phonz

Public Health Observatory
of New Zealand Inc
Te Rōpū Mātai Hauora Pāpori
O Aotearoa

Public Health Observatory of NZ Te Ropu Matai Hauora Papori O Aotearoa
2024 Virtual Conference

A Robust Machine Learning Model for Long-Term Survival Prediction of Breast Cancer Patients in New Zealand

Rooshan Ghous & Binh Thanh Dang

School of Information Technology

WhiteCliffe College, Auckland, New Zealand

Whitecliffe

Te Rēhita
Mate Ūtaetae
Breast Cancer
Foundation
National
Register

Agenda

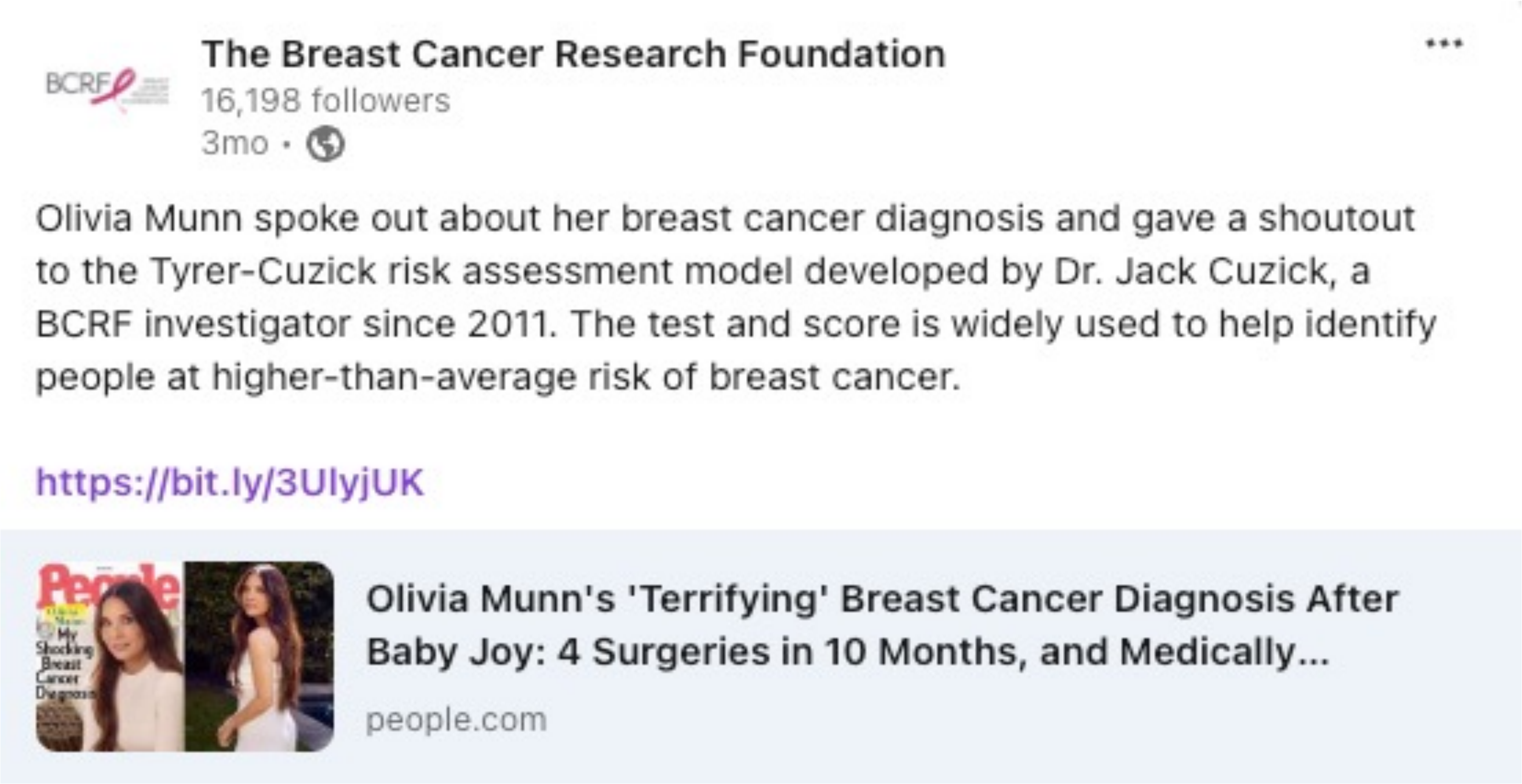
- Motivation
- Research Objectives
- Pilot Study – US SEER Dataset
- NZ Breast Cancer Dataset
- The Proposed Approach
- Preliminary Results and Discussion

Motivation

- Breast cancer is the most commonly affecting cancer in women & the third most prevalent cancer overall in NZ, resulting in a toll of 600 fatalities annually (MOH, 2023).

- A robust survival prediction model can help improve long term disease outcome in New Zealand by understanding its prognosis and targeting the risk factors.

Case Study



The Breast Cancer Research Foundation

16,198 followers

3mo •

Olivia Munn spoke out about her breast cancer diagnosis and gave a shoutout to the Tyrer-Cuzick risk assessment model developed by Dr. Jack Cuzick, a BCRF investigator since 2011. The test and score is widely used to help identify people at higher-than-average risk of breast cancer.

https://bit.ly/3UlyjUK

Olivia Munn's 'Terrifying' Breast Cancer Diagnosis After Baby Joy: 4 Surgeries in 10 Months, and Medically...

people.com

Why Develop a Local Survival Prediction Tool

Artificial Intelligence (AI) based tools for survival risk prediction are sensitive to training data and therefore need for a local prediction model.

- Help identify survival time precisely based on individual risk
- Selecting favorable treatment plan & targeting modifiable risk factors
- Triage for treatment & screening
- Assist in public health policymaking
- Reducing the ethnic disparity that exists in health outcomes of cancer patients

Research Objectives

1. Develop a robust 15-year survival prediction model for breast cancer using machine learning.
2. Identify temporal variations that may exist within risk factors during 5-year, 10-year & 15 years period post diagnosis.

Pilot Study: Surveillance, Epidemiology & End Result (SEER)Data

US Surveillance Epidemiology & End Result Data		
Patients diagnosed between 1/1/2010 -31/12/2014		
Total Number of Cases		**289,303**
Gender	Female	289,303
Cancer at Diagnosis	Localized to breast	271,763
	Distant Metastasis	17,540
Patient Status at 60 months	Alive	211,727
	Deceased	77,576

Comparison of ML Model Performance on SEER Data

Model Name	Area Under Curve (AUC)
Logistic Regression	0.8233
Decision Tree Classifier	0.6218
Random Forest Classifier	0.8000
XG Boost	0.8258
Deep Neural Network	0.8225

[Range for AUC 0-1]

Comparison of ML Model Performance on SEER Data

Model Name	Accuracy	Precision	Recall	F-1 Score
Logistic Regression	87.18%	88.60%	97.56%	92.86%
Decision Tree Classifier	86.61%	87.19%	98.85%	92.66%
Random Forest Classifier	86.65%	88.40%	97.11%	92.55%
XG Boost	87.37%	88.89%	97.39%	92.95%
K- Nearest Neighbor	85.68%	87.99%	96.40%	92.01%
Deep Neural Network	87.22%	87.80%	98.76%	92.96%

Te Rēhita
Mate Ūtaetae
Breast Cancer
Foundation
National
Register

NZ Breast Cancer Dataset

Data acquired from New Zealand Breast Cancer Foundation National Register

Inclusion criteria: Women registered between 01/01/2002 – 31/12/2017

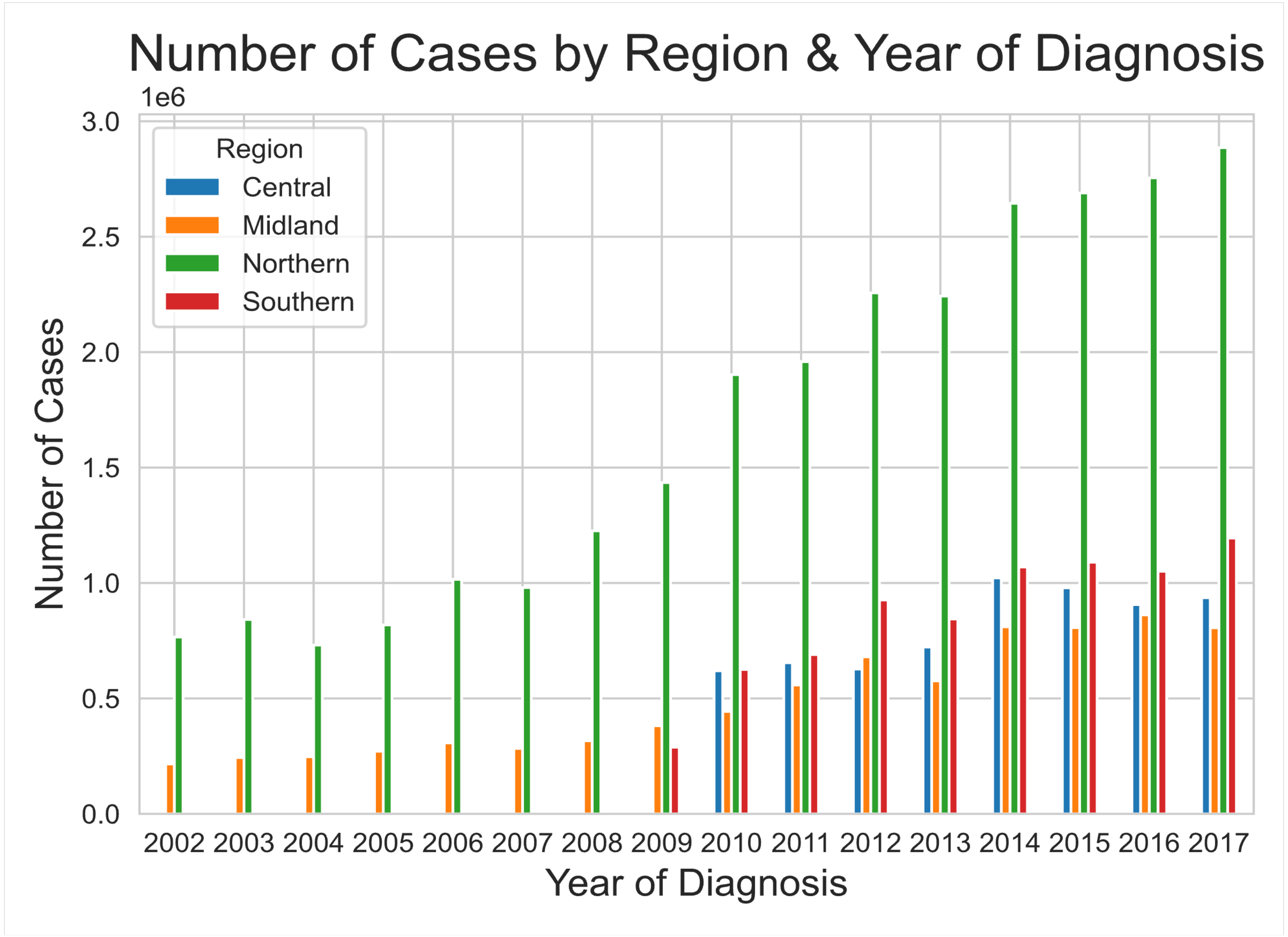
Exclusion criteria: Patients lost to follow-up, and those with metastasis at diagnosis

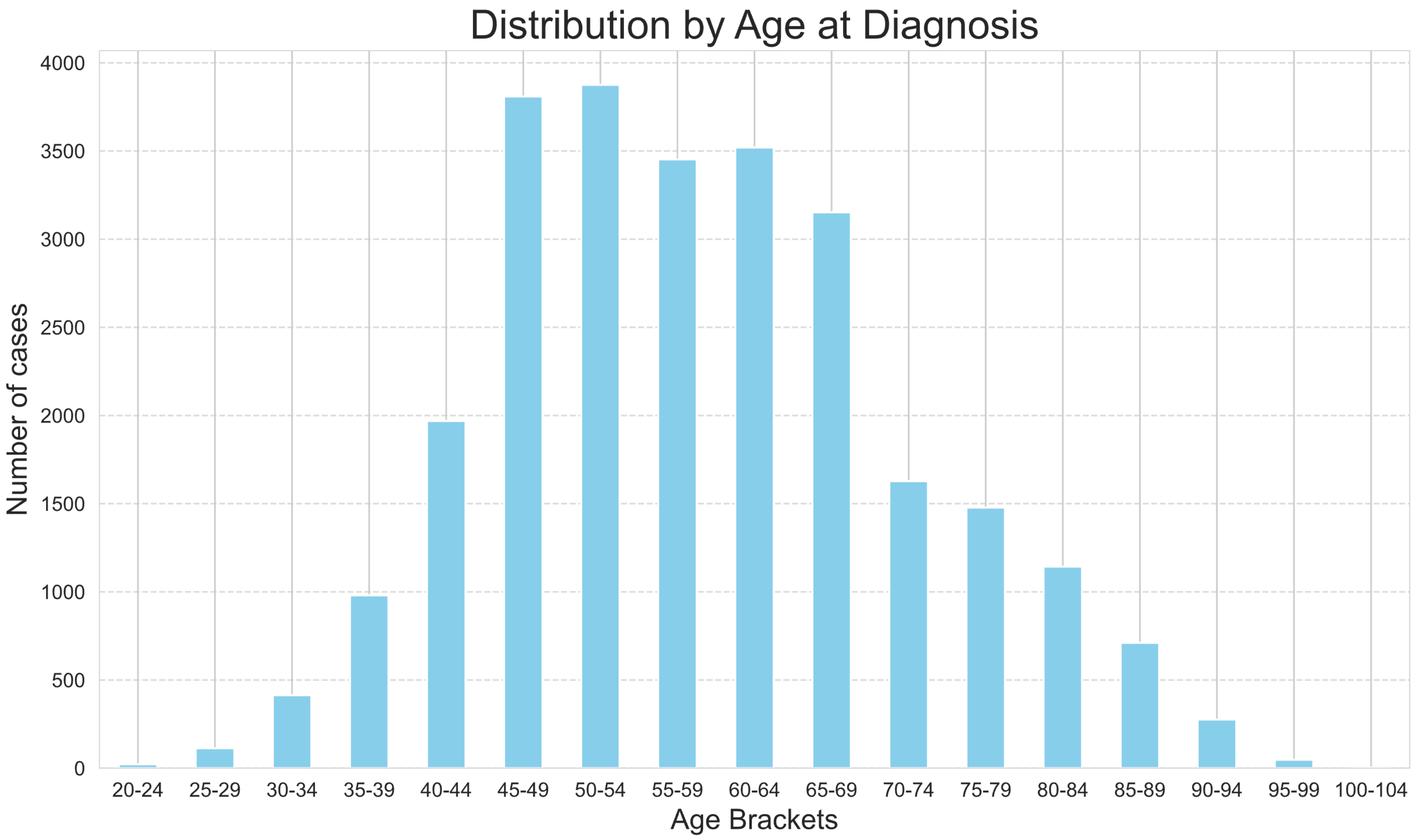
Cohort ID	Inclusion Period	Follow-up Period
Cohort 1	All patients diagnosed in 2002-2007	Followed up for 180 months to estimate 15-year survival
Cohort 2	All patients diagnosed in 2008-2012	Followed up for 120 months to estimate 10-year survival
Cohort 3	All patients diagnosed in 2013-2017	Followed up for 60 months to estimate 5-year survival

Dataset Summary

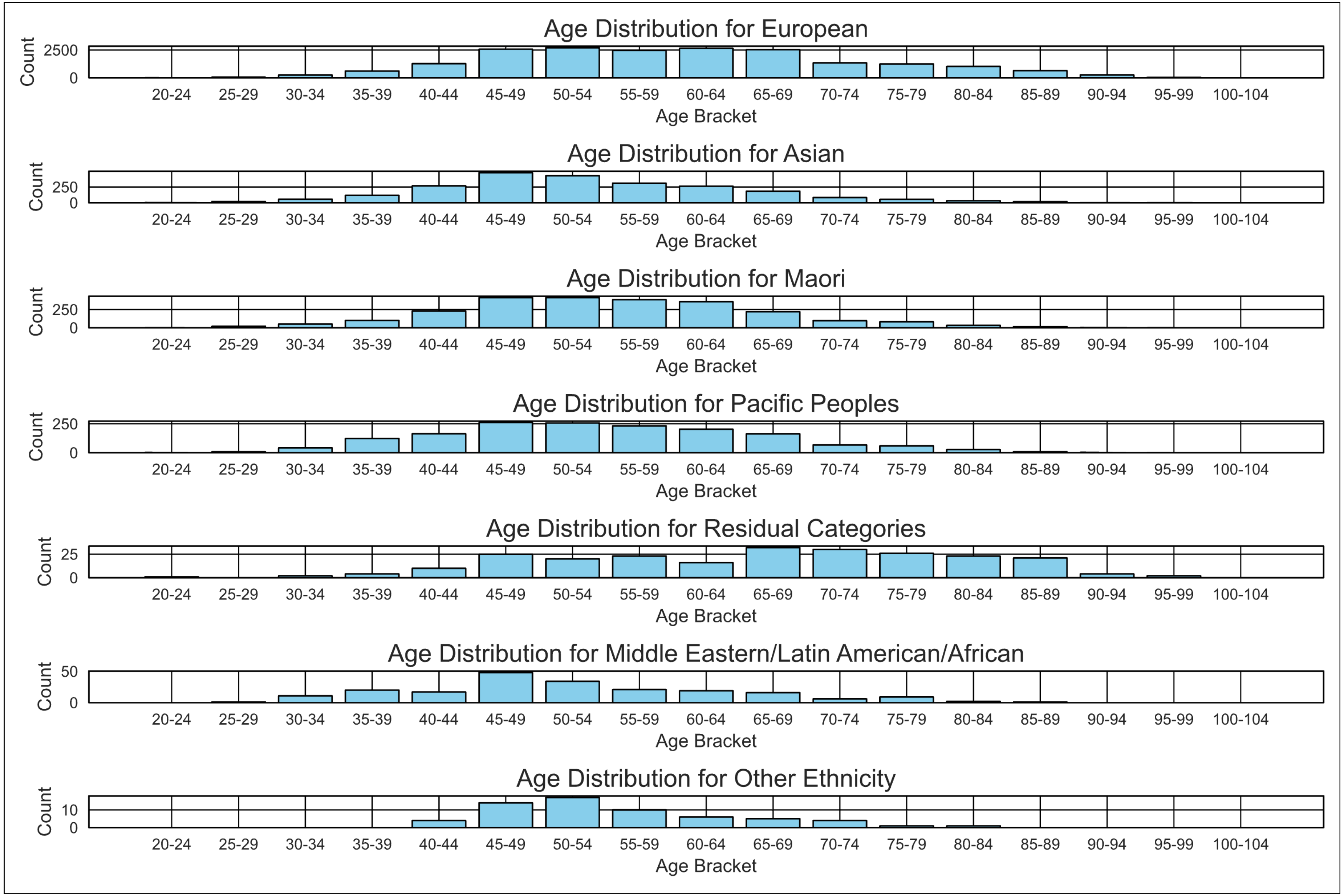
Te Rēhita Mate Ūtaetae - Breast Cancer Foundation National Register		
Total Number of Cases		**26,786**
Gender	Female	26,594
	Male	186
	Unknown	5
	Birth sex female	1
Cancer at Diagnosis	Localized to breast	22,443
	Distant Metastasis	4,124
	Unknown	27
Total Cases		22,470

About 230 predictor variables and 65 variables related to outcome and follow-up were requested for analysis


Number of Cases by Region & Year of Diagnosis
1e6
Number of Cases
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Year of Diagnosis
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Region
Central
Midland
Northern
Southern


Distribution by Age at Diagnosis
Number of cases
4000
3500
3000
2500
2000
1500
1000
500
0
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80-84
85-89
90-94
95-99
100-104
Age Brackets

Distribution of Diagnosis Age by Ethnicity



Multifactor Risk Model

Significant predictors of breast cancer survival across 7 fields including

1. Patient demographics
2. Clinical presentation and radiological features
3. Histological features including biomarkers
4. Stage & grade of tumor
5. Tumor characteristics including hormone receptor status
6. Treatment received
7. Histopathological details

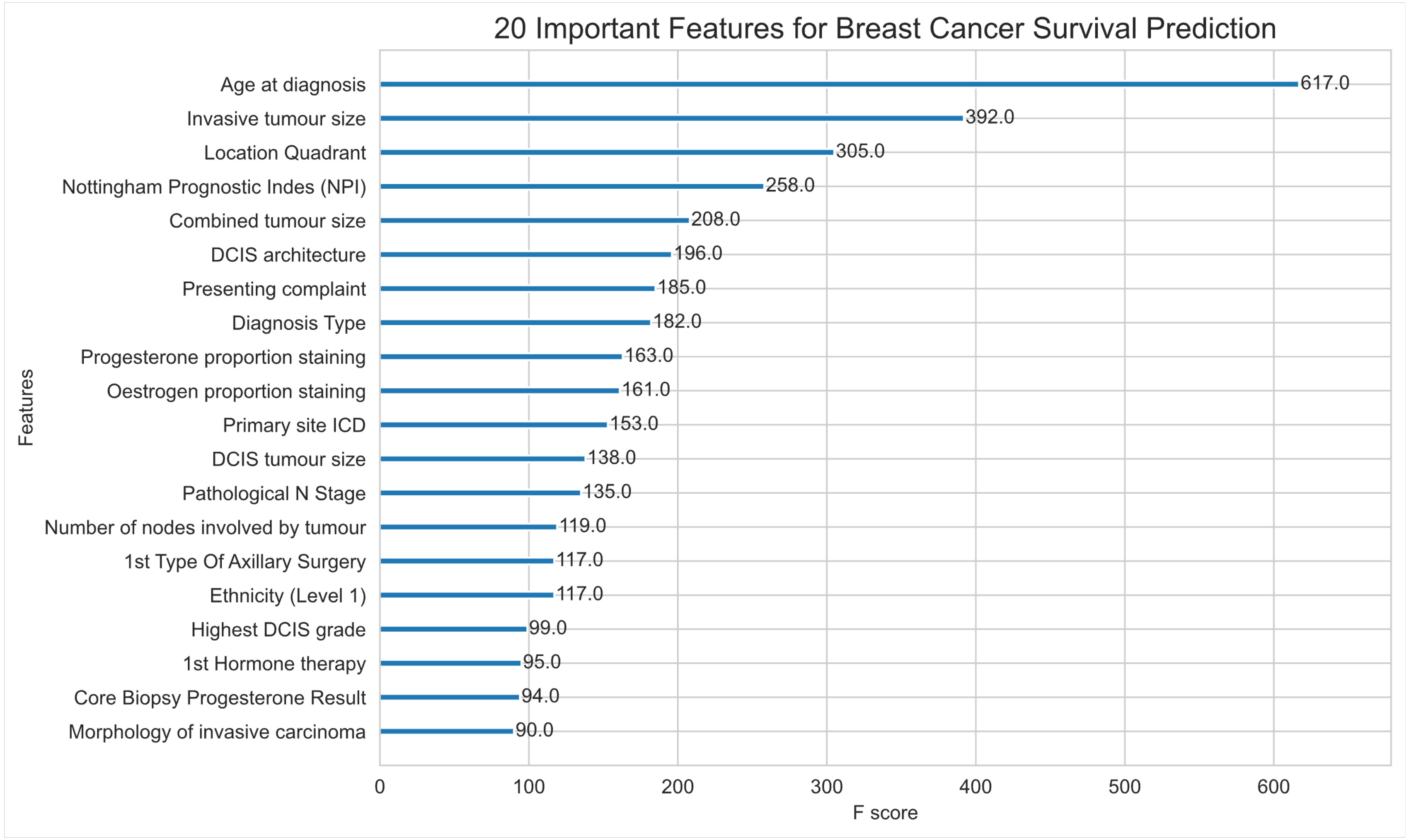
Feature Ranking Using XGBoost

- XGBoost aids in variable selection, to help select most significant predictors of survival in breast cancer patients. It has ability to integrate diverse clinical data types and analyze it to rate features by importance

- XGBoost is a robust modelling algorithm due to its ability to avoid overfitting and its ability to handle complex relationships in data, making it suitable for identifying important features in survival analysis

- Its tree-based structure allows it to capture nonlinear relationships between features and survival outcomes, which traditional linear models may struggle to capture non-linear relationships effectively- critical to understand cancer prognosis

Feature ranking of risk factors for 5-year survival using XG Boost



Results

The risk factors show temporal variation over time. This can be significant for accurate model development & later in treatment planning.

The Machine learning models are highly sensitive to data selected. Therefore, a population specific tool is critical for accurate prediction that includes relevant risk factors & representative population.

Conclusion

Cancer is a complex disease with multiple factors affecting its outcome. Understanding their role in the prognosis is complex and critical task, especially in the long term.

AI can help decipher these relationships by use of good data and team of multidisciplinary experts.

Clinical Advisors

Clinical Advisor:

Dr. Reuben Broom

Clinical Researcher & Oncologist

Dr. Vanessa Blair

Breast Surgeon & Researcher

Maori Integrity Advisor:

George Tongariro

Cultural Capacity & Capability Building for Indigenous Peoples (Māori & Pasifika)

Acknowledgements

We extend our gratitude to the Te Rēhita Mate Ūtaetae - Breast Cancer Foundation National Register for providing us with data funded by the New Zealand Breast Cancer Foundation NZ.

Special thanks to Jane Yates (Data Manager for the Northern Region) and the dedicated Data Managers at BCFNR for their ongoing support.

We sincerely appreciate the invaluable guidance from our Clinical Advisors.

Special thanks to George Tongariro for his input in our project, ensuring it aligns with principles of Maori and Pacifica sovereignty.

Thank you

Rooshan Ghous:
Email: 20231512@mywhitecliffe.com