

**Biological Factors Affecting the Breast Cancers among Sri
Lankan Women: A Case Study of Wathupitiwala Base
Hospital**

S.A.P.M. Senanayake

(179065M)

Degree of Master of Science in Business Statistics

Department of Mathematics

University of Moratuwa

Sri Lanka

October 2023

**Biological Factors Affecting the Breast Cancers among Sri
Lankan Women: A Case Study of Wathupitiwala Base
Hospital**

S.A.P.M. Senanayake

(179065M)

Dissertation submitted in partial fulfilment of the requirements of the
Degree of Master of Science in Business Statistics

Department of Mathematics
University of Moratuwa

Sri Lanka

October 2023

DECLARATION OF THE CANDIDATE

I declare that this is my own work, and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works.

Signature of the candidate: S.A.P.M. Senanayake

S.A.P.M. Senanayake

14th October 2023

Name:

Date:

The above candidate has carried out research for the Master's Dissertation under my supervision.

Signature

Date:

Ms. R.T. Jayasundara,

Senior Lecturer

Course Coordinator

M Sc / Post Graduate Diploma in Business Statistics

Department of Mathematics

Faculty of Engineering

University of Moratuwa.

Acknowledgement

I wish to express my gratitude to our Madam, D.R.T. Jayasundara, Senior Lecturer, Course Coordinator, Department of Mathematics, University of Moratuwa and the Course Coordinator of the M.Sc./Post Graduate Diploma in Business Statistics for supporting us, for the sound knowledge given and guiding us throughout this course and supporting us in very difficult times to finish our work without hesitation.

Finally, I wish to express my debt to my Husband Bhashana, Mother, Father, Sister, my friends Anjali and Lochanie for their continued and unfailing help.

Abstract

Breast cancer remains a significant public health concern worldwide, with varying risk factors across diverse populations. The incidence of breast cancer in Sri Lanka is also observed to be on the rise, with nearly 3000 women being diagnosed each year. Therefore, the objective of this research is to identify the underlying relationships between significant biological factors affecting the breast cancers among Sri Lankan female population and to uplift the awareness of risk factors among total population.

This study case study was carried out at one of the Base Hospitals in Sri Lanka. The data set was acquired from the patients admitted to the medicine ward of Base Hospital, Wathupitiwala, Sri Lanka. Data collection was carried out by a team of trained medical graduates and nurses between January 2018 and August 2018 period. In this study, data from a total of 144 participants were collected, and it is important to note that all individuals included in this research cohort had tested positive for breast cancer. Data were gathered across 17 distinct variables, encompassing both continuous and categorical variables. Specifically, these variables include Age, Incidental lump, Breast pain, Nipple discharge, Nipple retraction, Breast asymmetry, Age at menarche, Age at first childbirth, Number of pregnancies, Number of children, Breast feeding history, Age at menopause, Usage of Oral Contraceptives, Usage of depot Provera, Family history, Skin Nodule, and Skin ulceration. Principal Component Analysis (PCA) based on Factor Analysis was employed for the analysis of the dataset. The initial dataset underwent a rigorous cleaning process utilizing diverse methods within SPSS, followed by a comprehensive data validation step involving Normality testing. Subsequently, factors were derived, and the correlations between these factors and the initial variables, denoted as factor loadings, were established. Through this analysis, variables exhibiting factor loading values surpassing 0.5 were identified as highly correlated with their respective principal components.

Notably, the study pinpointed the relationship among most significant risk factors associated with breast cancer incidence among Sri Lankan women. These prominent relationships include Number of Children and Number of Pregnancies, Skin Ulceration, Skin Nodule, Breast Asymmetry and Nipple Retraction, Age, Age at menopause and Use of Depot Provera, Nipple Discharge and Breast Pain. While the risk factors of Breast Feeding and Age at Menarche are incorporated into distinct components, there is no discernible significant relationship between these two factors and any other risk factors.

Keywords: (Breast Cancer, PCA, Tumors, Biological factor)

Table of contents

Declaration of the candidate	iii
acknowledgement.....	iv
Abstract	v
Table of contents	vi
List of Figures	x
List of Table	xi
1. Introduction	x
1.1 Problem Identification	3
1.2 Problem Justification	4
1.3 Significance of the study	4
1.4 Limitations of the Study	5
1.5 Objectives of the Study	5
1.6 Summar of the introductiony.....	5
1.7 Chapter Outline	5
2. Litereature Review	6
2.1 Risk Factors of Breast Cancers.....	6
2.1.1 Age	6
2.1.2 Family History	7
2.1.3 Reproductive Factors	7
2.1.4 Estrogen.....	7
2.1.5 Lifestyle.....	8
2.1.6 Breast Density	9
2.1.7 Exposure to Radiation	9
2.1.8 Being Overweigt or Obese	9

2.1.9	Not being physicaly active	10
2.1.10	Not having Children	10
2.1.11	Not Breastfeeding	10
2.2	Symptoms of Breast Cancers	11
2.3	Scientific View of Breast Cancers.....	11
2.4	Importance of Statistical Analysis Techniques in Disease Diagnosis.....	12
2.5	Breast Cancer Analysis and Statistical Method	12
2.10	Summery of Literature Review	13
3.	Materials and Methodology	14
3.1	Consetual Framework.....	14
3.2	Data Acqision	14
3.3	Participants	14
3.4	Types of Variables.....	14
3.4.1	Descriptions of Variables in Study	16
3.5	Data Analysing	19
3.6	Statistical Methods for Analyzing.....	19
3.6.1	Principal Component Analysis.....	2
3.6.2	Factor Analysis	20
3.7	Methodology	21
3.7.1	Data Pre-processing	21
3.7.2	Data Validation	21
3.7.3	Data standerdization.....	21
3.7.3.1	Skewness and Kurtosis Measurements	21
3.7.3.2	Sharpoi-Wilk Test	22
3.7.3.3	Visual Inspection.....	22
3.7.3	Data standerdization.....	22

3.7.4	Principal Component Analysis.....	22
3.7.5	Factor Loading	22
4.	Results and Discussion.....	24
4.1	Descriptive Statistics of the Dataset	24
4.2	Data Wrangling and Transformation.....	28
4.2.1	Checking for Null Values	28
4.2.2	Checking for Duplicate Values	29
4.3	Data Validation.....	29
4.3.1.	Normality Test	30
4.3.1.1	Skewness and Kurtosis Measurements	30
4.3.1.2	Sharpio-Wilk Test	30
4.3.1.3	Visual Representation	31
4.4	Data Standerdization	34
4.5	Principal Component Analysis.....	35
4.6	Factor Loadings	39
5.	Conclusion	42
5.1	Conclusion.....	42
5.2	Recommondation.....	43
5.3	Future Research Directions	45
	References List	46
	Appendix I – Results of Visual Inspection	53

List of Figures

Figure 2.1 - The lobes and ducts inside the breast lymph nodes near breast	11
Figure 3.1 – Conceptual Framework.....	14
Figure 4.1 - Distribution of Age.....	32
Figure 4.2 - Distribution of Age at Menarche.....	32
Figure 4.3 - Distribution of Age at First Child	33
Figure 4.4 - Q-Q Plot of Age	33
Figure 4.5 - Q-Q Plot of Age at Menarche	34
Figure 4.6 - Q-Q Plot of Age at First Child	34
Figure 4.7 - Screening plot of all Initial Variables.....	38

List of Table

Table 3.1 - Data Types	15
Table 3.2 - Scale of Categorical Variables.....	15
Table 4.1 - Dataset Information	24
Table 4.2 - Parameters in Descriptive Statistics.....	25
Table 4.3 - Descriptive Statistics of all Variables.....	26
Table 4.4 - Summarized Correlation Matrix of the Dataset.....	27
Table 4.5 - Variables with Missing Values.....	28
Table 4.6 - Skewness and Kurtosis Statistics of Initial Variables	30
Table 4.7 - Sharpio-Willks Test Statistics.....	31
Table 4.8 - Summarized Significant Values of Total Correlation Matrix	35
Table 4.9 - KMO and Bartlett's Test	36
Table 4.10 - Communalities	36
Table 4.11 - Total Variance Explained	37
Table 4.12 - Correlation Between Initial Variables and Selected Factors	39
Table 4.13 - Factor Loadings of Selected Factors and Initial Variables(Varimax) ...	40