이학석사 학위논문

A Novel Prognostic Immune Index for the Stratification of High-Risk Early Breast Cancer

초기 유방암의 고위험군 분류를 위한 새로운 예후 면역지표

2019 년 2 월

서울대학교 대학원

협동과정 생물정보학

A Novel Prognostic Immune Index for the Stratification of High-Risk Early Breast Cancer

초기 유방암의 고위험군 분류를 위한 새로운 예후 면역지표

지도교수 신 영 기

이 논문을 이학석사 학위논문으로 제출함

2019년 2월

서울대학교 대학원
협동과정 생물정보학과
이 한 나

이한나의 석사학위논문을 인준함
2019년 2월

위원장	박 태 성	(인)
부위원장	신 영 기	(인)
위원	이 미 옥	(인)

Abstract

A Novel Prognostic Immune Index for the Stratification of High-Risk Early Breast Cancer

Hannah Lee

College of Natural Sciences

Interdisciplinary Program in Bioinformatics

Seoul National University

Proliferation markers, given their strong connection to prognosis in breast cancer, have been implemented in multi-gene assays to predict patient prognosis. Despite the current use of breast cancer prognostic and predictive models, improvement of these models is crucial. In addition, a wider application of these assays targeting molecular subtypes other than ER+ type is needed. Unlike the deteriorating effects of proliferation genes, immune response-related genes have shown protective role as well as predictive prognostics capabilities in breast cancer, especially in the ER- subtypes. This study set out to unravel the effect of immune response

signatures and their contribution to recurrence and survival which may aid the existing prognostic and predictive breast cancer algorithms.

A prognostic model was developed from the 5 most significant immune response genes related to survival revealed by Lasso feature selection. With the addition of lymph node status as a variable, the risk model revealed prognostic significance in homogenized risk groups stratified according to Adjuvant! Online and proliferation markers. This novel immunogenetic algorithm demonstrated strong prognostic ability by reclassifying patients into significant prognostic risk groups within the homogenized risk groups. Additionally, its prognostic ability was revealed in all molecular subtypes of breast cancer, especially robust in HR- breast cancer subtypes. Since limitations exist in the widely used prognostic algorithms in breast cancer, the model constructed in this study can potentially assist or better yet, improve them by selecting patients pool with a greater likelihood for survival without the administration of adjuvant chemotherapy.

Keywords: breast cancer, immune response-related genes, proliferation genes, breast cancer molecular subtypes, prognostic model, recurrence and survival risk, adjuvant chemotherapy

Student Number: 2017-26462

Table of Contents

Table

Figures

1. Introduction

Proliferation and cell cycle regulation markers are widely used prognostic and predictive factors in early stage breast cancer patients that estimate the risk of tumor recurrence or death from cancer [1]. Decades of research and studies have proven the prognostic and predictive value of proliferation and cell cycle markers in breast cancer [2]. Furthermore, given their contribution to patient survival outcome, proliferation markers have been implemented in prognostic assays such as Oncotype DX, [3] MammaPrint, [4] PAM50 Prosigna, [5] and Endopredict [6]. These gene-expression based algorithms predict the risk of recurrence or distant metastasis after breast cancer surgery. In addition, they predict patient response to adjuvant chemotherapy by monitoring markers of cell proliferation, hormone receptors (HR), human epidermal growth factor receptor 2 (HER2), and basal cytokeratin [7].

A recent update has been made to the American Joint Committee of Cancer (AJCC) to incorporate biologic factors and gene-expression prognostic panels in classifying tumors, lymph nodes, and metastasis (TNM) staging system [8]. However, there exists multiple limitations in these prognostics. First of all, they target node-negative, estrogen receptor (ER) positive early breast cancer patients [9]. This excludes HR- subtypes, HER2+ and Triple Negative Breast Cancer (TNBC) as well as progressed stages of breast cancer. Moreover, recent reports have claimed discordance between PAM50 expression assay and Oncotype DX in risk

classifying patients [10]. Studies have shown that PAM50 better predicts the risk of distance recurrence in ER-positive, node-negative patients compared to Oncotype DX [11]. Given the current situation, improvements to the existing prognostics is necessary for more accurate prediction of patient survival outcome and response to adjuvant chemotherapy. Additionally, a wider application of these multi-gene based prognostic tests targeting all breast cancer subtypes will be favorable.

While focus has been mainly on proliferation and cell cycle signatures as prognostic markers for breast cancer and implementing these markers into breast cancer prognostics, immune genes in certain subtypes of breast cancer have shown predictive prognostics capabilities [12–20]. Studies have repeatedly reported that the presence of immune genes in breast cancer patients impacts survival outcome favorably, especially in HER2+ and TNBC subtypes (HR- subtypes) [12–19]. A previous study by a research group confirmed that immune response prognosis indicated a beneficial defense mechanism in breast cancer in contrast to the damaging effect of proliferation genes in lymph node-negative breast cancer [16]. This study by Oh et al. revealed that with the increase in proliferation activity, the activity of immune response also increases in both ER+ and ER- subtypes.

Furthermore, a study by Yang et al., revealed through single cell RNA-seq analysis that increase in immune response may be the contribution of tumor infiltrating lymphocytes (TILs, hereafter), and not breast carcinomas [7]. Another study by Schmidt et al. stated that immune cell infiltration is a major contributor in

prohibiting tumor progression in breast carcinomas, irrespective of ER status [17]. TILs are known to infiltrate carcinomas and are associated with preventing development and progression of malignant tumors. Out of all breast cancer subtypes TNBC has manifested to be the most immunogenic, with highest TIL density associated with better prognosis, followed by HER2+ subtype [12, 13]. Prognostic significance of the immune response gene (*CD2*) has been expanded in a previous study as well, specifically in the HR-/HER2+ breast cancer [15]. In addition, correlation between immune infiltrates and response to chemotherapy post-treatment has been reported specifically in HR- breast cancer subtypes [20]. Consequently, recent research has focused largely on HR- breast cancer subtype to unravel the clinical role of immune biomarkers in breast cancer. Furthermore, Teschendorff et al., succeeded in finding association between immune response signatures in HR- breast cancer that consists of 7 gene immune response-module (*CIQA, IGL2, LY9, TNFRSF17, SPP1, XCL2*, and *HLA-F*) which highlighted the prognostic significance of immune genes [15,21].

In contrast to the HR- subtypes, overall TIL levels are lower in HR+ subtypes [22]. As a matter of fact, some reports have shown that TIL in HR+ and HR- have a negative prognostic effect in breast cancer patients [23]. Nevertheless, few immune cell types upregulated in HR+ appear to be associated with improved survival outcome, such NK cells, T and B cells [24]. In general, elevation of immune

response signatures is associated with reduced recurrence risk and prolonged survival in both HR- and HR+ populations [15].

Despite the prognostic ability of certain immune related markers, they have yet to be incorporated into prognostic assays. Because proliferative and cell cycle genes are often the most powerful factors concerning breast cancer progression, especially in the HR+ subtype, hypothesis was made that the total effect of immune response genes may be masked by these markers. Therefore, this study aims to unravel the prognostic effects of immune response signature in all breast cancer subtypes by initially stratifying patients into groups based on conventional clinicopathological factors and proliferation marker-based classification. Consequently, deeper insights on immune microenvironment and their contribution to patient survival can be integrated to breast cancer prognostics as a way to improve their accuracy in prediction. Thus, the need for bridging the gap between breast cancer and immune system is paramount.

2. Materials and methods

2.1. Data Selection

Acquiring appropriate and valid data relevant to this study was crucial. First public database, National Center for Biotechnology Information Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo), was searched thoroughly for the acquisition of five different breast cancer datasets for discovery analysis. Despite abundant breast cancer cohorts archived in GEO, datasets use in this study were strictly selected under the following criteria: 1). ER status or molecular subtype identified in the clinical data, 2). Patients did not receive chemotherapy, 3). Datasets had been probed with the Affymetrix platform, either [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array or [HG-U133A] Affymetrix Human Genome U133A Array, 4). Dataset contained survival information of some sorts, either disease free survival/distant-metastasis free survival (DFS/DMFS) or more preferable endpoint, overall survival (OS), 5). Datasets contained clinical information of lymph node status, tumor size, patient age and histological grade. Discovery set was selected out of GSE6532, GSE7390, GSE1121, GSE31519 and GSE4922 cohorts, all probe with the same platform, Affymetrix GPL96. In total, 967 patient samples were analyzed in this study. For further validation by a different cohorts and platforms, two micro-array datasets (GPL96 and GPL570) and METABRIC gene expression profiles were analyzed with the same criteria applied to the discovery cohorts. Data was accessed and

downloaded through cBioportal website (http://www.cbioportal.org/index.do) and were $\log_2$ normalized prior to analysis. Table 1 shows summary of conventional clinicopathological characteristics of all patients, organized by molecular subtypes and cohorts. Overall, summary of datasets and platforms used in the discovery and validation sets is displayed in Fig. 1.

Table1 Summary of clinical characteristics organized by molecular subtypes and cohorts of the discovery set

	Total (n= 967) No. (%)	**HR+/HER2-** (n= 619) No. (%)	**HR+/HER2+** (n= 99) No. (%)	**HR-/HER2+** (n= 47) No. (%)	**HR-/HER2-** (n= 202) No. (%)	
Age (years)						
<50	316 (32.7)	160 (25.8)	42 (42.4)	17 (36.2)	97 (48.0)	
≥50	651 (67.3)	459 (74.2)	57 (57.6)	30 (63.8)	105 (25.0)	
Tumor size (cm)						
≤2	522 (54.0)	321 (51.9)	48 (48.5)	14 (29.8)	139 (68.8)	
2~5	432 (44.7)	289 (46.7)	48 (48.5)	33 (70.2)	62 (30.7)	
>5	13 (1.3)	9 (1.4)	3 (3)	0 (0.0)	1 (0.5)	
Lymph node status						
Negative	736 (76.1)	457 (73.8)	75 (75.8)	30 (63.8)	174 (86.1)	
Positive	181 (18.7)	136 (22.0)	19 (19.2)	11 (23.4)	15 (7.4)	
NA	50 (5.2)	26 (4.2)	5 (5)	6 (12.8)	13 (6.4)	
Histologic grade						
1	289 (29.9)	164 (26.5)	70 (70.7)	14 (29.8)	41 (20.3)	
2	378 (39.1)	353 (57.0)	0 (0.0)	0 (0.0)	25 (12.4)	
3	299 (30.9)	102 (16.5)	28 (28.3)	33 (70.2)	136 (67.3)	
NA	1 (0.1)	0 (0.0)	1 (1)	0 (0.0)	0 (0.0)	
	Total (n= 967) No. (%)	**GSE6532** (n= 256) No.(%)	**GSE7390** (n= 161) No.(%)	**GSE11121** (n= 200) No.(%)	**GSE31519** (n=105) No.(%)	**GSE4922** (n=245) No.(%)
Age (years)						
<50	316 (32.7)	61 (23.8)	109 (67.7)	47 (23.5)	49 (46.7)	50 (20.4)
≥50	651 (67.3)	195 (76.2)	52 (32.3)	153 (76.5)	56 (53.3)	195 (79.6)
Tumor size (cm)						
≤2	522 (54.0)	115 (44.9)	66 (41.0)	112 (56)	105 (100)	124 (50.6)
2~5	432 (44.7)	137 (53.5)	95 (59.0)	85 (42.5)	0 (0.0)	115 (46.9)
>5	13 (1.3)	4 (1.6)	0 (0.0)	3 (1.5)	0 (0.0)	6 (2.4)
Lymph node status						
Negative	736 (76.1)	176 (68.8)	103 (64.0)	200 (100)	100 (95.2)	157 (64.1)
Positive	181 (18.7)	77 (30)	18 (11.2)	0 (0.0)	5 (4.8)	81 (33.1)
NA	50 (5.1)	3 (1.2)	40 (24.8)	0 (0.0)	0 (0.0)	7 (2.8)
Histologic grade						
1	289 (29.9)	88 (34.4)	27 (16.8)	55 (27.5)	35 (33.3)	84 (34.3)
2	378 (39.1)	109 (42.6)	53 (32.9)	110 (55)	0 (0.0)	106 (43.3)
3	299 (30.9)	58 (22.6)	81 (50.3)	35 (17.5)	70 (66.7)	55 (22.4)
NA	1 (0.1)	1 (0.4)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)

Discovery
Validation

Platform	Set Type	GEO Number	Year	Total	Number of Event	Survival Type	Treatment
GPL96	Discovery data set	GSE31519	2011	67	23	DFS	Some patients treated with adjuvant therapy
		GSE4922	2006	578	89	DFS	Some endocrine therapy, some chemotherapy
		GSE11121	2008	200	46	DMFS	No adjuvant therapy
		GSE6532	2006	284	62	DMFS	Tamoxifen only
		GSE7390	2006	198	62	DMFS	No adjuvant therapy
GPL570	Validation data set	GSE21653	2010	266	70	DFS	No adjuvant therapy
		GSE42568	2013	104	49, 36	DFS, OS	Some endocrine therapy, some chemotherapy
GPL96		GSE3494	2005	251	55	OS	Some endocrine therapy, some chemotherapy
Multi-Platform based	Validation METABRIC	--	2012	2509	1143	OS	Some endocrine therapy, some chemotherapy

Figure 1 Cohorts in the discovery and validation sets

Summary of cohorts used in the discovery and validation sets organized according to GEO number, year of data availability, total number of patient samples in the dataset, survival and treatment types.

2.2. Data Mining and preparation

Information on ER IHC was acquired to identify breast cancer molecular subtypes by HR status. This information alone was insufficient to divide the breast cancer types into four groups of HR+/HER2- (ER+ or PR+/HER2-), HR+/HER2+ (ER+ or PR+/HER2+), HR-/HER2+ (ER-/PR-/HER2+), or TNBC (ER-/PR-/HER2-). If a dataset did not contain information on ER IHC, PR and HER2 status, the expression levels of the corresponding genes of each marker were proportioned accordingly as indicated in Cheang et al. [25]. The downloaded datasets were $\log_2$ normalized in advance to analysis. Next to reduce bias in the discovery set, only genes that exceeded a threshold value using interquartile range were selected. In addition, to reduce non-biological variations present in the selected datasets, all the while retaining biological differences, batch effect correction was performed on the discovery and validation sets composed of multiple cohorts using ComBat algorithm. Validation of batch effect correction is shown by principal component analysis (Fig. 2). ComBat corrects technical batch effects by applying empirical Bayesian framework to the designated batches [26]. After adjustment and normalization was conducted, each of the molecular subtypes were stratified into four risk subcategories by clinical and gene risk classification schemes.

Figure 2 Batch effect correction by ComBat algorithm

The first two principal components of batch effect correction by ComBat algorithm. The left-hand panel is PCA plot of the five different cohorts prior to the batch effect removal. The right-hand panel displays PCA plot after the batch effect has been removed.

2.3. Survival analysis: Cox regression and Kaplan-Meier

Cox proportional hazard regression is a semi-parametric regression used to model association between patient survival time and predictor variables by using the hazard function:

$$\lambda(t|X) = \lambda(t)\exp(X\beta)$$

Cox proportional hazard regression is semi-parametric because this model makes no prior assumption of the underlying shapes for both hazards, λ(t), and survival functions, S(t). The most favorable endpoint in survival analysis is overall survival (OS), but due to time limitations, this information may not always be available. Consequently, surrogate endpoints such as DFS and DMFS are used in place of OS. In this study, OS and DFS/DMFS were selected as clinical endpoints. Both univariate and multivariate analyses of clinical and genetic variables was carried out by Cox proportional hazard regression. Multivariate analysis identified independent contribution of predictor variables. In addition, the Kaplan-Meier method and log-rank test were used to graph survival outcome and show survival differences between the groups, respectively. Log rank p-values of < 0.05 were assumed to be statistically significant.

2.4. Gene ontology and pathway analysis

Gene annotation and pathway analysis was conducted by breast cancer subtype. Annotation of gene pathway consisted of two parts. First, the top most significant genes with adjusted p-value of less than 0.01 from Cox regression were annotated in DAVID. For further analysis, gene annotation package topGO in R was used to annotate the pathways of the most significant genes from the regression analysis. topGO applies two types of statistics to find most significant pathways, Fisher's exact test and Kolmogorov-Smirnov test, to compute enrichment scores. In addition, two types of algorithms, classic and elim method, can be applied to each statistic. For the purpose of this study, two algorithms were applied to Kolmogorov-Smirnov test and in addition, classic Fisher were used to find most significant gene annotations. Prior to dividing datasets for analysis, three breast cancer subtypes were chosen out of the four common breast cancer subtypes for pathway analysis. Because survival outcomes were not statistically different between HR-HER+ and HR-HER- (data not shown), they were grouped into a larger subunit, making a total of three groups: HR-, HR+/HER+, HR+/HER-. Thus, for the purpose of gene annotation and pathway analysis, the HR- subtypes were grouped together.

2.5. Gene selection and risk stratification

Initially, Cox proportional hazard regression was applied to all the subtypes to find the strongest (most significant) genes related to proliferation of cells. A total of 37 proliferative genes and 110 immune response genes were selected with the following criteria: 1). High variance, 2). Gene ontology analysis with significant result, and 3). previous mention and use in literature.

For all the patient samples, the expression level of each proliferation/cell-cycle regulation gene was classified as binary category, “high” or “low”, according to the averaged expression of the gene. If the expression levels of five or more of the ten selected genes were classified as “low”, then the patient was classified into proliferation low-risk group and otherwise into proliferation-high risk group.

In addition to classification of each sample according to the gene-risk stratification, modified version of Adjuvant! Online was applied to classify patients into clinically high and low-risks group, as show in in Fig. 3 [29]. As shown in the figure, which displays clinical risk assessment of each of the four molecular subtypes of breast cancer, groups were classified based on histological grade, nodal status, and tumor size. Gene risk stratification based on proliferation/cell-cycle related genes combined with Adjuvant! Online clinical risk classification stratified patients into four risk subcategories: 1). Clinically high-risk/proliferation high-risk, 2). Clinically high-risk/proliferation low-risk, 3). Clinically low-risk/proliferation

high-risk, and 4). Clinically low-risk/proliferation low-risk. HR+/HER2- group was divided to three risks groups with clinical high- proliferation high group as All high-risk group (hereon, AH) and high-low and low-high clinical and proliferation risks as All intermediate-group (hereon, AI), and clinically low- proliferation low group as All low-risk group (hereon, AL). However, dividing each subtype into the four risk categories produced insufficient sample number in each risk subcategories and consequently, each risk subcategories within molecular subtypes were combined based on sample size and Cox regression outcomes (Fig. 4). Subsequently, HR+/HER2+ was finalized into two risk groups with clinically high and proliferation high-group as AH group and the rest as AL group. All but three samples in the HR-/HER2+ subtype initially classified into AH group and thus no further regrouping was necessary. Finally, TNBC subtype was regrouped into AH group containing clinically high- proliferation high, clinically high- proliferation low, and clinically low- proliferation high group and AL group contained clinical low and proliferation low-risk groups. Patient samples with missing clinical information were excluded from this study. The overall schematics of this study is displayed in Fig. 5. Furthermore, validation cohorts were classified according to clinical and gene risk scores as the discovery set prior to survival analysis. To maintain homogeneity of clinical and gene risks as conducted in the discovery set, only the samples that were classified as high-risk by both clinical and gene risk was selected.

ER status	HER2 status	Grade	Nodal status	Tumor Size	Clinical Risk
Positive	HER2 negative	well differentiated	N-	≤ 3 cm	C-low
				3.1-5 cm	C-high
			1-3 positive nodes	≤ 2 cm	C-low
				2.1-5 cm	C-high
		moderately differentiated	N-	≤ 2 cm	C-low
				2.1-5 cm	C-high
			1-3 positive nodes	Any size	C-high
		poorly differentiated OR undifferentiated	N-	≤ 2 cm	C-low
				2.1-5 cm	C-high
			1-3 positive nodes	Any size	C-high
	HER2 positive	well OR moderately differentiated	N-	≤ 1 cm	C-low
				1.1-5 cm	C-high
			1-3 positive nodes	Any size	C-high
		poorly OR undifferentiated	N-	≤ 2 cm	C-low
				2.1-5 cm	C-high
			1-3 positive nodes	Any size	C-high
Negative	HER2 negative	well differentiated	N-	≤ 2 cm	C-low
				2.1-5 cm	C-high
			1-3 positive nodes	Any size	C-high
		moderately OR poorly differentiated OR undifferentiated	N-	≤ 1 cm	C-low
				1.1-5 cm	C-high
			1-3 positive nodes	Any size	C-high
	HER2 positive	well OR moderately differentiated	N-	≤ 1 cm	C-low
				1.1-5 cm	C-high
			1-3 positive nodes	Any size	C-high
		poorly OR undifferentiated	Any	Any size	C-high

Figure 3 Classification scheme by modified version of Adjuvant! Online
Classification algorithm used to classify patients clinically into low and high-risk groups.

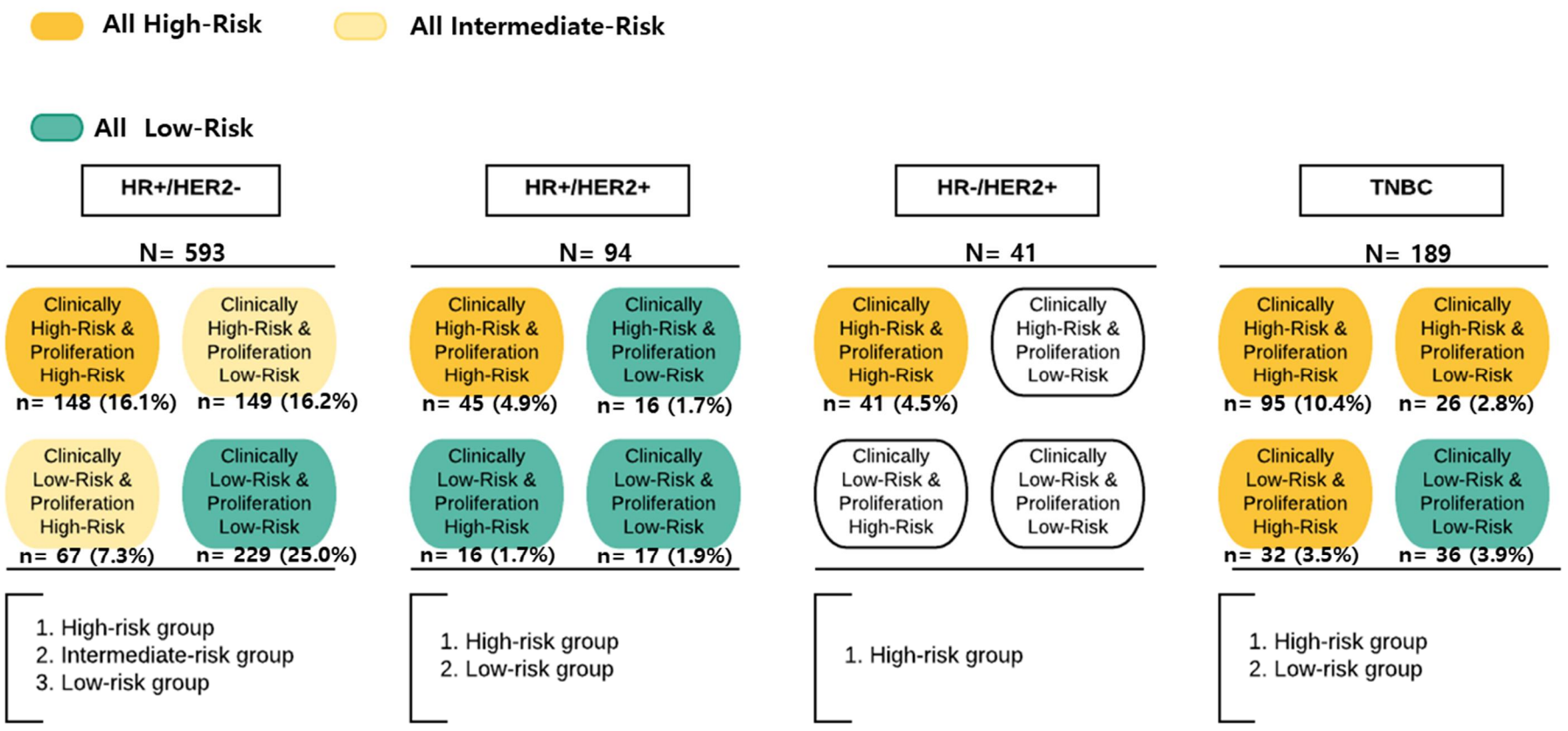


Figure 4 Homogenized risk groups by clinical and proliferation risks

Four different risk group based on both clinical (Adjuvant! Online) and proliferation gene risk criteria within each molecular subtype. Figure displays the final homogenized groups of the four different risk groups in each subtype based on Cox regression outcome and survival rates.

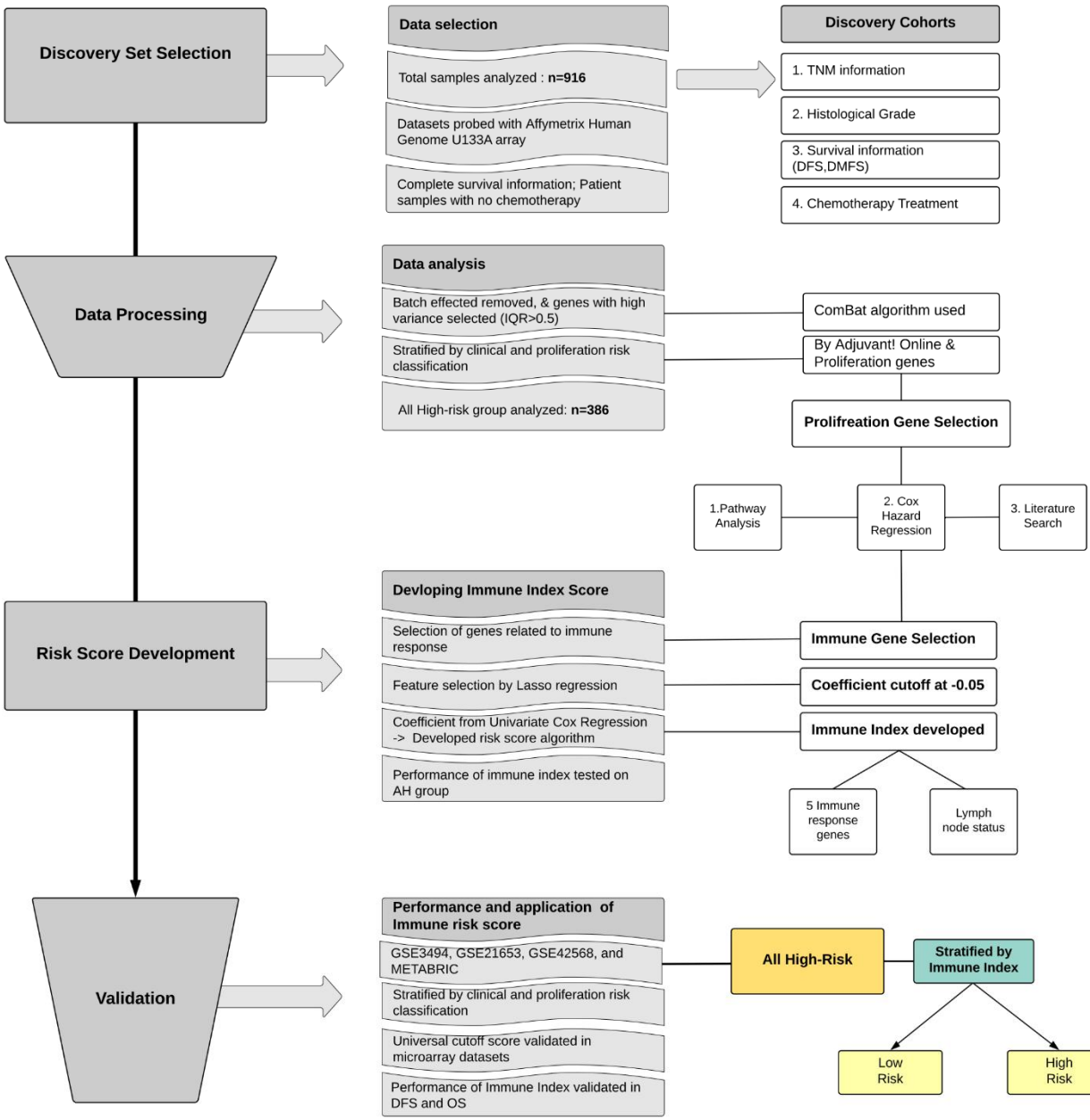


Figure 5 Overall workflow of the study

This study starts with data selection/mining and ends by validation of the developed algorithm. Criteria for dataset selection, gene selection and risk-group stratification are included in the workflow.

2.6. Prognostic model development by Lasso regression

The initial step of model selection began by identifying significant immune genes related to survival for each breast cancer subtype. First, Lasso feature selection was used to select most significant genes related to DFS/DMFS by applying 'coxnet' package in R to find optimal lambda value by 10,000-fold cross-validation and subsequently, found active covariate in the model. Lasso regression performs regularization and feature selection by penalizing coefficients of the variables in regression using optimal λ as a tuning parameter. Thus, Lasso minimizes the sum of squared errors bound by:

$$\text{Min } \sum_i (y_i - \sum_i x_{ij}\beta_j)^2 \text{ subject to } \sum_i |\beta_i| \leq s$$

where s is the upper bound for the sum of the coefficients and λ takes a reverse relationship compared to s, and as λ increases, so does the amount of shrinkage. Fig. 6 shows cross validated error plot, displaying the optimal lambda value. Left vertical line indicates minimum value reached by CV-error curve and the right vertical line indicates CV-error of regularized model within 1standard deviation from the minimum point. After Lasso regression was performed and most significant genes were selected, the result was validated by Cox proportional hazard univariate analysis.

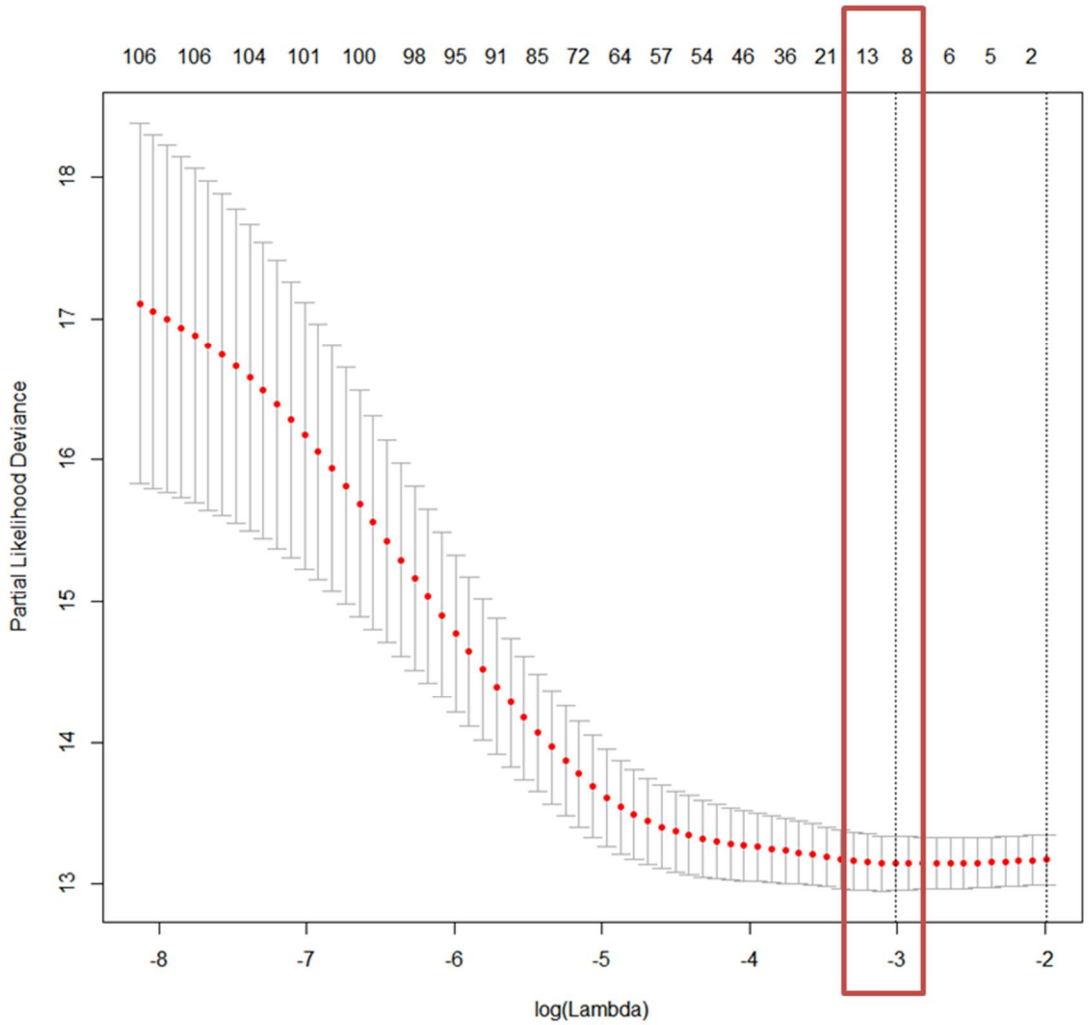


Figure 6 Cross validated error plot

Cross validated error plot showing the optimal lambda value. Left-vertical line indicates minimum value reached by CV-error curve and the right-vertical line indicates CV-error of regularized model within 1 standard deviation from the minimum point.

2.7. Assessment and validation of the prognostic model

Optimal cutoff point of the risk score was acquired by bootstrapping maximally selected statistics using the 'survminer' package. Fig. 7 shows two cutoff points identified by bootstrapping method. For validation and assessment of the risk model, concordance index was calculated from 'survcomp' package in R. The concordance index is a standard measurement which assess the performance of a predictive model in survival analysis. All computational calculations were conducted in R version 3.4.3.

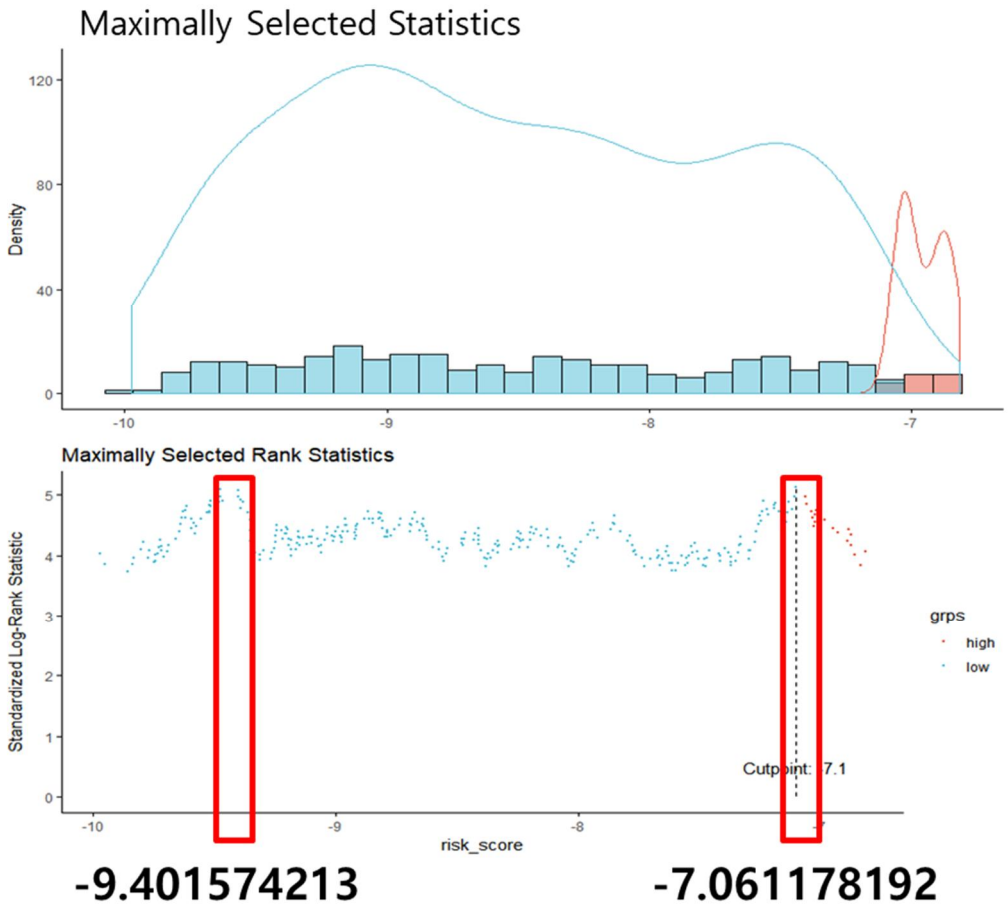


Figure 7 Optimal cutoff points by maximally selected statistics

Two optimal cutoff points determined by bootstrapping maximally selected statistics.

3. Results

3.1. Stratification of patient samples into risk groups

Pathway analysis results revealed that most of the genes significantly contributing to survival in the HR+ types were related to cell proliferation and cell cycle regulation while HR- types were related to locomotion and immune response (Table 2). Based on the pathway analysis, total of 37 genes related to proliferation and significantly contributing to survival outcome in the HR+ groups analyzed by Cox proportional hazard regression to find gene predictors with significant association with DFS/DMFS. Total of 10 genes were chosen as proliferation/cell-cycle regulation genes for gene risk classification. Multivariate Cox regression analysis is displayed by a forest plot in Fig. 8.

Log-rank p-values of the risk groups in HR+/HER2- and HR+/HER2+ subtypes classified by clinical and proliferation risk criteria in were $p < 0.0001$ each, and $p = 0.0018$ for TNBC subtype (Fig. 9). Since HR-/HER2+ only consisted of AH group, and thus no survival curve was estimated. Cox regression analysis according to different risk groups showed that in HR+/HER2- subtype, the AI group and AL group had hazard ratio of 0.613 ($p=0.003$, 95% CI:0.444-0.847) and 0.217($p<0.0001$, 95% CI: 0.145-0.327), respectively compared to the AH group. Similar observations were shown in the HR+/HER2+ and TNBC subtypes with hazard ratio of 0.255 ($p < 0.0001$, 95% CI: 0.134-0.483) and 0.377 ($p = 0.0182$, 95% CI: 0.162-0.873-0.0.8426) respectively, in the AL group compared to AH group.

Table 2 Gene annotation results of the top most significant biological pathways associated with each subtype

HR+/HER2-

GO.ID	Term	Annotated	Significant	Expected	Rank in classicKS	classicKS	elimKS
GO:0030154	cell differentiation	54	54	54	175	0.77513	0.016
GO:0050793	regulation of developmental process	40	40	40	107	0.32526	0.023
GO:0048869	cellular developmental process	55	55	55	174	0.75265	0.028
GO:2000026	regulation of multicellular organismal d...	35	35	35	113	0.33666	0.043
GO:0033554	cellular response to stress	55	55	55	108	0.32814	0.044
GO:0051173	positive regulation of nitrogen compound...	57	57	57	36	0.02073	0.055
GO:0010604	positive regulation of macromolecule met...	60	60	60	42	0.03029	0.093
GO:0031325	positive regulation of cellular metaboli...	54	54	54	49	0.03221	0.098
GO:0042981	regulation of apoptotic process	31	31	31	168	0.71457	0.098
GO:0006366	transcription from RNA polymerase II pro..	38	38	38	95	0.24381	0.115
GO:0043067	regulation of programmed cell death	31	31	31	169	0.71457	0.143
GO:0045935	positive regulation of nucleobase-contai...	31	31	31	124	0.38235	0.151

HR+/HER2+

GO.ID	Term	Annotated	Significant	Expected	Rank in classicKS	classicKS	elimKS
GO:0006260	DNA replication	38	38	38	7	3.50E-05	3.50E-05
GO:0044772	mitotic cell cycle phase transition	70	70	70	4	7.90E-06	0.00071
GO:0000070	mitotic sister chromatid segregation	30	30	30	10	0.0011	0.00111
GO:0051301	cell division	71	71	71	12	0.0015	0.00152
GO:0000082	G1/S transition of mitotic cell cycle	34	34	34	15	0.0021	0.00215
GO:0007346	regulation of mitotic cell cycle	63	63	63	19	0.0029	0.00287
GO:1901987	regulation of cell cycle phase transitio...	49	49	49	22	0.008	0.00805
GO:0090068	positive regulation of cell cycle proces...	33	33	33	23	0.0081	0.00807
GO:1903047	mitotic cell cycle process	110	110	110	2	4.30E-07	0.00858
GO:0006974	cellular response to DNA damage stimulus	71	71	71	25	0.011	0.01104
GO:1901990	regulation of mitotic cell cycle phase t...	48	48	48	26	0.0112	0.01123
GO:0006281	DNA repair	50	50	50	30	0.0151	0.01511

HR-

GO.ID	Term	Annotated	Significant	Expected	Rank in classicKS	classicKS	classicFisher
GO:0000902	cell morphogenesis	32	32	32	172	0.472	1
GO:0001525	angiogenesis	30	30	30	22	0.035	1
GO:0001568	blood vessel development	38	38	38	5	0.011	1
GO:0001775	cell activation	61	61	61	226	0.617	1
GO:0001816	cytokine production	36	36	36	302	0.899	1
GO:0001817	regulation of cytokine production	31	31	31	314	0.953	1
GO:0001932	regulation of protein phosphorylation	47	47	47	158	0.426	1
GO:0001934	positive regulation of protein phosphory..	30	30	30	124	0.351	1
GO:0001944	vasculature development	38	38	38	6	0.011	1
GO:0002250	adaptive immune response	30	30	30	237	0.647	1
GO:0002252	immune effector process	55	55	55	288	0.845	1
GO:0002376	immune system process	119	119	119	277	0.797	1

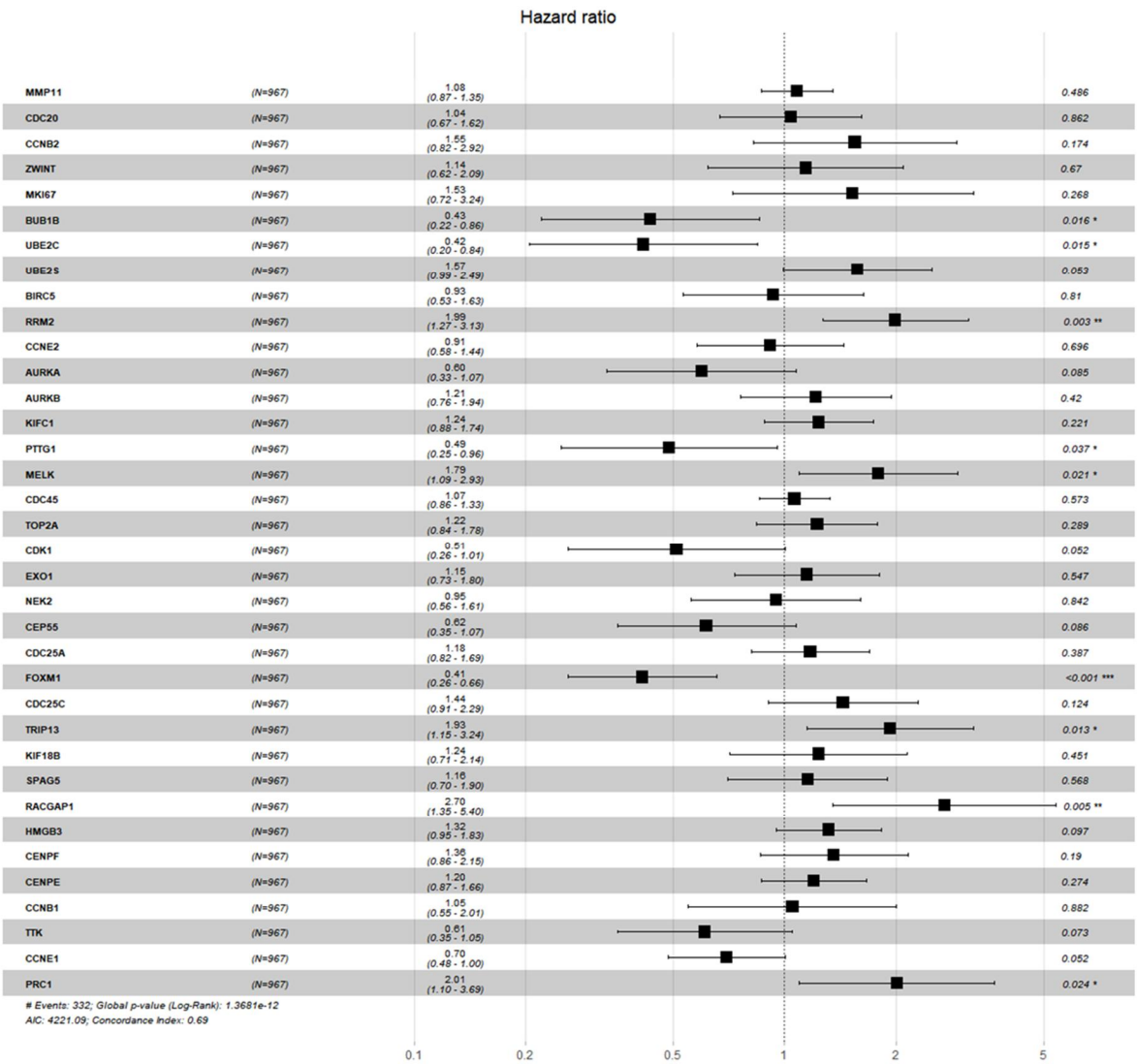
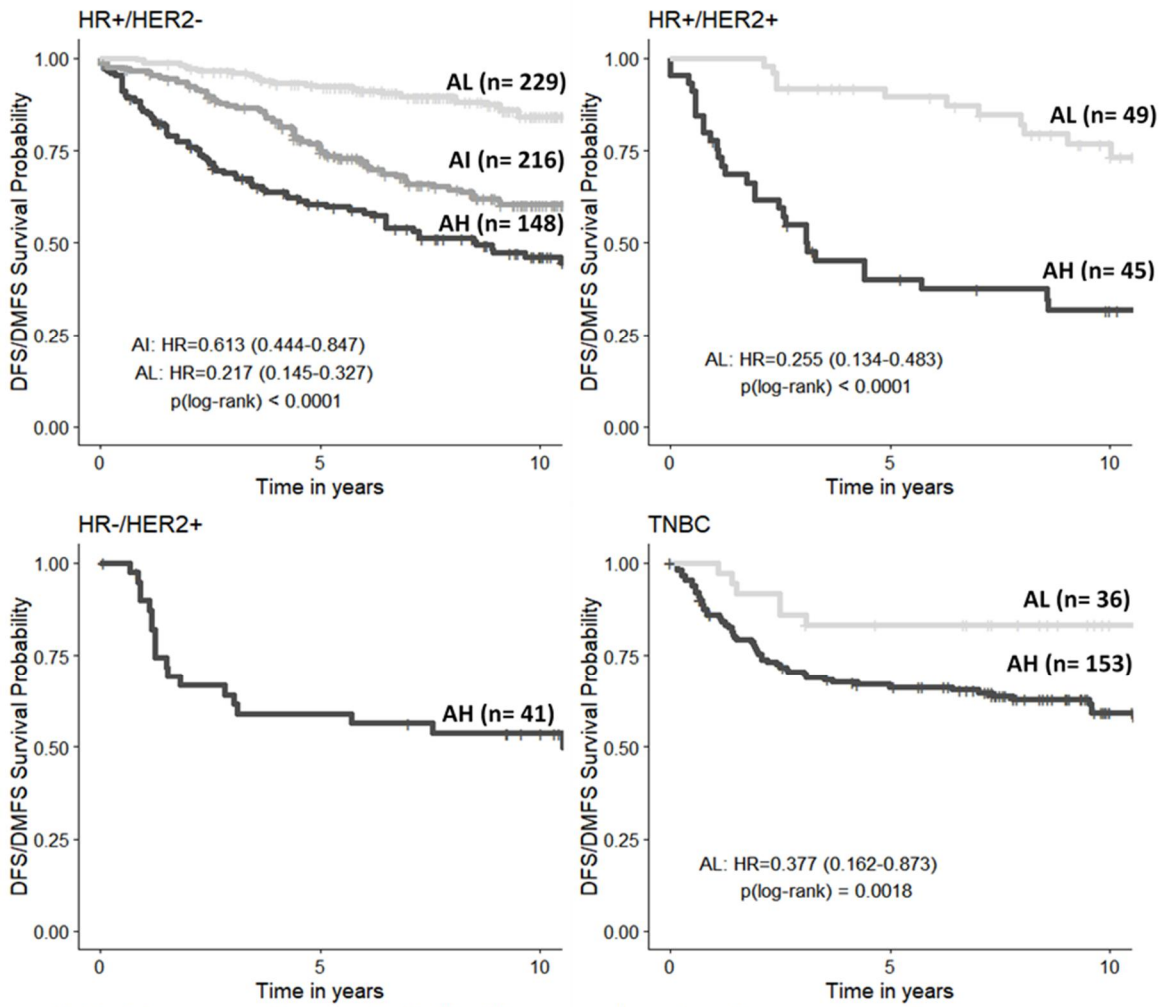


Figure 8 Multivariate Cox regression forest plot of 37 proliferative genes

Multivariate Cox regression forest plot of the 37 proliferative genes. Hazard ratio and 95% CI is on the horizontal axis. P-values are shown on the far right with statistically significant genes indicated by asterisk.



3.2. Prognostic model finding and selection using immune response genes

Prognostic value of the immune response genes on each of the subgroup

Figure 9 Survival curve of AI, AL and AH risk groups

Kaplan-Meier plots of DFS/DMFS differences by log-rank test of the finalized risk groups. Starting from clock-wise HR+/HER2-, HR+/HER2+, TNBC and HR-/HER2+. The p-values shown are based on log-rank test. HR+/HER2- showing total of 3 risk groups and HR+/HER2+ and TNBC both showing two final risk groups. HR-/HER2+ KM-graph is displaying one risk group (no p-value).

was analyzed within each subtype. Overall, 110 genes were selected from immune models in Rody et al, [27] associated with MHC-1,2, T-cell, and B-cells, and from Wolf et al., [28] associated with immune response. Prognostic gene-selection began with univariate analysis of the identified 110 immune response related genes and by observing their significance in each of molecular subtype (Table 3). All the AH groups in each subtype had elevated expression of the immune response genes significantly associated with positive prognostic outcomes. As shown in Table 3, univariate analysis revealed that HR+/HER2- group contained 55 immune response genes with significant p-value ($p<0.05$) and all had negative coefficient value which indicates their positive association with prolonged survival. In a similar manner, the AH group in HR+/HER2+, HR-/HER2+ and TNBC subtypes each had 96, 30 and 8 immune response genes with significant p-value ($p<0.05$), all with negative coefficients.

In contrast to the results shown in the AH groups, the effects of immune genes in the AI and AL groups were less prominent. HR+/HER2- AI group revealed not one significant immune response gene associated with survival, with the lowest p-value of all genes being greater than 0.09. AL group contained few significant genes, but their hazard ratios were not associated with positive DFS/DMFS outcomes. Based on the Cox regression outcomes, focus was made on the AH groups to further investigate and select prognostic immune response genes.

Lasso regression and Cox regression analysis was used to develop a prediction model. First, AH groups in each of the subtypes were combined. Patient data with missing or ambiguous clinical information were excluded in this analysis were analyzed for model selection and development. Lasso chose nine active genes (*CTLA4, CTSW, DOCK10, GPR18, IGHM, IL21R, IL2RB, TNFRSF9*, and *TRAT1*) with negative effect on hazard (Table 4), and Cox regression discovered five genes (*TRAT1, IGHM, IL21R, GZMB, GPR18*) with significant effect on hazard ($p < 0.0001$) (Table 5). Five genes (*TRAT1, IL21R, IGHM, CTLA4, IL2RB*) that exceeded negative coefficient value of -0.05 from lasso regression results and that have p-value less than 0.001 from the Cox regression results were selected. Furthermore, univariate and multivariate analysis on the clinical variables revealed lymph node status to have highest significance affecting survival all the while being an independent prognostic factor (data not shown). Thus, lymph node status was accounted for in the risk model. Based on the Cox regression coefficient values of the five selected genes, immune index was calculated according to the following formula.

$$\textbf{Immune Index} = \frac{\sum(\boldsymbol{Cox\ regression\ coefficeint\ of\ genes} * \boldsymbol{gene\ expression\ levels})}{\mathbf{2} * (\textbf{lymph node status})}$$

Table 3 Univariate results of 110 immune response genes for each molecular subtype

HR+/HER2-	coef	hr	se(coef)	z	pvalue
CD69	-0.68986	0.501646	0.211284	-3.26509	0.001094
CD55	-1.02617	0.358375	0.323082	-3.1762	0.001492
TRAF3IP3	-0.93552	0.392382	0.311055	-3.00757	0.002633
EVI2B	-0.91907	0.398891	0.30849	-2.97925	0.00289
IL21R	-0.80724	0.44609	0.273252	-2.95418	0.003135
IGHM	-0.42671	0.652655	0.148793	-2.86779	0.004134
IGJ	-0.38369	0.681341	0.135338	-2.83506	0.004582
CR2	-0.46309	0.629334	0.164106	-2.82192	0.004774
GZMB	-0.553	0.575221	0.20029	-2.761	0.005762
STAP1	-0.59319	0.552562	0.218993	-2.70871	0.006755

HR+/HER2+	coef	hr	se(coef)	z	pvalue
KLRB1	-1.52862	0.216835	0.325476	-4.69656	2.65E-06
PRKCB	-2.19827	0.110995	0.468355	-4.6936	2.68E-06
CD37	-1.93956	0.143767	0.413562	-4.68989	2.73E-06
GPR171	-2.04521	0.129353	0.436535	-4.6851	2.80E-06
CD3D	-1.92875	0.14533	0.417865	-4.61572	3.92E-06
PPP1R16B	-1.67262	0.187754	0.365964	-4.57047	4.87E-06
ITK	-2.57119	0.076444	0.564613	-4.5539	5.27E-06
SH2D1A	-2.41604	0.089274	0.531415	-4.54643	5.46E-06
TNFRSF1B	-3.84986	0.021283	0.847878	-4.54058	5.61E-06
CD48	-2.60977	0.073551	0.582633	-4.47927	7.49E-06

HR-/HER2+	coef	hr	se(coef)	z	pvalue
BANK1	-0.94469	0.3888	0.261254	-3.61599	0.000299
GIMAP6	-2.0162	0.13316	0.644425	-3.12868	0.001756
CD69	-1.61401	0.199087	0.542267	-2.97642	0.002916
GPR18	-0.94085	0.390294	0.335457	-2.80469	0.005036
LY9	-0.99188	0.370877	0.353694	-2.80436	0.005042
VNN2	-1.05607	0.347819	0.388099	-2.72115	0.006506
TCL1A	-2.0106	0.133909	0.742011	-2.70966	0.006735
CYTIP	-1.61108	0.199672	0.606694	-2.6555	0.007919
CTSW	-0.89989	0.406614	0.351074	-2.56325	0.01037
PTPRC	-1.2287	0.292673	0.483899	-2.53916	0.011112

TNBC	coef	hr	se(coef)	z	pvalue
PDCD1LG2	-0.67776	0.507751	0.258882	-2.61804	0.008844
LTA	-1.2035	0.300141	0.484417	-2.48444	0.012976
IGLV1-44	-0.39441	0.674078	0.1651	-2.38891	0.016898
CCR10	-0.75399	0.470485	0.329995	-2.28486	0.022321
TNFRSF9	-0.8434	0.430247	0.370424	-2.27684	0.022796
GPR18	-0.47547	0.621591	0.218	-2.18107	0.029178
IGLJ3	-0.49076	0.612164	0.232956	-2.10664	0.035148
IGHG1	-0.54498	0.579854	0.268435	-2.03021	0.042335
IGHM	-0.29906	0.741512	0.153638	-1.94655	0.051589
CD19	-0.49133	0.611815	0.254689	-1.92912	0.053716

Table 4 Lasso regression coefficient of the 11 most significant immune genes

Gene	Coefficient
TRAT1	-0.13118865
IL21R	-0.10504567
IGHM	-0.0997505
CTLA4	-0.09963025
IL2RB	-0.08664438
TNFRSF9	-0.04891361
CTSW	-0.04188042
CCR10	-0.01014675
GPR18	-0.00497377
CR2	-0.00196256
DOCK10	0.240217798

Table 5 Cox regression result showing top most significant immune genes

	coef	hr	se(coef)	z	pvalue
TRAT1	-0.38483917	0.68056	0.090824	-4.23721	2.26E-05
IGHM	-0.36913285	0.691334	0.08911	-4.14242	3.44E-05
IL21R	-0.63111572	0.531998	0.155136	-4.06814	4.74E-05
GZMB	-0.43362517	0.648155	0.109901	-3.94561	7.96E-05
GPR18	-0.51640599	0.596661	0.132572	-3.89528	9.81E-05
CTSW	-0.424037	0.6544	0.110839	-3.82572	0.00013
EVI2B	-0.69469314	0.499228	0.184028	-3.77492	0.00016
CORO1A	-0.65641984	0.518705	0.174807	-3.75512	0.000173
CTLA4	-0.50054211	0.606202	0.133582	-3.74706	0.000179
ITK	-0.61326229	0.541581	0.163868	-3.74242	0.000182
LTB	-0.50805218	0.601666	0.138251	-3.67486	0.000238
IGLJ3	-0.5058777	0.602976	0.137725	-3.67311	0.00024
IGLV1-44	-0.36467637	0.694421	0.099401	-3.66876	0.000244
AIM2	-0.70455447	0.494329	0.192936	-3.65175	0.00026
CXCL9	-0.32522115	0.722368	0.091125	-3.56895	0.000358
IL2RB	-0.76753583	0.464155	0.216092	-3.5519	0.000382
CXCL13	-0.23298293	0.792167	0.065837	-3.53879	0.000402
KIAA0125	-0.8382797	0.432454	0.237175	-3.53444	0.000409
IL2RG	-0.58234889	0.558585	0.165644	-3.51567	0.000439

3.3. Prognostic performance and validation of the Immune Index for DFS/DFMS

Risk score for each patient was calculated according to the immune index formula above. Based on the immune index score, patient samples were further stratified into risk groups within the AH group. The performance of the immune risk score was tested in two parts:1). the risk score as continuous variable, and 2). the risk score based on optimal cutoff points by bootstrapping maximally selected rank statistics using the 'survminer' package in R. Hypothesis was made that that lower (more negative) risk score will be associated with decreased probability of recurrence as well as prolonged survival.

The continuous risk score based on univariate analysis was highly and significantly associated with recurrence outcome ($p<0.0001$). Statistical significance was preserved in the multivariate analysis of the risk score and clinical factors, with the risk score being the most prominent variable associated with recurrence with hazard ratio of 1.46 ($p<0.0001$, 95% CI:1.30-1.65) as risk score increases (data not shown). This result indicates that lower risk score was associated with decreased probability of recurrence as well as prolonged survival. Two optimal points were selected via bootstrapping of maximally selected rank statistics. The first cutoff score stratified immune low and immune high-risk group, with immune low-risk group having hazard ratio of 0.35 ($p<0.0001$,95% CI: 0.25-0.50). The second optimal cutoff point also revealed significance recurrence differences with hazard

ratio of 0.35 ($p < 0.0001$, CI: 0.22-0.56) (Fig. 10). The two optimal cutoff points were then joined to create an intermediate group that was classified differently, either low or high, by the two optimal points. Joining the risk score created three immune index-based risk groups: high, intermediate and low (Table 6, Fig. 11). Fig. 11a shows the survival curves of the discovery set stratified into the three risk groups. All three risk groups displayed statistically significant differences, with intermediate-risk group having a hazard ratio of 0.42 ($p<0.0001$, CI: 0.29-0.56) and low-risk group with a hazard ratio of 0.17 ($p<0.0001$, CI:0.10-0.29) in comparison with the high-risk group. The 5-year survival rate for the immune low-risk group was 90.9%, 56.4% for the immune intermediate-risk group, and 32.5% for the immune high-risk group. Furthermore, the survival rate at 10-years dropped to 73.4% for the immune low-risk group, 51.3% for the immune intermediate-group and drastically dropped to 14.1% for the immune high-risk group. Fig. 11 shows survival curves of the discovery group with the addition of AL&AI groups that were not included in the development of the immune index. The result indicates that there exists no statistical difference between the AL&AI groups and immune low-risk group while statistical differences are retained in comparison to the immune intermediate and low-risk groups.

To observe the independence in recurrence prediction of the risk score, it was further validated by multivariate analysis (Table 6). The immune index manifested statistical significance when adjusted with conventional clinicopathological

parameters by multivariate analysis as well as to be the strongest variable related to DFS/DMFS. In addition to multivariate analysis, the Harrell's Concordance Index (C-index) was used to calculate goodness of fit of the immune index along with another clinical variables (Fig. 12). As shown in the bar plot of C-index measurement, the immune index had the highest concordance value of 0.64 compared to the conventional clinical variables of nodal status (C-index: 0.57), tumor size (C-index: 0.56), histological grade (C-index: 0.52), and age (C-index: 0.50). This result further validates the independence of the risk score as a predictive prognostic indicator in disease recurrence and metastasis, superior to the clinicopathological variables. Of note, immune index in combination with nodal status and tumor size had the highest concordance index, 0.66, followed by combination of immune index and size, 0.65, as shown in the C-index bar plot.

Moreover, the application of the risk score on each of the molecular subtypes was tested. As shown in Fig. 13, the survival curve of the three immune risk groups high, intermediate and low-risk, are shown with the addition to the AL & AI groups. Although HR-/HER2+ and HR+/HER2+ subtypes have immune-high risk group with sample size less than 10, the immune index preserved statistical significance in all four molecular subtypes ($p < 0.05$).

Table 6 Univariate and multivariate analysis of the immune index and clinicopathological variables

	Univariate Analysis			Multivariate Analysis		
	Hazard Ratio	95% CI	*P* value	Hazard Ratio	95% CI	*P* value
Number of patients n=386 Event = 181						
Risk Score: Risk Optimal						
High	1.00			1.00		
Intermediate	0.42	0.29 - 0.60	**<0.0001**	0.49	0.32 - 0.73	**0.0004**
Low	0.17	0.10 - 0.29	**<0.0001**	0.21	0.12 - 0.37	**<0.0001**
Clinical Variables:						
Lymph node infiltration						
0	1.00			1.00		
1	2.02	1.48 - 2.76	**<0.0001**	1.31	0914 - 1.88	0.14044
Histological grade						
High	1.00					
Low&Intermediate	1.24	0.93 - 1.67	0.146			
Tumor size						
A	1.00					
B	0.75	0.55 - 1.02	0.0679			
Age						
A	1.00					
B	1.10	0.81 - 1.48	0.55			

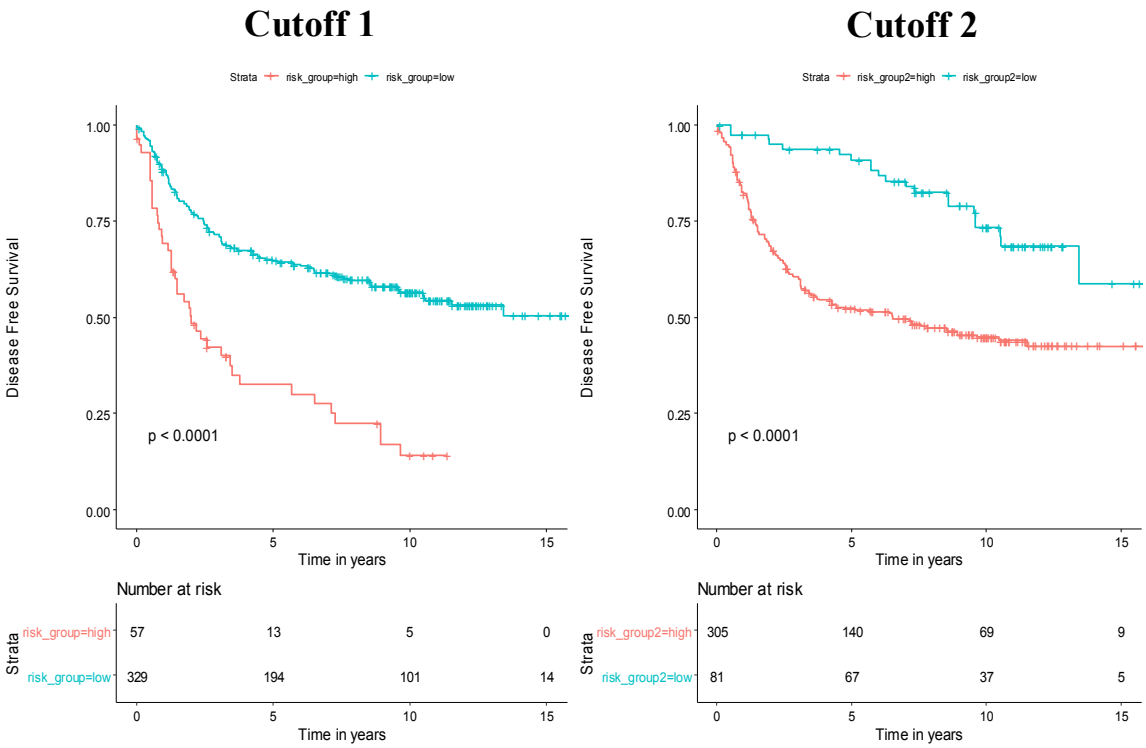


Figure 10 Survival curves of two cutoff points

Kaplan-Meier curves showing DFS/DMFS of low and high-risk groups defined by the immune index. The two different Kaplan-Meier plots are based on two optimal cutoff points of the immune index. Log-rank test used to acquire survival estimate p-value.

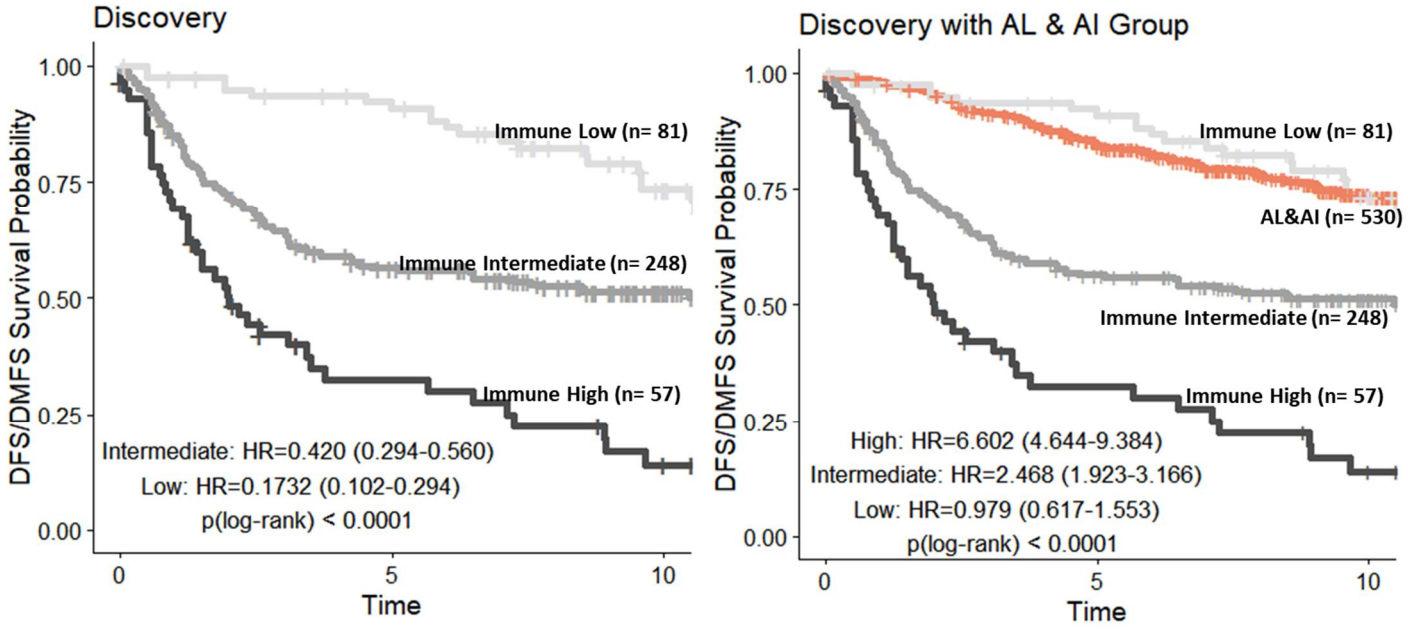


Figure 11 Kaplan-Meier curve of the immune-risk groups

KM curve showing DFS/DMFS of immune high, intermediate and low-risk groups as defined by the optimal cutoff score of the immune index. Significance is retained in all groups compared to the immune high-risk group (11a). KM curve was plotted with the addition of AL&AI groups. AL and immune low-risk group overlap, and no significance is shown between the two groups (11b).

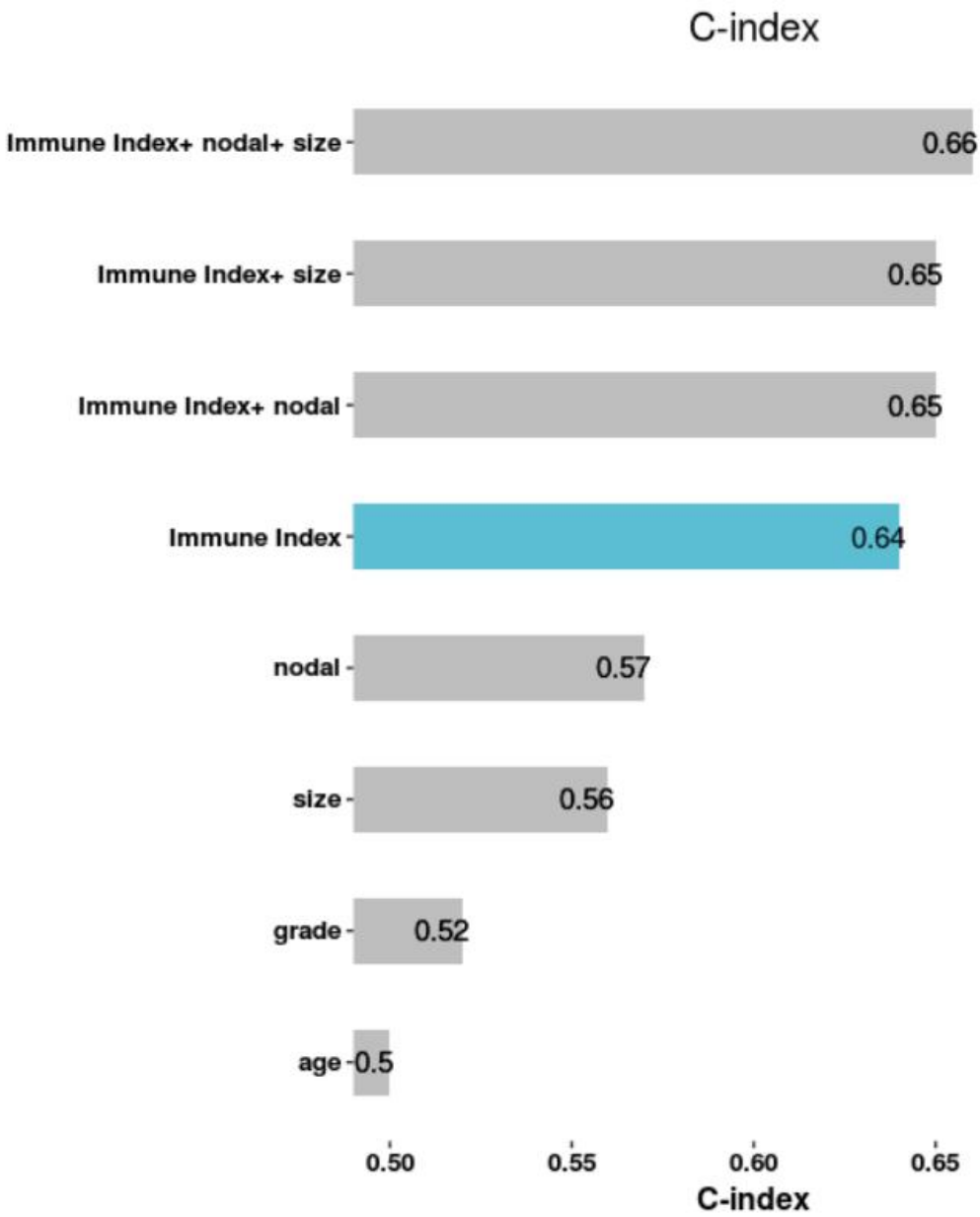


Figure 12 C-index bar plot

Prognostic performance of the immune index in predicting DFS/DMFS is represented as C-index along with other clinical characteristics and individual genes. The immune index with nodal status and size has the highest C-index score.

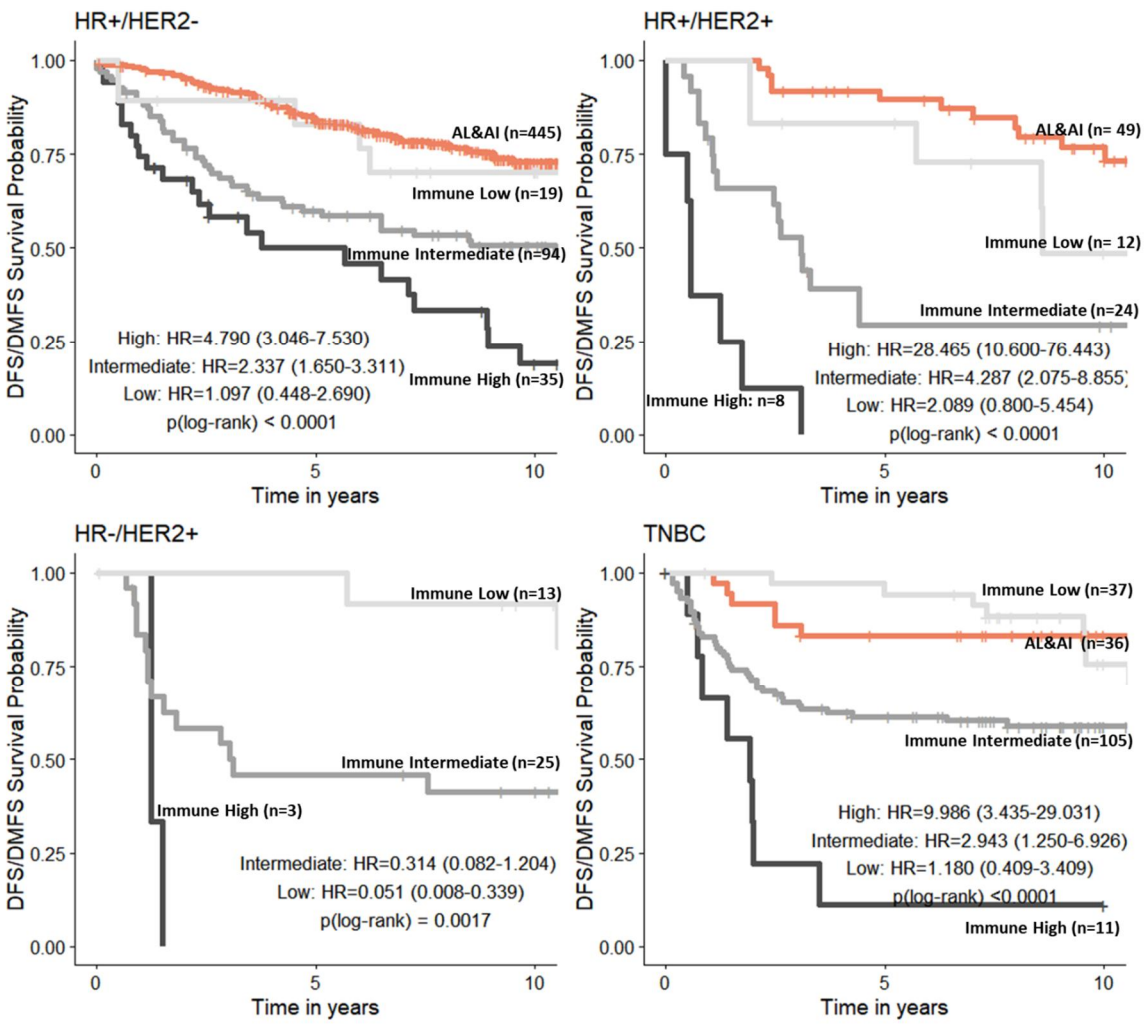


Figure 13 Survival curve of all molecular subtypes

Prognostic value of the immune index in all molecular subtypes. AL & AI groups were added for comparison. Clock-wise HR+/HER2-, HR+/HER2+, TNBC, and HR-/HER2+.

3.4. Validation of the prognostic model in microarray and METABRIC datasets

To further expand its application, the immune index developed from the discovery set was further validated by cohorts across various platforms. In total, the risk model was tested in three different test sets; two different microarray platform sets and the other validation set used METABRIC data. First validation set selected, GSE3494, was of the same platform as the discovery cohorts, (Affymetrix GPL96). Second validation dataset consisted of two cohorts GSE21653 and GSE42568 (Affymetrix GPL570).

The optimal cutoff value of -7.1 was applied to the validation sets to stratify patients into immune low and high-risk groups within the clinically high-risk and gene high-risk group (AH risk group). Fig. 14 displays survival curve of the three validation datasets showing significance difference in recurrence and survival between the low and high-risk groups. Fig. 14a shows OS differences between the two groups defined by the immune index of GSE3494 cohort with hazard ratio of 0.36 (p=0.0339, CI: 0.14-0.92) in the low-risk group and Fig. 14b displays recurrence probability of the two groups in GSE21653&GSE42568 validation set with hazard ratio of 0.23 (p=0.0137, CI:0.07-0.74). In both validation sets, immune index successfully classified low and high-risk groups, retaining statistical differences in survival outcomes.

In addition, the 5-year overall survival rate in the immune low and high-risk group was 90.0% and 60.9% respectively, for the first validation set (GSE3494), and 5-year DFS was 89.7% and 50.0% for the low and high-risk groups, respectively, in the second validation set (GSE42683, GSE21653). These rates change to 75.0% and 50.8% for low and high-risk groups respectively at 10-year overall survival in the first validation set, and changes to 79.8% and 33.7% of recurrence rate for low and high-risk groups respectively in the second validation set. Furthermore, univariate and multivariate analysis conducted on the combined Affymetrix GPL96 cohort showed the risk score to be the strongest prognostics after adjustment of other variables (Table 7). Overall, based on the findings from microarray validation sets, the risk model proved robustness in predicting overall survival and recurrence ($p<0.05$).

Finally, the immune index was validated by METABRIC cohort using overall survival as the primary endpoint. Due to the abundant clinical information including adjuvant therapies, only those patients who did not receive adjuvant chemotherapy were selected, as in the discovery set. In total, 370 patients were analyzed by the immune index. However, only three out of five genes of the immune index score was found in the METABRIC data set and consequently, two genes (*IGHM* and *IL2RB*) were excluded, and thus coefficient of the three genes acquired in the METABRIC dataset was used to make changes to the algorithm. Survival analysis conducted on the METABRIC cohort revealed that the risk score

preserved statistical significance defined by the optimal cutoff point ($p<0.0001$) as shown in Table 8. Cox regression performed on the validation set revealed that risk groups divided according to optimal cutoff point of the immune index successfully maintained significance between immune low and high-risk groups (Fig. 14c).

Once again in the METABRIC dataset, the immune index preserved significance in OS and had the strongest prognostic performance after adjustment of other variables (Table 8). The survival rate at 5-years was 97.0% for the immune low-risk group and 72.1% for the immune high-risk group. These rates change to 83.3% and 51.2% at 10-years for immune low-risk and high-risk respectively. Lastly, the immune index was applied to all subtypes in METABRIC dataset, and significance was retained in all breast cancer subtypes with the exception of HR+/HER2- ($p=0.093$) (Fig. 15).

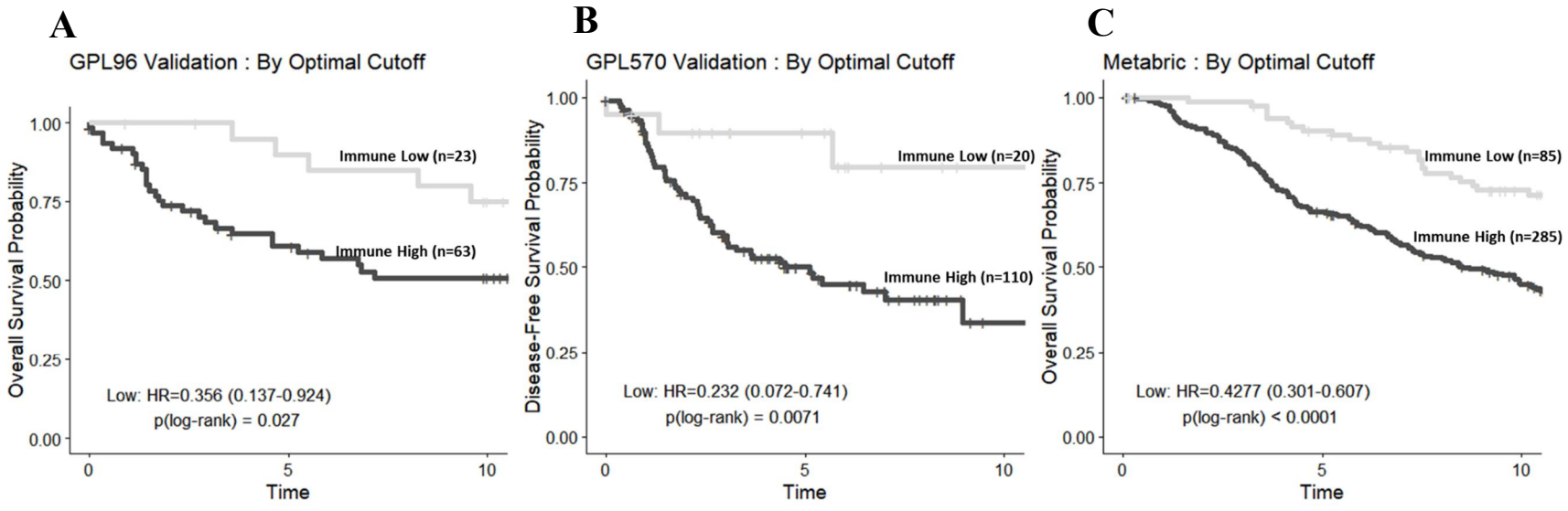


Figure 14 Survival curve of the validation cohorts

Validation of the immune index in two Affymetrix microarray platforms. (a) GPL96 showing OS and (b) GPL570 showing DFS. Patients were stratified into low and high groups by optimal cutoff points of each validation set. (c) Survival curve of Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort.

Table 7 Univariate and multivariate analysis in of the two microarray validation sets

	Univariate Analysis (GSE3494)				Multivariate Analysis (GSE3494)		
	Hazard Ratio	95% CI	*P* value		Hazard Ratio	95% CI	*P* value
Number of patients n=86 Events = 33				Number of patients n=86 Events = 33			
Risk Score:				**Risk Score:**			
Continous				Continous			
As score increases	2.24	1.41 - 3.57	**0.000664**	As score increases	2.73	1.13 - 6.58	**0.0252**
Risk Optimal							
High	1.00			**Clinical Variables:**			
Low	0.36	0.14 - 0.92	**0.0339**	Lymph node infiltration			
Clinical Variables:				0	1.00		
Lymph node infiltration				1	0.68	0.17 - 2.88	0.604
0	1.00						
1	2.74	1.30 - 5.80	**0.00824**				
Histological grade							
High	1.00						
Low&Intermediate	1.03	0.52 - 2.04	0.937				
Tumor size							
A	1.00						
B	2.36	0.83 - 6.73	0.107				
Age							
A	1.00						
B	1.39	0.66 - 2.93	0.382				

	Univariate Analysis (GSE42563 & GSE21563)			Multivariate Analysis (GSE42568 & GSE21653)		
	Hazard Ratio	95% CI	*P* value	Hazard Ratio	95% CI	*P* value
Number of patients						
n=130						
Events = 58						
Risk Score:						
Continous				1.00		
As score increase	2.47	1.42 - 4.30	**0.00139**	2.30	1.31 - 4.04	**0.00384**
Risk Optimal						
High	1.00					
Low	0.24	0.07 - 0.74	**0.0137**			
Clinical Variables:						
Lymph node infiltration						
0	1.00			1.00		
1	2.41	1.38 - 4.22	**0.00198**	2.19	1.25 - 3.83	**0.00619**
Histological grade						
High	1.00					
Low&Intermediate	0.65	0.35 - 1.21	0.173			
Tumor size						
A	1.00					
B	0.89	0.35 - 2.22	0.797			
Age						
A	1.00					
B	0.83	0.48 - 1.45	0.52			

Table 8 Univariate and multivariate analysis of the immune index and clinicopathological variables of METABRIC validation set

	Univariate Analysis				Multivaraite Analysis		
	Hazard Ratio	95% CI	*P* value		Hazard Ratio	95% CI	*P* value
Number of patients n=370 Events = 250				Number of patients n=370 Events = 250			
Risk Score:				**Risk Score:**			
Continous				High	1.00		
As score increase	1.7	1.36-2.13	**<0.0001**	Low	0.50	0.35-0.73	**0.0003**
Risk Optimal				**Clinical Variables:**			
High	1.00			Lymph node infiltration			
Low	0.43	0.30-0.61	**<0.0001**	0	1.00		
Clinical Variables:				1	1.22	0.84-1.77	0.2886
Lymph node infiltration				Tumor size			
0	1.00			A	1.00		
1	1.86	1.37 - 2.53	**<0.0001**	B	1.40	1.02-1.92	**0.0361**
Histological grade				Age			
High	1.00			A	1.00		
Low&Intermediate	1.14	0.88 - 1.50	0.3080	B	0.99	0.64-1.55	0.9994
Tumor size							
A	1.00						
B	1.74	1.32 - 2.30	**0.0001**				
Age							
A	1.00						
B	0.86	0.56-1.32	0.4890				

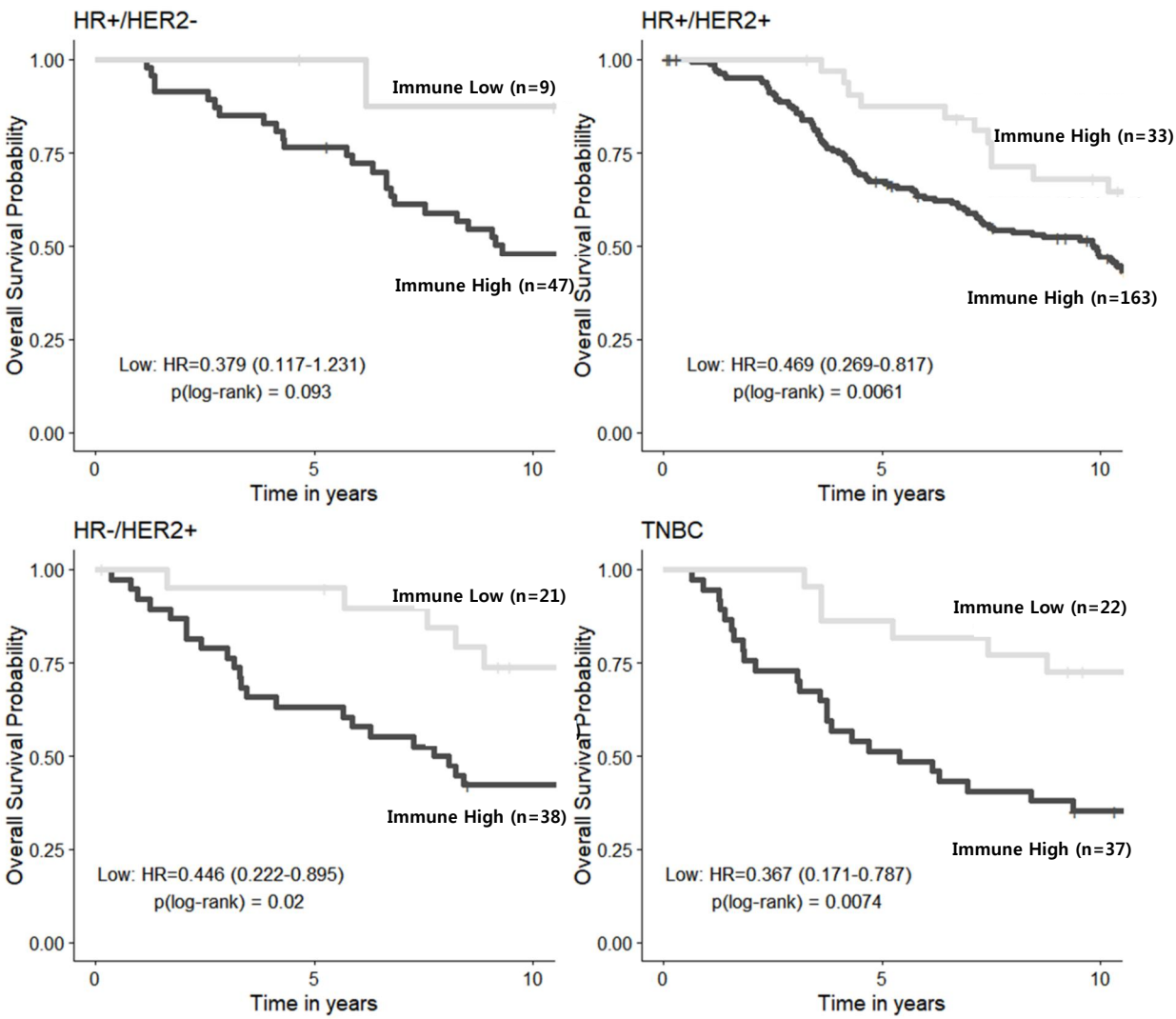


Figure 15 Survival curve of all molecular subtype in METABRIC

Prognostic value of the immune index in all molecular subtypes of the METABRIC validation set. Based the immune index, patients were stratified into immune high and low-risk groups. Significance was retained in all but HR+/HER2-.

intermediate and high-risk groups. Prognostic significance of the immune index in the HR+ subtype is important given that most previous findings have concentrated on the significance of immunity found in HR- subtypes. Interestingly, the immune model demonstrated robustness in both HR+ and HR- subtypes, signifying that elevated expression of immune genes correlates with better prognosis in the HR+/HER2- and HR+/HER2+ molecular subtypes. This result indicates that immune response signatures are associated with positive prognosis in all breast cancer types at a specific progression stage of breast cancer, thereby supporting the findings from Oh et al., which confirmed that rise immunity is beneficial in all subtypes in early breast cancer. In addition, the immunogenetic model successfully categorized low-risk groups within the clinically-high, proliferative-high group (AH group). Furthermore, the low-risk group stratified by the immune index may have favorable survival outcome without further intervention of adjuvant chemotherapy.

While appropriate adjuvant therapies may significantly increase the lifespan of a patient, there are still limitations and numerous side effects that accompany these therapies, and thus should only be given to high risk patients [4,5]. Immunotherapy, on other hand, may suggest a new treatment paradigm for breast cancer patients by boosting a patient's immune system. Tumor specific immune cells when activated, will only target dividing tumor cells in contrast to chemotherapy and radiation therapy which destroys all dividing cells, including non-tumor cells [5]. Many

researches in the past decade and ongoing have found breast cancers infiltrated by immune cells with prognostic and predictive value. Consequently, deeper insights on immune microenvironment and their contribution to patient survival can induce adequate immunotherapy for breast cancer patients [6].

This study found that the elevated expression of five genes, *TRAT1, IL21R, CTLA4, IGHM*, and *IL2RB*, with the addition of lymph node status was associated with better prognosis in the AH group, classified according to predefined clinical and proliferation risk classification. Nodal status is strongly associated with unfavorable prognostic outcome, and thus it was incorporated into the risk-model. The five genes selected for the genetic model have shown prognostic value. *TRAT1* plays a role in cellular defense response [30]. Although its association with positive survival has been suggested in metastatic melanoma, its role in breast cancer progression and survival has been elusive. Yet, deducing from the findings of the present research, *TRAT1* may enhance anti-tumorigenic immune activity in early breast cancer patients.

On the contrary, the positive prognostic effect of *IL21R*, type 1 cytokine receptor, in breast cancer has been well-reported. *IL21R* is negatively associated with distant disease-free survival (DDFS) when its expression increases in HER2+. *IL21R* is known to be widely expressed by immune cells. Importantly, its expression on CD8+ T cells is required for optimal anti-ErbB2 monoclonal antibody therapy [31]. *IGHM* is a recognition antigen of B cells. Like *IL21R*, function of IgM heavy chain

in breast cancer has been reported. IgM heavy chain transcripts are present in low-proliferating breast cancer [32]. However, no definitive association with survival was found in previous breast cancer studies.

IL2RB is a gene mainly involved in T cell-mediated immune response and regulates the proliferation and differentiation of cytotoxic T cells and NK cells. *IL2RB* tethered to *IL2* will enhance NK-mediated immunity. Previous findings have reported that *IL2RB* increases NK cytotoxic activity against cancer cells [33,34].

CTLA4, Cytotoxic T lymphocyte antigen 4, is highly expressed in breast cancer cells. Although *CTLA4* is well known to be inhibitor of anti-tumor response and down-regulator of T-cell activation, findings from present research contrastingly suggest that *CTLA4* at early stages of breast cancer is associated with positive prognosis. This may perhaps be due to breast cancer microenvironment. Yu et al. [35] reported that *CTLA4* in TIL are associated with longer DFS and OS, however, higher expression of *CLTA4* has contrasting effects in carcinoma and results in diminished OS.

In summary, *TRAT1* and *IGHM* that have not been previously reported to have any prognostic association with breast cancer, however their prognostic value in breast cancer recurrence and survival has been revealed for the first time through this study. While *IL21R* and *CTLA4* have been reported to negatively affect survival outcomes in breast cancer, intriguingly, both of these genes were revealed to act

otherwise in the pN0/1 early breast cancer dataset. Since the analyzed datasets were highly homogenized by risk groups, this may suggest the differing role of immune-related genes in different stages of cancer. The landscape of cancer immunity may change in regards to survival as breast cancer progresses.

Although few limitations in this research exists, the immune index proved to be the most robust prognostic variable of DFS/DMFS compared to clinicopathological variables. Additionally, the relatively small sample size used in the discovery set was due to lack of complete demographic information in numerous public datasets. To categorize patients into risk groups, we needed TNM, histological grade, as well as chemotherapy status, for adjuvant chemotherapy may bias survival outcomes significantly. In the same light, there was disparity between sample numbers in each sample subtype. This disparity became more drastic when each subtype was subdivided and homogenized according to clinical and genetic risk criteria. Future research will be strengthened with a greater sample size in each of the homogenized risk groups and evenly proportioned sample sizes will be favorable in each risk groups for unbiased and robust statistical validations of the immune index.

Furthermore, discrepancy in the optimal cutoff score between the discovery dataset and validation sets exists. While two optimal cutoff points of the genomic algorithm was selected for the discovery set, in the two microarray validation sets, only one optimal cutoff point (cutoff score = -7.1) was applied to stratify patients

into low and high-risk groups. The validation set did not have score ranges that were as low as the second cutoff point (cutoff score= -9.4). This may be the effect of differing chemotherapy status between patients in the discovery and validation dataset. In future studies validating the prognostic performance of the immune index, validation sets that have not received chemotherapy may be more favorable to determine exact prognostic performance of the immunogenetic algorithm. Finally, based on the prognostic performance of the model in this study, immune index may prove to be not only prognostic, but predictive model in prospective clinical trial predicting outcomes of chemotherapy administration.

5. Conclusion

Overall, this study was successful in developing a novel prognostic model constructed with immune response genes related to recurrence and survival in breast cancer. It was noteworthy that the model composed of only immune response genes with the addition of lymph node status was robust and independent prognostic factor in the subdivided and homogenized patient pools in both discovery and validation datasets. Since limitations to the current prognostic algorithms exist, the immune index can potentially assist or better yet, improve by selecting patients pool with greater likelihood for survival without the administration of adjuvant chemotherapy.

6. References

[1] Colozza M, Azambuja E, Cardoso F, Sotiriou C, Larsimont D, Piccart MJ. Proliferative markers as prognostic and predictive tools in early breast cancer: where are we now? Ann Oncol 2005;16:1723–39. doi:10.1093/annonc/mdi352.

[2] van Diest PJ, van der Wall E, Baak JPA. Prognostic value of proliferation in invasive breast cancer: a review. J Clin Pathol 2004;57:675–81. doi:10.1136/jcp.2003.010777.

[3] Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. N Engl J Med 2004;351:2817–26. doi:10.1056/NEJMoa041588.

[4] van de Vijver MJ, He YD, van 't Veer LJ, Dai H, Hart AAM, Voskuil DW, et al. A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. N Engl J Med 2002;347:1999–2009. doi:10.1056/NEJMoa021967.

[5] Bernard PS, Parker JS, Mullins M, Cheung MCU, Leung S, Voduc D, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009;27:1160–7. doi:10.1200/JCO.2008.18.1370.

[6] Filipits M, Rudas M, Jakesz R, Dubsky P, Fitzal F, Singer CF, et al. A new molecular predictor of distant recurrence in ER-positive, HER2-negative

breast cancer adds independent information to conventional clinical risk factors. Clin Cancer Res 2011;17:6012–20. doi:10.1158/1078-0432.CCR-11-0926.

[7] Yang B, Chou J, Tao Y, Wu D, Wu X, Li X, et al. An assessment of prognostic immunity markers in breast cancer. Npj Breast Cancer 2018;4:35. doi:10.1038/s41523-018-0088-0.

[8] Giuliano AE, Connolly JL, Edge SB, Mittendorf EA, Rugo HS, Solin LJ, et al. Breast Cancer-Major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. CA Cancer J Clin 2017;67:290–303. doi:10.3322/caac.21393.

[9] Tiberi D, Masucci L, Shedid D, Roy I, Vu T, Patocskai E, et al. Limitations of Personalized Medicine and Gene Assays for Breast Cancer. Cureus 2017;9:e1100. doi:10.7759/cureus.1100.

[10] Alvarado MD, Prasad C, Rothney M, Cherbavaz DB, Sing AP, Baehner FL, et al. A Prospective Comparison of the 21-Gene Recurrence Score and the PAM50-Based Prosigna in Estrogen Receptor-Positive Early-Stage Breast Cancer. Adv Ther 2015;32:1237–47. doi:10.1007/s12325-015-0269-2.

[11] Dowsett M, Sestak I, Lopez-Knowles E, Sidhu K, Dunbier AK, Cowens JW, et al. Comparison of PAM50 Risk of Recurrence Score With Onco *type* DX and IHC4 for Predicting Risk of Distant Recurrence After Endocrine

Therapy. J Clin Oncol 2013;31:2783–90. doi:10.1200/JCO.2012.46.1558.

[12] Disis ML, Stanton SE. Immunotherapy in breast cancer: An introduction. The Breast 2018;37:196–9. doi:10.1016/J.BREAST.2017.01.013.

[13] Alexe G, Dalgin GS, Scanfeld D, Tamayo P, Mesirov JP, DeLisi C, et al. High expression of lymphocyte-associated genes in node-negative HER2+ breast cancers correlates with lower recurrence rates. Cancer Res 2007;67:10669–76. doi:10.1158/0008-5472.CAN-07-0539.

[14] Schmidt M, Böhm D, Von Törne C, Steiner E, Puhl A, Pilch H, et al. The Humoral Immune System Has a Key Prognostic Impact in Node-Negative Breast Cancer 2008. doi:10.1158/0008-5472.CAN-07-5206.

[15] Teschendorff AE, Caldas C. A robust classifier of high predictive value to identify good prognosis patients in ER-negative breast cancer. Breast Cancer Res 2008;10:R73. doi:10.1186/bcr2138.

[16] Oh E, Choi Y-L, Park T, Lee S, Seok •, Nam J, et al. A prognostic model for lymph node-negative breast cancer patients based on the integration of proliferation and immunity. Breast Cancer Res Treat 2012;132:499–509. doi:10.1007/s10549-011-1626-8.

[17] Schmidt M, Hengstler JG, von Törne C, Koelbl H, Gehrmann MC. Coordinates in the universe of node-negative breast cancer revisited. Cancer Res 2009;69:2695–8. doi:10.1158/0008-5472.CAN-08-4013.

[18] Bedognetti D, Hendrickx W, Marincola FM, Miller LD. Prognostic and predictive immune gene signatures in breast cancer. Curr Opin Oncol 2015;27:433–44. doi:10.1097/CCO.0000000000000234.

[19] Han J, Choi Y-L, Kim H, Young Choi J, Kyung Lee S, Eon Lee J, et al. MMP11 and CD2 as novel prognostic factors in hormone receptor-negative, HER2-positive breast cancer. Breast Cancer Res Treat 2017;10:41–56. doi:10.1007/s10549-017-4234-4.

[20] Nagarajan D, McArdle S, Nagarajan D, McArdle SEB. Immune Landscape of Breast Cancers. Biomedicines 2018;6:20. doi:10.3390/biomedicines6010020.

[21] Teschendorff AE, Miremadi A, Pinder SE, Ellis IO, Caldas C. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. Genome Biol 2007;8:R157. doi:10.1186/gb-2007-8-8-r157.

[22] Dieci MV, Radosevic-Robin N, Fineberg S, van den Eynden G, Ternes N, Penault-Llorca F, et al. Update on tumor-infiltrating lymphocytes (TILs) in breast cancer, including recommendations to assess TILs in residual disease after neoadjuvant therapy and in carcinoma in situ: A report of the International Immuno-Oncology Biomarker Working Group on Breast Cancer. Semin Cancer Biol 2018;52:16–25. doi:10.1016/J.SEMCANCER.2017.10.003.

[23] Calabrò A, Beissbarth T, Kuner R, Stojanov M, Benner A, Asslaber M, et al. Effects of infiltrating lymphocytes and estrogen receptor on gene expression and prognosis in breast cancer. Breast Cancer Res Treat 2009;116:69–77. doi:10.1007/s10549-008-0105-3.

[24] Hammerl D, Smid M, Timmermans AM, Sleijfer S, Martens JWM, Debets R. Breast cancer genomics and immuno-oncological markers to guide immune therapies. Semin Cancer Biol 2018;52:178–88. doi:10.1016/J.SEMCANCER.2017.11.003.

[25] Cheang MCU, Voduc D, Bajdik C, Leung S, McKinney S, Chia SK, et al. Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. Clin Cancer Res 2008;14:1368–76. doi:10.1158/1078-0432.CCR-07-1658.

[26] Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 2012;28:882–3. doi:10.1093/bioinformatics/bts034.

[27] Rody A, Karn T, Liedtke C, Pusztai L, Ruckhaeberle E, Hanker L, et al. A clinically relevant gene signature in triple negative and basal-like breast cancer. Breast Cancer Res 2011;13:R97. doi:10.1186/bcr3035.

[28] Wolf DM, Lenburg ME, Yau C, Boudreau A, Van ’t Veer LJ. Gene Co-

Expression Modules as Clinically Relevant Hallmarks of Breast Cancer Diversity n.d. doi:10.1371/journal.pone.0088309.

[29] Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. N Engl J Med 2016;375:717–29. doi:10.1056/NEJMoa1602253.

[30] Bogunovic D, O'Neill DW, Belitskaya-Levy I, Vacic V, Yu Y-L, Adams S, et al. Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in predicting patient survival. Proc Natl Acad Sci 2009;106:20429–34. doi:10.1073/pnas.0905139106.

[31] Mittal D, Caramia F, Michiels S, Joensuu H, Kellokumpu-Lehtinen P-L, Sotiriou C, et al. Microenvironment and Immunology Improved Treatment of Breast Cancer with Anti-HER2 Therapy Requires Interleukin-21 Signaling in CD8 þ T Cells 2016. doi:10.1158/0008-5472.CAN-15-1567.

[32] Whiteside TL, Ferrone S. For breast cancer prognosis, immunoglobulin kappa chain surfaces to the top. Clin Cancer Res 2012;18:2417–9. doi:10.1158/1078-0432.CCR-12-0566.

[33] Marra P, Mathew S, Grigoriadis A, Wu Y, Kyle-Cezar F, Watkins J, et al. IL15RA drives antagonistic mechanisms of cancer development and immune control in lymphocyte-enriched triple-negative breast cancers. Cancer Res 2014;74:4908–21. doi:10.1158/0008-5472.CAN-14-0637.

[34] Jounaidi Y, Cotten JF, Miller KW, Forman SA. Therapeutics, Targets, and Chemical Biology Tethering IL2 to Its Receptor IL2Rb Enhances Antitumor Activity and Expansion of Natural Killer NK92 Cells 2017. doi:10.1158/0008-5472.CAN-17-1007.

[35] Yu H, Yang J, Jiao S, Li Y, Zhang W, Wang J. Cytotoxic T lymphocyte antigen 4 expression in human breast cancer: implications for prognosis. Cancer Immunol Immunother 2015;64:853–60. doi:10.1007/s00262-015-1696-2.

Abstract

요약 (국문초록)

초기 유방암의 고위험군 분류를 위한 새로운 예후 면역지표

서울대학교 대학원

협동과정 생물정보학

이 한 나

유방암에서 증식 관련 유전자들의 예후와의 강력한 연관성은 많은 선행 연구들을 통해 입증 되었고 다중 유전자 기반의 유전자 예후 진단에서 증식 관련 유전자 또한 광범위하게 분석 되었다. 하지만 현존하고 있는 유방암 예후 진단은 더 높은 정확도와 주요 환자군인 ER+ 환자분자 아형 외에도 더 넓은 적용이 필요하다.

증식 유전자들이 유방암에 미치는 손상효과와는 달리, 면역반응관련 유전자들의 방어적 역할과 유방암 예후 예측 능력이 보고되어져 왔다.

따라서 본 연구는 면역 반응 관련 유전자들이 유방암에 미치는 영향과 재발 및 생존에 대한 기여가 기존의 유방암 유전자 분석법에 어떻게 적용 될 수 있는 지 연구해 보았다. 라쏘 회귀분석(Lasso regression)으로 생존의 제일 유의미한 영향을 끼치는 5 가지의 유전자를 선별하여 예후 알고리즘을 만들고 암 예후의 제일 큰 영향을 끼친다고 알려져 있는 임상 병리학적 변수인 림프절 상태를 변수로 추가하여 예후 모델의 유의

미성을 Adjuvant! Online 과 증식 유전자들에 따라 층화된 위험 그룹에서 확인하였다.

본 연구를 통해 확인된 새로운 면역 유전자 알고리즘은 층화된 그룹 안에서도 환자들을 유의미한 예후 위험 그룹으로 재분류함으로써, 강력한 예후 분류 능력을 입증하였다. 또한 본 면역 알고리즘은 유방암의 모든 분자 아형, 특히 HR− 유방암 아형에서의 유의미한 예후와의 연관성을 나타냈다. 현존하고 있는 다중 유전자 분석에는 한계가 존재하므로 본 연구에서 개발한 새로운 면역 예후 모델은 보조 화학 요법의 시행 없이도 생존 가능성이 더 높은 환자군을 선택함으로써 기존의 유전자 분석을 보조할 수 있는 예후 알고리즘으로써의 가능성을 보였다.

주요어: 유방암, 면역 반응 관련 유전자, 증식 유전자, 분자 아형, 예후 모델, 재발 및 생존 위험, 보조 화학 요법

학번: 2017−26462

감사의 글

저게는 많은 경험과 배움 그리고 인내의 시간이었던 석사 과정이 이렇게 마무리가 될 수 있어서 참 감사합니다. 제 혼자 힘으로는 절대로 이 과정을 마무리할 수 없었습니다. 석사학위를 무사히 마칠 수 있도록 도움 주신 분들에게 진심으로 감사의 말을 전합니다.

우선 신교수님 항상 이끌어 주시고 연구자로서의 역량을 쌓을 수 있게 해주셔서 감사합니다. 교수님 밑에서 과학자로서의 기본을, 그리고 연구에 꼭 필요한 질문을 끊임없이 하는 법을 배울 수 있었습니다. 앞으로도 이런 사고방식과 질문들로 연구에 임하겠습니다.

연구실 선배들에게도 감사의 인사드립니다. 항상 건강하시고 앞으로의 연구를 비롯하여 모든 일 다 잘되시길 바랄게요. 응원하겠습니다. 또한, 동기들에게 감사의 인사 전하고 싶습니다. 너무 훌륭한 동기들을 만나서 연구실 생활 동안 많은 위로를 받고 버틸 수 있었던 것 같습니다. 앞으로의 미래가 더 기대되는 친구들이기에 졸업 이후에도 각자의 삶에서 멋있고 의미 있는 일들을 해내리라 믿습니다. 연구실 후배들, 그리고 현재 열심히 학위 과정중이신 분들에게도 응원의 말 전하고 싶어요. 더 많이 챙겨주고 고민도 들어줬어야 했는데 그러지 못해 아쉬움이 남습니다. 앞으로의 남은 과정 거뜬하게 마칠 수 있기를 기도하겠습니다. 특히 부사수

범모에게 너무 많은 도움을 받아서 빚진 것 같은 기분이 듭니다. 학위 마지막까지 저보다도 더 잘 해내리라 의심하지 않습니다. 마지막으로 졸업과 각자의 이유로 지금은 연구실에 없지만 함께하는 동안 맺은 인연들, 같이 한 시간 동안 행복했습니다. 항상 건강하시고 각자의 자리에서 맡은 바 잘 해내실 거라 생각합니다.

연구실 밖에서 저를 응원하고 지켜줬던 친구들 그리고 언니 오빠들. 무너지려는 많은 순간 저를 붙들어 주고 위로해줘서 계속 나아갈 수 있었습니다. 같이 웃을 수 있어 따듯했습니다. 진심으로 감사드려요.

부모님, 항상 미끄러지려고 하는 저를 잡아주시고 바로 세워 주셔서 감사합니다. 엄마 아빠의 응원과 기도가 없었다면 여기까지 오지 못했을 거예요. 항상 감사하고 사랑해요.

And finally, thank-you Lord for where you have led me. Thank-you for all that I've been through and for all that's to come. Through it all, may you be glorified!

제 삶 곳곳에 다른 분들의 도움을 받지 않은 흔적이 없습니다. 지난 2년 넘는 시간 동안의 배움과 경험을 감사히 기억하며, 앞으로도 가치 있는 것을 알고, 가꾸며 지키는 사람으로 살겠습니다. 다시 한 번 깊이 감사드립니다.