DEVELOPMENT OF MALAYSIAN BREAST CANCER SURVIVAL PROGNOSTIC TOOL (myBeST) FOR PREDICTION OF SURVIVAL PROBABILITY AMONG WOMEN WITH BREAST CANCER IN MALAYSIA

DR. MOHD NASRULLAH BIN NIK AB. KADIR

UNIVERSITI SAINS MALAYSIA

2023

DEVELOPMENT OF MALAYSIAN BREAST CANCER SURVIVAL PROGNOSTIC TOOL (myBeST) FOR PREDICTION OF SURVIVAL PROBABILITY AMONG WOMEN WITH BREAST CANCER IN MALAYSIA

By

DR. MOHD NASRULLAH BIN NIK AB. KADIR

Dissertation submitted in partial fulfilment of the requirements for the degree of

Doctor of Public Health (Epidemiology)

MARCH 2023

DECLARATION

I declare that this dissertation is my own work and has not been submitted for the award of a higher degree elsewhere. However, the information which has been derived from other sources is indicated in this dissertation.

Mohd Nasrullah Bin Nik Ab. Kadir

Date: 2nd March 2023

ACKNOWLEDGEMENTS

Firstly, I want to offer this endeavour to Allah S.W.T., the almighty, for the wisdom and resources he bestowed upon me to complete this project. I want to extend my gratitude to the individuals that assist me throughout the journey. To my family, for making my life easier with abundant love, space, and time.

To the lecturers of the Department of Community Medicine, School of Medical Sciences, Universiti Sains Malaysia, thank you for providing essential public health research skills. To my supervisor, Dr Suhaily Mohd Hairon and co-supervisor, Assoc. Prof Dr Najib Majdi Yaacob, I admire their constant input, guidance, and knowledge in order to complete this demanding Doctor of Public Health research project as it finally comes together.

To my research team members who had been instrumental in every phase of the study, I am grateful for their constant contribution. They are:

1. Dato' Dr Imi Sairi Ab Hadi, Senior Consultant Breast and Endocrine Surgeon, Breast and Endocrine Surgery Unit, Hospital Raja Perempuan Zainab II, Kota Bharu
2. Dr Siti Norbayah Yusof, Public Health Medicine Specialist, Head of the Malaysian National Cancer Registry Department, National Cancer Institute, Putrajaya
3. Assoc. Prof Dr Kamarul Imran Musa, Medical Epidemiologist and Senior Medical Lecturer, Department of Community Medicine, School of Medical Sciences, Universiti Sains Malaysia

4. Dr Maya Mazuwin Yahya, Breast and Endocrine Surgeon, Department of Surgery, School of Medical Sciences, Universiti Sains Malaysia
5. Dr Farzaana Adam, Public Health Medicine Specialist, Public Health Division, Penang State Health Department
6. Dr Seoparjoo Azmel Mohd Isa, Medical Pathologist, Department of Pathology, School of Medical Sciences, Universiti Sains Malaysia
7. Dr Balqis Bahtiar, Medical Officer and Cancer Registrar, Malaysian National Cancer Registry Department, National Cancer Institute, Putrajaya
8. Dr Siti Maryam Muhamat, Medical Officer and Cancer Registrar, Malaysian National Cancer Registry Department, National Cancer Institute, Putrajaya
9. Mastulu Wahab, Cancer Registrar, Sarawak Cancer Registry, Non-Communicable (NCD) Unit, Sarawak State Health Department
10. Dr Nabihah Ali, Doctor of Public Health Candidate, Public Health Medicine Department, Faculty of Medicine and Health Sciences, Universiti Malaysia Sabah
11. Dr Ahmad Nabeil Alias, Doctor of Public Health Candidate, Public Health Medicine Department, Faculty of Medicine and Health Sciences, Universiti Malaysia Sabah
12. Dr Noor ‘Adilah Kamarudin, Medical Officer, Johor Bahru District Health Office, Johor State Health Department
13. Dr Muhammad Hafizuddin Mamat Azlan, Medical Officer, Hospital Sibu, Sarawak
14. Dr Ahmad Nadzmi Ibrahim, Medical Officer, Hospital Sultanah Bahiyah, Alor Setar
15. Dr Sim Hui Song, Medical Officer, Hospital Pulau Pinang

To the medical record staffs, officers and directors of participating hospitals, National Cancer Institute, Hospital Raja Perempuan Zainab II, Hospital Pulau Pinang, Hospital Sultan Ismail, Hospital Umum Sarawak, Hospital Queen Elizabeth II, and Hospital Sultanah Bahiyah, I could not thank them enough for facilitating the research data collection process.

To dear colleagues and others who had been involved indirectly in putting these pieces together, may Allah bless you all. Finally, to the Ministry of Health, Malaysia, which gave me the opportunity and sponsorship to upgrade my public health competency.

TABLE OF CONTENTS

LIST OF PAPERS

This Doctor of Public Health (DrPH) dissertation contains three papers:

1. Mohd Nasrullah Nik Ab Kadir, Najib Majdi Yaacob, Siti Norbayah Yusof, Imi Sairi Ab Hadi, Kamarul Imran Musa, Seoparjoo Azmel Mohd Isa, Balqis Bahtiar, Farzaana Adam, Maya Mazuwin Yahya and Suhaily Mohd Hairon. *Development of predictive models for survival among women with breast cancer in Malaysia.* (This paper was published in the International Journal of Environmental Research and Public Health, https://doi.org/10.3390/ijerph192215335)
2. Mohd Nasrullah Nik Ab. Kadir, Suhaily Mohd Hairon, Siti Norbayah Yusof, Siti Maryam Muhamat, Imi Sairi Ab Hadi and Najib Majdi Yaacob. *A comparison of online prognostic tool PREDICT breast cancer and myBeST among women with breast cancer in Malaysia.* (This paper was submitted to the Cancers journal)
3. Mohd Nasrullah Nik Ab. Kadir, Suhaily Mohd Hairon, Najib Majdi Yaacob, Siti Norbayah Yusof, Kamarul Imran Musa, Maya Mazuwin Yahya, Seoparjoo Azmel Mohd Isa, Muhammad Hafizuddin Mamat Azlan and Imi Sairi Ab Hadi. *myBeST – A web-based survival prognostic tool for women with breast cancer in Malaysia: Development process and preliminary validation study.* (This paper was published in the International Journal of Environmental Research and Public Health, https://doi.org/10.3390/ijerph20042985)

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

Adj. HR	Adjusted Hazard Ratio
ANN	Artificial neural network
ASR	Age-standardised incidence rate
AUC	Area under the receiver operating characteristic curve
CI	Confidence interval
Cox PH	Cox proportional hazard regression
CVI	Content validation index
DT	Decision tree
ER	Oestrogen receptor
FVI	Face validation index
HER2	Human epidermal growth factor 2
IQR	Interquartile range
KM	Kaplan-Maier
MNCR	Malaysian National Cancer Registry
MOH	Ministry of Health
myBeST	Malaysian Breast Cancer Survival Prognostic Tool

LIST OF ABBREVIATIONS

NCI	National Cancer Institute, Ministry of Health Malaysia
NPI	Nottingham Prognostic Index
OS	Overall survival
PR	Progesterone receptor
PREDICT	PREDICT breast cancer tool
ROC	Receiver characteristic curve
SD	Standard deviation
SDLC	Software Development Life Cycle
SE	Standard error
TNM	Tumour Node Metastases
TRIPOD	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
URL	Uniform Resource Locator

LIST OF SYMBOLS

>	More than
<	Less than
=	Equal to
≥	More than and equal to
≤	Less than and equal to
%	Percentage
+	Plus

ABSTRAK

PEMBINAAN INSTRUMEN PROGNOSTIK SURVIVAL KANSER PAYUDARA (myBeST) UNTUK MERAMAL KEBARANGKALIAN SURVIVAL DALAM KALANGAN WANITA DENGAN KANSER PAYUDARA DI MALAYSIA

Latar belakang: Kanser payudara adalah penyumbang utama diagnosis kes baharu kanser. Instrumen prognostik telah dibangunkan untuk memaklumkan pesakit mengenai survival penyakit tersebut. Prestasi instrumen dari negara Barat didapati kurang tepat apabila diaplikasikan dalam kalangan pesakit tempatan dengan instrumen PREDICT breast cancer (PREDICT) mempunyai ketepatan ramalan yang memuaskan.

Objektif: Kajian ini bertujuan membangunkan model ramalan survival dalam kalangan wanita Malaysia yang mengidap kanser payudara, membandingkan prestasinya dengan PREDICT, dan algoritma model tersebut dipilih untuk membina laman web Instrumen Prognostik Survival Kanser Payudara (myBeST).

Kaedah: Kajian ini terdiri daripada dua fasa. Fasa 1 ialah kajian kohort retrospektif menggunakan data yang diambil di tujuh pusat rujukan kanser payudara di Malaysia. Kami mengumpul 13 pemboleh ubah peramal dan status survival. Analisis regresi bahaya berkadaran Cox (Cox PH) dan dua kaedah pengelas pembelajaran mesin yang diselia (pepohon keputusan (DT) dan rangkaian neural buatan (ANN)) digunakan untuk memodelkan dan meramalkan kebarangkalian survival lima tahun. Model dengan indeks prestasi terbaik dibandingkan dengan instrumen PREDICT. Selepas itu, dalam Fasa 2, model tersebut telah dibangunkan dalam bentuk web berserta kandungan berkaitan untuk penghuraian instrumen tersebut. Laman web tersebut menjalani

beberapa peringkat pembangunan berulang termasuk penilaian kandungan (n = 8) dan kesahan muka (n = 20) oleh pakar perubatan dan pegawai perubatan.

Keputusan: Sejumlah 1,006 pesakit dilibatkan dalam analisis derivasi dan pengesahan model. Kebanyakan mereka adalah berbangsa Melayu, dengan karsinoma saluran, sensitif terhadap hormon, negatif HER2, pada peringkat T2, N1, tanpa metastasis, menerima pembedahan dan kemoterapi. Mengikut model Cox PH, etnik India mempunyai risiko kematian yang lebih tinggi berbanding kaum Melayu (Nisbah Bahaya Terlaras (Adj. HR): 1.77, 95% CI: 1.19, 2.63). Jenis histologi, gred kanser, tahap diagnosis kanser mengikut tumor, nodus limfa, dan metastasis adalah berkait secara signifikan dengan prognosis kematian. Mereka yang menerima sebarang rawatan pembedahan (Adj. HR: 0.49, 95% CI: 0.28, 0.87), kemoterapi (Adj. HR: 0.59, 95% CI: 0.44, 0.79), dan radioterapi (Adj. HR: 0.70, 95% CI: 0.51, 0.96) mempunyai risiko kematian yang lebih rendah. Model Cox PH mengatasi model DT dan ANN dari segi ketepatan (Cox PH: 0.841, DT: 0.811, ANN: 0.821), skor F1 (Cox PH: 0.879, DT: 0.859, ANN: 0.870) dan kawasan di bawah lengkung ciri operasi penerima (AUC: Cox PH: 0.891, DT: 0.39, ANN: 0.877). Model Cox PH adalah lebih tepat dalam meramalkan kebarangkalian survival lima tahun dengan nilai AUC yang lebih tinggi (0.78, 95% CI: 0.73, 0.82) berbanding PREDICT (AUC: 0.75, 95% CI 0.70, 0.80). Oleh itu, model itu digunakan sebagai ciri utama instrumen prognostik berasaskan web kami. Laman web tersebut dibangunkan dan ditambah baik pada setiap peringkat pembangunan. Indeks kesahan kandungan ialah ≥0.88 dan indeks kesahan muka ialah >0.90, dalam menghasilkan instrumen prognostik yang berfungsi dan mesra pengguna.

Kesimpulan: Instrumen berasaskan web yang dibangunkan daripada model Cox PH dengan prestasi yang teguh menunjukkan penemuan yang memberangsangkan. Kajian

pengesahan lanjut, kebolehgunaan dan kebolehlaksanaan adalah perlu kerana instrumen tersebut berpotensi digunakan oleh penyedia penjagaan kesihatan dalam menyampaikan ramalan survival individu kepada pesakit kanser payudara yang baharu didiagnosis.

Kata kunci*:* neoplasma payudara, model ramalan, analisis survival, instrumen prognostik, wanita Malaysia

ABSTRACT

DEVELOPMENT OF MALAYSIAN BREAST CANCER SURVIVAL PROGNOSTIC TOOL (myBeST) FOR PREDICTION OF SURVIVAL PROBABILITY AMONG WOMEN WITH BREAST CANCER IN MALAYSIA

Background: Breast cancer accounts for a sizeable portion of newly diagnosed cancer. Prognostic tools were developed to inform patients regarding their outcomes. Performance of Western-centric tools found to be less accurate when applied in our setting with PREDICT breast cancer (PREDICT) had an acceptable accuracy.

Objective: The study aimed to develop predictive models for survival among women with breast cancer in Malaysia, to compare its performance with PREDICT and the model's algorithm was incorporated to develop a web-based Malaysian Breast Cancer Survival Prognostic Tool (myBeST).

Methodology: This study consists of two phases. Phase 1 is a retrospective cohort study using data abstracted from seven regional breast cancer referral centres in Malaysia. We collected 13 predictors and survival outcomes. Time-to-event Cox proportional hazard (PH) analysis and two supervised machine learning classifiers (decision tree (DT) and artificial neural networks (ANN)) were employed to model and predict five-year survival probability. The model with the best performance indices was compared with the PREDICT tool. Subsequently, in Phase 2, the model was deployed in a web-based format with accompanying content to describe the tool. The website underwent several user-centred iterative development stages, including

content (n = 8) and face validity (n = 20) assessments by medical specialists and medical officers.

Results: There were 1,006 patients included for model derivation and validation. They were mostly Malay, with ductal carcinoma, hormone-sensitive, HER2-negative, at T2, N1-stage, without metastasis, received surgery and chemotherapy. The five-year survival was 60.5% (95% CI: 57.6, 63.6). By the Cox PH model, Indians had a higher hazard of death compared to Malay (Adjusted HR (Adj. HR): 1.77, 95% CI: 1.19, 2.63). Histological type, cancer grade, tumour, node, and metastasis stage at diagnosis significantly associated with death. Those who received surgery (Adj. HR: 0.49, 95% CI: 0.28, 0.87), chemotherapy (Adj. HR: 0.59, 95% CI: 0.44, 0.79), and radiotherapy (Adj. HR: 0.70, 95% CI: 0.51, 0.96) had a lower risk of death. Cox PH model outperformed the DT and ANN model in terms of accuracy (Cox PH: 0.841, DT: 0.811, ANN: 0.821), F1-score (Cox PH: 0.879, DT: 0.859, ANN: 0.870) and the area under the receiver operating characteristic curve (AUC; Cox PH: 0.891, DT: 0.39, ANN: 0.877). The Cox PH was more accurate in predicting five-year survival probability with a higher AUC (0.78, 95% CI: 0.73, 0.82) than PREDICT (AUC: 0.75, 95% CI 0.70, 0.80). Thus, the model was deployed as the main feature of our web-based prognostic tool. The website was developed and improved at every iterative stage. The content validity indices were ≥0.88 and face validity indices were >0.90, resulting in a functioning and user-centred prognostic tool.

Conclusion: The web-based tool derived from robust Cox PH model showed promising results. Further validation, usability, and feasibility studies are necessary as the tool could potentially be used by care providers to convey individualised survival prediction for newly diagnosed breast cancer patients.

Keywords: breast neoplasm, predictive model, survival analysis, prognostic tool, Malaysian women

CHAPTER 1

INTRODUCTION

1.1 Background

1.1.1 Breast cancer as a public health priority

Breast cancer is the most common cancer among women worldwide making it a global public health priority. Nearly all countries and territories (Figure 1.1) rank breast cancer as the leading type of female cancer (Bray *et al.*, 2018; Sung *et al.*, 2021).

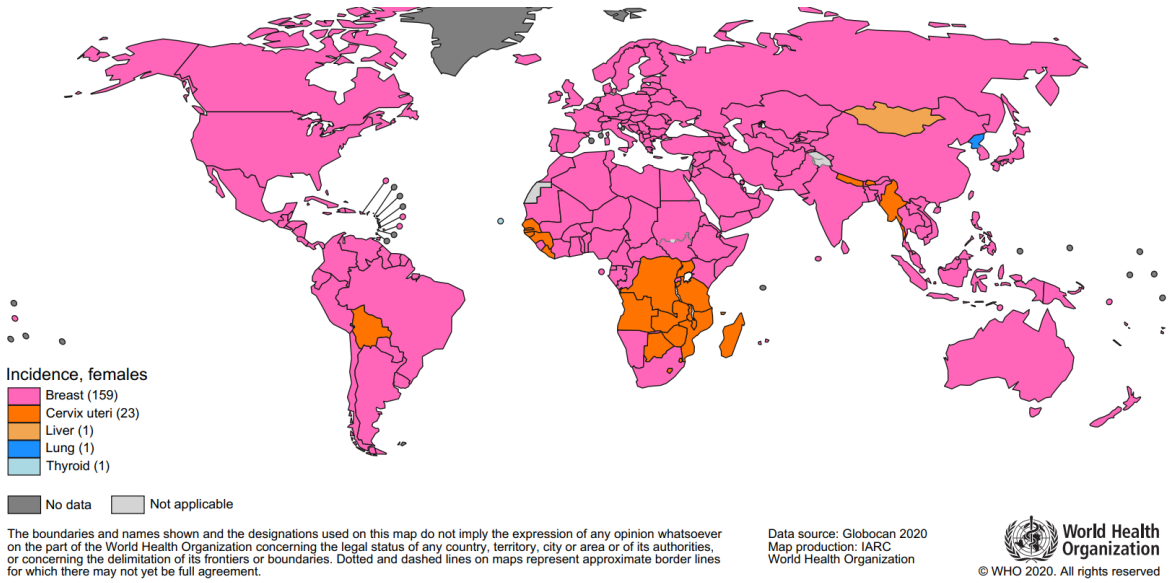

Figure 1.1: Global Maps Presenting the Most Common Type of Cancer Incidence in 2020 in Each Country Among Women (Sung *et al.*, 2021)

A comparable situation occurred in Malaysia with more than 21,000 new cases detected within five years between 2012 and 2016. The age-standardised incidence rate (ASR) for female breast cancer is 34.1 per 100,000 residents leaving colorectal

cancer far behind as the second commonest cancer (National Cancer Institute Ministry of Health (NCI MOH) Malaysia, 2019).

The number of newly-diagnosed breast cancer cases is increasing, albeit reducing trend of mortality rates and disability-adjusted life-years, DALY (Ji *et al.*, 2020). Recently diagnosed breast cancer patients confront immense challenges such as life-changing experiences, uncomfortable long-term treatments and follow up as well as financial difficulties.

1.1.2 Breast cancer survival

One out of four cancer mortality annually is estimated due to breast cancer (Bray *et al.*, 2018; Sung *et al.*, 2021). Cancer disproportionately affects women in low- and middle-income countries such as Malaysia. Five-year relative breast cancer survival for breast cancer patients diagnosed between 2007 and 2011 in Malaysia is 66.8%. The survival rate (Figure 1.2) was markedly lower than in other developed countries such as Japan, Australia, and neighbouring Singapore (NCI MOH Malaysia, 2018).

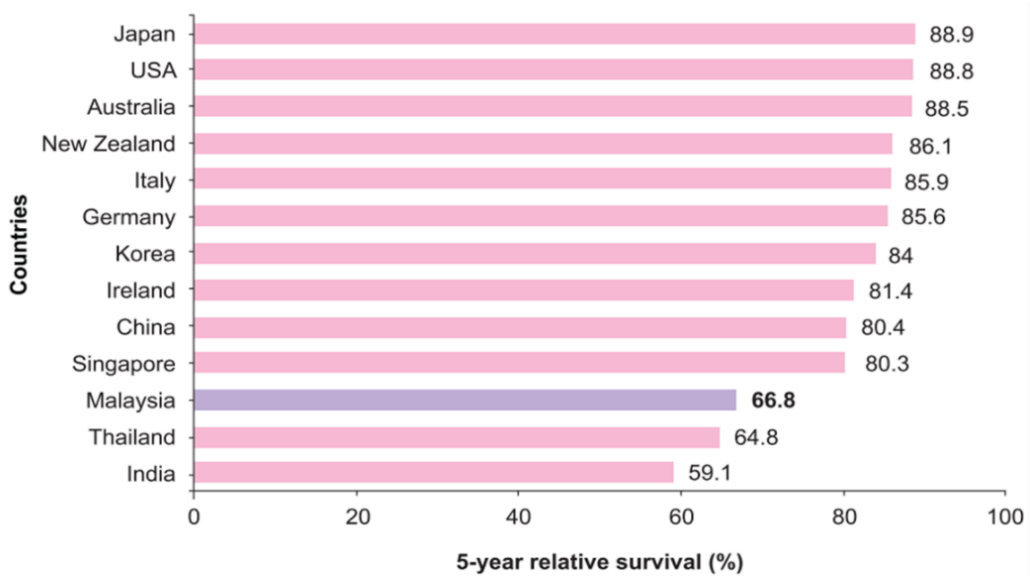

Figure 1.2: Female Breast: International comparison of 5-year relative survival (NCI MOH Malaysia, 2018)

The country-level indicators such as cancer incidence, mortality, and survival based on population-based cancer registries are essential for evaluating healthcare performance. Policymakers and researchers are interested in comparing the findings between population and changes over time. However, patients and care providers are more concerned with individualised survival based on specific patients' characteristics to provide more precise survival estimations (Moons *et al.*, 2009).

Individualised survival will inform the stakeholders of their disease's possible outcome and offer evidence to decide on treatment options (Moons *et al.*, 2009). Predicting breast cancer prognosis evolves from relying on physician experience to using clinical and pathological parameters to group patients into Tumour Node Metastasis (TNM) stages as the basis for communicating survival information. In addition, recent studies incorporate biomarkers such as genes and hormonal receptors to develop a prognostic model to increase the accuracy of estimating survival (Moons *et al.*, 2009; Hippisley-Cox & Coupland, 2017).

1.1.3 Breast cancer prognostic model

A prognostic model is a prediction model to estimate the probability that a specific event will occur (Figure 1.3). Patients or individuals in a health state, taking into account their characteristics as predictors, were monitored longitudinally to observe the event's occurrence (Collins *et al.*, 2015).

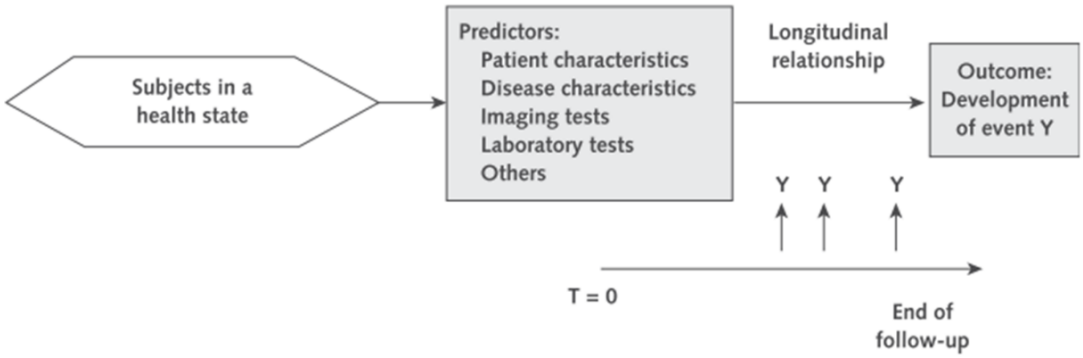

Figure 1.3: Schematic representation of prognostic prediction modelling studies (Collins *et al.*, 2015)

Breast cancer survival prognostic models developed before 2017 were mainly based on the western hemisphere and developed countries' patients (Phung *et al*., 2019). Studies conducted among the Asian population were from China, including Taiwan and South Korea. The frequently validated and helpful models were the Nottingham Prognostic Index (NPI), Adjuvant! Online, Cancer Math and PREDICT breast cancer (PREDICT) (Shachar & Muss, 2016; Phung *et al*., 2019). The three later models were deployed as online tools.

PREDICT was the latest model to be developed, and the up-to-date version was published in 2017 (Wishart *et al.*, 2010; Candido dos Reis *et al.*, 2017). Furthermore, among the three validation studies of existing online tools conducted among the Malaysian population, PREDICT had the best performance compared to Adjuvant! Online and CancerMath. However, these tools generally overestimate the survival probability, and the models' overall performances were modest (Bhoo-Pathy *et al*., 2012b; Wong *et al.*, 2015; Miao *et al.*, 2016).

1.2 Problem statement

Besides overall well-being and physical functioning, overall survival was rated as the most important outcome for Malaysian women with breast cancer (Rajaram *et al.*, 2019). Survival estimates based on the overall TNM stage at diagnosis alone are useful but less accurate in communicating survival estimates than the combination of clinical and pathological characteristics (Hippisley-Cox & Coupland, 2017).

Most of the prognostic models derived from developed countries' populations showed mixed performance if applied to the independent populations. Hence, prognostic models are preferably tailored to a specific population if the existing models are inconsistent with the settings' experience (Yip *et al.*, 2014; Phung *et al*., 2019).

Furthermore, breast cancer survival studies among Malaysian consistently showed the association between ethnicity and risk of death, which necessitate the inclusion of the variable into a new predictive model. The published research in Malaysia is primarily interested in identifying and determining the magnitude of the association between the independent variables and mortality. They are not translated into developing a predictive model to determine the survival probability (Abdullah *et al.*, 2013; NCI MOH Malaysia, 2018; Nordin *et al.*, 2018; Ganggayah *et al.*, 2019). To the best of our knowledge, there was no accessible online individualised survival prognostic calculator for women with breast cancer in Malaysia.

1.3 The rationale of the study

This study developed new prognostic models taking into consideration local experience in an attempt to provide a more precise individualise survival prediction. An online tool known as the Malaysian Breast Cancer Survival Prognostic Tool (myBeST) was developed based on the best-performing model. The tool is readily accessible for stakeholder use (i.e., care providers, patients, and families). At this stage, the tool was developed primarily for healthcare providers to assist them in conveying cancer prognosis to the patients upon diagnosis and prior receiving treatments.

This tool allows healthcare professionals to provide survival probability based on personalised patient parameters upon diagnosis. Malaysia's breast cancer care centres routinely collected the predictors' information for straightforward application in the clinical setting. On the other hand, patients and relatives were more informed regarding the disease trajectory and better prepared mentally and emotionally using evidence-based local experience. Hence, it could subsequently manage their expectations and improve adherence to the providers' recommendations, which aligns with the goal of tertiary prevention to maximise the outcomes (such as survival and quality of life) and prevent further morbidity and mortality of cancer.

Continuous feedback, recalibration, and validation by independent researchers are encouraged by deploying the tool as a web-based application for further improvement. The newly developed prognostic model could act as a baseline reference for future breast cancer survival prediction modelling studies in Malaysia as nationwide electronic medical records become available.

1.4 Research questions

1. How the newly developed predictive models for survival perform in predicting breast cancer survival probability among women with breast cancer in Malaysia?
2. How good is the newly developed prognostic model's performance compared to PREDICT breast cancer, a well-known web-based breast cancer survival prognostic tool?
3. How the contents and user interface of the web-based prognostic tool (myBeST) developed?

1.5 Objectives

1.5.1 General objective

To develop predictive models for survival among women with breast cancer in Malaysia, to compare a selected best-performing model with PREDICT breast cancer and the model's algorithm was used to develop a web-based Malaysian Breast cancer Survival prognostic Tool (myBeST).

1.5.2 Specific objectives

Phase I:

1. To develop and validate the survival probability predictive models for newly diagnosed women with breast cancer in Malaysia.
2. To compare a selected best-performing model's performance with PREDICT breast cancer, a well-known web-based breast cancer survival prognostic tool.

Phase II:

3. To develop content and user interface of a web-based Malaysian Breast Cancer Survival Prognostic Tool (myBeST) using the selected model's algorithm

Specific objective 1, 2 and 3 are addressed in CHAPTER 4, 5 and 6, respectively.

1.6 Research hypotheses

1. The newly developed predictive models have good performances in predicting breast cancer survival probability among breast cancer in Malaysia.
2. The selected model performance is comparable with PREDICT breast cancer.
3. myBeST's content and user interface are developed with good content and face validity resulting in functioning web-based prognostic tool.

CHAPTER 2

LITERATURE REVIEW

Literature related to this study is discussed and organised into these subheadings:

- Breast cancer survival study
- Breast cancer survival prognostic models
- Developing a prognostic model
- Presenting the prognostic model

2.1 Breast cancer survival

The pooled five-year survival rate for women with breast cancer was 0.74 (95% CI: 0.66, 0.80) as found by a meta-analysis study involving 14 studies published between 2010 and 2017 (Maajani *et al.*, 2019). There were marked disparities for five-year net survival within each continent for those diagnosed between 2010 and 2014 with patients in the North America and Oceania had the survival of almost 90% (Allemani *et al.*, 2018; Zaidi & Dib, 2019).

Within Asian region, women with breast cancer in Malaysia and Thailand has comparable five-year relative survival of 66.8% (95% CI: 66.0, 67.6) and 68.7% (95% CI: 66.6, 70.8), respectively. However, the survival was lower than those in Singapore (80.3% , 95% CI: 78.4, 82.2), South Korea (86.6% , 95% CI: 85.8, 87.5) and Japan (89.9% , 95% CI: 88.9, 89.9) (Allemani *et al.*, 2018; NCI MOH Malaysia, 2018). The disparities within the same continent could be attributed to differential socioeconomic

level between middle- and high-income countries which influence early detection and optimal access to medical care (Saxena *et al.*, 2012; Coughlin, 2019).

2.1.1 Malaysian breast cancer survival study

Several studies have examined breast cancer survival among Malaysians. These studies found a significant association between ethnicity and the hazard of death. Existing prognostic models often overlooked the local ethnic group as one of the required predictors (Abdullah *et al.*, 2013; NCI MOH Malaysia, 2018; Nordin *et al.*, 2018; Ganggayah *et al.*, 2019).

The studies which used population-based cancer registry data contain limited variables with a high proportion of missing data for developing an accurate predictive model (Abdullah *et al.*, 2013; NCI MOH Malaysia, 2018; Nordin *et al.*, 2018; Tan *et al.*, 2021). Likewise, the hospital-based studies are limited to one centre that could not represent diverse Malaysian breast cancer patients (Abdullah *et al.*, 2016; Azman *et al.*, 2019; Ganggayah *et al.*, 2019). The summary literature review of recent breast cancer survival studies among Malaysian patients as displayed in Table 2.1.

Table 2.1: Summary literature of Malaysian breast cancer survival studies

Author, Year	**Patient**	**Factors associated with survival**	**Analysis**
Population-based cancer registry			
(NCI MOH Malaysia, 2018)	n = 17,009 Year of diagnosis: 2007-2011	Stage at diagnosis, age group at diagnosis, ethnicity	Cox Proportional Hazard regression
(Nordin *et al.*, 2018)	n = 549 Year of diagnosis: 2007-2011	Ethnicity, stage at presentation, history of surgical treatment	Cox Proportional Hazard regression

Table 2.1 continued.

Author, Year	Patient	Factors associated with survival	Analysis
Population-based cancer registry			
(Abdullah *et al.*, 2013)	n = 10,230 Year of diagnosis: 2000-2005	Age at diagnosis, ethnicity	The log-rank test
(Tan *et al.*, 2021)	n = 2,166 Year of diagnosis: 2010-2014	Age groups, disease stage, treatment receipt	Cox proportional hazard regression
Hospital-based studies			
(Ganggayah *et al.*, 2019)	n = 8,066 Year of diagnosis: 1993-2016	23 predictors (marital status, menopausal status, presence of family history, race, method of diagnosis, classification of breast cancer, laterality, cancer stage classification, grade of differentiation in tumour, oestrogen receptor (ER) status, progesterone receptor (PR) status, c-er-b2 status, primary treatment type, surgery status, type of surgery, method of axillary lymph node dissection, radiotherapy, chemotherapy, hormonal therapy)	Six machine learning algorithms (decision tree, random forest, neural networks, extreme boost, logistic regression, and support vector machine)
(Azman *et al.*, 2019)	n = 214 Year of diagnosis: 2008-2012	Marital status, lymph node involvement	Cox Proportional Hazard regression
(Abdullah *et al.*, 2016)	n = 675 Year of diagnosis: 2008-2012	Stage at diagnosis	Relative survival analysis

2.2 Breast cancer survival prognostic models

Existing prognostic models that were frequently validated include Nottingham Prognostic Index (NPI), Adjuvant! Online and PREDICT breast cancer (Phung *et al*., 2019). NPI was the earliest breast cancer prognostic model developed since 1982. The model used datasets from breast cancer patients treated at Nottingham City Hospital, United Kingdom. It is a simple model with the initial model consisting of tumour size, pathological grade, and nodal stage as the predictors. The model was found to be more discriminating from the lymph-node stage alone that was used prior. Five new models were derived from NPI (Haybittle *et al.*, 1982; Phung *et al*., 2019).

Adjuvant! Online used meta-analysis results for treatment efficacy and Surveillance, Epidemiology, and End-results (SEER) data comprised of the United States patients. The tool calculates individual patient 10-year survival probability and risks of relapse. The model showed varied performance, i.e., good performance among the Dutch population but otherwise poor performance among patients in the United Kingdom and Asia. However, the model ceased to be available online (Ravdin *et al.*, 2001; Shachar & Muss, 2016; Phung *et al*., 2019).

In addition to Adjuvant! Online, Cancer Math for breast cancer was developed using the SEER database. The tool used a different mathematical equation based on the binary biological model of cancer metastasis. The information on tumour size, nodal status and other prognostic factors was used to calculate cancer mortality, life expectancy, and treatment benefit. The model can stratify patients into groups with merely 2% differences between observed and predicted risk of death in the validation dataset (Michaelson *et al.*, 2011).

2.2.1 PREDICT breast cancer tool

PREDICT breast cancer was developed based on 5,694 women treated in East Anglia, the United Kingdom, between 1999 and 2003. The model employed Cox proportional hazards regression to calculate cumulative hazard to determine overall and breast cancer-specific survival (Wishart *et al.*, 2010). The tool is readily accessible online (https://breast.predict.nhs.uk). It is ongoing maintenance and planned for an update to improve the current version. The latest version, i.e., version 2, was published in 2017 (Candido dos Reis *et al.*, 2017).

The tool received endorsement from the American Joint Commission on Cancer (AJCC) and the United Kingdom medical professional body (Candido dos Reis *et al.*, 2017; NICE, 2018). Table 2.3 and Figure 2.1 summarise the aforementioned tools, their performances and other prognostic models.

Table 2.2: Summary of predictors, analysis, outcome and performance of the existing breast cancer survival prognostic models

Prognostic tool/model	**Patients' data source**	**Predictors**	**Analysis**	**Outcome/ Performance**
PREDICT breast cancer (Wishart *et al.*, 2010; Candido dos Reis *et al.*, 2017)	Develop based on 5,694 women treated in East Anglia from 1999-2003	Age, menopausal status, ER, HER2, Ki-67 status, tumour size, grade, positive nodes, hormone treatment, chemotherapy	Cox proportional hazards regression	overall survival and breast cancer-specific survival v1 AUC: 0.79; v2 AUC: 0.75
Cancer Math (Michaelson *et al.*, 2011)	Surveillance, Epidemiology and End Results (SEER) database, USA	Age, menopausal status, ER, HER2, tumour size, grade, positive nodes, hormone treatment, chemotherapy	the SNAP (Size + Nodes + Prognostic markers) method based on the binary biological model of cancer metastasis	cancer mortality, life expectancy, therapy benefit c-index: 0.93
Adjuvant! Online (Ravdin *et al.*, 2001)	Surveillance, Epidemiology and End Results (SEER) database and treatment efficacy data from meta-analyses	Age at diagnosis, lymph node status, tumour size, ER, Comorbidity, Menopausal status	Bayesian method	10-survival probabilities and risks of relapse
Nottingham prognostic index (NPI) (Haybittle *et al.*, 1982)	Patient in Nottingham City Hospital	Tumour size, pathological grade, and nodal stage	Cox proportional hazards	Overall survival
Commission on Cancer's "Cancer Survival Prognostic Calculator" (Asare *et al.*, 2016)	National Cancer Data Base (NCDB), US; Year of diagnosis: 2003–2006	T, N, and M stage, age, chemotherapy, surgery, radiation therapy, histologic grade, ER and PR status	multivariable Cox proportional hazards regression model	Overall survival c-index: 0.78-0.79

Table 2.2, continued.

Prognostic tool/model	Patients' data source	Predictors	Analysis	Outcome/ Performance
New Zealand Model (NZM) (Elwood *et al.*, 2018)	Breast cancer registry in New Zealand Year of diagnosis: 2000–2014	Age, ethnicity, tumour size, number of positive lymph nodes, tumour grade, presence of metastasis, ER and PR status, HER2 status, histological type, lymphovascular invasion (LVI)	Multivariable Cox proportional hazards regression model	10-year breast cancer-specific survival c-index: 0.83
Prediction model, Taiwan (Huang *et al.*, 2019)	Taiwan Cancer Registry (TCR); Year of diagnosis: 2011-2015	Clinical and pathological factors; treatments	Cox proportional hazards model	overall survival and breast cancer-specific survival AUC > 0.85
Proview (Seow *et al.*, 2020)	Ontario Cancer Registry, Canada; Year of diagnosis: 2008-2013	Demographic characteristics, clinical data, treatment received, patient-reported outcomes, health care use	Cox proportional hazards regression model	Conditional survival probabilities c-index: 0.91
Predictbcos nomogram (Ji, *et al.*, 2020)	Surveillance, Epidemiology and End Results (SEER) database; Year of diagnosis: 2007- 2012	Clinicopathological factors and socioeconomic factors (marital status and education level)	Cox proportional hazards regression model	Breast cancer-specific survival and overall survival c-index: 0.77

AUC, area under receiver operating characteristics (ROC) curve; c-index, concordance index; ER, Oestrogen receptor; PR, Progesterone receptor; HER2, Human epidermal growth factor receptor 2, USA, the United States of America; T, Tumour; N, Node; M, Metastasis; v1, version 1; v2, version 2

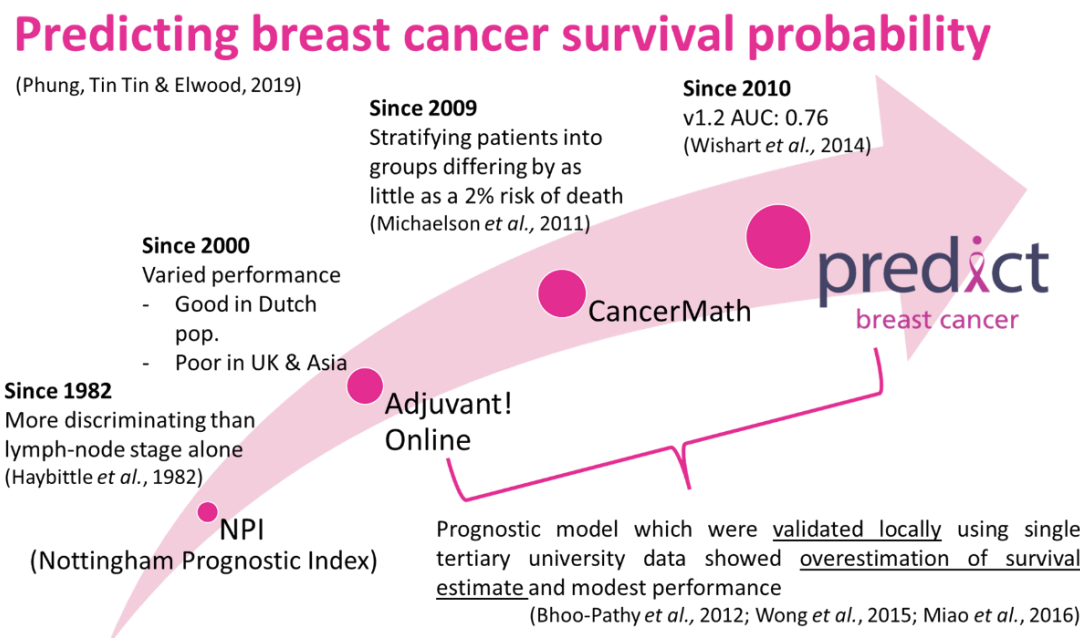

Figure 2.1: Summary of the commonly used prognostic tools and their performance

2.2.2 Validation studies of local patients

Previously mentioned tools comprised Adjuvant! Online, Cancer Math, and PREDICT breast cancer was validated locally. However, the validation dataset was limited to those attending one urban academic centre that might differ from the rest of Malaysian breast cancer patients' experiences. The patients were diagnosed between 1990 and 2011 (Bhoo-Pathy *et al.*, 2012b; Wong *et al.*, 2015; Miao *et al.*, 2016).

These tools showed modest performance with discriminant indices of the area under receiver operating characteristics (ROC) curve between 0.73 and 0.78. The overall model's calibration measures in terms of predicted survival were found to deviate between 1.3% (PREDICT) and 8.0% (Cancer Math) from the observed survival (Bhoo-Pathy *et al.*, 2012b; Wong *et al.*, 2015; Miao *et al.*, 2016). The summary of the validation studies is presented in Table 2.3.

Table 2.3: Summary of validation studies conducted among Malaysian patients

Prognostic tools and study (Author, year)	**Malaysian data sources**	**Findings**
PREDICT breast cancer (Wong *et al.*, 2015)	Breast cancer registry of University Malaya Medical Centre (n = 1,480, year of diagnosis 1998-2006)	Outcome: 5-year survival - AUC: 0.78 (95% CI: 0.74 – 0.81) - predicted (86.3%) vs observed (87.6%) Outcome: 10-year survival - AUC: 0.73 (95% CI: 0.68–0.78) - predicted (77.5%) vs observed (74.2%) - Overestimated survival in patients aged <40 years
Cancer Math (Miao *et al.*, 2016)	Singapore Malaysia Hospital-Based Breast Cancer Registry (n for Malaysian = 2,143, year of diagnosis 1990-2011)	Outcome: 5-year survival - AUC: 0.77 (95 % CI,0.75–0.79) - predicted (86.1%) vs observed (80.3%) Outcome: 10-year survival - AUC: 0.74 (95 % CI,0.71–0.76) - predicted (73.3%) vs observed (65.3%)
Adjuvant! Online (Bhoo-Pathy *et al*., 2012b)	University Malaya Hospital-Based Breast Cancer Registry (n = 641, year of diagnosis 1993-2000)	Outcome: 10-year survival - AUC: 0.73 (95% CI: 0.69 – 0.77) - predicted (70.3%) vs observed (63.6%)

AUC: area under receiver operating characteristics (ROC) curve

2.2.3 Selection of PREDICT tool as a comparison tool

PREDICT breast cancer is chosen to be compared with our newly developed tool due to its recent development and endorsement by well-known professional bodies. The prognostic tool had the best performance in comparison to the Adjuvant! Online and Cancer Math when applied in previous validation studies conducted in Malaysia (Bhoo-Pathy *et al*., 2012b; Wong *et al.*, 2015; Miao *et al.*, 2016). It consistently

undergoes improvements, and the future version (v2.3) will include PR status as one of the predictors (Grootes *et al.*, 2022). PREDICT breast cancer is accessible online, available as an R package "nhspredict" for analysis and shares a similar outcome of the interest in this study, five-year overall survival.

2.3 Developing a prognostic model

A prognostic model or tool is also known as a prediction model, prediction rule, or risk score. Developing a prognostic model involves building an accurate and discriminating model from multiple variables, as rarely one variable can accurately predict the outcome. The model must be relevant to the intended purpose and setting with routinely available predictors for seamless application (Moons *et al.*, 2009; Royston *et al.*, 2009; Collins *et al.*, 2015; Harrell, 2015).

The critical process of developing a prognostic model consists of selecting the relevant or clinically important candidate predictors, appropriate data collection method, judging data quality including missing data, appropriate modelling approaches (i.e., statistical model specifications or machine learning algorithm) and measuring the prediction performance (Moons *et al.*, 2009; Royston *et al.*, 2009; Collins *et al.*, 2015; Harrell, 2015).

Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guideline is a widely used recommendation to report research related to prognostic modelling in the medical domain. Experts comprised of methodologists, health care professionals, and journal editors developed the guideline to address the previous studies' poor reporting. Thus,

allowing the assessment of the prediction models' usefulness and biases (Collins *et al.*, 2015).

The guideline covers prediction model studies that use:

- Only a single dataset for model development and validation
- A portion of a single dataset to develop the model and the other portion for validation
- A data set for model development and separate data set to validate the model

The TRIPOD Statement contains 22 items (Appendix F). The statement is essential to produce fair quality reporting of studies involving prognostic modelling. This study described the findings as per the checklist.

A comprehensive systematic review of breast cancer prognostic model studies published between 1982 and 2016 analysed 96 eligible articles and found the commonest methods for model development are Cox proportional hazards (PH) regression, followed by artificial neural network and decision trees algorithm. The review includes studies that employ multivariable prognostic modelling of clinical and pathological variables. The review excluded genetic studies. A summary of the review's findings is presented in Table 2.4. Beyond 2016, Cox PH continues to be the dominant method for model development in addition to the supervised machine learning method (Table 2.4).

Table 2.4: Summary of findings from the systematic review of breast cancer prognostic models, 1982-2016 (Phung *et al*., 2019)

Characteristics	Number of models
Total	58 models
Outcomes	
Mortality	28 models
Recurrence	23 models
Both	7 models
Methods for development	
Cox PH regression	32 models
Artificial neural networks	6 models
Decision tree	4 models
Logistic regression	3 models
Bayesian method	3 models
Multistate model	2 models
Support vector machine	2 models
Others	6 models
Predictors	
Nodal status	49 models
Tumour size	42 models
Tumour grade	29 models
Age at diagnosis	24 models
ER status	21 models
Treatment	17 models
HER2 status	13 models
PR status	10 models
Lymphovascular invasion (LVI)	8 models
Stage	8 models
Others	Mitotic activity index (MAI), histological subtypes, comorbidity, menopausal status
Presentation of model	
Regression formula	13 models
Online tool	8 models
Decision tree	5 models
Nomogram	4 models
Score chart	1 model
No report	27 models

PH, proportional hazard; ER, Oestrogen receptor status; PR, Progesterone receptor; HER2, Human epidermal growth factor receptor

2.3.1 Predictors of breast cancer survival

The survival of women with breast cancer depends on multiple clinical and pathological predictors. They could be categorised into sociodemographic predictors,

cancer characteristics including the extension of the disease, treatment receipt, and other predictors such as comorbidities and mode of detection. The commonest predictors used to develop prognostic models are nodal status, tumour size, tumour grade, age at diagnosis, and oestrogen receptor (ER) (Phung *et al*., 2019). The predictors used in the existing prognostic model are summarised in Table 2.4.

2.3.1 (a) Sociodemographic predictors

Age at diagnosis is an essential sociodemographic and clinical predictor in determining breast cancer survival (Phung *et al*., 2019). Women with breast cancer at a younger age had lower survival and a higher hazard of death due to aggressive cancer type, especially those younger than 40 years old (Copson *et al.*, 2013; Brandt *et al.*, 2015; Chen *et al.*, 2016). On the other hand, extremely elderly patients are expectedly to have a higher risk of mortality due to the frailty of ageing and multiple comorbidities (Brandt *et al.*, 2015; McGuire *et al.*, 2015; Chen *et al.*, 2016).

Health outcomes disparities, including breast cancer survival, have been recognised globally, mainly albeit indirectly attributed to each individual's social circumstances (Coughlin, 2019; Kurani *et al.*, 2020). Ethnicity or race is among the most important social determinants of health. The relationship between ethnicity and mortality risk was documented across several studies involving women with breast cancer in Malaysia (Abdullah *et al.*, 2013; NCI MOH Malaysia, 2018; Nordin *et al.*, 2018; Ganggayah *et al.*, 2019). In addition, local ethnic groups or races were considered valuable predictors to be included in developing a new prognostic model in Taiwan and New Zealand (Elwood *et al.*, 2018; Huang *et al.*, 2019).

Other social determinants of health for predicting breast cancer mortality include marital status and education level. Single or divorced patients have a higher risk of

death than married patients (Ji *et al.*, 2020). Meanwhile, patients with better educational opportunities have better cancer survival (Ji *et al.*, 2020). However, a systematic review revealed inconclusive findings regarding the educational level and mortality risk (Lundqvist *et al.*, 2016).

Breast cancer is mainly a female disease, with male breast cancer merely representing 1% of the incidence. Male breast cancer is considered a distinct disease entity and is usually studied separately. Males had lower survival compared to females (Liu *et al.*, 2018). The situation could be attributed to the limited screening practice recommendations, low male breast tissue volume and diagnosis at a later stage (Ferzoco & Ruddy, 2016; Liu *et al.*, 2018; Gucalp *et al.*, 2019).

2.3.1 (b) Cancer Characteristics

The cancer stage refers to the extension of the disease upon diagnosis based on local tumour infiltration (T), involvement of the regional lymph nodes (N) and presence of distant metastasis (M). The American Joint Committee on Cancer, AJCC 7th edition Cancer Staging was introduced in 2010 (AJCC, 2010; Edge & Compton, 2010) and endorsed by Malaysia breast cancer clinical guideline (MOH Malaysia, 2010). The stage at diagnosis remained the principal predictor of cancer survival.

In addition to that, the extension of the disease includes lymphovascular invasion. Other crucial cancer characteristics predictors are histological type, cancer grade, oestrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2) and Ki-67 status. These histopathological predictors have grown in importance as more women are diagnosed early and the discovery of novel prognostic biomarkers. Recently, these predictors have been a routinely-documented in pathological reports of any breast cancer specimen in Malaysia (MOH Malaysia, 2019;

Phung *et al*., 2019; Li *et al.*, 2021). These predictors were incorporated into the eighth edition of the AJCC cancer staging manual as the clinical prognostic stage group to provide more individualised prediction outcomes (Giuliano *et al*., 2018; Kalli *et al*., 2018).

2.3.1 (c) Treatment receipt

Surgery is the mainstay treatment which includes mastectomy and breast-conserving surgery in conjunction with adjuvant (or neoadjuvant) chemotherapy. For early breast cancer, adjuvant radiotherapy is indicated for those who have breast-conserving surgery with a clear margin and post-mastectomy patients with positive lymph nodes and/or surgical margins that are not amenable to surgery (MOH Malaysia, 2019). Thus, the management of any breast cancer patient requires a multidisciplinary team approach. Patients who receive any treatment substantially improve their survival and prognosis.

Women with HER2-positive breast cancer who are receiving adjuvant chemotherapy should be offered to receive targeted therapy, such as trastuzumab. In cases of hormone receptor-positive breast cancer, adjuvant hormonal therapy may be recommended depending on the patient's risk of disease recurrence and potential side effects (MOH Malaysia, 2019).

2.3.1 (d) Other predictors

Other predictors that could influence breast cancer survival include the mode of cancer diagnosis and the presence of comorbidities. As a diagnosis by screening approach may add to bias in survival prediction, the mode of detection is considered a crucial predictor to be included in multivariable predictive modelling (Hofvind *et al.*, 2016).

Breast cancer patients with comorbidities such as peripheral vascular disease, dementia, chronic pulmonary disease, liver and renal diseases have a significantly increased risk of dying (Ewertz *et al.*, 2018). Those with a Charlson Comorbidity Index (CCI) of more than three are associated with a higher hazard of death (Parés-Badell *et al.*, 2017).

2.3.2 Modelling approaches

2.3.2 (a) Survival analysis and cox proportional hazard regression analysis

Survival analysis considers both the outcome (i.e., the event) and the time to the event or censored observation (i.e., the observation that does not develop the event). Cox proportional hazard (PH) regression is the commonly used modelling approach in multivariable survival studies. It is a semi-parametric survival analysis as it does not require the specification of baseline hazard function compared to the parametric survival analysis (Hosmer, Lemeshow & May, 2008; Kleinbaum & Klein, 2012).

2.3.2 (b) Supervised machine learning classifiers for survival status

Machine learning has emerged as a promising technique for handling high-dimensional data capable of identifying nonlinear patterns and optimising outcome prediction. It is increasingly applied in clinical decision support, such as cancer prediction and prognosis. In machine learning, several algorithms are trained and tested on other datasets. Then, the performance indices are used to select the best-performing algorithm (Ganggayah *et al.*, 2019; Das *et al.*, 2020; Senders *et al.*, 2020).

Two widely used supervised machine learning techniques in healthcare research are artificial neural networks and tree-based algorithms such as decision tree classification analysis. Artificial neural networks are frequently described as black