UNIVERSITI TEKNOLOGI MARA

UNDERSTANDING THE OCCURRENCE OF METASTATIC BREAST CANCER THROUGH CLINICAL, PHENOTYPE AND GENOTYPE DATA, AND THE EMPLOYMENT OF MACHINE LEARNING

NADIA JALALUDIN

PhD

August 2023

UNIVERSITI TEKNOLOGI MARA

UNDERSTANDING THE OCCURRENCE OF METASTATIC BREAST CANCER THROUGH CLINICAL, PHENOTYPE AND GENOTYPE DATA, AND THE EMPLOYMENT OF MACHINE LEARNING

NADIA JALALUDIN

Dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Pharmacy

August 2023

CONFIRMATION BY PANEL OF EXAMINERS

I certify that a Panel of Examiners has met on 20 December 2023 to conduct the final examination of Nadia Jalaludin on her **Doctor of Philosophy** thesis entitled "Understanding the Occurrence of Metastatic Breast Cancer Through Clinical, Phenotype and Genotype Data, and the Employment of Machine Learning" in accordance with Universiti Teknologi MARA Act 1976 (Akta 173). The Panel of Examiner recommends that the student be awarded the relevant degree. The Panel of Examiners was as follows:

Meor Mohd Redzuan Meor Mohd
Affandi, PhD
Associate Professor
Faculty of Pharmacy
Universiti Teknologi MARA
(Chairman)

Yuslina Zakaria, PhD
Senior Lecturer
Faculty of Pharmacy
Universiti Teknologi MARA
(Internal Examiner)

Noraida Mohamed Shah, PhD
Associate Professor
Faculty of Medicine
Universiti Kebangsaan Malaysia
(External Examiner)

PROFESSOR IR. DR. ZUHAINA
HAJI ZAKARIA
Dean
Institute of Graduates Studies
Universiti Teknologi MARA
Date: 28 August 2023

AUTHOR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the regulations of Universiti Teknologi MARA. It is original and is the results of my own work, unless otherwise indicated or acknowledged as referenced work. This thesis has not been submitted to any other academic institution or non-academic institution for any degree or qualification.

I, hereby, acknowledge that I have been supplied with the Academic Rules and Regulations for Postgraduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

Name of Student	:	Nadia Jalaludin
Student I.D. No.	:	2014274662
Programme	:	Doctor of Philosophy – PH990
Faculty	:	Pharmacy
Thesis Title	:	Understanding the Occurrence of Metastatic Breast Cancer Through Clinical, Phenotype and Genotype Data, and the Employment of Machine Learning

Signature of Student : ……………………………………………..

Date : August 2023

ABSTRACT

Approximately 2.3 million women were diagnosed with breast cancer (BC) in 2020 and nearly 30% of women diagnosed with early-stage breast cancer will later develop metastatic disease. Despite the development and discovery of drugs and pharmacotherapy for breast cancer, the 5-year survival rate for people with metastatic breast cancer (MBC) remains low. Therefore, the objective of this study is to: a) mine and integrate clinical, phenotype and genotype data that contributes to the occurrence of MBC, b) build a prediction model that can predict possibility of occurrence to metastatic state of breast cancer based on factors previously determined in (a), and c) to validate findings from (a) and (b) through systematic review of randomised controlled trials of MBC. For objective (a), genotype and clinical data was mined from databases such as cBioportal and Genomic Data Common (GDC) portal, and was analysed using principal component analysis (PCA; after feature selection) and multiple correspondence analysis (MCA) in R. The data was then subjected to subsequent pathway mapping, Gene Ontology (GO) mapping and protein-protein interaction (PPI) to investigate its connection to the metastatic phenotype. The odds ratio of mutated gene, disease similarities and hierarchical clustering were also done before all the result was consolidated. For objective (b), prediction model was generated based on the outcome of (a) by using the Random Forest (RF) algorithm and validated by 5-fold cross validation. Additionally, the sensitivity and specificity of each model was also calculated. Meanwhile, for objective (c), six keywords: "metastatic breast cancer, chemotherapy, hormonal therapy, targeted therapy, gene and progression free survival" were used in PubMed, Scopus, Web of Science, Cochrane and Science Direct to further validate the previous findings. Based on all of these evaluations, the findings suggest that mRNA and genetic profiling can differentiate between breast cancer and metastatic breast cancer patients and more attention should be paid to YAP1 and SP7 genes. It was also found that the most important factors to predict MBC are age, and mutations in OR5T2 and SCGB1D1 genes with an importance of 0.37, 0.93 and 0.95 respectively (scale of importance from 0 to 1, 1 being the highest). Moreover, clinical factors such as age, chemotherapy, hormone therapy, ER, PR and HER2 status can also differentiate between BC and MBC to a certain extent. This was further validated by the systematic review that found with the latest systemic treatment, HR-negative/HER2-positive and HR- positive/HER2-negative MBC patients seemed to have higher median progression free survival (PFS) and overall survival (OS) compared to its counterpart. This is in line with the previous hierarchical clustering result that aligned HER2-, ER+, PR+, age_5, hormonal therapy and patient BC all together while HER2+ and HR- were also clustered together. Given all of these revelations, this work showed persuasive evidence that by identifying the mRNA, gene and clinical profiling, the occurrence of MBC can eventually be predicted and elucidated. Future work should include the combination of these domains in one predictive model, or combining the results of both clinical and genotype predictive models, as this prediction model have the potential to aid in the clinical management of the disease.

ACKNOWLEDGEMENT

Firstly, I wish to thank God for giving me the opportunity to embark on my PhD and for completing this long and challenging journey successfully. My gratitude and thanks go to my supervisor Dr. Fazlin Mohd Fauzi for her endless patience, guidance and support and my co-supervisor Dr. Mahmathi Karuppannan, whose commitments made this journey possible.

My appreciation goes to my family for their continued support throughout this journey. Special thanks to my colleagues and friends in UiTM for helping me with this project.

Thank you everyone.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviations	
ASR	Age Standardise Rate
BC	Breast Cancer
ER	Oestrogen Receptor
GDC	Genomic Data Common
GLOBOCAN	Global Cancer Observatory
HR	Hormone Receptor
MBC	Metastatic Breast Cancer
OS	Overall Survival
PFS	Progression Free Survival
PR	Progesterone Receptor
SNP	Single Nucleotide Polymorphism
TCGA	The Cancer Genome Atlas
TNBC	Triple Negative Breast Cancer
WHO	World Health Organisation

LIST OF NOMENCLATURE

Nomenclatures

BRCA1	Breast Cancer Gene 1
BRCA2	Breast Cancer Gene 2
CNA	Copy Number Alteration
CSC	Cancer Stem Cell
DO	Disease Ontology
EMT	Epithelial–Mesenchymal Transition
GO	Gene Ontology
HER2	Human Epidermal Growth Factor Receptor 2
KEGG	Kyoto Encyclopaedia of Genes and Genomes
MAPK	Microtubule Associated Protein Kinase
MCA	Multiple Correspondence Analysis
OR	Odds Ratio
PARP	Poly Adenosine Diphosphate-Ribose Polymerase
PCA	Principal Component Analysis
PMN	Pre-Metastatic Niche
PPI	Protein-Protein Interaction
TDM-1	Trastuzumab Emtansine
TKI	Tyrosine Kinase Inhibitors

CHAPTER ONE
INTRODUCTION

1.1 Breast Cancer

In 2020, 2.3 million women were diagnosed with breast cancer and there were 685,000 death cases reported worldwide (WHO, 2021). Meanwhile, there were 7.8 million women alive who were diagnosed with breast cancer in the past 5 years, making it the world's most prevalent cancer by the end of 2020 (WHO, 2021).

In Malaysia, breast cancer ranked first in terms of number of new cases and 5-year prevalence in 2020 (GLOBOCAN, 2020), while number of deaths for breast cancer ranked second after lung cancer. This high mortality rate of breast cancer usually stemmed from the late stage of breast cancer, also known as metastatic breast cancer.

1.2 Metastatic Breast Cancer

Metastatic breast cancer is also known as Stage IV breast cancer, which signifies the cancer cells has already migrated to other organs in the body. Several steps are involved in the biological process of metastasis which involves the local invasion of primary tumour cells into surrounding tissues; intravasation of these cells into the circulatory system and survival during hematogenous transit; arrest and extravasation through vascular walls into the parenchyma of distant tissues; formation of micrometastatic colonies in this parenchyma; and the subsequent proliferation of microscopic colonies into overt, clinically detectable metastatic lesions, this last process being termed colonisation. (Lambert et al, 2017; Chambers et al, 2002; Fidler, 2003).

Major alterations in genomic deoxyribonucleic acid (DNA) were once viewed as an exclusive trait of advanced cancers. However, it is now recognised that DNA damage and genomic instability are also the underlying features of human cancer from the earliest stages of tumourigenesis (Gaorav & Joan, 2006). A defect in regulation of DNA damage checkpoint, DNA repair machinery and mitotic checkpoint could result in genomic instability which in turn might lead to malignant transformation (Yao & Dai, 2014). Damage to genomic DNA can be seen even in apparently normal cells and becomes more apparent as tumours emerge (Gorgoulis et al, 2005). Research of molecular cancer profiling using genomic- level approaches have already revealed genes whose expression in primary tumours correlates strongly with the likelihood of metastatic recurrence (Weigelt et al, 2005).

Besides genomic factor, clinical factors such as age, gender and hormone receptor status can also be a determinant factor of cancer recurrence and metastasis. Most breast cancer transpire in women and the number of occurrences is 100 times higher in women compared to men and the estimated number of deaths for women are 98% as compared to only 2% for men (Siegel et al, 2022). The high prevalence of metastasis is also seen in asymptomatic patients which is high in large tumours (diameter more than 5 cm [15%]) or in patients with extensive nodal disease (more than three involved lymph nodes [4%]) (Gerber et al, 2003).

1.3 Problem Statement

Despite recent advances made in surgical techniques, radiotherapy and the development of molecularly targeted therapies, most cancer-related deaths (more than 90%) are the result of progressive growth of therapy-resistant metastasis (Langley and Fidler, 2007). Approximately 30 – 40% of breast cancer patients were reported to suffer from recurrence (Cheng et al, 2012) and approximately 10 –15% of them were reported to die of cancer metastasis or recurrence (van den Hurk et al, 2011). Dillekas et al (2019) reported that the majority of deaths (at least 2/3) from solid tumours are caused by metastases. Over the years, there were some prediction models developed to predict the risk of getting breast cancer, for targeted intervention, as well as for enrolment into prevention trials (Palmer et al, 2021). However, the models focused more on early detection, rather than predicting the risk of occurrence of metastatic breast cancer. Furthermore, most of the prediction models only incorporated clinical data and genomic data in their prediction and only predicts risk of breast cancer, not metastatic breast cancer. Since there is a lack of MBC prediction models, it is unclear which factors are most likely to contribute to the occurrence of MBC. Therefore, it is essential to know the contributing factors before the prediction model can be built. Since predicting the risk of developing metastasis in individual patients is difficult, more than 80% of breast cancer patients received adjuvant chemotherapy, even though only approximately 40% of these patients relapsed and ultimately die of metastatic breast cancer (Weigelt et al, 2005). Hence, many women who should be cured by undergoing only surgery and radiotherapy, will be ‘over-treated’ and suffer the toxic side effects of chemotherapy needlessly. Therefore, improving our understanding on the factors that lead to metastatic process of breast cancer might improve the prognosis and clinical management of the disease.

1.4 Research Questions

- What are the clinical, phenotype and genotype factors that lead to metastatic breast cancer and how do they contribute to it?
- Can a prediction model be built based on these factors?
- Will the prediction model be able to predict metastatic occurrence?

1.5 Significance of Study

The mining and integration of clinical, genomic and phenotypic data of breast cancer patients may reveal the interplay between these entities, consequently revealing several pathways and mechanisms of metastatic breast cancer. Furthermore, the prediction models based on these three factors can predict the metastasis-prone breast cancer patients which in turn might lead to better clinical management of the disease.

1.6 Objectives

1. To mine and integrate clinical, phenotype and genotype data that contribute to the occurrence of metastatic breast cancer.
2. To build a prediction model that can predict the possibility of occurrence of metastatic state of breast cancer based on factors determined in (1).
3. To validate findings from (1) and (2) through systematic review of randomised controlled trials of metastatic breast cancer.

CHAPTER TWO
LITERATURE REVIEW

2.1 Background Information on Breast Cancer

Breast cancer (BC) is the most common cancer affecting women worldwide, with nearly 19.3 million cases diagnosed annually (GLOBOCAN, 2020). Breast cancer is caused by uncontrolled growth of cells in any of the tissues, or parts of, the breast. It can start in different parts of the breast, either the lobules, ducts, or connective tissues. Specifically, breast cancers arise from the epithelial cells, which line the terminal duct lobular unit. If the cancer cells have not passed through the basement membrane, they are *in situ* or non-invasive breast cancers. Meanwhile, an invasive cancer is when cancer cells have passed through the basement membrane of the ducts and lobules, invading the surrounding adjacent normal breast tissue and thus have the potential to metastasise (Sibbering and Courtney, 2019).

Breast cancer is the most common cause of cancer death among women (10.1 million deaths in 2020) and the most frequently diagnosed cancer among women worldwide (GLOBOCAN, 2020). In 2020, there were 7.8 million women alive who had been diagnosed with breast cancer in the previous five years (WHO, 2020). Meanwhile in Malaysia, breast cancer ranked first in number of new cases (8 418 new cases in 2020) compared to other cancer sites with number of deaths reaching 3 503, seconded only to lung cancer (GLOBOCAN, 2020). According to Malaysian National Cancer Registry Report 2012-2016, the lifetime risk of getting breast cancer in Malaysia is 1 in 27 for all females, with the breakdowns of 1 in 30 Malays, 1 in 22 Chinese and 1 in 23 Indians, with higher rates are seen after the age of 50 years old.

2.1.1 Signs and Symptoms

Breast lump is the most common symptom in women with breast cancer (83%) and has relatively high predictive value for malignancy (Koo et al, 2017). There are also non-lump breast symptoms such as nipple abnormalities (7%), breast pain (6%) and breast skin abnormalities (2%) (Koo et al, 2017). Apart from that, there are also non-breast symptoms such as back pain (1%) and weight loss (0.3%). Cancer can also be present when there is nipple discharge, skin changes at the breast and if there are abnormalities of the nipple and areola such as retraction or elevation of the nipple (Zhang et al, 2012).

2.1.2 Diagnosis

Breast cancer is generally diagnosed through either screening or a symptom(s) (e.g., pain or a palpable mass) that prompts a diagnostic exam (McDonald et al, 2016). Most early breast cancers are asymptomatic and discovered on a screening mammography (Jordan et. al, 2019). Screening mammography leads to a 19% overall reduction in breast cancer mortality (Pace and Keating, 2014), with less benefit for women in their 40s (15%) and more benefit for women in their 60s (32%). As a result, screening mammography is recommended by the American Cancer Society beginning at the age of 45, or sooner depending on individual preference (McDonald et al, 2016). A meta-analysis of 14 studies of high-risk women found that magnetic resonance imaging (MRI) had a higher sensitivity for malignancy (84.6%) than mammography (38.6%) or ultrasound (39.6%) (Lehman, 2012). Furthermore, the use of MRI as an adjunct to mammography had a higher sensitivity for malignancy (92.7%) than the use of ultrasound as an adjunct to mammography (52%) (Berg, 2009). As a result, for women who have a lifetime risk of breast cancer of greater than 20%, breast MRI as an adjunct to mammography is recommended by the American Cancer Society (McDonald et al, 2016).

2.1.3 Breast Cancer Risk Factors

A risk factor is defined as factors that affect an individual’s chance of getting a disease. In the case of breast cancer, female gender and increasing age are the biggest risk factors for developing the disease (Sibbering and Courtney, 2019). Research shows that a woman’s risk of getting breast cancer is about 100 times more likely than a man, and an aging woman is likely to get breast cancer as most cases are diagnosed in women aged 55 and older (Feng et al, 2018). Besides gender and age, lifestyle and genetic predispositions are also the contributing factors.

Genetic predispositions play a huge role whereby a woman’s risk of developing breast cancer nearly doubles if she has a first-degree relative (mother, sister, or daughter) diagnosed with breast cancer. Close to 15% of US women who suffer from breast cancer also have a family member who has been diagnosed (Colditz et al, 2012). Approximately 15% of breast cancer patients report family history of breast and ovarian cancer, and the most significant genetic predisposition genes identified are Breast Cancer gene 1 and 2 (BRCA1 and BRCA2) (Yip et. al, 2014). Statistically, women with a BRCA1 mutation have a 55 to 65% lifetime risk of developing breast cancer and women with a BRCA2 mutation, has a lifetime risk of 45%. On average, a woman with a BRCA1 or BRCA2 gene mutation has about 70% chance of getting breast cancer by the age of 80 (Feng et al, 2018). Women with one of these two mutations are also more likely to be diagnosed with breast cancer at a younger age, as well as having cancer in both breasts (American Cancer Society, 2019).

Other factors include increased oestrogen exposure through birth control and contraceptives, race and ethnicity, lack of physical activities, significant overweight or obese, excessive alcohol consumption, history of benign breast disease as well as chest/breast exposure to radiation (Jordan et. al, 2019). In Malaysian women, well-known risk factors such as nulliparity, family history, not breastfeeding and use of oral contraceptives are observed to be associated with an increased risk of breast cancer (Yip et. al, 2014).

Summary of risk factors for breast cancer are listed in Table 2.1.

Table 2.1
Risk Factors for Breast Cancer

Risk Factors	**Description**
Age	More than 50 years old
Genetics	BRCA1 and BRCA2 mutations
Personal history of breast cancer	Higher likelihood of recurrence
Family history of breast cancer	First-degree female or male relative, multiple family members
Reproductive	Menarche before age 12 Menopause after age 55 Nulliparity Late age of first pregnancy (after age 30)

Lifestyle	Obesity Sedentary lifestyle Alcohol consumption (>1 drink per day)
Dense breast tissue	More tissue can obscure lesions on mammography
History of benign breast disease	Proliferative lesions such as atypical hyperplasia or lobular carcinoma in situ
Radiation	Chest/breast exposure to radiation at young age

Source: (Jordan et. al, 2019)

2.1.4 Stages of Breast Cancer

The TNM staging system is currently the most used classification guideline to describe the stages of breast cancer. T refers to the primary tumour site and size, N refers to regional lymph node involvement while M refers to the metastatic state of the disease (Union for International Cancer Control, 2022).

In 2018, the American Joint Committee on Cancer (AJCC) updated the breast cancer staging guidelines by adding other cancer characteristics to the TNM system to determine a cancer’s stage. One such characteristics is tumour grade, which is a measurement of how much the cancer cells look like normal cells. Then the oestrogen- and progesterone-receptors statuses were considered to describe the stages of breast cancer. Oestrogen and progesterone are hormones that are involved in the proliferation of breast cells. Thus, presence of receptors for these two hormones in breast cancer cells will fuel the growth of the cells. Knowing the oestrogen- and progesterone-receptor status is useful for clinicians to predict breast cancer response to hormonal treatments.

Human epidermal growth factor receptor 2 (HER2) is also one of the determinants of unfavourable prognostic factor that is associated with high-grade tumours, high rate of cell proliferation, and lymph node involvement (Taucher et al, 2003). HER2 overexpression is present in approximately 20 –30% of breast cancer tumours (Mitri, 2012). Its overexpression is associated with a more aggressive disease, higher recurrence rate, and shortened survival (Hudis, 2007).

Another tumour classification that has been introduced is the Oncotype DX (Genomic Health Inc., Redwood City, CA), which is a clinically validated twenty-one-gene genomic assay that can quantify the risk of breast cancer recurrence (McVeigh et al, 2014). The gene panel includes five reference genes and sixteen cancer-related genes, including those associated with cell proliferation, invasion and hormone response (McVeigh et al, 2014). The 21 genes consist of proliferation-related genes (Ki67, STK15, BIRC5, CCNB1, MYBL2), metastasis-related genes (MMP11, CTSL2), HER2-related genes (GRB7, HER2), sex hormone-related genes (ER, PGR, BCL2, SCUBE2, GSTM1, BAG1, CD68) and internal control genes (ACTB, GAPDH, GUS, RPLPO and TFRC) (Huang et al, 2020). The test generates a recurrence score between 0 and 100 that correlates to the likelihood of disease recurrence within 10 years of diagnosis (McVeigh et al, 2014).

Adding information about tumour grade, hormone-receptor status, HER2 status, and possibly Oncotype DX test results has made determining the stage of a breast cancer more complex, but also more accurate (Koh and Kim, 2019). Based on these characteristics, the stage of the breast cancer can be determined. The stages are usually expressed as a number on a scale of 0 to IV. Stage 0 is non - invasive cancer that remain in their local region, stage I, II and III are invasive cancers characterised by the size of the tumour and the lymph nodes that are affected while stage IV are invasive cancers that have spread outside the breast to other parts of the body and also known as metastatic breast cancer (American Cancer Society, 2021).

2.1.5 Treatment of Breast Cancer

Traditionally, there are three major types of cancer treatment, which are surgery, chemotherapy and radiotherapy. The type of cancer treatment is dependent on factors such as type of tumour and the stage of tumour development. Also, depending on the type and stage, there are three general goals for cancer treatment: removing entire neoplasm, controlling the recurrence or spread of the primary cancer, and relieving pain if all therapeutic methods have been exhausted (National Cancer Institute, 2017).

Surgical intervention is the primary means of local and regional breast cancer treatment (McDonald et al, 2016). A study shows that removing the primary tumour can reduce the mortality rate of patients with primary distant metastatic disease by up to 40%, with median survival of 16 months longer than those who did not undergo surgery (Ruiterkamp et al, 2009).

Chemotherapy is a form of treatment which involves the use of drugs to destroy cancer cells and to shrink tumours to prevent it from further growth and metastasising to other parts of the body. Chemotherapy can also be administered to ease symptoms caused by the cancer, thereby improving quality of life especially in the advanced stages. Several broad classes of drugs for treating breast cancer are available which largely depend on the tumour characteristics and disease extent. Chemotherapy is recommended in most triple-negative (patients with ER negative, PR negative and HER2 negative receptors), HER2-positive breast cancers and in high-risk luminal HER2-negative tumours (Senkus et al, 2015). It is usually administered for 12– 24 weeks (four to eight cycles), depending on the individual recurrence risk and the selected regimen.

Chemotherapy is also categorised based on time of administration, which are before (neoadjuvant chemotherapy) and after (adjuvant chemotherapy) surgery. Adjuvant chemotherapy after definitive surgery is generally recommended for patients with disease at high risk of recurrence (McDonald et al, 2016). While for neoadjuvant therapy, the aim is to reduce the size of the breast cancer (tumour) if it is too big to be removed in an operation. Indications such as a tumour larger than 5 cm in a patient desiring breast conservation or a tumour fixed to the chest wall will be suitable for the treatment. Also, indications such as locally advanced disease, and inflammatory breast cancer are suitable for neoadjuvant therapy (McDonald et al, 2016). Based on a meta-analysis done by Early Breast Cancer Trialists' Collaborative Group (2018), neoadjuvant chemotherapy allows more breast-conserving therapy than adjuvant chemotherapy and provides information about an individual patient's response to a particular chemotherapy regimen. However, it does not reduce breast cancer mortality, and it is associated with moderately increased local recurrence risk, which persists for at least 10 years.

Besides chemotherapy, other treatments for breast cancer include radiation therapy. Radiation therapy for breast cancer uses high-energy X-rays, protons or other particles to kill cancer cells. This therapy is recommended after a patient has undergone surgery. Whole breast radiation therapy (WBRT) reduces the 10-year risk of any first recurrence (including locoregional and distant) by 15% and the 15 -year risk of breast cancer-related mortality by 4% (Darby et al, 2011). While radiation after mastectomy in node-positive patients reduces the 10-year risk of any recurrence (including locoregional and distant) by 10% and the 20-year risk of breast cancer-related mortality by 8% (EBCTCG, 2014).

2.1.6 Prognosis and Survival Rate of Breast Cancer

Prognosis and survival rate of breast cancer depends on many factors and differ between individuals. Numerous prognostic factors for breast cancer have been identified, which include nodal status, tumour size, presence of distant metastasis, and hormone receptor status. The most important prognostic factor in breast cancer is the involvement of axillary nodes (Kim et al, 2016). Most clinical trials classify patients based on four nodal groups that are based on National Surgical Adjuvant Breast and Bowel Project (NSABP) data: negative nodes, 1–3 positive nodes, 4–9 positive nodes, and 10 or more positive nodes. The node-positive group showed significantly worse 5-year and 10-year disease free survival (DFS) compared to the node-negative group (87% vs. 92.4%, 79.3% vs. 85.1%, P = 0.005) (Kim et al, 2016)

Another important prognostic factor is the size of the tumour. Tumour size correlates with the presence and number of involved axillary lymph nodes. Rosen et al. (1993) examined the relationship between tumour size and 20 -year recurrence-free survival and found a significant association, with a 20-year recurrence-free survival of 88% for tumours ≤1 cm, 72% for tumours 1.1 cm to 3 cm, and 59% for tumours between 3.1 cm and 5 cm. However, in a recent study by Liu et. al. (2021), it was found that T4 tumours (tumour size more than 6.1cm) exhibited worse outcomes than N3 tumours (involved 10 or more axillary lymph nodes) independent of other prognostic factors.

Other prognostic factors include the status of the oestrogen- and progesterone-receptors. Oestrogen receptor α (ERα) is an important biomarker, with approximately 70% of all primary breast cancers being ERα-positive (Cao and Lu, 2016). In women with ERα-positive tumours, targeting ERα is effective, reducing the risk of recurrence by 50% for the first 5 years and by a third the following 5 years when tamoxifen is administered (Davies et al, 2011).

2.2 Metastatic Breast Cancer (MBC)

Metastases of cancer account for a vast majority of morbidity and mortality of cancer patients and are associated with about 90% of all cancer-associated deaths (Mehlen and Puisieux, 2006). Despite the advances in the diagnosis and therapy of breast cancer, more than 44 000 women die of metastatic disease in the United States (Slamon et al, 2011). In an Australian study, 7% of patients with breast cancer had metastatic disease at diagnosis and 10% of patients with a diagnosis of early-stage breast cancer (EBC) were found to have metastatic disease within 5 years (Lord et al, 2012).

Cancer metastasis is defined as the formation of new tumours (secondary and tertiary tumour nests) in tissues and organs away from the primary site of tumour origin (Zubair and Ahmad, 2017). Breast cancer can metastasise to bones, lungs, regional lymph nodes, the liver and brain. Approximately 70% of patients with advanced breast cancer have bone metastasis origin (Otaghfar et al, 2015). Bone is the most commonly observed site for distant metastases and is the location of 30 – 40% of first tumour recurrence (Shaffrey et al, 2004).

Metastasis requires a careful choreography of chain-of-events to be completed for successful colonisation, which otherwise can lead to the elimination of emigrating cells at any stage of metastasis (Van et al, 2011). The major events involved in the "metastatic cascade" involved several stages as shown in Figure 2.1, which are as follows: (a) the activation of epithelial–mesenchymal transition (EMT), during which cancer cells lose all cell–cell contact, such as substrate adhesion, acquiring ownership of movement; (b) local invasion, whereby malignant cells degrade the basal lamina, the special extracellular matrix that organises and separates epithelial tissues from the stroma, which plays an important role in both cell signalling and being a reservoir of growth factors released by tumour cells; (c) intravasation, during which tumour cells pass through the walls of blood vessels and enter the bloodstream; (d) the ability to survive in the bloodstream; (e) extravasation, whereby tumour cells exit the bloodstream, passing through the walls of blood vessels into the tissue of a particular organ; (f) establishment of tumour cells in the tissues of the organ where metastasis will form; in other words, the establishment of a pre-metastatic niche (PMN) to create a favourable environment for the growth of cancer cells (Arvelo, 2016).

Identification of underlying mechanism(s) during these crucial events can lead to the design of novel targeted therapies that can limit cancer invasion and result in better management and treatment of cancer (Zubair and Ahmad, 2017).

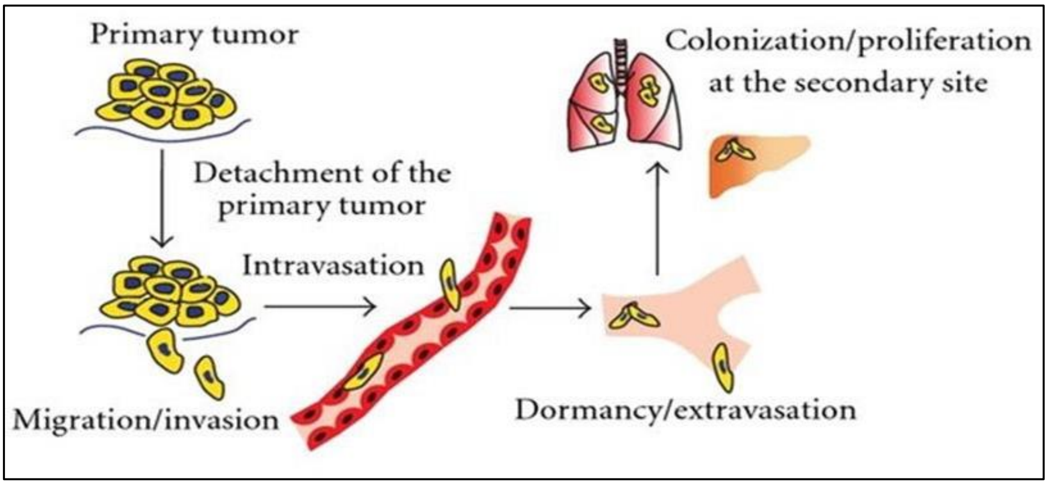

Primary tumor
Colonization/proliferation at the secondary site
Detachment of the primary tumor
Intravasation
Migration/invasion
Dormancy/extravasation

Figure 2.1 The Steps of the Metastatic Cascade (Alsarraj and Hunter, 2002). The steps involved detachment of the primary tumour and migration or intravasation of these cells into the blood stream and colonization or proliferation of the tumour at the secondary site

2.3 Genetics of Metastatic Breast Cancer

Since metastatic breast cancer accounts for majority of breast cancer deaths, it is crucial to understand the underlying mechanism(s) of the metastatic disease. Recent evidence indicates that inherited susceptibility affects not only the development of the primary tumour, but it is also an important factor in progression and metastasis (Shukla et al, 2014). Some genetic changes that help initiate metastasis are mutations, genomic instability and epigenetic alterations which lead to proliferation, angiogenesis, survival as well as the invasion of the tumour (Chatterjee et al, 2018).

Gene-expression studies of cells that are able to establish themselves as distant metastases reveal genes that are co-expressed in the primary tumour as well as some that are not detected in the primary tumour (Minn et al, 2007). A 54-gene signature associated with breast cancer metastasis to the lung includes epiregulin, *CXCL1, COX2* and *MMP1* that are expressed in both primary tumour and metastatic cells, as well as other genes such as *SPARC, MMP2, VCAM1* and *IL13RA2* that are largely expressed in lung metastatic populations (Minn et al, 2005).

Over the years, new metastasis susceptibility genes continued to be identified with the current trend appears to be associated with chromatin interactors and modifiers (Shukla et al, 2014). Changes in the state of chromatin could alter the ratio of heterochromatin and euchromatin, moving the cells into a more ‘stem’ form, affect DNA repair and gene transcription and alter physicochemical properties of the cells that would lead to abnormal proliferation, invasiveness and survival (Shukla et al, 2014). Thus, chromatin modification might emerge as a common theme for metastasis modifiers and become potential biomarkers for metastatic breast cancer.

Understanding the genetic of breast cancer has changed the landscape of how we treat breast cancer in clinical settings. By identifying targetable mutations, better drugs can be selected that are efficacious in specific intrinsic tumour types. Since metastatic breast cancer remained incurable, determining and predicting the genotype of this metastatic state is important in handling this.

2.4 Phenotypes of Metastatic Breast Cancer

Phenotype is defined as the state of an organism resulting from interactions between genes, environment, disease, molecular mechanisms, and chance (Cheng et al, 2016). It is observable and represents the consequences of unique interactions between genetic background and environmental factors (Cheng et al, 2016).

Phenotypically, metastatic tumour cells are supposed to fit well in the concepts of "cancer stem cell" (CSC), which describes a cancer cell's ability to self -renew and establish secondary tumours, and the epithelial-to-mesenchymal transition (EMT), which represents a mechanism that confers migration/invasion abilities to cells (Gao et al, 2018). Many genes strictly related to cell stemness and EMT phenotypes such as enhanced migration, invasion, anti-apoptosis, and self-renewal have been characterised as metastasis-related phenotypes (Gao et al, 2018). However, there is very limited overlap among these metastatic signatures, indicating that perhaps there are certain requirements for metastasis formation in different organ microenvironments that affects the cells' abilities to metastasised (Gao et al, 2018).

The epithelial–mesenchymal transition (EMT) is a process where epithelial cells lose their cell polarity and cell adhesion ability, which will lead to cancer metastasis (Brabletz et al, 2018). Besides genetic alterations, the presence of cancer cells with more mesenchymal, stem cell like features has been associated with increased risk of metastatic disease (Polyak and Weinberg, 2009).

Tumour cell adhesion, migration, and invasion of metastatic cells require the involvements of integrin, a family of transmembrane adhesion receptors, composed of noncovalently linked α and β subunits (Ruoslahti, 1999). In breast and ovarian cancer, as well as in melanoma and glioma, malignant progression is associated with expression of tumour cell integrin αvβ3 (Pignatelli et al, 1992). Rolli et al (2003) found that migration of MBC cells toward fibrinogen is mediated by integrin αvβ3, and it is very strongly enhanced if the receptor is activated. This indicates that breast cancer cell migration depends on the endogenous control of αvβ3 functionality and perhaps on other supporting factors.

Trophinin-associated protein was found to participate in the proliferation, invasion, and migration of many cancers (Li et al, 2019). Trophinin is a member of the Melanoma Antigen Gene (MAGE) family, expressed by human trophoblastic cells. It functions as a unique adhesion molecule that mediates the initial attachment of the blastocyst to the uterine epithelial cells during embryo implantation (Cai et al, 2021) The upregulation of trophinin promoted the metastatic potential in human gallbladder cancer cells, which was correlated with high expression of integrin alpha3, matrix metalloproteinase-7 (MMP-7), MMP-9, and a transcription factor, Ets-1 (Chang, 2009).

Several studies have revealed the role of apoptosis resistance in metastasis, linking development of a metastatic phenotype to the onset of apoptosis loss in cells (Fernandez et al, 2000). Mendez et al (2005) observed an increased in tumourigenesis in tumours over-expressing anti-apoptotic proteins, Bcl-2 and Bcl-xL. This process was paralleled by increased metastatic activity, in terms of its pervasiveness, affecting such organs as the bones and lymph nodes, in addition to the lungs.

Metastatic breast cancer has been shown to display distinct characteristics according to metastatic site. For example, brain metastasis is associated with young age, oestrogen receptor (ER) negativity, prior lung metastasis, HER-2 overexpression, EGFR overexpression, and the basal subtype (Hicks et al, 2006); while bone metastasis is associated with lower histologic grade, ER positivity, ER positivity/progesterone receptor (PR) negativity, and the presence of fibrotic foci in invasive ductal carcinoma (Wei et al, 2008).

One of the challenges in disease studies is to determine the genotype-phenotype relationship in the disease progression because different genetic aberration might still lead to the same disease phenotypes. One explanation of this is that different genetic alterations can dysregulate the same pathways, resulting in similar disease phenotypes (Gao et al, 2018). Therefore, a network-centric view of MBC might help to overcome the challenges posed by the complex genotype-phenotype relationships and facilitates finding genotypic causes of this disease.

2.5 Treatment for Metastatic Breast Cancer

MBC is not a curable disease, but with the right treatment, patients can live longer. In the late 1990s, docetaxel has emerged as the most effective single agent drug for the treatment of metastatic breast cancer (Vogel and Nabholtz, 1999). Then there are also the liposomal-encapsulated anthracyclines, losoxantrone, gemcitabine, capecitabine, uracil plus tegafur (UFT), ethynyluracil (GW 76C85) plus fluorouracil, raltitrexed, pemetrexed disodium (LY 231514), and edatrexate that have significant activity in metastatic breast cancer (Vogel and Nabholtz, 1999).

Aside from that, trastuzumab, a recombinant humanised monoclonal antibody has also become a major treatment for MBC because of its activity and lack of subjective toxicity in most patients. It was the first biological drug approved for the treatment of HER2-positive BC (Maximiano, 2016).

Moving on to the 21st century, treatment of breast cancer is now dependent on the molecular subtypes of breast cancer which are commonly extrapolated into clinical subtypes based on receptor status (Santa-Maria and Gradishar, 2015). The specific receptors that are assessed in standard clinical practice are the oestrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor 2 -neu (HER2) receptor (Santa-Maria and Gradishar, 2015). Whenever metastasis is suspected, it is essential to confirm the receptor status so that the right treatment can be given.

For patients diagnosed with hormone receptor positive (HR+), human epidermal growth factor receptor 2 negative (HER2-) metastatic breast cancer (MBC), which was previously treated with non-steroidal aromatase inhibitors (NSAI), exemestane in combination with everolimus represents an important treatment option (Riccardi et al, 2018).

The Breast Cancer Trials of Oral Everolimus-2 (BOLERO-2) study in 2012 showed that by combining everolimus with exemestane, progression free survival (PFS) in patients with ER-positive metastatic breast cancer previously treated with a nonsteroidal aromatase inhibitor (AI) improved, as compared to treatment with exemestane alone (Baselga et al, 2012)

A new strategy in treating patients with ER-positive breast cancer is to target cyclin-dependent kinases 4 and 6 (CDK4/6), a key pathway involved in regulation of the G1/S transition of the cell cycle. Preclinical studies combining tamoxifen with the CDK4/6 inhibitor, palbociclib, proclaimed synergistic antitumour effects, which led to a phase 2 study randomising 165 women with ER-positive metastatic breast cancer to front-line letrozole alone or in combination with palbociclib. This study showed a significant difference in PFS, and palbociclib in combination with an AI is now an option for front-line therapy in postmenopausal patients with metastatic ER-positive breast cancer (Santa-Maria and Gradishar 2015).

For HER2-positive breast cancer patients, aside from trastuzumab, a few other drugs targeting HER2 were developed and approved, namely lapatinib, pertuzumab and trastuzumab emtansine (TDM-1). Since clinical trials of these drugs showed positive results, these drugs are now considered a standard of care for HER2-positive breast cancer patients (Santa-Maria and Gradishar, 2015). However, in metastatic setting these drugs are not effective due to the mechanism of anti-HER2 therapy resistance. Therefore, drug development for metastatic breast cancer is currently focusing on tackling these mechanisms of resistance.

For this purpose, next-generation tyrosine kinase inhibitors (TKIs) of HER2 were being developed, such as afatinib and neratinib. The small molecule HER2 inhibitor, ONT-380, is being studied in 3 phase 1 studies in combination with other anti-HER2 therapies and results are needed before moving forward with next-phase studies (Santa-Maria and Gradishar, 2015).

Besides TKIs, nanotechnology has also been used to transport cytotoxic drugs specifically to HER2-positive cells. MM-302, which is a HER2-targeted anthracycline-loaded nanoparticle, is currently under investigation in a multicentre phase 2 study in patients whose disease has progressed during treatment with trastuzumab and TDM-1 (Lee et al, 2015). Further studies utilised preclinical models to identify a minimum critical concentration threshold of MM-302 tumour delivery required to control tumour growth. Their data suggested that patients most likely to benefit from treatment with therapeutic nanoparticles can be identified by the use of pre-treatment imaging of nanoparticle deposition in tumours (Lee et al, 2017).

Patients with ER-, PR-, and HER2-negative tumours, the so-called triple-negative breast cancers (TNBCs) however, currently lack targeted therapy options and have only a limited amount of cytotoxic agents available to treat their disease (Foulkes et al, 2010). Thus, new drugs such as iniparib, olaparib and veliparib has been developed that targeted poly (adenosine diphosphate-ribose) polymerase (PARP) inhibitors that showed promising prospect of treating TNBC (Audeh, 2014).

Another strategy targeting TNBC is through the glycoprotein NMB (gpNMB), a transmembrane protein expressed in approximately 40% to 60% of breast cancers (Santa-Maria and Gradishar, 2015). A study using a fully humanised anti- gpNMB monoclonal antibody conjugated tomonomethyl auristatin in patients with metastatic breast cancer showed improvements of PFS in gpNMB-positive and TNBC and phase 2 studies are under way (Bendell et al, 2014).

Despite the development and discovery of drugs and pharmacotherapy for breast cancer, a common finding of many metastatic studies is that long-term survivors are rare. What is severely lacking is an in-depth molecular insight into metastatic breast cancer that could aid in better therapy and consequently higher survival rate (Sledge, 2016).

In recent years, a large amount of data has been generated in different ecosystem of health care. Additionally, the analysis and manipulation of these data are also advancing with increasing computational powers and sophisticated statistical and bioinformatics tools, creating unprecedented opportunities for researchers to not only augment existing knowledge of various diseases but also produce predictive models to elucidate factors influencing these patients' survival and patients’ risks. The research into metastatic breast cancer may benefit from analysing not just molecular data such as genotypic and phenotypic data but also patient-level data, using data mining and artificial intelligence approaches.

2.6 Utilising Data Mining to Predict Occurrence of MBC

2.6.1 Background Information

Data mining is defined as a process of uncovering anomalies, patterns and correlations within large data sets to predict a particular outcome. In the context of health and health care, the purpose of data mining may range from diagnosing to treatment, with the end goal of improving Health Care Output (HCO), or the quality of care that healthcare can provide to patients (Herland et al, 2014). Mining data for these purposes not only involved a large amount of data, but also a diverse set of data which may range from tissue-level to population level data where the different data answers different set of questions. However, of late, these data are integrated and analysed together in order to provide a more holistic answer to the question in hand, as some of the data may overlap and relying on one domain may be insufficient. There are different aims or tasks of data mining, and each of these utilises different quantitative and/or qualitative analyses that involve not just statistical analysis but also artificial intelligence such as machine learning.

The subsequent subsection is arranged as such: a) types of data used for data mining in health care, b) data mining tasks and its methodology, and c) current predictive model in breast cancer.

2.6.2 Types of Data Used for Data Mining in Health Care

There are a few types of data used for data mining. These include tissue-level data, molecular level data, patient-based data and population-level data (Herland et al, 2014) Each data represents a different subfield. Tissue-level data is being utilised in Neuroinformatics, molecular level data is being used in the field of Bioinformatics, patient-based data in Clinical Informatics while population-level data is utilised in Public Health Informatics. However, in the subfield of Translational Bioinformatics, all these data, from molecular level to population-level are being used.

2.6.2.1 Tissue-Level Data

Tissue-level data incorporates imaging data such as images from the Magnetic Resonance Imaging (MRI) to answer some biological questions which can yield information on prognosis, diagnosis and treatments of a disease. The Neuroinformatics field utilises this tissue-level data to represent the broader domain of Medical Image Informatics by limiting the scope to brain images so that more in-depth research may be performed (Herland, et al, 2014). There is a huge project known as the Human Connectome Project (HCP) where the goal is to map the human brain by making a comprehensive connectivity diagram. HCP is looking to find a map of the neural pathways that make up the brain in order to advance current knowledge of how the brain functions and behaves region-to-region (Van Essen et al, 2013). Creating a full connectivity map of the brain could lead to information that could help in determining the reasons why people have certain brain disorders at a level previously unattainable, giving physician a possibility for easier diagnosis, early detection of future illnesses or maybe even prevention of mental or physical ailments (Herland et al, 2014).

2.6.2.2 Molecular-Level Data

Bioinformatics is a field that utilises molecular level data such as the gene expression data and DNA sequencing data to analyse how the human body works. People in this field also developed methods of effectively handling these data since this type of data tends to have thousands (or tens of thousands) of possible molecules, configurations of molecules, or molecule-molecule interactions that needed to be analysed (Herland et al, 2014). An example of the usage of these data is shown by the work of Haferlach et al (2010) who uses gene expression profiling to categorise leukaemia into two different subclasses, by formulating a gene expression profiling classifier. The authors chose to use an all-pairwise classification design using the trimmed mean of the difference between perfect match and mismatch intensities with quantile normalisation, all to handle the multiclass nature of this research. This enables them to place patients into 18 different subclasses of either myeloid or lymphoid leukaemia with a median specificity of 99.8% and a median sensitivity of 95.6%.

2.6.2.3 Patient-Based Data

Patient-based data is used in Clinical Informatics to help physicians make better, faster and more accurate decisions about their patients through analysis of patient data. These types of data, such as age, physiology features and disease type features are useful in making predictions on what will happen to the patient in the future. For example, to predict Intensive Care Unit (ICU) readmission, mortality rate after ICU discharge as well as predicting a 5-year life expectancy rate. These can all be achieved by using patient-based data.

2.6.2.4 Population-Level Data

Data in Public Health Informatics is from the population, gathered either from “traditional” means (experts or hospitals) or gathered from the population (social media) in order to gain medical insight (Herland et al, 2014). This data could be from Twitter, internet query data (e.g., Google search data), message boards, or anywhere else people put information on the internet and it could be useful in answering both clinical questions and epidemic-scale questions. There is much that can be learnt by employing research on social media data including real-time tracking of a harmful and infectious diseases, increasing the knowledge of global distribution for various diseases, and creating an extremely accessible way of letting people get information about any medical questions they might have (Herland et al, 2014). However, the challenge with these social media data is that it could be unreliable and the authenticity might not be real.

To address that issue, data mining can be implemented so that the useful data can be extracted and utilised to gain medical insight of the big data.

2.6.3 Data Mining Tasks and Its Corresponding Methodology

There are several algorithms in data mining that serves different purposes. The purposes may range from classification, regression, association and clustering, all of which will be explained in more detail.

2.6.3.1 Classification

Classification involves assigning a class or property to a new observation or object, based on the characteristics shared with a set of data whose class or property is known. Machine learning is usually used in classification. It has been applied to cancer patients to predict diagnosis as well as survival rate of the patients. The algorithms used include Random Forest (RF), Naïve Bayes and Support Vector Machine (SVM). Delen (2009) has predicted survival of prostate cancer patients using a classification model. The model utilised a public database-SEER (Surveillance, Epidemiology, and End Result) and applied a stratified ten-fold sampling approach. The algorithms used were Decision Tree (DT), Artificial Neural Network (ANN) and Support Vector Machine (SVM). Their result showed that SVM outperformed other algorithms with 92.85% classification accuracy wherein DT and ANN achieved 90% and 91.07% accuracy respectively.

In a work published by Anisha et. al. (2021), they tested the Random Forest algorithm on 14 parameters namely age, gender, BMI, glucose, HOMA, insulin, adiponectin, leptin, resistin, MCP.1, family history, genetic factors, lumps and position to predict the risk of getting breast cancer and the findings came up with area under curve (AUC) value of 0.98. This indicates that the accuracy of the model is 98%, which is considered very desirable.

2.6.3.2 Prediction and Regression

Prediction involves predicting values for new data vectors. Regression models are used to predict a continuous value by predicting the value of dependent variable (Y) from the input of independent variable (X). This enables us to identify the relationship between those variables and generate predictions from them.

There are a few types of regression, namely simple linear regression, logistic regression and polynomial regression. Simple linear regression is the most common type of regression where a linear relationship should exist between target variable and predictor.

Zhou et al (2004) used logistic regression model to relate gene expression with the class labels. They used published microarray data to separate BRCA1 and BRCA2 mutation-positive breast cancers and tested it on data sets of hereditary breast cancer data, small round blue-cell tumour data, and acute leukaemia tumour data. The experimental results showed that the proposed method can effectively find genes that are consistent with the existing biological knowledge and can predict cancer with high accuracy.

2.6.3.3 Association

Association is a way of searching for relationships among members of a data set. It can find features that occur together or features that are correlated. Methods of association include correlation analysis (simple) to counter-propagation or back-propagation neural networks (complex). Ramasamy and Nirmala (2017) used the weighted association rule mining and keyword-based clustering algorithms to detect the accurate disease based on the user symptoms from the hospital information database and obtained the highest accuracy and efficiency as compared to using other algorithms such as the decision tree and K-Nearest neighbour (KNN). K-Nearest Neighbour is an algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. Basically, it classifies a data point based on how its neighbours are classified (Subramaniam, 2019).

2.6.3.4 Clustering

Clustering algorithms are procedures for partitioning data into groups or clusters such that the clusters are distinct, and members of each cluster belong together based on intrinsic similarities. In health care, self-organising maps and k-means are the most commonly used clustering algorithm. One of the main uses of clustering in health care is for microarray data analysis (Yoo, 2011). Genes with unknown functions might be identified if it is clustered together with some known genes.

2.6.3.5 Data Mining and Machine Learning

Data mining and machine learning are similar in the sense that both are trying to make use of big data, to learn something from it and make better decisions from it. However, while data mining might need some human intervention to make decisions, machine learning lets the algorithm do it. By providing a ‘training’ data set, which teaches the computer how to make sense of the data, the machine learning will be able to make predictions about a new data set. While data mining is simply looking for patterns that already exist in the data, machine learning goes beyond what has happened in the past to predict future outcomes based on the pre-existing data (Marr B, 2021).

2.6.4 Current Predictive Model for Breast Cancer

Several models have been developed to predict the risk of breast cancer over time (for example, 5-year, 10-year or lifetime risks). One of the earliest models is the Gail Model developed by Dr. Mitchell Gail and colleagues in 1989. This model considered age at menarche, age at first live birth, number of previous breast biopsies, benign breast disease, and number of first-degree relatives of breast cancer for its prediction (Gail et al., 1989). But studies indicated that this model might under-predict actual risk because of limited family history and because it does not include age of onset of cancer in the family (Amir et al, 2003; Jacobi et al, 2009; Quante et al, 2012).

Another well-known model is the Tyrer-Cuzick model, which is a well-studied, widely available model for predicting breast cancer risk. Besides all the variables used in Gail Model, it also incorporates second degree family history, age of onset of cancer and use of hormone replacement therapy (HRT), as well as the presence of BRCA gene mutations in its prediction (Tyrer et al, 2004). However, this model has its limitation, whereby it only accounts for hereditary breast and ovarian cancer (Vianna et al, 2019).

Due to the limitations of these models, some modifications have been made to further increase its prediction capability. Tice and Colleagues (2008) added the Breast Imaging Reporting and Data System (BI-RADS) classification to the Gail Model, and the C-statistics (measure of improvement of risk assessment), increased significantly. Then there's Shepherd and colleagues (2011) who used a technique called single x-ray absorptiometry and added log fibroglandular volume to the standard risk factor and the C-statistics increased significantly as well.

Besides clinical data, some researchers also incorporated single nucleotide polymorphisms (SNPs) to the models to increase its predictive capability. SNPs can be responsible for a large percentage of cancers in the population. Validated SNPs associated with breast cancer risk prediction is currently over 170, but there may be hundreds more (Mavaddat et al, 2019).

Wacholder et al (2010) added 10 SNPs to the Gail Model and the C-statistics increased significantly from 0.58 to 0.61. Then, in 2013 Dite and colleagues included seven SNPs and reported an increase from 0.58 to 0.61. In 2018, Wu et al, developed a statistical model to estimate the probability of breast cancer using diagnostic data from the Electronic Health Record (EHR) and observed predictive capability of 0.648.

Since most breast cancer deaths are due to its metastatic state, prediction of metastatic breast cancer would be a very beneficial tool. Previous studies had applied machine learning technology to predict breast cancer recurrence by using demographic, pathological, and genetic data. Tseng et al (2019) established models that can effectively predict breast cancer metastasis at least 3 months in advance by using eight clinical features including demographic data (age), pathological data (TNM stage, ER, PR, and HER2), and serum biomarkers (CA15-3, CEA, and sHER2). With 19 metastasis and 125 non-metastasis patients' data, the optimal prediction model for the test set was the model constructed using the random forest classifier. The area under the curve (AUC) value for this model was 0.75 ($p < 0.001$). The accuracy of this model optimised based on Youden Index was 0.75, with sensitivity 0.80, specificity 0.71, positive predictive value 0.36, and negative predictive value 0.96.

Meanwhile, Nicolo et al (2020) used random survival forest analysis to select a minimal set of five covariates with the best predictive power to predict MBC. Out of 21 of clinical and pathologic data that they investigated, the 5 most statistically significant variables are tumour size, age at diagnosis, and biomarkers of Ki67, EGFR and CD44. The model achieved a c-index of 0.65 (95% CI, 0.60 to 0.71) in cross-validation and had predictive performance similar to that of random survival forest (95% CI, 0.66 to 0.69) and Cox regression (95% CI, 0.62 to 0.67) as well as machine learning classification algorithms (Nicolo et al, 2020).

Even though some gene mutations have been associated with MBC, there were also concern on whether the mutation occurred due to therapy-induced mutations since a large proportion of breast cancer patients with early-stage tumours receive postoperative adjuvant therapy with hormonal therapy, chemotherapy, or both (Razavi et al, 2018). Razavi et al (2018) identified mutations in the MAPK pathway and the oestrogen receptor transcriptional program in 22% of hormone receptor-positive breast cancers after hormone therapy. These mutations are mutually exclusive with ESR1 (gene coding the oestrogen receptor) mutations and correlate with a shorter response duration to subsequent hormone therapies. Even though these mutations can be detected by genetic tests, but the high cost may be a limiting factor. Thus, having a high accuracy prediction model might solve this issue as we can possibly predict which patients are likely to have these mutations.

With comprehensive electronic medical records and advanced machine learning technologies, a breast cancer metastasis prediction model incorporating all three elements of clinical symptoms, genotype and phenotypic data might be a useful decision support tool for care intervention and increase the overall survival rate of metastatic breast cancer patients.

CHAPTER THREE
RESEARCH METHODOLOGY

3.1 Introduction

This chapter will be divided to three sections, which are: (i) data mining of genotype, phenotype and clinical data of metastatic breast cancer patients, (ii) construction of predictive model using genotype, phenotype and clinical data, as well as the (iii) validation using systematic review. Each section was meant to achieve each objective of the study.

3.2 Data Mining of Genotype, Phenotype and Clinical Data of Metastatic Breast Cancer (MBC) Patients

This part of the study addressed the first objective of this study, which is to mine and integrate clinical, phenotype and genotype data to elucidate the mechanism behind metastatic breast cancer. For this purpose, a few databases containing the data mentioned with regards to breast cancer patients were retrieved and analysed using unsupervised machine learning such as principal component analysis (PCA) and hierarchical clustering, as well as other statistical analyses like multiple correspondence analysis (MCA) and odds ratio. Figure 3.1 summarised the steps involved.

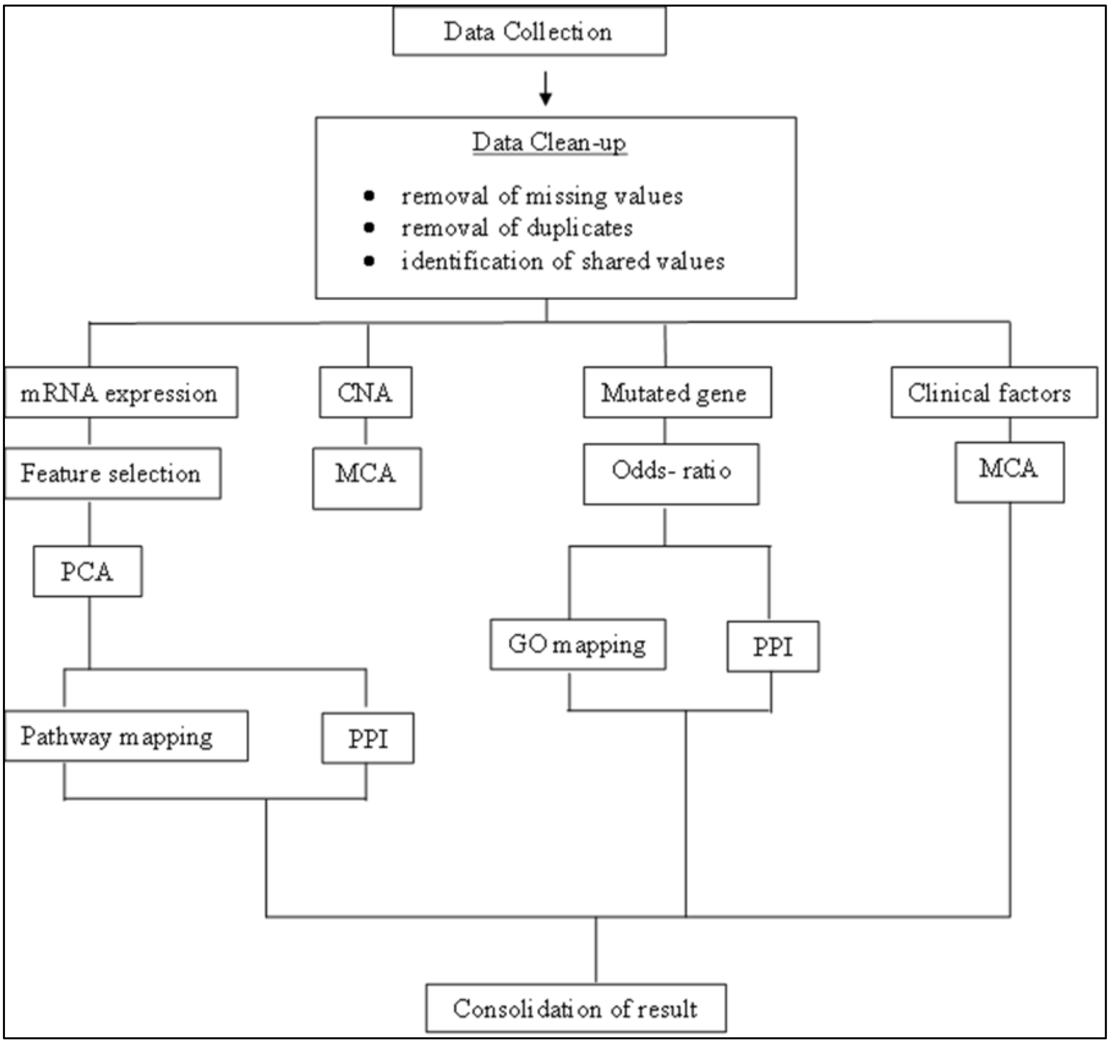


Figure 3.1Workflow of the Data Mining and Mapping Methodology. CNA stands for copy number alteration, while PCA stands for principal components analysis and MCA is multiple correspondence analysis. PPI stands for protein-protein interaction and GO stands for gene ontology

3.2.1 Datasets

Five datasets were utilised in this study. A dataset is an organised collection of data that can be in various formats, such as a table, a spreadsheet or a database (in this case a database) and the summary of these datasets can be found in Table 3.1. All the datasets were obtained from cBioPortal for Cancer Genomics (http://www.cbioportal.org/) and only data of female patients were used in this study. cBioportal is an open-access, open-source resource for interactive exploration of multidimensional cancer genomics data sets. It gave researchers access to molecular profiles and clinical attributes from large-scale cancer genomics projects, so that these datasets can be translated into biologic insights and clinical applications (Cerami et al, 2012).

From the five datasets that were obtained, the following data types were used for further analysis:

- mRNA gene expression
- Gene mutation
- Clinical profile of patients
- Copy number alteration (CNA)

Data types refer to the category or variable that the data represents. The breakdown for each of the data types and its analyses can be found in Table 3.2. These data were chosen because it contained both breast cancer and metastatic breast cancer data and it represents clinical and genotype data. Any duplicates found were removed; however, the same patients may have more than one sample tested. For these patients, information on the samples was retained if the samples were not from the same metastatic site. For example, if one of the patients has samples for both metastatic site lungs and lymph nodes, then both samples will be retained even though it is from the same patient. For clinical data, ER, PR and HER2 status of both primary and metastatic sites is considered important. Hence, if more than one of these values are missing, the patient's data were excluded.

The mRNA gene expression was presented in z-scores as it indicates the number of standard deviations away from the mean of expression in the reference. The Cancer Genome Atlas (TCGA) states that for mRNA and microRNA expression data, relative expression of an individual gene and tumour to the gene's expression distribution were computed in a reference population. That reference population is either all tumours that are diploid for the gene in question, or, when available, normal adjacent tissue. Although z-scores do not explicitly show the under- or over-expression of genes, it is suitable to be used for profiling (Cheadle et. al., 2003), such as this situation. This is also why this data set is chosen as the z-scores means the data has already undergone z-score normalisation which is suitable to be used in PCA as it helps to account for differences in gene expression magnitudes.

As for the copy number alteration (CNA), it is defined as copy number variations (CNVs), including duplication, amplification, deletion, and homozygous deletion, in a specific genomic region in somatic cells (Beroukhim et al, 2010). Genes were defined as amplified if the mean segment fold-change (ratio of normalised sequencing depth in tumour to normal) is more than 1.8 with adjusted p-value of less than 0.05. Meanwhile for homozygous deletions, a mean whole-gene fold-change of less than −2 or at least one exon within the affected gene having a fold-change of less than −2 were required, with both criteria requiring adjusted statistical significance of p-value less than 0.05 (Thiesen et al, 2017).

Table 3.1

Information on the Five Datasets that were Used in This Study. All of the data were obtained from cBioPortal (http://www.cbioportal.org/)

Dataset	Number of data	Data type(s) available	Reference
MSK 2018	Data of 1 756 patients (1 261 diagnosed with metastatic BC *vs* 495 Breast Cancer) and 1 918 samples (1 000 metastatic *vs* 918 primary samples)	• CNA • Gene Mutation	Razavi et al, 2018
INSERM	Data of 216 Metastatic Breast Cancer patients and 216 samples	• Gene Mutation • CNA	Lefebvre et al, 2016
The Metastatic Breast Cancer Project (MBCP)	Data for 180 Metastatic Breast Cancer patients and 237 Samples	• mRNA gene expression • Clinical • CNA • Mutation	Cerami et al, 2012
Metabric	Data of 2 509 Breast Cancer data and 2 509 samples	• mRNA gene expression • Clinical • CNA • Gene Mutation	Curtis et al, 2012 & Pereira et al, 2016
TCGA BRCA	Data of 1 084 patients	• Clinical	Hoadley et al, 2018

Table 3.2
Complete Information of the Data Types Used and the Analyses Performed on the Data Types

Data Type	Variables	Analysis performed	Comparison/ subgroups	Number of data points
mRNA gene expression	Expression of 16 374 genes, value is expressed as Z-score	• PCA and feature selection	• Gene expression in breast sample of Breast Cancer (BC) *vs* Metastatic Breast Cancer (MBC) patient	• 1 895 BC *vs* 29 MBC
			• Gene expression between primary vs metastatic sample	• 2 024 primary sample *vs* 26 metastatic sample
Copy number alteration (CNA)	Copy number alteration of 1 616 genes	• PCA and feature selection	• CNA of breast sample of BC *vs* MBC patient	• 2 668 BC *vs* 607 MBC
			• CNA between primary *vs* metastatic sample	• 3 275 primary sample *vs* 1 255 metastatic sample
Gene mutation		• Odds-ratio	• Mutations in breast sample of BC *vs* MBC patient	• 17 223 BC vs 20 607 MBC
			• Mutations between primary *vs* metastatic sample	• 34 946 primary vs 32 020 metastatic
Clinical profile	• Age • PR status of primary • ER status of primary • HER2 status of primary	• MCA and hierarchical clustering	• Clinical profile of BC vs MBC • patients (without OncoTree Code and Adjuvant Radiation variables)	• 2 100 BC *vs* 36 MBC

	- Chemotherapy (Yes/No) Hormone Therapy (Yes/No)		
	Same as the above but with the addition of the following variables - OncoTree Code - Adjuvant Radiation	- Clinical profile of BC vs MBC patient	- 1 936 BC *vs* 33 MBC
	- Age - PR status of primary - ER status of primary - HER2 status of primary - ER, PR and HER2 status of metastatic sample	- Sample profile of metastatic breast cancer patients	- 911 MBC patients

3.2.2 Data Mining and Statistical Analysis

3.2.2.1 Principal Component Analysis

Principal component analysis (PCA) is a method of reducing the dimensionality of robust datasets, increasing its interpretability while preserving as much variability and minimising information loss (Jolliffe and Cadima, 2016). This statistical technique creates new uncorrelated variables, the principal components, that successively maximise variance and these new variables are defined by the dataset we have. In this study, PCA was chosen since PCA helps mitigate challenges in analysing high dimensional data like the gene expression data. PCA finds patterns without reference to prior knowledge about whether the samples come from different treatment groups or have phenotypic differences (Lever et al, 2017). This analysis was carried out by using the scikit-learn package for principal component analysis in Python (Pedregosa et al, 2011). To ensure smooth running of the package and all its required dependencies, Anaconda Navigator (Version 1.9.12) was used.

Before PCA was carried out, a few assumptions were tested. For instance, PCA assumes that the relationships between variables (genes in the case of gene expression data) are linear and the variables are independent of each other. PCA is also sensitive to outliers and missing data, that is why pre-processing of the data were conducted before PCA. To check for linearity, multivariate regression was performed by using the 'lm()' function in R. With F-statistic: 1503, and p-value: < 2.2e-16, it was concluded that at least one independent variable in the model is significantly related to the dependent variable and we can conclude that the regression model as a whole is statistically significant. After making sure the data met the assumptions, PCA was then performed.

Given a data matrix, ***X*****,** of $n \times p$, where n is the number of rows of instances and p is the number of features, the principal component for each variable, $\boldsymbol{x}$, is calculated as the weighted average of the original variables. The matrix containing the principal components of the data is referred to as matrix $\boldsymbol{Y}$ and can thus be calculated as:

$$Y = W.\ X \tag{3.1}$$

where W is a matrix of coefficients that is obtained from the calculation of covariance, eigenvalues and eigenvector. Eigenvalues and eigenvectors are the linear algebra concepts that needed to be computed from the covariance matrix in order to determine the principal components of the data (Jaadi, 2021):

$$y_{ij} = w_{1i} x_{1j} + w_{2i} x_{2j} + \ldots + w_{pi} x_{pj} \tag{3.2}$$

The covariance between two variables, x_i and x_j can be calculated as:

$$Cov(xi, xj) = \frac{1}{n-1} \sum_{i=1}^{n} (xi - \overline{xi})(xj - \overline{xj}) \tag{3.3}$$

The eigenvalues and eigenvectors are then determined from the covariance matrix. The eigenvectors (principal components) determine the directions of the new feature space, and the eigenvalues determine their magnitude.

3.2.2.2 Multiple Correspondence Analysis (MCA)

Multiple correspondence analysis (MCA) is a method for summarising and visualising a data table containing more than two categorical variables. MCA was used in this study since it is considered as analogues to PCA for categorical variables (Mueller, 2019). In this study, MCA was used to analyse the CNA and clinical data. This analysis was done in R software using FactoMineR for the analysis and factoextra for data visualisation (Le et al, 2008). MCA is obtained by using a standard correspondence analysis on an indicator matrix (i.e., a matrix whose entries are 0 or 1) (Abdi and Valentin, 2007). MCA is used to represent and model datasets as “clouds” of points in a multidimensional Euclidean space; this means that it is distinctive in describing the patterns geometrically by locating each variable/unit of analysis as a point in a low- dimensional space (Costa et al, 2013). It can uncover a cluster of variable categories providing key insights on relationships between categories.

Besides determining the distribution of patients from categorical variables, the following analyses were also performed with the FactoMineR package such as:

i. associations between the variables

ii. association between the variable categories

Just like PCA, MCA also has some assumptions to be met. Firstly, MCA assumes that the categorical variables are mutually independent. MCA also assumes that the associations between categories are homogeneous across groups or subpopulations. Thirdly, MCA assumes that there are no empty cells in the contingency table formed by the variables and there is no ordinal relationship among the categories.

3.2.2.3 Feature Selection

Feature selection is the process of selecting features which contribute most to the prediction variable or output by reducing the number of input variables. Feature selection helps in improving the performance of the model and also reducing computational cost and training time (Shaikh, 2018). In this study, feature selection was employed to manage the 'curse of dimensionality' (Kuo and Sloan, 2005) associated with having a large amount of features or variables, particularly relating to the mRNA and CNA data types. The term curse of dimensionality was introduced by Bellman (Bellman, 1957) to describe the problem caused by the exponential increase in volume associated with adding extra dimensions to Euclidean space (Keogh et al, 2017). It basically refers to problems that could arise due to working with high-dimensional data. In addition, multicollinearity, or near-linear dependence, is a statistical phenomenon in which two or more predictors' variables in a multiple regression model are highly correlated (Daoud, 2017). However, PCA can effectively eliminate multicollinearity between features since it combines the highly correlated variables into a set of uncorrelated variables (Pramoditha, 2021). The feature selection used in this study was chi squared, statistical feature selection, forward and backward.

As Random Forest was used as the classification algorithm, feature importance was employed as the feature selection method. The average of each decision tree in Random Forest was pooled to build a final prediction. Feature importance can be calculated by dividing the number of samples that reach the node with the total number of samples. Feature importance was calculated in Python using the sklearn module (Brownlee, 2020). High score indicates the importance of the feature. Assuming only two child nodes or a binary tree, a nodes importance was calculated using Gini importance for each decision tree.

Nodes' importance was calculated as follow:

$$ni_j = w_jC_j - w_{left(j)}C_{left(j)} - w_{right(j)}C_{right(j)} \qquad (3.4)$$

where ni_j is the importance of node j, w_j is the weighted number of samples reaching node, C_j is the impurity value of node j, $_{left(j)}$ is the child node from left split on node j and $_{right(j)}$ is the child node from right split on node j.

The importance for each feature on a decision tree is then calculated as:

$$fi_j = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k} \qquad (3.5)$$

Where fi_j is the importance of feature i and ni_j is the importance of node j.

Then, these can be normalised to a value between 0 and 1 by dividing with the sum of all feature importance values:

$$normfi_j = \frac{fi_j}{\sum_{j \in all\ features} fi_j} \qquad (3.6)$$

The average of all the trees is the final feature importance. The sum of the feature's importance value on each tree was calculated and divided by the total number of trees:

$$RFfi_i \ \frac{\sum_{j \in all\ tress} normfi_{ij}}{T} \qquad (3.7)$$

Where $RFfi_i$ is the importance of feature i calculated from all trees in the Random Forest model, $normfi_{ij}$ is the normalised feature importance for i in tree j and T is the total number of trees.

Each feature will be assigned a value between 0 to 1, where the higher the value, the more important the feature is. The relative importance of a feature was calculated by comparing its value to the highest scoring feature. Therefore, the calculated importance is now:

$$feature\ importance_{\ i} = \frac{RFfi_i}{RFfi_{max}} \times 100 \qquad (3.8)$$

3.2.2.4 Odds Ratio

Odds ratio (OR) represents the probability that an outcome will occur given a particular exposure, compared to the probability of the outcome occurring in the absence of that exposure (Szumilas, 2010). In the context of gene mutations, it is employed to compare the frequency of mutations in two different groups. In this study, we measured OR to see the differences in mutation of breast cancer compared to metastatic breast cancer. Thus, the OR for the following were computed:

- Mutation of a particular gene in breast samples of metastatic breast cancer over breast cancer
- Mutations of a particular gene in metastatic samples over primary samples

The odds ratio for a certain gene, *g*, is calculated as such:

$$OR_g = \frac{(n_g^a/N^a)}{(n_g^b/N^b)} \tag{3.9}$$

where n_g^a is the frequency of mutation in group *a*, N^a is the total number of samples in group *a*, n_g^b is the frequency of mutation in group *b* and N^b is the total number of samples in group *b*. An OR of more than 1 indicates that the mutation is more frequently observed in group *a* compared to *b*. p-value and 95% confidence interval (CI) were calculated as indication of statistical significance for each gene. After calculating the OR value, confidence interval was calculated using the formula below:

$$\text{Upper 95\% CI} = \mathrm{e}\,[\ln(\mathrm{OR}) + 1.96\sqrt{}(1/\,\mathrm{nga} + 1/\,\mathrm{ngb} + 1/\,\mathrm{Na} + 1/\,\mathrm{Nb})] \tag{3.10}$$

$$\text{Lower 95\% CI} = \mathrm{e}\,[\ln(\mathrm{OR}) - 1.96\sqrt{}(1/\,\mathrm{nga} + 1/\,\mathrm{ngb} + 1/\,\mathrm{Na} + 1/\,\mathrm{Nb}) \tag{3.11}$$

3.2.2.5 Hierarchical Clustering

Hierarchical clustering is an unsupervised clustering algorithm which involves creating clusters that have predominant ordering from top to bottom. The endpoint is a set of clusters or groups, that is represented in a dendrogram. A dendrogram is a type of tree diagram showing hierarchical relationships between different sets of data. It has clades which are branches that are arranged according to how similar the variables are. Each clade has leaves and the distance between each leave represent similarities while the length of the clades represents the distance between clusters.

In this study, hierarchical clustering was used to cluster the variable categories of clinical factors and the dendrogram was used to visualise the distance between the variable categories better. Hierarchical clustering was performed in R, using the ‘hclust’ method with the default algorithm which is ‘complete’ and the default distance measure which is the square root of the sum of the square differences or known as Euclidean distance.

3.2.3 Similarities of Metastatic Sites Based on Genes

Given that the metastatic samples used in this study were from several sites with some sites being underrepresented, the similarities between the different sites based on their genetic profile were analysed.

Top 10 metastatic sites (phenotypes) that were observed from the metastatic sample data were determined. The top 20 mutated genes for each of the sites were obtained from Genomic Data Commons (GDC) portal at https://portal.gdc.cancer.gov. The portal is a robust data-driven platform that allows users to search and download cancer data for analysis. The similarity of the metastatic sites based on their gene profile was calculated using BOG method and Sun’s Annotation-based measure. Both are frequently used to calculate disease-disease similarity. The similarities were then visualised using a heat map, which was constructed in R, using the ‘heatmap’ function. The following subsections explain the BOG method and Sun’s Annotation -based measure, which was done in R using the *dSimer* package (Li & Ni, 2018).

3.2.3.1 BOG Method

BOG is a method used for analysing disease similarity. Its similarities are calculated using disease-gene associations. Given two vectors of diseases and a list of disease-gene associations, this function will calculate disease similarity (Peng and Min, 2015). This was done by finding over-represented extracted Disease Ontology (DO) terms using the hypergeometric distribution and the Benjamini-Hochberg correction for multiple tests (Mathur and Dinakarpandian, 2010). To account for random or rare occurrences, a similarity metric called BV (Mathur and Dinakarpadian, 2007) that is based on both co-annotation and hierarchy (Equations 3.12 and 3.13) was used. A p-value was calculated for similarity scores using 100,000 randomly generated pairs of diseases.

Given DO terms A & B (where A & B are disease vectors), n(A) = number of genes annotated with A, n(A ∩ B) = number of genes annotated with both A and B, and N = total genes, similarity is given by:

$$sim(A,B) = \frac{\frac{n(A \cap B)}{n(A \cup B)}}{n\frac{A}{N} \cdot \frac{nB}{N}} \qquad (3.12)$$

The value obtained is normalised by the average of the maximum scores for A and B, and multiplied by the average surprisal of the terms as follows:

$$Score(A,B) = \frac{sim\ (A,B)}{\max\ \_sim(A,i) + \frac{\max\ \_sim(B,j)}{2 + sup(B)}} \cdot Avg(Sup\ A) \qquad (3.13)$$

Max_sim(A,i) is the maximum similarity score for DO terms A and ‘i.’ Sup(A) is the surprisal of A.

3.2.3.2 Sun’s Annotation-based Measure

Sun’s annotation-based measure is another method of calculating disease-disease similarities by using the disease-gene association data. This method was also used in this study to calculate the similarities of mutated genes from other cancers to the frequently mutated genes in breast cancer. A known standard method for comparing the similarity between two sets, the Jaccard index was used to estimate the similarity score between diseases as follows (Sun et al, 2014). Let G_{Di} be the set of genes associated with a disease D_i. The annotation-based similarity score of two diseases D_i and D_j were computed as the Jaccard index (or Jaccard similarity coefficient) of GD_i and GD_j:

$$sim_{annotation}(D_{i,}D_j) = \frac{|GD_i \cap GD_j|}{|GD_i \cup GD_j|} \qquad (3.14)$$

3.2.4 Protein-Protein Interaction Prediction

Protein-Protein Interaction (PPI) prediction using STRING was employed to see whether two proteins may interact. STRING measures both direct (physical) and indirect (functional) interactions between two proteins, based on experimental data of protein-protein interactions (Szklarczyk., 2018).

A score is provided for each protein-protein association. The scores represent confidence scores, ranging from 0 to 1, indicating estimated likelihood that the association is biologically significant, given the supporting evidence. The supporting evidence is based on seven factors, which are neighbourhood in genome, gene fusions, cooccurrence across genomes, co-expression, experimental/biochemical data, association in curated databases and co-mentioned in PubMed abstracts. Based on the seven factors, a combined and final confidence score is computed. A good interaction should not only have a high combined score, but also having more than one factor contributing to the score. In the visualisation of PPI in STRING, proteins are represented as nodes and an edge connects the two nodes if interaction is predicted. The edge is colour coded where the colours indicate different evidence. The red edge indicates the presence of fusion evidence, green edge indicates neighbourhood evidence, blue edge is the cooccurrence evidence, purple edge represents experimental evidence, yellow edge indicates text mining evidence, light blue edge indicates database evidence, while the black edge represents co-expression evidence.

In this study, PPI was conducted to elucidate the interactions between the protein corresponding to each investigated genes with oestrogen receptor, progesterone receptor and HER2.

3.2.5 Pathway Analyses

To see which pathways the important mRNA genes are a member of, the genes were mapped to pathways in the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database (https://www.genome.jp/kegg/pathway.html). KEGG is a database resource that provides all knowledge about genomes and their relationships to biological systems such as cells and whole organisms as well as their interactions with the environment (Aoki-Kinoshita and Kanehisa, 2007).

To curate potential pathway for the occurrence of metastatic breast cancer, SIGNOR2.0 (SIGnaling Network Open Resource) was employed (Licata, 2019). SIGNOR2.0 contains more than 23 000 manually annotated causal relationships between proteins. SIGNOR2.0 calculate the causal relationship between two proteins, or between a protein and a phenotype from the following factors: (i) number of annotated articles reporting the interaction, (ii) number of occurrences of specific relationship in pathways annotated in SIGNOR2.0, (iii) number of occurrences of the pair from Reactome database and (iv) whether the pair are mentioned in each protein's UniProtKB page. SIGNOR2.0 will also annotate the type of relationship between two proteins (e.g. up-regulation, down-regulation, direct, indirect), as well as mechanism underlying the relationship (e.g. phosphorylation, ubiquitination).

In this study, SIGNOR 2.0 was applied to elucidate the potential pathway of MBC occurrence based on factors determined from the previous results.

3.2.6 GO Mapping

The top 20 mutated genes from breast samples of MBC patients and from metastatic samples were mapped to GO Biological Process using Cytoscape, through the ClueGO application. By using visual style group functional analysis, with medium network specificity, redundant groups with more than 50% overlap were merged and mapped.

3.3 Construction of Predictive Model Using Genotype, Phenotype and Clinical Data

This part of the study aimed to build a predictive model that can predict metastatic status of breast cancer patients based on factors determined from data mining. Figure 3.2 showed the workflow for the construction of prediction models where two sets of different data were obtained for internal and external validation.

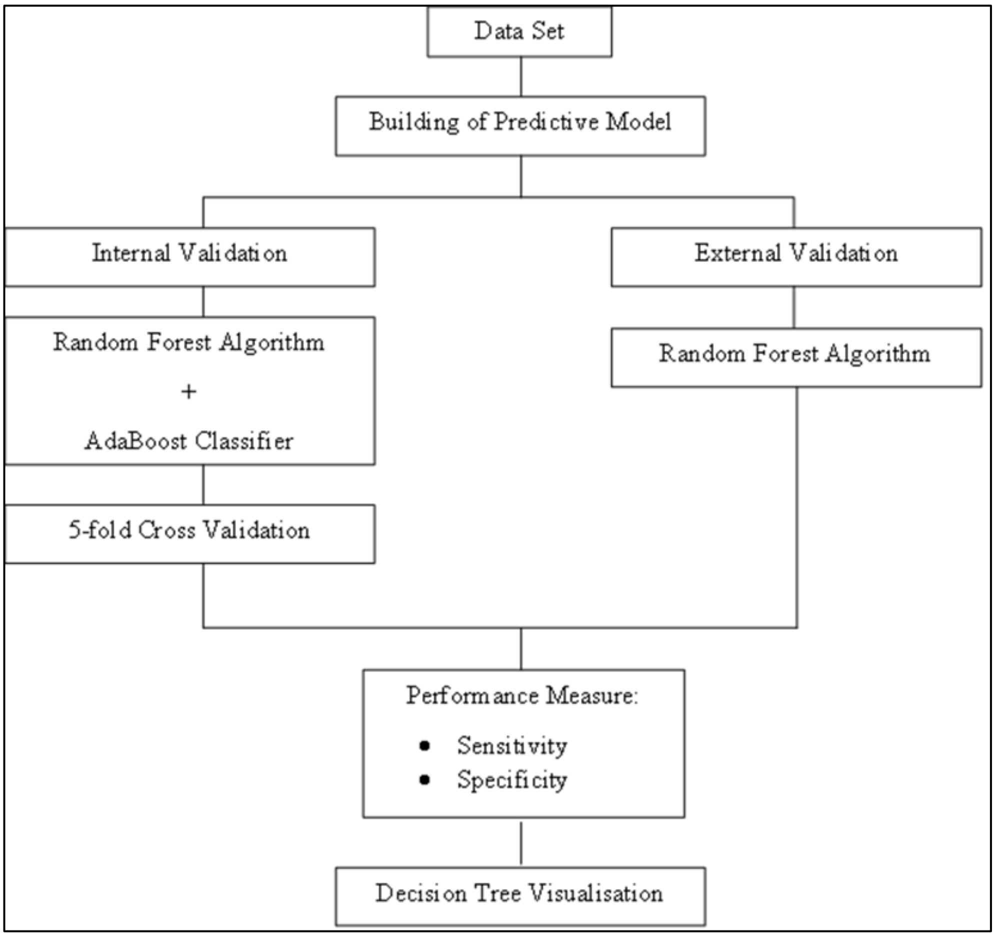


Figure 3.2 Workflow of the Method

3.3.1 Training Set

From the data mining section, the data that showed clear separation (seen as less overlap) between breast cancer and metastatic breast cancer patients were used to construct a predictive model. The details of the training sets are shown in Table 3.3 below. It should be noted that for training set labelled mRNA_15, the data were divided into six-fold, where one- fold was hold off for external validation while the rest were used for internal validation.

Table 3.3
Details of the Training Set Used to Build the Prediction Models

Training set	List of Variables	Description	Number of data points
Clinical_7	Age ER PR HER2 Chemotherapy Hormonetherapy AdjuvantRadiation OncoTree Code	This training set contains two groups, breast cancer and metastatic breast cancer patients.	There are 1 936 breast cancer patients and 33 metastatic patients in this dataset
Clinical_5	Age ER PR HER2 Chemotherapy Hormonetherapy	This training set contains two groups, breast cancer and metastatic breast cancer patients but without the variables OncoTree code and Adjuvant radiation seen in Clinical_7 data set	There are 2 100 breast cancer patients and 36 metastatic patients in this dataset
mRNA_15	DBIL5P2 FGF4 KRT76 KRTAP25-1 LINC00943 LINC01091 LINC01107 MAGEA9B MT4 OR5J2 OR5T2 OR9G4 SCGB1D1 TMEM207 TUSC7	This training set contains two groups, which are breast cancer and metastatic breast cancer patients. The variables are genes that were identified to be important based on mRNA profiling from the previous chapter.	There are 1 895 breast cancer patients and 129 metastatic patients in this dataset

mRNA_7	FGF4 MT4 OR5T2 KRT76 OR9G4 SCGB1D1 OR5J2	This training set was derived from mRNA_15 but with only 7 out of the 15 genes used. This is due to the difficulty in finding an external data set that has the complete 15 genes.	There are 1 894 breast cancer patients and 129 metastatic patients in this dataset
CNA	ARID5B CCND1 CDKN1B CTCF DAXX ERBB2 FGF3 FGF4 FGFR1 FH FOXA1 GPS2 HIST3H3 MAP2K4 MCL1 MYC NBN PAK1 PARP1 PDPK1 PLCG2	This training set contains two groups, which are breast cancer and metastatic breast cancer patients. The variables are genes that were identified to be important based on CNA profiling from the previous chapter.	There are 2 668 breast cancer patients and 607 metastatic patients in this dataset

RAD51C
RARA
RECQL4
SPOP
TCEB1
WHSC1L1
ZFHX3

3.3.2 Random Forest Algorithm

Random Forest algorithm is a supervised classification algorithm (Polamuri, 2017). As the name suggests, this algorithm will create a forest with a lot of trees (also called the decision trees). The trees were built using training sets consisting of multiple feature or characteristics for each of the instance in the training set. Then, output results were produced from the variables of the training set of interest. The result was obtained by aggregating all the outputs from different trees. The greater number of trees in the forest will lead to higher accuracy results (Polamuri, 2017).

There are two stages in Random Forest which are: (1) random forest creation and (2) prediction from the random forest classifier created in the first stage.

Firstly, the algorithm built m amount of decision trees. Each of the decision trees were initiated with a single node (denoted as the top of each tree in Figure 3.3) where a number of randomly selected samples served as the data set. Then, a bootstrap sample of n number of variables of the training data were drawn and selected at random.

From the random selected subset, the variable that provides the best split, measured using the Gini index, will split the node into two daughter nodes, specifying possible outcomes. The tree was further split until a maximum size is reached without pruning.

Gini index (S) is calculated as follow:

$$Gini\ (S)\ =\ 1 - \sum_{j}^{2} P \qquad (3.15)$$

where P_j is the relative frequency of class j in S. Each time, the split then was divided into two subsets of S_1 and S_2 in which gini (S) data was divided into:

$$Ginisplit\ (S) = \frac{n1}{n}\ gini\ (S1) + \frac{n2}{n}\ gini\ (S2) \qquad (3.16)$$

This process will repeat until the tree has reached the specified number of branches, in this case nodes are expanded, and a path was established (indicated by orange coloured node in Figure 3.3). As illustrated in Figure 3.3, the progression of the branches was gradually expanded until all training data were assigned to a terminal leaf node which is represented by green and red square boxes. At the end of the tree, class probability was calculated. The outcome can be calculated as either the mean of the class probability from each decision trees or the highest votes.

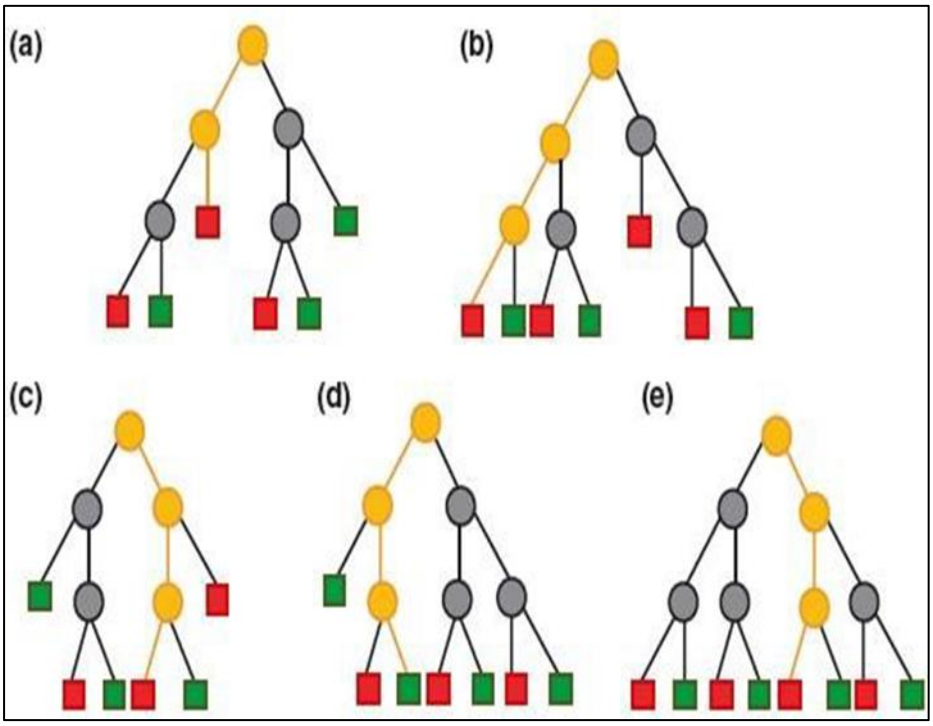


Figure 3.3 Example of Five Illustrative Trees of Random Forest Learning Algorithm. The terminal leaf nodes are shown as squares and coloured red or green according to class of interest. The path taken through each tree is shown in orange. Trees (a), (b), (c), and (e) predict that the object belongs to the red class, meanwhile tree (d) dissenting, so that the Random Forest will assign it to the red class by a 4:1 majority vote

In this study, Random Forest classifier was used to determine which features are the most important in predicting the occurrence of MBC (based on previous results.) Beside Random Forest classifier, an additional classifier called AdaBoost classifier was also incorporated in this study (Pedregosa et al, 2011). The core principle of AdaBoost is to fit a series of weak learners (i.e., models that are only slightly better than random guessing, such as small decision trees) on repeatedly modified versions of the data. All of the predictions are then combined through a weighted majority vote (or sum) to produce the final prediction (Pedregosa et al, 2011).

3.3.3 Validation

3.3.3.1 Internal Validation

For validation of the predictive models, *k*-fold cross validation procedure was conducted. Cross-validation is a statistical method used to estimate the skill of machine learning models. It is commonly used in applied machine learning to compare and select a model for a given predictive modelling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods (Brownlee, 2018).

In *k*-fold cross-validation, the available learning set is partitioned into *k* disjoint subsets of approximately equal size. Here, fold refers to the number of resulting subsets. This partitioning is performed by randomly sampling cases from the learning set without replacement. The model is trained using *k–1* subset, which, together, represent the training set. Then, the model is applied to the remaining subset, which is denoted as the validation set, and the performance is measured. This procedure is repeated until each of the *k* subsets has served as validation set. The average of the *k* performance measurements on the *k* validation sets is the cross-validated performance. Figure 3.4 illustrates this process for $k = 10$, i.e., 10-fold cross-validation. In the first fold, the first subset serves as validation set and the remaining nine subsets serve as training set. In the second fold, the second subset is the validation set and the remaining subsets are the training set, and so on. Every data point gets to be in a validation set exactly once and gets to be in a training set *k-1* times.

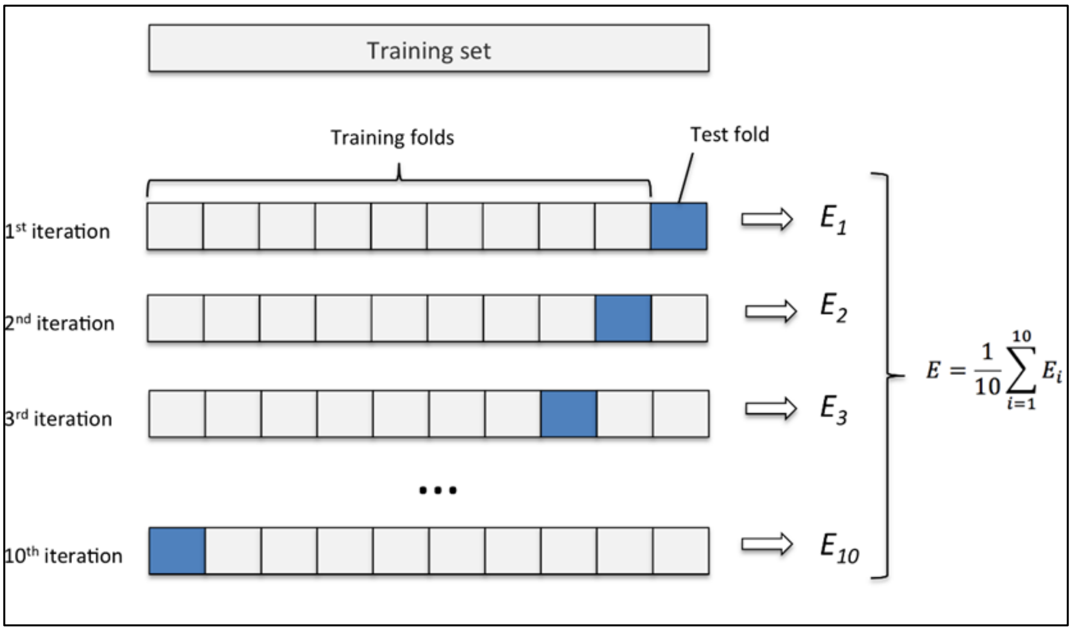


Figure 3.4 10-fold Cross-validation. The data set is randomly split into ten disjoint subsets, each containing (approximately) 10% of the data. The model is trained on the training set and then applied to the validation set (Kumar, 2020)

3.3.3.2 External Validation

External validation is performed to evaluate the ability of the model to predict outcome for instances that they have never encountered, or outside of the training set. The Cancer Genome Atlas (TCGA) Breast Invasive Carcinoma Project dataset was used as the external data set for Clinical_5 and mRNA_7. These data were initially screened to ensure that it is not in the training set. The external data set used for each training set can be found in Table 3.4 below.

Table 3.4
Details of the External Data Set Used for External Validation of Each Training Set

Training set	Details
Clinical_7	No external data set containing all 8 variables could be found, hence external validation was not performed for this training set.
Clinical_5	External data set contain data of 81 breast cancer patients and 2 metastatic breast cancer patients
mRNA_15	As previously mentioned, the training set was divided into six folds where one-fold was hold of as external data set. It contained 102 metastatic breast cancer patients and 1 584 non-metastatic breast cancer patients
mRNA_7	External data set contain data of 80 breast cancer patients and 2 metastatic breast cancer patients
CNA	Not performed

3.3.4 Performance Measure

Performance measure evaluates the ability of the predictive model to accurately predict the class label and this was measured through calculating sensitivity and specificity of the models.

3.3.4.1 Sensitivity

Sensitivity or true positive rate (TPR) measures the proportion of actual positives that are correctly identified. The calculation of sensitivity is as follows:

$$\text{Sensitivity, TPR} = \frac{TP}{TP + FN} \qquad (3.17)$$

where TP is true positive; FN is number of false negative.

3.3.4.2 Specificity

Specificity or true negative rate (TNR) measures the proportion of actual negatives that are correctly identified. The calculation of specificity is as follows:

$$\text{Specificity, TNR} = \frac{TN}{TN + FP} \qquad (3.18)$$

where TN is true negative; FP is number of false positive. The prediction model is deemed acceptable when the value of specificity and sensitivity are at least 50% (0.5) (Power et al, 2013).

3.4 Validation using Systematic Review

For benchmarking of the results, the initial plan was to use datasets from the local hospital to validate our models. However, due to limitations of the COVID-19 outbreak, the benchmarking had to be improvised. Therefore, a systematic review on the treatments of metastatic breast cancer were conducted to further validate the previous clinical findings. By going through research from 20 years ago up until now, it is postulated that some of the factors that were found to be associated with metastatic breast cancer in the previous sections can also be found in literatures.

3.4.1 Search Strategy for Identification of Studies

A search was conducted on 9th December 2020 in PubMed, Scopus, Web of Science and Science Direct. Keywords used in the search strategy were "metastatic breast cancer, chemotherapy, hormonal therapy, targeted therapy, gene and progression free survival". The keyword "metastatic breast cancer" was first utilised to find studies focused on MBC. Then the search was narrowed down to "chemotherapy, hormonal therapy and targeted therapy" to eliminate other interventions such as surgery and radiotherapy. Then the keyword "gene" was inserted within these searches to see whether there were studies that connected any gene with the efficacies of these treatments. Finally, "progression free survival" was added to make sure this endpoint was mentioned in each article reviewed since the discussion will be based on this endpoint.

The search strategy was developed using Boolean operators and slightly differed for each database based on their respective formats, as shown in Table 3.5.

Table 3.5
Search String for Databases

Database	Search string
Scopus	(TITLE-ABS-KEY ("metastatic breast cancer")) AND (((hormon* OR chemotherapy OR "targeted AND therapy")) AND (gene*)) AND ("progress* free survival") AND (LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2015) OR LIMIT-TO (PUBYEAR , 2014) OR LIMIT-TO (PUBYEAR , 2013) OR LIMIT-TO (PUBYEAR , 2012) OR LIMIT-TO (PUBYEAR , 2011) OR LIMIT-TO (PUBYEAR , 2010) OR LIMIT-TO (PUBYEAR , 2009) OR LIMIT-TO (PUBYEAR , 2008) OR LIMIT-TO (PUBYEAR , 2007) OR LIMIT-TO (PUBYEAR , 2006) OR LIMIT-TO (PUBYEAR ,

	2005) OR LIMIT-TO (PUBYEAR , 2004) OR LIMIT-TO (PUBYEAR , 2003) OR LIMIT-TO (PUBYEAR , 2002) OR LIMIT-TO (PUBYEAR , 2001)) AND (LIMIT-TO (DOCTYPE , "ar"))
Web of science	"metastatic breast cancer" Refined by: TOPIC: (Hormon* OR chemotherapy OR “targeted therapy”) AND TOPIC: (gene*) AND TOPIC: (progres* free survival) AND PUBLICATION YEARS: (2020 OR 2019 OR 2018 OR 2017 OR 2016 OR 2015 OR 2014 OR 2013 OR 2012 OR 2011 OR 2010 OR 2009 OR 2008 OR 2007 OR 2006 OR 2005 OR 2004 OR 2003 OR 2002 OR 2001) AND DOCUMENT TYPES: (ARTICLE)
Science Direct	"metastatic breast cancer" AND (hormone OR chemotherapy OR "targeted therapy") AND (gene) AND ("progression free survival") Refine by Publication Years: (2001-2020) AND Refine by Document Types: (Articles) AND Refine by: (Subscribed Journals)
Pubmed	"metastatic breast cancer" AND (hormon* OR chemotherapy OR "targeted therapy") AND (gene*) AND ("progression free survival") AND Limit to: (Full Text) AND (English Language) AND Publication Year (January 2001 to December 2020) AND Source Type: (Academic Journals)
Cochrane	"metastatic breast cancer" AND (hormon* OR chemotherapy OR "targeted therapy") AND (gene*) AND ("progression free survival") in Title Abstract Keyword

3.4.2 Study Selection

3.4.2.1 Inclusion Criteria

Randomised controlled trial (RCT) on the treatments of metastatic breast cancer patients from year 2001 till year 2020 were included in this review. Criteria chosen were based on different treatments of metastatic breast cancer which are chemotherapy, hormonal therapy and/or targeted therapy, while the outcome measured were progression free survival and overall survival.

3.4.2.2 Exclusion Criteria

For this review, studies with male patients, other treatments such as surgery and radiotherapy, any non-English literature and unavailable full text were excluded. Non-randomised controlled trials, review articles, book chapters and proceedings were also excluded.

3.4.3 Data Extraction

Titles and abstracts of each paper were screened according to the inclusion and exclusion criteria. The selected citations were then downloaded into EndNote^{TM} x8.2. The citations were organised, and the duplicates were removed to avoid redundancy. Information extracted from each article included study design, median age of participants, publication year, types of metastatic breast cancer treatments as well as the outcome measured which are progression free survival and overall survival.

3.4.4 Quality Assessment

The papers were screened independently by two independent researchers using the Jadad Scale (Jadad et al, 1996) as shown in Appendix 1. The Jadad Scale was used because it presented the best validity evidence and has been tested for reliability in different research areas such as medicine, dentistry, psychology, and physical therapy (Olivo et al, 2008). The scores were based on randomisation, blinding, as well as the withdrawals and dropouts of the participants for each study. A score of 3 and above means these three criteria were adequately mentioned in the research paper, thus accepted as a reliable study. When a consensus was not met, a third reviewer was consulted.

CHAPTER FOUR
RESULTS AND DISCUSSION

4.1 Introduction

This chapter is organised as such: Section 4.2 will address the (i) results and discussion of mRNA profile, followed by (ii) CNA profile, (iii) genetic mutations, (iv) clinical profile and (v) pathway curation to analyse whether there is a correlation between the mRNA expression and clinical profile. Section 4.3 discusses on construction of predictive models: by first discussing on (i) internal validation, followed by (ii) external validation and (iii) visualisation of decision trees. Meanwhile, section 4.4 is on validation by systematic review.

4.2 Data Mining of Genotype, Phenotype and Clinical Data of Metastatic Breast Cancer (MBC) Patients

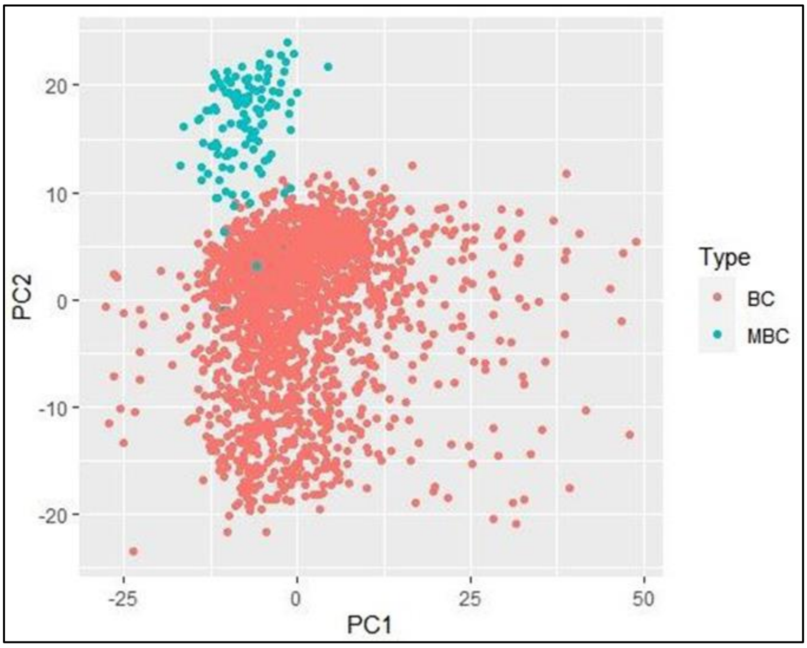
4.2.1 mRNA Profile of Breast Samples of Breast Cancer *vs* Metastatic Breast Cancer

Figure 4.1 shows the Principal Component Analysis (PCA) plot of mRNA expression of breast samples of breast cancer (BC) and metastatic breast cancer (MBC) patients, using z-scores. PCA was applied after feature selection, to avoid the curse of dimensionality. The total number of PC generated from this analysis was 2, with eigenvalues for PC1 was 20482304 and eigenvalues of PC2 was 38.13347. As for the percentage of variance, PC1 accounts for 100% of the variance, while PC2 does not contribute significantly to the variance. However, for the sake of only 2 PCs produced, both PCs were used for the subsequent analysis. In the PCA plot, principal component 1 (PC1) was plotted against principal component 2 (PC2) as the first two PCs contain the most information. Figure 4.1 (A) is when features with a score of above 0 were retained, where 1 616 out of 16 374 genes were used to calculate the principal components and later, the PC1 and PC2 was plotted. However, 1,616 genes were a bit much to analyse, thus we narrowed it down to a feature selection score of above 30 which carries 70% of the feature importance with only 15 genes involved. Figure 4.1

(B) is when features with a score of above 30 were retained, where 15 out of 16 374 genes were used to construct the PCA. Since better separation was seen between BC and MBC patients with these 15 genes, thus these 15 genes were retained for further analysis. The information of the 15 genes that were used in calculating the principal components can be found in Table 4.1, the membership of the genes in KEGG pathways can be found in Figure 4.2 and PPI interactions between the proteins can be found in Figure 4.3.

Several observations can be made in regard to the mRNA expression profile of breast samples of BC and MBC patients. First, it can be seen that by first identifying important features through feature selection, the data avoided the curse of dimensionality as well as showing a good separation between the two groups, BC and MBC (see Figure 4.1 (A)). By further reducing the variables to 15, where the feature importance was more than 30, the separation between the two groups was clearer (see Figure 4.1 (B)).

A



B

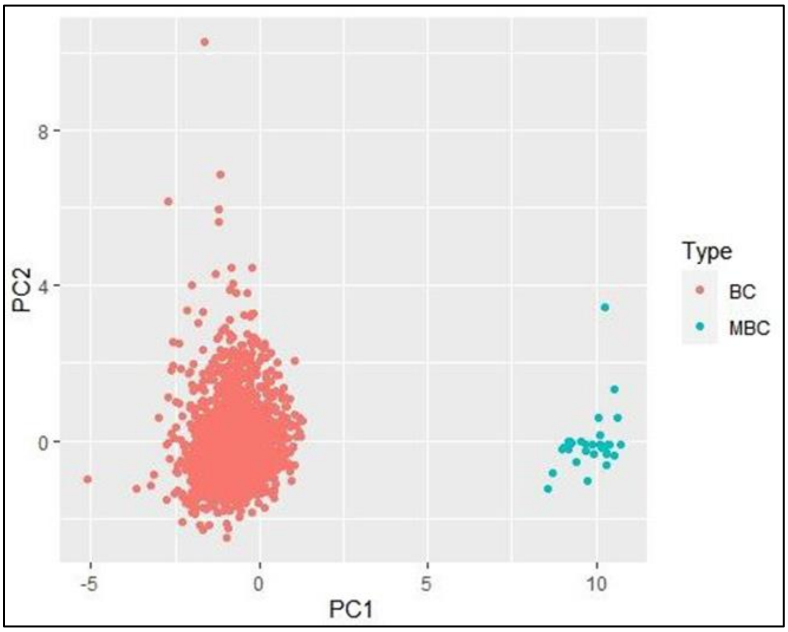


Figure 4.1 The PCA Plot of mRNA Expression Profile of Breast Samples of Breast Cancer (BC) and Metastatic Breast Cancer Patients (MBC) where (A) is when features with a score of above 0 were used to calculate the principal components and (B) is when features with a score of 30 and above were used to calculate the principal components. PC1 refers to principal component 1 (x-axis) while PC2 refers to principal component 2 (y-axis)

Secondly, three of the 15 genes can be connected to breast cancer based on literature review (see Table 4.1). These include Melanoma Antigen Family A, 9B (MAGEA9B) and Fibroblast Growth Factor (FGF-4), where MAGEA9B is pro-tumourigenic and have been reported to affect the biological characteristic of cancer cells such as migration, metastasis and invasion. A study by Parish et al (2015), showed that 56 out of 139 patients showed aberrant FGF receptor (FGFR) or FGF ligand (FGF) genes, where the FGF/FGFR aberrations were most frequently found in patients with breast cancer. Furthermore, aberrations in FGF/FGFR were also correlated with a diagnosis of breast cancer and liver metastasis.

Table 4.1
List and Details of the 15 Genes Used to Calculate the Principal Components where there was clear separation between breast samples of breast cancer and metastatic breast cancer. FI is feature importance

Gene symbol	Gene name	FI	Description
TUSC7	Tumour Suppressor Candidate 7	100.0	TUSC7 is an RNA gene and was initially discovered to suppress tumour in osteosarcoma (Pasic et al, 2010). Although several studies have reported that TUSC7 expression is lower in cancer tissues compared to normal tissues (Cong et a l, 2016; Ding et al, 2014), its role in cancer pathophysiology is far from being fully elucidated. A meta-analysis by Shi et al (2017), highlighted that low expression of TUSC7 is associated with poor survival. Additional, bioinformatics analysis suggests that TUSC7 is involved in protein ubiquitin pathway, which leads to protein degradation.
MAGEA9B	Melanoma Antigen Family A, 9B	79.2	The MAGE-A family is pro-tumourigenic and have been reported to affect the biological characteristic of cancer cells such as migration, metastasis and invasion. MAGE-A2 have been shown to induce chemoresistance in breast cancer cells by inhibiting the transactivation of p53-responsive genes, which are involved in cell cycle arrest and apoptosis in response to tamoxifen. MAGE-A2 can also form a complex with oestrogen receptor (ER)-α to enhance transcriptional activity (Wong et al, 2014). A meta-analysis done by Poojari et al (2020), showed that overexpression of MAGEA9 showed prognostic association in breast cancer.
FKRTAP25-1	Keratin Associated Protein 25-1	75.3	Keratins have been extensively used in diagnostic tumour pathology as immunohistochemical markers (Miettinen and Fetsch, 2000). Epithelial keratins such as K8, K18 and K19 are expressed in most adenocarcinomas. Although no studies have been found on KRTAP25-1 and breast cancer,

			K17 expression in breast tumours have been shown to have poor clinical outcome, independent of tumour grade and size (van de Rijn et al, 2002). Furthermore, reduced overall survival in ER-negative, triple negative and HER2-positive breast tumours have been associated with K19 expression (Ignatiadis et al, 2007).
KRT76	Keratin 76	59.0	No studies have shown the association of KRT76 with breast cancer or metastasis.
LINC00943	Long Intergenic Non-Protein Coding RNA 943	54.1	Limited studies available on LINC00943, although Zhu et al (2020), showed a possible involvement of LINC00943 in the tumourigenesis of melanoma.
SCGB1D1	Secretoglobin Family 1D Member 1	52.1	The role of the SCGB1D family is largely unknown although Heinonen et al (2015), have demonstrated the up- regulation of SCGB1D1 expression in HOXB7-knockdown cells.
MT4	Metallothionein 4	48.1	No studies have been found that associates MT4 with breast cancer or metastasis. However, overexpression of MT2A has been demonstrated in cell proliferation and invasion of breast cancer cells (Hyung et al, 2011).
DBIL5P2	Diazepam Binding Inhibitor-Like 5 Pseudogene 2	43.6	No studies found on the association between DBIL5P2 and breast cancer or metastasis.
OR9G4	Olfactory Receptor Family 9	41.0	Olfactory receptors (ORs) belong to the G protein-coupled receptors family where it is expressed mainly in the olfactory sensory neurons. Several studies have reported on the potential of some ORs

	Subfamily G Member 4		to be tumour markers. Sanz et al (2014), demonstrated that ORs activate PI3Kγ signalling pathway, consequently activating cell invasiveness. Li et al (2019), demonstrated that OR2T6 family initiates MAP/ERK pathway and Epithelial-Mesenchymal Transition, which shows its involvement in breast cancer progression. However, the involvement of OR9G4 in breast cancer or metastasis has not been found in the literature.
LINC01107	Long Intergenic Non-Protein Coding RNA 1107	39.7	No studies found on the association between LINC01107and breast cancer or metastasis.
FGF4	Fibroblast Growth Factor 4	37.9	A study by Parish et al (2015), showed that 56 out of 139 patients showed aberrant FGF receptor (FGFR) or FGF ligand (FGF) genes, where the FGF/FGFR aberrations were most frequently found in patients with breast cancer. Furthermore, aberrations in FGF/FGFR were also correlated with a diagnosis of breast cancer and liver metastasis.
TMEM207	Transmembrane Protein 207	36.5	TMEM207 expression was found to be downregulated in clear cell renal cell carcinoma but no studies were found in breast cancer cells (Wrzesinski et al, 2015). However, high expression of TMEM45A was seen in breast cancer, which suggests that TMEM45A as a potential biomarker (Flamant et al, 2012).
OR5J2	Olfactory Receptor Family 5 Subfamily J	34.4	ORs and OR2T6 have been demonstrated to play a role in breast cancer pathogenesis (see above) but the involvement of OR5J2 is not currently available.

	Member 2		
OR5T2	Olfactory Receptor Family 5 Subfamily T Member 2	34.2	ORs and OR2T6 have been demonstrated to play a role in breast cancer pathogenesis (see above) but the involvement of OR5T2 is not currently available.
LINC01091	Long Intergenic Non-Protein Coding RNA 1091	30.2	No studies found on the association between LINC01091 and breast cancer or metastasis.

When these 15 genes were mapped to KEGG pathway (see Figure 4.2), two genes were shown to be involved in key cancer signalling pathways such as MAPK and PI3K-Akt pathways. When growth factor receptors on the cell membrane such as HER2 are stimulated, this leads to the activation of two different but connected pathways, which are the mitogen activated protein kinase (MAPK) and PI3K-Akt pathways where the former promotes proliferation and invasion and the latter is involved in anabolism. Under normal circumstances, the pathways are highly regulated but in cancer, it is deregulated (Burotto, 2014). The MAPK pathway produces pro-oncogenic effects, however, in certain cancers it can produce tumour suppressor effects (Burotto, 2014). Activation of the Akt pathway in breast tumour samples have been associated with resistance to endocrine therapy, and activation of MAPK pathway was observed in tamoxifen and fulvestrant-resistant breast cancer cells (Ghayad et al, 2010).

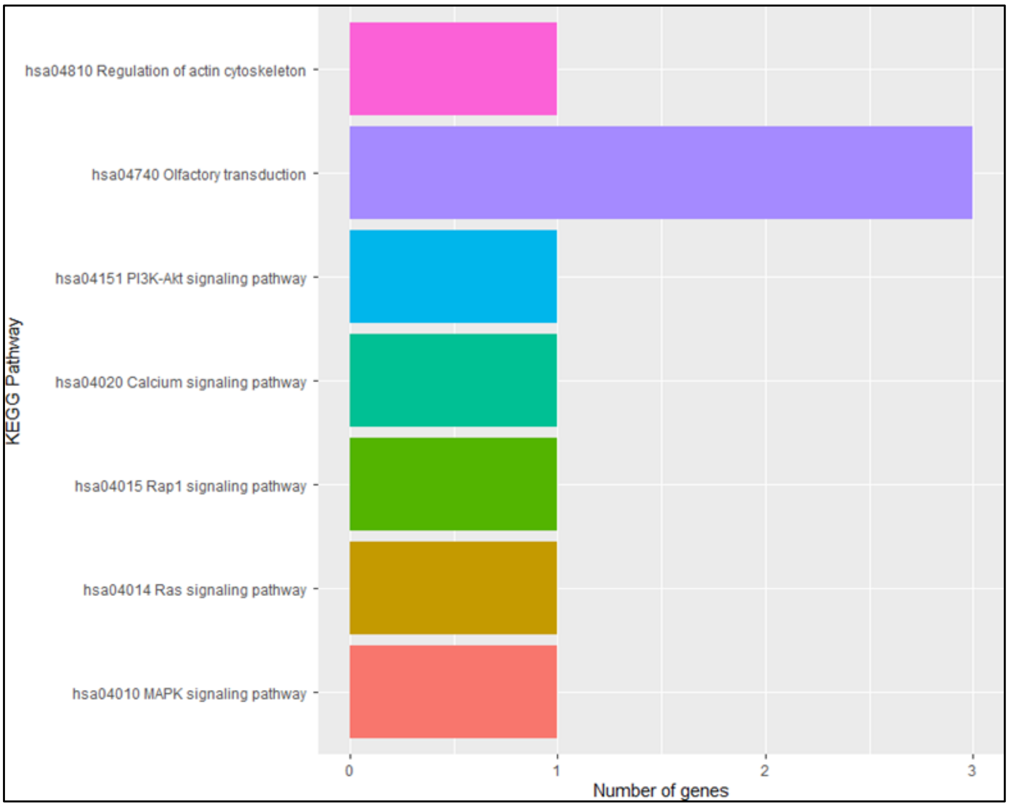


Figure 4.2 Mapping of the 15 Genes Deemed Important to KEGG Pathways

Three of the 15 genes were mapped to the Olfactory Signal Transduction Pathways. There have been several studies linking this pathway to breast cancer and metastasis. Olfactory receptors (OR) have been shown to be expressed in some cancer tissues. For instance, OR51E2 shows a high, tumour-specific expression in prostate cancer cells in comparison to non-cancer tissues (Neuhaus et al, 2012). Meanwhile, Weber et al (2018) found that transcripts of OR2B6 were detected in 73% of all breast carcinoma cell lines that they worked on whereas no expression of the gene was observed in healthy tissues. However, even though there were no published evidence of connection to metastasis at the moment, the occurrence of three types of OR namely OR9G4, OR5J2 and OR5T2 in this work showed there could be an underlying mechanism connecting metastasis to the olfactory transduction as seen in the KEGG pathway mapping. STRING analysis (Figure 4.3) however, showed no interaction between the ORs, as well as with other genes and there were no literature published regarding this as well.

To see potential connections between the 15 genes as well as with oestrogen receptor, progesterone receptor and HER2, PPI interaction prediction via STRING was conducted (see Figure 4.3). FGF4 was predicted to interact with HER2 (denoted as ERBB2 in STRING and in Figure 4.3) and the two keratin proteins (KRTAP25 and KT76) were also predicted to interact with each other.

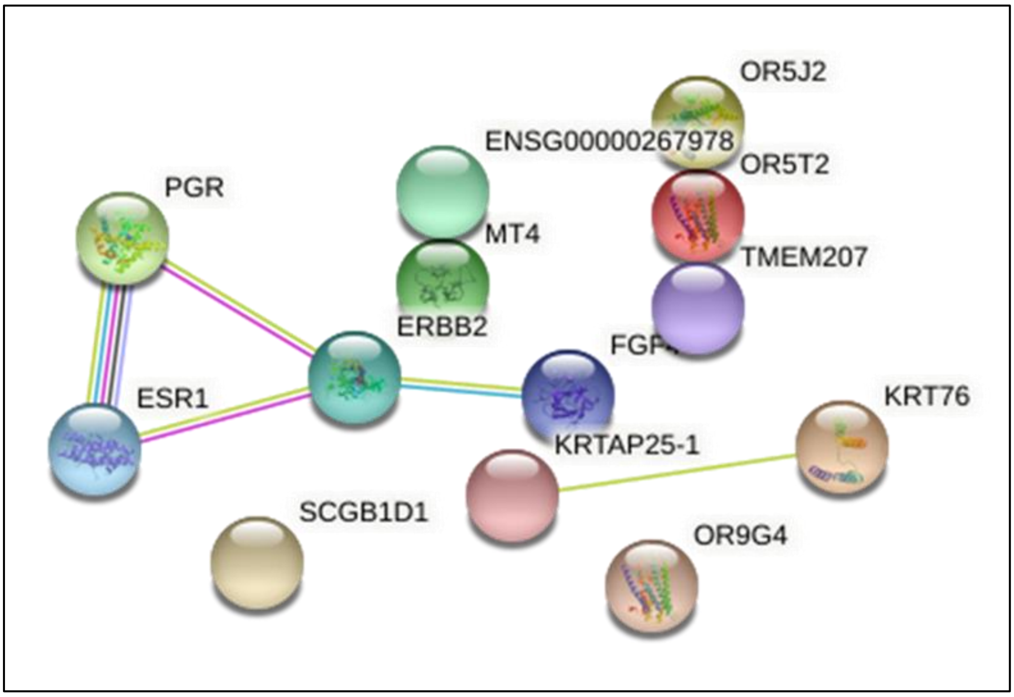


Figure 4.3 Protein-protein Interaction (PPI) Analysis Using STRING. The proteins corresponding to the genes are represented by nodes and if an interaction is predicted between 2 genes, there is an edge connecting the two nodes. The colour of the edges represents the factors covered. Green line indicates neighbourhood evidence, blue indicates cooccurrence evidence, purple indicates experimental evidence, light blue indicates database evidence and black line indicates co-expression evidence

The separation between the two groups when these 15 genes were chosen suggests that the genes could be used to differentiate the breast samples between BC and MBC. However, collectively, the interactions between the genes, and consequently the biological outcome or phenotype, relating to its pro- or anti-metastatic outcome is still unclear as publications in this area is scarce.

4.2.2 mRNA Profile of Primary *vs* Metastatic Samples

Figure 4.4 shows the PCA plot of mRNA expression of breast and metastatic samples, using z-scores. Figure 4.4 (A) is when features with a score of above 0 were retained, where 1 223 out of 16 374 genes were used to calculate the principal components and later, the PC1 and PC2 was plotted. Figure 4.4 (B) is when features with a score of above 30 were retained, where 26 out of 16 374 genes were used to construct the PCA. The total number of PC generated from this analysis was 2, with eigenvalues for PC1 was 21646016 and eigenvalues of PC2 was 59.95058. As for the percentage of variance, PC1 accounts for 100% of the variance, while PC2 does not contribute significantly to the variance. However, for the sake of only 2 PCs produced, both PCs were used for the subsequent analysis. The information of the 26 genes that were used in calculating the principal components can be found in Table 4.2, the membership of the genes in KEGG pathways can be found in Figure 4.5 and PPI interactions between the proteins can be found in Figure 4.6.

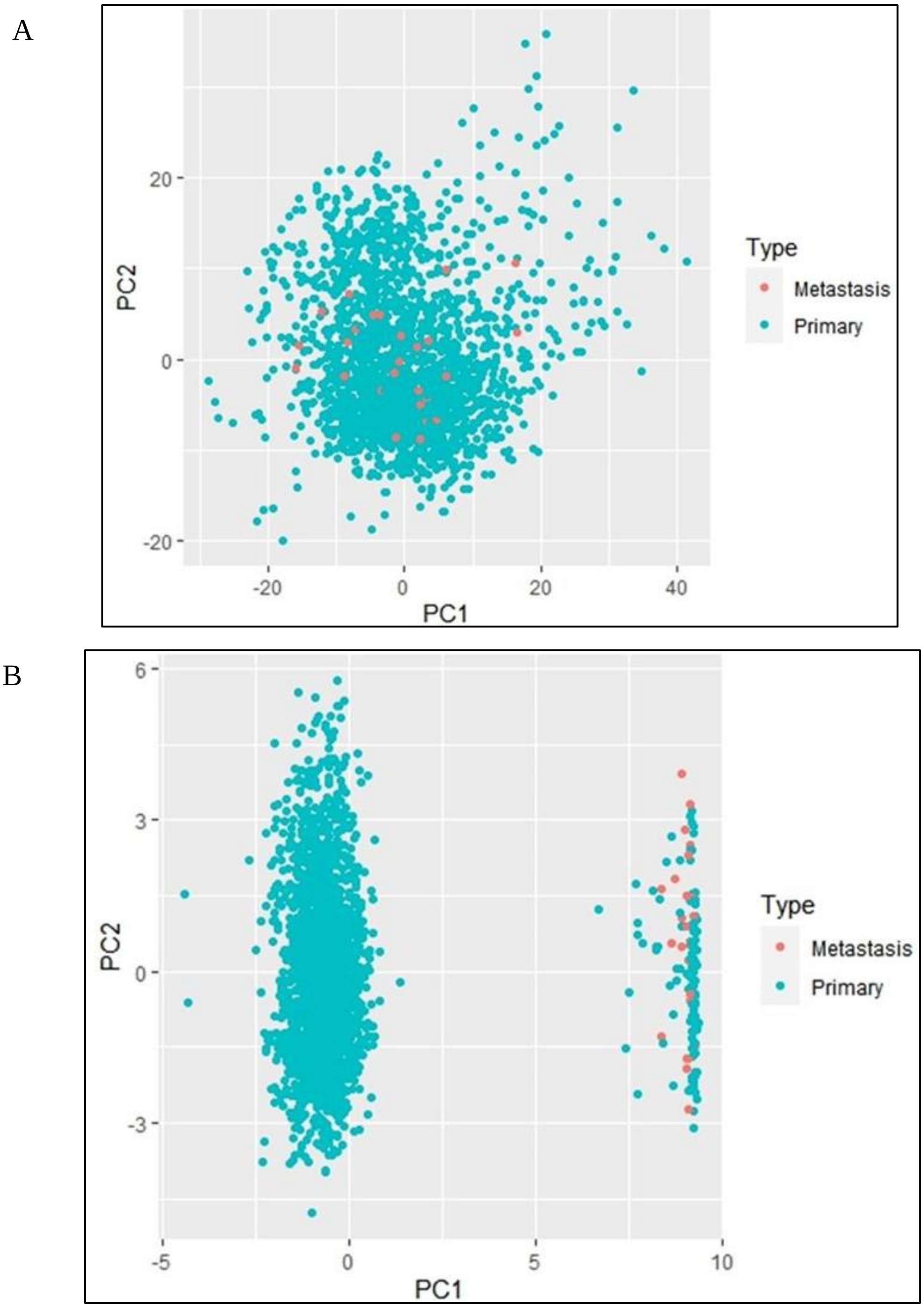


Figure 4.4 The PCA Plot of mRNA Expression Profile of Breast Samples of Primary

and Metastatic Breast Samples where (A) is when features with a score of above 0 were used to calculate the principal components and (B) is when features with a score of 30 and above were used to calculate the principal components. PC1 refers to principal component 1 (x-axis) while PC2 refers to principal component 2 (y-axis)

Several observations can be made in regard to the mRNA expression profile of breast and metastatic samples. Firstly, by employing feature selection, it avoids the curse of dimensionality. However, as seen in Figure 4.4 (A), there are no separation between the two samples. By retaining variables with scores of 30, the separation between the two samples is better (see Figure 4.4 (B)), although some of the data between the groups still overlaps. As the metastatic samples are not site-specific, this may suggest that only certain metastatic sites share similar mRNA profile with the primary samples.

Secondly, several of the 26 genes can be connected to breast cancer, metastasis and/or other cancer types (see Table 4.2). One of the genes is Matrix Metallopeptidase 13 (MMP13) where it has been shown to potentially be involved in tumour invasion and metastasis. Additionally, high expression of MMP13 has been found in breast cancer patients (Kotepui et al, 2016). Besides MMP13, FBLN1 is also correlated to breast cancer as it is found to be upregulated in breast and ovarian cancers (Moll et al, 2002). However, a study done by Gallagher et al (2005) showed that FBLN1 gene exhibit both pro-oncogenic as well as tumour suppressive effects which could be due to its ability to suppress epidermal growth factor receptors (EGFR) activation (Harikrishnan et. al, 2020). It is worth noting that several genes in Table 4.2 are also listed in Table 4.1 such as FGF4, KRT6B, KRTAP25-1, LINC00943, MAGEA9B, OR5B, OR9G4 and SCGB1D1. The shared genes may suggest potential link between breast cancer and its metastatic sites.

Table 4.2
List and Details of the 26 Genes Used to Calculate the Principal Components where there was clear separation between primary and metastatic samples. FI is feature importance

Gene symbol	Gene name	FI	Description
KRTAP25-1		100	See Table 4.1
TUSC7		73.7	See Table 4.1
MMP13	Matrix metallopeptidase 13	57.4	MMP13 have been shown to potentially be involved in tumour invasion and metastasis. Additionally, high expression of MMP13 has been found in breast cancer patients (Kotepui et al, 2016).
MMP3	matrix metallopeptidase 3	54.8	The expression of MMP3 has been found to be high in Egyptian breast cancer patients (Ibrahim, et a l, 2020). However, another study showed that MMP3 might inhibit cancer (Martin and Matrisian, 2007).
SCGB1D1		53.7	See Table 4.1
FGF4		53	See Table 4.1
OR9G4		52.5	See Table 4.1
XIST	X Inactive Specific Transcript	51.3	XIST was down regulated significantly in brain metastatic tumour of breast cancer patients. XIST expression was inversely correlated with brain metastasis, but not with bone metastasis in patients (Xing et al, 2018). Decreased expression of XIST seems to stimulate epithelial-mesenchymal transition and activated c-Met via MSN-mediated protein stabilisation, promoting stemness in the

			tumour cells.
COLEC12	collectin subfamily member 12	47.7	COLEC12, which plays a role in innate immunity, has been shown to participate in leukocyte recruitment and cancer metastasis. In cancer-associated stromal cells, increased levels of COLEC12 expression have been observed, which suggests involvement in tumour inflammation, possible through TLR4 (Li et. al, 2020).
INSC	INSC Spindle Orientation Adaptor Protein	47.5	No information on the involvement of INSC in breast cancer or metastasis.
OR5B3		47.3	See Table 4.1
BMI1	BMI1 proto-oncogene, polycomb ring finger	45.7	Several studies have shown the involvement of BMI1 in cancer stem cells where it exerts properties such as tumour initiation and differentiation of cancer stem cells. Althobiti et el. (2020) found that high BMI1 expression is associated with clinicopathological variables and outcome in breast cancer. High expression of BMI1 was associated with longer breast cancer-specific survival (BCSS) independent of other prognostic variables. It is considered a potential cancer target.
MYO1G	myosin IG	41.2	No information on the involvement of MYO1G in breast cancer or metastasis. However, the involvement MYO1E has been found to promote tumour cell de-differentiation and proliferation. Additionally, poor breast cancer patient outcome is associated with MYO1E expression, and regulation of tumour metastasis (Ouderkirk-Pecone et al, 2016).

ZNF521	Zinc Finger Protein 521	38.2	No information on the involvement of SUN5 in breast cancer or metastasis. However, its expression has been shown to promote invasion, motility and proliferation of gastric cancer cells (Huan et al, 2019).
SUN5	Sad1 And UNC84 Domain Containing 5	36.9	No information on the involvement of SUN5 in breast cancer or metastasis.
MAGEA9B		36.3	See Table 4.1
TLR8	Toll-like Receptor 8	34.1	TLRs have been implicated in invasion of breast cancer, promoting metastasis. However, TLR8 particularly have not been implicated in this (Yang et al, 2014).
KRT6B		34.1	See Table 4.1
SP7	Sp7 transcription factor also known as Osterix (Osx)	33.2	SP7 expression was significantly correlated with lymph node metastasis. Studies showed that SP7 facilitates bone metastasis of breast cancer by upregulating the expression of a cohort of genes that contribute to steps in the metastatic cascade (Yao et al, 2019).
ITGB6	Integrin subunit beta-6	32.8	The function of Integrin includes invasion, migration, adhesion, survival, growth and differentiation (Moore et. al, 2015). Dysregulation of integrin expression and/or signalling have been found to correlate with development of cancer through inappropriately regulating the processes mentioned, in addition to mediating invasion and metastasis. High expression of either the mRNA or protein in breast cancer for the integrin subunit β6 was associated with very poor survival and increased metastases to distant sites (Cantor et al, 2015).
FBLN1	fibulin 1	32.5	Higher expression of FBLN1 in normal- than cancer-associated fibroblastic stroma was

			confirmed by immunohistochemistry of breast tissues was observed. More specifically, stromal expression of FBLN1 was higher in oestrogen receptor α-positive cancers. Additionally, low stromal expression of FBLN1 was correlated with higher proliferation of cancer epithelial cells (Sadlonova et al, 2009). FBLN1 was also significantly downregulated in melanomas, and its high expression level in melanoma patients were significantly associated with having better overall survival (Liu et al, 2021), therefore suggesting connections to metastasis.
LINC00943		31.8	See Table 4.1
TGM2	Transglutaminase 2	31.8	TGM2 promotes epithelial to mesenchymal transition and was shown to promote bone metastasis of breast cancer cells possibly through down regulation of microRNA-205 (Seo et al, 2019).
FREM1	FRAS1 related extracellular matrix 1	31.2	A significantly low expression of FREM1 has been observed in breast cancer tissues. More specifically, decreased FREM1 expression was often associated with oestrogen receptor (ER)/progesterone receptor (PR) negative and triple negative breast carcinoma status (Li et al, 2020). It has been found that FREM1 is involved in the infiltration of immune cells.
CD7	CD7 molecule, T-Cell Leukaemia Antigen	30.8	No information relating to breast cancer was found, however CD7 expression was significantly correlated with tumour metastasis in NK/T-cell lymphoma patients (Fu et al, 2020).
CACNA1G-	CACNA1G antisense	30.5	CACNA1G-AS1 was found to be involved in enhancing the proliferative and invasive abilities

AS1	RNA 1	of colorectal cancer cells (Wei et al, 2020). In non-small cell lung cancer cell lines, it was found that the expression of CACNA1G-AS1 was significantly higher than in normal tissues (Yang et al, 2021). The high expression was shown to be correlated to distant metastasis, migration, cell invasion and increased epithelial-mesenchymal transition.

When mapped to KEGG pathways (see Figure 4.5), some of the genes was shown to be involved in key cancer signalling pathways such as 1 gene for MAPK and 2 genes for PI3K-Akt pathways, which was discussed previously. The olfactory transduction pathway seen in Figure 4.5 was also discussed previously. Several pathways such as regulation of actin cytoskeleton, IL-7 signalling and ECM-receptor interaction pathways may shed some light on the metastatic progression. There are five main steps to metastasis, which are detachment, cell migration and invasion, intravasation, extravasation and growth of secondary tumour. Cytoskeleton, which is made of three protein, actin, microtubules and intermediate filaments, is responsible for cell migration. Cell migration is a tightly controlled process, which is important in several biological processes such as tissue repair and embryonic morphogenesis. Aberrant cell migration in turn promotes progression of many diseases such as cancer invasion and metastasis. Cell migration can be divided into four main steps which are protrusion, adhesion, contraction and retraction (Fife et al, 2014). Cell migration is initiated in response to growth factors, which is then followed by the protrusion of the cell membranes. The protrusion is stabilised by adhesion proteins, linking actin cytoskeleton to the extracellular matrix (ECM) (Fife et al, 2014). The process continues with contraction, where the adhesions are disassembled at the rear of the cell, allowing retraction of trailing cell body towards the direction of cell movement (Ridley et al, 2003). While aberrant cell migration promotes metastasis, the process must be further aided by tumour microenvironment where inflammation may be an important contributor. Dysregulated inflammatory processes may increase influx of angiogenic cytokines from neighbouring immune cells, promoting a metastatic state (Finger and Giaccia, 2010).

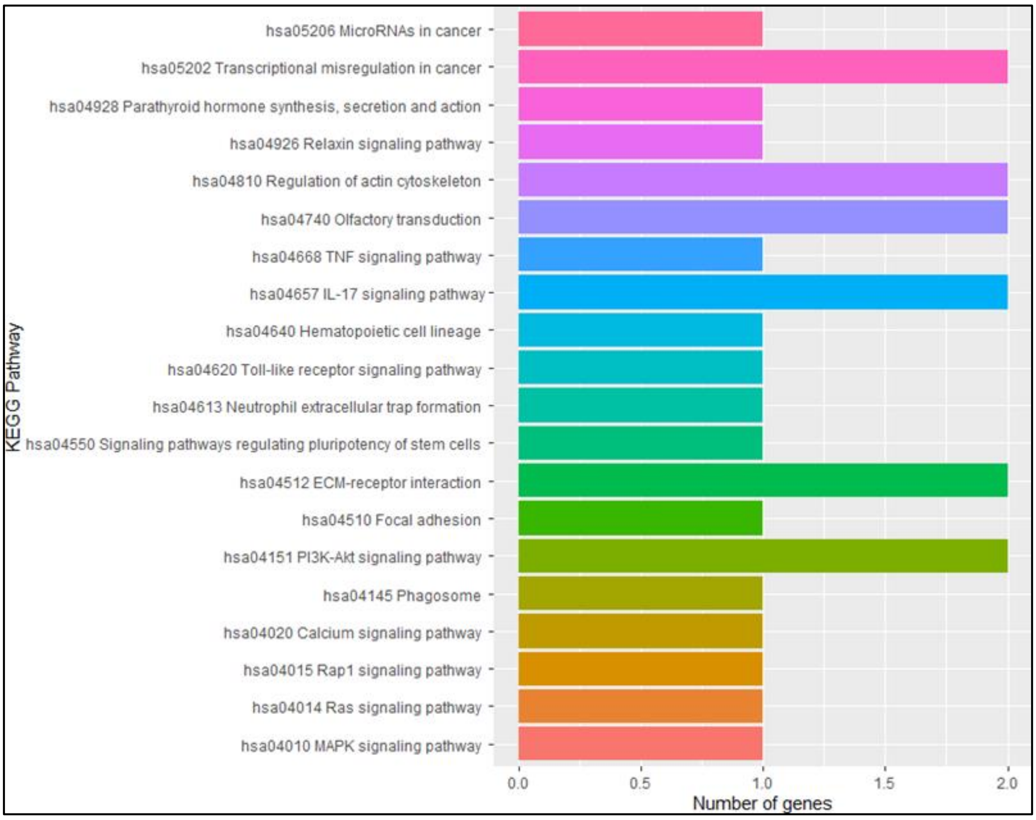


Figure 4.5 Mapping of the 26 Genes Deemed Important to KEGG Pathway

PPI prediction using STRING (see Figure 4.6) shows that several of the genes are connected to oestrogen receptor, progesterone receptor and HER2, which includes TLR8, SP7, MMP13, MMP3 and BM1. MMPs (Matrix Metalloproteinases) are a family of zinc-dependent endopeptidases that degrade various proteins in the ECM such as collagen and elastin and are important in cell migration and proliferation. TLR8 is a member of the toll-like receptor family, where TLR-induced inflammation has been shown to support tumour microenvironment. SP7 expression was significantly correlated with lymph node metastasis. Studies showed that SP7 facilitates bone metastasis of breast cancer by upregulating the expression of a cohort of genes that contribute to steps in the metastatic cascade (Yao et al, 2019). Genes involved are MMP9, MMP13, VEGF, IL-8, and PTHrP which were downregulated when SP7 were knockdown and leads to inhibition of the invasive capacity of breast cancer cells and osteolytic metastasis (Yao et al, 2019). However, overexpression of SP7 had the inverse effect which shows the involvement of SP7 in bone metastasis of breast cancer.

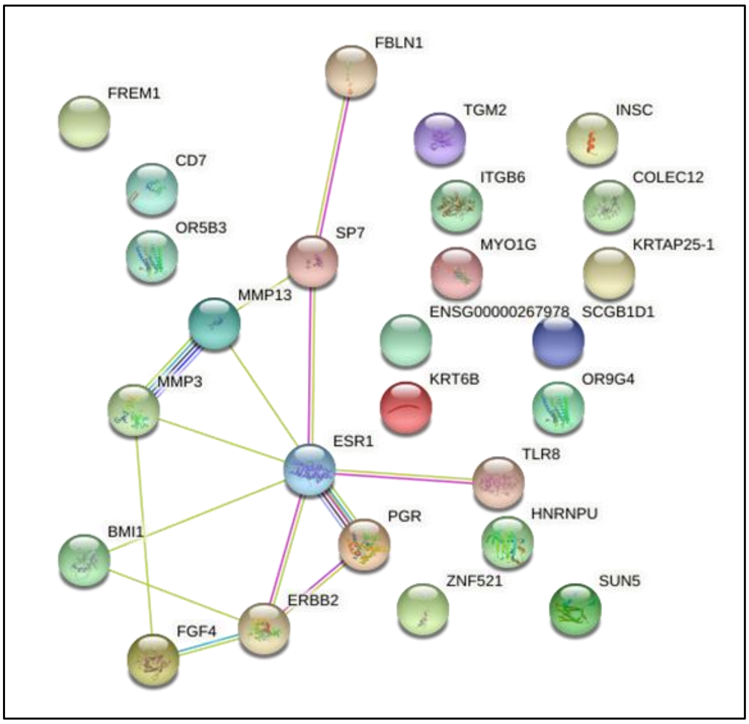


Figure 4.6 Analysis Using STRING. The genes are represented by nodes and if an interaction is predicted between 2 proteins encoded by the genes, there is an edge connecting the two nodes. The colour of the edges represents the factors covered. Red line indicates the presence of fusion evidence, green line indicates neighbourhood evidence, blue indicates cooccurrence evidence, purple indicates experimental evidence, yellow indicates text mining evidence, light blue indicates database evidence and black line indicates co- expression evidence

As previously mentioned, there is still some overlap between the groups even when only variables with a score of 30 and above were retained. To see whether this is true, similarity profile of different cancer based on the metastatic sites, with breast cancer were conducted based on their top 10 mutated genes obtained from TCGA database. Two similarity measures were used, which are the Sun's method and BOG's method (see Figure 4.7 (A) and (B) respectively). Based on Sun's method, brain cancer is the most similar type of cancer with breast cancer, with a similarity value of only 0.212. Based on BOG's method, brain cancer is also the most similar type of cancer with breast cancer, with a similarity value of 0.182. This suggests that certain types of cancer such as. brain cancer, and possibly metastatic sites, do share some similarity with breast cancer with similarity percentage of between 18 to 21%.

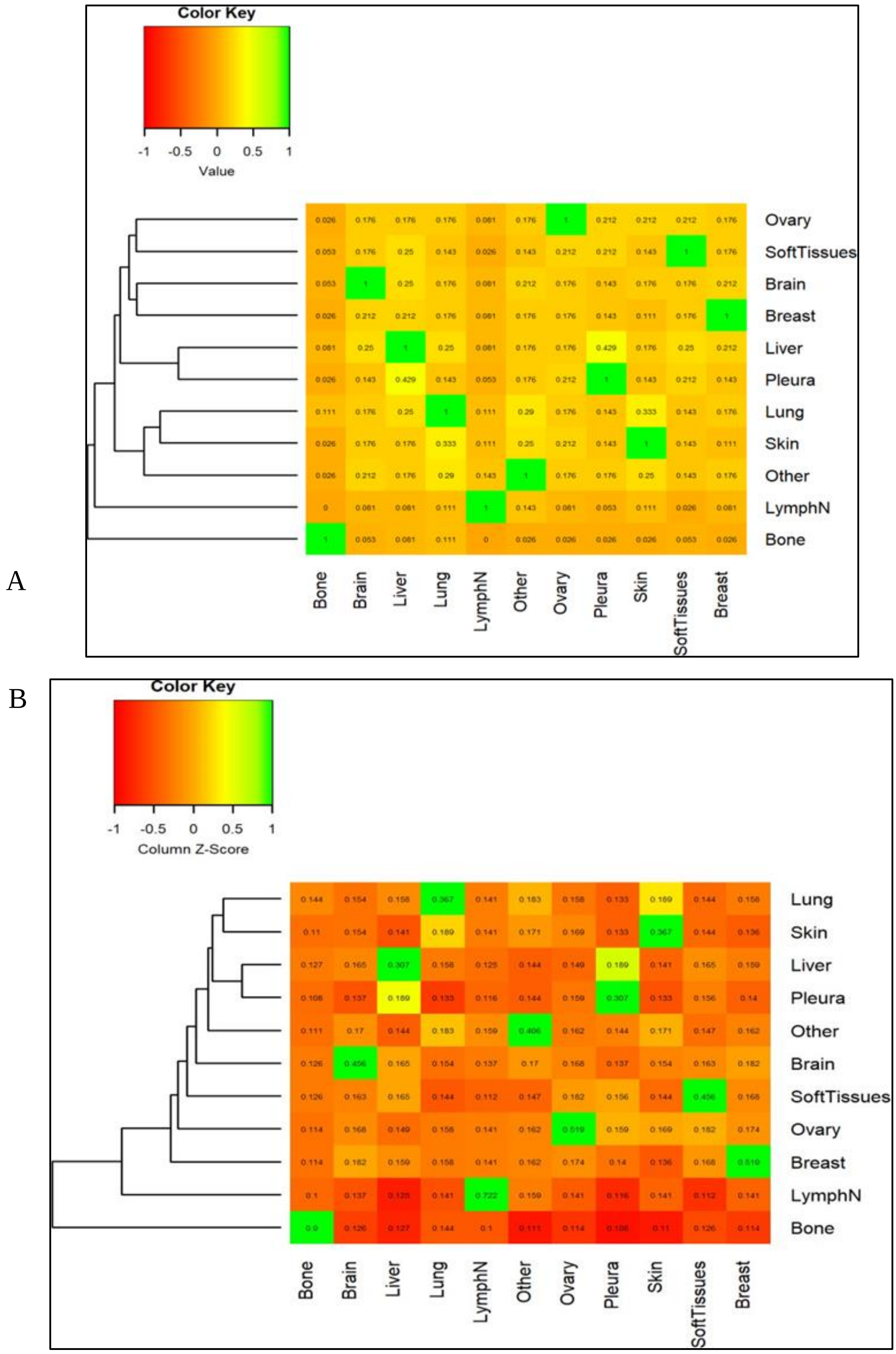


Figure 4.7 Heat Map Showing the Similarities Between the Different Cancer Types, which represents the different metastatic sites, based on top 10 mutated genes for each cancer types. (A) is when similarity was calculated using Sun's method and (B) is when similarity was calculated using BOG's method

4.2.3 CNA Profile of Breast Samples of Breast Cancer *vs* Metastatic Breast Cancer

Figure 4.8 shows the MCA plot of Copy Number Alteration (CNA) profile of breast samples of breast cancer (BC) and metastatic breast cancer (MBC). The figure showed CNA profiling was not able to separate the two groups, therefore no further analysis was performed.

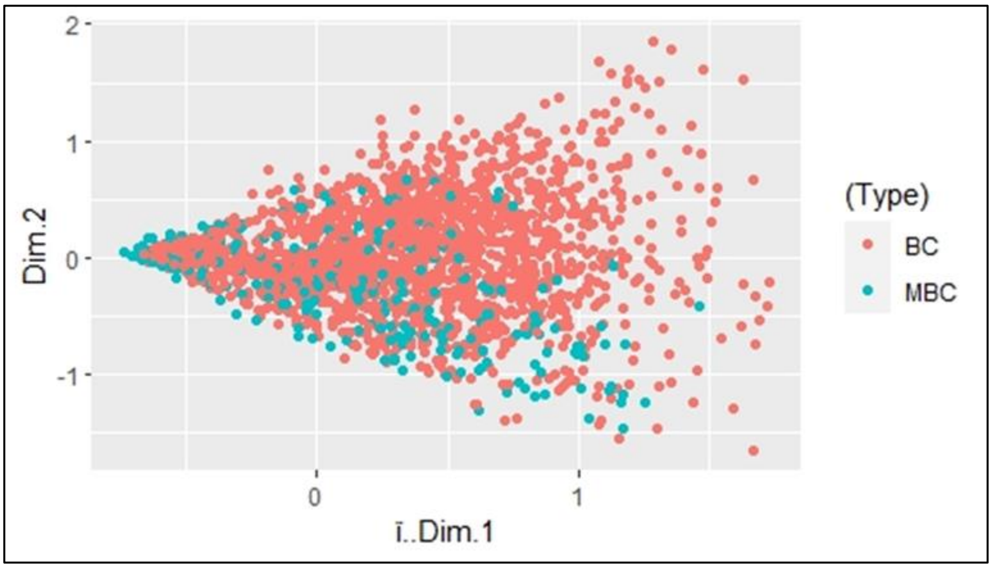


Figure 4.8 The MCA Plot of CNA Profile of Breast Samples of Breast Cancer (BC) and Metastatic Breast Cancer Patients (MBC)

4.2.4 Odds Ratio (OR) of Genetic Mutations in Breast Samples of Breast Cancer *vs* Metastatic Breast Cancer

From the data, odds ratio (OR) of genetic mutations in breast samples were calculated and its 95% confidence interval were calculated. Only 2 genes scored an OR value of above 1 with p-value < 0.05 which are AMER1 and DDR2. AMER1 which is the APC Membrane Recruitment Protein 1 was found to upregulate transcriptional activation by the Wilms tumour protein. While DDR2, also known as Discoidin Domain Receptor Tyrosine Kinase 2 facilitates cell migration and tumour cell invasion by up-regulation of the collagenases MMP1, MMP2 and MMP13. However, as shown in Figure 4.9, STRING analysis showed no recorded interactions between these two genes with ER, PR and HER2 genes.

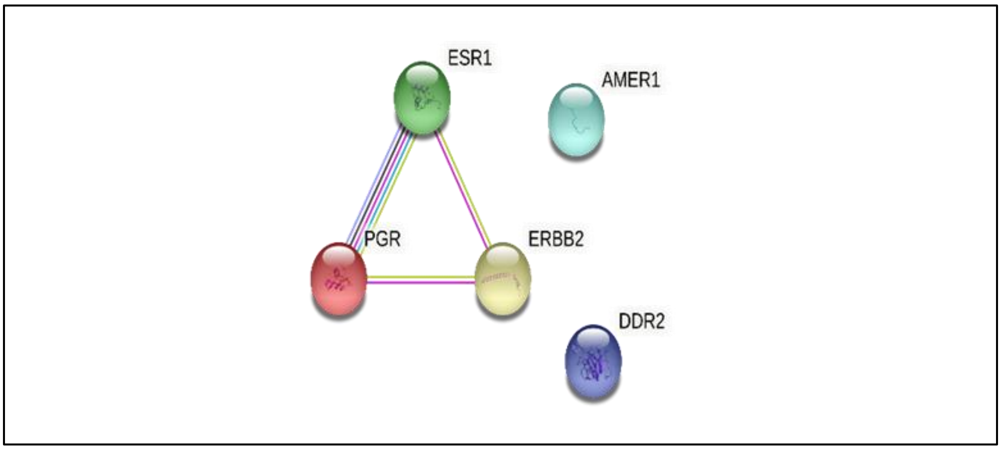


Figure 4.9 PPI of the Mutated Genes in Breast Samples of Breast Cancer vs Metastatic Breast Cancer. The protein of each gene are represented by nodes and if an interaction is predicted between 2 genes, there is an edge connecting the two nodes. The colour of the edges represents the factors covered. Green line indicates neighbourhood evidence, blue indicates cooccurrence evidence, purple indicates experimental evidence, yellow line indicates text mining evidence light blue indicates database evidence and black line indicates co-expression evidence

4.2.5 Odds Ratio (OR) of Genetic Mutations in Primary *vs* Metastatic Samples

Odds ratio and 95% confidence interval of genetic mutations in primary and metastatic samples were also calculated and the results are as shown in Figure 4.10. A total of 30 genes scored an OR value of above 1 with p-value < 0.05. However, only the top 20 were retained for further analysis. The full details of the 20 genes can be found in Table 4.3.

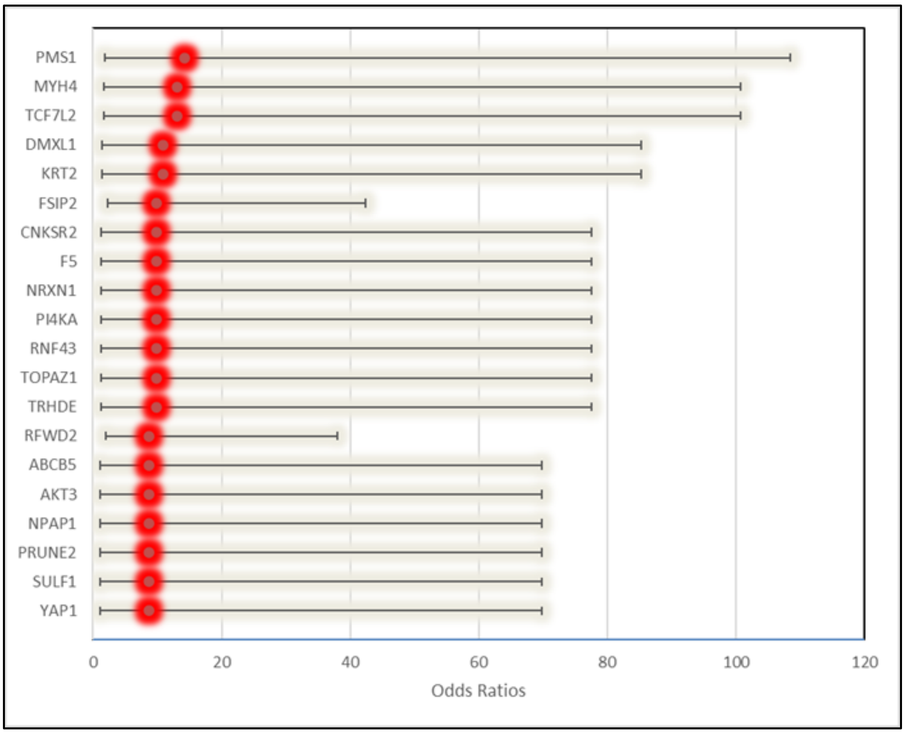


Figure 4.10 Odds Ratio and 95% Confidence Interval of Top 20 Mutated Genes in Primary and Metastatic Samples

From Table 4.3, it can be seen that some of the genes has connection to metastasis, such as DMXL1, NRXN1, PI4KA, RFWD2, ABCB5, AKT3 and YAP1. RFWD2 has been demonstrated to play a vital role in the regulation of cell proliferation, apoptosis and DNA repair and can play tumour suppressive and oncogenic roles in human malignancies (Song et al, 2020). Meanwhile, ABCB5 can significantly enhance metastasis and epithelial–mesenchymal transition (EMT), while knockdown of ABCB5 inhibited these processes (Yao et al, 2017). Furthermore, two genes have been associated with poor relapse-free survival in breast cancer patients which are MYH4 and KRT2. Phylogenetic analysis of a triple-negative ductal cancer patient, which later relapse with CNS metastases and breast metastases found that there is a mutation in the FSIP2 gene (Mattos-Arruda et al, 2019) which suggested that mutation in the FSIP2 gene might lead to breast metastasis.

Table 4.3

Top 20 Significantly Mutated Genes in Primary vs Metastatic Samples

Gene Symbol	Gene Name	Description
PMS1	DNA Mismatch Repair Protein PMS1	PMS1 is likely to be involved in the repair of mismatched DNA. Even though mismatch repair genes (MMR) have been shown to play a role in tumour control and progression, however, the role of PMS1 in this process is still poorly understood (Silva-Fernandes et al, 2021).
MYH4	Myosin Heavy Chain 4	MYH4 is involved in muscle contraction and it enables double-stranded RNA binding activity. It is associated with poor relapse free survival in breast cancer patients (Gerashchenko et al, 2020).
TCF7L2	Transcription Factor 7 Like 2	This gene encodes a transcription factor that plays a key role in the Wnt signalling pathway. The increased risk of breast cancer was found associated with TCF7L2 polymorphisms (Wang et al, 2015).
DMXL1	DmX-Like Protein	Lan et al (2019) has identified DMXL1 as one of the hub genes in triple negative breast cancer patients in which this hub genes were considered to serve important roles in the underlying mechanisms of malignancy.
KRT2	Keratin 2	Han et al (2021) found that the mRNA levels of KRT2 were expressed significantly different between primary melanoma and metastatic melanoma. Even though there is no study relating KRT2 with MBC, however, in silico analyses revealed association between KRT16 expression and shorter relapse-free survival in metastatic breast cancer (Joose et al, 2012).
FSIP2	Fibrous Sheath	Phylogenetic analysis of a triple-negative ductal cancer patient, which later relapse with CNS

	Interacting Protein 2	metastases and breast metastases found that there is a mutation in the FSIP2 gene (Mattos-Arruda et al, 2019).
CNKSR2	Connector Enhancer of Kinase Suppressor of Ras 2	David et al (2018) found that Smurf2-CNKSR2 interaction may serve as a common strategy to control proliferation of human breast cancer cells by modulating CNKSR2 protein stability.
F5	Coagulation Factor V	Tinholt et al (2018) found that F5 was expressed higher in human breast tumours as compared to normal tissues. This high expression of F5 was not only associated with aggressive tumours but also associated with survival in breast cancer.
NRXN1	Neurexin 1	Alkhathami et al (2021) observed an increase of 11.61-fold of NRXN1 expression in breast cancer patients compared to healthy controls. They found that NRXN-1 expression was significantly associated with menopausal status, lymph node involvement, oestrogen receptor (ER) status, progesterone receptor status, TNM stages, and distant metastases. They concluded that increased expression of NRXN-1 was linked with disease advancement and distant metastases (Alkhathami et al, 2021).
PI4KA	Phosphatidylinositol 4-Kinase Alpha	Waugh (2012) reported that alterations to phosphatidylinositol 4-kinase expression levels can modulate MAP kinase and Akt signalling, and are important for chemoresistance, tumour angiogenesis and the suppression of apoptosis and metastases.
RNF43	Ring Finger Protein 43	In right-sided colorectal cancer (RCRC), RNF43 mutation is associated with aggressive tumour biology along with BRAF V600E mutation (Matsumoto et al, 2020).

TOPAZ1	Testis And Ovary Specific TOPAZ 1	The expression pattern of TOPAZ1 suggests that it may play an important role in germ cell development (Baillet et al, 2011). However, no connection to metastasis was found at this moment.
TRHDE	Thyrotropin Releasing Hormone Degrading Enzyme	Hu et al (2021) stated that low TRHDE- antisense RNA1 expression is associated with poor outcomes in patients with breast cancer and potentially contributes to the aggressive tumour biology of breast cancer.
RFWD2	RING-Type E3 Ubiquitin Transferase RFWD2	RFWD2 is also known as COP1, has been demonstrated to play a vital role in the regulation of cell proliferation, apoptosis and DNA repair and can play tumour suppressive and oncogenic roles in human malignancies (Song et al, 2020).
ABCB5	ATP Binding Cassette Subfamily B Member 5	Yao et al (2017) found that ABCB5 can significantly enhance metastasis and epithelial–mesenchymal transition (EMT), while knockdown of ABCB5 inhibited these processes. They also found that ABCB5 expression was increased in metastatic tissues when compared with nonmetastatic tissues in a number of cancer subtypes, including breast cancer (Yao et al, 2017).
AKT3	AKT Serine/Threonine Kinase 3	Grottke et al (2016) found that depletion of AKT3, but not AKT1 or AKT2, resulted in increased migration in vitro while combined downregulation of AKT2 and AKT3, as well as AKT1 and AKT3 significantly increased metastasis formation in vivo. Their results showed that knockdown of AKT3 can increase the metastatic potential of triple negative breast cancer cells (Grottke et al, 2016).
NPAP1	Nuclear Pore Associated Protein 1	Even though mutations of NPAP1 in non-small cell lung cancer (NSCLC) patients were significantly associated with better progression free survival, objective response rate and durable

		clinical benefit compared with those with wild-type NPAP1 (Yang et al, 2022), there were no available studies found to associate it with metastatic breast cancer.
PRUNE2	Prune Homolog 2 With BCH Domain	Li et al (2022) investigated the expression level of PRUNE2 in colorectal cancer (CRC) cell lines and found that PRUNE2 overexpression leads to decreased cell survival, proliferation, invasion and tumourigenicity and promoted apoptosis, suggesting that PRUNE2 may function as a tumour-suppressive gene in CRC. However, there were no available studies found to associate it with metastatic breast cancer.
SULF1	Sulfatase 1	A study conducted by Fattahi et al (2021) found that there was statistically significant increase of SULF1 expression levels in advanced stages of colorectal cancer (CRC) patients and the expression level also increased in the metastatic stage as well.
YAP1	Yes1 Associated Transcriptional Regulator	Qadir et al (2021) reported that YAP1 expression was nine folds higher in tumours compared to controls and significantly associated with metastasis ($p < 0.05$) and poor survival in Pakistani breast cancer patients. These findings establish the role of YAP1 overexpression in tumourigenesis and metastasis of breast cancer.

To further elucidate the interaction between these 20 genes with ER, PR and HER2 receptor, PPI analysis was conducted. Based on Figure 4.11, it can be seen that two genes have connections with the 3 hormone receptors. AKT3 has connection with ER, while YAP1 showed connection to HER2. This may suggest not just potential mechanism of metastatic breast cancer, but also the interplay between hormone receptor status, metastatic pathways and phenotypes (see later).

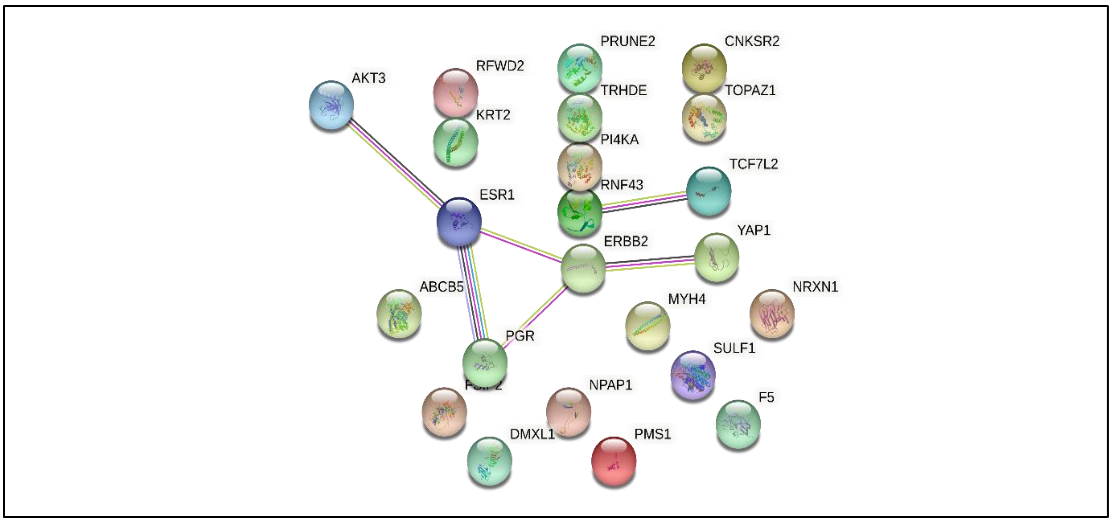


Figure 4.11 PPI of the 20 Mutated Genes in Metastatic Samples. The protein of the genes is represented by nodes and if an interaction is predicted between 2 genes, there is an edge connecting the two nodes. The colour of the edges represents the factors covered. Green line indicates neighbourhood evidence, blue indicates cooccurrence evidence, purple indicates experimental evidence, yellow line indicates text mining evidence light blue indicates database evidence and black line indicates co-expression evidence

Figure 4.12 shows the mapping of the top 20 mutated genes in metastatic samples to GO Biological Process. It can be seen that almost half of the genes are either involved in epithelial cell proliferation or regulation of it (31.15% and 24.59%, respectively).

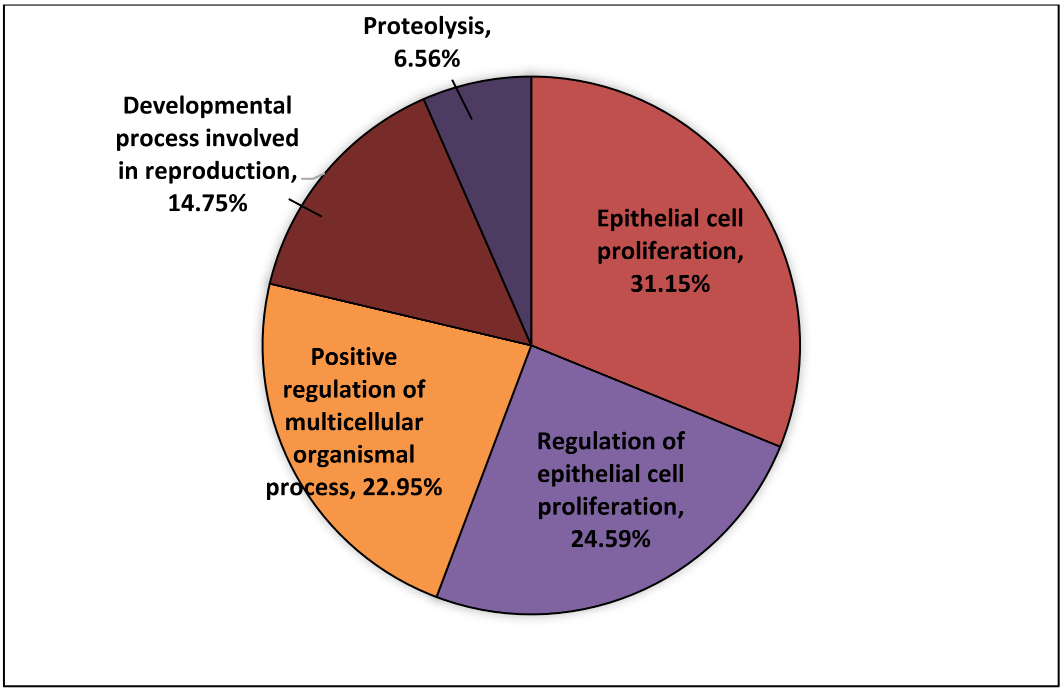


Figure 4.12 GO Annotation of the Top 20 Mutated Genes in Metastatic Samples

4.2.6 Clinical profile of Metastatic Breast Cancer Patients

4.2.6.1 Profile of Metastatic Breast Cancer Patients

In analysing the profile of metastatic breast cancer patients, multiple correspondence analysis (MCA) was performed on 911 metastatic breast cancer patients. Two analyses were performed, which were (i) associations between the variable (Figure 4.13) and (ii) association between the variable categories (Figure 4.14). To better elucidate (ii), hierarchical clustering was performed and can be seen in Figure 4.15. The MCA biplot helps to identify variables that are the most correlated with each dimension. The squared correlations between variables and the dimensions are used as coordinates. For example, in Figure 4.13, the PR and ER status of the primary site (PR_primary and ER_primary respectively) are the ones that contributed the most to dimension 1. Meanwhile, the HER2 status of the primary site and HER2 status of breast samples (HER2_primary and HER2_sample respectively) are the ones that contributed the most to dimension 2.

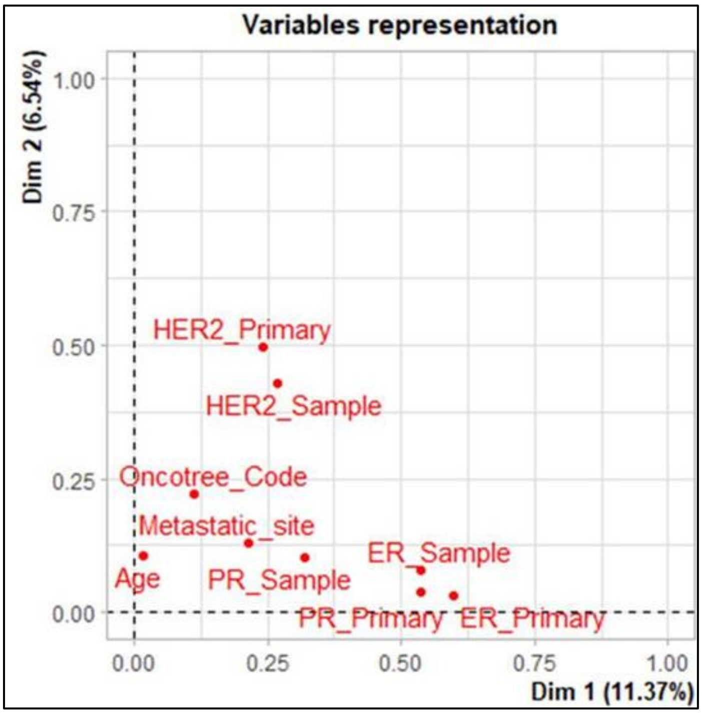


Figure 4.13 Correlation Between the Variables Collected from Metastatic Breast Cancer Patients. The closer the distance between two variables, the more correlated they are

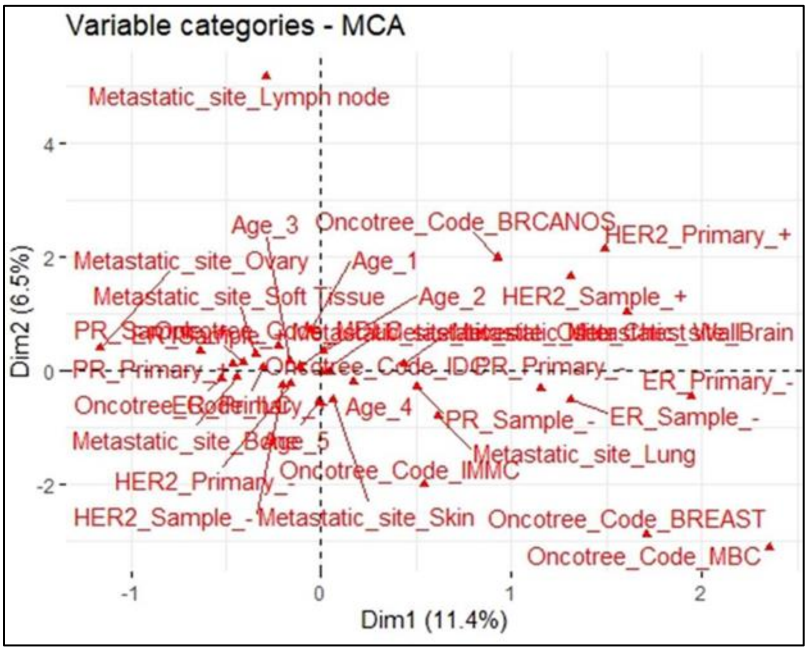


Figure 4.14 Correlation Between the Variable Categories Collected from Metastatic Breast Cancer Patients. The closer the distance between two variable categories, the more correlated they are

However, given the many variable categories in Figure 4.14, it is difficult to determine the most contributing factor for dimension 1 and 2 thus this can be better seen in the hierarchical clusters in Figure 4.15. Here, the data was divided into 5 clusters, based on the Elbow Method by plotting the within-cluster sum of squares (WCSS) or the average silhouette width against the number of clusters. This plot gives rise to the 'elbow' which suggest the number of clusters which is 5. (Adding more clusters than the 'elbow' does not significantly improve the clustering quality). Each cluster is depicted by the different colour dendrogram. Two variable categories are highly similar if they belong in the same cluster, and the distance between them is as close to 0. The distance is shown as the x - axis of the cluster. Two variables can belong in the same cluster but having a notable distance between them such as the case with OncoTree_breast and Age_1, which represents patients with OncoTree Code Breast and patients aged 30 years or below. Although both belong in the same cluster (denoted in blue), the distance is around 5, and hence are not closely similar. In contrast, ER_sample_+ and PR_primary_+, which signify ER+ status of the sample and PR+ status of primary site respectively, belongs to the same cluster (denoted in purple) and has a distance close to 0. The negative counterparts of both variable categories were also clustered together, but at a higher distance. To validate the hierarchical clustering, the cophenetic correlation coefficients were calculated using the cophenetic() function in the R 'stats' package. The result showed a mean of 4.163, median of 3.742 and the mode of 6.481 repeating 1504 times. A cophenetic coefficient ranges from -1 to 1, with 1 indicating a better fit between the original dissimilarity matrix and the cophenetic distance matrix. A higher cophenetic correlation coefficient suggests that the dendrogram provides a more accurate representation of the underlying data. Thus, based on these statistical measures, it can be said that our hierarchical clustering result is indeed reliable. Hence, it can be said that both variable categories are highly similar. PR modulates ER function and has been used as a biomarker for ER+ breast cancer. The analysis here suggest that PR may also modulate ER function in non-breast tissues.

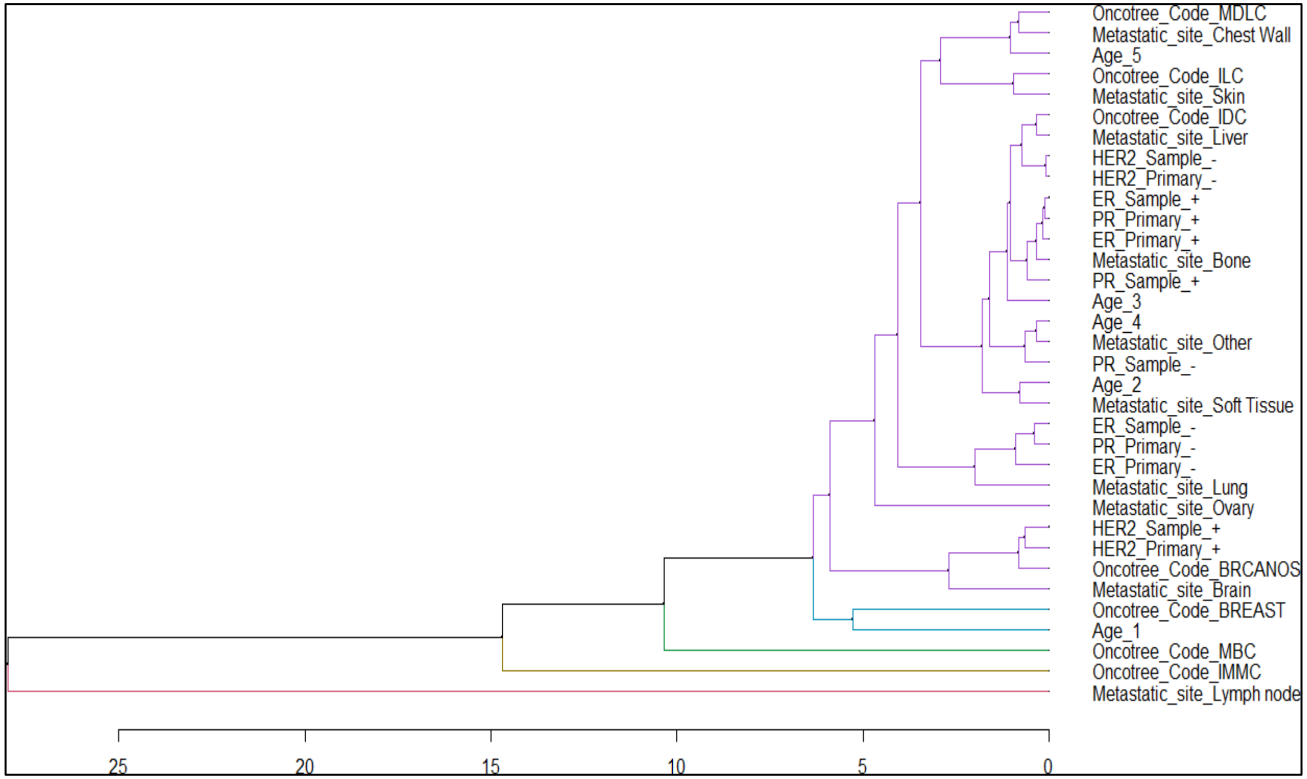


Figure 4.15 Hierarchical Clustering of the Variable Categories. The data was clustered into 5 clusters, which was colour coded. Two variable categories are similar if they belong in the same cluster and has a distance of closer to 1 (shown by the x-axis)

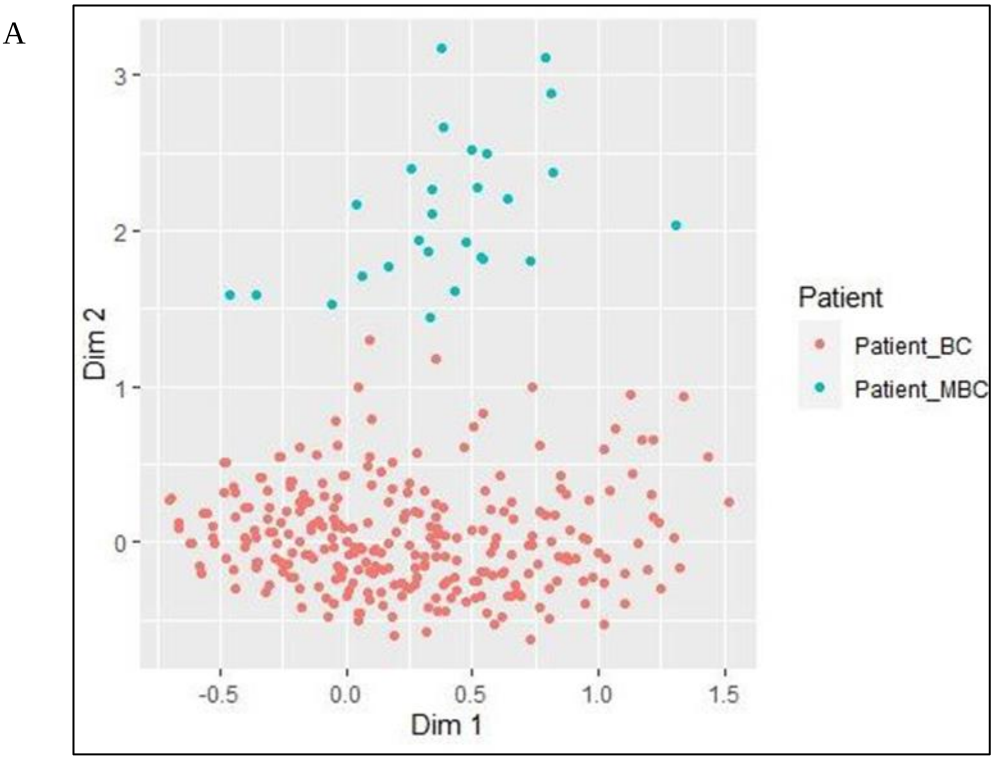
Another notable observation is between HER2 status of both primary and metastatic sites. There seems to be a high similarity between both statuses as HER2_sample_+ and HER2_primary_+ are in the same cluster, as well as having a distance close to 0. The same can be said with HER2_sample_- and HER2_primary_-. These observations closely follow information from the literature where a high concordance is observed between HER2 status of primary and metastatic sites (Bozzetti et al, 2011). Recent analysis has shown that in 33.2% of the cases, there is a discordance between the HER2 status (Arslan et al, 2011). A study done by Lower et al (2009) reported a discordance in HER-2 status in 127 patients (33.2%): 90 cases (23.6%) changed from positive to negative whereas 37 cases (9.6%) changed from negative to positive in primary and metastatic site. This study found that HER-2 negative primary lesion patients and HER- 2 positive metastasis group showed the best survival. These changes could be related to some of these mechanisms; intratumoural heterogeneity, genetic alteration during tumour progression, selective effect of prior or adjuvant chemotherapy, endocrine treatment or targeted agents and selection of resistant tumour clones (Edgerton et al, 2003).

Meanwhile, studies done in 2002 by Vincent-Salomon and colleagues found that after preoperative chemotherapy, none of the HER-2 negative tumours had changed to positive status. This means that patients with HER-2 negative primary lesions showed concordant results in the metastasis as well. The same goes for the study by Gong et al (2005) where they found the concordance rate between primary and metastatic was 97%. However, in the last few years, there were more reports on discordances in the HER2 status between primary and metastatic patients which shows that there might still be underlying factors such as epigenetics or environmental factors that have yet to be determined.

Several other observations that can be made were that IDC (invasive ductal carcinoma) and metastasis to the liver are in the same cluster, as well as ILC (invasive lobular carcinoma) and metastasis skin. Currently, there are no established relationship between subtypes of breast cancer and metastatic sites. However, there are several studies that explore this. For instance, Matthew et al (2017) found that patients with IDC had greater liver and lung/pleura involvement, which is in line with what we found, while patients with ILC had a greater tendency to develop GI and ovarian metastases. They also found that HER2-positive disease was more frequent in patients with IDC compared to the ILC group.

4.2.6.2 Comparison between Breast Cancer vs Metastatic Breast Cancer Patients using MCA

MCA was also performed to compare the clinical profile between breast cancer and metastatic breast cancer patients. Two different data sets were used, where one contains 2 additional variables, which are OncoTree code and Adjuvant Radiation (Figure 4.16 (A)) and the other does not contain these two variables (Figure 4.16 (B)).



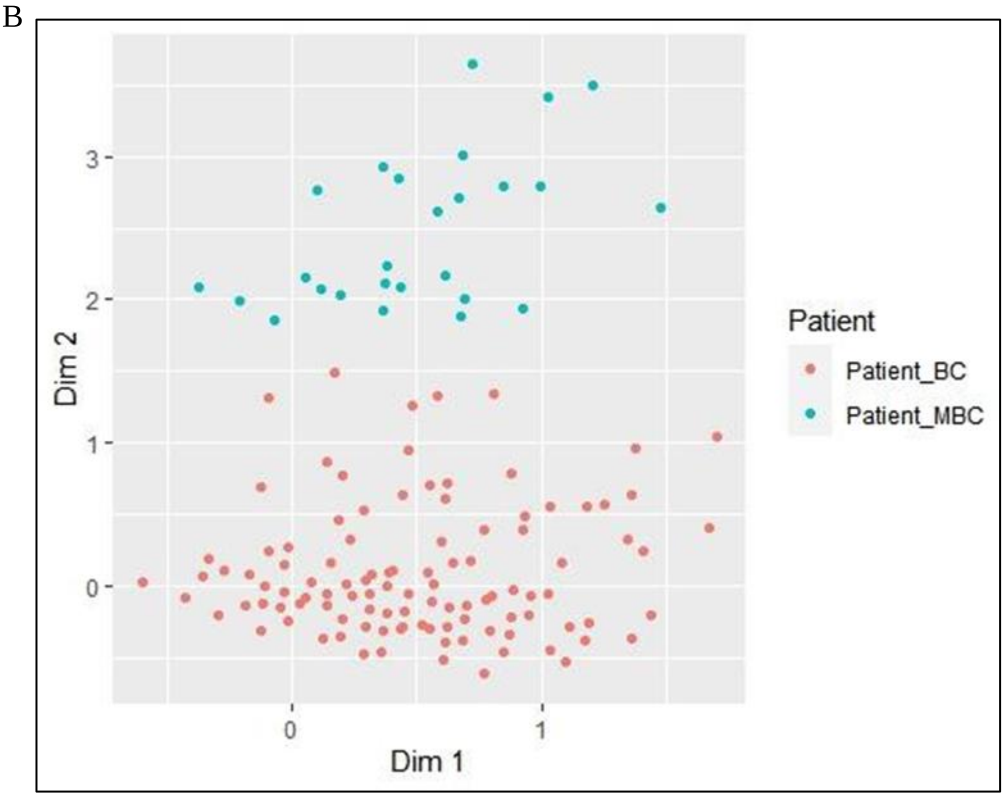


Figure 4.16 MCA Plot of Clinical Profiles Between Breast Cancer and Metastatic Breast Cancer Patients. (A) is when variables Adjuvant Radiation and OncoTree Code were used and (B) is when the variables were excluded from the profiling

From the MCA plot, both data sets showed clear separation between the two groups despite the difference in variables. This suggests that the variables used can determine the two patient subtypes. Looking at Figure 4.17, the variable Adjuvant Radiation does not seem close to any of the other variables while the OncoTree code is closest to HER2. The distance of other variables seemed almost similar to one another, making it hard to make a deduction. However, after taking out these two variables (OncoTree code and Adjuvant Radiation), it seemed hormone therapy stood in between HER2 and PR in terms of closeness and chemotherapy seemed to be closer to PR than ER as seen in Figure 4.18.

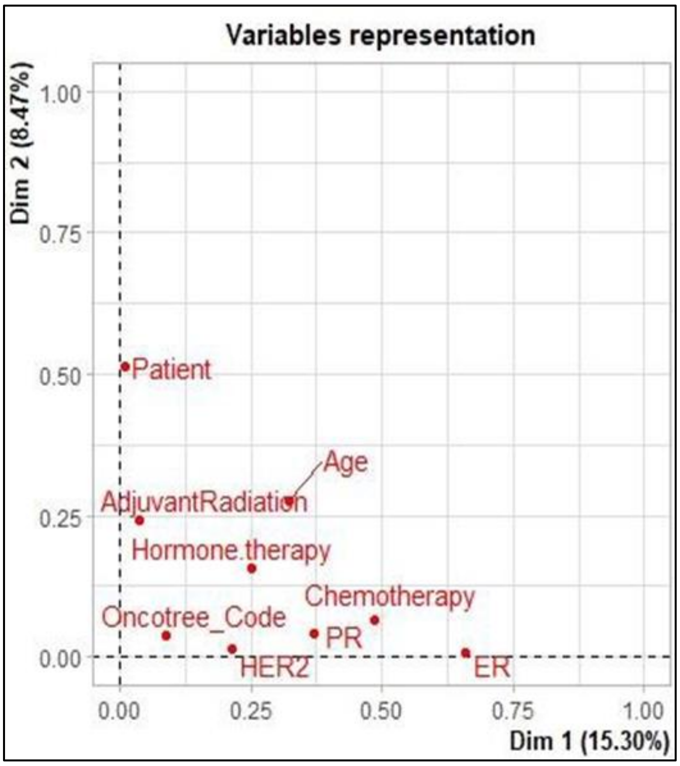


Figure 4.17 Correlation Between the Variables Used in the Profiling where Variables Adjuvant Radiation and OncoTree Code were used

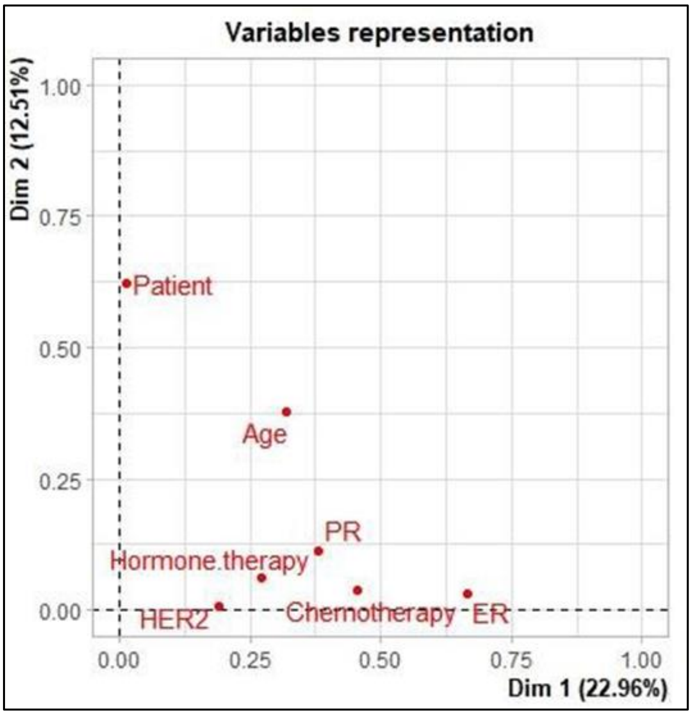


Figure 4.18 Correlation Between the Variables Used in the Profiling when the Variables Adjuvant Radiation and OncoTree Code were Excluded from the Profiling

To further elucidate the relationship between all these variables, hierarchical clustering was conducted, with inclusion of OncoTree code and Adjuvant Radiation as shown in Figure 4.19, and without these two variables as shown in Figure 4.20. From Figure 4.19, the distance of close to zero is between variable Patient_BC and HER2 -negative. This raised the question of whether patients with HER2-positive are more prone to metastasis since its negative counterpart is closed to the variable of primary breast cancer (Patient_BC). HER2-positive however is clustered together with ER-negative and PR-negative which raised the question of whether ER-negative and PR-negative is also prone to metastasis since they are all clustered together.

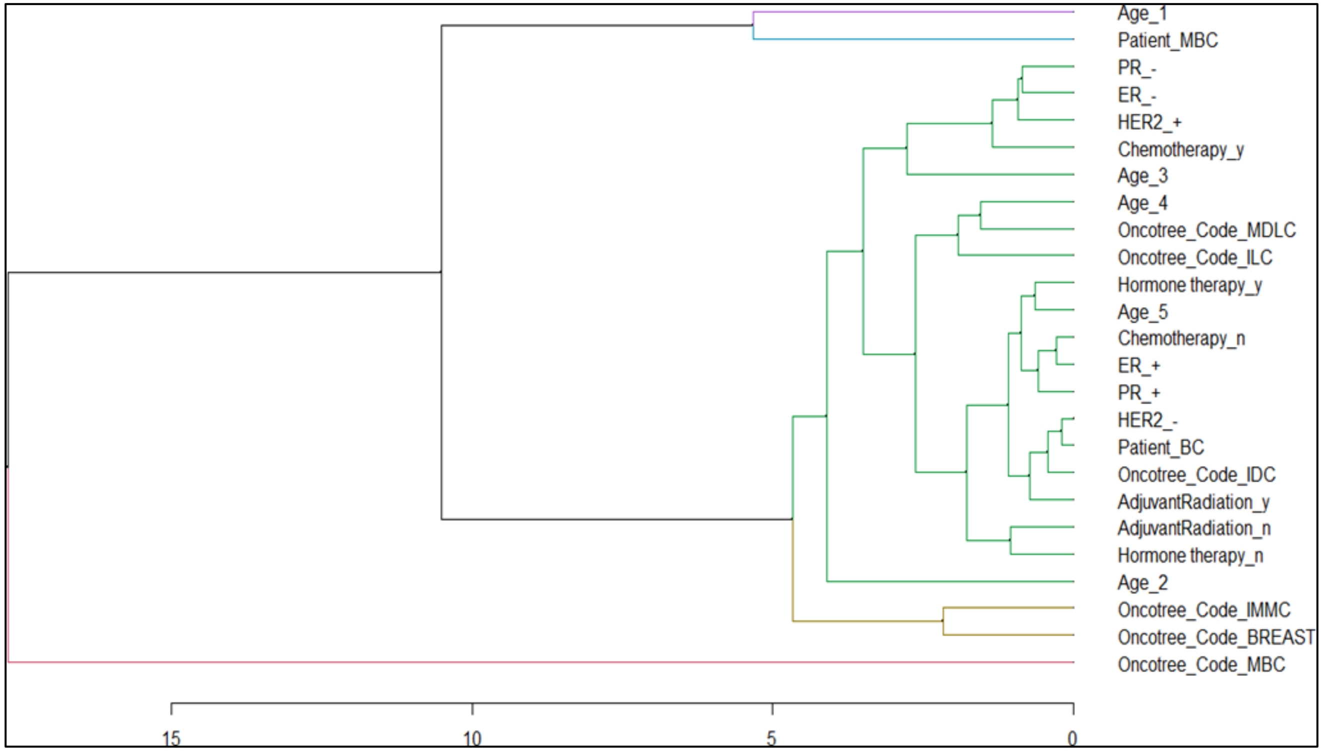


Figure 4.19 Hierarchical Clustering of the Variable Categories, which Include OncoTree Code and Adjuvant Radiation. The data was clustered into 5 clusters, which was colour coded. Two variable categories are similar if they belong in the same cluster and has a distance of closer to 1 (shown by the x-axis)

Another close relationship of close to zero is between the variable category Chemotherapy_no (patients did not receive chemotherapy for more than 2 years) and ER+ as can be seen in both Figure 4.19 and Figure 4.20. This is expected since patients with oestrogen receptor positive are usually treated with hormonal therapy.

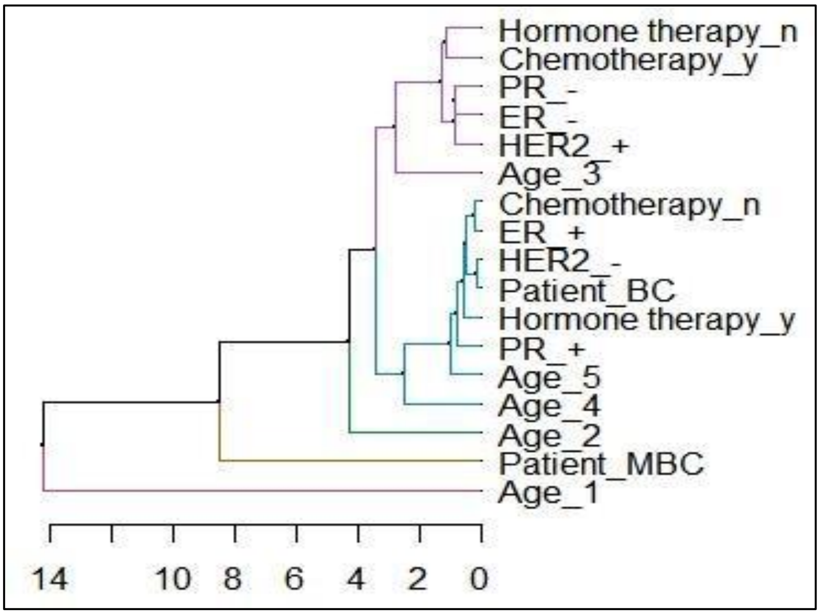


Figure 4.20 Hierarchical Clustering of the Variable Categories, without OncoTree Code and Adjuvant Radiation. The data was clustered into 5 clusters, which was colour coded. Two variable categories are similar if they belong in the same cluster and has a distance of closer to 1 (shown by the x-axis)

4.2.7 Pathway Curation of Potential Events Involved in the Occurrence of Metastatic Breast Cancer

To consolidate the previous results, with the aim of elucidating potential mechanism of occurrence of metastatic breast cancer, we first mapped the top 20 mutated genes from metastatic samples to GO Biological Process using Cytoscape, through the ClueGO application. We then proceeded to analyse whether there is a correlation between the mRNA expression and clinical profile, given that both showed good separation between breast cancer and MBC patients. Lastly, using SIGNOR 2.0, we manually curated potential pathways of MBC occurrence. Steps involved include: (i) identifying the phenotypes/ outcome of the genes using the ‘all’ option, (ii) from results identified in (i), find the published interactions between the genes with ESR1, PGR and ERBB2, as well as genes identified in the mRNA profiling and top 20 mutated genes using the ‘shortest path’ function. Regarding the latter, the interaction between 2 genes is only explored if PPI was predicted by STRING.

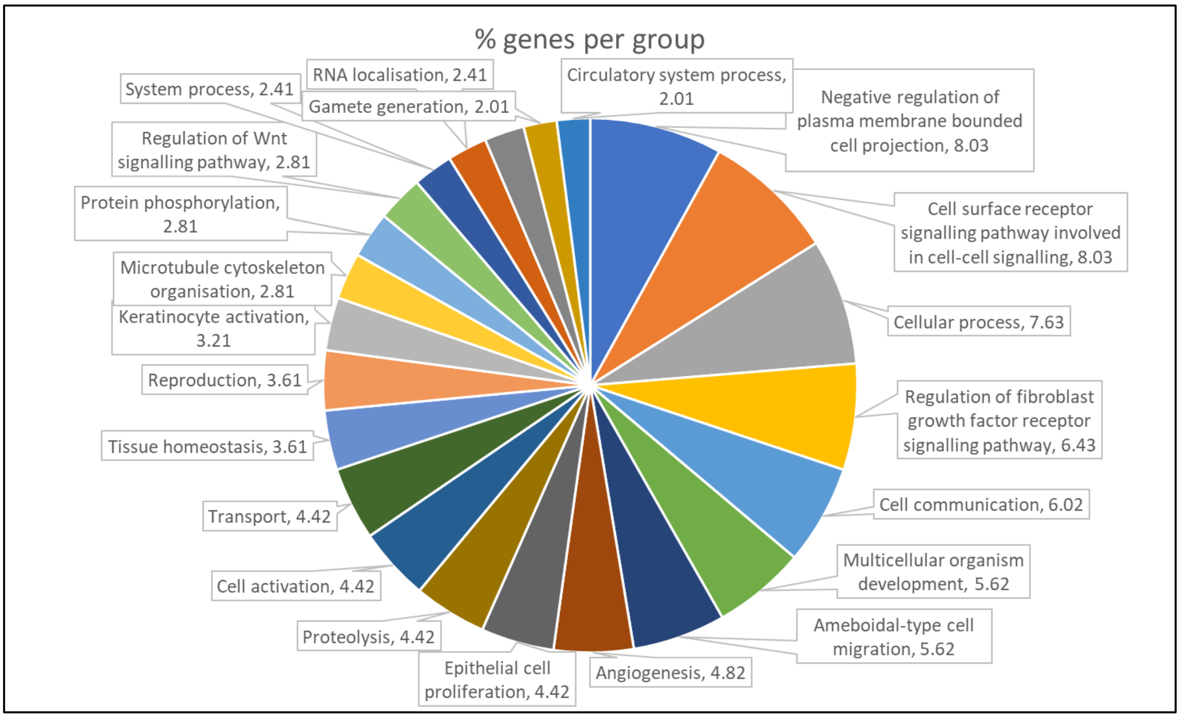


Figure 4.21 GO Annotation of Genes from Breast Samples of MBC Patients and Metastatic Samples

From Figure 4.21, the top three GO Biological Processes involved are: negative regulation of plasma membrane bounded cell projection assembly (8.03%), cell surface receptor signaling pathway involved in cell-cell signaling (8.03%) and cellular process (7.63%). Negative regulation of plasma membrane bounded cell projection assembly is referring to any process that stops, prevents or reduces the frequency, rate or extent of plasma membrane bounded cell projection assembly. It can be said that most of the genes are involved in biological processes at the cellular level, with the first two happening at the cell surface.

In the attempt to see whether the mRNA and clinical profiles are correlated, given the clear separation in the PCA and MCA respectively, Pearson’s correlation was calculated between PC1 and PC2 from PCA and Dimension 1 and 2 from MCA (see Table 4.5). From Table 4.4, it can be seen that a linear correlation between the two profiles is very weak (less than 0.3). A non-linear correlation was then calculated (see Table 4.5).

Table 4.4
Pearson's Linear correlation of PC1 and PC2 from PCA and Dimension 1 and 2 from MCA

	Dim 1	Dim 2	PC1	PC2
Dim 1	1.0***	-0.11***	-0.09***	0.06***
Dim 2	-0.11***	1.0***	0.02	0.04*
PC1	-0.09***	0.02	1.0***	0.08***
PC2	0.06***	0.04*	0.08***	1.0***

**** indicates p-value <0.01 * indicates p-value <0.1*

Table 4.5
Non-linear Correlation of PC1 and PC2 from PCA and Dimension 1 and 2 from MCA

	Dim 1	Dim 2	PC1	PC2
Dim 1	1.0***	0.36***	0.12**	0.17***
Dim 2	0.35***	1.0***	0.04**	0.16**
PC1	0.09***	0.02***	1.0***	0.08***
PC2	0.09***	0.03***	0.25**	1.0***

**** indicates p-value <0.01** indicates p-value <0.05*

A non-linear correlation means the ratio of change between two variables is not constant. If plotted on a graph, a straight line might not occur, instead the graph will have a curve and it is hard to determine the value of one variable given the value of the other variable. However, after calculating the non-linear correlation for these variables, the correlation coefficient is still very low (less than 0.3) which shows the two profiles has significantly low correlation in a non-linear manner either (Table 4.5).

In an attempt to explain possible mechanism on the occurrence of metastatic breast cancer based on the results obtained, a potential pathway was curated (see Figure 4.22).

When subjected to SIGNOR 2.0, two genes were mapped to particular phenotypes which are SP7 and YAP1. YAP1 was mapped to, and upregulates proliferation while inhibiting apoptosis and cell death. SP7 was mapped to, and upregulates osteoblast differentiation.

Apoptosis is a form of programmed cell death and is essential in inhibiting metastasis by killing misplaced cells. Resistance to apoptosis is imperative for each step of metastatic progression, but the most crucial step may be the resistance to cell death initiated by the loss of cell-cell and cell-ECM contacts (Zornig et al, 2001). The detachment of cells from the ECM leads to a type of apoptosis called anoikis, and anoikis resistance have been shown to be commonly detected in metastatic cells (Su et al, 2015).

Osteoblasts are involved in bone construction in which they synthesise the bone extracellular matrix (osteogenesis) (Rutkovskiy et al, 2016). In a study performed by Kolb et al (2019), it was found that osteoblasts suppressed both triple-negative and oestrogen receptor-positive breast cancer cell proliferation, suggesting a functional role in retarding breast cancer cell growth.

From here, we worked our way back and traced which genes lead to the up- or down-regulation of the two genes, as well as finding any potential interactions with ESR1, PGR, ERBB2, as well as genes from mRNA expression profile and OR analysis. This is done by referring to the PPI previously done, and searching for the shortest path between two genes through SIGNOR 2.0. It should be noted that not all PPI presented in STRING could be found in SIGNOR 2.0.

From Figure 4.22, it can be seen that YAP1 upregulates ESR1 (oestrogen receptor). TEAD (TEA DNA binding domains) as can be seen in the figure, are transcription factors of YAP. A study by Ma et al (2022) found that YAP-TEAD binding increases local chromatin accessibility to stimulate transcription of nearby genes.

As for SP7, SIGNOR 2.0 did not compute any interaction with ER, PR and HER2 receptor.

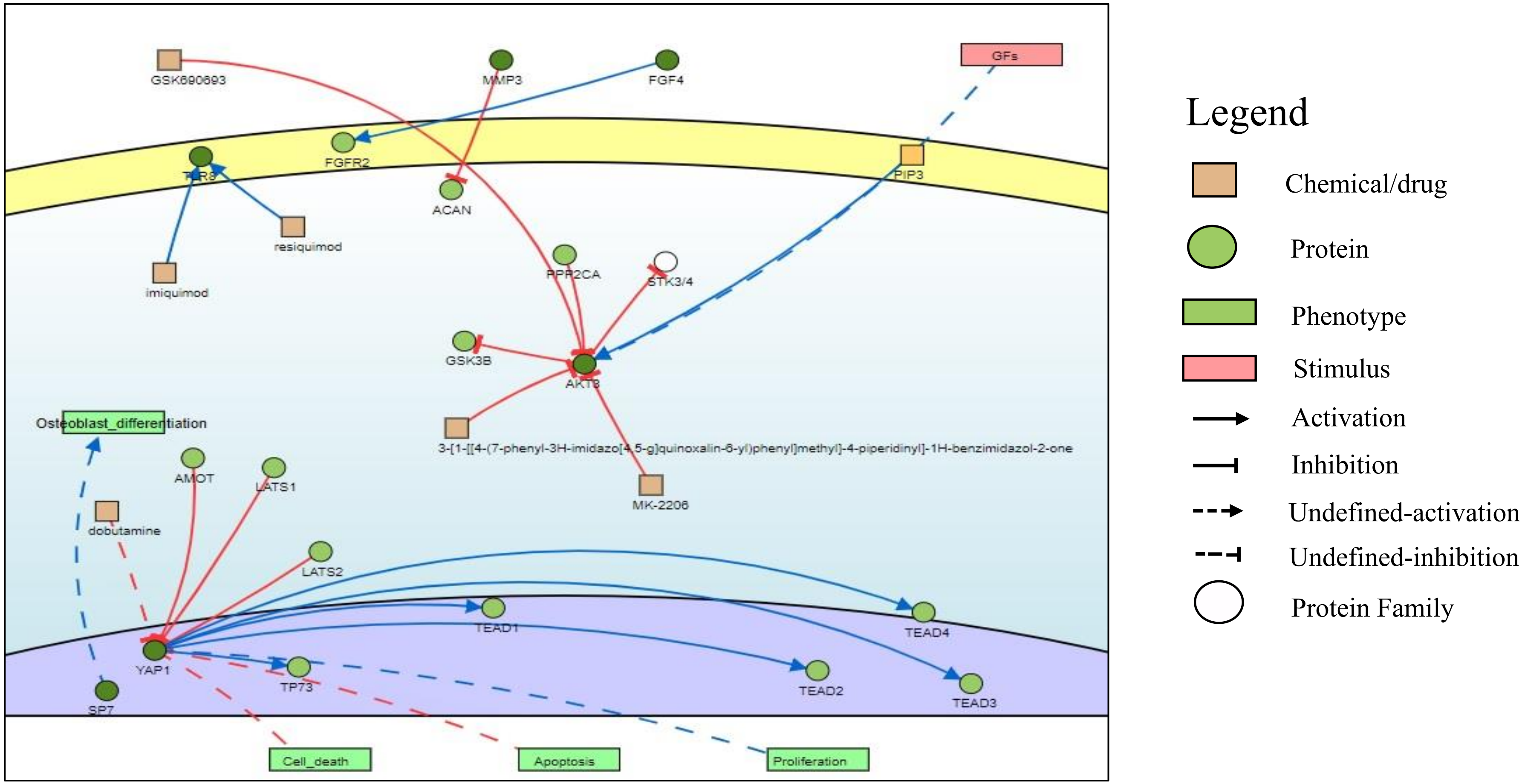


Figure 4.22 Curated Pathways Based on mRNA Expression Profile and OR Mutation of Breast Samples of MBC Patients and Metastatic Sites

4.3 Construction of Predictive Model Using Genotype, Phenotype and Clinical Data

To further validate whether the 15 genes and clinical factors identified can predict metastatic status of breast cancer patients, prediction models were constructed where two sets of different data were obtained for internal and external validation.

4.3.1 Internal Validation

The results of the internal validation for all of the training sets used can be seen in Table 4.7. As previously mentioned, the Clinical_7 contained 8 variables which are age, ER, PR and HER2 status, chemotherapy, hormone therapy, adjuvant radiation and OncoTree code; while Clinical_5 has the same variables minus adjuvant radiation and OncoTree code.

For the Clinical_7 training set, the sensitivity value was low, which is 0.303 but the specificity value is high, which is 0.996. Given the difference between the data points of the two classes, where more data were available for breast cancer patients compared to metastatic breast cancer patients, the values were expected. This scenario is commonly encountered where there are more data or results available on negative results compared to positive results. To see whether the model could be improved, AdaBoost was employed. However, as can be seen in Table 4.6, there is no significant improvement, which could mean that the AdaBoost classifier cannot capture the underlying patterns in the data effectively. To proceed, the number of breast cancer data points was reduced so that the ratio between breast cancer and metastatic breast cancer is now 5:1. The change in ratio did increase the sensitivity two-fold, while the specificity did not change significantly. The ratio was then further reduced to to 3:1, which then produce a slight reduction of both sensitivity and specificity.

In the Clinical_5 training set, it showed a low sensitivity value of 0.083 but a high specificity at 0.998. Applying the AdaBoost only produced a modest increase in sensitivity but the specificity was not significantly affected. By setting the ratio of breast cancer and metastatic breast cancer to 5:1, the sensitivity value increased six-fold to 0.611, with a slight decrease in specificity at 0.95. Further reducing the ratio to 3:1 produced a slight increase of sensitivity to 0.722 and a slight decrease in specificity.

When comparing the results between Clinical_7 and Clinical_5, reducing the ratio improved the performance of the model. The ratio of the training set, as well as the number of variables influences the performance of the model.

Meanwhile, mRNA_15 has all 15 genes that were identified previously, while mRNA_7 contained only 7 of the genes which are FGF4, MT4, OR5T2, KRT76, OR9G4, SCGB1D1 and OR5J2. For both the mRNA_15 and mRNA_7, both showed a high specificity value of 1.0 as well as high sensitivity values of 0.922 and 0.93 respectively. This could be due to the value of false positive that turned out to be 0 for both data sets.

For the CNA training set, the model showed sensitivity and specificity values of 0.346 and 0.959 respectively. Given the poor separation between the two groups shown in the previous chapter, the low sensitivity was expected. This somewhat confirms the result of the previous chapter.

Table 4.6
Results of the Internal Validation for All of the Training Sets

Data	Error Rate	Precision	F-measure	AUC	Sensitivity	Specificity
Clinical_7	0.015	0.988	0.992	0.897	0.303	0.996
Clinical_7 + Ada Boost	0.016	0.988	0.991	0.908	0.303	0.995
Clinical_7_ 5:1	0.066	0.941	0.960	0.931	0.688	0.982
*Ratio of BC: MBC is 5:1						
Clinical_7 _3:1	0.138	0.877	0.911	0.882	0.594	0.949
*Ratio of BC: MBC is 3:1						
Clinical_5	0.017	0.984	0.991	0.795	0.083	0.998
Clinical_5 + Ada Boost	0.016	0.984	0.991	0.814	0.139	0.997
Clinical_5_5:1	0.111	0.918	0.934	0.834	0.611	0.95
*Ratio of BC: MBC is 5:1						
Clinical_5_3:1	0.017	0.977	0.988	0.959	0.722	0.942
*Ratio of BC: MBC is 3:1						
mRNA_15	0.004	0.994	0.997	0.950	0.922	1.0
mRNA_7	0.004	0.995	0.997	0.962	0.93	1.0
CNA	0.154	0.865	0.909	0.855	0.346	0.959

4.3.2 External Validation

The results of the external validation can be found in Table 4.7 below. As previously mentioned, since there were no external data set containing all 7 variables for Clinical_7 could be found, hence external validation was not performed for this training set. External validation for CNA was also not performed because of its low sensitivity of 0.34 and because CNA profiling by MCA previously showed no clear separation of CNA between BC and MBC.

Table 4.7
Result of the External Validation

Training set	Error Rate	Precision	F-measure	AUC	Sensitivity	Specificity
Clinical_5	0.036	0.975	0.981	0.665	0.0	0.988
mRNA_15	0.004	0.994	0.997	0.950	0.963	1.0
mRNA_7	0.002	0.996	0.998	0.977	0.93	1.0

In terms of specificity, all 3 models showed good value. As previously mentioned, there are usually more negative data available than positive, hence this was expected. However, the same could not be said about the sensitivity value. While mRNA_15 and mRNA_7 showed high sensitivity values of 0.922 and 0.93 respectively, Clinical_5 produced a sensitivity of 0. It should be stressed that the external data set for Clinical_5 only contain 2 data points for metastatic breast cancer, and the model failed to correctly predict for both. Given these circumstances, the ability of the model to predict positive value should not be based on the prediction of these 2 data points. This model should be further tested once more data are available in public repositories (at least in the ratio of 5:1 for BC and MBC, respectively) or in future studies.

4.3.3 Visualisation of Decision Trees

To understand how Random Forest predicts the outcome based on the variables, the decision tree was visualised using the graphviz function in sklearn. Figure 4.23 is the decision tree for the Clinical_7 training set where estimator = 5. Here, the most important feature is Age, with an importance of 0.37. The top box, or node, is known as root node where the depth is 0. At each node, a question is being asked, which then divides the data into smaller subsets. There are also several characteristics in each node, which are gini, samples, value and class. Gini quantifies the purity of the node where a value of zero signifies that the node is pure, and all the samples belong to one class. If gini is more than zero, this indicate that the samples belong to a different class. Sample indicate the number of samples used in the predictive model. Value then represents how many samples belong to each class. For example [50, 50] indicates that 50 samples belong to NMBC (non-MBC) class and 50 samples belong to MBC class. Lastly, class indicate the prediction of class membership. The terminal node, where there are no further nodes, indicate the final prediction of the samples based on all questions asked. To illustrate the decision tree, the path marked in red arrow in Figure 4.23 will be explained here. In Figure 4.23, the first question is, is the patient's age in category 1 or 2 (30 years old or younger or between 31-40 years old, respectively) Based on the answer, the path will follow either the True or False paths. Further questions will be asked until it reaches a terminal node. If the patient age is in category 1 or 2, the next question that was asked was whether progesterone receptor status is negative. If the answer is true, the next question will ask whether the oestrogen receptor is negative. A false answer will be followed up on a question on whether patient is on hormone therapy. A true answer will lead to a terminal node where class is MBC and a false answer will lead to a terminal node where class is NMBC. From the decision tree, the combination of factors that leads to a particular outcome can be visualised, as well as provide an insight into metastatic breast cancer.

It should be noted that this only indicate one decision tree. The Random Forest here generates a final prediction based on 100 decision tree and hence the path of this decision tree here should not be taken as the 'end all'. However, it could be used to observe how the combination of different variable values contribute to an outcome. As for Figure 4.23 which is the decision tree for clinical factors, the most important feature in this decision tree is Age (as compared to the hormone receptor status and treatments received), with an importance of 0.37.

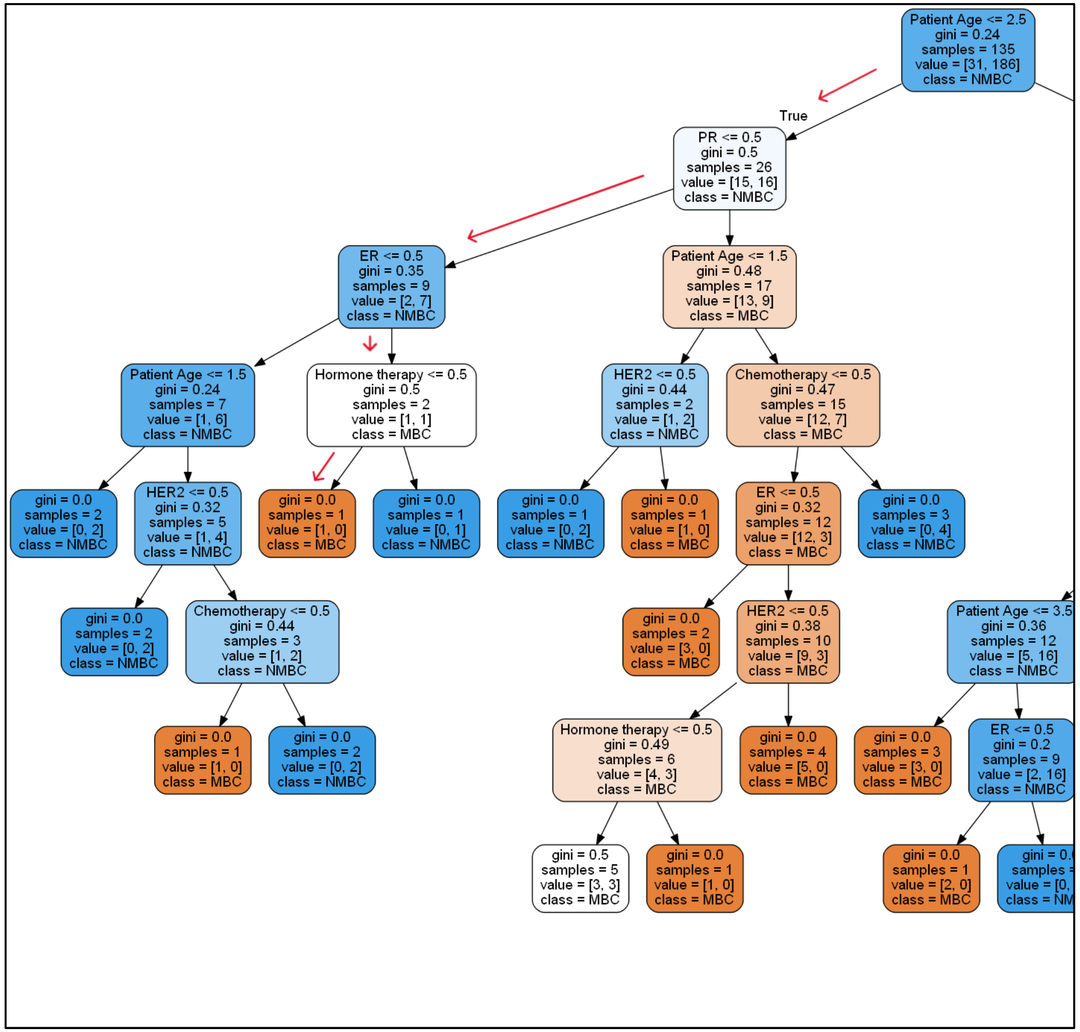


Figure 4.23 Decision Tree for Clinical_7 when Estimator = 5. The most important feature in this decision tree is Age, with an importance of 0.37

Figure 4.24 shows the decision tree for the mRNA_15 training set when estimator = 5. Here, the most important feature is the gene OR5T2, with a score of 0.93. From the decision tree, it can be seen that a z-score of -4.06 or less for OR5T2 is a strong indicator of MBC. However, there may be some exception to this, where 5 terminal nodes containing the class membership MBC exists when z-score of OR5T2 is higher than -4.06. The decision tree here shows the combination of values based on gene expression that can lead to a potential MBC diagnosis.

To illustrate this decision tree, the path marked in asterisk in Figure 4.24 will be explained here. In Figure 4.24, the first question is, whether there is an overexpression of OR5T2. Based on the answer, the path will follow either the True or False paths. Further questions will be asked until it reaches a terminal node. If there is no overexpression of OR5T2, the next question that was asked was whether there is an overexpression of KRTAP25. A false answer will be followed up on a question on overexpression of KRT76. Another false answer will be followed up on a question on overexpression of LINC01091. A true answer will lead to a terminal node where class is MBC and a false answer will lead to a terminal node where class is NMBC. From here, the combination of factors that leads to a particular outcome can be visualised and we can see how the combination of different variable values contribute to an outcome.

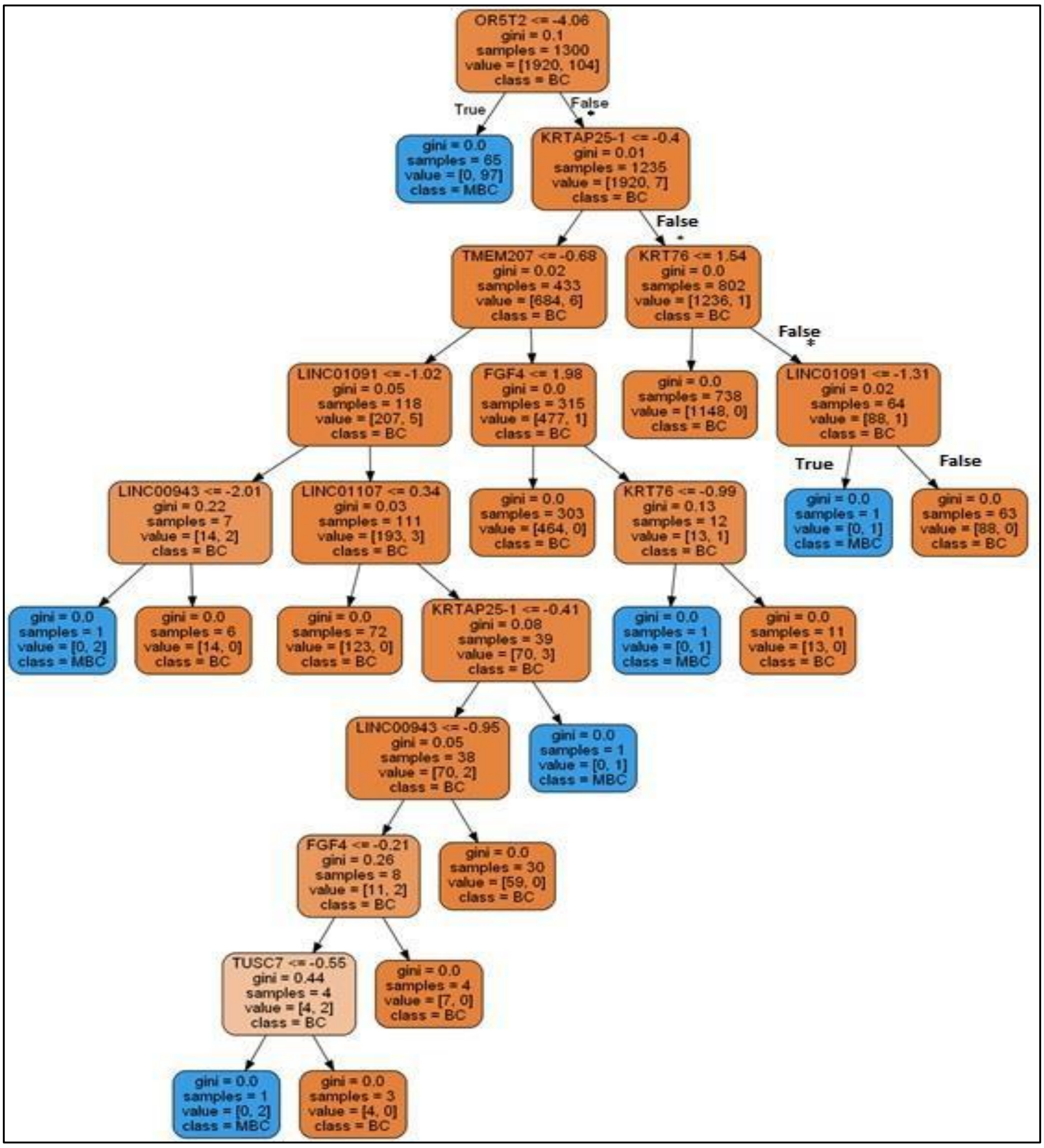


Figure 4.24 Decision Tree for mRNA_15 when Estimator = 5. The most important feature in this decision tree is OR5T2, with an importance of 0.93

Figure 4.25 shows the decision tree for the mRNA_7 training set when estimator = 5. Here, the most important feature is the gene SCGB1D1, with a score of 0.95. From the decision tree, it can be seen that a z-score of -2.97 or less for SCGB1D1 is a strong indicator of MBC. However, there may be some exception to this, where 5 terminal nodes containing the class membership MBC exists when z-score of SCGB1D1 is higher than -2.97. The decision tree here shows the combination of values based on gene expression that can lead to a potential MBC diagnosis.

For example, if expression of SCGB1D1 is not -2.97 or less, the next question is the expression of DBIL5P2. If it is true, the next question is again on expression of SCGB1D1 whether it’s 0.35 or less or -0.81 or less and it will continue on until a terminal node is reached.

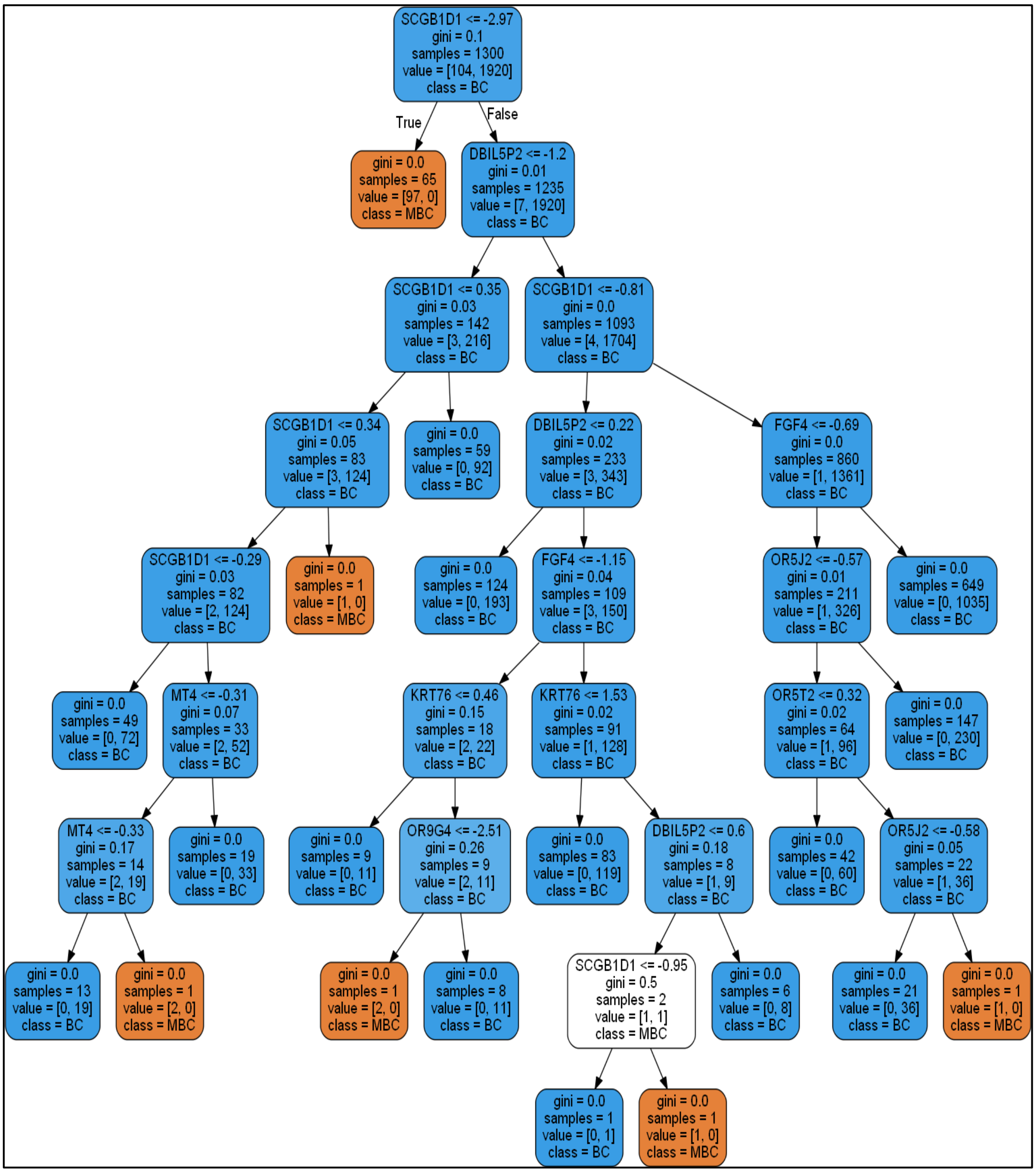


Figure 4.25 Decision Tree for mRNA_7 when Estimator = 5. The most important feature in this decision tree is SCGB1D1, with an importance of 0.95

4.4 Validation using Systematic Review

This systematic review was done to further validate the previous findings on clinical factors that could possibly lead to metastatic breast cancer. Ideally, the review should also cover the 15 genes that were previously identified. However, since not all of the parameters we looked for can be found in literature, this review only focuses on the effect of treatments for metastatic breast cancer patients with regards to their HER2 and hormone receptor status.

In the previous findings (section 4.2.6), the followings were found:

HER2- patients were clustered together with primary breast cancer (Figure 4.23), which raised the question on whether HER2+ patients are more prone to metastatic disease. HER2+ patients were clustered together with ER- and PR-. Additionally, patients who received hormone therapy for more than 2 years are clustered together with primary breast cancer along with HER2- and ER+ parameters which raised the question of whether treatment with hormone therapy for more than 2 years reduced the risk of HER2 -, ER+ patients to progress to metastatic state. Predictive modelling showed that combination of factors such as mRNA and genetic profiling was able to classify BC and MBC.

Therefore, this review was conducted to validate these findings. In this review, median PFS was used as an indicator for occurrence of metastatic breast cancer. High PFS was viewed to mean that the combination of the statuses has lower probability to contribute to MBC occurrence. In contrast, low PFS could mean the combination of statuses has higher probability to contribute to metastatic occurrence. However, it has to be noted that not all parameters used in the classification model in the previous chapter were included in the articles, as well as in the systematic review.

4.4.1 Search Results and Characteristics of Included Study

The search result was summarised in Figure 4.26. A total of 5301 articles were identified through Web of Science (366 articles), Science Direct (452 articles), PubMed (3172 articles) and Scopus (1311 articles).

After exclusion of articles published before year 2001, non-English articles and articles in the form of review, proceedings, chapters in books, book series and books, a total of 725 of articles were selected. The titles and abstracts of the articles were then screened for relevancy which narrowed down the number to 261. Another 205 articles were removed due to duplicates (151) and lack of access to full text articles (54). Remaining articles were then assessed for eligibility based on the inclusion criteria and a total of 56 articles met the criteria. However, after quality assessments were conducted, the final number of articles for this review is 13.

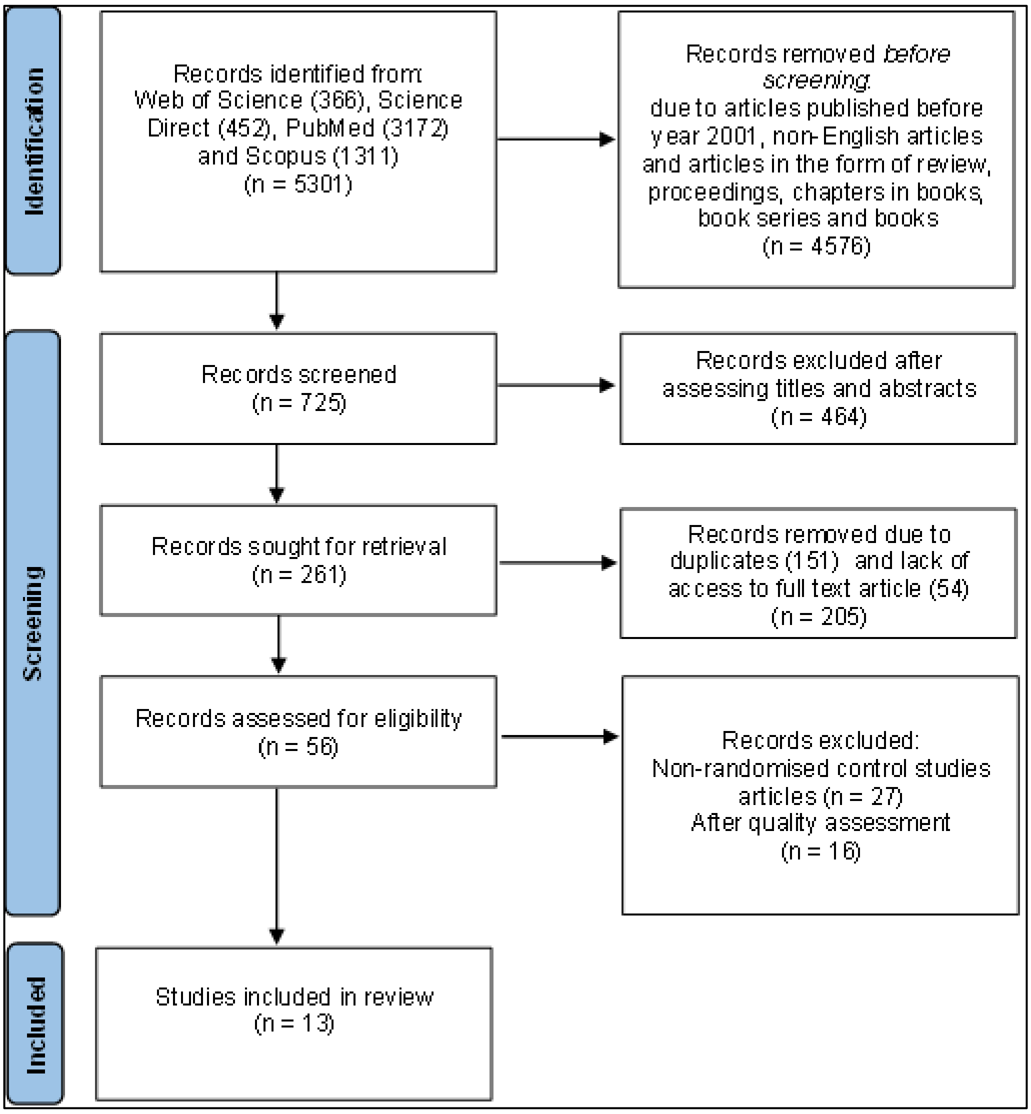

Identification
Records identified from Web of Science (366), Science Direct (452), PubMed (3172) and Scopus (1311) (n = 5301)
Records removed *before screening* due to articles published before year 2001, non-English articles and articles in the form of review, proceedings, chapters in books, book series and books (n = 4576)
Screening
Records screened (n = 725)
Records excluded after assessing titles and abstracts (n = 464)
Records sought for retrieval (n = 261)
Records removed due to duplicates (151) and lack of access to full text article (54) (n = 205)
Records assessed for eligibility (n = 56)
Records excluded: Non-randomised control studies articles (n = 27) After quality assessment (n = 16)
Included
Studies included in review (n = 13)

Figure 4.26 Overview of Study Selection

4.4.2 Quality Assessment

After quality assessment from 3 independent researchers, 13 studies that fulfilled the inclusion criteria were identified and reviewed in this study. The scoring for each identified paper is listed in Table 4.8. The score was given based on the items listed in the JADAD Scoring list (see Appendix 1) where 2 marks were on randomisation, 2 marks on blinding and 1 mark on withdrawals and dropouts. Only papers with scores of 3 and above were reviewed (13 articles).

Table 4.8
JADAD Scoring for the Shortlisted Randomised Controlled Trial. 13 papers with scores of 3 and above were agreed to be reviewed

Author, year (ref no)	Randomisation	Blinding	Withdrawals and dropouts	Total
Baselga et al, 2012 (42)	2	1	0	3
Han et al, 2018 (200)	2	1	0	3
Hortobagyi et al, 2018 (201)	2	1	0	3
Im et al, 2020 (202)	2	0	1	3
Johnston et al, 2013 (203)	2	0	1	3
Kornblum et al, 2018 (204)	2	1	1	4
Krop et al, 2016 (205)	2	1	1	4
Masuda et al, 2017 (206)	1	1	1	3
Murthy et al, 2020 (207)	1	1	1	3
Pallis et al, 2012 (208)	2	1	0	3
Swain et al, 2015 (209)	1	1	1	3
Toi et al, (2017) (210)	1	1	1	3
Vuylsteke et al, 2016 (211)	1	1	1	3

The selected 13 papers with scores of 3 and above were then classified into different groups depending on the treatment used in the study. Four types of treatments were identified where one study compared treatment of chemotherapy versus other chemotherapy (Pallis et al, 2012). Seven studies studied the combination between targeted therapy and chemotherapy (Baselga et al 2012; Swain et al 2015; Vuylsteke et al 2016; Masuda et al 2017; Toi et al 2017; Han et al 2018 and Murthy et al 2020) while Im et al, 2020 is the only one comparing targeted therapy and chemotherapy. Three studies compared the treatment of targeted therapy to placebo (Krop et al 2016; Hortobagyi et al 2018; Kornblum et al 2018), while the last one compared the treatment of hormonal therapy versus another hormonal therapy (Johnston et al, 2013). The summary for each paper is listed in Table 4.9.

Table 4.9

Research Findings of the 13 Shortlisted Paper. T1 represents the main treatment being investigated while T2 is the control treatment. OS represents overall survival and PFS is the progression free survival (in months)

Reference	**Target Population**	**T1**	**T2**	**Type of Treatment T1**	**Type of Treatment T2**	**Primary endpoint of efficacy**	**Patients on T1, *n***	**Patients on T2, *n***	**Average age**	**OS T1**	**OS T2**	**Hazard Ratio (95% CI), *p* value**	**PFS T1**	**PFS T2**	**Hazard Ratio (95% CI), *p* value**
Chemotherapy only															
Pallis; et al (2012)	Women with MBC pretreated with anthracyclines and taxanes.	capecitabine (Cap arm: 1250 mg/m2 twice daily, on days 1 through 14)	vinorelbine/ gemcitabine doublet (VG arm: vinorelbine 25 mg/m2; gemcitabine 1000 mg/m2; both drugs on days 1 and 15).	chemotherapy	chemotherapy	PFS; OS	78	80	60	22.4	20.4	NA	5.2	5.4	NA
Targeted therapy plus chemotherapy															
Baselga; et al (2012)	HER2-positive metastatic breast cancer.	8 mg/kg of trastuzumab, followed by 6 mg/kg every 3 weeks; 75mg/m2 docetaxel every 3 weeks; 840 mg of Pertuzumab followed by 420 mg every 3 weeks until disease progression	placebo plus trastuzumab plus docetaxel	Targeted therapy: anti-HER2 monoclonal antibodies	chemotherapy	PFS; OS	402	406	54	69 [17. 2%]	96 [23. 6%]	0.64 (0.47 to 0.88)P = 0.005)	18.5	12.4	0.62 (0.51 to 0.75); P<0.001

Swain; et al (2015)	MBC patients who had not received previous chemotherapy or anti-HER2 therapy for their metastatic disease.	Pertuzumab 840mg on day 1 of cycle 1, followed by 420 mg on day 1 of each subsequent cycle; Trastuzumab 8 mg/kg on day 2 of cycle 1, followed by 6 mg/kg on day 1 of the remaining cycle; docetaxel 75 mg/m2 on day 2 of cycle 1 and on day 1 of the remaining cycles.	Placebo 840mg on day 1 of cycle 1, followed by 420 mg on day 1 of each subsequent cycle; Trastuzumab 8 mg/kg on day 2 of cycle 1, followed by 6 mg/kg on day 1 of the remaining cycle; docetaxel 75 mg/m2 on day 2 of cycle 1 and on day 1 of the remaining cycles	Targeted therapy: anti-HER2 monoclonal antibodies	Placebo	PFS; OS	402	406	NA	56.5	34.7	0.55; (0.45 to 0.67); P<0.001	56.5	40.8	0.68; (0.56 to 0.84); P<0.001
Toi; et al (2017)	Asian postmenopausal women with HER2+ advanced breast cancer, who had not received systemic therapy for advanced disease	10 mg everolimus once a day orally plus weekly trastuzumab intravenously at 4 mg/kg loading dose on day 1 with subsequent weekly doses of 2 mg/kg of each 4 weeks cycle plus paclitaxel intravenously at a dose of 80 mg/m2 on days 1, 8, and 15 of each 4 weeks cycle	10 mg placebo once a day orally plus weekly trastuzumab intravenously at 4 mg/kg loading dose on day 1 with subsequent weekly doses of 2 mg/kg of each 4-week cycle plus paclitaxel intravenously at a dose of 80 mg/m2 on days 1, 8, and 15 of each 4	Targeted therapy: mTOR inhibitor	Placebo	PFS, ORR	30	14	53	NA	NA	NA	18.4	18.2	0.82; (0.61–1.11)

			weeks cycle												
Murthy; et al (2020)	Patients with HER2–positive metastatic breast cancer who has disease progression after therapy with multiple HER2-targeted agents	tucatinib (300 mg orally twice daily throughout the treatment period) in combination with trastuzumab (6 mg/kg intravenously once every 21 days, with an initial loading dose of 8 mg/kg) and capecitabine (1000 mg/m^2) orally twice daily on days 1 to 14 of each 21-day cycle	placebo (300mg orally twice daily) (6 mg/kg intravenously once every 21 days, with an initial loading dose of 8 mg/kg and capecitabine (1000 mg/m^2 orally twice daily on days 1 to 14 of each 21-day cycle	Targeted therapy: kinase inhibitor	Placebo	PFS, OS	320	160	54	21.9	17.4	0.66; (0.50 to 0.88) P = 0.005	7.8	5.6	0.54; (0.42 to 0.71) P<0.001
Vuylsteke; et al (2016)	Patients with hormone receptor-positive, HER2-negative locally recurrent or metastatic BC (mBC).	28-day cycles of intravenous paclitaxel (90 mg/m2 weekly for 3 of every 4 weeks each cycle) with 260 mg pictilisib given orally (pictilisib arm) daily on days 1–5 every week	28-day cycles of intravenous paclitaxel (90 mg/m2 weekly for 3 of every 4 weeks each cycle) with 260 mg placebo given orally (placebo arm) daily on days 1–5 every week	Targeted therapy: kinase inhibitor	Placebo	PFS	91	92	56	NA	NA	NA	8.2 With PIK3CA mutation: 7.3	7.8 With PIK3CA mutation: 5.8	0.95; (0.62–1.46) P = 0.83]. With PIK3CA mutation: 1.06 (0.52–2.12) P = 0.88).

Masuda; et al (2017)	Japanese patients with HER2-negative metastatic breast cancer.	paclitaxel 90mg/m2 on Days 1, 8 and 15 with bevacizumab 10mg/kg on Days 1 and 15, repeated every 4 weeks	paclitaxel 90mg/m2 on Days 1, 8 and 15 with placebo 10mg/kg on Days 1 and 15, repeated every 4 weeks	Targeted therapy: anti-HER2 monoclonal antibodies	Placebo	PFS, OS	24	30	56	25.1	NA	0.67 (0.25–1.81).	12.7	9.2	0.64 (0.29–1.40)
Han; et al (2018)	Patients more than 18 years old with locally recurrent or metastatic breast cancer and a deleterious BRCA1/2 germline mutation	120 mg veliparib BID orally on days 1–7 (21-day cycle). Carboplatin (area under the curve 6 mg/ml/min) and paclitaxel (175 mg/m2) were administered intravenously on day 3.	40 mg veliparib BID orally on days 1–7. Temozolomide started at 150 mg/m2 QD orally on days 1–5 (28-day cycle), and was escalated to 200 mg/m2 at cycle 2 if well-tolerated during the first cycle	Targeted therapy: PARP inhibitor	Chemotherapy	PFS, OS	97	99	46	28.3	25.9	0.750; (0.503–1.117) P=0.156	14.1	12.3	0.789 (0.536–1.162) P=0.227

Targeted therapy versus chemotherapy															
Im; et al (2020)	Asian patients with a germline BRCA1/2 mutation (gBRCAm) and human epidermal growth factor receptor 2 (HER2)-negative metastatic breast cancer who had received ≤2 chemotherapy lines in the metastatic setting.	olaparib tablets (300 mg twice daily)	single agent chemotherapy TPC (21-day cycles of either capecitabine, eribulin or vinorelbine)	Targeted therapy: PARP inhibitor	Chemotherapy	PFS, OS	205	97	45	20.5	20.9	0.98 (0.54–1.78)	5.7	4.2	0.53 (0.29–0.97)
Targeted therapy plus hormonal therapy															
Krop; et al (2016)	Postmenopausal women aged 18 years or older with ER+, HER2-negative breast cancer, resistant to treatment with an aromatase inhibitor in the adjuvant or metastatic setting Part 1: included patients with or without PIK3CA mutations Part 2: included only patients with PIK3CA mutations.	pictilisib (340 mg in part 1 and 260 mg in part 2) starting on day 15 of cycle 1, plus intramuscular fulvestrant 500 mg on day 1 and day 15 of cycle 1 and day 1 of subsequent cycles in both groups	placebo (340 mg in part 1 and 260 mg in part 2) starting on day 15 of cycle 1, plus intramuscular fulvestrant 500 mg on day 1 and day 15 of cycle 1 and day 1 of subsequent cycles in both groups	Targeted therapy: kinase inhibitor	Placebo	PFS	Part 1: 89 Part 2: 41	Part 1: 79 Part 2: 20	63	NA	NA	NA	Part 1: 6.6 Part 2: 5.4	Part 1: 5.1 Part 2: 10.0	Part 1: 0.74 (0.52-1.06) p=0.096 Part 2: 1.07 (0.53–2.18) p=0.84
Hortobagyi; et al (2018)	Post-menopausal women with hormone receptor	Ribociclib 600mg/day; 3 weeks on, 1 week off in 28 days treatment	Placebo 600mg/day; 3 weeks on, 1 week off in 28 days	Targeted therapy: kinase inhibitor	Placebo	PFS, OS	131	88	NA	NA	NA	NA	25.3	16.0	0.568 (0.457-0.704) log-rank P=9.63 x

	positive (HR+), human epidermal growth factor-2 negative (HER2-) advanced breast cancer who had no prior treatment for advanced disease	cycles plus letrozole 2.5mg/day on a continuous schedule	treatment cycles plus letrozole 2.5mg/day on a continuous schedule												10^{-8}
Kornblum; et al (2018)	Postmenopausal women with ER-positive, human epidermal growth factor receptor 2–negative, AI-resistant metastatic breast cancer.	Fulvestrant 500mg intramuscularly days 1, 15 in cycle 1, and day 1 of each subsequent 28-day cycle) and everolimus 10 mg orally once per day	Fulvestrant 500mg intramuscularly days 1, 15 in cycle 1, and day 1 of each subsequent 28-day cycle) and placebo 10 mg orally once per day	Targeted therapy: mTOR inhibitor	Placebo	PFS	66	65	62	28.3	31.4	1.31 (0.72 to 2.38) stratified log-rank test P value = 0.37	10.3	5.1	0.61 (0.40 to 0.92) stratified log-rank P = 0.02
Hormonal therapy only															
Johnston; et al (2013)	Postmenopausal women with HR+ breast cancer ([ER+], [PR+], or both) who relapsed or progressed with locally advanced or metastatic disease on a non-steroidal aromatase inhibitors (NSAI).	Treatment group: fulvestrant (500 mg intramuscular injection on day 1, followed by 250 mg doses on days 15 and 29, and then every 28 days) plus daily oral anastrozole (1 mg); fulvestrant plus anastrozole-matched placebo	Exemestane (25mg)	Hormonal therapy	Hormonal Therapy: Aromatase Inhibitor (AI)	PFS	243+231	249	64	NA	NA	NA	4.4	3.4	1·00, (0·83 to 1·21);p= 0·98)

4.4.3 Chemotherapy

Out of 13 papers, only 1 focused on chemotherapy while the others were focusing on targeted therapy and hormonal therapy. This study by Pallis et al (2012) examined the effect of capecitabine alone versus the combination of vinorelbine and gemcitabine (VG) in MBC patients pre-treated with anthracyclines and taxanes. The outcome of this trial did not demonstrate superiority of vinorelbine/gemcitabine doublet over single agent capecitabine in terms of PFS. In fact, single agent capecitabine was recommended due to lesser toxicity and convenience of oral administration. This study involved 82 patients with hormone receptor (HR) positive (HR+), 27 HR negative (HR-) and 39 patients with unknown receptor status. Median age of the patient was 60 years old and more than 80% of them were post-menopausal. Furthermore, 20 of the treated patients were HER2-positive (HER2+), 82 were HER2-negative (HER2-) and 46 were unknown. It can be seen that most of the studied patients were those with HER2- and HR+ receptors and they found that patients with HR- tumour had significantly longer PFS when treated with capecitabine compared with patients treated with VG. However, the author noted that these analyses were exploratory and not pre-planned by study design thus the conclusion could be invalid due to small sample size.

4.4.4 Targeted Therapy Plus Chemotherapy

A total of 7 papers investigated the combination of targeted therapy with chemotherapy and 4 out of the 7 studies examined these combinations in HER2+ MBC patients. Baselga et al (2012) assessed the efficacy of a combination treatment between double anti-HER2 pertuzumab and trastuzumab plus a chemotherapy docetaxel in HER2+ metastatic breast cancer patients. The median PFS was 12.4 months in the control group, as compared to 18.5 months in the pertuzumab group while overall survival showed more deaths occurred in the control group as compared to the pertuzumab group. This study showed that the combination therapy with two monoclonal antibodies had promising effects in median PFS and overall survival of HER2+ metastatic breast cancer patients. The study involved 388 patients with HR+, 408 HR- and 12 patients with unknown receptor status. The median age for the patient was 54 years old. PFS in prespecified group showed that the hazard ratio for patients with HR+ was 0.72 (95% CI: 0.55–0.95) compared to 0.55 (95% CI: 0.42–0.72) of the

HR- patients. The hazard ratio is the ratio of a chance of an event occurring in the treatment arm compared to the chance of an event occurring in the control arm (Brody, 2016). The hazard ratio of less than 1 means the experimental treatment is better than control; while a hazard ratio of more than 1 means the treatment is worse than control (Barraclough et al, 2011). Thus, in this study, the combination therapy of two monoclonal antibodies was in favour of the HR- patients as the hazard ratio is lower compared to those with HR+.

Another study that involved HER2+ MBC patients, Swain et al (2015) studied the effect of anti-HER2 monoclonal antibody pertuzumab combination versus placebo combination. Both groups were combined with trastuzumab and docetaxel. In this study, the median progression free survival was 56.5 months in the group receiving the pertuzumab combination, as compared with 40.8 months in the group receiving the placebo combination, with a significant difference of 15.7 months. There were 388 HR+ patients favouring pertuzumab with hazard ratio of 0.71 (95% CI: 0.53–0.96) and 408 HR- patients with hazard ratio of 0.61 (95% CI: 0.47–0.81). This showed that the treatment is more effective for HR- patients compared to those with HR+, similar to the finding from the previous author.

Meanwhile, Toi et. al (2017) assessed the efficacy and safety of everolimus (an mTOR inhibitor) for treatment of postmenopausal HER2+ advanced breast cancer in the Asian subset of patients in the BOLERO-1 trial. They investigated the efficacy of everolimus plus trastuzumab and paclitaxel and found that median PFS of this combination treatment showed no significant difference compared to placebo. In this study, the addition of everolimus to trastuzumab and paclitaxel did not improve PFS compared to placebo, however it prolonged PFS in the HR- subpopulations. In the HR-Asian subset, median PFS in the everolimus arm was 25.46 months compared to 14.49 months in the placebo arm (hazard ratio = 0.48; 95% CI 0.29–0.79) which was a 10.97-month improvement in the median PFS. In the HR- subpopulation of the non-Asian subset, median PFS in the everolimus arm was 16.20 months compared to 12.29 months in the placebo arm (hazard ratio = 0.76; 95% CI 0.51–1.15). This trial involved 311 HR-patients, and 406 HR+ patients while there were 1 missing value. This study showed that this treatment might benefit patients with HR- as compared to HR+ patients, which is in line with the previous two studies.

Another study involving HER2+ MBC patients was conducted by Murthy et al (2020) which investigated a kinase inhibitor, tucatinib in combination with an anti-

HER2 agent trastuzumab plus a chemotherapy drug capecitabine in patients with HER2+ metastatic breast cancer. They found that in patients with HER2+ metastatic breast cancer, including those with brain metastases who has previously been treated with trastuzumab, pertuzumab, and trastuzumab emtansine, adding tucatinib to trastuzumab and capecitabine resulted in better PFS and overall survival with median PFS of 7.8 months compared to 5.6 months in placebo. In this study, there were 559 patients with HR+ while 423 patients were HR- with hazard ratio of 0.58 (95% CI: 0.42–0.80) and 0.54 (95% CI: 0.34–0.86) respectively. This showed that this treatment might give similar benefit for both patients either with positive or negative hormone receptors.

Meanwhile, there were two studies that involved HER2 - patients. One study was conducted by Vuylsteke et al (2016) which investigated the effect of a P13K inhibitor, pictilicib combined with a chemotherapy drug paclitaxel versus placebo. The study involved patients with HR+, HER2- locally recurrent or metastatic breast cancer. However, no significant differences were observed in the median PFS for this combination treatment. Since it specifically targets patients with HR+ receptors, no comparison was made among HR- and HR+ patients.

The other study was done by Masuda et al (2017) which investigated another anti-HER2 agent which is bevacizumab in combination with a chemotherapy drug paclitaxel in Japanese patients with HER2- metastatic breast cancer. Median PFS was 9.2 months with placebo–paclitaxel versus 12.7 months with bevacizumab–paclitaxel.

This is consistent with the MERiDiAN ITT population where the median PFS for placebo-paclitaxel was 8.8 months versus 11.0 months with bevacizumab-paclitaxel (Miles et al, 2017). This showed that adding bevacizumab to paclitaxel for HER2- metastatic breast cancer patients significantly improved progression free survival. In this study, number of HR+ were 218 while HR- patients were 36. However, no comparisons were made between the two types of hormone receptor and the hazard ratio between the two receptors was not calculated.

Lastly, for combination of targeted therapy plus chemotherapy, Han et al (2018) investigated the safety and efficacy of a combination of PARP inhibitor (veliparib) with chemotherapy drug carboplatin/paclitaxel (VCP) versus the combination of veliparib with chemotherapy drug temozolomide (VT) in patients with BRCA1/2 -mutated breast cancer. Median PFS for VCP and placebo showed significant difference of 14.1 months and 12.3 months respectively in favour of the PARP inhibitor. Demographically there were 167 patients with HR+, 120 triple negative, 15 HER2+, 153 with BRCA1 mutations and 133 BRCA2 mutations, but no correlations between these criteria were mentioned corresponding to the result.

From these studies, it can be seen that the treatment for HER2+ MBC patients showed longer median PFS in HR- patients as compared to HR+ patients. This is consistent in 3 out of 4 studies involving HER2+ MBC while the other study showed no significant difference between both receptors.

As for studies involving HER2- patients, both studies did not investigate the median PFS between HR+ and HR- patients. However, while the combination of pictilicib (kinase inhibitor) with paclitaxel did not show any significant difference in HER2- patients, the combination of bevacizumab (anti-HER2 monoclonal antibody) plus paclitaxel significantly improved progression free survival for these patients. Meanwhile, for MBC patients with BRCA1/2 mutations, the combination of veliparib (PARP inhibitor) with carboplatin/paclitaxel (VCP) showed better median PFS compared to the combination of veliparib with temozolomide (VT) regardless of the BRCA1/2 mutation status.

4.4.5 Targeted Therapy Versus Chemotherapy

There was only one study in this category, which was done by Im et al (2020) that also assessed the efficacy of a PARP inhibitor, oliparib over chemotherapy treatment of physician’s choice (either capecitabine, eribulin or vinorelbine). This study was conducted in Asian patients with a germline BRCA1/2 mutation (gBRCAm) and HER2- metastatic breast cancer. Their result showed that olaparib achieved longer median PFS compared to the chemotherapy treatment with 5.7 months compared to 4.2 months respectively. This is consistent with the previously mentioned study (Han et al, 2018). In this study, 194 patients were HR+, while 195 of them were triple negative breast cancer patients. However, the author did not measure the PFS in relation to the receptors. It can be concluded that in patients with BRCA1/2 mutations, the oliparib tablet monotherapy achieved longer median PFS compared to the chemotherapy treatment.

4.4.6 Hormonal Therapy Plus Targeted Therapy

There were 3 studies that studied the effects of combination therapy between hormonal and targeted therapy. The first one was conducted by Krop et al (2016) which studied the effect of a kinase inhibitor, pictilicib in combination with a hormonal therapy fulvestrant in postmenopausal women with ER+, HER2- breast cancer. They also investigated whether the presence or absence of PIK3CA mutation affect the efficacy of this treatment. Since the study focused on ER+ patients, no HR- patients were enrolled. However, there were a subgroup of patients who were PR+ (116 patients) and some were PR- (35 patients). Even though the result showed no significant difference in median PFS regardless of the PIK3CA mutational status, there were improvement in median PFS of the PR+ patients with hazard ratio of 0.44 (95% CI 0.28-0.69).

Then, in 2018, Hortobagyi et al, assessed the efficacy of a cyclin dependent kinase (CDK) inhibitor ribociclib plus a hormonal agent letrozole in postmenopausal women with ER+, HER2–, aromatase inhibitor (AI) -resistant metastatic breast cancer. Median PFS for ribociclib plus letrozole was 25.3 months which is significantly different as compared to 16.0 months of PFS for placebo. This treatment showed significant result regardless of PIK3CA or TP53 mutation status, or CDKN2A, CCND1 and ESR1 mRNA levels. Since the study focused on ER+ patients, there were no comparisons between the positive and negative receptors.

The last one in this category was done by Kornblum et al in 2018. This study included 131 postmenopausal women with ER+, HER2-, AI-resistant metastatic breast cancer who were randomly assigned to fulvestrant plus everolimus (mTOR inhibitor) or fulvestrant plus placebo. As compared to previous study which combined everolimus with trastuzumab and paclitaxel (Toi et al, 2017), this study was done by combining everolimus to a hormonal therapy, fulvestrant. Their study showed that the addition of everolimus to fulvestrant improved the median progression free survival from 5.1 to 10.3 months and this proved that the addition of everolimus enhances the efficacy of fulvestrant in AI-resistant, ER+ metastatic breast cancer. Since the study focused on ER+ patients, there were no comparisons between the positive and negative receptors. From these three studies, it can be seen that the combination of hormonal therapy and targeted therapy were used on ER+, HER2- patients. Two of the studies combined targeted therapy with fulvestrant while the other one used letrozole. Krop et al (2016) showed that the combination of a kinase inhibitor pictilicib with fulvestrant did not show any significance difference in median PFS of the patients. However, the combination of an mTOR inhibitor, everolimus with fulvestrant showed a significant PFS difference (Kornblum et al, 2018). Meanwhile, the combination of a kinase inhibitor, ribociclib with hormonal agent, letrozole showed a significant difference in the median PFS of the patients regardless of mutation status of PIK3CA or TP53, or CDKN2A, CCND1 and ESR1 mRNA levels (Hortobagyi et al, 2018).

However, since these studies focused on ER+ patients, no comparison with the negative receptors can be concluded.

4.4.7 Hormonal Therapy

The only study comparing hormonal therapy versus another hormonal therapy was conducted by Johnston et al (2013). They investigated the combination of fulvestrant with an aromatase inhibitor (AI) anastrozole compared to treatment with an AI (exemestane) alone in postmenopausal women with HR+ breast cancer ([ER+], [PR+], or both). This study revealed that there were no significant differences in median PFS between combination of hormonal therapy and AI versus AI alone. However, they found that patients with known ER+ and PR+ tumours seemed to show the greatest benefit for fulvestrant plus anastrozole.

4.4.8 Summary of Findings

The summary of the systematic review findings with regards to HER2 and hormone receptor status is as stated in Table 4.10.

Table 4.10
Summary of the Treatments Based on HER2 and Hormone Receptor Status

HER2 status	Summary
HER2+	longer median PFS in HR- patients as compared to HR+ patients.
HER2-	improved median PFS in PR+ ER+ patients

As can be seen from the summary table, in the treatment of chemotherapy (alone and combined with targeted therapy), both treatments favoured the HR- patients.

Meanwhile, for hormonal therapy alone and combined with targeted therapy, the treatments were in favour of ER+ PR+ patients. However, comparison with HR- receptors was not conducted as it was not discussed in the literature filtered.

4.4.9 Discussion

4.4.9.1 HER2-Positive

Based on the previous findings, the first thing to address in this systematic review is whether HER2+, ER- and PR- patients are more prone to metastatic disease. Generally, high PFS meant lower probability to contribute to MBC occurrence since it took longer to progress and vice versa for low PFS. Findings from the systematic review showed that HER2+/HR- patients treated with targeted therapy plus chemotherapy showed higher PFS compared to placebo as shown by Baselga et al (2012), Swain et al (2015) and Toi et al (2017). This finding is similar to Sim et al (2019) who found in HER2+ MBC patients, the treatment of targeted therapy plus chemotherapy is more favourable towards HR- patients, as opposed to HR+ patients. The findings from Saura et al (2020) is also consistent with this in which treatment with targeted therapy plus chemotherapy showed that in patients with HER2+ MBC, HR- derived the greatest PFS benefit if compared to HR+. This is in line with our hierarchical clustering result (Figure 4.20) that clustered the three hormone receptors (HER2+, ER- and PR-) together with chemotherapy. However, even though this clustering without the adjuvant radiation and OncoTree code did not cluster these factors with MBC patients, the hierarchical clustering that contained the two variables (Figure 4.19) actually clustered HER2+, ER- and PR- and MBC patient in a same cluster albeit at the distance of more than 1. Since the distance is quite far (almost 5) in our hierarchical clustering, it can be said that the connection of HER2+, ER- and PR- with MBC is not too strong, which is in line with the systematic review findings (Baselga et al, 2012; Swain et al, 2015; Toi et al 2017) that showed these combinations has higher PFS (viewed as less contribution towards MBC).

Moreover, despite the fact that these findings reached to a similar conclusion that HER2+/HR- patients showed higher survival rate, there are still a lot of other parameters that needed to be considered. For instance, the type of therapy used, duration of the treatment, as well as previous treatments received by the patients. Baselga and

Swain used two anti-HER2 antibody (trastuzumab and pertuzumab) with combination of docetaxel and the treatment was given every 3 weeks for a median of 8 weeks while the median duration of study treatment was 18.1 months. Meanwhile, the study by Saura investigated a kinase inhibitor e.g neratinib in combination with capecitabine. The treatment was a 21-day cycle with no break between the cycles for a duration of average 9.5 months of the active treatment phase. While both of these studies involved different kind of drugs, the duration of treatment was also different and both treatments on average did not exceed 2 years. As shown in the ExteNET trial (Martin et al, 2017), patients treated with neratinib for additional one year after being treated with trastuzumab-based adjuvant therapy showed better disease-free survival in HER2+/HR+ with hazard ratio of 0·60 (95% CI 0·43–0·83), compared to HER2+/HR- patients that showed hazard ratio of 0·95 (95% CI 0·66–1·35). Thus, it can be said that a different combination of factors will potentially lead to a different outcome.

Although both the clinical profiling and systematic review findings are agreeable, the type and duration of each treatment in the review was different and the data used in the data mining did not include targeted therapy. Furthermore, there were also the matter of discordance and concordance of HER2 status as discussed previously, and none of this was investigated in the articles that were reviewed. Therefore, even though the survival rate showed better PFS for HER2+/HR- patients in the systematic review, and these receptors were clustered together in the hierarchical clustering, other parameters still needed to be considered, as combination of different factors might change the outcome of MBC or non-MBC. Nevertheless, result of the clinical profiling does corroborate with the systematic review to a certain extent.

4.4.9.2 HER2-Negative

As for the HER2- patients, in this systematic review, it was found that the treatment of hormonal therapy plus targeted therapy showed a slightly different result between median PFS in HER2-negative, HR-positive (HER2-/HR+) MBC patients. While Krop et al (2016) found no significant differences in the PFS, Hortobagyi et al (2018) and Kornblum et al (2018) showed a significant increase in PFS when HER2-/HR+ MBC patients were treated with hormonal therapy plus targeted therapy.

However, similar to the HER2+ outcome, the difference could be due to the different types of drugs used as well as the duration of the treatment. Krop et al (2016) tested the kinase inhibitor (pictilisib) together with fulvestrant for a median exposure of 2.9 months, while Hortobagyi et al (2018) tested kinase inhibitor (ribociclib) in combination with letrozole for a median duration treatment of 20.2 months. There was a huge gap in terms of months of exposure between the two studies. Meanwhile Kornblum et al (2018) tested the mTOR inhibitor everolimus in combination with fulvestrant for a median duration of 22 weeks. This showed that the duration of treatment could also affect the efficacy of treatment. The longer the duration of treatment, the median progression free survival seemed to improve. This is in line with the hierarchical clustering previous findings (Figure 4.20) where HER2- were clustered together with ER+, HR+ and hormonal therapy more than 2 years.

Aside from that, in the hierarchical clustering as seen in Figure 4.20, HER2- was clustered together with ER+, PR+, age_5, hormonal therapy and patient BC. This seemed in line with the findings from the systematic review especially since all 3 studies on HER2- patients were those in the post-menopausal age and showed higher median PFS in HR+ receptors (less prone to progress to MBC). However, it has to be noted that during the data analysis, the data provided were only on the class of the treatment (hormonal or chemotherapy), not on the detailed treatment itself. Plus, data of targeted treatment was not available during the analysis thus it cannot be ascertained to which cluster this treatment are clustered with. Therefore, it can be said that the findings from this systematic review do corroborate with the previous findings, since HER2- and HR+ were indeed clustered together with patient BC in the hierarchical clustering, however HER2 and HR receptors are not the only determining factors for MBC occurrence. Other factors such as the duration of the treatment, type of the treatment, as well as other factors that was not discussed in the review such as the mRNA profile and gene mutations should not be neglected. The combination of all of these factors might give a clearer picture on which combination of factors will lead to MBC.

CHAPTER FIVE
CONCLUSION

The first objective of this study was to mine and integrate clinical, phenotype and genotype data to identify factors that contributed to the occurrence of metastatic breast cancer. To achieve this, first PCA was done for the mRNA expression and MCA for CNA profiles between: (i) breast samples of breast cancer vs metastatic breast cancer and (ii) breast samples vs metastatic samples. In addition, odds-ratio calculation was also performed for the two comparisons. The rationale of this was to identify and connect the genes involved for the migration and colonisation of breast cancer cells from primary to distant or metastatic sites. Additionally, clinical profile of metastatic breast cancer was also conducted to provide a holistic view on the occurrence of metastatic breast cancer. Then, predictive models were built based on the mRNA expression and clinical profile of metastatic breast cancer patients that were identified. This is to fulfil the second objective which was to build a prediction model that can predict possibility of occurrence of the metastatic state of breast cancer based on factors previously determined. Finally, a systematic review was carried out to achieve the third objective which was to further validate the findings of this study by comparing it with what has been published in literature. Hence, based on our findings, several conclusions can be made.

Firstly, the 15 genes identified through feature selection could provide insights into the occurrence of metastatic breast cancer, and differentiate between breast cancer and metastatic breast cancer patients. When analysing the mRNA expression profile of breast cancer samples of breast cancer and metastatic breast cancer patients, a clear separation was observed between the groups. This was further supported by the predictive models built which shows a specificity and sensitivity of 1.0 and 0.922 respectively in the internal validation. Further external validation showed the same specificity and sensitivity of 1.0 and 0.922 respectively. As previously mentioned, there are five main steps to metastasis, which are detachment, cell migration and invasion, intravasation, extravasation and growth of secondary tumour. From conducting the mRNA profiling, feature selection and odds-ratio calculation, the 15 genes identified could be connected to all five processes, although the majority of the genes seem to be involved in cell migration and invasion. For example, the Fibroblast

Growth Factor (FGF-4) gene that was involved in various cell biology processes such as cell differentiation, morphogenesis and cell proliferation while another gene called Melanoma Antigen Family A, 9B (MAGEA9B) have been reported to affect the biological characteristic of cancer cells such as migration, metastasis and invasion.

Secondly, when curating the path to metastasis, two genes which are YAP1 and SP7, was mapped to phenotypes that are pro- metastasis such as apoptosis. The path to these genes could be traced back to ESR1 (oestrogen receptor).

Thirdly, clinical factors could be used to differentiate between breast cancer and metastatic breast cancer to a certain extent. When profiling the two groups based on clinical factors such as age, chemotherapy, hormone therapy, ER, PR and HER2 status using MCA, a clear separation could be seen. This was further supported by the predictive model where the internal validation shows a specificity and sensitivity of 0.942 and 0.722 respectively. Further external validation showed a specificity and sensitivity of 0.988 and 0.0 respectively. It has to be noted that the 0 sensitivity of the external validation was due to lack of metastatic data which affect its predictive capability. However, the significance of these factors was later validated in the systematic review. As an alternative to strengthen the findings, a systematic review (SR) was conducted to compare what was found, with what has been published in literature. Based on the SR findings, the HER2-/HR+ patients aged more than 60 who received hormonal therapy showed higher PFS (viewed as less contribution towards MBC) which was in line with the hierarchical clustering that clusters all of these factors together with non-MBC patient. Thus, even though lack of data hinders the ability of the prediction model to predict the occurrence of metastatic breast cancer, findings from the SR showed supporting evidence towards the clinical profiling.

Fourthly, **the occurrence of three types of OR namely OR9G4, OR5J2 and OR5T2 in this work showed there could be an underlying mechanism connecting metastasis to the olfactory transduction as seen in the KEGG pathway mapping.** This was further strengthened when the decision tree for the mRNA_15 training set found the most important feature is the gene OR5T2, with a score of 0.93. Also several genes such as FGF4, KRT6B, KRTAP25-1, LINC00943, MAGEA9B, OR5B, OR9G4 and SCGB1D1 may suggest potential link between breast cancer and its metastatic sites. As shown by the mRNA_7 training set, the most important feature is the gene SCGB1D1, with a score of 0.95. Several pathways such as regulation of actin cytoskeleton, IL-7 signaling and ECM-receptor interaction pathways may shed some light on the metastatic occurrence as well.

Based on all of these evidence, one significant conclusion could be drawn which is**, by identifying key genes and clinical factors, the metastatic state of breast cancer patients can be predicted.** Thus, this work showed persuasive evidence that by identifying these key factors, the occurrence of metastatic breast cancer can eventually be predicted which could potentially aid in the clinical management of the disease in the future. For example, by identifying key contributing factors, clinicians can personalise treatment plans based on individual patient characteristics which allows for tailored treatment strategies. Furthermore, some contributing factors can serve as predictive markers for treatment response. Thus, by identifying these factors, clinicians can tailor treatment choices and avoid ineffective treatments, minimising unnecessary toxicity and optimising treatment outcomes. Monitoring key contributing factors over time can also aid in disease surveillance and response evaluation which allows for timely adjustments to treatment plans and the consideration of alternative strategies to better manage the evolving disease.

CHAPTER SIX
LIMITATIONS AND FUTURE DIRECTION

6.1 Limitation

To build a prediction model, a vast amount of data is needed to ensure its specificity and sensitivity. However, in this study, since genetic data outnumbers clinical data, the combination of both domain in one predictive model or combining the results of both predictive models cannot be done. Initially, the plan was to get data from local hospitals to validate the models since ethics approval has already been obtained (attached as Appendix 2). However, this plan was hindered due to the pandemic, thus there was a limitation on validating the findings. As an alternative, a systematic review was conducted to validate the findings, but not all of the parameters covered in the first two phases could be included in the systematic review, for example genes and radiation therapy. Though 'gene' was also used as a keyword to search for articles, not many papers integrated this with clinical parameters. Therefore, only some of our clinical factors can be validated by the systematic review.

6.2 Future Direction

Given that this disease is multifactorial, the use of a predictive model could be used as a diagnostic tool. Future work should include the combination of both domain (genotype and clinical) in one predictive model or combining the results of both predictive model. This could not be done currently as the genetic data outnumbers clinical data, and hence combining both data would result in a small number of data points and the model may suffer from the curse of dimensionality. Additionally, real data from hospitals should also be retrieved to build a more dynamic model with high sensitivity and specificity to predict the occurrence of metastatic breast cancer.

REFERENCES

Adjuvant therapy: Treatment to keep cancer from returning - Mayo Clinic [Internet]. [cited 2018 Nov 17]. Available from: https://www.mayoclinic.org/diseases-conditions/cancer/in-depth/adjuvant-therapy/art-20046687.

Abdi, H. and Valentin, D. (2007) Multiple Correspondence Analysis. In: Salkind, N.J., Ed., Encyclopedia of Measurement and Statistics, SAGE Publications, Thousand Oaks, CA, 1-13.

Ahmed A., Ali A., Ali S., Ahmad A., Philip S. F. (2012). Breast Cancer Metastasis and Drug Resistance. *Breast Cancer Metastasis Drug Resist*, 2012, 1–18.

Alsarraj J., & Hunter K. W. [Internet]. [cited 2019 Oct 21]. Bromodomain- Containing Protein 4: A Dynamic Regulator of Breast Cancer Metastasis through Modulation of the Extracellular Matrix - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/The- steps-of-the-metastatic-cascade_fig1_221796415.

Althobiti M., Muftah A. A., Aleskandarany M. A., Joseph C., Toss M. S., Green A., & Rakha, E. (2020). The prognostic significance of BMI1 expression in invasive breast cancer is dependent on its molecular subtypes. *Breast cancer research and treatment, 182*(3), 581–589.

Alkhathami A. G., Verma A. K., Alfaifi M., Kumar L., Alshahrani M. Y., Hakami A. R., Alshehri O. M., Asiri M., Beg M. M. A. Role of miRNA-495 and NRXN-1 and CNTN-1 mRNA Expression and Its Prognostic Importance in Breast Cancer Patients. (2021) *Journal of Oncology*, vol. 2021, Article ID 9657071, 10 pages, 2021.

American Cancer Society. Breast Cancer Stages [Internet]. [cited 2018 Sept 20]. Available from: https://www.cancer.org/cancer/types/breast-cancer/understanding-a-breast-cancer-diagnosis/stages-of-breast-cancer.html

Amir E., Evans D. G., Shenton A., Lalloo F., Moran A., Boggis C., Wilson M., & Howell A. (2003). Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. *J Med Genet* 40:807-814.

Anaconda Navigator (Version 1.9.12) [Computer software]. (2016). Retrieved from https://www.anaconda.com/products/navigator

Analytics Vidhya Content Team. (2020). PCA: A Practical Guide to Principal Component Analysis in R & Python.

https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/. Retrieved on 27th July 2022.

Aoki-Kinoshita K. F. & Kanehisa M. (2007). Gene Annotation and Pathway Mapping in KEGG. In: Bergman N.H. (eds) Comparative Genomics. Methods In Molecular Biology™, vol 396. Humana Press.

Aranda S., Schofield P., Weih L., et al. (2005). Mapping the quality of life and unmet needs of urban women with metastatic breast cancer. *Eur J Cancer Care*, 14, 211-22.

Arslan C., Sari E., Aksoy S., & Altundag K. (2011). Variation in hormone receptor and HER-2 status between primary and metastatic breast cancer: review of the literature, *Expert Opinion on Therapeutic Targets*, 15(1), 21-30.

Arvelo F., Sojo F., & Cotte C. (2016). Tumour progression and metastasis. *Ecancer*, 10, 617.

Asri H., Mousannif H., Al Moatassime H., & Noel T. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science*, 83, 1064 – 1069.

Audeh M. W. (2014). Novel treatment strategies in triple-negative breast cancer: specific role of poly (adenosine diphosphate-ribose) polymerase inhibition. *Pharmgenomics Pers Med*, 7, 307-316.

Baillet A, Le Bouffant R, Volff JN, Luangpraseuth A, Poumerol E, Thépot D, Pailhoux E, Livera G, Cotinot C, Mandon-Pépin B. TOPAZ1, a novel germ cell-specific expressed gene conserved during evolution across vertebrates. PLoS One. 2011;6(11):e26950. doi: 10.1371/journal.pone.0026950. Epub 2011 Nov 1. PMID: 22069478; PMCID: PMC3206057.

Barraclough H., Simms L., & Govindan R. (2011). Biostatistics primer: what a clinician ought to know: hazard ratios. *Journal of thoracic oncology: official publication of the International Association for the Study of Lung Cancer*, 6(6), 978–982.

Baselga J., Campone M., Piccart M., et. al. (2012). Everolimus in postmenopausal hormone-receptor positive advanced breast cancer. *N Engl J Med*, 366(6), 520-529.

Belle L., Ali N., Lonic A., Li X., Paltridge J. L., Roslan S., Herrmann D., Conway J. R., Gehling F. K., Bert A. G., Crocker L. A., Tsykin A., Farshid G., Goodall G. J., Timpson P., Daly R. J., & Khew-Goodall Y. (2015). The tyrosine phosphatase PTPN14 (Pez) inhibits metastasis by altering protein trafficking. *Sci Signal*, 8(364), ra18.

Bellman R. E. (1957) Dynamic programming. Princeton University Press, Princeton.

Bendell J., Saleh M., Rose A. A., et. al. (2014). Phase I/II study of the antibody- drug conjugate glembatumumab vedotin in patients with locally advanced or metastatic breast cancer. *J Clin Oncol*, 32(32), 3619-3625.

Berg W. A. (2009). Tailored supplemental screening for breast cancer: what now and what next? *AJR*, 192, 390–399.

Beroukhim R., Mermel C. H., Porter D., Wei G., Raychaudhuri S., Donovan J., Barretina J., Boehm J. S., Dobson J., Urashima M., Mc Henry K. T., Pinchback R. M., Ligon A. H., Cho Y. J., Haery L., Greulich H., et. al. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature*, 463, 899–905.

Blood Marker Tests: Diagnosing and Monitoring Breast Cancer [Internet]. [cited 2018 Oct 23]. Available from: https://www.breastcancer.org/symptoms/testing/types/blood_marker.

Boulding T., McCuaig R. D., Tan A., Hardy K., Wu F., Dunn J., Kalimutho M., Sutton C. R., Forwood J. K., Bert A. G., et al. (2018). LSD1 activation promotes inducible EMT programs and modulates the tumour microenvironment in breast cancer. *Sci. Rep*, 8, 73.

Bozzetti, C., Negri, F. V., Lagrasta, C. A., Crafa, P., Bassano, C., Tamagnini, I., Gardini, G., Nizzoli, R., Leonardi, F., Gasparro, D., Camisa, R., Cavalli, S., Silini, E. M., & Ardizzoni, A. (2011). Comparison of HER2 status in primary and paired metastatic sites of gastric carcinoma. British journal of cancer, 104(9), 1372–1376. https://doi.org/10.1038/bjc.2011.121.

Brabletz T., Kalluri R., Nieto M. A., & Weinberg R. A. (2018). EMT in cancer. *Nat Rev Cancer*, 18(2), 128–34.

Breast Cancer Stages. Breastcancer.org [Internet]. [cited 2019 Feb 18]. Available from: https://www.breastcancer.org/symptoms/diagnosis/staging.

Brody T. (2016) Chapter 9 - Biostatistics—Part I, Editor(s): Tom Brody, Clinical Trials (Second Edition), Academic Press, Pages 203-226, ISBN 9780128042175.

Brownlee J. How to Calculate Feature Importance With Python [Internet]. [cited 2020 Jul 8]. Available from: https://machinelearningmastery.com/calculate- feature-importance-with-python/

Buonadonna A., Crivellari D., Frustaci S., et. al., (1997). Vinorelbine (VRL) as palliative treatment in elderly patients with metastatic breast cancer. *Breast Cancer Res Treat*, 46, 96a.

Burotto M., Chiou V. L., Lee J. M., & Kohn E. C. (2014). The MAPK pathway across different malignancies: a new perspective. *Cancer, 120*(22), 3446-56.

Cai P., Lu Y., Yin Z., Wang X., Zhou X., & Li Z. (2021). Trophinin Is an Important Biomarker and Prognostic Factor in Osteosarcoma: Data Mining from Oncomine and the Cancer Genome Atlas Databases. Biomed Research International, 2021, 6885897. doi:10.1155/2021/6885897

Cantor DI, Cheruku HR, Nice EC, Baker MS. Integrin alphavbeta6sets the stage for colorectal cancer metastasis. Cancer MetastasisRev. 2015;34(4):715–734.

Cao S. S. & Lu C. T. (2016). Recent perspectives of breast cancer prognosis and predictive factors (Review). *Oncology Letters*, 12, 3674-3678.

Cerami E., Gao J., Dogrusoz U., Gross B. E., Sumer S. O., Aksoy B. A., Jacobsen A., Byrne C. J., Heuer M. L., Larsson E., Antipin Y., Reva B., Goldberg A. P., Sander C., & Schultz N. (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. ;2(5):401-4.

Chambers A. F., Groom A. C., & MacDonald I. C. (2002). Dissemination and growth of cancer cells in metastatic sites. *Nat. Rev. Cancer*, 2, 563–572.

Chang X. Z., Yu J., Zhang X. H., Yin J., Wang T., & Cao X. C. (2009). Enhanced expression of trophinin promotes invasive and metastatic potential of human gallbladder cancer cells. J*ournal of Cancer Research and Clinical Oncology*, 135(4), 581–590.

Chatterjee A., Rodger E. J., Eccles M. R. (2018). Epigenetic drivers of tumourigenesis and cancer metastasis. *Seminars in Cancer Biology*, 51, 149-159.

Cheadle C., Vawter M. P., Freed W. J., & Becker K. G. (2003). Analysis of microarray data using Z score transformation. The Journal of molecular diagnostics: JMD, 5(2), 73–81.

Chen B., Wang J., Dai D., et. al. (2017). AHNAK suppresses tumour proliferation and invasion by targeting multiple pathways in triple-negative breast cancer. *J Exp Clin Cancer Res, 36*, 65.

Chen D., Sun Y., Wei Y., Zhang P., Rezaeian A. H., Teruya-Feldstein J., Gupta S., Liang H., Lin H. K., Hung M. C., & Ma L. (2012). LIFR is a breast cancer metastasis suppressor upstream of the Hippo-YAP pathway and a prognostic marker. *Nat Med, 18*(10), 1511-7.

Chen X., Zhang G., Chen B., Wang Y., Guo L., Cao L., Ren C., Wen L., & Liao N. (2019). Association between histone lysine methyltransferase KMT2C mutation and clinicopathological factors in breast cancer. *Biomed Pharmacother*, 116, 108997.

Cheng K. C., Katz S. R., Lin A. Y., Xin X., & Ding Y. (2016). Chapter Four - Whole-Organism Cellular Pathology: A Systems Approach to Phenomics. *Advances in Genetics*, 95, 89-115.

Cheng L., Swartz M. D., Zhao H., Kapadia A. S., Lai D., Rowan P. J., Buchholz T. A., & Giordano S. H. (2012). Hazard of recurrence among women after primary breast cancer treatment–A 10-year follow-up using data from SEER-Medicare. *Cancer Epidemiol. Biomarkers,* 21, 800–809.

Chia J., Kusuma N., Anderson R., Parker B., Bidwell B., Zamurs L., Nice E., Pouliot N. (2007). Evidence for a role of tumor-derived laminin-511 in the metastatic progression of breast cancer. *Am J Pathol*, 170(6), 2135-48.

Chia S. K., Speers C. H., Kang D. Y., Malfair-Taylor A., Barnett S., Coldman A., Gelmon K. A., O'Reilly S. E., & Olivotto I. A. (2007), The impact of new chemotherapeutic and hormone agents on survival in a population-based cohort of women with metastatic breast cancer. *Cancer*, 110, 973-979.

Cianfrocca M. & Goldstein L. J. (2004). Prognostic and predictive factors in early-stage breast cancer. *Oncologist,* 9, 606-616.

Colditz G. A., Kaphingst K. A., Hankinson S. E., & Rosner B. (2012). Family history and risk of breast cancer: nurses' health study. *Breast Cancer Res Treat*, 133(3), 1097-1104.

Cong M., Li J., Jing R., et al. (2016). Long non-coding RNA tumor suppressor candidate 7 functions as a tumor suppressor and inhibits proliferation in osteosarcoma. *Tumour Biol*, 37, 9441-50.

Costa P. S., Santos N. C., Cunha P., Cotter J., & Sousa N. (2013). The Use of Multiple Correspondence Analysis to Explore Associations between Categories of Qualitative Variables in Healthy Ageing. *Journal of Aging Research*, 2013, 302163.

Curtis C., Shah S., Chin, SF., et. al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486, 346–352.

Daoud J. I. (2017). Multicollinearity and Regression Analysis. *J. Phys.: Conf. Ser*. 949, 012009.

Darby S., McGale P., et. al. (2011). Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta- analysis of individual patient data for 10 801 women in 17 randomised trials. *Lancet,* 378, 1707–1716.

Davies C., Godwin J., Gray R., Clarke M., Cutter D., Darby S., McGale P., Pan H. C., Taylor C., Wang Y. C., et. al. (2011). Early Breast Cancer Trialists' Collaborative Group (EBCTCG): Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: Patient- level meta-analysis of randomised trials. *Lancet,* 378, 771-784.

De Blasio A., Di Fiore R., Morreale M., Carlisi D., Drago-Ferrante R., Montalbano M., Scerri C., Tesoriere G., & Vento R. (2019). Unusual roles of caspase-8 in triple-negative breast cancer cell line MDA-MB-231. *Int J Oncol* 48, 2339-2348.

De Groot J. S., Ratze M. A., van Amersfoort M., Eisemann T., Vlug E. J., Niklaas M. T., Chin S. F., Caldas C., van Diest P. J., Jonkers J., de Rooij J., & Derksen P. W. (2018). αE-catenin is a candidate tumor suppressor for the development of E-cadherin-expressing lobular-type breast cancer. *J Pathol,* 245(4), 456-467.

Delen, D. (2009). Analysis of cancer data: A data mining approach. *Expert Syst,* 26, 100–112.

Demsar J., Zupan B., Aoki N., Wall M. J., Granchi T. H., & Beck J. R. (2001). Feature mining and predictive model construction from severe trauma patient's data. *International Journal of Medical Informatics,* 63, 41–50.

Detection and screening methods for breast cancer [Internet]. [cited 2018 Oct 23]. Available from: https://nbcf.org.au/about-national-breast-cancer-foundation/about-breast-cancer/what-you-need-to- know/detection/detection-and-screening-methods/.

Diana D., Surendran A., Jissa V., & Nair A. (2018). Regulation of CNKSR2 protein stability by the HECT E3 ubiquitin ligase Smurf2, and its role in breast cancer progression. *BMC Cancer*. 18. 10.1186/s12885-018-4188-x.

Dillekas H., Rogers M. S., & Straume O. (2019). Are 90% of deaths from cancer caused by metastases? Cancer Med. 2019 Sep;8(12):5574-5576.

Ding Y. C., Yu W., Ma C., et. al. (2014). Expression of long non- coding RNA LOC285194 and its prognostic significance in human pancreatic ductal adenocarcinoma. *Int J Clin Exp Pathol*, 7, 8065-70.

Dite G. S., Mahmoodi M., Bickerstaffe A., Hammet F., Macinnis R. J., Tsimiklis H., Dowty J. G., Apicella C., Phillips K. A., Giles G. G., Southey M. C., & Hopper J. L. (2013). Using SNP genotypes to improve the discrimination of a simple breast cancer risk prediction model. *Breast cancer research and treatment*, 139(3), 887–896.

Early Breast Cancer Trialists' Collaborative Group (EBCTCG). (2018). Long-term outcomes for neoadjuvant versus adjuvant chemotherapy in early breast cancer: meta-analysis of individual patient data from ten randomised trials. *Lancet Oncol*, 19(1), 27–39.

Early Breast Cancer Trialists' Collaborative Group (EBCTCG). (2014). Effect of radiotherapy after mastectomy and axillary surgery on 10-year recurrence and 20-year breast cancer mortality: meta-analysis of individual patient data for 8 135 women in 22 randomised trials. *Lancet*, 383, 2127–2135.

Edgerton S. M., Moore D., Merkel D., & Thor A. D. (2003). ErbB-2 (HER-2) and breast cancer progression. *Appl Immunohistochem Mol Morphol*, 11, 214- 21.

Fattahi F, Kiani J, Alemrajabi M. et al. (2021). Overexpression of DDIT4 and TPTEP1 are associated with metastasis and advanced stages in colorectal cancer patients: a study utilizing bioinformatics prediction and experimental validation. Cancer Cell Int 21, 303

Feng Y., Spezia M., Huang S., Yuan C., Zeng Z., Zhang L., Ji X., Liu W., Huang B., Luo W., Liu B., Lei Y., Du S., Vuppalapati A., Luu H. H., Haydon R. C., He T. C., & Ren G. (2018). Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *Genes & Diseases*, 5, 77-106.

Ferlay J., Soerjomataram I., & Dikshit R. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*, 136, E359-E86.

Fernandez Y., Espana L., Manas S., et. al. (2000). Bcl-xL promotes metastasis of breast cancer cells by induction of cytokines resistance. *Cell Death Differ*, 7, 350–9.

Fidler I. J. (2003). The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. *Nat. Rev. Cancer,* 3, 453–458.

Fife C. M., McCarroll J. A., & Kavallaris, M. (2014). Movers and shakers: cell cytoskeleton in cancer metastasis. *British journal of pharmacology, 171*(24), 5507–5523.

Finger E. C. & Giaccia A. J. (2010). Hypoxia, inflammation, and the tumor microenvironment in metastatic disease. *Cancer Metastasis Rev* 29, 285– 293.

Fisher B., Bauer M., Wickerham D. L., Redmond C. K., Fisher E. R., Cruz A. B., et. al. (1983). Relation of number of positive axillary nodes to the prognosis of patients with primary breast cancer. An NSABP update. *Cancer*, 52, 1551-7.

Fitzmaurice G. M. & Laird N. M. (2001). Multivariate Analysis: Discrete Variables (Logistic Regression). *International Encyclopedia of the Social & Behavioral Sciences*, 10221–10228.

Flamant L., Roegiers E., Pierre M., Hayez A., Sterpin C., De Backer O., et. al. (2012). TMEM45A is essential for hypoxia-induced chemoresistance in breast and liver cancer cells. *BMC Cancer,* 12, 391.

Foulkes W. D., Smith I. E., & Reis-Filho J. S. (2010). Triple-negative breast cancer.

N Engl J Med, 363(20), 1938-1948.

Freddie B., Jacques F., Isabelle S., Rebecca L., Siegel, Lindsey A., & Torre A. J. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *Cancer J Clin 2018,* 68, 394–424.

Fu X. R., Wan W. J., Sun Z. C., Zhang X. D., Nan F. F., Ge J. R., Xia Y. Q., & Zhang, M. Z. (2020). Expression of CD7 and its correlation with prognosis in patients with NK/T-cell lymphoma. *Zhonghua xueyexue zazhi*, 41(11), 921–926.

Gail M. H., Brinton L. A., Byar D. P., Corle D. K., Green S. B., Schairer C., & Mulvihill J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*, 81(24), 1879–1886.

Gala K., Li Q., Sinha A., Razavi P., Dorso M., Sanchez-Vega F., Chung Y. R., Hendrickson R., Hsieh J. J., Berger M., Schultz N., Pastore A., Abdel- Wahab O., & Chandarlapaty S. (2018). KMT2C mediates the estrogen dependence of breast cancer through regulation of ERα enhancer function. *Oncogene,* 37(34), 4692-4710.

Gallagher W. M., Currid C. A., & Whelan, L. C. (2005). Fibulins and cancer: friend or foe? *Trends Mol. Med,* 11, 336–340.

Gaorav P. G. & Joan M. (2006). Cancer Metastasis: Building a Framework. *Cell*, 127, 679-695.

Gao D., Mittal V., Ban Y., Lourenco A. R., Yomtoubian S., & Lee S. (2018). Metastatic tumor cells – genotypes and phenotypes. *Front Biol*, 13(4), 277– 286.

General principles of neoadjuvant therapy for breast cancer - UpToDate [Internet]. [cited 2018 Oct 23]. Available from: https://www.uptodate.com/contents/general-principles-of-neoadjuvant- therapy-for-breast-cancer.

Gerashchenko T. S., Zolotaryova S. Y., Kiselev A. M., Tashireva L. A., Novikov N. M., Krakhmal N. V., Cherdyntseva N. V., Zavyalova M. V., Perelmuter V. M., & Denisov E. V. (2020). The Activity of KIF14, Mieap, and EZR in a New Type of the Invasive Component, Torpedo-Like Structures, Predetermines the Metastatic Potential of Breast Cancer. Cancers, 12(7), 1909.

Gerber B., Seitz E., & Müller H. (2003). Perioperative screening for metastatic disease is not indicated in patients with primary breast cancer and no clinical signs of tumor spread. *Breast Cancer Res Treat*, 82**,** 29–37.

Ghayad S. E., Vendrell J. A., Ben Larbi S., Dumontet C., Bieche I., & Cohen P. A. (2010). Endocrineresistanceassociated with activated ErbB system in breast cancer cells is reversed by inhibiting MAPK or PI3K/Akt signaling pathways. *Int J Cancer*, 126(2), 545-62.

Ghoncheh M., Pournamdar Z., & Salehiniya H. (2016). Incidence and Mortality and Epidemiology of Breast Cancer in the World. *Asian Pac J Cancer Prev*, 17, *Cancer Control in Western Asia Special Issue 2016*, 43-46.

GLOBOCAN 2020: Malaysia - Global Cancer Observatory [Internet]. [cited 2021 July 26]. Available from: https://gco.iarc.fr/today/data/factsheets/populations/458-malaysia-fact- sheets.pdf.

GLOBOCAN 2012: Estimated cancer incidence, mortality and prevalence world in 2012 [Internet]. [cited 2018 Oct 26]. Available from: https://www.http://gco.iarc.fr/.

Gogia A., Deo S. V. S., Sharma D., et. al. (2019). Clinicopathologic Characteristics and Treatment Outcomes of Patients with Up-Front Metastatic Breast Cancer: Single-Center Experience in India. *J Glob Oncol*, 5, 1-9.

Goicoechea S. M., Bednarski B., García-Mata R., Prentice-Dunn H., Kim H. J., & Otey C. A. (2009). Palladin contributes to invasive motility in human breast cancer cells. *Oncogene*. 28(4), 587-98.

Golbraikh A. & Tropsha A. (2002). Beware of q^2! *Journal of Molecular Graphics and Modelling*, 20(4), 269-276.

Gong Y., Booser D. J., & Sneige N. (2005). Comparison of HER-2 status determined by fluorescence in situ hybridization in primary and metastatic breast carcinoma. *Cancer*, 103, 1763-9.

Gorgoulis V., Vassiliou L.V., Karakaidos P. et. al. (2005). Activation of the DNA damage checkpoint and genomic instability in human precancerous lesions. *Nature,* 434, 907–913.

Grace-Martin K. [Internet]. [cited 2019 Dec 09]. The analysis factor: Assessing the Fit of Regression Models. https://www.theanalysisfactor.com/assessing-the- fit-of-regression-models/

Gradek F., Lopez-Charcas O., Chadet S., Poisson L., Ouldamer L., Goupille C., Jourdan M. L., Chevalier S., Moussata D., Besson P., & Roger S. (2019). Sodium Channel Nav1.5 Controls Epithelial-to-Mesenchymal Transition and Invasiveness in Breast Cancer Cells Through its Regulation by the Salt-Inducible Kinase-1. *Sci Rep*, 9(1), 18652.

Groot J. S., Ratze M. A., van Amersfoort M., Eisemann T., Vlug E. J., Niklaas M. T., Chin S. F., Caldas C., van Diest P. J., Jonkers J., de Rooij J., & Derksen P. W. (2018). αE-catenin is a candidate tumor suppressor for the development of E-cadherin-expressing lobular-type breast cancer. *The Journal of pathology*, 245(4), 456–467.

Grottke A., Ewald F., Lange T., Nörz D., Herzberger C., Bach J., Grabinski N., Gräser L., Höppner F., Nashan B., Schumacher U., Jücker M. (2016) Downregulation of AKT3 Increases Migration and Metastasis in Triple Negative Breast Cancer Cells by Upregulating S100A4. *PLoS One*. 2016 Jan 7;11(1):e0146370.

Haferlach T., Kohlmann A., Wieczorek L., Basso G., Kronnie G. T., Bene M. C., De Vos, J., Hernández J. M., Hofmann W. K., Mills K. I., Gilkes A., Chiaretti S., Shurtleff S. A., Kipps T. J., Rassenti L. Z., Yeoh A. E., Papenhausen P. R., Liu W. M., Williams P. M., & Foa, R. (2010). Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: report from the International Microarray Innovations in Leukemia Study Group.

Journal of clinical oncology: *official journal of the American Society of Clinical Oncology*, 28(15), 2529–2537.

Han H. S., Diéras V., Robson M., Palácová M., Marcom P. K., Jager A., Bondarenko I., Citrin D., Campone M., Telli M. L., Domchek S. M., Friedlander M., Kaufman B., Garber J. E., Shparyk Y., Chmielowska E., Jakobsen E. H., Kaklamani V., Gradishar W., Ratajczak C. K., Nickner C., Qin Q., Qian J., Shepherd S. P., Isakoff S. J., & Puhalla S. (2018). Veliparib with temozolomide or carboplatin/paclitaxel versus placebo with carboplatin/paclitaxel in patients with BRCA1/2 locally recurrent/metastatic breast cancer: randomized phase II study. *Ann Oncol*, 29(1), 154-161.

Han W., Hu C., Fan ZJ. et al. (2021). Transcript levels of keratin 1/5/6/14/15/16/17 as potential prognostic indicators in melanoma patients. *Sci Rep* 11, 1023.

Harikrishnan K., Joshi O., Madangirika S., & Balasubramanian N. (2020) Cell Derived Matrix Fibulin-1 Associates with Epidermal Growth Factor Receptor to Inhibit Its Activation, Localization and Function in Lung Cancer Calu-1 Cells. *Frontiers in Cell and Developmental Biology, 8*, 522.

Heinonen H., Lepikhova T., Sahu B., Pehkonen H., Pihlajamaa P., Louhimo R., Gao P., Wei G. H., Hautaniemi S., Jänne O. A, & Monni O. (2015) Identification of several potential chromatin binding sites of HOXB7 and its downstream target genes in breast cancer. *Int J Cancer*, 137(10), 2374-83.

Henseler J., Ringle C. M., & Sinkovics R. R. (2009). The use of partial least squares path modeling in international marketing. *Advances in International Marketing*, 20, 277-319.

Herland M., Khoshgoftaar T. M. & Wald R. (2014). A review of data mining using big data in health informatics. *Journal of Big Data*, 1, 2.

Hicks D. G., Short S. M., Prescott N. L., Tarr S. M., Coleman K. A., Yoder B. J., Crowe J. P., Choueiri T. K., Dawson A. E., Budd G. T., Tubbs R. R., Casey G., & Weil R. J. (2006). Breast cancers with brain metastases are more likely to be estrogen receptor negative, express the basal cytokeratin CK5/6, and overexpress HER2 or EGFR. *Am J Surg Pathol*, 30(9), 1097-104.

Hoadley K. A., Yau C., Wolf D. M., Cherniack A. D., Tamborero D., Ng S., Leiserson M., Niu B., McLellan M. D., Uzunangelov V., Zhang J., Kandoth C., Akbani R., Shen H., Omberg L., Chu A., Margolin A. A., Van't Veer L. J., Lopez-Bigas N., Laird P. W., & Stuart, J. M. (2014). Multiplatform analysis of

12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4), 929–944.

Hortobagyi G. N., Stemmer S. M., Burris H. A., Yap Y. S., Sonke G. S., Paluch- Shimon S., Campone M., Petrakova K., Blackwell K. L., Winer E. P., Janni W., Verma S., Conte P., Arteaga C. L., Cameron D. A., Mondal S., Su F., Miller M., Elmeliegy M., Germa C., & O'Shaughnessy J. (2018). Updated results from MONALEESA-2, a phase III trial of first-line ribociclib plus letrozole versus placebo plus letrozole in hormone receptor-positive, HER2- negative advanced breast cancer. *Ann Oncol*, 29(7), 1541-1547.

How to Calculate Feature Importance With Python [Internet]. [cited 2020 Jul 8]. Available from: https://machinelearningmastery.com/calculate-feature-importance-with-python/

Hu S., Zheng W., Jin L. (2021). Astragaloside IV inhibits cell proliferation and metastasis of breast cancer via promoting the long noncoding RNA TRHDE-AS1. J Nat Med. 2021 Jan;75(1):156-166.

Huan C., Xiaoxu C., & Xifang R. (2019). Zinc Finger Protein 521, Negatively Regulated by MicroRNA-204-5p, Promotes Proliferation, Motility and Invasion of Gastric Cancer Cells. *Technol Cancer Res Treat*, 18, 1533033819874783.

Huang T., Sun L., Yuan X., & Qiu H. (2017). Thrombospondin-1 is a multifaceted player in tumor progression. *Oncotarget*, 8(48), 84546-84558.

Huang Y., Li G., Wang K., et. al. (2018). Collagen Type VI Alpha 3 Chain Promotes Epithelial-Mesenchymal Transition in Bladder Cancer Cells via Transforming Growth Factor β (TGF-β)/Smad Pathway. *Med Sci Monit*, 24, 5346-5354.

Huang Z., Qin Q., Xia L., Lian B., Tan Q., Yu Y., Mo Q. (2020). Significance of Oncotype DX 21-Gene Test and Expression of Long Non-Coding RNA MALAT1 in Early and Estrogen Receptor-Positive Breast Cancer Patients. *Cancer Manag Res*, 13, 587-593.

Hudis C. A. (2007). Trastuzumab—mechanism of action and use in clinical practice. *The New England Journal of Medicine*, 357(1), 39–51.

Hyung K., Jin K. Y., Han E. H., Hwang Y., Choi J., Park B., & Jeong H. G. (2011). Metallothionein-2A overexpression increases the expression of matrix metalloproteinase-9 and invasion of breast cancer cells. *FEBS letters*, 585, 421-8.

Ibrahim F. A., Elfeky S. E., Haroun M., Ahmed M. A., Elnaggar M., Ismail N. A., & Abd El Moneim N. A. (2020). Association of matrix metalloproteinases 3 and 9 single nucleotide polymorphisms with breast cancer risk: A case-control study. *Molecular and Clinical Oncology*, 13, 54-62.

Ignatiadis M., Xenidis N., Perraki M., Apostolaki S., Politaki E., Kafousi M., et al. (2007). Different prognostic value of cytokeratin-19 mRNA positive circulating tumor cells according to estrogen receptor and HER2 status in early-stage breast cancer. *J Clin Oncol*, 25, 5194–5202.

Im S. A., Xu B., Li W., Robson M., Ouyang Q., Yeh D. C., Iwata H., Park Y. H., Sohn J. H., Tseng L. M., Goessl C., Wu W., & Masuda N. (2020). Olaparib monotherapy for Asian patients with a germline BRCA mutation and HER2-negative metastatic breast cancer: OlympiAD randomized trial subgroup analysis. *Sci Rep*, 10, 8753.

Jacobi C. E., de Bock G. H., Siegerink B., & van Asperen C. J. (2009). Differences and similarities in breast cancer risk assessment models in clinical practice: Which model to choose? *Breast Cancer Res Treat* 115:381-390.

Jin M. L., Kim Y. W., Jin H. L., Kang H., Lee E. K., Stallcup M. R., & Jeong K. W. (2018). Aberrant expression of SETD1A promotes survival and migration of estrogen receptor α-positive breast cancer cells. *Int J Cancer,* 143(11), 2871-2883.

Jiramongkol Y. & Lam E. W. (2020). FOXO transcription factor family in cancer and metastasis. *Cancer Metastasis Rev*, 39(3), 681-709.

Johnson J. M., Dalton R. R., Wester S. M., Landercasper J., & Lambert P. J. (1999). Histological Correlation of Microcalcifications in Breast Biopsy Specimens. *Arch Surg American Medical Association*; 134(7), 712.

Johnston S. R., Kilburn L. S., Ellis P., Dodwell D., Cameron D., Hayward L., Im Y. H., Braybrooke J. P., Brunt A. M., Cheung K. L., Jyothirmayi R., Robinson A., Wardley A. M., Wheatley D., Howell A., Coombes G., Sergenson N., Sin H. J., Folkerd E., Dowsett M., Bliss J. M. & SoFEA investigators. (2013). Fulvestrant plus anastrozole or placebo versus exemestane alone after progression on non-steroidal aromatase inhibitors in postmenopausal patients with hormone-receptor-positive locally advanced or metastatic breast cancer (SoFEA): a composite, multicentre, phase 3 randomised trial. *Lancet Oncol*, 14(10), 989-98.

Jolliffe I.T., & Cadima J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374.

Joosse S. A., Hannemann J., Spötter J., Bauche A., Andreas A., Müller V., & Pantel K. (2012). Changes in Keratin Expression during Metastatic Progression of Breast Cancer: Impact on the Detection of Circulating Tumor Cells. Clinical Cancer Research, 18(4), 993-1003.

Jordan V., Khan M., & Prill D. (2019). Breast Cancer Screening Why Can't Everyone Agree? *Primary Care: Clinics in Office Practice*, 46(1), 97-115.

Jung S. N., Lim H. S., Liu L., et. al. (2018). LAMB3 mediates metastatic tumor behavior in papillary thyroid cancer by regulating c-MET/Akt signals. *Sci Rep,* 8, 2718.

Keogh E, Mueen A. (2017) Curse of Dimensionality. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA.

Kim H., Cho J., Kwon S. Y., & Kang S. H. (2016). Biologic subtype is a more important prognostic factor than nodal involvement in patients with stages I and II breast carcinoma. *Annals of surgical treatment and research*, 90(1), 1–9.

Koh J, Kim MJ. (2019) Introduction of a New Staging System of Breast Cancer for Radiologists: An Emphasis on the Prognostic Stage. Korean J Radiol. 2019 Jan;20(1):69-82. doi: 10.3348/kjr.2018.0231.

Kong Y. C., Bhoo-Pathy N., Subramaniam S., Bhoo-Pathy N., Taib N. A., Jamaris S., Kaur K., See M. H., Ho G. F., & Yip C. H. (2017). Advanced Stage at Presentation Remains a Major Factor Contributing to Breast Cancer Survival Disparity between Public and Private Hospitals in a Middle-Income Country. *Int. J. Environ. Res. Public Health*, 14, 427.

Koo M. M., von Wagner C., Abel G. A., McPhail S., Rubin G. P., & Lyratzopoulos G. (2017). Typical and atypical presenting symptoms of breast cancer and their associations with diagnostic intervals: Evidence from a national audit of cancer diagnosis. *Cancer Epidemiology,* 48, 140–146.

Kornblum N., Zhao F., Manola J., Klein P., Ramaswamy B., Brufsky A., Stella P. J., Burnette B., Telli M., Makower D. F., Cheema P., Truica C. I., Wolff A. C., Soori G. S., Haley B., Wassenaar T. R., Goldstein L. J., Miller K. D., & Sparano J. A. (2018). Randomized Phase II Trial of Fulvestrant Plus Everolimus or Placebo in Postmenopausal Women With Hormone Receptor-Positive, Human Epidermal Growth Factor Receptor 2 -Negative Metastatic Breast

Cancer Resistant to Aromatase Inhibitor Therapy: Results of PrE0102. *J Clin Oncol*, 36(16), 1556-1563.

Kotepui M., Punsawad C., Chupeerach C., Songsri A., Charoenkijkajorn L., & Petmitr S. (2016). Differential expression of matrix metalloproteinase-13 in association with invasion of breast cancer. *Contemp Oncol*, 20(3), 225-8.

Krop I. E., Mayer I. A., Ganju V., Dickler M., Johnston S., Morales S., Yardley D. A., Melichar B., Forero-Torres A., Lee S. C., de Boer R., Petrakova K., Vallentin S., Perez E. A., Piccart M., Ellis M., Winer E., Gendreau S., Derynck M., Lackner M., Levy G., Qiu J., He .J, & Schmid P. (2016). Pictilisib for oestrogen receptor-positive, aromatase inhibitor-resistant, advanced or metastatic breast cancer (FERGI): a randomised, double-blind, placebo-controlled, phase 2 trial. *Lancet Oncol*, 17(6), 811-821.

Kumar S. (2020). Cross-Validation: Estimator Evaluator [Internet]. [cited 2020 April 20]. https://medium.com/swlh/cross-validation-estimator-evaluator-897d28afb4ff.

Kuo F.Y. & Sloan I.H., (2005). Lifting the curse of dimensionality. *Notices of the AMS*, 52(11), 1320-1328.

Lambert, A. W., Pattabiraman, D. R., & Weinberg, R. A. (2017). Emerging Biological Principles of Metastasis. Cell, 168(4), 670–691.

Lan L., Xu B., Chen Q., Jiang J., & Shen Y. (2019). Weighted correlation network analysis of triple-negative breast cancer progression: Identifying specific modules and hub genes based on the GEO and TCGA database. *Oncology Letters*, 18, 1207-1217.

Langley R. R. & Fidler I. J. (2007). Tumor cell-organ microenvironment interactions in the pathogenesis of cancer metastasis. *Endocr Rev*, 28,297–321.

Le S., Josse J. & Husson F. (2008). FactoMineR: A Package for Multivariate Analysis. *Journal of Statistical Software*, 25(1), 1–18.

Lee H., Zheng J., Gaddy D., et. al. (2015). A gradient-loadable (64) Cu-chelator for quantifying tumor deposition kinetics of nanoliposomal therapeutics by positron emission tomography. *Nanomedicine*, 11(1), 155-165.

Lee H, Shields AF, Siegel BA, Miller KD, Krop I, Ma CX, LoRusso PM, Munster PN, Campbell K, Gaddy DF, Leonard SC, Geretti E, Blocker SJ, Kirpotin DB, Moyo V, Wickham TJ, Hendriks BS. (2017). 64Cu-MM-302 Positron Emission Tomography Quantifies Variability of Enhanced Permeability and Retention of

Nanoparticles in Relation to Treatment Response in Patients with Metastatic Breast Cancer. *Clin Cancer Res*. Aug 1;23(15):4190-4202.

Lefebvre C., Bachelot T., Filleron T., Pedrero M., Campone M., Soria J-C., et. al. (2016) Mutational Profile of Metastatic Breast Cancers: A Retrospective Analysis. *PLoS Med*, 13 (12): e1002201.

Lehman C. D. (2012). Clinical indications: what is the evidence? *Eur J Radiol*. 81 (suppl 1), S82–S84.

Lever J., Krzywinski M. & Altman N. Principal component analysis. Nat Methods 14, 641–642 (2017). https://doi.org/10.1038/nmeth.4346

Li G. Z., Deng J. F., Qi Y. Z., Liu R., & Liu Z. X. (2020). COLEC12 regulates apoptosis of osteosarcoma through Toll-like receptor 4 –activated inflammation. *J Clin Lab Anal*, 34, e23469.

Li H. & Zheng B. (2019). Overexpression of the Ubiquitin-Specific Peptidase 9 X-Linked (USP9X) Gene is Associated with Upregulation of Cyclin D1 (CCND1) and Downregulation of Cyclin-Dependent Inhibitor Kinase 1A (CDKN1A) in Breast Cancer Tissue and Cell Lines. *Med Sci Monit*, 25, 4207-4216.

Li H. N., Li X. R., Cai M. M., Wang G., & Yang, Z. F. (2020). Elevated expression of FREM1 in breast cancer indicates favorable prognosis and high-level immune infiltration status. *Cancer Med*, 9, 9554-9570.

Li K., Zhang R., Wei M., Zhao L., Wang Y., Feng X., Yang Y., Yang S., & Zhang L. (2019). TROAP Promotes Breast Cancer Proliferation and Metastasis. *BioMed Research International,* Volume 2019, Article ID 6140951.

Li M. & Ni P. (2018). dSimer: Integration of Disease Similarity Methods. R package version 1.8.0.

Li M., Wang X., Ma R., Shi D. B., Wang Y. W., Li X. M., He J. Y., Wang J., & Gao P. (2019). The Olfactory Receptor Family 2, Subfamily T, Member 6 (OR2T6) Is Involved in Breast Cancer Progression via Initiating Epithelial- Mesenchymal Transition and MAPK/ERK Pathway. *Frontiers in Oncology*, 9, 1210.

Li T, Huang S, Yan W, Zhang Y, Guo Q. (2022). PRUNE2 inhibits progression of colorectal cancer in vitro and in vivo. *Exp Ther Med*. 2022 Feb;23(2):169.

Liberman L., Dershaw D. D., Rosen P. P., Abramson A. F., Deutch B. M., & Hann L. E. (1994). Stereotaxic 14-gauge breast biopsy: how many core biopsy specimens are needed? *Radiology*, 192(3), 793–5.

Licata L., Lo Surdo P., Iannuccelli M., Palma A., Micarelli E., Perfetto L., Peluso D., Calderone A., Castagnoli L., & Cesareni G. (2019). SIGNOR 2.0, the SIGnaling Network Open Resource 2.0: update. *Nucleic Acids Research,* 48(D1), D504–D510.

Liu J., Shen J. X., Wu H. T., Li X. L., Wen X. F., Du C. W., & Zhang G. J. (2018). Collagen 1A1 (COL1A1) promotes metastasis of breast cancer and is a potential therapeutic target. *Discov Med*, 25(139), 211-223.

Liu Y., Li L., Liu X., et. al. (2020). Arginine methylation of SHANK2 by PRMT7 promotes human breast cancer metastasis through activating endosomal FAK signalling. *Elife*, 9, e57617.

Liu Y., He M., Zuo W.-J., Hao S., Wang Z.-H., & Shao Z.-M. (2021). Tumor Size Still Impacts Prognosis in Breast Cancer With Extensive Nodal Involvement. *Frontiers in Oncology*, 11.

Liu X. T., Liu T. T., Wu M. Y., Chen Q. X., Zhuang J. X., & Qin W. (2021). Identifying FBLN1 (Gene ID: 2192) as a Potential Melanoma Biomarker for Melanoma based on an Analysis of microRNA Expression Profiles in the GEO and TCGA Databases. *Genetic Testing and Molecular Biomarkers,* 68- 78.

Lord S. J., Marinovich M. L., Patterson J. A , et. al. (2012). Incidence of metastatic breast cancer in an Australian population-based cohort of women with non-metastatic breast cancer at diagnosis. *Med J Aust*, 196, 688-692.

Lower E. E., Glass E., Blau R., & Harman S. (2009). HER-2/neu expression in primary and metastatic breast cancer. Breast Cancer Res Treat, 113, 301-6.

Malaysian National Cancer Registry Report 2012-2016 [Internet]. [cited 2023 May 05]. Available from: https://www2.moh.gov.my/moh/resources/Penerbitan/Laporan/Umum/2012-2016%20(MNCRR)/MNCR_2012-2016_FINAL_(PUBLISHED_2019).pdf.

Mangasarian O. L., Street W. N., & Wolberg W. H. (1995). Breast Cancer Diagnosis and Prognosis Via Linear Programming. *Oper Res*, 43(4), 570–7.

Mani S. A., Guo W., Liao M. J., Eaton E. N., Ayyanan A., Zhou A. Y., Brooks M., Reinhard F., Zhang C. C., Shipitsin M., Campbell L. L., Polyak K., Brisken C., Yang J., & Weinberg R. A. (2008). The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell*, 133, 704–715.

Marr B. (2021). What Is The Difference Between Data Mining And Machine Learning? https://bernardmarr.com/what-is-the-difference-between-data-mining-and-machine-learning. Retrieved 5th July 2022

Martin M. D., & Matrisian L.M. (2007). The other side of MMPs: protective roles in tumor progression. *Cancer Metastasis Rev*, 26, 717-724.

Martin M., Holmes F. A., Ejlertsen B., Delaloge S., Moy B., Iwata H., von Minckwitz G., Chia S. K. L., Mansi J., Barrios C. H., Gnant M., Tomašević Z., Denduluri N., Šeparović R., Gokmen E., Bashford A., Borrego M. R., Kim S. B., Jakobsen E. H., Ciceniene A., Inoue K., Overkamp F., Heijns J. B., Armstrong A. C., Link J. S., Joy A. A., Bryce R., Wong A., Moran S., Yao B., Xu F., Auerbach A., Buyse M., & Chan A. (2017). Neratinib after trastuzumab-based adjuvant therapy in HER2-positive breast cancer (ExteNET): 5-year analysis of a randomised, double-blind, placebo- controlled, phase 3 trial, *The Lancet Oncology*, 18(12), 1688-1700.

Masuda N., Takahashi M., Nakagami K., Okumura Y., Nakayama T., Sato N., Kanatani K., Tajima K., & Kashiwaba M. (2017). First-line bevacizumab plus paclitaxel in Japanese patients with HER2-negative metastatic breast cancer: subgroup results from the randomized Phase III MERiDiAN trial. *Jpn J Clin Oncol*, 47(5), 385-392.

Mathew A., Rajagopal P. S., Villgran V., et. al. (2017). Distinct Pattern of Metastases in Patients with Invasive Lobular Carcinoma of the Breast. *Geburtshilfe Frauenheilkd*, 77(6), 660-666.

Mathur S. & Dinakarpandian, D. (2010). Automated ontological gene annotation for computing disease similarity. *Summit on translational bioinformatics*, 12–16.

Mathur S., & Dinakarpandian D. A New Metric to Measure Gene Product Similarity. (2007) IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2007), 2007, pp. 333-338.

Matsumoto A., Shimada Y., Nakano M., Oyanagi H., Tajima Y., Nakano M., Kameyama H., Hirose Y., Ichikawa H., Nagahashi M., Nogami H., Maruyama S., Takii Y., Ling Y., Okuda S., Wakai T. (2020). RNF43 mutation is associated with aggressive tumor biology along with BRAF V600E mutation in right-sided colorectal cancer. *Oncol Rep*. 2020 Jun;43(6):1853-1862.

Mattos-Arruda L. D., Sammut S., Ross E. M., Bashford-Rogers R., Greenstein E., Markus H., Morganella S., Teng Y., Maruvka Y., Pereira B., Rueda O. M., Chin S., Contente-Cuomo T., Mayor R., Arias A., Ali H. R., Cope W., Tiezzi D., Dariush A., Amarante T. D., Reshef D., Ciriaco N., Martinez-Saez E., Peg V., Cajal S. R., Cortes J., Vassiliou G., Getz G., Nik-Zainal S., Murtaza M., Friedman N., Markowetz F., Seoane J., & Caldas C. (2019). The Genomic and Immune Landscapes of Lethal Metastatic Breast Cancer, *Cell Reports,* Volume 27, Issue 9, Pages 2690-2708.e10.

Mavaddat N., Michailidou K., Dennis J., Lush M., Fachal L., Lee A., Tyrer J. P., Chen T-H., Wang Q., et al. (2019). Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am J Hum Genet*., 104:21–34.

Maximiano S, Magalhaes P, Guerreiro MP, Morgado M. (2016). Trastuzumab in the Treatment of Breast Cancer. BioDrugs. 2016 Apr;30(2):75-86.

McDonald E. S., Clark A. S., Tchou J., Zhang P. & Freedman G. M. (2016). Clinical Diagnosis and Management of Breast Cancer. *The Journal of Nuclear Medicine*, Vol. 57, No. 2 (Suppl. 1).

McVeigh T. P., Hughes L. M., Miller N., Sheehan M., Keane M., Sweeney K. J., & Kerin M. J. (2014). The impact of Oncotype DX testing on breast cancer management and chemotherapy prescribing patterns in a tertiary referral centre. *Eur J Cancer,* 50(16), 2763–2770.

Mehlen P. & Puisieux A. (2006). Metastasis: a question of life or death. *Nat. Rev. Cancer,* 6, 449–458.

Mendez O., Fernández Y., Peinado M. A., Moreno V., & Sierra A. (2005). Anti-apoptotic Proteins Induce Non-random Genetic Alterations that Result in Selecting Breast Cancer Metastatic Cells. *Clinical & Experimental Metastasis,* 22(4), 297–307.

Miettinen M. & Fetsch, J. F. (2000). Distribution of keratins in normal endothelial cells and a spectrum of vascular tumors: implications in tumor diagnosis. *Human pathology*, 31(9), 1062-1067.

Minn A. J, Gupta G. P., Siegel P. M., et. al. (2005). Genes that mediate breast cancer metastasis to lung. *Nature,* 436(7050), 518–524.

Minn A. J., Gupta G. P., Padua D., et. al. (2007). Lung metastasis genes couple breast tumor size and metastatic spread. *Proc. Natl Acad. Sci. USA*, 104(16), 6740–6745.

Mitri Z., Constantine T., & O’Reagan R. (2012.) The HER2 Receptor in Breast Cancer: Pathophysiology, Clinical Use, and New Advances in Therapy. *Chemother Res Pract*, 2012, 743193.

Moll F., Katsaros D., Lazennec G., Hellio N., Roger P., Giacalone P. L., et. al. (2002). Estrogen induction and overexpression of fibulin-1C mRNA in ovarian cancer cells. *Oncogene* 21, 1097–1107.

Moore K. M., Thomas G. J., Duffy S. W., Warwick J., Gabe R., Chou P., Ellis I. O., Green A. R., Haider S., Brouilette K., Saha A., Vallath S., Bowen R., Chelala C., Eccles D., Tapper W. J., Thompson A. M., Quinlan P., Jordan L., Gillett C., Brentnall A., Violette S., Weinreb P. H., Kendrew J., Barry S. T., Hart I. R., Jones J. L., & Marshall J. F. (2015). Therapeutic targeting of integrin αvβ6 in breast cancer. *J Natl Cancer Inst*, 106(8), 169.

Mueller S. T. (2019). Correspondence Analysis and Multiple Correspondence Analysis. https://pages.mtu.edu/~shanem/psy5220/daily/Day23/CAMCA.html. Retrieved on 28th July 2022

Murthy R. K., Loi S, Okines A., Paplomata E., Hamilton E., Hurvitz S. A., Lin N. U., Borges V., Abramson V., Anders C., Bedard P. L., Oliveira M., Jakobsen E., Bachelot T., Shachar S. S., Müller V., Braga S., Duhoux F. P., Greil R., Cameron D., Carey L. A., Curigliano G., Gelmon K., Hortobagyi G., Krop I., Loibl S., Pegram M., Slamon D., Palanca-Wessels M. C., Walker L., Feng W., & Winer EP. (2020). Tucatinib, Trastuzumab, and Capecitabine for HER2-Positive Metastatic Breast Cancer. *New England Journal of Medicine*, 382(7), 597-609.

Neoadjuvant chemotherapy | Breast Cancer Network Australia [Internet]. [cited 2018 Oct 23]. Available from: https://www.bcna.org.au/understanding- breast-cancer/treatment/neoadjuvant-chemotherapy/.

Neuhaus E. M., Zhang W., Gelis L., Deng Y., Noldus J., & Hatt H. (20 12). Activation of an olfactory receptor inhibits proliferation of prostate cancer cells. *J Biol Chem, 284(24), 16218-16225.*

Nicolo C., Perier C., Prague M., Bellera C., MacGrogan G., Saut O., & Benzekry S. (2020). Machine Learning and Mechanistic Modeling for Prediction of Metastatic Relapse in Early-Stage Breast Cancer. JCO Clinical Cancer Informatics (4), 259-274.

Olivo, S. A., Macedo, L. G., Gadotti, I. C., Fuentes, J., Stanton, T., & Magee, D. J. (2008). Scales to Assess the Quality of Randomized Controlled Trials: A Systematic Review. Physical Therapy, 88(2), 156-175.

Otaghfar H. A., Hosseini M., Tizmaghz A., Shabestanipour G., & Noori H. (2015). A review on metastatic breast cancer in Iran. Asian Pacific Journal of Tropical Biomedicine, 5(6), 429-433.

Ouderkirk-Pecone J. L., Goreczny G. J., Chase S. E., Tatum A. H., Turner C. E., & Krendel M. (2016). Myosin 1e promotes breast cancer malignancy by enhancing tumor cell proliferation and stimulating tumor cell de-differentiation. *Oncotarget*, 7(29), 46419-46432.

Pace L E., & Keating N. L. (2014). A systematic assessment of benefits and risks to guide breast cancer screening decisions. *JAMA*, 311, 1327–1335.

Page M. J., McKenzie J. E., Bossuyt P. M., Boutron I., Hoffmann T. C., Mulrow C. D., et. al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, n71.

Pallis A. G., Boukovinas I., Ardavanis A., Varthalitis I., Malamos N., Georgoulias V., & Mavroudis D. (2012). A multicenter randomized phase III trial of vinorelbine/gemcitabine doublet versus capecitabine monotherapy in anthracycline- and taxane-pretreated women with metastatic breast cancer. *Ann Oncol*, 23(5), 1164-1169.

Palmer J. R., Zirpoli G., Bertrand K. A., Battaglia T., Bernstein L., Ambrosone C. B., Bandera E. V., Troester M. A., Rosenberg L., Pfeiffer R. M., & Trinquart L. A Validated Risk Prediction Model for Breast Cancer in US Black Women. (2021). *J Clin Oncol.* Dec 1;39(34):3866-3877.

Parish A., Schwaederle M., Daniels G., Piccioni D., Fanta P., Schwab R., Shimabukuro K., Parker B. A., Helsten T., & Kurzrock R. (2015). Fibroblast growth factor family aberrations in cancers: clinical and molecular characteristics. *Cell Cycle,* 14(13), 2121-8.

Pasic I., Shlien A., Durbin A. D., et. al. (2010). Recurrent focal copy-number changes and loss of heterozygosity implicate two noncoding RNAs and one tumor suppressor gene at chromosome 3q13.31 in osteosarcoma. *Cancer Res*, 70, 160-71.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., & Duchesnay E. (2011) Scikit-learn: Machine Learning in Python, *JMLR*, 12, 2825-2830.

Peinado H., Zhang H., Matei I. R., Costa-Silva B., Hoshino A., Rodrigues G., Psaila B., Kaplan R. N., Bromberg J. F., Kang Y., Bissell M. J., Cox T. R., Giaccia A. J., Erler J. T., Hiratsuka S., Ghajar C. M., & Lyden D. (2017). Pre- metastatic niches: organ-specific homes for metastases. *Nat Rev Cancer*, 17(5), 302-317.

Peng N. & Min L. (2019). BOG: calculate disease similarity by BOG [Internet]. [cited 2020 September 20]. Available from: https://rdrr.io/bioc/dSimer/man/BOG.html

Pereira B., Chin S. F., Rueda O., et. al. (2016). The somatic mutation profiles of 2,433 breast cancers refinetheir genomic and transcriptomic landscapes. *Nat Commun*, 7, 11479.

Pignatelli M., Cardillo M. R., Hanby A., & Stamp G. W. (1992). Integrins and their accessory adhesion molecules in mammary carcinomas: loss of polarization in poorly differentiated tumors. *Hum Pathol*, 10, 1159-66.

Polamuri S. (2017). How the random forest algorithm works in machine learning [Internet]. [cited 2020 September 20]. Available from: https://dataaspirant.com/random-forest-algorithm-machine-learing/

Polednak A. P. (2003). Survival of lymph node-negative breast cancer patients in relation to number of lymph nodes examined. *Ann Surg*, 237: 163-167.

Polyak K. & Weinberg R. A. (2009). Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. *Nat Rev Cancer*, 9, 265-73.

Poojary M., Jishnu P. V., & Kabekkodu S. P. (2020). Prognostic Value of Melanoma-Associated Antigen-A (MAGE-A) Gene Expression in Various Human Cancers: A Systematic Review and Meta-analysis of 7 428 Patients and 44 Studies. *Mol Diagn Ther*, 24(5), 537-555.

Power M, Fell G, Wright M. (2013). Principles for high-quality, high-value testing. *BMJ Evidence-Based Medicine*, 18:5-10.

Pramoditha R. (2021). How do you apply PCA to Logistic Regression to remove Multicollinearity? https://towardsdatascience.com/how-do-you-apply-pca-to-logistic-regression-to-remove-multicollinearity-10b7f8e89f9b. Retrieved on 6th July 2022.

Qadir J, Riaz S. K., Taj K., Sattar N., Sahar N. E., Khan J. S., Kayani M. A., Haq F., Arshad Malik M. F. (2021). Increased YAP1 expression is significantly associated with breast cancer progression, metastasis and poor survival. *Future Oncol*. 2021 Jul;17(21):2725-2734.

Quante A. S., Whittemore A. S., Shriver T., Strauch K., & Terry M. B. (2012). Breast cancer risk assessment across the risk continuum: genetic and nongenetic risk factors contributing to differential model performance *Breast Cancer Res*. 14:R144.

Radiation therapy for breast cancer - Mayo Clinic [Internet]. [cited 2018 Oct 22]. Available from: https://www.mayoclinic.org/tests-procedures/radiation-therapy-for-breast-cancer/about/pac-20384940.

Ramasamy S. & Nirmala K. (2017). Disease prediction in data mining using association rule mining and keyword-based clustering algorithms. *International Journal of Computers and Applications*, 1–8.

Razavi P., Chang M. T., Xu G., Bandlamudi C., Ross D. S., Vasan N., Cai Y., Bielski C. M., Donoghue M., Jonsson P., Penson A., Shen R., Pareja F., Kundra R., Middha S., Cheng M. L., Zehir A., Kandoth C., Patel R., Huberman, K., Baselga J. (2018). The Genomic Landscape of Endocrine- Resistant Advanced Breast Cancers. *Cancer cell*, 34(3), 427–438.e6.

Rehman A., Kim Y., Kim H., Sim J., Ahn H., Chung M. S., Shin S. J., & Jang K (2018). FOXO3a expression is associated with lymph node metastasis and poor disease-free survival in triple-negative breast cancer. *J Clin Pathol*, 71(9), 806-813.

Riccardi F., Colantuoni G., Diana A., Mocerino A., Cartenì G., Lauria R., Febbraro A., Nuzzo F., Addeo R., Marano O., Incoronato P., De Placido S., Ciardiello F., & Orditura M. (2018). Exemestane and Everolimus combination treatment of hormone receptor positive, HER2 negative metastatic breast cancer: A retrospective study of 9 cancer centers in the Campania Region (Southern Italy) focused on activity, efficacy and safety. *Mol Clin Oncol*, 9(3), 255–263.

Richter K., Paakkola T., Mennerich D., Kubaichuk K., Konzack A., Ali-Kippari H., Kozlova N., Koivunen P., Haapasaari K. M., Jukkola-Vuorinen A., Teppo H. R., Dimova E. Y., Bloigu R., Szabo Z., Kerkelä R., & Kietzmann T. (2018). USP28 Deficiency Promotes Breast and Liver Carcinogenesis as well as Tumor Angiogenesis in a HIF-independent Manner. *Mol Cancer Res*, 16(6), 1000-1012.

Ridley A. J., Schwartz M. A., Burridge K., Firtel R. A., Ginsberg M. H., Borisy G., Parsons J. T., & Horwitz A.R. (2003). Cell migration: integrating signals from front to back. *Science, 302*, 1704-1709.

Rolli M., Fransvea E., Pilch J., Saven A., & Felding-Habermann, B. (2003). Activated integrin alphavbeta3 cooperates with metalloproteinase MMP-9 in regulating migration of metastatic breast cancer cells. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16), 9482–9487.

Rongxian J., Vincent T. K., Puay-Hoon T., Thameem D., Wei D., & Boon-Huat B. (2002). Metallothionein 2A expression is associated with cell proliferation in breast cancer, *Carcinogenesis*, 23(1), 81–86.

Rosen P. P., Groshen S., Kinne D. W. et. al. (1990). Factors influencing prognosis in node-negative breast carcinoma: analysis of 767 T1N0M0/T2N0M0 patients with long-term follow-up. *J Clin Oncol*, 11, 2090–2100.

Rudland P. S., Platt-Higgins A. M., Davies L. M., Rudland S. D. S., Wilson J. B., Aladwani A., Winstanley J. H. R., Barraclough D. L., Barraclough R., West C. R., & Jones N. J., (2010). Significance of the Fanconi Anemia FANCD2 Protein in Sporadic and Metastatic Human Breast Cancer. *The American Journal of Pathology*, 176(6), 2935-2947.

Ruiterkamp J., Ernst M. F., van de Poll-Franse L. V., Bosscha K., Tjen-Hyejnen V. C. G., & Voogd A. C. (2009). Surgical resection of the primary tumour is associated with improved survival in patients with distant metastatic breast cancer at diagnosis. *EJSO*, 35, 1146-1151.

Ruoslahti E. (1999). Fibronectin and its integrin receptors in cancer. *Adv. Cancer Res,* 76**,** 1-20.

Rutkovskiy A., Stensløkken K. O & Vaage IJ. Osteoblast Differentiation at a Glance. Med Sci Monit Basic Res. 2016 Sep 26;22:95-106. doi: 10.12659/msmbr.901142. PMID: 27667570; PMCID: PMC5040224.

Sadlonova A., Bowe D. B., Novak Z., Mukherjee S., Duncan V. E., Page G. P., Frost A. R. (2009). Identification of molecular distinctions between normal breast-associated fibroblasts and breast cancer-associated fibroblasts. *Cancer Microenviron*, 2(1):9-21.

Santa-Maria C. A. & Gradishar W. J. (2015). Changing Treatment Paradigms in Metastatic Breast Cancer Lessons Learned. *JAMA Oncology*, Volume 1, Number 4.

Sanz G., Leray I., Dewaele A., Sobilo J., Lerondel S., Bouet S., et al. (2014) Promotion of Cancer Cell Invasiveness and Metastasis Emergence Caused by Olfactory Receptor Stimulation. *PLoS One*, 9(1), e85110.

Saura C., Oliveira M., Feng Y. H., Dai M. S., Chen S. W., Hurvitz S. A., Kim S. B., Moy B., Delaloge S., Gradishar W., Masuda N., Palacova M., Trudeau M. E., Mattson J., Yap Y. S., Hou M. F., Laurentiis M. D., Yeh Y. M., Chang H. T., Yau T., Wildiers H., Haley B., Fagnani D., Lu Y. S., Crown J., Lin J., Takahashi M., Takano T., Yamaguchi M., Fujii T., Yao B., Bebchuk J., Keyvanjah K., Bryce R., Brufsky A., & the NALA Investigators. (2020). Neratinib Plus Capecitabine Versus Lapatinib Plus Capecitabine in HER2 - Positive Metastatic Breast Cancer Previously Treated With ≥ 2 HER2 - Directed Regimens: Phase III NALA Trial. *Journal of Clinical Oncology*, 38(27), 3138-3149.

Segni M. T., Tadesse D. M., Amdemichael R., & Demissie H. F. (2016). Breast Self-examination: Knowledge, Attitude, and Practice among Female Health Science Students at Adama Science and Technology University, Ethiopia. *Gynecology and Obstetrics*, 6, 4.

Senkus E., Kyriakides S., Ohno S., Penault-Llorca F., Poortmans P., Rutgers E., Zackrisson S., Cardoso F., & ESMO Guidelines Committee (2015). Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of oncology : official journal of the European Society for Medical Oncology*, 26 (5), v8–v30.

Seo S., Moon Y., Choi J., et. al. (2019). The GTP binding activity of transglutaminase 2 promotes bone metastasis of breast cancer cells by downregulating microRNA-205. *Am J Cancer Res*, 9(3), 597-607.

Sethuraman A., Brown M., Seagroves T. N., et. al. (2016). SMARCE1 regulates metastatic potential of breast cancer cells through the HIF1A/PTK2 pathway. *Breast Cancer Res*, 18, 81.

Shaffrey M. E., Mut M., Asher A. L., Burri S. H., Chahlavi A., Chang S. M., et al. (2004). Brain metastases. Curr Probl Surg, 41 (8), 665-741.

Shaikh R. Feature Selection Techniques in Machine Learning with Python [Internet]. [cited 2020 Jun 23]. Available from: https://towardsdatascience.com/feature-selection-techniques-in-machine- learning-with-python-f24e7da3f36e.

Shepherd J. A., Kerlikowske K., Ma L., Duewer F., Fan B., Wang J., Malkov S., Vittinghoff E., & Cummings S. R. (2011). Volume of mammographic density and risk of breast cancer. *Cancer epidemiology, biomarkers & prevention*: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 20(7), 1473– 1482.

Shi X., Chen Y., Chen A., Le X., Huang K., Chen J., Wen S., Zeng H., Chen C., Li J. (2017). LncRNA TUSC7 affects malignant tumor prognosis by regulating protein ubiquitination: a genome-wide analysis from 10 237 pan-cancer patients. *Translational Cancer Research*, 6(4).

Shukla A. A. & Hunter K. (2014). Understanding susceptibility to breast cancer metastasis: the genetic approach. *Breast Cancer Manag*, 3(2), 165–172.

Sibbering M. and Courtney C. A. (2019). Management of breast cancer: basic principles. *Surgery*, https://doi.org/10.1016/ j.mpsur.2019.01.004.

Siegel R. L., Miller K. D., Fuchs H. E., & Jemal A. Cancer statistics, 2022. *CA Cancer J Clin*. 2022.

Silva-Fernandes I. J., Picanço-Albuquerque C. G., Claudia dos Santos Luciano M., Wong D. V. T., Bezerra M. J. B., Bitencourt F. S., Lima M. V. A. (2021). Abstract PS16-30: Pms1 gene: A new risk-mutation description? Cancer research, 81(4_Supplement), PS16-30-PS16-30.

Sim S.H., Park I.H., Jung K.H. et al. (2019). Randomised Phase 2 study of lapatinib and vinorelbine vs vinorelbine in patients with HER2 + metastatic breast cancer after lapatinib and trastuzumab treatment (KCSG BR11 -16). *Br J Cancer*, 121, 985–990.

Slamon D. J., Jones B. L., Shak S., Fuchs H., Paton V., Bajamonde A., et. al. (2011). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med, 344(11), 783-792

Sledge G. W. (2016). Curing Metastatic Breast Cancer. *Journal of Oncology Practice*, Volume 12, Issue 1.

Song Y., Liu Y., Pan S., Xie S., Wang Z.-w., & Zhu X. (2020). Role of the COP1 protein in cancer development and therapy. *Seminars in Cancer Biology*, 67, 43-52.

Su Z., Yang Z., Xu Y., Chen Y., & Yu Q. (2015). Apoptosis, autophagy, necroptosis, and cancer metastasis. *Mol Cancer*, 14, 48.

Subramaniam D. (2019). A Simple Introduction to K-Nearest Neighbors Algorithm [Internet]. [cited 2023 May 8]. Available from: https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-algorithm-.

Sun K., Gonçalves J. P., Larminie C., et. al. (2014). Predicting disease associations via biological network analysis. *BMC Bioinformatics*, 15, 304.

Sutcliffe E. L., Bunting K. L., He Y. Q., Li J., Phetsouphanh C., Seddiki N., Zafar A., Hindmarsh E. J., Parish C. R., Kelleher A. D., et. al. (2011). Chromatin-associated protein kinase C-theta regulates an inducible gene expression program and microRNAs in human T lymphocytes. *Mol. Cell*, 41, 704–719.

Swain S. M., Baselga J., Kim S. B., Ro J., Semiglazov V., Campone M., Ciruelos E., Ferrero J. M., Schneeweiss A., Heeson S., Clark E., Ross G., Benyunes M. C., Cortés J.; CLEOPATRA Study Group. (2015). Pertuzumab, trastuzumab, and docetaxel in HER2-positive metastatic breast cancer. *N Engl J Med*, 372(8), 724-34.

Szklarczyk D., Morris J. H., Cook H., Kuhn M., Wyder S., Simonovic M., Santos A., Doncheva N. T., Roth A., Bork P., Jensen L. J., & von Mering C. (2017). The STRING database in 2017: quality-controlled protein-protein ssociation networks, made broadly accessible. *Nucleic Acids Res*, 45(D1), D362-D368.

Szumilas M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 19(3), 227–229.

Taucher S., Rudas M., Mader R. M., Gnant M., Dubsky P., Bachleitner T., Roka S., Fitzal F., Kandioler D., Sporn E., Friedl J., Mittlbock M., & Jakesz R. (2003). Do we need HER-2/neu testing for all patients with primary breast carcinoma? *Cancer*, 98, 2547-2553.

Thiesen H. J., Steinbeck F., Maruschke M., Koczan D., Ziems B., & Hakenberg O. W. (2017). Stratification of clear cell renal cell carcinoma (ccRCC) genomes by gene-directed copy number alteration (CNA) analysis. *PLoS One*, 12(5), e0176659.

Tice J. A., Cummings S. R., Smith-Bindman R., Ichikawa L., Barlow W. E., & Kerlikowske K. (2008). Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. *Annals of internal medicine*, 148(5), 337–347.

Tinholt M., Garred Ø., Borgen E., Beraki E., Schlichting E., Kristensen V., Sahlberg K. K., & Iversen N. (2018) Subtype-specific clinical and prognostic relevance of tumor-expressed F5 and regulatory F5 variants in breast cancer: the CoCaV study. *J Thromb Haemost*. Jul;16(7):1347-1356.

Toi M., Shao Z., Hurvitz S., Tseng L. M., Zhang Q., Shen K., Liu D., Feng J., Xu B., Wang X., Lee K. S., Ng T. Y., Ridolfi A., Noel-Baron F., Ringeisen F., & Jiang Z. (2017). Efficacy and safety of everolimus in combination with trastuzumab and paclitaxel in Asian patients with HER2+ advanced breast cancer in BOLERO-1. *Breast Cancer Res*, 19(1), 47.

Tseng Y. J., Huang C. E., Wen C. N., Lai P. Y., Wu M. H., Sun Y. C., Wang H. Y., & Lu J. J. (2019). Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. *International Journal of Medical Informatics*, 128, 79–86.

Turaka A., Freedman G. M., Li T., et. al. (2009). Young age is not associated with increased local recurrence for DCIS treated by breast-conserving surgery and radiation. *J Surg Oncol*, 100, 25–31.

Tyrer J., Duffy S. W., & Cuzick J. (2004). A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in medicine*, 23(7), 1111–1130.

Union for International Cancer Control (UICC). TNM Classification of Malignant Tumours [Internet]. [cited 2022 Nov 20th]. Available from: https://www.uicc.org/what-we-do/sharing-knowledge/tnm#49500.

U.S. Breast Cancer Statistics | Breastcancer.org [Internet]. [cited 2019 Feb 20th]. Available from: https://www.breastcancer.org/symptoms/understand_bc/statistics

van de Rijn M., Perou C. M., Tibshirani R., Haas P., Kallioniemi O., Kononen J., et al. (2002). Expression of cytokeratins 17 and 5 identifies a group of breast carcinomas with poor clinical outcome. *Am J Pathol*. 161, 1991–1996.

van den Hurk C. J., Eckel R., van de Poll-Franse L. V., Coebergh J. W., Nortier J. W., Hölzel D., Breed W. P., & Engel J. (2011). Unfavourable pattern of metastases in M0 breast cancer patients during 1978 -2008: a population- based analysis of the Munich cancer registry. *Breast Cancer Res. Treat*, 128, 795–805.

Van Essen D. C., Smith S. M., Barch D. M., Behrens T. E., Yacoub E., & Ugurbil K. (2013). The WU-Minn human connectome project: an overview. *NeuroImage,* 80 (0), 62–79.

Van Z. F., Krupitza G., & Mikulits., W. (2011). Initial steps of metastasis: cell invasion and endothelial transmigration. Mutat. Res, 728, 23 –34.

Veerasamy R., Rajak H., Jain A., Sivadasan S., Varghese C. P., & Agrawal R. K. (2011). Validation of QSAR models – strategies and importance. *International Journal of Drug Design and Discovery,* 2(3), 511-519.

Venning F. A., Wullkopf L., & Erler J. T. (2015). Targeting ECM Disrupts Cancer Progression. *Frontiers in Oncology*, 5, 224.

Verma S., Bakshi D., Sharma V., Sharma I., Shah R., Bhat A., Bhat G. R., Sharma B., Wakhloo A., Kaul S., Heer V., Bhat A., Abrol D., Verma V., Kumar R. (2020). Genetic variants of DNAH11 and LRFN2 genes and their association with ovarian and breast cancer. *Int J Gynaecol Obstet,* 148(1), 118-122.

Vianna F. S. L., Giacomazzi J., Oliveira Netto C. B., Nunes L. N., Caleffi M., Ashton-Prolla P., Camey S. A. (2019). Performance of the Gail and Tyrer-Cuzick breast cancer risk assessment models in women screened in a primary care setting with the FHS-7 questionnaire. *Genet Mol Biol*., 42(1 suppl 1):232-237.

Vincent-Salomon A., Jouve M., Genin P., et al. (2002). HER2 status in patients with breast carcinoma is not modified selectively by preoperative chemotherapy and is stable during the metastatic process. *Cancer*, 94, 2169-73.

Vogel C. L., & Nabholtz J. M. (1999). Monotherapy of Metastatic Breast Cancer: A Review of Newer Agents. *The Oncologist***,** 4, 17-33

Vogel C., O'Rourke M., Winer E., et. al. (1996). A clinical trial of intravenous (IV) Navelbine (NVB) (vinorelbine tartrate) for first line treatment of women 60 years of age with advanced breast cancer (ABC). *Proc Am Soc Clin Oncol,* 15, 101a.

Vuylsteke P., Huizing M., Petrakova K., Roylance R., Laing R., Chan S., Abell F., Gendreau S., Rooney I., Apt D., Zhou J., Singel S., & Fehrenbacher L. (2016). Pictilisib PI3Kinase inhibitor (a phosphatidylinositol 3 -kinase [PI3K] inhibitor) plus paclitaxel for the treatment of hormone receptor- positive, HER2-negative, locally recurrent, or metastatic breast cancer: interim analysis of the multicentre, placebo-controlled, phase II randomised PEGGY study. *Ann Oncol*, 27(11), 2059-2066.

Wacholder S., Hartge P., Prentice R., Garcia-Closas M. F., Spencer H., Diver W., Thun R., Cox M. J, Hankinson D. G., Kraft S. E., Rosner P., Berg B., Christine D. (2010). Performance of Common Genetic Variants in Breast- Cancer Risk Models. *New England Journal of Medicine,* 362(11), 986-993.

Wan L., Pantel K., & Kang Y. (2013). Tumor metastasis: moving new biological insights into the clinic. *Nat. Med*, 19, 1450–1464.

Wang B., Qi X., Liu J., Zhou R., Lin C., Shangguan J., et. al. (2019). MYH9 promotes growth and metastasis via activation of MAPK/AKT signaling in colorectal cancer. *Journal of Cancer,* 10(4), 874-884.

Wang F., Jiang L., Li J., Yu X., Li M., Wu G., Yu Z., Zhou K., Chu H., & Zhai H. (2015). Association between TCF7L2 polymorphisms and breast cancer susceptibility: a meta-analysis. *International journal of clinical and experimental medicine*, 8(6), 9355–9361.

Wang M., Li X., Zhang J., Yang Q., Chen W., Jin W., Huang Y. R., Yang R., & Gao W. Q. (2017). AHNAK2 is a Novel Prognostic Marker and Oncogenic Protein for Clear Cell Renal Cell Carcinoma. *Theranostics*, 7(5), 1100-1113.

Wang Y., Cho S. G., Wu X., Siwko S., & Liu M. (2014). G-Protein Coupled Receptor 124 (GPR124) in Endothelial Cells Regulates Vascular Endothelial Growth Factor (VEGF)-Induced Tumor Angiogenesis. *Current Molecular Medicine*, 14, 10.2174/1566524014666140414205943.

Waugh M. G. (2012). Phosphatidylinositol 4-kinases, phosphatidylinositol 4-phosphate and cancer. *Cancer Lett*. 2012 Dec 28;325(2):125-31.

Weber L., Maberg D., Becker C., et. al. (2018). Olfactory Receptors as Biomarkers in Human Breast Carcinoma Tissues. *Front Oncol*, 8, 33.

Wei B., Wang J., Bourne P., Yang Q., Hicks D., Bu H., & Tang P. (2008). Bone metastasis is strongly associated with estrogen receptor- positive/progesterone receptor-negative breast carcinomas. *Hum Pathol*, 39(12), 1809-15.

Weide R., Feiten S., Friesenhahn V., et. al. (2014). Metastatic breast cancer: prolongation of survival in routine care is restricted to hormone-receptor- and Her2-positive tumors. *Springerplus*, 3, 535.

Weigelt B., Peterse J. L., & van Veer L. J. (2005). Breast cancer metastasis: markers and models. *Nat. Rev. Cancer,* 5, 591–602.

Weigelt B., Peterse J. L., & van Veer L. J. (2005). Breast Cancer Metastasis: Markers and Models. *Nature reviews Cancer,* September 2005.

Wong M. M., Guo C., & Zhang J. (2018). Nuclear receptor corepressor complexes in cancer: mechanism, function and regulation. *Am J Clin Exp Urol*, 2(3), 169-187.

Wong P. P., Yeoh C. C., Ahmad A. S., Chelala C., Gillett C., Speirs V., et al. (2014). Identification of MAGEA antigens as causal players in the development of tamoxifen-resistant breast cancer. *Oncogene*, 33(37), 4579–88.

World Health Organization Fact Sheets on Breast Cancer (2021) [Internet]. [cited 2021 July 26]. Available from: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

Wrzesinski T., Szelag M., Cieślikowski W.A. et. al. (2015). Expression of pre- selected TMEMs with predicted ER localization as potential classifiers of ccRCC tumors. *BMC Cancer*, 15, 518.

Wu Y., Fan J., Peissig P., Berg R., Tafti A. P., Yin J., Yuan M., Page D., Cox J., & Burnside E. S. (2018). Quantifying predictive capability of electronic health records for the most harmful breast cancer. *Proceedings of SPIE--the International Society for Optical Engineering*, 10577, 105770J.

Xing F., Liu Y., Wu S. Y., Wu K., Sharma S., Mo Y. Y., Feng J., Sanders S., Jin G., Singh R., Vidi P. A., Tyagi A., Chan M. D., Ruiz J., Debinski W., Pasche B. C., Lo H. W., Metheny-Barlow L. J., D'Agostino R. B. Jr., & Watabe K. (2018). Loss of XIST in Breast Cancer Activates MSN-c-Met and Reprograms Microglia via Exosomal miRNA to Promote Brain Metastasis. *Cancer Res*. 78(15), 4316-4330.

Yang H., Wang B., Wang T., Xu L., He C., Wen H., et. al. (2014) Toll-Like Receptor 4 Prompts Human Breast Cancer Cells Invasiveness via Lipopolysaccharide Stimulation and Is Overexpressed in Patients with Lymph Node Metastasis. *PLoS One*, 9(10), e109980.

Yang H, Gao L, Zhang M, Ning N, Wang Y, Wu D, Li X. Identification and Analysis of An Epigenetically Regulated Five-lncRNA Signature Associated With Outcome and Chemotherapy Response in Ovarian Cancer. *Front Cell Dev Biol*. 2021 Feb 23;9:644940.

Yang N., Huang L., Liu J., Guo J., Zhao L., Chai H., Qi C. (2022). NPAP1 mutation as an indicator stratified patients benefit from immune checkpoint inhibitors in NSCLC. *Journal of Clinical Oncology*, 40(16_suppl), e14574-e14574.

Yao B., Wang J., Qu S., et. al. (2019). Upregulated osterix promotes invasion and bone metastasis and predicts for a poor prognosis in breast cancer. *Cell Death Dis,* 10(1), 28.

Yao J., Yao X., Tian T., Fu X., Wang W., Li S., Shi T., Suo A., Ruan Z., Guo H., Nan K., Huo X. (2017). ABCB5-ZEB1 Axis Promotes Invasion and Metastasis in Breast Cancer Cells. *Oncol Res*. 2017 Mar 13;25(3):305-316.

Yip C. H., Mohd Taib N. A., & Mohamed I. (2006). Epidemiology of Breast Cancer in Malaysia. *Asian Pacific Journal of Cancer Prevention*, 7, 369-374.

Yip C. H., Pathy N. B., & Teo S. H. (2014). A Review of Breast Cancer Research in Malaysia. *Med J Malaysia*, Vol 69, Supplement A.

Yoo I., Alafaireet P., Marinov M. et al. (2012) Data Mining in Healthcare and Biomedicine: A Survey of the Literature. *J Med Syst,* 36, 2431–2448.

Zhang B. N., Cao X. C., Chen J.Y., Chen J., Fu L., Hu X. C., Jiang Z. F., Li H. Y. Liao N., Liu D. G., Tao O., Shao Z. M., Sun Q., Wang S., Wang Y. W., Xu B. H., & Zhang J. (2012). Guidelines on the diagnosis and treatment of breast cancer (2011 edition). *Gland Surgery*, 1(1), 39–61.

Zhang H., Pan Y. Z., Cheung M., et al. (2019). LAMB3 mediates apoptotic, proliferative, invasive, and metastatic behaviors in pancreatic cancer by regulating the PI3K/Akt signaling pathway. *Cell Death*, 10, 230.

Zhang S., Lu Y., Qi L., Wang H., Wang Z., & Cai Z. (2020). AHNAK2 Is Associated with Poor Prognosis and Cell Migration in Lung Adenocarcinoma. *Biomed Research International*, 2020, 8571932.

Zhao R., Fu L., Yuan Z., Liu Y., Zhang K., Chen Y., Wang L., Sun D., Chen L., Liu B., & Zhang L. (2021). Discovery of a novel small-molecule inhibitor of Fam20C that induces apoptosis and inhibits migration in triple negative breast cancer. *Eur J Med Chem*, 210, 113088.

Zhou X., Liu K. Y, & Wong S. T. C. (2004). Cancer classification and prediction using logistic regression with Bayesian gene selection, *Journal of Biomedical Informatics*, 37(4), 249-259.

Zhu J., Deng J., Zhang L., et. al. (2020). Reconstruction of lncRNA-miRNA-mRNA network based on competitive endogenous RNA reveals functional lncRNAs in skin cutaneous melanoma. *BMC Cancer*, 20, 927.

Zlobin A., Wyatt D., Varsanik M., Dingwall A., & Osipo C. (2018). Roles for MLL2/ KMT2D or MLL3/ KMT2C in HER+ breast cancer stem cells. *Cancer Res*, 78 (13), 5845.

Zohrap N., Saatci Ö., Ozes B., Coban I., Atay H. M., Battaloglu E., Şahin Ö., & Bugra K. (2018). SIK2 attenuates proliferation and survival of breast cancer cells with simultaneous perturbation of MAPK and PI3K/Akt pathways. *Oncotarget*, 9(31), 21876-21892.

Zörnig M., Hueber A., Baum W., & Evan G. (2001). Apoptosis regulators and their role in tumorigenesis. *Biochim Biophys Acta*, 1551(2), F1-37.

Zubair H., & Ahmad A. (2017). Cancer Metastasis. Introduction to Cancer Metastasis, 3–12.doi:10.1016/b978-0-12-804003-4.00001-3

APPENDICES

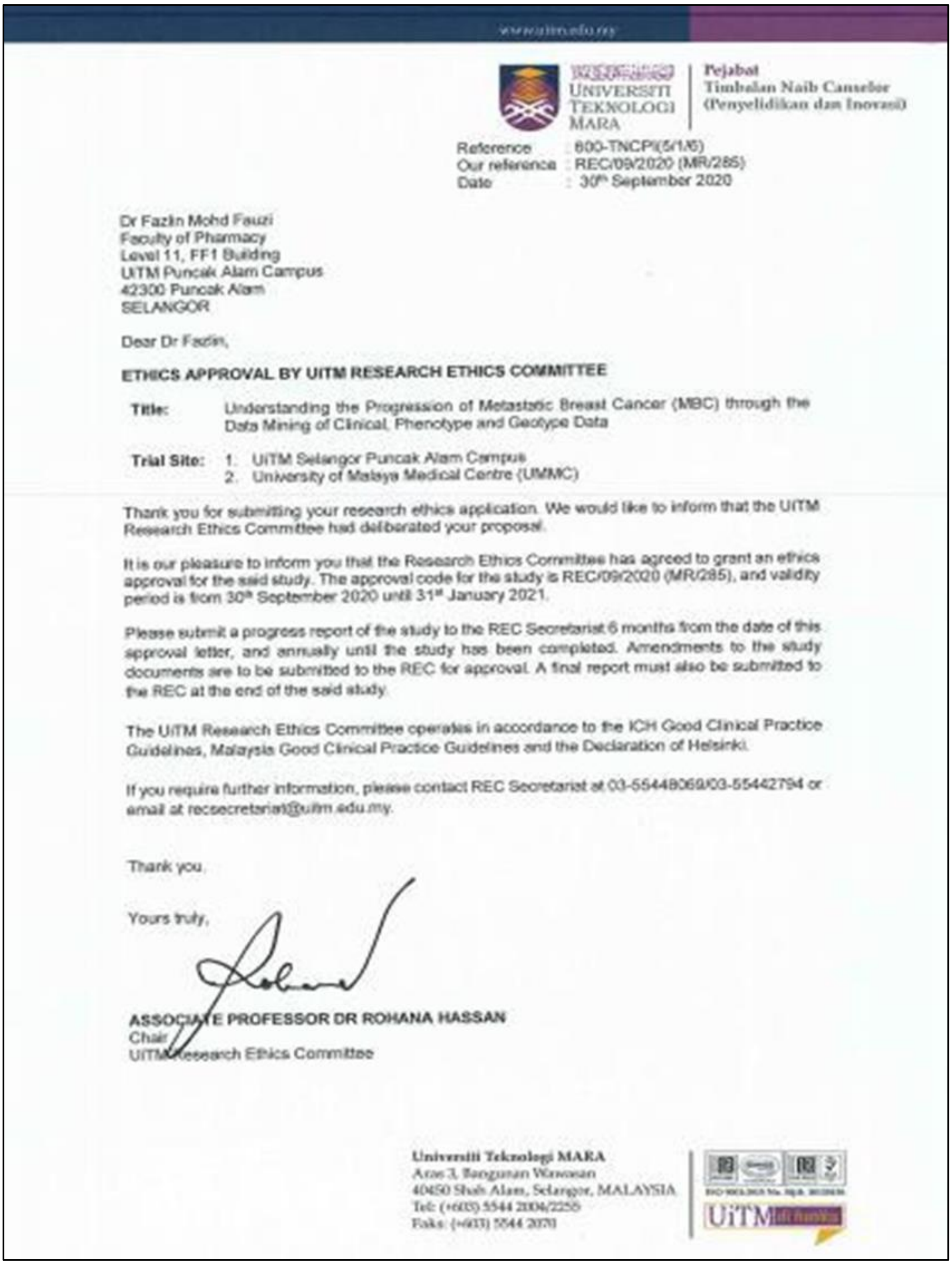
APPENDIX 1

Jadad Score Calculation

Item	**Score**
Was the study described as randomised (this includes words such as randomly, random, and randomisation)?	0/1
Was the method used to generate the sequence of randomisation described and appropriate (table of random numbers, computer-generated, etc)?	0/1
Was the study described as double blind?	0/1
Was the method of double blinding described and appropriate (identical placebo, active placebo, dummy, etc)?	0/1
Was there a description of withdrawals and dropouts?	0/1
Guidelines for Assessment	
Randomisation A method to generate the sequence of randomisation will be regarded as appropriate if it allowed each study participant to have the same chance of receiving each intervention and the investigators could not predict which treatment was next. Methods of allocation using date of birth, date of admission, hospital numbers, or alternation should not be regarded as appropriate.	
Double blinding A study must be regarded as double blind if the word "double blind" is used. The method will be regarded as appropriate if it is stated that neither the person doing the assessments nor the study participant could identify the intervention being assessed, or if in the absence of such a statement the use of active placebos, identical placebos, or dummies is mentioned.	
Withdrawals and dropouts Participants who were included in the study but did not complete the observation period or who were not included in the analysis must be described. The number and the reasons for withdrawal in each group must be stated. If there were no withdrawals, it should be stated in the article. If there is no statement on withdrawals, this item must be given no points.	

APPENDIX 2

Ethics Approval



www.uitm.edu.my

UNIVERSITI
TEKNOLOGI
MARA

Pejabat
Timbalan Naib Canselor
(Penyelidikan dan Inovasi)

Reference : 600-TNCPI(5/1/6)
Our reference : REC/09/2020 (MR/285)
Date : 30th September 2020

Dr Fazlin Mohd Fauzi
Faculty of Pharmacy
Level 11, FF1 Building
UiTM Puncak Alam Campus
42300 Puncak Alam
SELANGOR

Dear Dr Fazlin,

ETHICS APPROVAL BY UITM RESEARCH ETHICS COMMITTEE

Title: Understanding the Progression of Metastatic Breast Cancer (MBC) through the Data Mining of Clinical, Phenotype and Geotype Data

Trial Site:
1. UiTM Selangor Puncak Alam Campus
2. University of Malaya Medical Centre (UMMC)

Thank you for submitting your research ethics application. We would like to inform that the UiTM Research Ethics Committee had deliberated your proposal.

It is our pleasure to inform you that the Research Ethics Committee has agreed to grant an ethics approval for the said study. The approval code for the study is REC/09/2020 (MR/285), and validity period is from 30th September 2020 until 31st January 2021.

Please submit a progress report of the study to the REC Secretariat 6 months from the date of this approval letter, and annually until the study has been completed. Amendments to the study documents are to be submitted to the REC for approval. A final report must also be submitted to the REC at the end of the said study.

The UiTM Research Ethics Committee operates in accordance to the ICH Good Clinical Practice Guidelines, Malaysia Good Clinical Practice Guidelines and the Declaration of Helsinki.

If you require further information, please contact REC Secretariat at 03-55448069/03-55442794 or email at recsecretariat@uitm.edu.my.

Thank you.

Yours truly,

ASSOCIATE PROFESSOR DR ROHANA HASSAN
Chair
UiTM Research Ethics Committee

Universiti Teknologi MARA
Aras 3, Bangunan Wawasan
40450 Shah Alam, Selangor, MALAYSIA
Tel: (+603) 5544 2004/2258
Faks: (+603) 5544 2070

UiTM

AUTHOR'S PROFILE

Nadia Jalaludin obtained Bachelor of Science (Genetics) (Hons) from Universiti Kebangsaan Malaysia, Selangor in 2005, Master of Science (Molecular Biology) in 2008 from the University of Queensland, Australia and completed her PhD in Pharmacoinformatic in 2023 from the Faculty of Pharmacy, Universiti Teknologi MARA, Selangor. She is currently a lecturer in the Faculty of Pharmacy UiTM Selangor.

List of Publication:

Mohd Mutalip S. S., Kadir B. I., Singh G. K. S., Mohamad M., Mohamed R., Jalaludin N., Hussin S. N., Mohd Jofrry S., Abdul Hadi M. F. and Anuar A. Histological Changes in Male Reproductive System of Time Response BPA Treated SD Rats. Proceeding for 1st International Conference on Art, Social Science & Technology (iCAST2010), 24th – 26th February 2010 at Hotel Gurney, Penang, Malaysia.

Mohd Mutalip S. S., Abdul Hadi M. F., Singh G. K. S., Mohamad M., Mohamed R., Jalaludin N., Hussin S. N., Mohd Jofrry S., Kadir B. I. and Anuar A. Histological Changes in Male Reproductive System of Dose Response BPA Treated SD Rats. Proceeding for Regional Seminar on Science, Technology and Social Science (STSS2010), 1st – 2nd June 2010 at MS Garden Hotel, Kuantan, Pahang, Malaysia.

Singh G. K. S., Mohamad M., Mohd Mutalip S. S., Mohamed R., Jalaludin N., Hussin S. N., Mohd Jofrry S. and Baharuddin M. S. Effects of the Estrogenic Pollutant Bisphenol A on Estrogen Receptor alpha in Time Response Treated Female Sprague-Dawley Rats. Book of Abstract: Proceedings of the International Conference and Exhibition on Pharmaceutical, Nutraceutical and Cosmeceutical: Formulation and Applications (Pharmatech2010), 25th-26th May 2010 at Kuala Lumpur Convention Centre, Kuala Lumpur, Malaysia.

Anuar A., Mohd Mutalip S. S., Singh G. K. S., Mohamad M., Mohamed R., Jalaludin N., Hussin S. N., Mohd Jofrry S., Mohd Faiz Abdul Hadi and Bob Iskandar Kadir. Detection of Protein Changes in Dose-Response Bisphenol A Treated Female Sprague-Dawley Rats. Book of Abstract: Proceedings of the International Conference and Exhibition on Pharmaceutical, Nutraceutical and Cosmeceutical: Formulation and Applications (Pharmatech2010), 25th-26th May 2010 at Kuala Lumpur Convention Centre, Kuala Lumpur, Malaysia.

Hussin S. N., Fatimah A., Singh G. K. S., Mohamad M., Mohd Mutalip S. S., Mohd Jofrry S., Mohamed R. and Jalaludin N.. Hormonal Changes in Bisphenol A Treated Sprague-Dawley Rats At Different Dose Response. Book of Abstract: Proceedings of the International Conference and Exhibition on Pharmaceutical, Nutraceutical and Cosmeceutical: Formulation and Applications (Pharmatech2010), 25th-26th May 2010 at Kuala Lumpur Convention Centre, Kuala Lumpur, Malaysia.

Mohamad M., Singh G. K. S., Mohd Mutalip S. S., Mohamed R., Jalaludin N., Hussin S. N., Mohd Jofrry S. and Zulkifli S. Effects of Bisphenol A on Estrogen Receptor Genes of Male Juvenile Sprague-Dawley Rats. Book of Abstract: Proceedings of the International Conference and Exhibition on Pharmaceutical, Nutraceutical and Cosmeceutical: Formulation and Applications (Pharmatech2010), 25th-26th May 2010 at Kuala Lumpur Convention Centre, Kuala Lumpur, Malaysia.

Mohd Jofrry S., Mohd Marzuki N., Singh G. K. S., Mohamad M., Jalaludin N., Mohamed R., Hussin S. N. and Mohd Mutalip S. S. Bisphenol A Exposure and Induction of Oxidative Stress: A Time Response Study. Book of Abstract: Proceedings of the International Conference and Exhibition on Pharmaceutical, Nutraceutical and Cosmeceutical: Formulation and Applications

(Pharmatech2010), 25th-26th May 2010 at Kuala Lumpur Convention Centre, Kuala Lumpur, Malaysia.

Mohamad M., Singh G. K. S., Mohd Mutalip S. S., Mohamed R., Jalaludin N., Hussin S. N., Mohd Jofrry S. and Mahmood F. H. Expression of Estrogen Receptor Genes in Dose Response, Bisphenol A Treated Female Juvenile Rats. Gene Expression Changes in Male Reproductive System of Time Response Bisphenol A Treated Rats. Book of Abstract: Proceedings of the 22nd Scientific Meeting of the Malaysian Society of Pharmacology and Physiology (MSPP), 2-3 June, 2010, Shah Alam Convention Centre, Selangor, Malaysia.

Hussin S. N., Siti Fatimah M.F., Singh G. K. S., Mohamad M., Mohd Mutalip S. S., Mohd Jofrry S., Mohamed R. and Jalaludin N.. Antioxidant Activity Study in the Brain of Male Bishenol A Treated Rats At Different Dose Response. Book of Abstract: Proceedings of the 22nd Scientific Meeting of the Malaysian Society of Pharmacology and Physiology (MSPP), 2-3 June 2010, Shah Alam Convention Centre, Selangor, Malaysia.