RESEARCH METHODILOGY

3.1 Introduction

This section presents the research methodology and processes adopted for the research to the intent of providing data driven answers to the research objectives.

3.2 Study Design

The study adopted a cross-sectional design, also called a social survey design. A cross-sectional study is a type of research design in which you collect data from many different individuals at a single point in time. In cross-sectional research, you observe variables without influencing them.

3.3 Study Area

The recurrent breast cancer dataset is a comprehensive collection of data from the Oncology Institute in the United State of America (USA), specifically curated to facilitate research and analysis in the domain of breast cancer recurrence. The dataset comprises 279 instances, each characterized by a set of attributes providing insights into various factors associated with breast cancer cases.

3.4 Data Sources

The data was downloaded from the Kaggle website which is an opensource data platform for collaboration amongst statisticians, machine learning experts, data scientists, data analyst and researchers. (https://www.kaggle.com/datasets/tanshihjen/recurrent-breastcancerdataset)

Data Dictionary

The attributes of the data are presented below:

Attributes	Description	Type
Age	Age	Ordinal
Menopause	Menopause status	Nominal; 1=Premeno,2=Ge40
Deg_Malig	Degree of Malignant	Ordinal; 1= Grade 1, 2= Grade 2. 3= Grade 3
Breast	Breast affected	Nominal; 1 = left, 2 = right
Irradiation	Irradiation treatment	Nominal: 1 = Yes, 0 = No

Target	Recurrence status of Breast cancer	Nominal: 1 = recurrent event, 0 = No recurrent event

Method of Data Analysis

To achieve the research objectives of this study, the study carried out using the Chi-square test, odd ratio and risk ratio.

The Chi-square test also known as test for goodness-of-fit and test of independence are his most important contribution to the modern theory of statistics. The importance of Pearson's Chi-square distribution was that, the statisticians could use the statistical methods that did not depend on the normal distribution to interpret its findings.

Chi-square test is a nonparametric test used for two specific purposes:

(a) To test the hypothesis of no association between two or more groups, population or criteria (i.e. to check independence between two variables);

(b) and to test how likely the observed distribution of data fits with the distribution that is expected (i.e., to test the goodness-of-fit).

Assumptions of Chi-Square test

I. The data are randomly drawn from a population

II. The values in the cells are considered adequate when expected counts are not <5 and there are no cells with zero count [4,5]

III. The sample size is sufficiently large.

IV. The variables under consideration must be mutually exclusive. It means that each variable must only be counted once in a particular category and should not be allowed to appear in other category. In other, words no item shall be counted twice.

The Chi-Square can be expressed as;

$$x^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$$

Where;

O = the observed frequency,

e = the expected frequency**.**

Risk Ratio

Risk Ratio (RR), also known as Relative Risk, is a statistical measure used in epidemiology and clinical research to compare the likelihood of an event occurring in an exposed group versus an unexposed group. It provides insight into the strength of association between a specific exposure (such as a treatment, lifestyle factor, or environmental condition) and an outcome (such as disease occurrence or recurrence).

Mathematically, the Risk Ratio is expressed as:

$$RR=\frac{\text{Risk in Exposed Group}}{\text{Risk in Unexposed Group}}$$

Where:

- Risk in Exposed Group = $\frac{A}{A+B}$
- Risk in Unexposed Group = $\frac{C}{C+D}$

Therefore, RR = $\frac{A/_{A+B}}{C/_{C+D}}$

Outcome	Exposed Group	Unexposed Group
Recurrent event	A	C
No Recurrent event	B	D

In the table above, A and C represent the number of individuals who experience the outcome (e.g., disease) in the exposed and unexposed groups, respectively, while B and D represent those who do not experience the outcome.

Odd Ratio

The Odds Ratio (OR) is a statistical measure used to determine the strength of association between an exposure (such as a risk factor or treatment) and an outcome (such as disease occurrence). It is commonly used in epidemiological and medical research, especially in case-control studies and logistic regression models.

$$\text{Odd Ratio (OR)} = \frac{A \times D}{B \times C}$$

A = Number of exposed individuals with the outcome

B = Number of exposed individuals without the outcome

C = Number of unexposed individuals with the outcome

D = Number of unexposed individuals without the outcome

DATA ANALYSIS AND INTERPRETATION OF RESULTS

1.1 Introduction

The recurrent breast cancer dataset is a comprehensive collection of data from the Oncology Institute in the United State of America (USA), specifically curated to facilitate research and analysis in the domain of breast cancer recurrence. The dataset comprises 279 instances, each characterized by a set of attributes providing insights into various factors associated with breast cancer cases.

1.2 Descriptive Statistics

This section will describe the features contained in the data set with percentages and appropriate visualization to help and provide accurate understanding of the dataset.

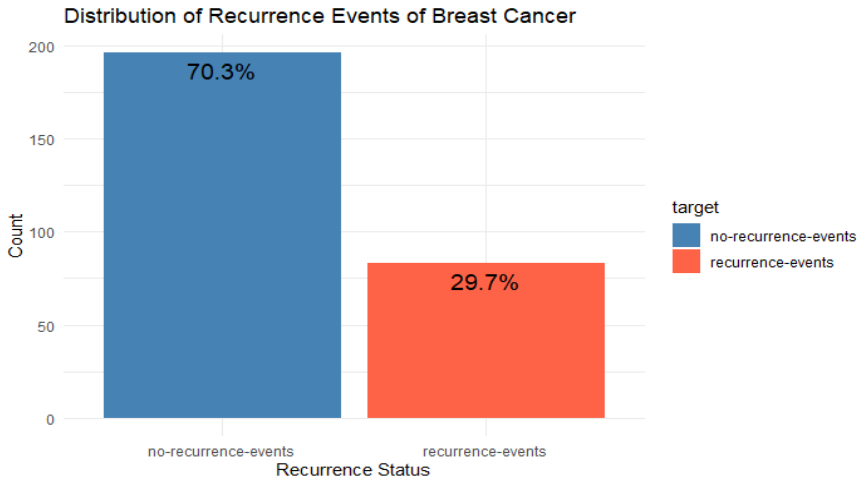

Figure 1: Distribution of the recurrent status of breast cancer

This chart shows the distribution of breast cancer recurrence events. 70.3% of patients experienced no recurrence, while 29.7% experienced a recurrence.

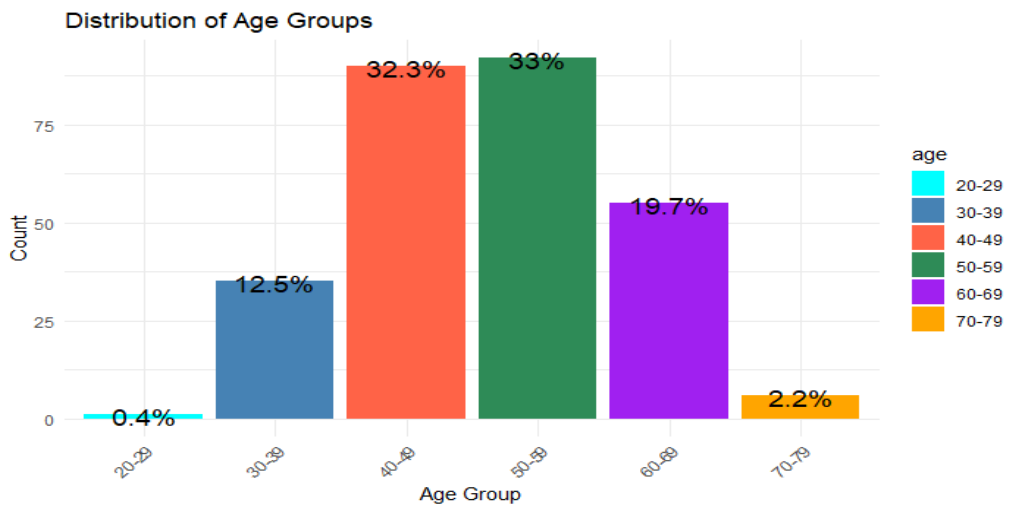

Figure 2: Age distribution of the patients

Figure 2 shows the distribution of age groups within the dataset reveals that the largest represented categories are those aged 50-59 and 40-49, comprising 33% and 32.3% of the total, respectively. The oldest and youngest age brackets are significantly less represented, with those aged 70-79 comprising only 2.2% and those aged 20-29 making up a mere 0.4% of the overall distribution.

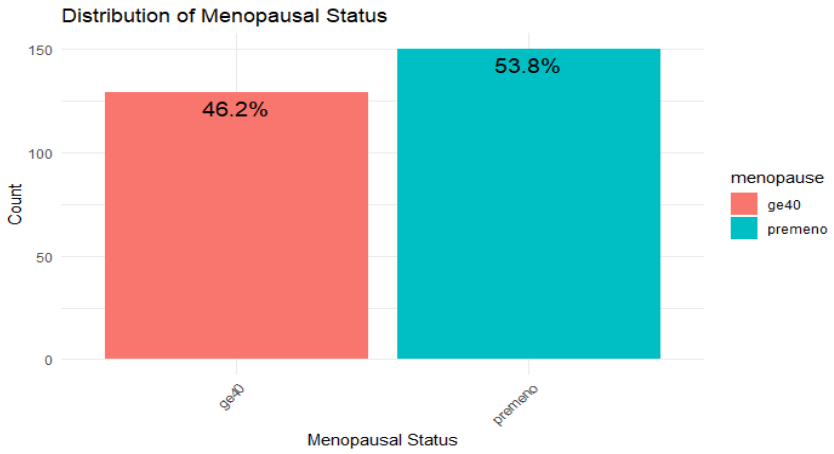

Figure 3: Distribution of Menopause status

Figure 3 displays the distribution of menopausal status, revealing that pre-menopausal women constitute the majority at 53.8%, while women aged 40 and above (ge40) make up the

remaining 46.2%. This indicates a higher prevalence of pre-menopausal status within the studied population.

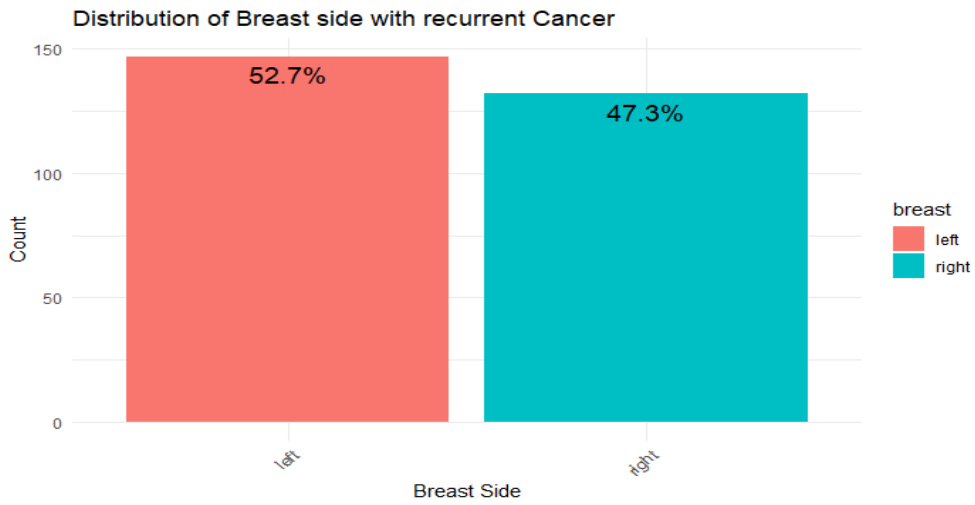

Figure 4: Distribution of breast side with recurrent cancer

Figure 4 illustrates the distribution of breast cancer recurrence based on the affected side. The data reveals a slightly higher occurrence of recurrence on the left side, representing 52.7% of cases, while the right side accounts for 47.3%.

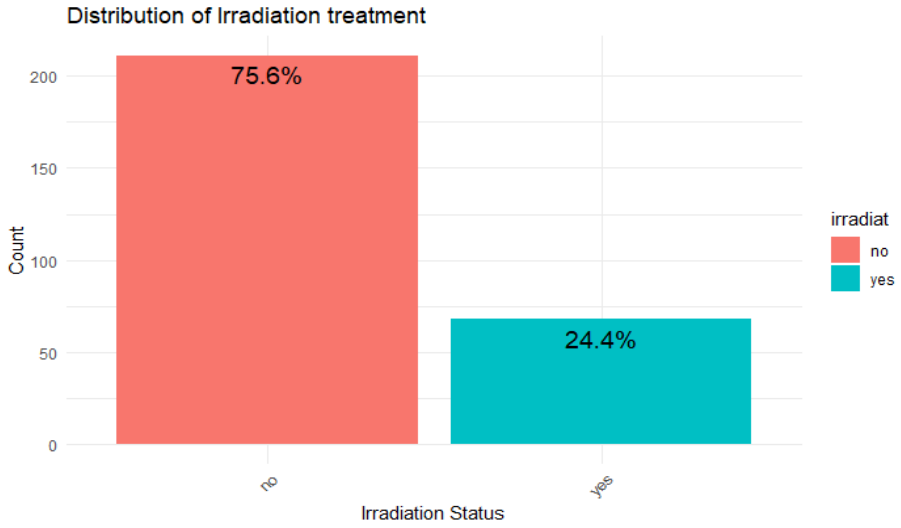

Figure 5: Irradiation treatment status distribution

Figure 5 displays the distribution of irradiation treatment, categorized as "no" or "yes." The majority of cases (75.6%) did not receive irradiation treatment, while 24.4% did. This indicates a substantially lower prevalence of irradiation treatment within the dataset.

Hypothesis Testing

H0: There is no significant association between menopause status and the recurrence of breast cancer using risk ratios and odds ratios.

HA: There is a significant association between menopause status and the recurrence of breast cancer using risk ratios and odds ratios.

Table 1

	No recurrent event (Outcome -)	**Recurrent event (Outcome +)**	**Total**
Ge40 (Post Menopausal)	94	35	129
Premeno (Pre-Menopausal)	102	48	150
Total	196	83	279

```
Point estimates and 95% CIs:
-------------------------------------------------------------------
Inc risk ratio                                 1.07 (0.92, 1.25)
Inc odds ratio                                 1.26 (0.75, 2.12)
Attrib risk in the exposed *                   4.87 (-5.84, 15.57)
Attrib fraction in the exposed (%)            6.68 (-8.65, 19.85)
Attrib risk in the population *                2.25 (-6.94, 11.44)
Attrib fraction in the population (%)         3.20 (-4.13, 10.02)
-------------------------------------------------------------------
Uncorrected chi2 test that OR = 1: chi2(1) = 0.786 Pr>chi2 = 0.375
Fisher exact test that OR = 1: Pr>chi2 = 0.431
 Wald confidence limits
 CI: confidence interval
 * Outcomes per 100 population units
```

Authors computation: R output

The incidence risk ratio (RR) was calculated to be 1.07, with a 95% confidence interval (CI) of [0.92, 1.25]. This explains that the risk of breast cancer recurrence events in post-menopausal women is approximately 1.07 times higher than those in pre-menopausal women.

The incidence odds ratio (OR) was calculated as 1.26, with a 95% CI of [0.75, 2.12], explains that the odds of recurrence events are 1.26 times higher in post-menopausal women compared to pre-menopausal women.

The chi-squared test for the null hypothesis yielded a chi-squared value of 0.786 with a p-value of 0.375. This result indicates that the observed data does not provide sufficient evidence to reject the null hypothesis, suggesting no significant association between menopause status and the recurrence of breast cancer.

Hypothesis two

H_0: There is no significant association between the side of the breast affected by cancer (right or left) and the recurrence of breast cancer using risk ratios and odds ratios.

Table 2

	No recurrent event (Outcome -)	**Recurrent event (Outcome +)**	**Total**
Left breast	100	47	147
Right breast	96	36	132
Total	**196**	**83**	**279**

```
Point estimates and 95% CIs:
-------------------------------------------------------------------
Inc risk ratio                                 0.94 (0.80, 1.09)
Inc odds ratio                                 0.80 (0.48, 1.34)
Attrib risk in the exposed *                   -4.70 (-15.40, 6.00)
Attrib fraction in the exposed (%)            -6.91 (-24.50, 8.19)
Attrib risk in the population *                -2.48 (-11.78, 6.82)
Attrib fraction in the population (%)         -3.53 (-11.90, 4.22)
-------------------------------------------------------------------
Uncorrected chi2 test that OR = 1: chi2(1) = 0.735 Pr>chi2 = 0.391
Fisher exact test that OR = 1: Pr>chi2 = 0.432
 Wald confidence limits
 CI: confidence interval
 * Outcomes per 100 population units
```

Authors computation: R output

The incidence risk ratio (RR) was calculated as 0.94, with a 95% confidence interval (CI) of [0.80, 1.09]. This indicates that the risk of recurrence events is 0.94 times lower in the left breast compared to the right breast. The incidence odds ratio (OR) was calculated as 0.80, with

a 95% CI of [0.48, 1.34], indicating that the odds of recurrence events are 0.80 times lower in the left breast compared to the right breast.

The chi-squared test for the null hypothesis yielded a chi-squared value of 0.735 with a p-value of 0.391. This result is greater than 0.05, meaning that the data does not provide sufficient evidence to reject the null hypothesis. Therefore, there is no statistically significant association between the type of event and recurrence of breast cancer.

Hypothesis three

H0: There is no significant association between the use of irradiation treatment and the recurrence of breast cancer using risk ratios and odds ratios.

Table 3

	No recurrent event (Outcome -)	**Recurrent event (Outcome +)**	**Total**
No (Didn't Receive Irradiation Treatment)	159	52	211
Yes (Received Irradiation Treatment)	37	31	68
Total	196	83	279

```
Point estimates and 95% CIs:
-------------------------------------------------------------------
Inc risk ratio                                 1.38 (1.10, 1.74)
Inc odds ratio                                 2.56 (1.45, 4.53)
Attrib risk in the exposed *                   20.94 (7.76, 34.13)
Attrib fraction in the exposed (%)            27.79 (9.04, 42.68)
Attrib risk in the population *                15.84 (2.84, 28.84)
Attrib fraction in the population (%)         22.55 (6.54, 35.81)
-------------------------------------------------------------------
Uncorrected chi2 test that OR = 1: chi2(1) = 10.794 Pr>chi2 = 0.001
Fisher exact test that OR = 1: Pr>chi2 = 0.001
 Wald confidence limits
 CI: confidence interval
 * Outcomes per 100 population units
```

Authors computation: R output

The incidence risk ratio (RR) was calculated as 1.38, with a 95% confidence interval (CI) of [1.10, 1.74]. This indicates that the risk of recurrence events is 1.38 times higher for those who received irradiation treatment (Yes) compared to those who did not (No).

The incidence odds ratio (OR) was calculated as 2.56, with a 95% CI of [1.45, 4.53]. This means that the odds of experiencing recurrence events are 2.56 times higher for individuals who received irradiation treatment compared to those who did not (No).

The chi-squared test for the null hypothesis that the odds ratio equals 1 produced a chi-squared value of 10.794 with a p-value of 0.001. This result is statistically significant ($p < 0.05$); Thus, we reject the null hypothesis and conclude that there is a statistically significant association between receiving irradiation treatment and the recurrence of breast cancer.

SUMMARY OF FINDINGS, CONCLUSION AND RECOMMENDATIONS

5.1 Summary of findings

- **Menopause Status and Breast Cancer Recurrence**: The analysis found no statistically significant association between menopause status and the recurrence of breast cancer. The incidence risk ratio (RR) was 1.07 (95% CI: [0.92, 1.25]), and the incidence odds ratio (OR) was 1.26 (95% CI: [0.75, 2.12]), indicating a slightly higher risk and odds of recurrence in post-menopausal women compared to pre-menopausal women. However, the chi-squared test yielded a p-value of 0.375, suggesting no significant evidence to reject the null hypothesis.
- **Side of the Breast Affected and Breast Cancer Recurrence**: There was no significant association between the side of the breast affected (left vs. right) and the recurrence of breast cancer. The incidence risk ratio (RR) was 0.94 (95% CI: [0.80, 1.09]), and the incidence odds ratio (OR) was 0.80 (95% CI: [0.48, 1.34]), both suggesting a slightly lower recurrence risk and odds in the left breast. The chi-squared test resulted in a p-value of 0.391, indicating no statistically significant relationship.
- **Irradiation Treatment and Breast Cancer Recurrence**: A statistically significant association was observed between irradiation treatment and breast cancer recurrence. The incidence risk ratio (RR) was 1.38 (95% CI: [1.10, 1.74]), and the incidence odds ratio (OR) was 2.56 (95% CI: [1.45, 4.53]), indicating that individuals who received irradiation treatment had a higher risk and higher odds of recurrence compared to those who did not receive treatment. The chi-squared test revealed a significant p-value of 0.001, supporting the conclusion that irradiation treatment is significantly associated with the recurrence of breast cancer.