradox. *Probabilistic and Causal Inference: The Works of Judea Pearl*. 399-412.

Pratt, D., Burneikis, T., Tu, C., and Grobmyer, S. (2021). Time to completion of breast cancer treatment and survival. *Annals of Surgical Oncology*, **28**, 8600-8608.

Pusic, A., Thompson, T. A., Kerrigan, C. L., Sargeant, R., Slezak, S., Chang, B. W., Helzlsouer, K. J., and Manson, P. N. (1999). Surgical options for early-stage breast cancer: factors associated with patient choice and postoperative quality of life. *Plastic and Reconstructive Surgery*, **104**, 1325-1333.

Ranstam, J. (2012). *Why The p-value Culture Is Bad And Confidence Intervals A Better Alternative*. Elsevier.

Rosenthal, R., Cooper, H., and Hedges, L. (1994). Parametric measures of effect size. *The Handbook of Research Synthesis*, **621**, 231-244.

Rusthoven, C., Rabinovitch, R., Jones, B., Koshy, M., Amini, A., Yeh, N., Jackson, M., and Fisher, C. (2016). The impact of postmastectomy and regional nodal radiation after neoadjuvant chemotherapy for clinically lymph node-positive breast cancer: a National Cancer Database (NCDB) analysis. *Annals of Oncology*, **27**, 818-827.

Solla, F., Tran, A., Bertoncelli, D., Musoff, C., and Bertoncelli, C. M. (2018). Why a *p*-value is not enough. *Clinical Spine Surgery*, **31**, 385-388.

Sullivan, G. M. and Feinn, R. (2012). Using effect size—or why the *p*-value is not enough. *Journal of Graduate Medical Education*, **4**, 279-282.

Wald, A. (1947). *Sequential Analysis*. J. Wiley & Sons.

Weiser, R., Haque, W., Polychronopoulou, E., Hatch, S. S., Kuo, Y.-f., Gradishar, W. J., and Klimberg, V. S. (2021). The 21-gene recurrence score in node-positive, hormone receptor-positive, HER2-negative breast cancer: a cautionary tale from an NCDB analysis. *Breast Cancer Research and Treatment*, **185**, 667-676.

Wrubel, E., Natwick, R., and Wright, G. P. (2021). Breast-conserving therapy is associated with improved survival compared with mastectomy for early-stage breast cancer: a propensity score matched comparison using the national cancer database. *Annals of Surgical Oncology*, **28**, 914-919.

Zhang, Z. (2016). Naïve Bayes classification in R. *Annals of Translational Medicine*, **4,** 241.

Zhu, L. and Chen, K. (2019). Clinicopathological features, treatment patterns, and prognosis of squamous cell carcinoma of the breast: an NCDB analysis. *BMC Cancer*, **19**, 1-9.