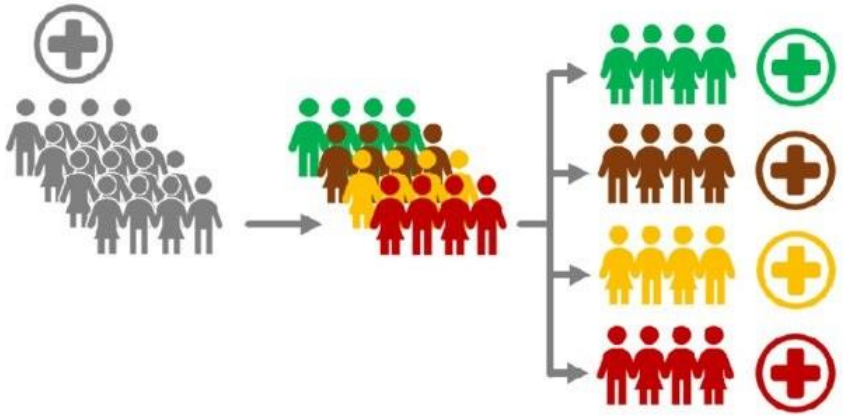
基因與臨床資料的整合運用



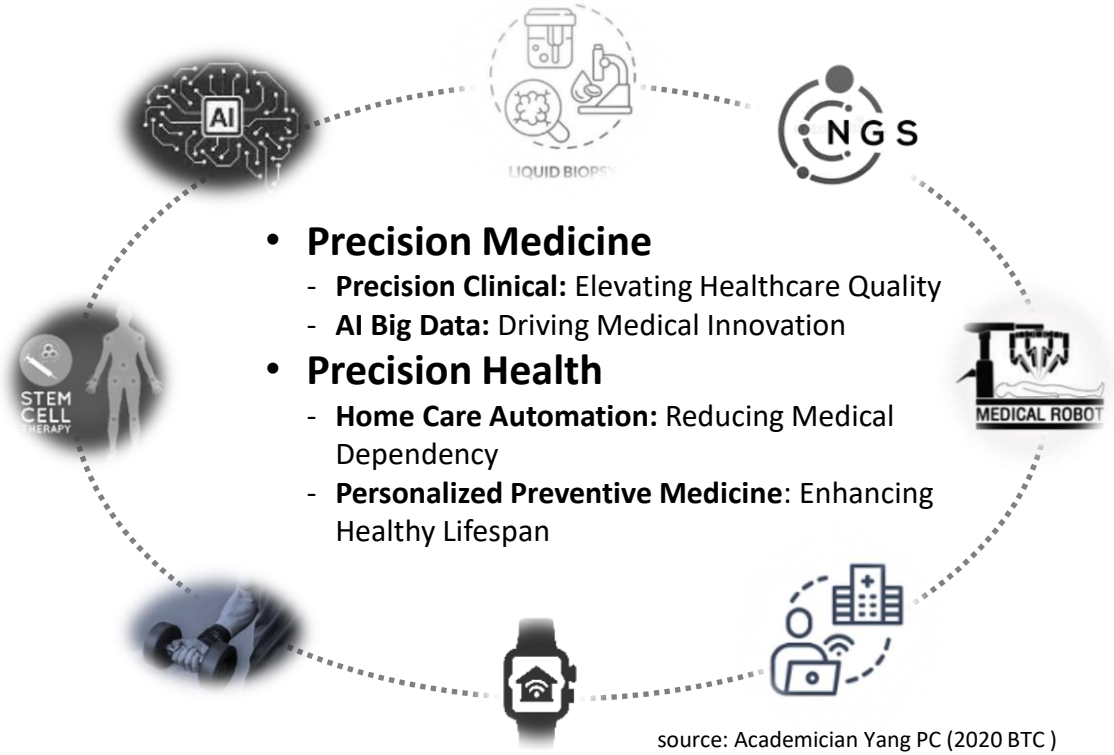
盧子彬 Tzu-Pin Lu
tplu@ntu.edu.tw
Institute of Health Data Analytics and Statistics,
National Taiwan University

05/15/2025

Sep 2018 Quest: Science for South Africa
https://www.shutterstock.com/zh/image-vector/big-data-medicine-digitization-all-sorts-1363311101

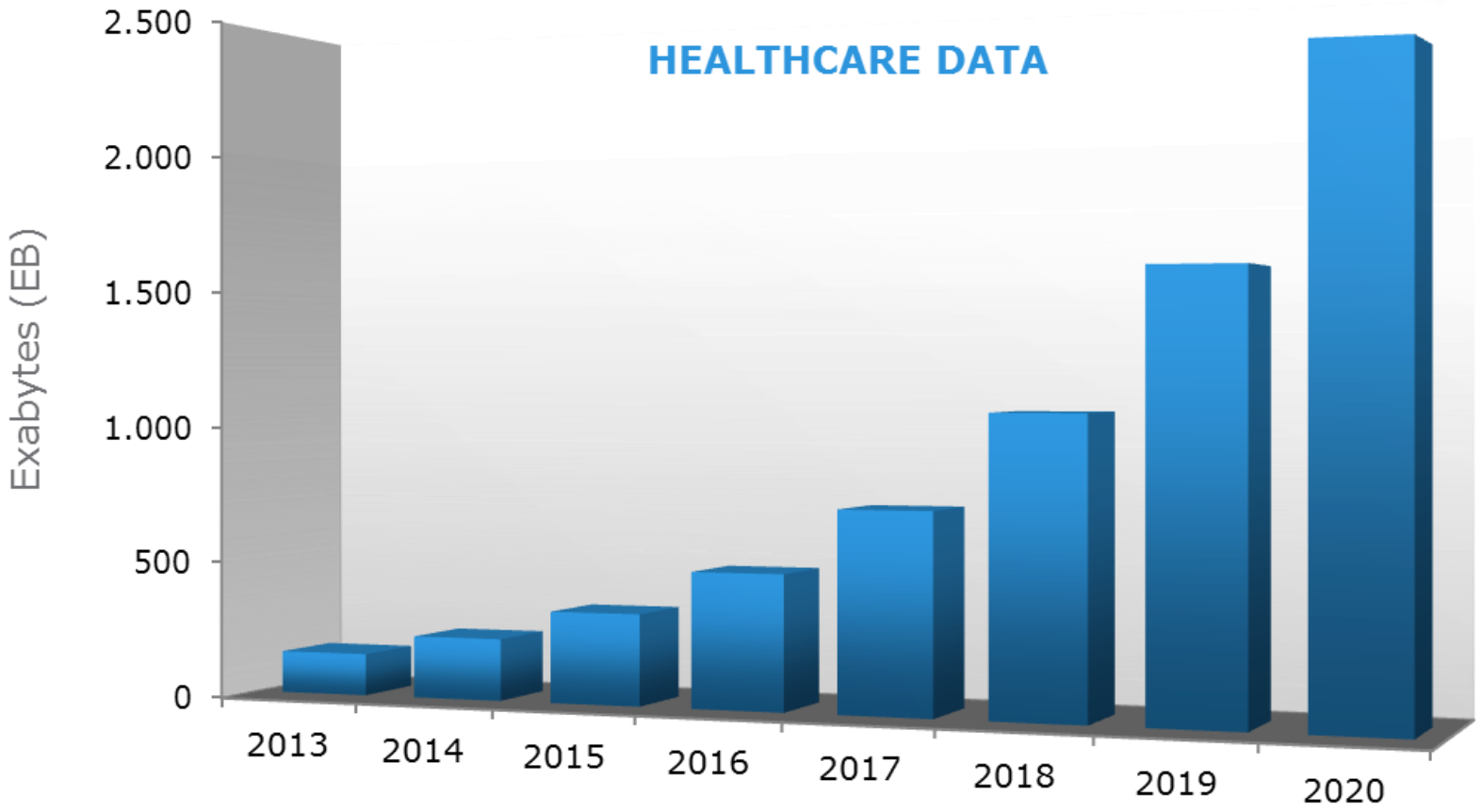
Precision Health
Start from a healthy state before illness

- **Predict, prevent, diagnose, and treat diseases accurately based on genetic, environmental, lifestyle, and molecular differences.**



source: Academician Yang PC (2020 BTC)

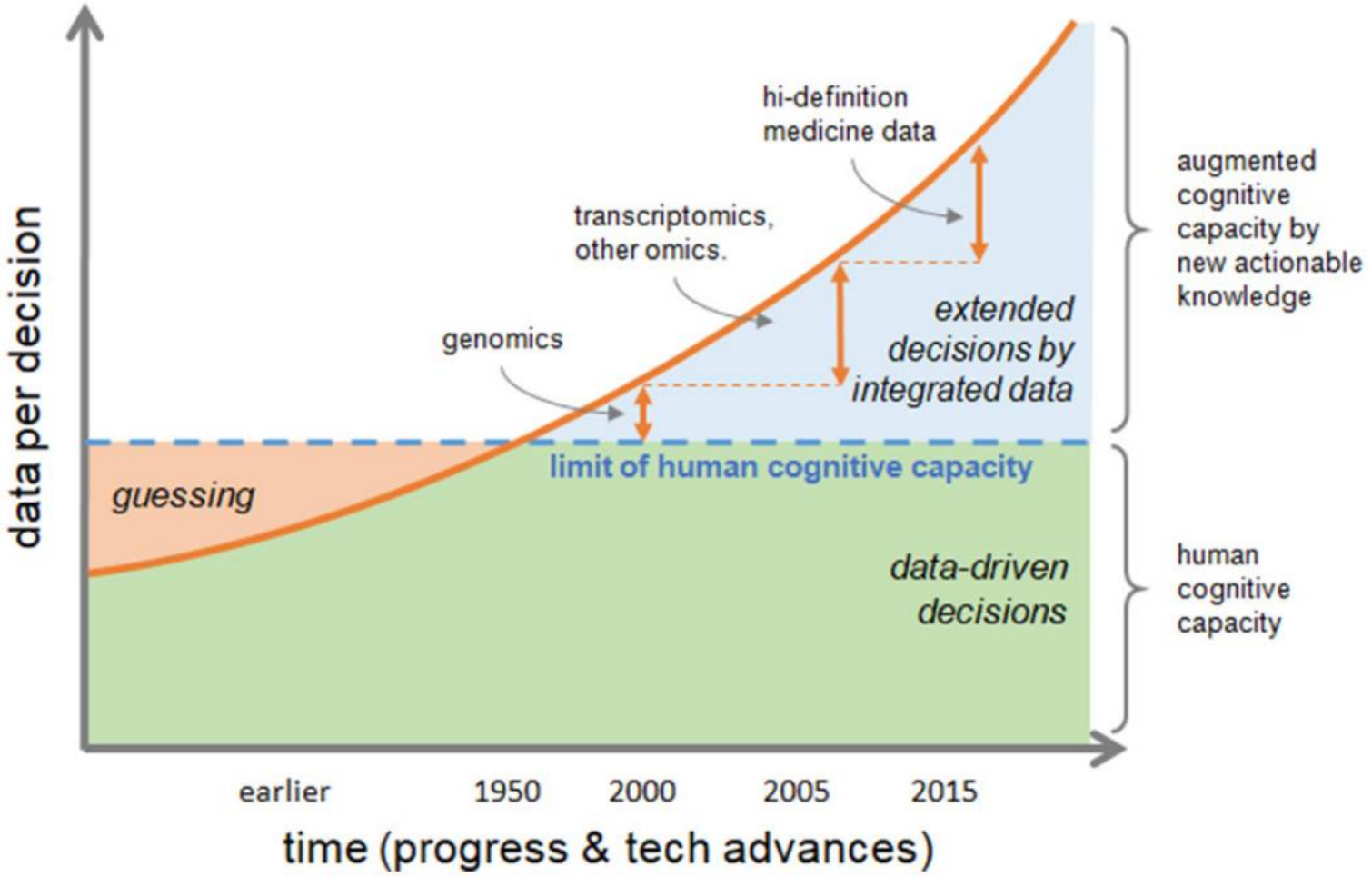
Big data in biomedical field



1 EB = 10^6 TB

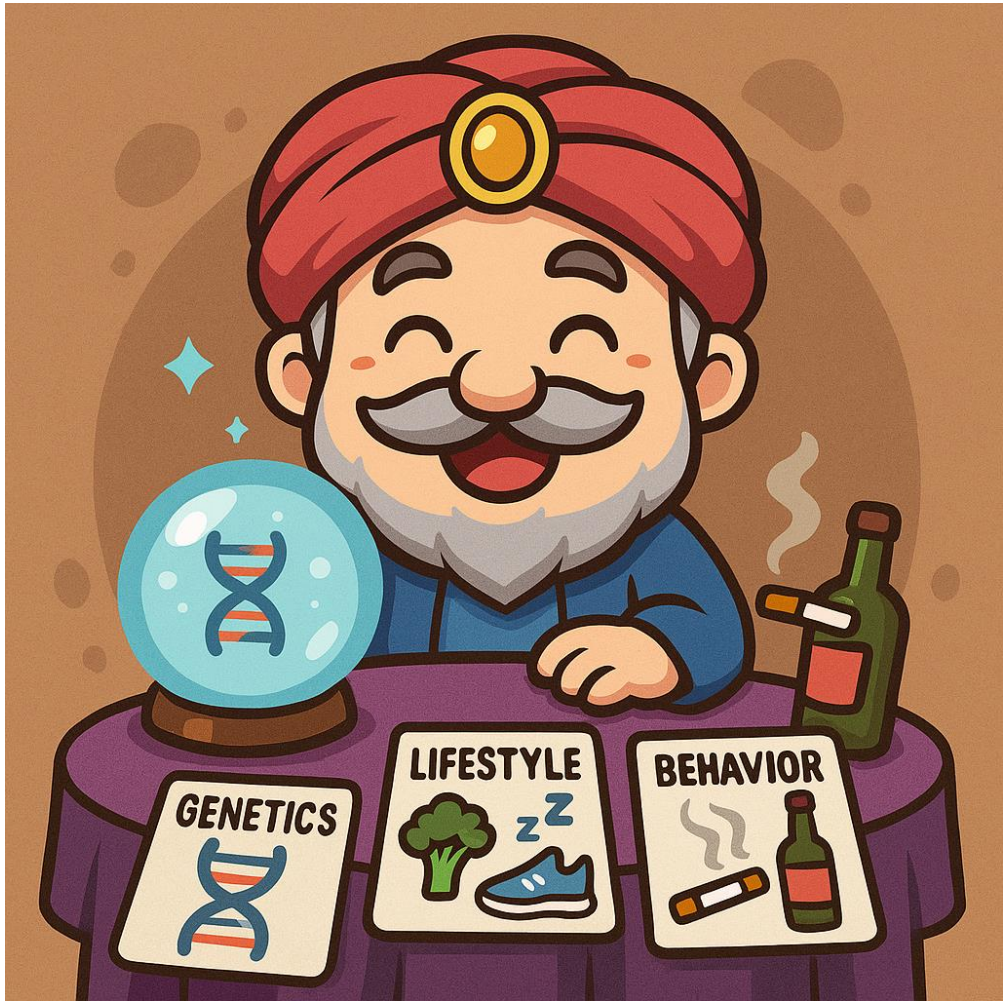
https://www.seekpng.com/ima/u2y3w7q8i1e6a9i1/

From big data to decision



https://www.frontiersin.org/articles/10.3389/fdigh.2018.00013/full

Our role-> Fortune telling



Our goal

Medical and Health Big Data

Medical Data

- ***Complex biomedical data*, inspiring innovative medical device and new service models**
- ***Retrospective biological data* (such as specimens, genetic information), studying new drug and cell therapy targets**

Medical Biological Imaging

- **3D MRI** (150MB)
- **3D CT**(1GB)
- **Mammography**(120 MB)
- **X-ray**(30MB)
- **Pathological slides**(1GB)

Physiological signals/ Lab Values

- **Electrocardiogram**(1MB/month)
- **Blood oxygen**(9KB/month)
- **Blood pressure**(9KB/month)
- **Body temperature**(6KB/month)
- **Blood sugar**(18KB/month)
- **Body weight**(6KB/month)

- **Smart Medical**
  - Image diagnostics
  - Precision health management
  - Personalized early screening
- **Smart Healthcare**
  - Home-based chronic disease care
  - Long-term support management
- **Smart Health**
  - Preventive lifestyle management
  - Enhancing health
- **Biobanks**
  - Precision cancer innovative therapies

Biobanks

- **Biological Database**(31)
- **NHI Research Database**
- **EMR Database**
- **Taiwan Cancer Registry Database**
- **Taiwan Rare Disease Database**

Multiomics Big data

- **Base calling**(5TB)
- **Data source disaggregation**(300GB)
- **Transcriptome assembly of unknown species**(200GB)
- **Sequence alignment/merging**(40GB)
- **Functional annotation(**150GB)
- **Gene differential expression profiling**(2GB)
- **Gene mutation sites**(300GB)

Taiwan Cancer Registry

Taiwan Cancer Registry

- **Nationwide population-based cancer registry (population : 23+ million)**
- **since 1979 (the first nationwide PBCR in Asia)**
- **+100,000 cancer patients per year**
- **Case reporting : 219 hospitals with beds >50 in Taiwan**
- **Legally required to report : Cancer Control Act (enacted 2003)**

衛生福利部國民健康署
Health Promotion Administration, Ministry of Health and Welfare

衛生福利部國民健康署/臺灣大學公共衛生學院
臺灣癌症登記中心
Taiwan Cancer Registry Center

Cancer Registration Items (total 115)

Item name (example)
CASE IDENTIFICATION
Reporting hospital code
Medical record number
Personal identity number
Sex
Date of birth
Residence code
CANCER IDENTIFICATION
Age at diagnosis
Sequence number
Class of case
Date of diagnosis
Primary tumour site (ICD-O)
Laterality
Histology (ICD-O)
Grade – Clin / Path
Diagnostic confirmation
Date of first microscopic confirmation

Item name (example)
STAGE OF DIEASE AT INITIAL DIAGNOSIS
Date of surgical diagnostic and staging procedure
Surgical diagnosis & staging procedure
TNM – Clin T /N /M
TNM – Clin stage group
TNM – Path T /N /M
TNM – Path stage group
Stage (prefix/suffix) descriptor - Clin / Path
AJCC staging edition
Other staging system
Other staging – Clin / Path
Size of tumour
Perineural invasion
Lymph-vascular invasion
Number of examined nodes
Number of positive nodes

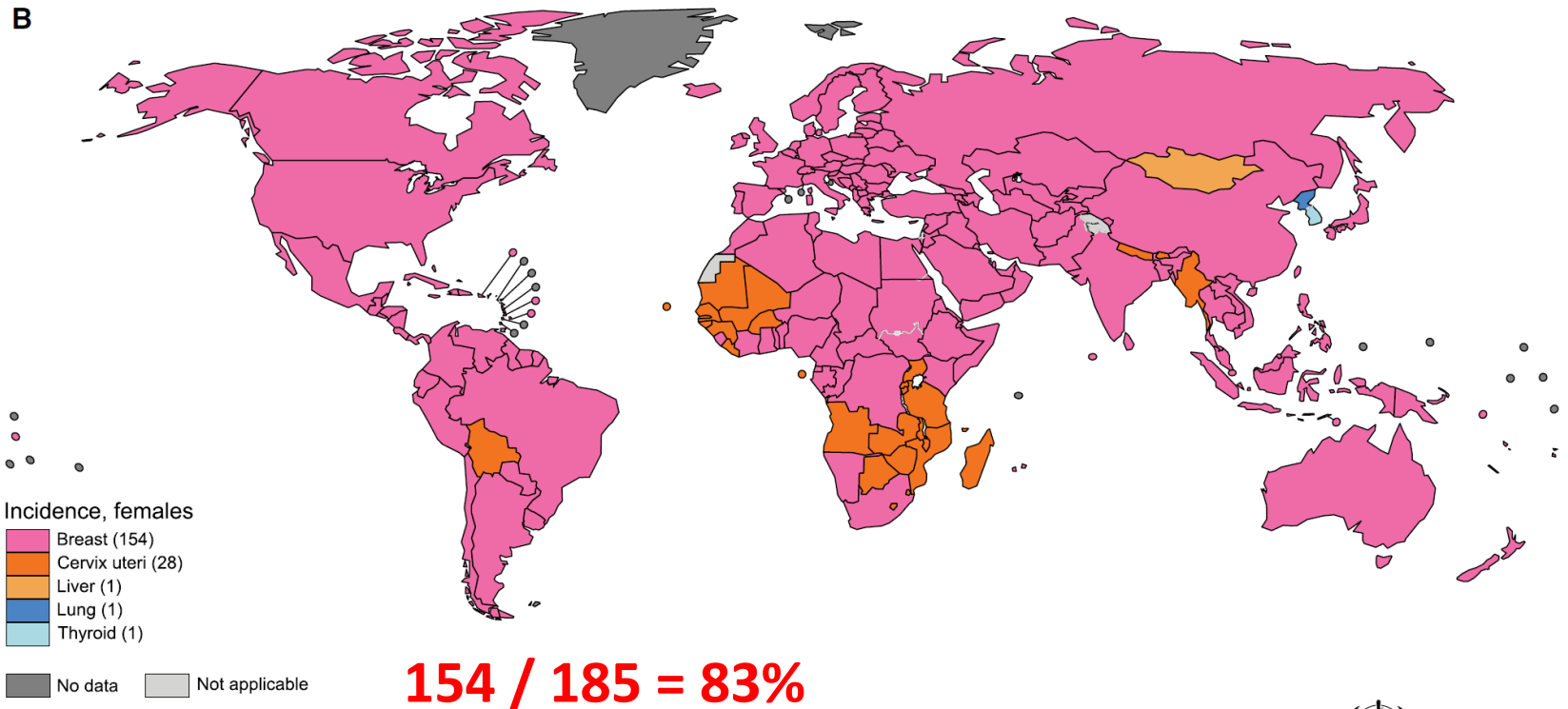
Quality Indicators for Taiwan Cancer Registration Database

Criterion	Year									
	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Completeness,%	97.7	98.4	98.4	98.2	98.2	98.4	98.3	98.2	98.4	98.1
DCO %	0.8	0.7	0.9	0.9	0.8	0.9	0.8	0.8	0.7	0.7
M/I %	44.8	44.3	44.0	43.8	43.5	43.9	41.7	40.9	40.7	41.1
MV %	91.2	91.4	91.8	92.3	92.6	92.8	93.2	93.5	94.1	94.4
Timeliness, month	17	17	17	16	16	14	14	14	14	14

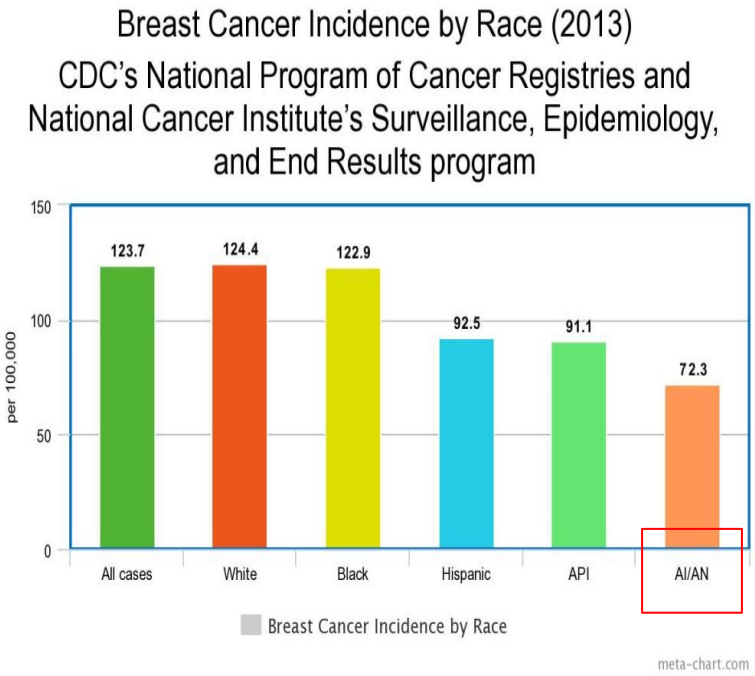
Note: DCO %: death certificate only (DCO) percentage
M/I: mortality verse incidence ratio (only included invasive cancer cases)
MV%: microscopically verified percentage
Timeliness: duration from the date of diagnosis to the date when the database close

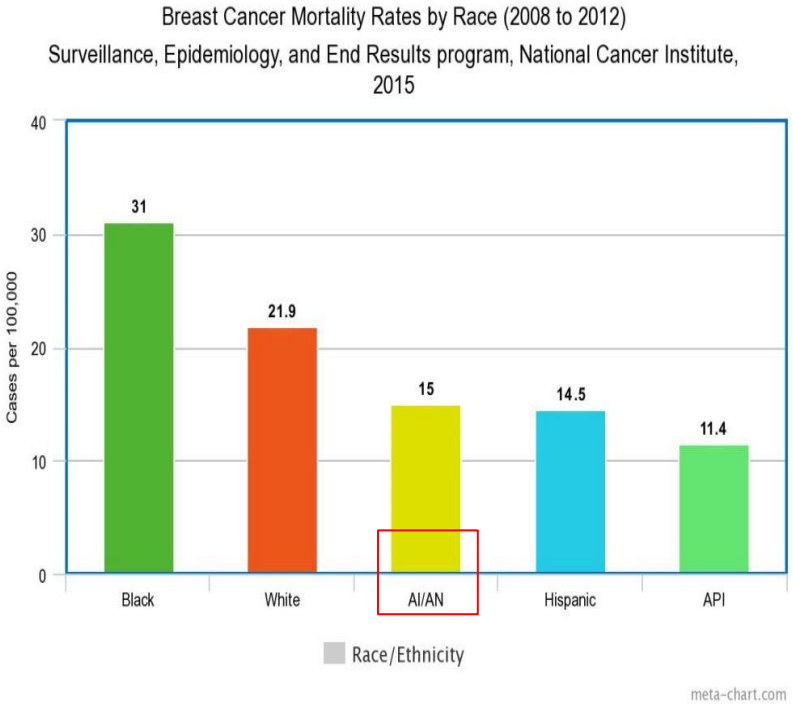
2021 annual TCR report

Most common type of cancer incidence in 2018 among women

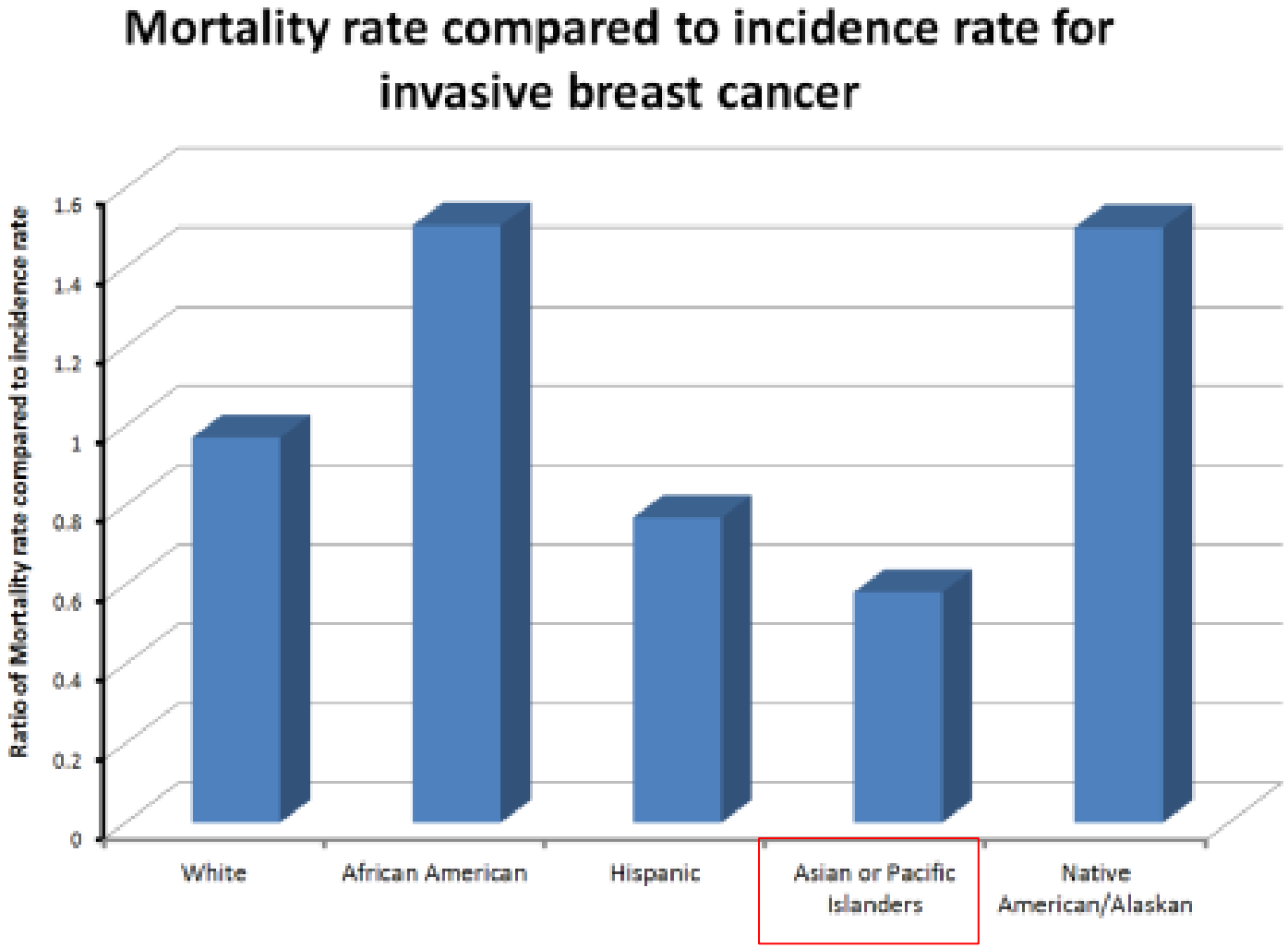


Incidence and mortality rates of breast cancer patients in different races





Ratio of mortality rate to incidence rate for breast cancer



Study aims

- To develop the prognostic models to predict the breast caner-specific survival and overall survival for breast cancer patients

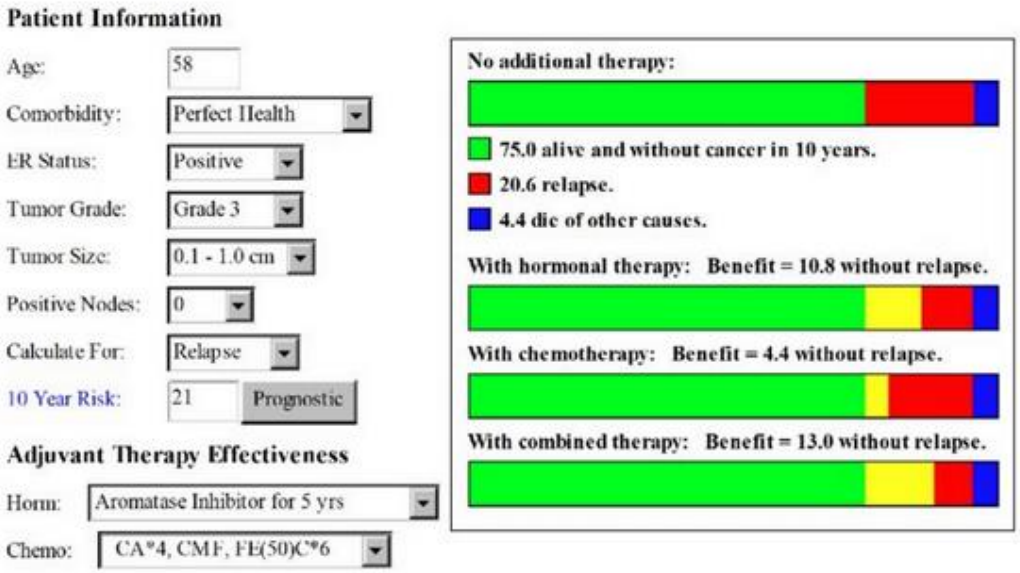
- To investigate the racial difference between the Asian population and non-Asian population

Adjuvant online

Imagine the case:

40 years old, chemobenefit : 3%

Receive treatment or not ?



Figure

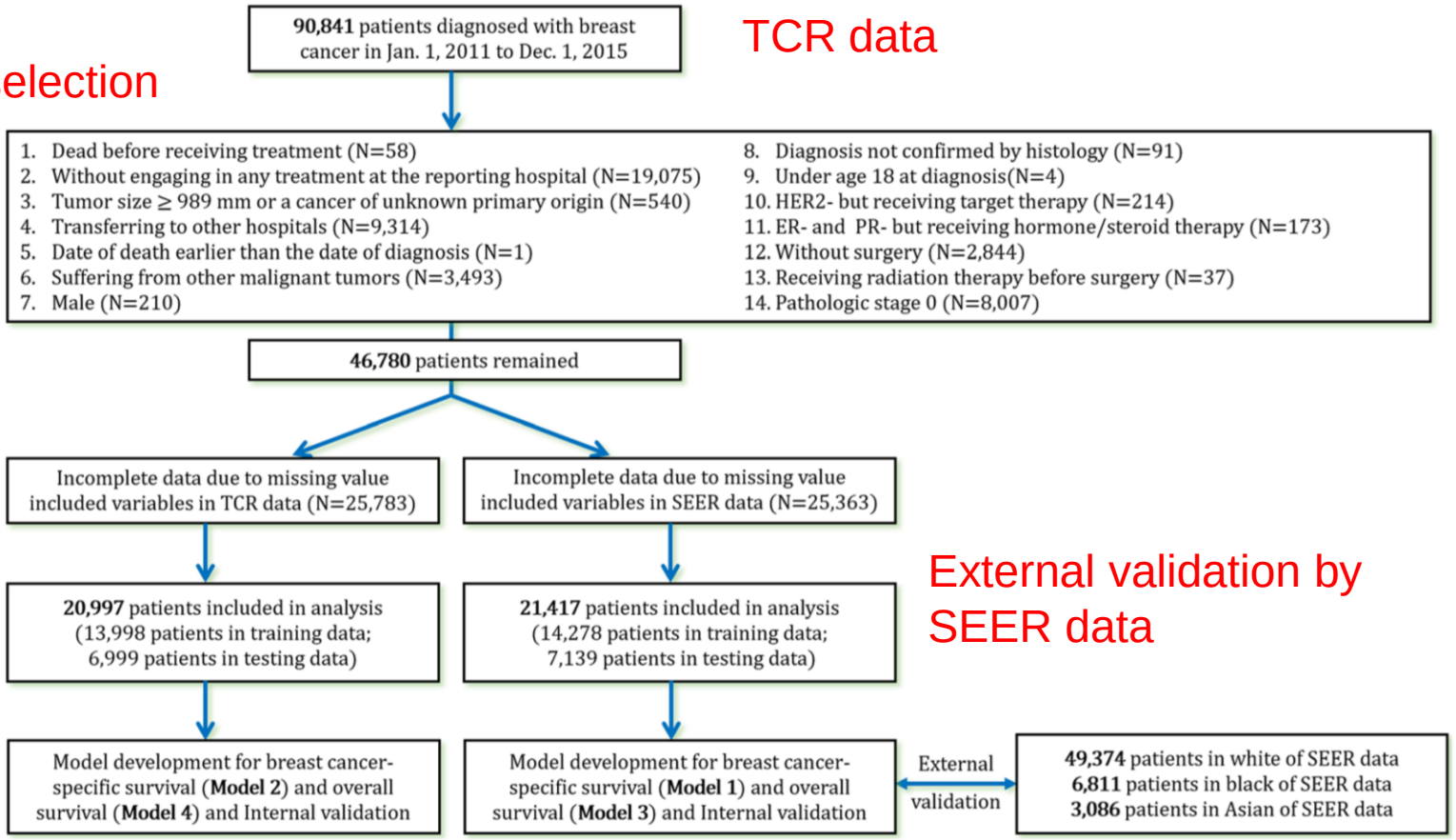
Caption

Figure 1. Prognostic data based on population-based analysis with Adjuvant! Online version 8.0 [4].

Analysis flow chart

Patients selection

TCR data



External validation by SEER data

Fig. 1 Flow diagram of exclusion criteria and statistical analysis

1. Cox proportional hazard model
2. Multivariable fractional polynomial (MFP) approach

Calibration of prediction performances

Table 4 Model calibration of breast cancer-specific mortality regressed on prognostic variables in SEER data

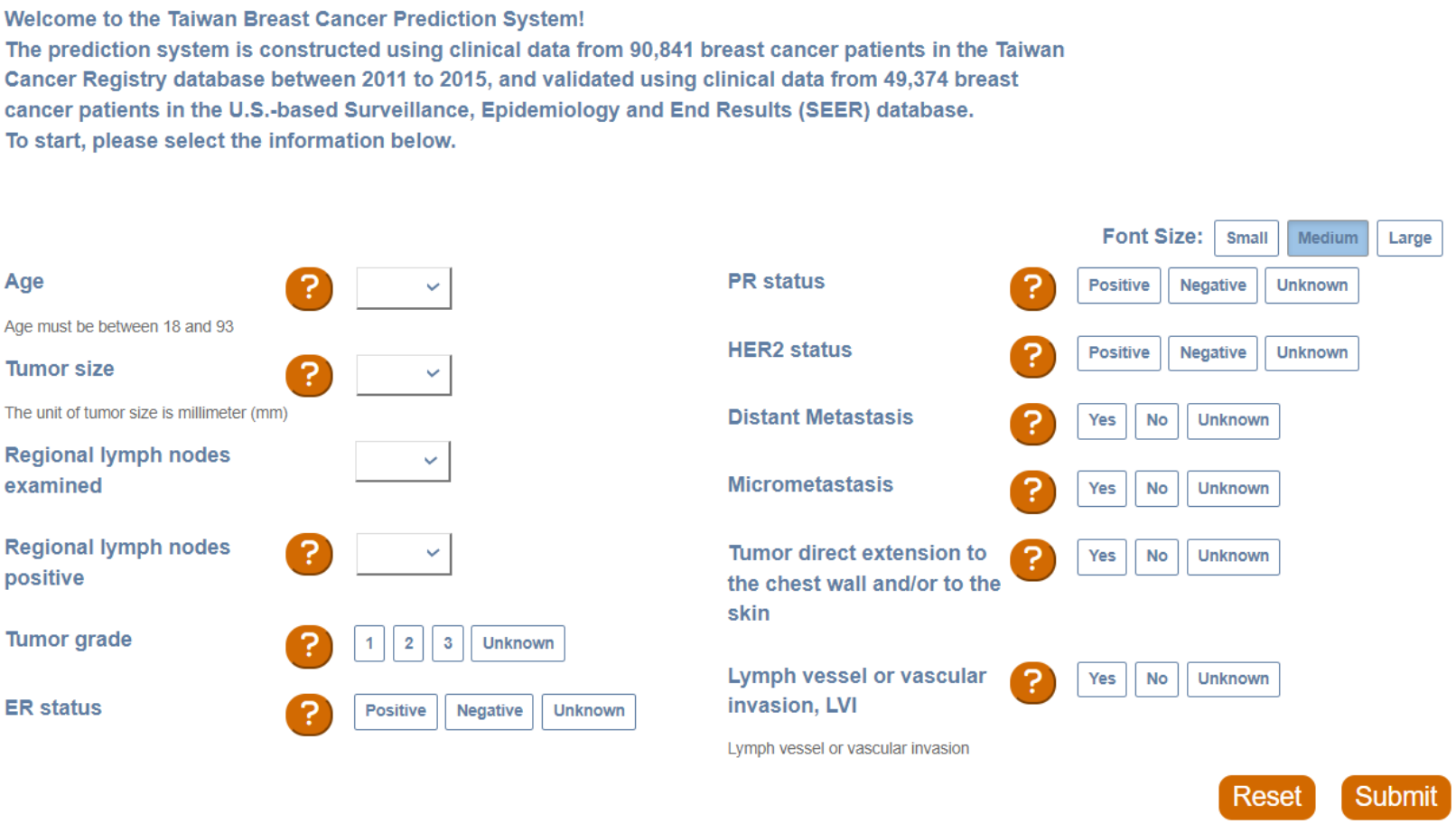
	Breast cancer-specific survival			Overall survival		
Calibration year	Observed	Predicted	*P* value	Observed	Predicted	*P* value
Training data						
1	51	52.64	0.821	83	83.87	0.924
2	229	236.75	0.615	316	317.83	0.918
3	363	373.45	0.589	481	475.71	0.808
4	425	422.94	0.920	569	544.60	0.296
5	348	350.04	0.913	466	456.30	0.650
6	220	215.38	0.753	290	275.41	0.379
Testing data						
1	33	25.28	0.125	47	41.21	0.367
2	109	114.89	0.583	142	157.12	0.228
3	177	182.81	0.668	237	237.48	0.975
4	202	206.35	0.762	266	270.31	0.793
5	176	172.28	0.777	227	226.75	0.987
6	96	104.32	0.416	129	133.50	0.697

Calibration of prediction performances in different populations

Table 4 Model calibration of breast cancer-specific mortality regressed on prognostic variables in SEER data

	Breast cancer-specific survival			Overall survival		
Calibration year	Observed	Predicted	*P* value	Observed	Predicted	*P* value
White						
1	236	132.35	< 0.001	416	245.11	< 0.001
2	472	432.07	0.055	805	666.75	< 0.001
3	554	570.83	0.481	905	834.99	0.015
4	378	426.84	0.018	617	628.31	0.652
5	–	–	–	–	–	–
6	–	–	–	–	–	–
Black						
1	47	29.32	0.001	87	43.98	< 0.001
2	132	90.05	< 0.001	194	115.66	< 0.001
3	138	115.04	0.032	195	141.00	< 0.001
4	99	86.07	0.164	146	106.57	< 0.001
5	–	–	–	–	–	–
6	–	–	–	–	–	–
Asian						
1	10	7.57	0.377	23	13.93	0.015
2	22	24.51	0.612	43	38.03	0.421
3	21	29.84	0.106	43	44.16	0.862
4	17	23.31	0.191	29	33.42	0.445
5	–	–	–	–	–	–
6	–	–	–	–	–	–

Breast cancer survival prediction by Taiwan Cancer Registry data[1]



https://preparetaiwan.org/zh_tw/prediction/breast_cancer

Breast cancer survival prediction by Taiwan Cancer Registry data[2]

Treatment options

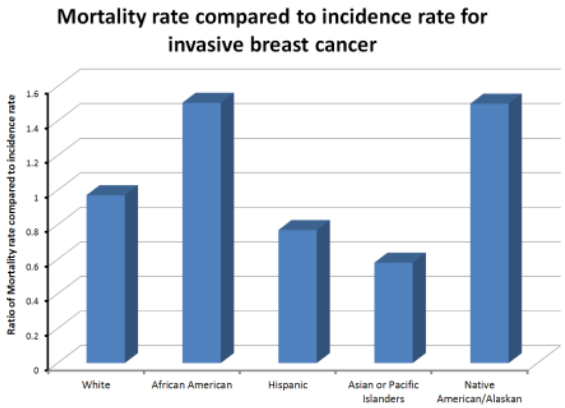
Hormone/Steroid therapy: No | Yes

Hormone/ steroid therapy is available when ER status is positive

Chemotherapy: No | Yes

Radiotherapy: No | Yes

Targeted therapy: No | Yes



Results

Please note that the patients need to consult with their medical doctors before making any decision.

Table | Texts

The analysis is for women who had undergone surgery.The table shows the 1-, 3- and 5-year survival rates,based on the treatment you have selected.

1 | 3 | 5 | 7

Treatment	Additional Benefit	Overall Survival(%)
Surgery only	-	33.72%
+Hormone/Steroid therapy	27.27%	60.99%
+Chemotherapy	11.46%	72.45%
+Radiotherapy	6.72%	79.17%
+Targeted therapy	7.79%	86.96%

https://preparetaiwan.org/zh_tw/prediction/breast_cancer

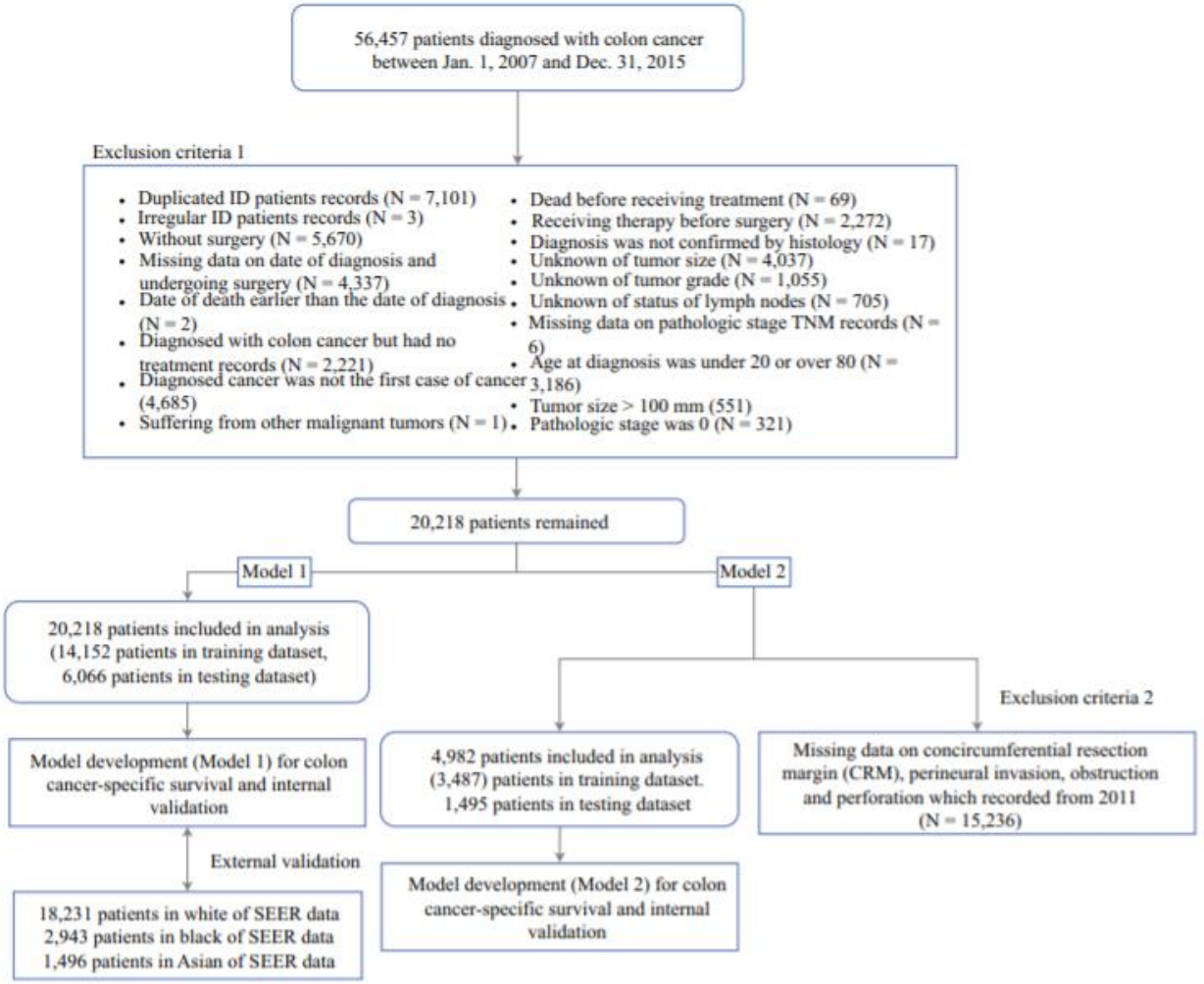
Colon cancer

Ann Surg Oncol
https://doi.org/10.1245/s10434-021-1064

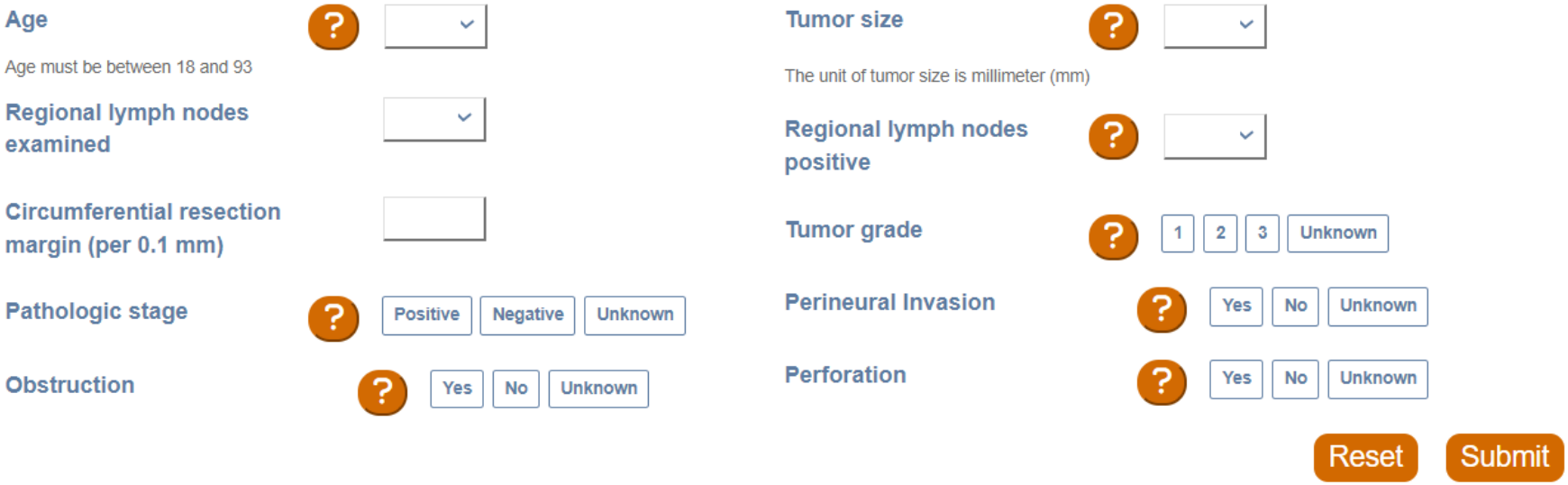
ORIGINAL ARTICLE – GLC

**Predicting Colon Ca
Population Using Na**

**Han-Ching Chan, MS[1], Chi-Che
Kuan-Hung Yeh, BS[2], Wen-Chu
Skye Hung-Chun Cheng, MD[5],**



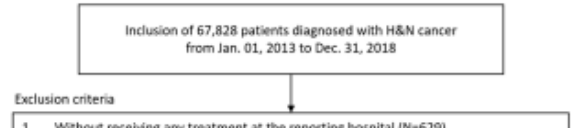
Colon cancer survival prediction by Taiwan Cancer Registry data



Head and neck cancer

Journal of Epidemiology and Global Health
https://doi.org/10.1007/s44197-024-00196-7

RESEARCH ARTICLE



Population-Based Prognostic Models fc

Welcome to the Taiwan Head and Neck Cancer Prediction System!
The prediction system is constructed using clinical data from 49,137 head and neck cancer patients in the Taiwan Cancer Registry database between 2013 to 2018.
To start, please select the information below.

Font Size: Small | Medium | Large

Age
Age must be between 20 and 80

Tumor size
The unit of tumor size is millimeter (mm)

Height(cm)

Weight(kg)

Pathological stage
1 | 2 | 3 | 4 | Unknown

Sex
Male | Female

Tsumor site
Oral Cavity | Oropharynx | Hypopharynx | Salivary Glands | Larynx

Maximum lymph node diameter
No Metastasis | 1-9 mm | 10-19 mm | 20-29 mm | 30-39 mm | 40-49 mm | 50-59 mm | >= 60 mm | Unknown

Alcohol consumption
Never | Sometime | Usually

Tobacco smoking
false | true

Reset | Submit

44,410 White patients in the SEER data
4,588 Black patients in the SEER data
2,001 Asian patients in the SEER data

Ovarian cancer

ORIG

Pre
Car
A R

Amrita
Wen-C

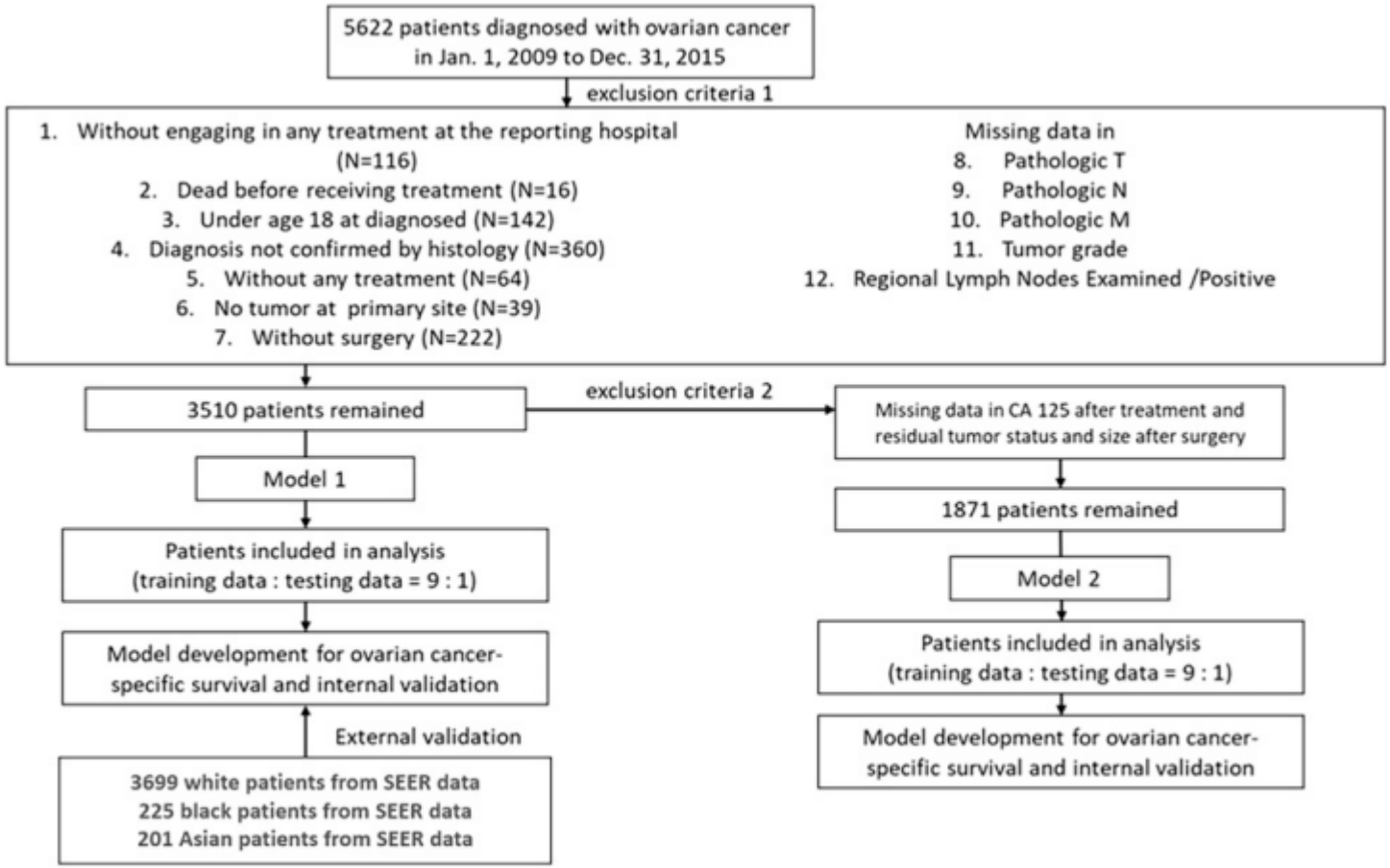
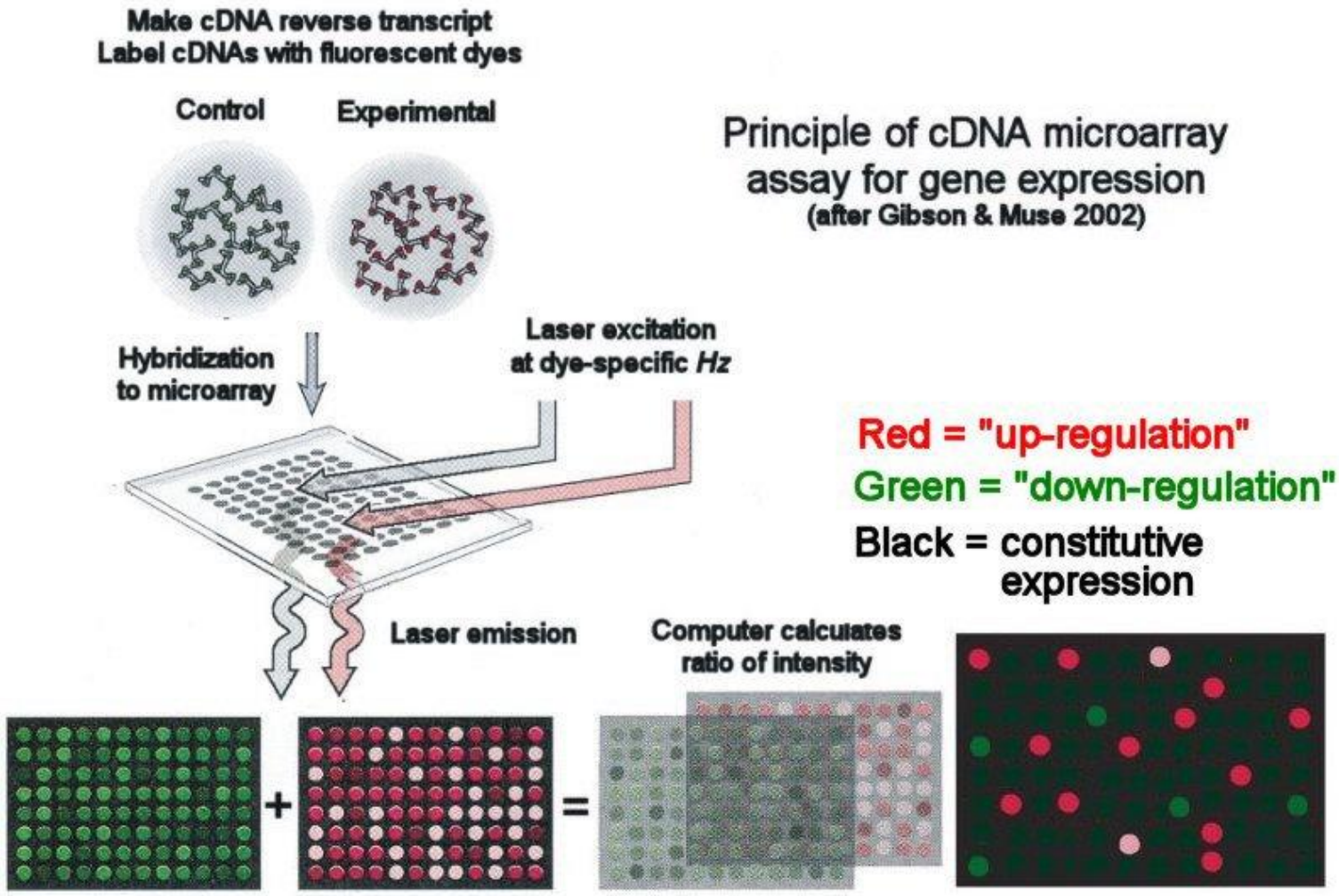


FIG. 1. Inclusions and exclusion criteria for study subjects.

Can we do better?

- In addition to clinical variables, what can we have?
  - Genetics (DNA)
  - Gene expression (RNA)

Microarray



http://thunder.biosci.umbc.edu/classes/biol414/spring2007/index.php

Gene expression microarray data

Samples

Probes

	ID	GSM5401(	GSM5401(	GSM54011	GSM54011	GSM54011	GSM54011	GSM54011
2	1007_s_at	12.61478	11.84253	12.52451	12.55463	12.9567	12.24632	13.00839
3	1053_at	7.903758	8.143369	8.830648	9.397681	8.765905	8.194851	8.622835
4	117_at	7.983277	8.125691	7.849639	7.965634	8.694988	8.272472	8.231823
5	121_at	10.00636	10.05648	10.46377	9.743123	9.803373	9.160232	9.037198
6	1255_g_at	1.52182	2.766601	3.275003	2.216496	2.06613	2.642343	1.766505
7	1294_at	10.05307	9.986836	11.6857	9.058527	9.671435	10.54355	10.07772
8	1316_at	8.796633	9.524082	6.998067	6.309369	6.489771	6.672373	6.020401
9	1320_at	5.952997	6.743085	6.398354	6.88609	6.493907	6.956016	6.456581
10	1405_i_at	9.002697	9.904427	12.43649	10.83128	8.803543	10.77742	11.47862

Affymetrix u133plus2.0 : 54675 probes

PAM50

J Clin Oncol . 2009 Mar 10;27(8):1160-7. doi: 10.1200/JCO.2008.18.1370.

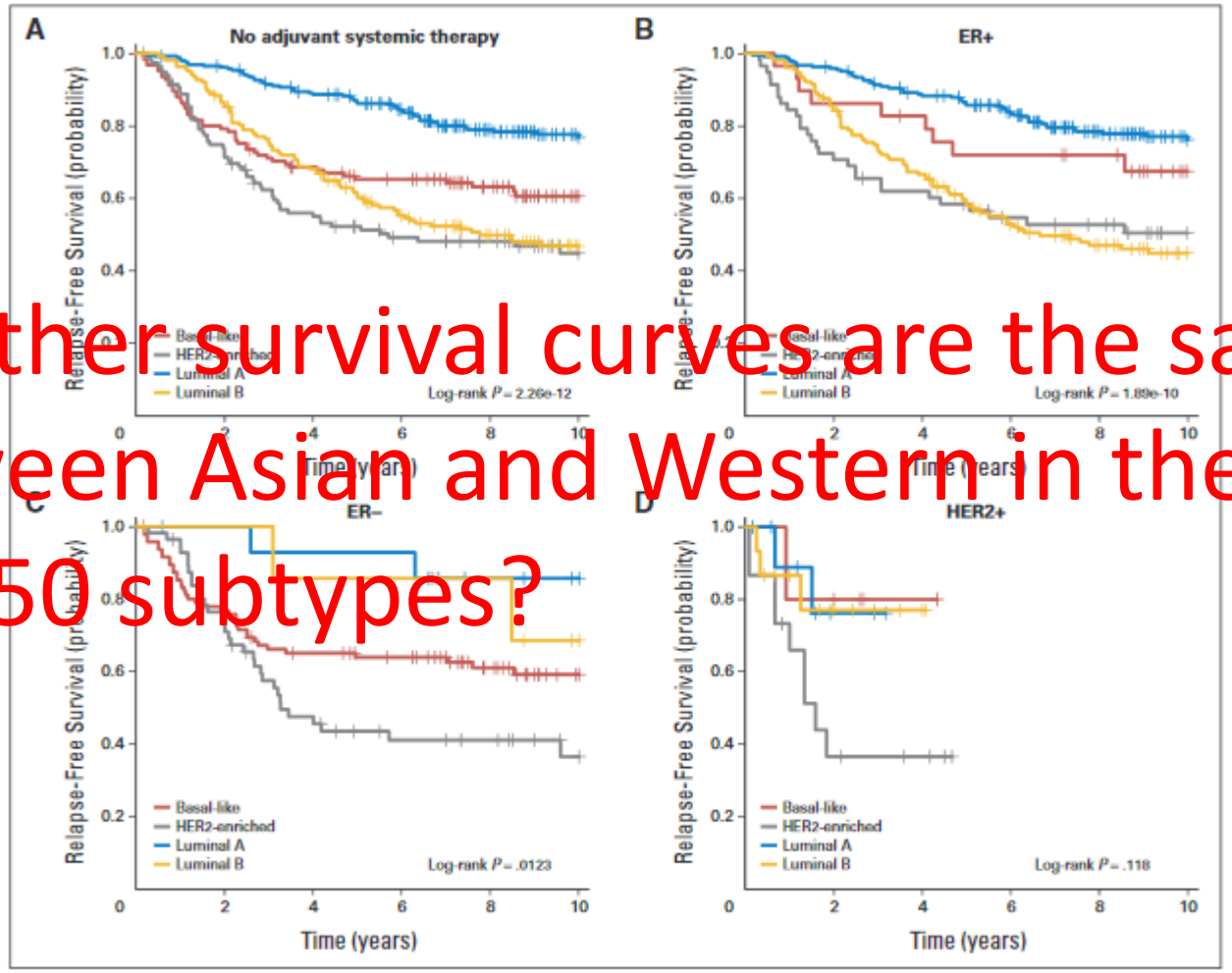


Fig 1. PAM50 intrinsic subtype prognosis for relapse-free survival (RFS). (A) Outcome predictions according to the four tumor subtypes in a test set of 710 node-negative, no systemic adjuvant therapy patients. (B) Outcome by subtype in the subset of patients with estrogen receptor (ER) –positive disease from Figure 1A. (C) Outcome by subtype in patients with ER-negative disease. (D) Outcome by subtype in HER2*clin*-positive patients.

Whether survival curves are the same between Asian and Western in the PAM50 subtypes?

GEO datasets in this study

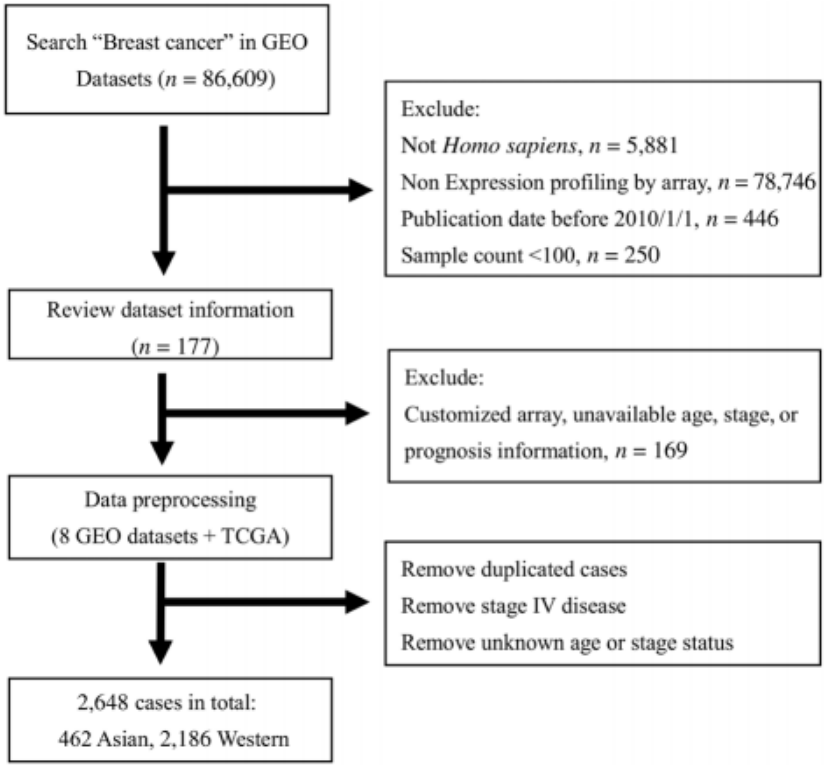


Figure 1. Study flow diagram. By applying searching filter, 177 studies were found. A total of eight data sets met selection criteria in further evaluation.
Abbreviations: GEO, Gene Expression Omnibus; TCGA, The Cancer Genome Atlas.

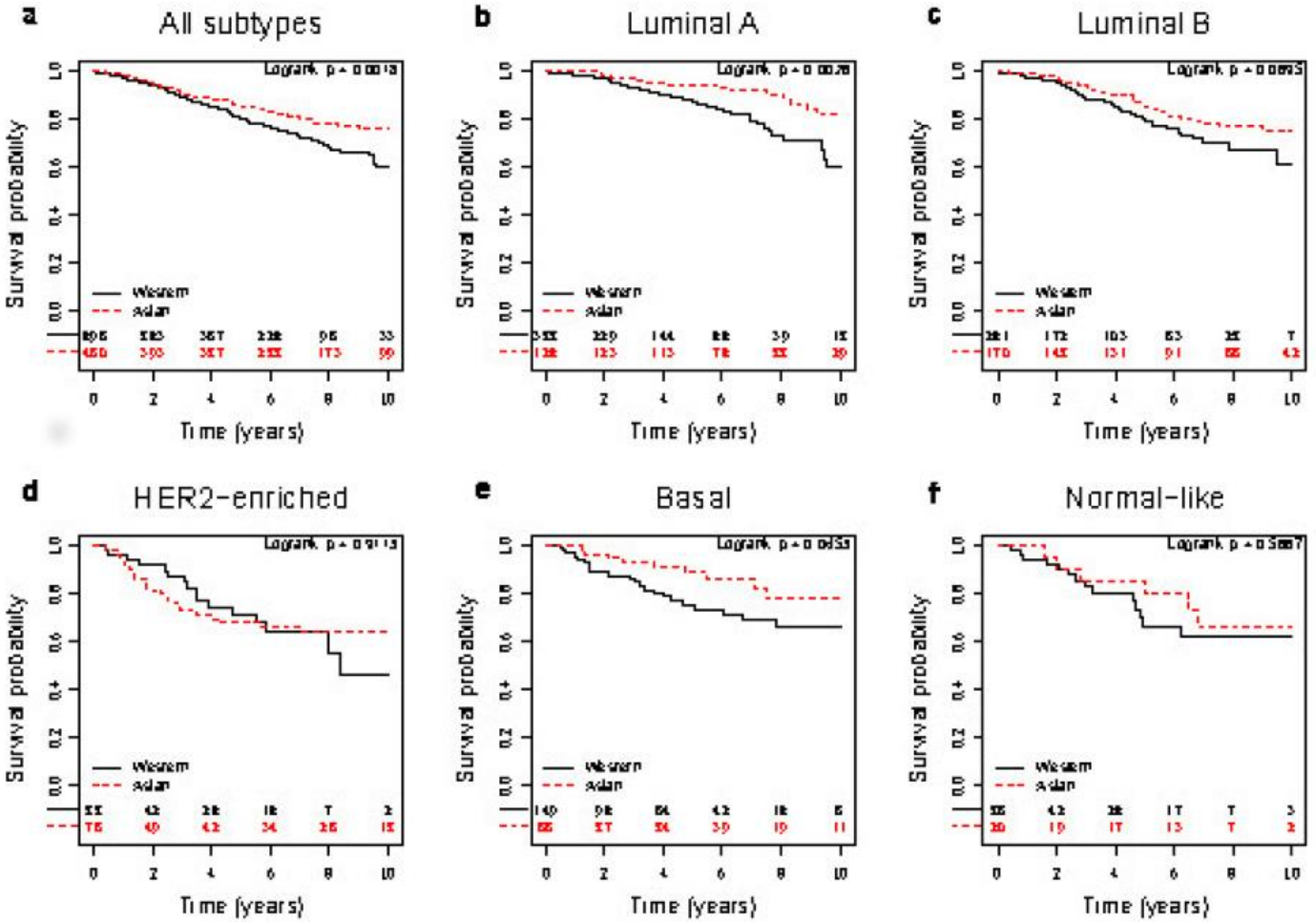
Sample characteristics

Table 1. Comparison of Estimation of Stromal and Immune Cells in Malignant Tumours Using Expression Data (ESTIMATE) immune scores of breast tumors between Asian and Western patients

Tumor subtype	Asian patients	Western patients	*p* value
All subtypes	1,187	1,014	<.0001
Luminal A	958	866	.2009
Luminal B	1,091	823	.0001
HER2-enriched	1,513	1,275	.0593
Basal	1,477	1,381	.4242
Normal-like	1,317	1,187	.5053

Abbreviation: HER2, human epidermal growth factor receptor 2.

Overall survival curves between Asian and Western populations

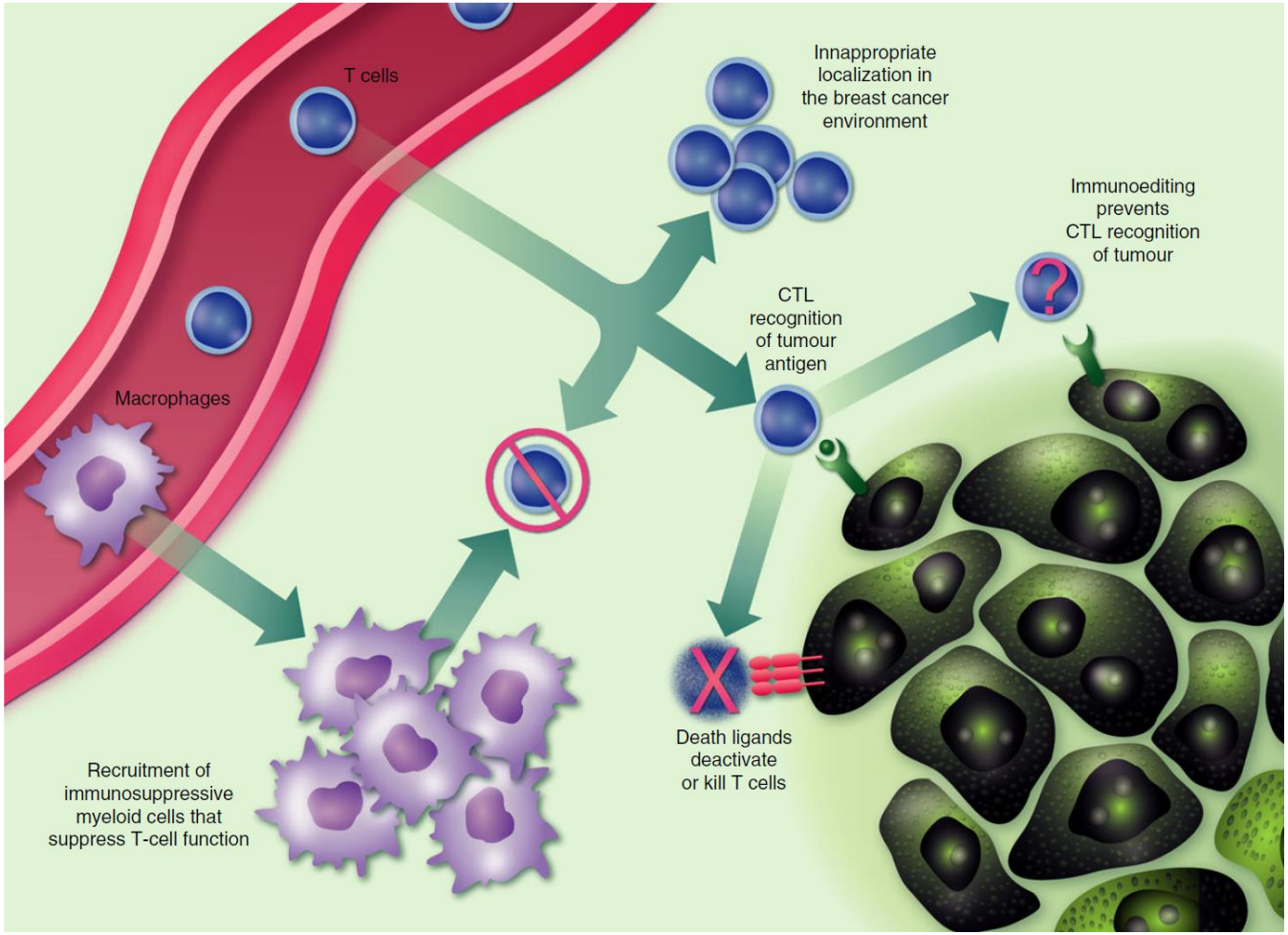


Study aims

- To further divide the breast cancer samples based on PAM50 subtypes into different subgroups based on the tumor infiltration lymphocytes (TILs)

- To elucidate the prognostic impact of TILs in breast cancer patients between Asian and Western populations

Tumor infiltrating lymphocytes (TILs)

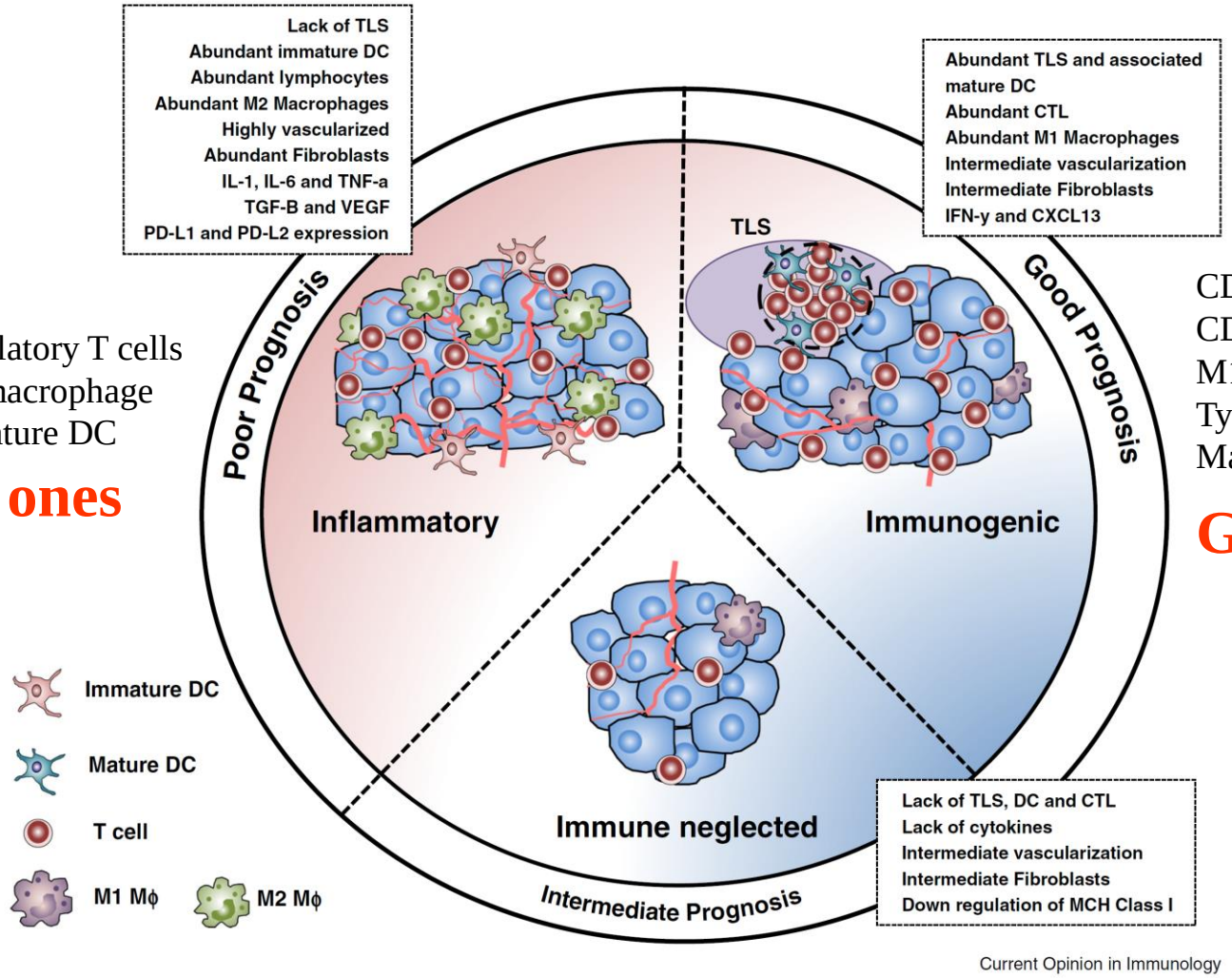


Immunology, 150, 25–34

Impact of TILs in prognosis in cancer patients

Regulatory T cells
M2 macrophage
Immature DC

Bad ones



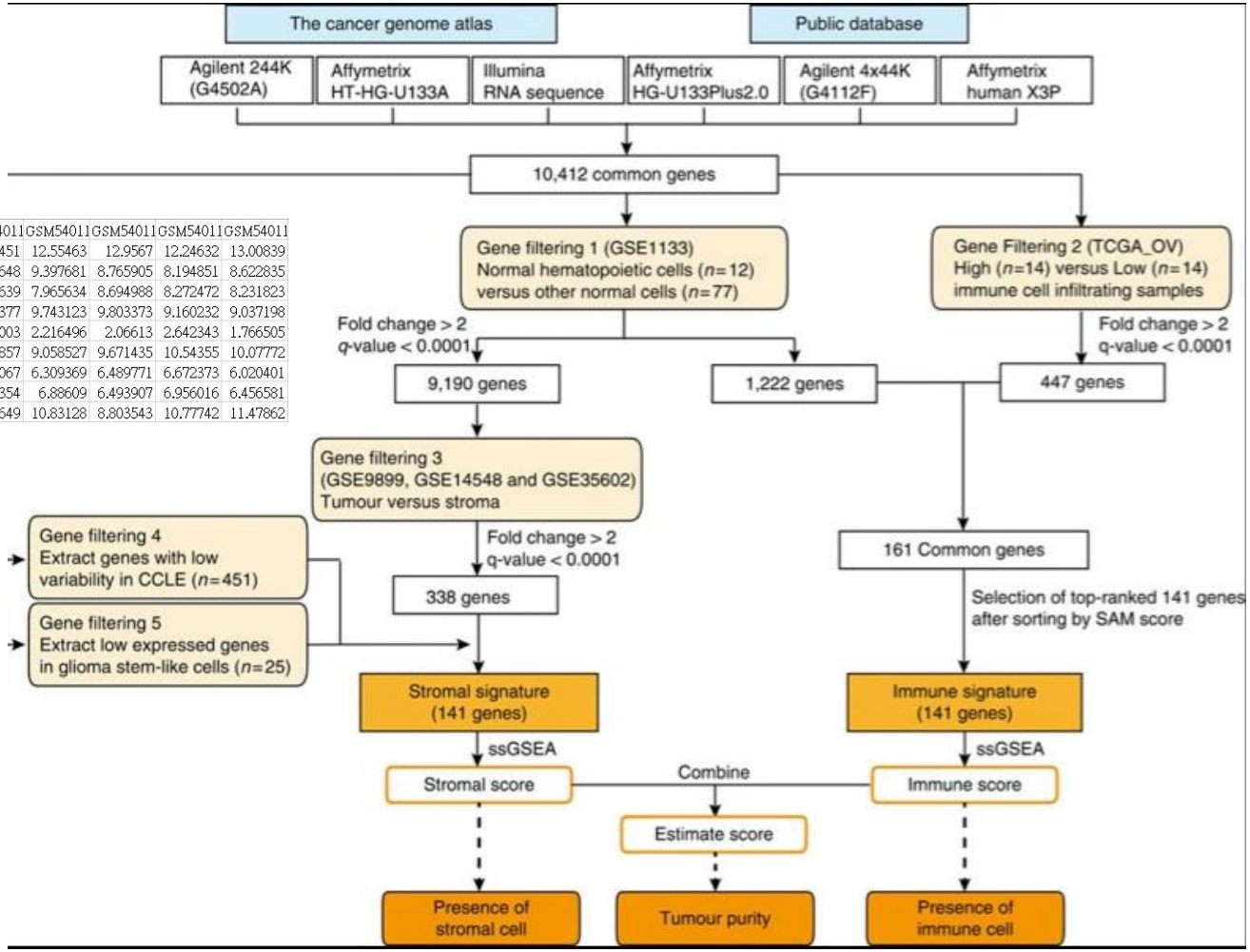
CD4+ T cells
CD8+ T cells
M1 macrophage
Type 1 helper T cells
Mature DC

Good ones

Current Opinion in Immunology 2016, 39:7–13

ESTIMATE algorithm

	ID	GSM5401(	GSM5401(	GSM54011	GSM54011	GSM54011	GSM54011	GSM54011
2	1007_s_at	12.61478	11.84253	12.52451	12.55463	12.9567	12.24632	13.00839
3	1053_at	7.903758	8.143369	8.830648	9.397681	8.765905	8.194851	8.622835
4	117_at	7.983277	8.125691	7.849639	7.965634	8.694988	8.272472	8.231823
5	121_at	10.00636	10.05648	10.46377	9.743123	9.803373	9.160232	9.037198
6	1255_g_at	1.52182	2.766601	3.275003	2.216496	2.06613	2.642343	1.766505
7	1294_at	10.05307	9.986836	11.6857	9.058527	9.671435	10.54355	10.07772
8	1316_at	8.796633	9.524082	6.998067	6.309369	6.489771	6.672373	6.020401
9	1320_at	5.952997	6.743085	6.398354	6.88609	6.493907	6.956016	6.456581
10	1405_i_at	9.002697	9.904427	12.43649	10.83128	8.803543	10.77742	11.47862



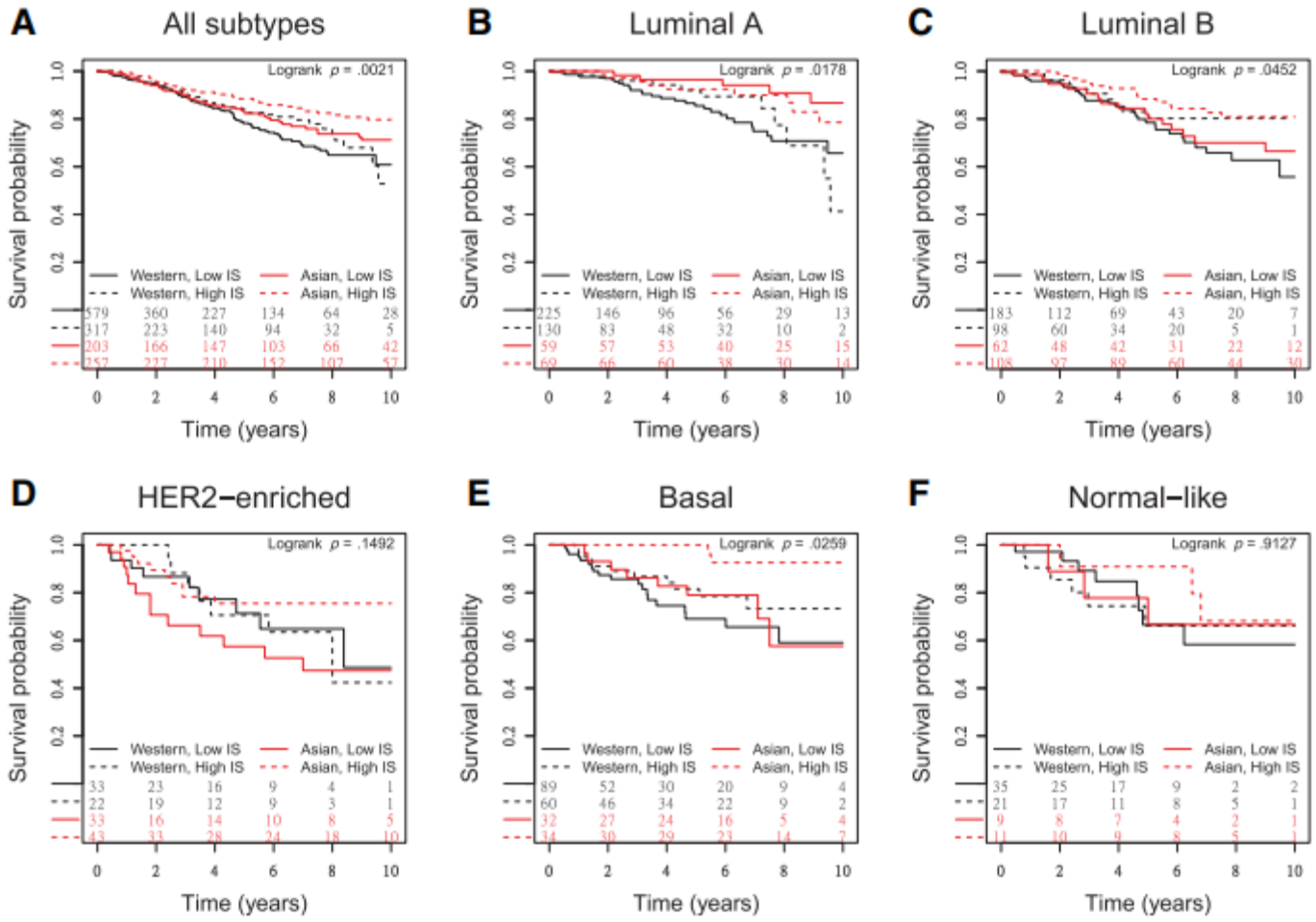
Immune cell scores in different PAM50 subtypes

Table 2. Associations of Estimation of Stromal and Immune Cells in Malignant Tumours Using Expression Data (ESTIMATE) immune score (high versus low) with OS and DFS, stratified by race and PAM50 subtype (adjusted for age and stage)

	OS		DFS	
Race, PAM50 subtype	**HR (95% CI)**	***p* value**	**HR (95% CI)**	***p* value**
Asian				
All subtypes[a]	0.45 (0.28–0.71)	<.0001	0.51 (0.34–0.77)	.0014
Luminal A	1.34 (0.46–3.91)	.5979	0.83 (0.35–1.95)	.6713
Luminal B	0.51 (0.24–1.09)	.0831	0.76 (0.39–1.48)	.4200
HER2-enriched	0.53 (0.21–1.39)	.1995	0.44 (0.19–1.04)	.0610
Basal	0.05 (0.01–0.37)	.0037	0.06 (0.01–0.44)	.0056
Normal-like	0.46 (0.03–7.98)	.5925	NR[b]	
Western				
All subtypes[a]	0.64 (0.44–0.94)	.0206	0.75 (0.60–0.92)	.0061
Luminal A	0.80 (0.39–1.61)	.5273	2.09 (1.11–3.94)	.0226
Luminal B	0.66 (0.29–1.52)	.3304	0.92 (0.65–1.29)	.6245
HER2-enriched	0.66 (0.23–1.94)	.4544	0.40 (0.24–0.69)	.0010
Basal	0.58 (0.28–1.22)	.1532	0.60 (0.42–0.85)	.0040
Normal-like	0.61 (0.20–1.84)	.3781	1.33 (0.38–4.65)	.6525

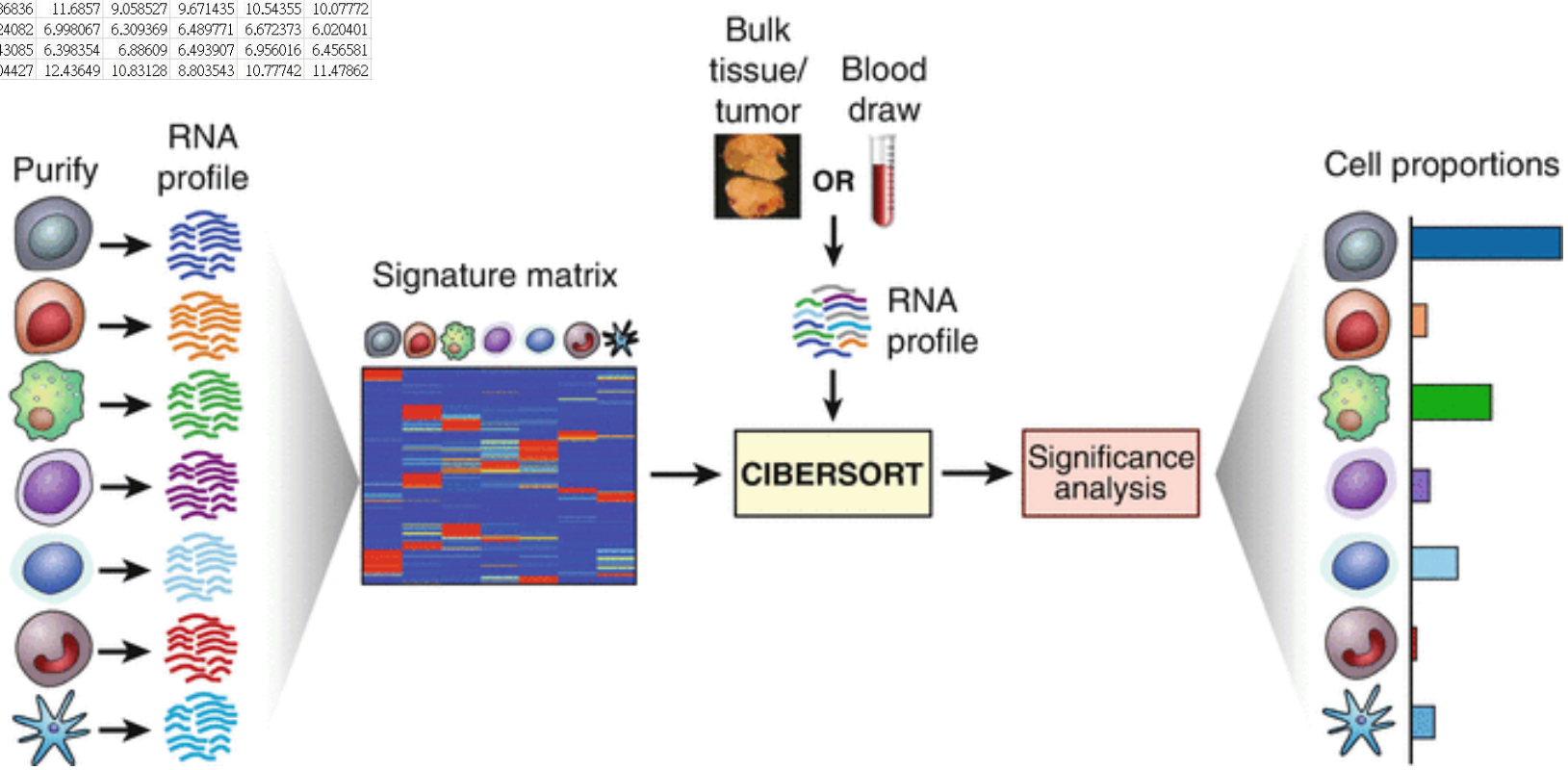
[a]Additional adjustment for PAM50 subtype.
[b]Cox regression failed because of collinearity.
Abbreviations: CI, confidence interval; DFS, disease-free survival; HER2, human epidermal growth factor receptor 2; HR, hazard ratio; NR, not reached; OS, overall survival; PAM50, Prediction Analysis of Microarray 50.

KM curves further divided by the abundances of TILs



CIBERSORT algorithm

	ID	GSM5401(	GSM5401(	GSM5401	GSM5401	GSM5401	GSM5401	GSM5401
2	1007_s_at	12.61478	11.84253	12.52451	12.55463	12.9567	12.24632	13.00839
3	1053_at	7.903758	8.143369	8.830648	9.397681	8.765905	8.194851	8.622835
4	117_at	7.983277	8.125691	7.849639	7.965634	8.694988	8.272472	8.231823
5	121_at	10.00636	10.05648	10.46377	9.743123	9.803373	9.160232	9.037198
6	1255_g_at	1.52182	2.766601	3.275003	2.216496	2.06613	2.642343	1.766505
7	1294_at	10.05307	9.986836	11.6857	9.058527	9.671435	10.54355	10.07772
8	1316_at	8.796633	9.524082	6.998067	6.309369	6.489771	6.672373	6.020401
9	1320_at	5.952997	6.743085	6.398354	6.88609	6.493907	6.956016	6.456581
10	1405_i_at	9.002697	9.904427	12.43649	10.83128	8.803543	10.77742	11.47862



Nature methods | VOL.12 NO.5 | MAY 2015 | 453

Impact of prognosis in breast cancer patients between Asian and Western populations

Asian **Western**

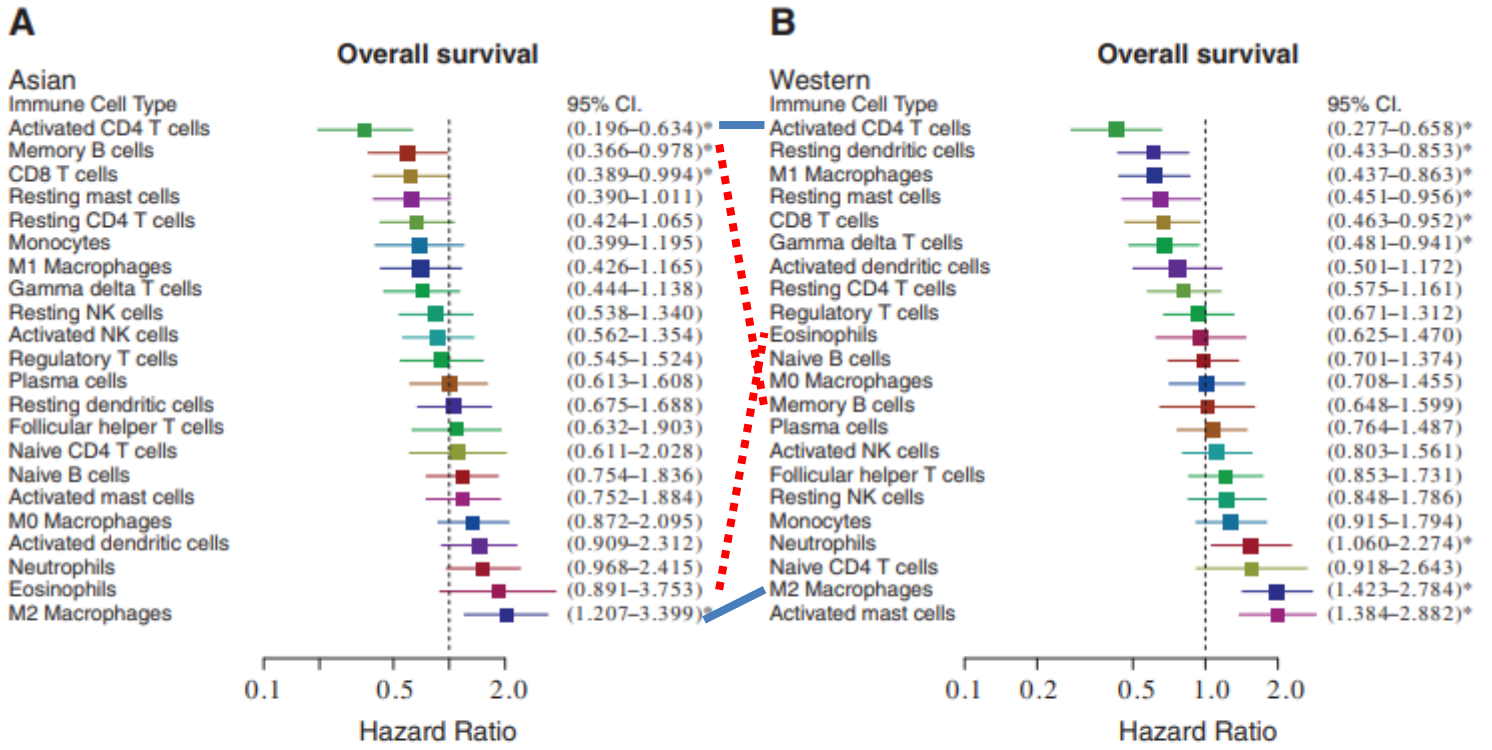


Figure 4. Prognostic effects of individual immune cell subsets on overall survival. **(A):** Effects of cell subsets on Asian patients. Adjusted for age, stage, and Prediction Analysis of Microarray 50 (PAM50) subtype, activated CD4 T cells, memory B cells, and CD8 T cells were associated with better prognosis, whereas M2 macrophage was associated with worse prognosis. **(B):** Effects of cell subsets on Western patients. The effects of activated CD4 T cells, CD8 T cells, and M2 macrophages were similar in Western patients with breast cancer. * indicates $p < .05$.
Abbreviations: CI, confidence interval; NK, natural killer.

Ovarian cancer

	New case	Death
2020 Global	**313,959**	**207,252**
2019 Taiwan	**1,559**	**683**

Stage	Taiwan	USA (SEER)
I	**43.60%**	**22%**
II	**8.20%**	**8%**
III	**25.80%**	**37%**
IV	**12.40%**	**28%**
Unknown	**10.00%**	**6%**

5-year survival for all stages

- Taiwan: 63.68%
- SEER (USA): 56%

Taiwan cancer registry (TCR) and Global Cancer Statistics 2020

Chemotherapy NCCN guideline in OV

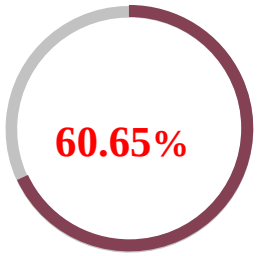
- Stage IA or Stage IB
  - Grade 1
  - Grade 2
- Else Stage I
- Stage II to Stage IV

check

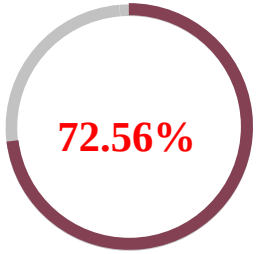
chemotherapy

Taiwan

Receiving chemotherapy



Stage I



All Stages

NCCN guideline clinical practice guideline of gynecological oncology

Development of a GE model

After surgery

Predicted by a model

Gene expression

Follow up

Low risk

High risk

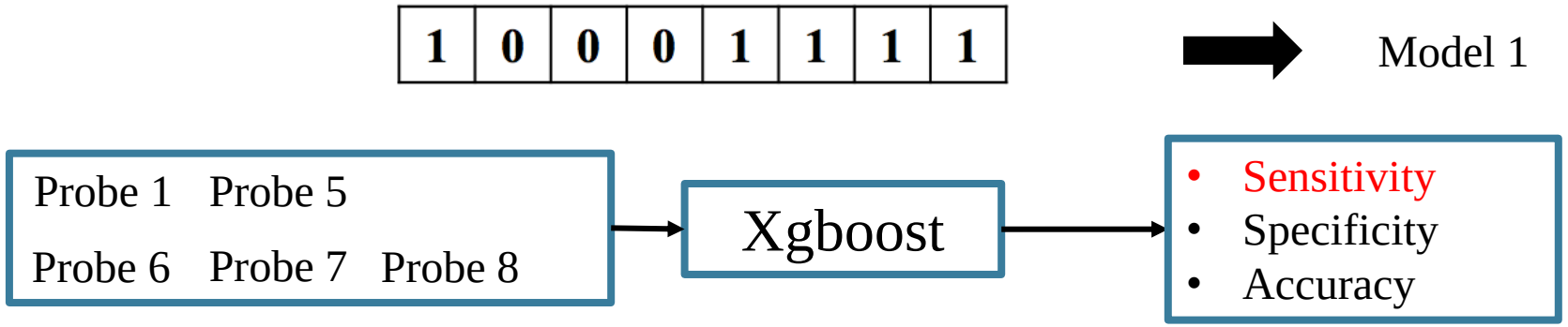
Chemotherapy

ID	GSM5401(	GSM5401(	GSM54011	GSM54011
1007_s_at	12.61478	11.84253	12.52451	12.55463
1053_at	7.903758	8.143369	8.830648	9.397681
117_at	7.983277	8.125691	7.849639	7.965634
121_at	10.00636	10.05648	10.46377	9.743123
1255_g_at	1.52182	2.766601	3.275003	2.216496
1294_at	10.05307	9.986836	11.6857	9.058527
1316_at	8.796633	9.524082	6.998067	6.309369

54675 probes

$$C_{10}^{54675} = 6.6 * 10^{40}$$

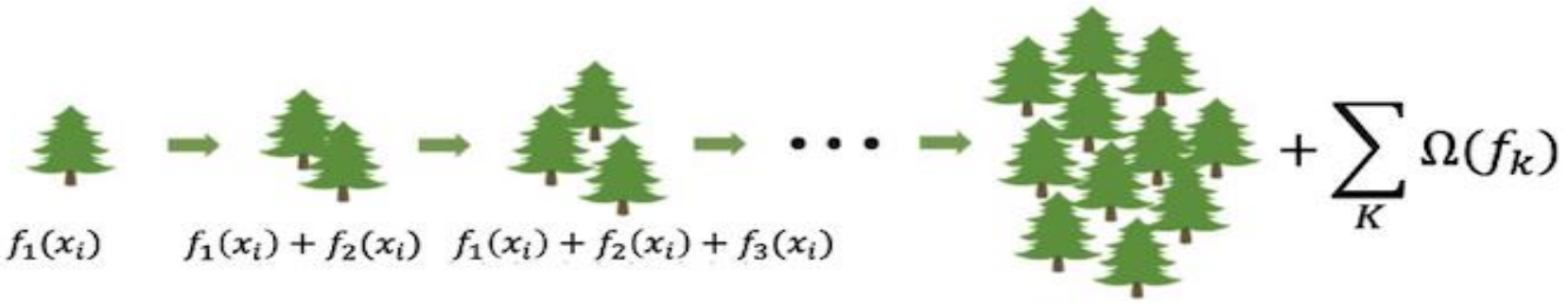
Genetic algorithm (GA)-Xgboost modeling (optimization)



Sample	**Probe 1**	**Probe 2**	**Probe 3**	**Probe 4**	**Probe 5**	Probe 6	Probe 7	Probe 8
	8.21	6.74	11.10	3.76	6.34	6.57	7.36	4.59
	6.09	6.61	10.37	3.65	6.38	5.47	8.30	4.45
	8.95	7.11	10.94	4.54	6.08	5.88	7.60	5.38
	9.22	7.05	10.48	3.87	6.09	6.38	5.02	4.90
	9.27	6.73	11.28	3.69	6.15	6.32	6.95	5.30
	8.21	7.21	11.29	3.94	6.51	5.86	7.10	4.70

Hsiao et al. JARE 2020

GA-xgboost modeling (optimization)



- Model: assuming we have K trees

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

- Objective

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Training loss — Complexity of the Trees

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

Number of leaves — **L2 norm of leaf scores**

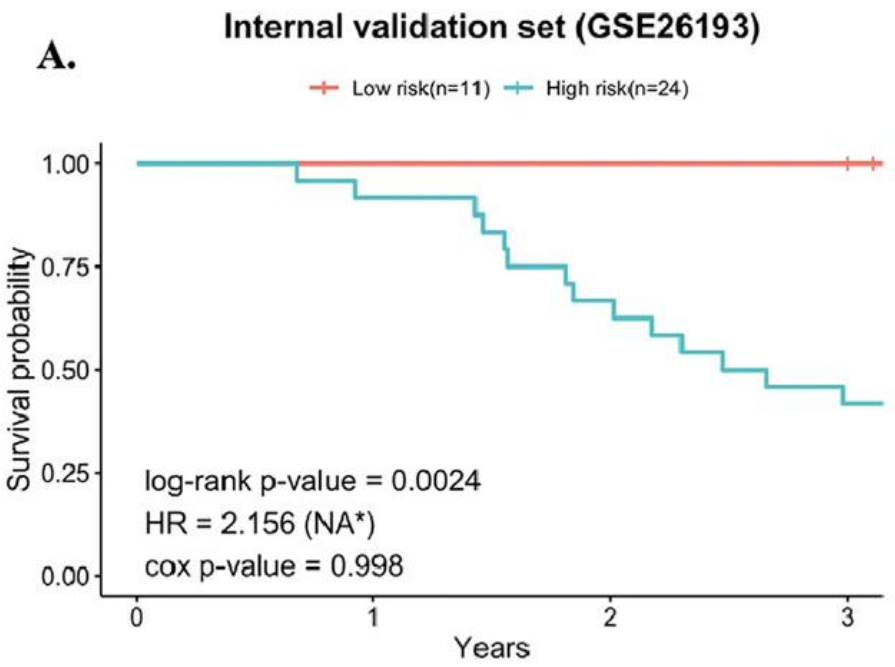
Hsiao et al. JARE 2020

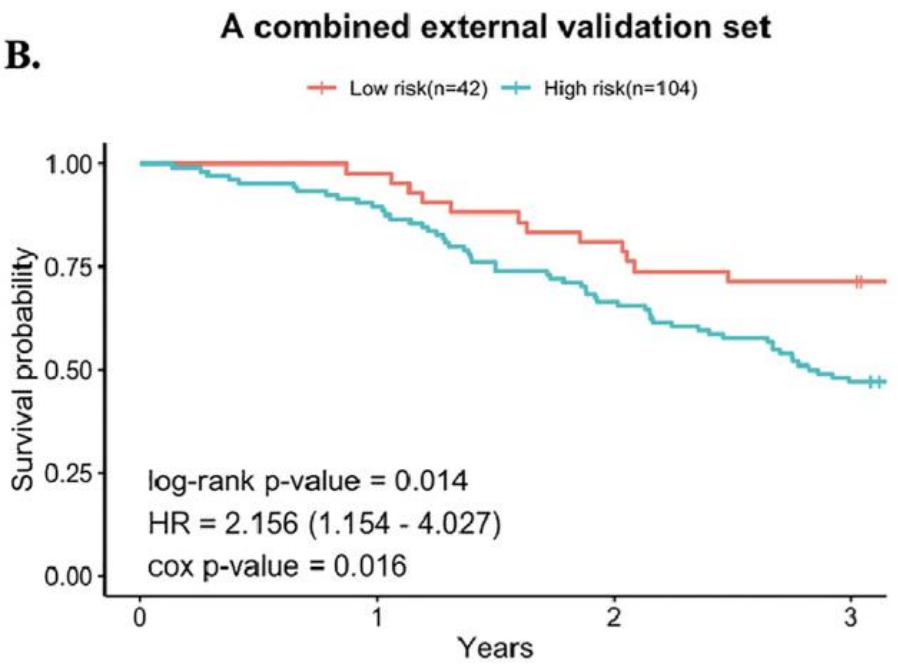
Prediction performances in external datasets

	Accuracy	Sensitivity	Specificity	F1-score
GA-XGBoost				
GSE26193 (internal validation set)	**0.714**	**1**	0.524	0.737
Combined external validation set*	**0.589**	**0.824**	0.385	0.651
GSE30161	**0.580**	**0.739**	0.444	0.618
GSE19829	**0.478**	**1**	0.143	0.600
GSE63885	**0.630**	**0.833**	0.432	0.690
Forward logistic regression				
GSE26193 (internal validation set)	0.600	0.533	0.650	0.533
Combined external validation set*	0.514	0.456	0.564	0.466
GSE30161	0.620	0.652	0.593	0.612
GSE19829	0.478	0.556	0.429	0.455
GSE63885	0.452	0.306	0.595	0.355
LASSO regression				
GSE26193 (internal validation set)	0.543	0.643	0.476	0.529
Combined external validation set*	0.555	0.544	0.564	0.532
GSE30161	0.520	0.478	0.556	0.478
GSE19829	0.565	0.889	0.357	0.615
GSE63885	0.575	0.500	0.649	0.537

Hsiao et al. JARE 2020

Survival differences between patients with high and low risks





* NA: The range of 95% of confident interval is between 0 and infinity because no event occurs in low risk group during the observed period.

20,000 * 1,500 * 0.05 = 1,500,000

Chemo Patient Percentage Cost per month

Hsiao et al. JARE 2020

Can we do better? (a little bit earlier)

- In addition to clinical variables, what can we have?
  - Genetics (DNA)
  - Gene expression (RNA)

Angelina Jolie Has Ovaries Removed After Doctor Detects Possible Sign of Early Cancer

Angelina Jolie

JAMES DEVANEY/GETTY

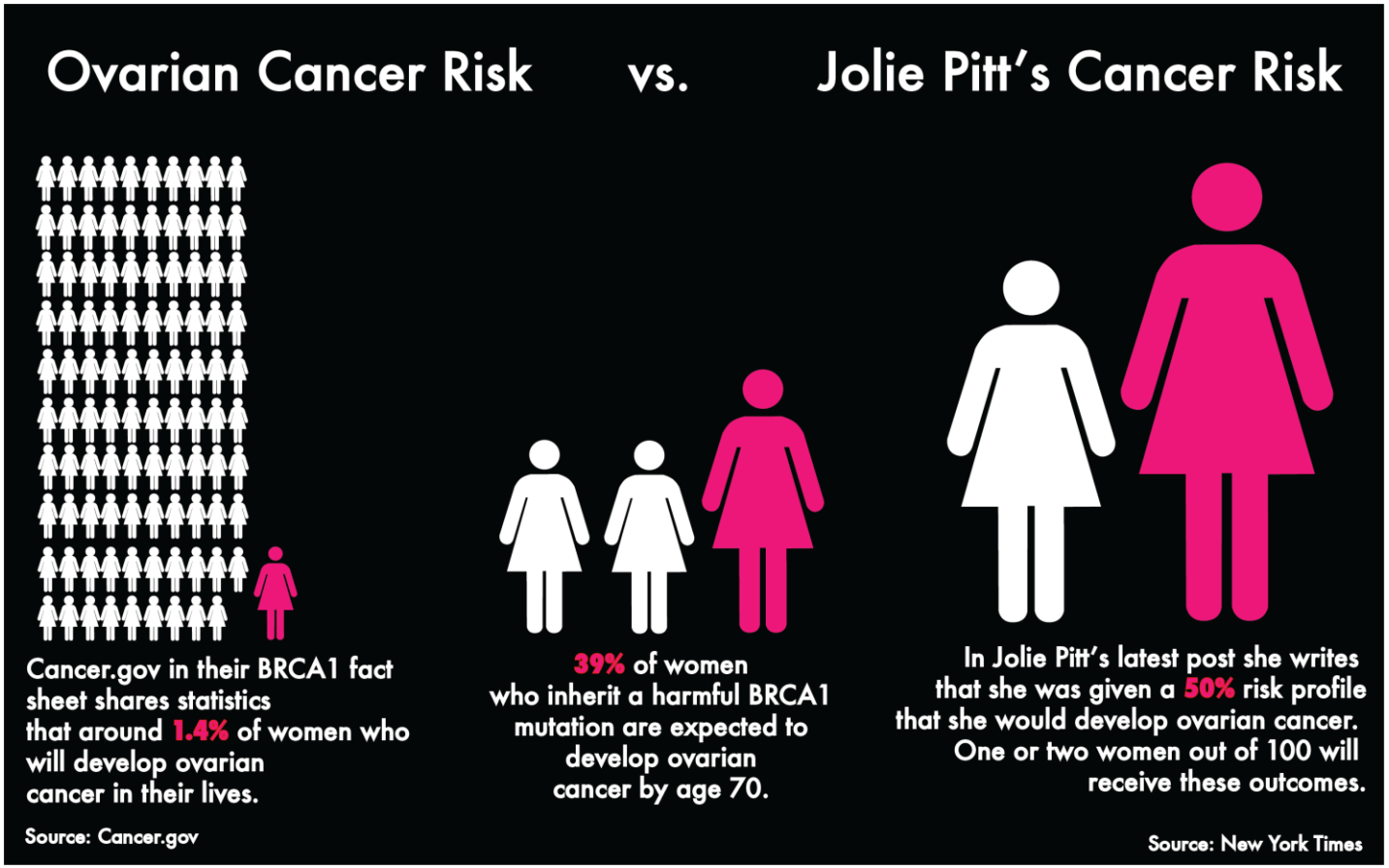
BY K.C. BLUMM 03/24/2015 AT 01:05 AM EDT

It was the phone call every woman hopes she never gets.

After having a preventive double mastectomy two years ago to reduce her risks of getting cancer, Angelina Jolie – who lost her mother, grandmother and aunt to the disease – got a call from her doctor two weeks ago with results from a recent blood test. As she writes in a *New York Times* op-ed piece published early Tuesday, the doctor told her that the test showed markers that could be a sign of early cancer.

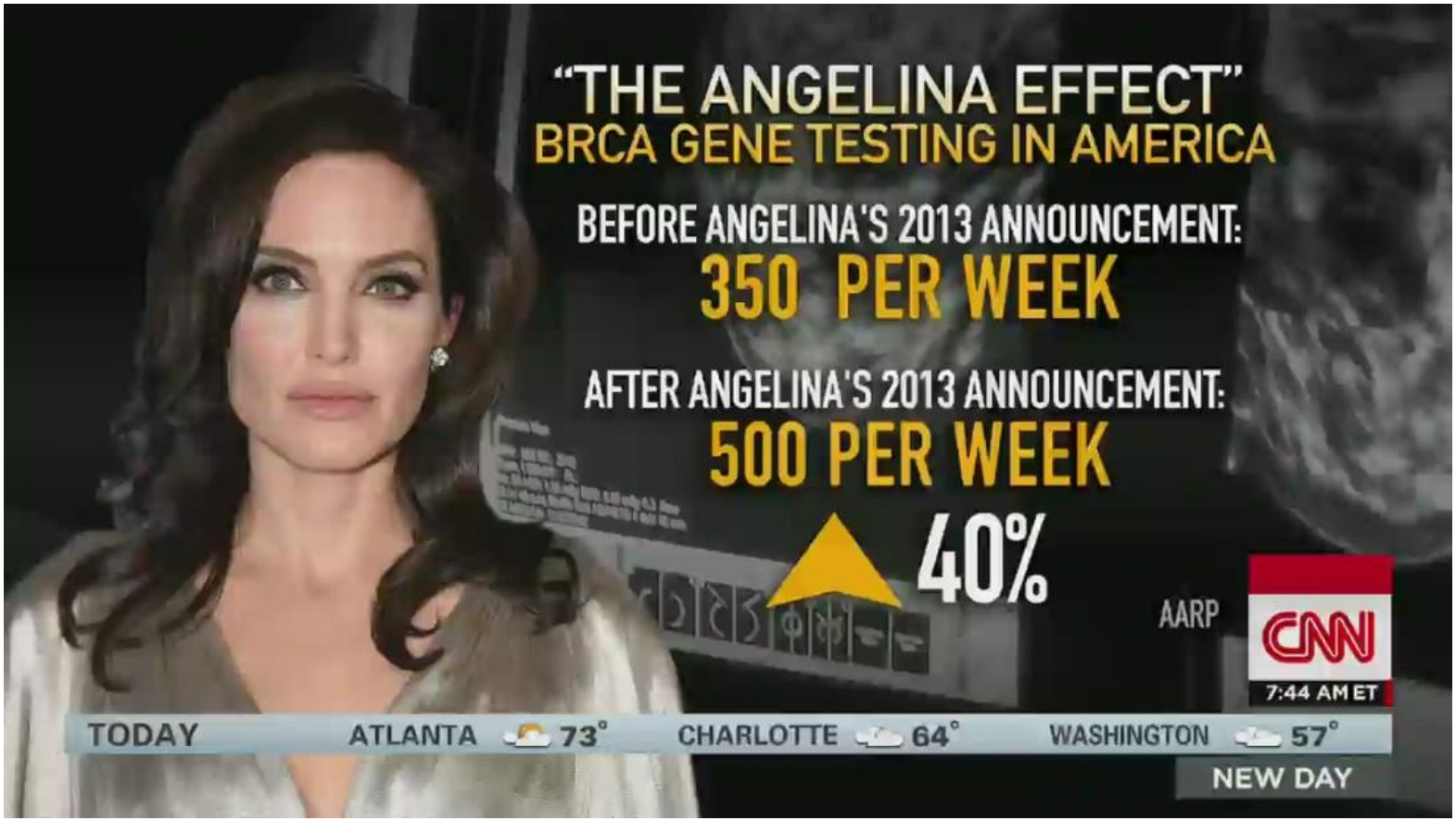
"I went through what I imagine thousands of other women have felt," Jolie, 39, writes. "I told myself to stay calm, to be strong, and that I had no reason to think I wouldn't live to see my children grow up and

Angelina Jolie's cancer risk



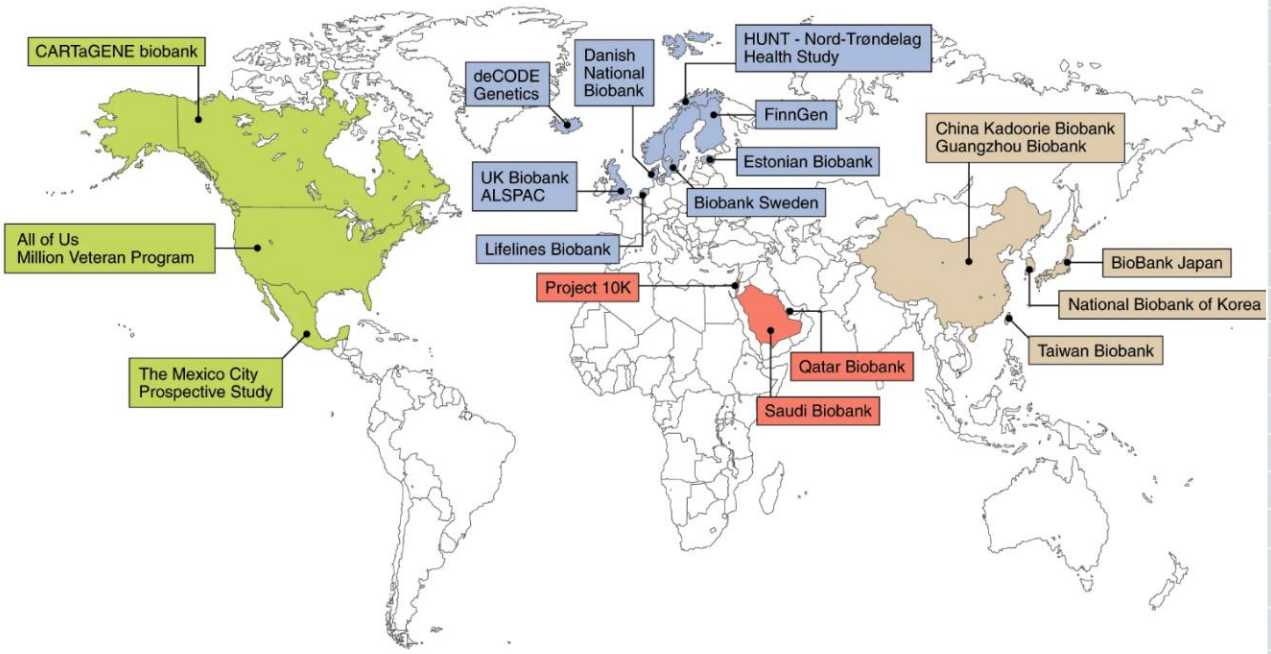
https://innovatemedtec.com/content/the-angelina-jolie-pitt-cancer-story

Angelina effect -> more people get screening



https://edition.cnn.com/2015/03/24/health/angelina-jolie-did-the-right-thing/index.html

Global Biobanks



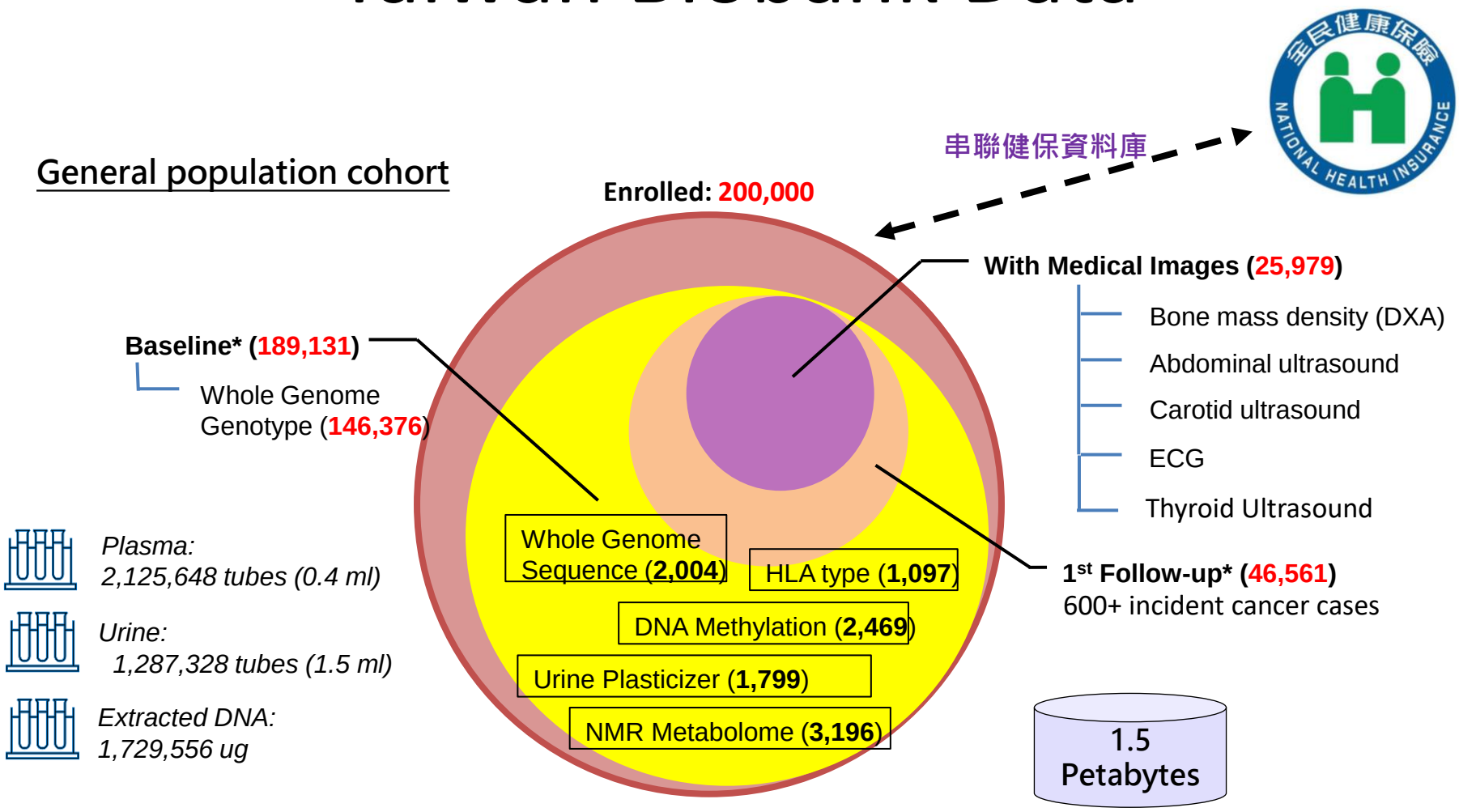
Location	Biobank	N (goal)
Canada	CARTaGENE biobank[119]	43,000
USA	All of Us[33] Million Veteran Program[49]	1,000,000 > 600,000
Mexico	The Mexico City Prospective Study[52]	150,000
Iceland	deCODE Genetics	500,000
UK	UK Biobank[38] Avon Longitudinal Study of Parents and Children (ALSPAC)[20]	500,000 > 15,000
Netherlands	Lifelines Biobank[120]	> 167,000
Denmark	Danish National Biobank[121]	
Norway	HUNT - Nord-Trøndelag Health Study[122]	125,000
Sweden	Biobank Sweden	
Finland	FinnGen	500,000
Estonia	Estonian Biobank[123]	52,000
Israel	Project 10K	10,000
Saudi Arabia	Saudi Biobank	200,000
Qatar	Qatar Biobank[124]	60,000
China	China Kadoorie Biobank[51] Guangzhou Biobank[125]	> 500,000 30,000
Japan	BioBank Japan[126]	200,000
Korea	National Biobank of Korea[127]	500,000
Taiwan	Taiwan Biobank[128]	200,000

Nat Med . 2020 Jan;26(1):29-38.

Taiwan Biobank

Formulated 80 SOPs + 200 documents to implement

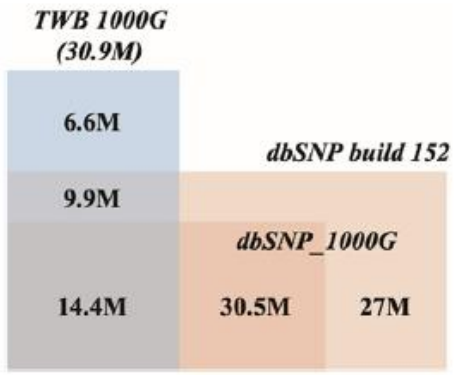
Taiwan Biobank Data

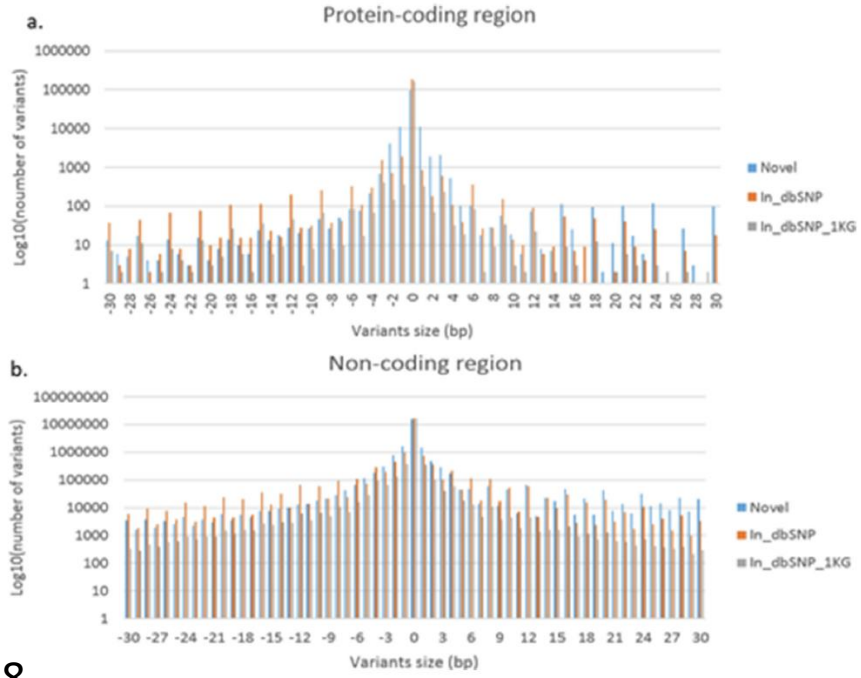


Whole Genome Sequencing (Taiwan biobank)

Rare variants discovery by extensive whole-genome sequencing of the Han Chinese population in Taiwan: Applications to cardiovascular medicine

Jyh-Ming Jimmy Juang [a], Tzu-Pin Lu [b], Ming-Wei Su [c], Chien-Wei Lin [c], Jenn-Hwai Yang [d], Hou-Wei Chu [c], Chien-Hsiun Chen [d], Yi-Wen Hsiao [b], Chien-Yueh Lee [e], Li-Mei Chiang [e], Qi-You Yu [b], Chuhsing Kate Hsiao [b], Ching-Yu Julius Chen [a], Pei-Ei Wu [d], Chien-Hua Pai [c], Eric Y. Chuang [e,*], Chen-Yang Shen [c,d,*]





Journal of Advanced Research 30 (2021) 147–158

Variants related to pharmacogenomics in different populations

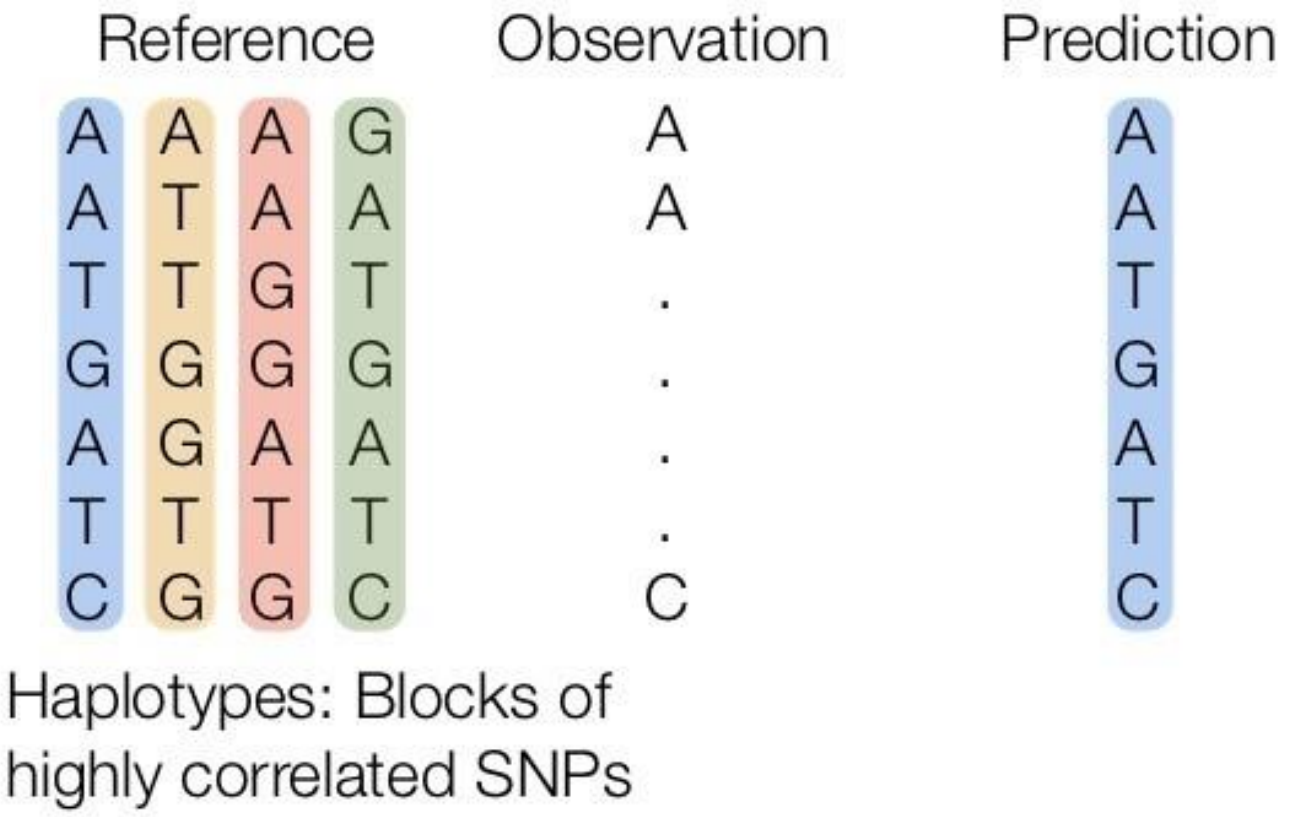
Table 4
Examples of pharmacogenetics in clinical scenarios and comparisons of minor allele frequency across different populations.

Drug name	Gene	Genetic variants	Ref.	Alt.	Minor Allele Frequency										Clinical application or impact
					1KTWWGS	1000GCHB	1000GCHS	1000G JPT	gnomADEAS	gnomAD SAS	1000GCEU	gnomAD non-Finish	gnomAD AFR	ChinaMap	
Warfarin	VKORC1	rs9923231	C	T	0.865	0.956*	0.890	0.903	0.901*	NA	0.429*	0.368*	0.102*	NA	Weekly warfarin dosage
	CYP2C9	rs1057910	A	C	0.034	0.038	0.047	0.019	0.033	0.109*	0.066	0.068*	0.012*	0.0459	
		rs1057910	A	G	NA	NA	NA	NA	NA	NA	NA	<0.000	0.00007	NA	
		rs1799853	C	T	0.001	<0.000	0.004	<0.000	0.0004	0.047*	0.152*	0.127*	0.021*	0.0012	
Clopidogrel	CYP2C19*2	rs4244285	G	A	0.348	0.335	0.352	0.3220	0.308*	0.325	0.131*	0.147*	0.178*	0.3117	Affects the metabolism of clopidogrel and the risk of coronary stenting thrombosis or cardiovascular events
		rs3758580	C	T	0.305	0.335	0.352	0.3220	0.310	0.327	0.131	0.147*	0.179*	0.3121	
		rs181297724	G	C	0.0048	<0.000	0.004	0.009	0.004	0.00007*	<0.000	0.0012*	0.00004*	0.0053	
		rs181297724	G	A	NA	NA	NA	NA	NA	NA	NA	<0.000	0.00007	NA	
		rs17878459	G	C	<0.000	<0.000	<0.000	<0.000	<0.000	0.0078*	0.0250	0.033	0.008	0.00024	
		rs778258371	G	A	<0.000	NA	NA	NA	NA	NA	NA	<0.000	<0.000	0.00005	
		rs144036596	G	A	NA	<0.000	<0.000	<0.000	0.0001	0.0001	<0.000	0.0003	0.0002	0.00014	
		rs144036596	G	C	NA	NA	NA	NA	NA	NA	NA	0.00005	<0.000	NA	
		rs144036596	G	T	NA	NA	NA	NA	NA	NA	NA	0.00001	<0.000	NA	
		rs550527959	A	T	<0.000	<0.000	<0.000	<0.000	<0.000	<0.000	<0.000	<0.000	<0.000	NA	
	CYP2C19*3	rs4986893	G	A	0.055	0.044	0.048	0.072	0.063	0.004*	0*	0.00025*	0.00038*	0.0485	
		rs763625282	T	A	<0.000	NA	NA	NA	NA	NA	NA	<0.000	<0.000	0.0005	
		rs144036596	G	A	<0.000	<0.000	<0.000	<0.000	0.00011	0.0001	<0.000	0.00029	0.00021	0.00014	
		rs144036596	G	C	NA	NA	NA	NA	NA	NA	NA	0.00005	<0.000	NA	
		rs144036596	G	T	NA	NA	NA	NA	NA	NA	NA	0.00001	<0.000	NA	

* Indicates *P* value < 0.01 compared with 1KTW-WGS; AFR: African/African American; CEU: European; CHB: Chinese-Bejing; CHS: Chinese-South; JPT: Japanese; EAS: East Asian; NA: not available; SAS: South Asian; Ref: reference; Alt: alternate; CHC: chronic hepatitis C; <0.000 means the actual number is not provided in the databases, but it should be less than < 0.000001.

DNA imputation

- To get genotypes of loci without being sequenced



https://zhuanlan.zhihu.com/p/53564008

Multi-ethnic Imputation System (MISystem)

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

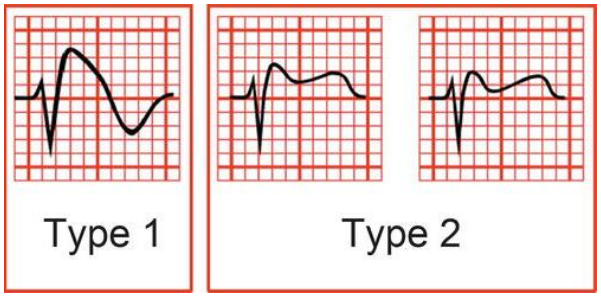
Multi-ethnic Imputation System (MI-System): A genotype imputation server for high-dimensional data

Amrita Chattopadhyay [a], Chien-Yueh Lee [b,1], Ying-Cheng Shen [c,1], Kuan-Chen Lu [d], Tzu-Hung Hsiao [e], Ching-Heng Lin [e], Liang-Chuan Lai [a,f], Mong-Hsun Tsai [a,g,h], Tzu-Pin Lu [a,d,i,j,*], Eric Y. Chuang [a,c,k,*]

Multi-ethnic Imputation System

https://misystem.cgm.ntu.edu.tw/

Brugada Syndrome (BrS)



- 4% - 20% sudden cardiac death
- Incidence
  - 1 to 5 per 10,000 in white people
  - 12 per 10,000 in southeast Asian
  - 10 per 10,000 in Taiwanese
- Normal heart structure, normal heart ECG
- Normal image and biochemical test

Juang et al. J Arrhythmia. 2016 Oct; 32(5):418-425

Validations of the 3 risk loci in Taiwanese

Table 2. Validation of the Previously Reported SNPs in the Taiwanese BrS Patients

	Chromosome Position (Build 37)	Imputed or Original	Gene or Nearest Gene	Risk Allele	Protective Allele	RAF (190 Cases /15 981 Controls) in Taiwanese	MAF (Case/ Control) in Whites	All BrS Patients in Taiwan vs Healthy Controls		OR in the Previous GWAS[14]
SNPs causing susceptibility to BrS (set 1)								*P* value*	OR	OR
rs11708996	Chr3:38633923	Original	*SCN5A*	C	G	0.011/0.007	0.23/0.15	4.27×10^{-01}	1.5 (0.46–3.57)	1.64 (1.30–2.07)
rs10428132	Chr3:38777554	Imputed	*SCN10A*	T	G	0.305/0.193	0.69/0.41	5.92×10^{-08}†	1.84 (1.47–2.29)	3.00 (2.45–3.69)
rs9388451	Chr6:126090377	Original	*HEY2, NCOA7*	C	T	0.818/0.742	0.65/0.50	7.94×10^{-04}†	1.56 (1.21–2.05)	1.83 (1.51–2.22)

Table 3. Comparisons of Disease Risk Using the PRS Models Generated by the 3 Sets of SNPs

PRS Range	Odds Ratio
Set 1 SNPs	
0%–20%*	1
21%–40%	0.16 (0.06–0.42)
41%–60%	2.10 (1.36–3.22)†
61%–80%	2.03 (1.32–3.13)†
81%–100%	0.77 (0.45–1.32)

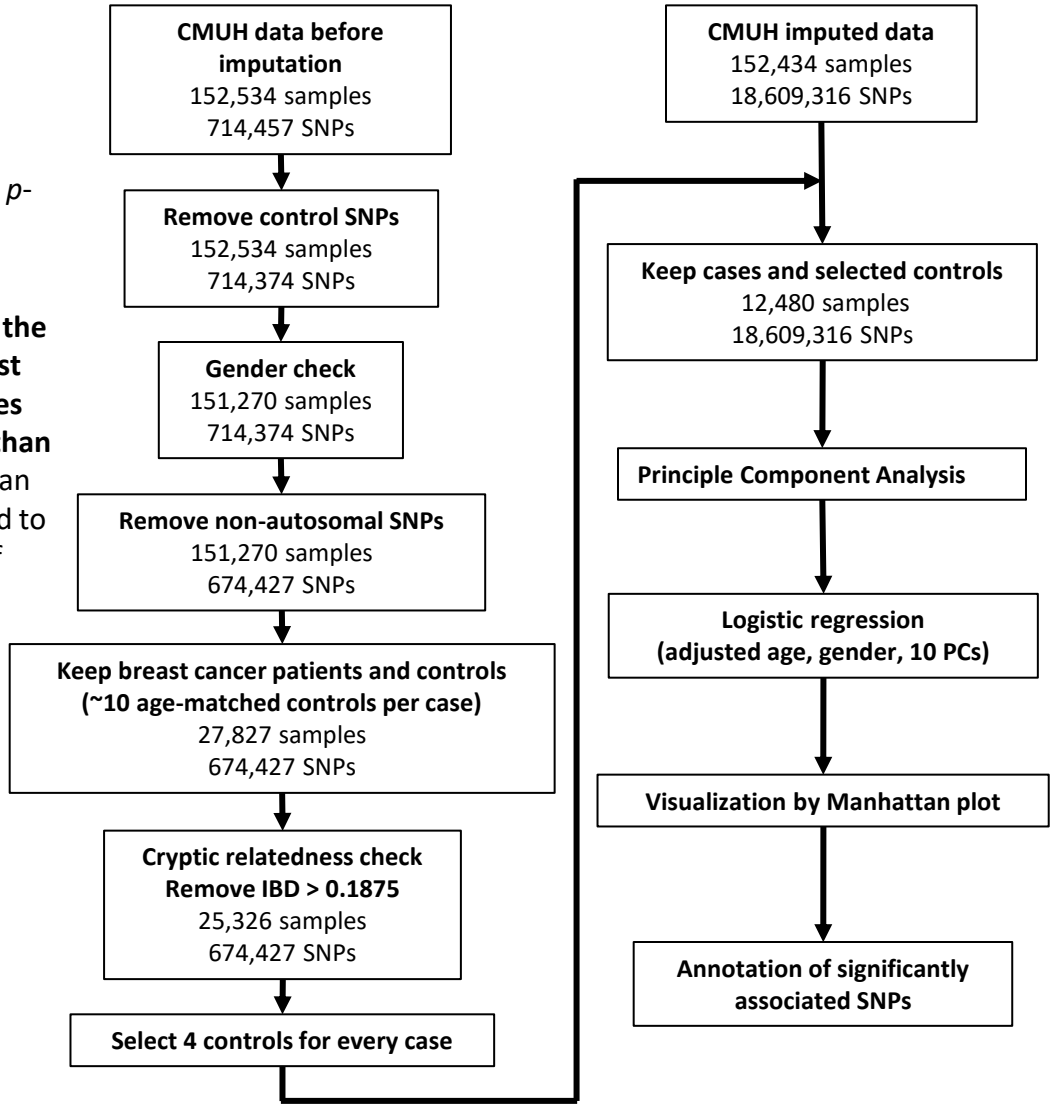
Mayo Clinic Genetic Heart Rhythm Clinic, USA

Juang et al. Cir Genom Precis Med. 2020 Aug 13(4):e002797

A GWAS in breast cancer

Data before imputation:
Already filtered by
SNP call rate < 0.95 &
genotype call rate < 0.95 &
Hardy-Weinberg equilibrium *p*-value < 10^{-6}

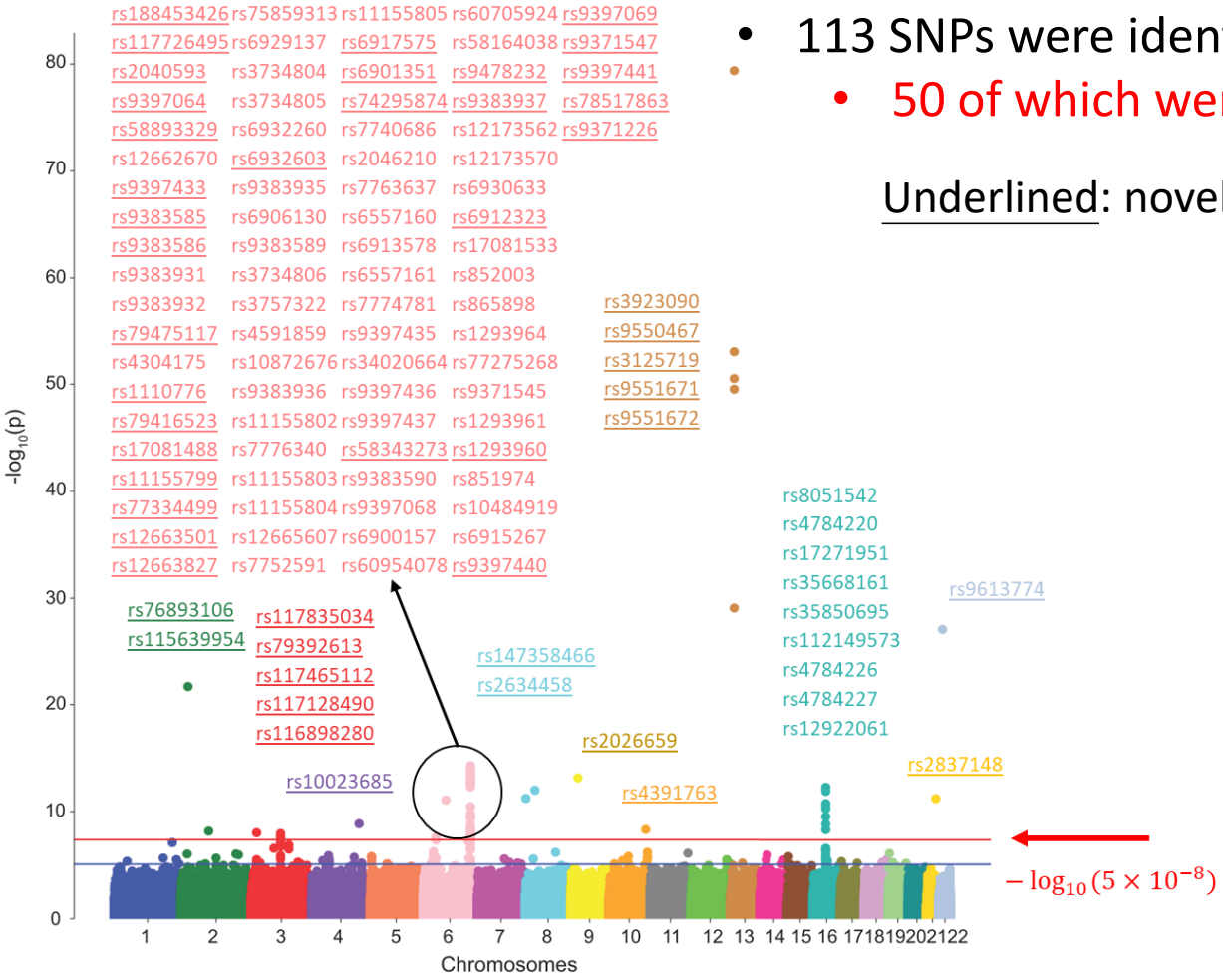
- **age at sample collection of the controls** vs. **the age at breast cancer diagnosis of the cases**
- controls with age no **more than 3 years older or younge**r than the cases were also included to ensure sufficient controls of matched age



Imputed data:
Already filtered by posterior likelihood of called genotypes > 0.9 and info score > 0.7

2,496 breast cancers
9,984 controls

113 Significant SNPs



- 113 SNPs were identified
  - 50 of which were novel SNPs

Underlined: novel SNPs

Racial differences

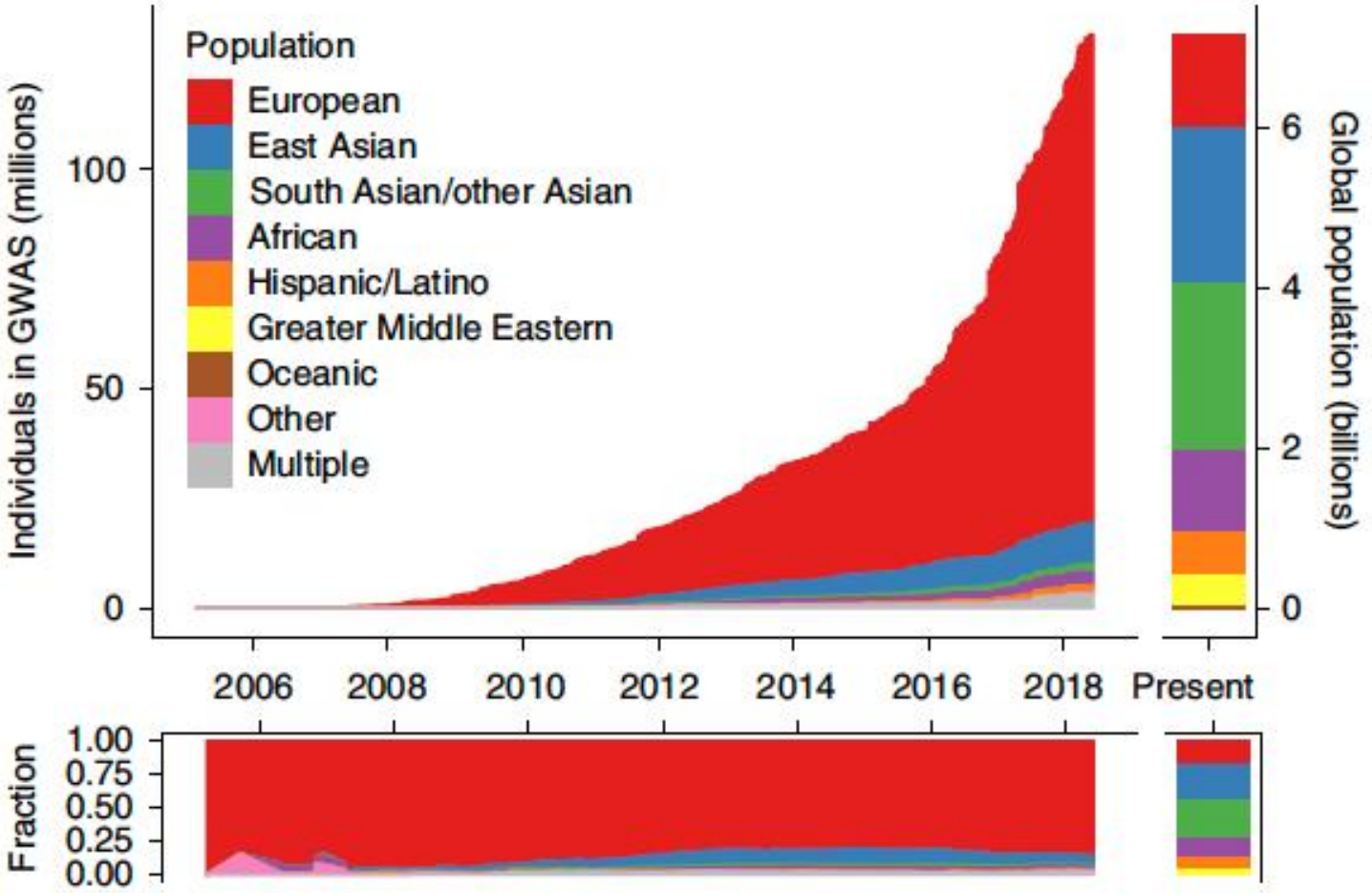


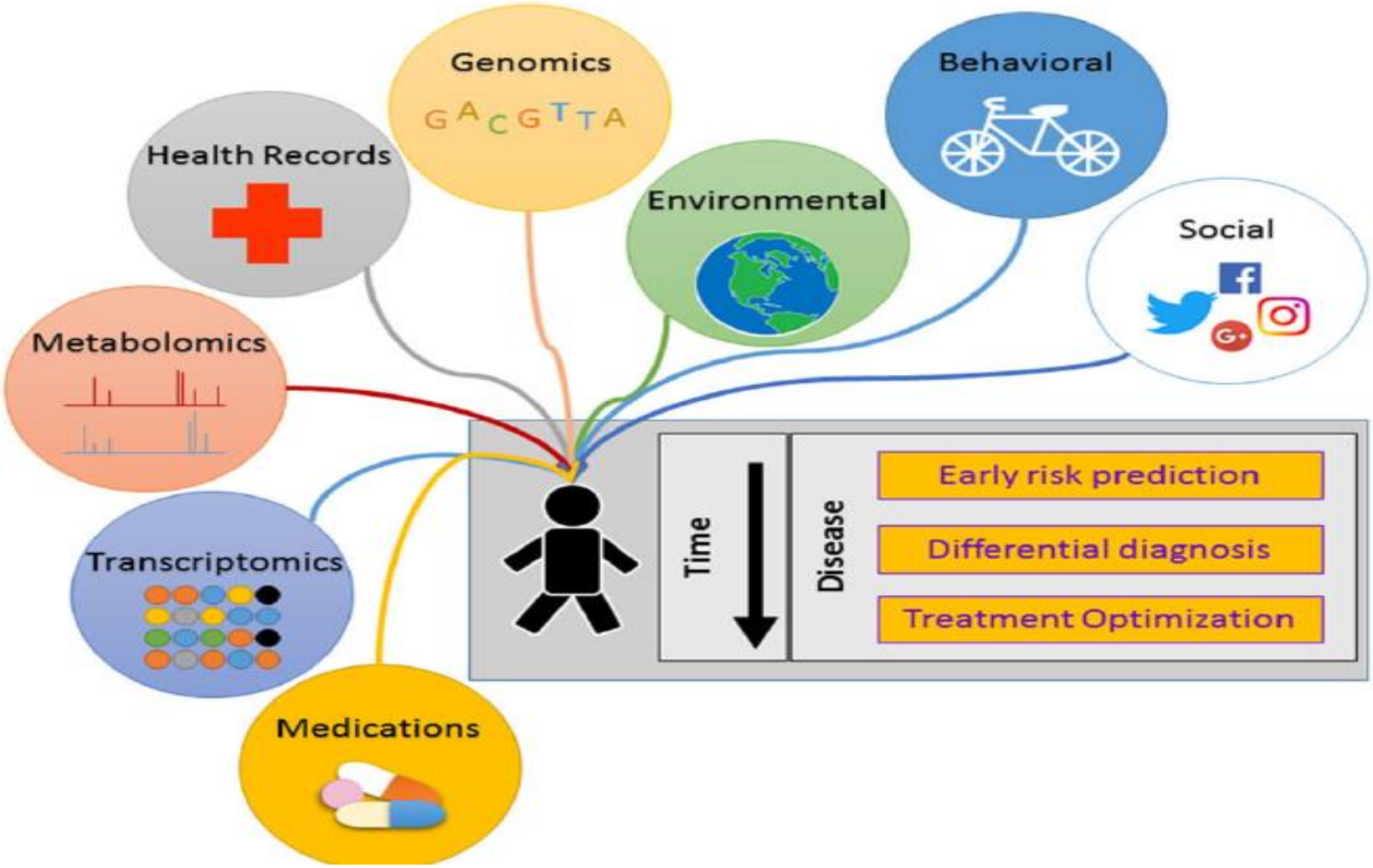
Fig. 1 | Ancestry of GWAS participants over time, as compared with the global population. Cumulative data, as reported by the GWAS catalog76. Individuals whose ancestry is 'not reported' are not shown.

Odds ratios of PRS in different PAM50 subtypes

		OR	95% CI lower	95% CI upper	*P* values
All breast cancers	20-40%	1.25	0.82	1.91	0.29
	40-60%	1.98	1.35	2.94	4.7×10^{-4}
	60-80%	1.77	1.20	2.65	4×10^{-3}
	80-100%	5.33	3.79	7.66	1.3×10^{-23}
Luminal A	20-40%	1.28	0.70	2.38	0.42
	40-60%	1.85	1.05	3.32	0.031
	60-80%	3.22	1.92	5.59	7×10^{-6}
	80-100%	3.55	2.13	6.14	7.7×10^{-7}
Luminal B	20-40%	3.91	1.12	19.10	0.046
	40-60%	2.31	0.58	12.00	0.32
	60-80%	4.55	1.34	21.93	0.016
	80-100%	4.23	1.23	20.51	0.027

		OR	95% CI lower	95% CI upper	*P* values
Basal-like	20-40%	1.00	0.28	3.57	1.00
	40-60%	1.48	0.48	4.94	0.48
	60-80%	1.64	0.54	5.40	0.37
	80-100%	2.28	0.81	7.21	0.11
Her2-enriched	20-40%	1.00	0.28	3.61	1.00
	40-60%	1.00	0.28	3.61	1.00
	60-80%	1.32	0.40	4.53	0.63
	80-100%	2.28	0.80	7.28	0.12

Take home message



Prosperi et al. BMC Medical Informatics and Decision Making (2018) 18:139

Thank you for your attentions!

Tzu-Pin Lu
Institute of Health Data Analytics and Statistics,
National Taiwan University
tplu@ntu.edu.tw

CAPPA
台灣癌症準備暨資源分享協會

衛生福利部國民健康署/臺灣大學公共衛生學院
臺灣癌症登記中心
Taiwan Cancer Registry Center