Supplementary data

A 23 gene–based molecular prognostic score precisely predicts overall survival of breast cancer patients

Hideyuki Shimizu, Keiichi I. Nakayama

Tables:

Table S1. List of international cohorts integrated in meta-analysis.

Table S2. Prognosis-related genes of the TCGA discovery cohort.

Table S3. Prognosis-related genes identified by meta-analysis.

Table S4. Characteristics of mPS in the METABRIC training cohort.

Supplementary figures:

Fig. S1. Workflow for computational calculations.

Fig. S2. Comprehensive validation of all prognosis-related genes by meta-analysis.

Fig. S3. Representative calculation of mPS.

Fig. S4. Characteristics of mPS bins.

Fig. S5. mPS stratifies DFS.

Fig. S6. Stratification of patients according to mPS for intrinsic subtypes of breast cancer.

Fig. S7. Kaplan-Meier curves according to mPS for OS of patients in the METABRIC test cohort in their 50s or 60s.

Fig. S8. Kaplan-Meier curves according to mPS for OS of patients in the METABRIC test cohort with IDC or MDLC.

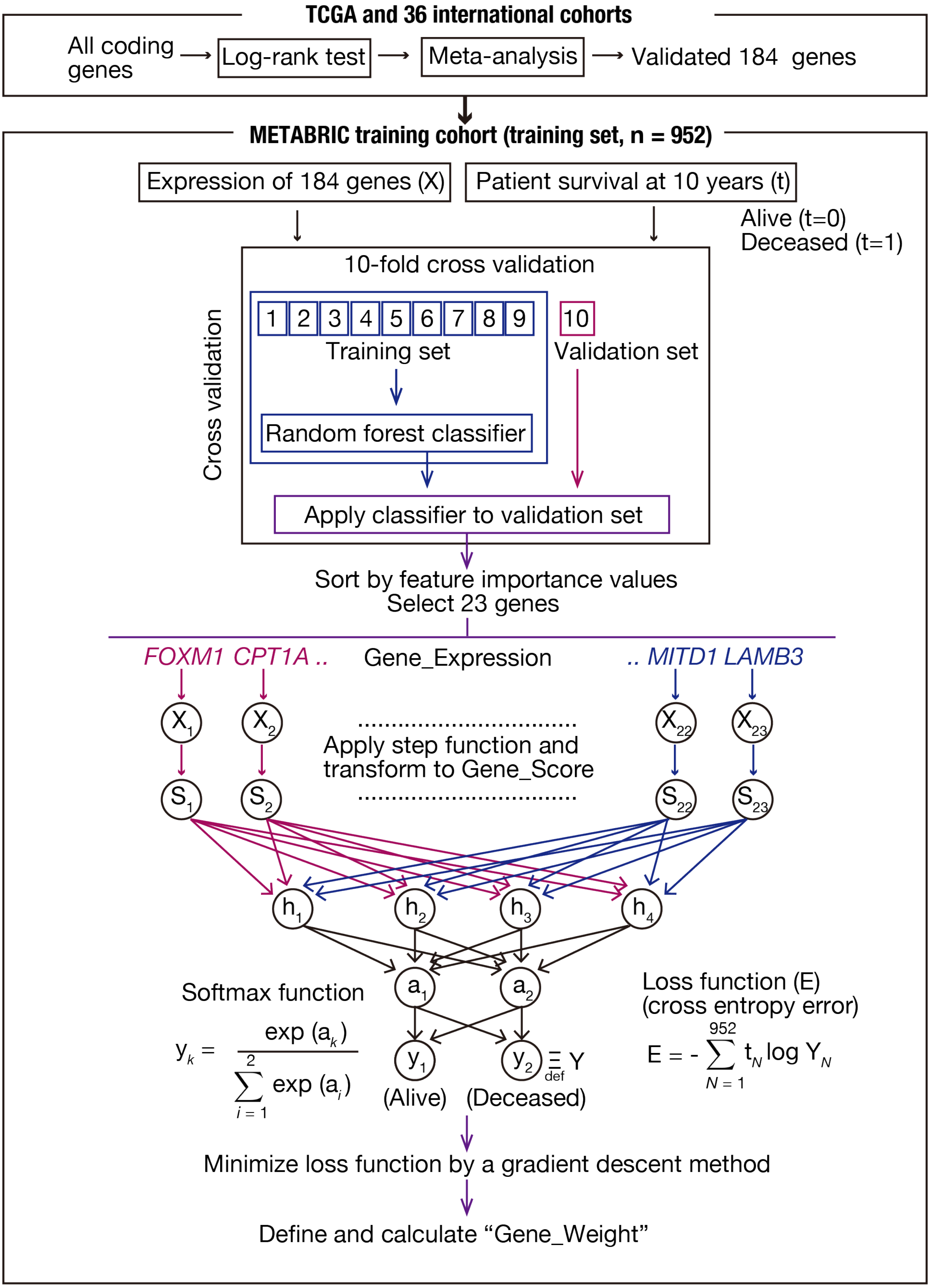
Fig. S9. Stratification of breast cancer patients of different races according to mPS.

Fig. S10. Kaplan-Meier curves according to mPS for OS of patients in the METABRIC test cohort at clinical TNM stage I or III.

Fig. S11. Stratification of patients according to mPS regardless of NPI.

Fig. S12. Relation of chemotherapy to OS in the METABRIC cohort.

Shimizu et al. Figure S1



TCGA and 36 international cohorts
All coding genes
Log-rank test
Meta-analysis
Validated 184 genes
METABRIC training cohort (training set, n = 952)
Expression of 184 genes (X)
Patient survival at 10 years (t)
Alive (t=0)
Deceased (t=1)
10-fold cross validation
Cross validation
1 2 3 4 5 6 7 8 9
10
Training set
Validation set
Random forest classifier
Apply classifier to validation set
Sort by feature importance values
Select 23 genes
FOXM1 CPT1A ..
Gene_Expression
.. MITD1 LAMB3
X_1 X_2 X_{22} X_{23}
Apply step function and transform to Gene_Score
S_1 S_2 S_{22} S_{23}
h_1 h_2 h_3 h_4
a_1 a_2
Softmax function
$y_k = \frac{\exp(a_k)}{\sum_{i=1}^{2} \exp(a_i)}$
y_1 $y_2 \underset{\text{def}}{\equiv} Y$
(Alive) (Deceased)
Loss function (E)
(cross entropy error)
$E = -\sum_{N=1}^{952} t_N \log Y_N$
Minimize loss function by a gradient descent method
Define and calculate "Gene_Weight"

Fig. S1. Workflow for computational calculations. Expression status (X) of the 184 validated prognosis-related genes in the METABRIC training set (*n* = 952) was first entered into a machine learning AI algorithm known as a random forest classifier. Twenty-three genes were selected on the basis of feature importance values. On the basis of the binary expression status of these 23 genes (S, designated Gene_Score), the probability for patient survival status at 10 years (y_1, alive; y_2, deceased) was predicted with the use of a softmax function. By comparison with the actual status (t), cross entropy error was calculated as a loss function. Each weight was optimized with the Adam method (learning rate, 0.001; epochs, 1000).

Shimizu et al. Figure S2

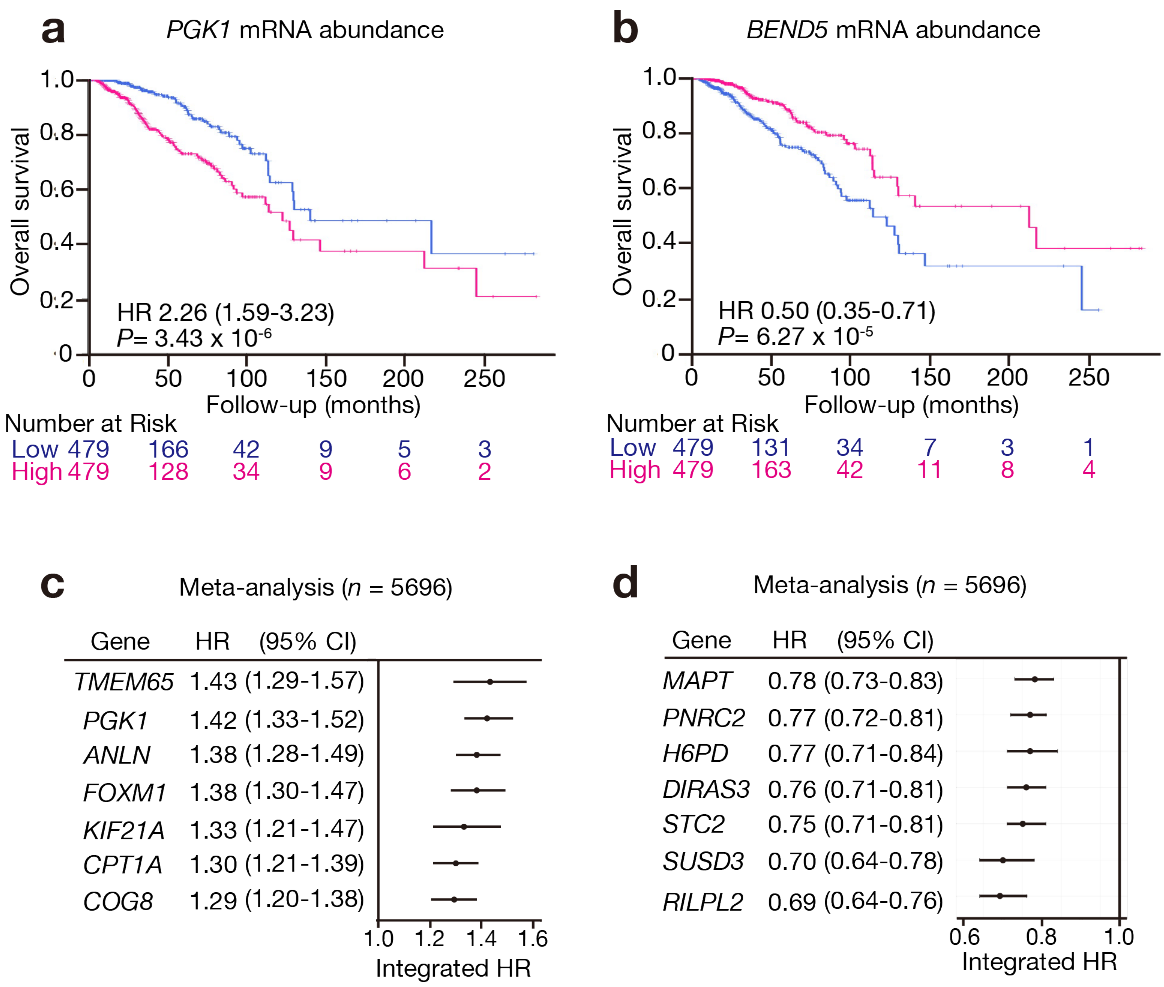
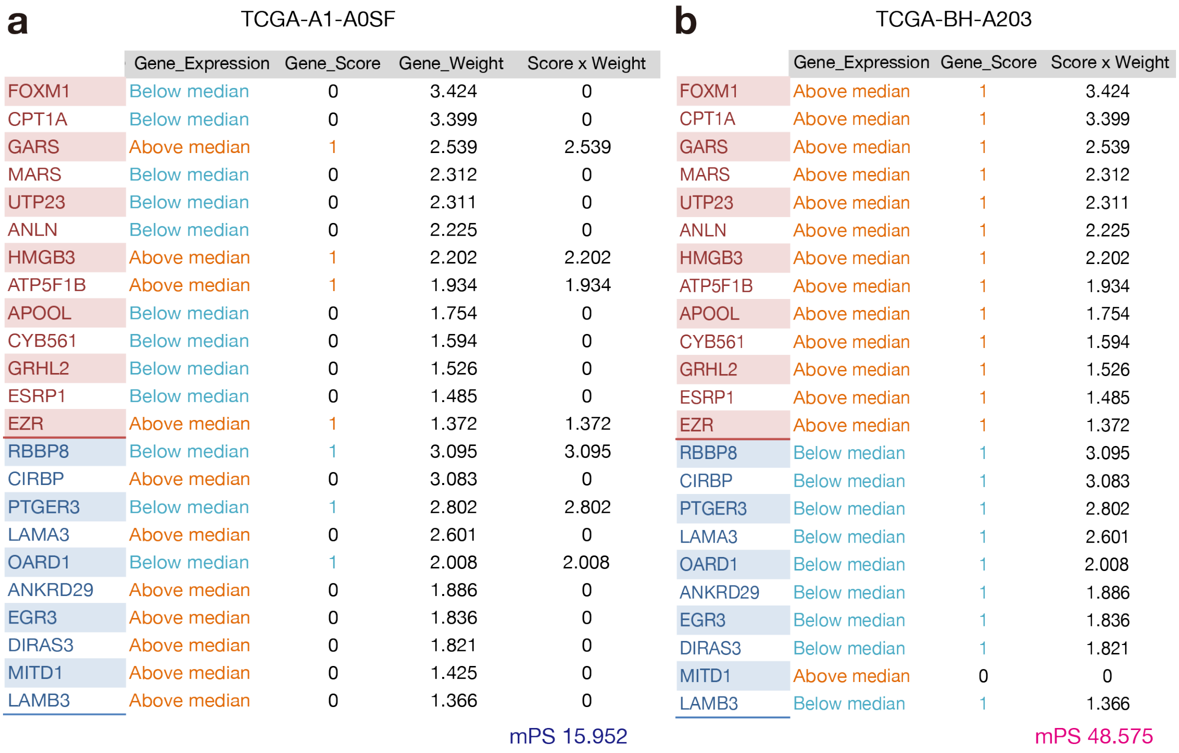


Fig. S2. Comprehensive validation of all prognosis-related genes by meta-analysis. (**a** and **b**) Kaplan-Meier curves for OS according to the expression level of *PGK1* (**a**) or *BEND5* (**b**) in the TCGA cohort. The HR, its 95% CI, the log-rank *P* value, and the number at risk are shown. (**c** and **d**) Top seven genes among the 184 validated prognosis-related genes for which high (**c**) or low (**d**) expression levels are associated with poor survival.

Shimizu et al. Figure S3



a TCGA-A1-A0SF

	Gene_Expression	Gene_Score	Gene_Weight	Score x Weight
FOXM1	Below median	0	3.424	0
CPT1A	Below median	0	3.399	0
GARS	Above median	1	2.539	2.539
MARS	Below median	0	2.312	0
UTP23	Below median	0	2.311	0
ANLN	Below median	0	2.225	0
HMGB3	Above median	1	2.202	2.202
ATP5F1B	Above median	1	1.934	1.934
APOOL	Below median	0	1.754	0
CYB561	Below median	0	1.594	0
GRHL2	Below median	0	1.526	0
ESRP1	Below median	0	1.485	0
EZR	Above median	1	1.372	1.372
RBBP8	Below median	1	3.095	3.095
CIRBP	Above median	0	3.083	0
PTGER3	Below median	1	2.802	2.802
LAMA3	Above median	0	2.601	0
OARD1	Below median	1	2.008	2.008
ANKRD29	Above median	0	1.886	0
EGR3	Above median	0	1.836	0
DIRAS3	Above median	0	1.821	0
MITD1	Above median	0	1.425	0
LAMB3	Above median	0	1.366	0

mPS 15.952

b TCGA-BH-A203

	Gene_Expression	Gene_Score	Score x Weight
FOXM1	Above median	1	3.424
CPT1A	Above median	1	3.399
GARS	Above median	1	2.539
MARS	Above median	1	2.312
UTP23	Above median	1	2.311
ANLN	Above median	1	2.225
HMGB3	Above median	1	2.202
ATP5F1B	Above median	1	1.934
APOOL	Above median	1	1.754
CYB561	Above median	1	1.594
GRHL2	Above median	1	1.526
ESRP1	Above median	1	1.485
EZR	Above median	1	1.372
RBBP8	Below median	1	3.095
CIRBP	Below median	1	3.083
PTGER3	Below median	1	2.802
LAMA3	Below median	1	2.601
OARD1	Below median	1	2.008
ANKRD29	Below median	1	1.886
EGR3	Below median	1	1.836
DIRAS3	Below median	1	1.821
MITD1	Above median	0	0
LAMB3	Below median	1	1.366

mPS 48.575

Fig. S3. Representative calculation of mPS. Actual calculation of mPS is shown for two patients (**a** and **b**) enrolled in the TCGA breast cancer cohort.

Shimizu et al. Figure S4

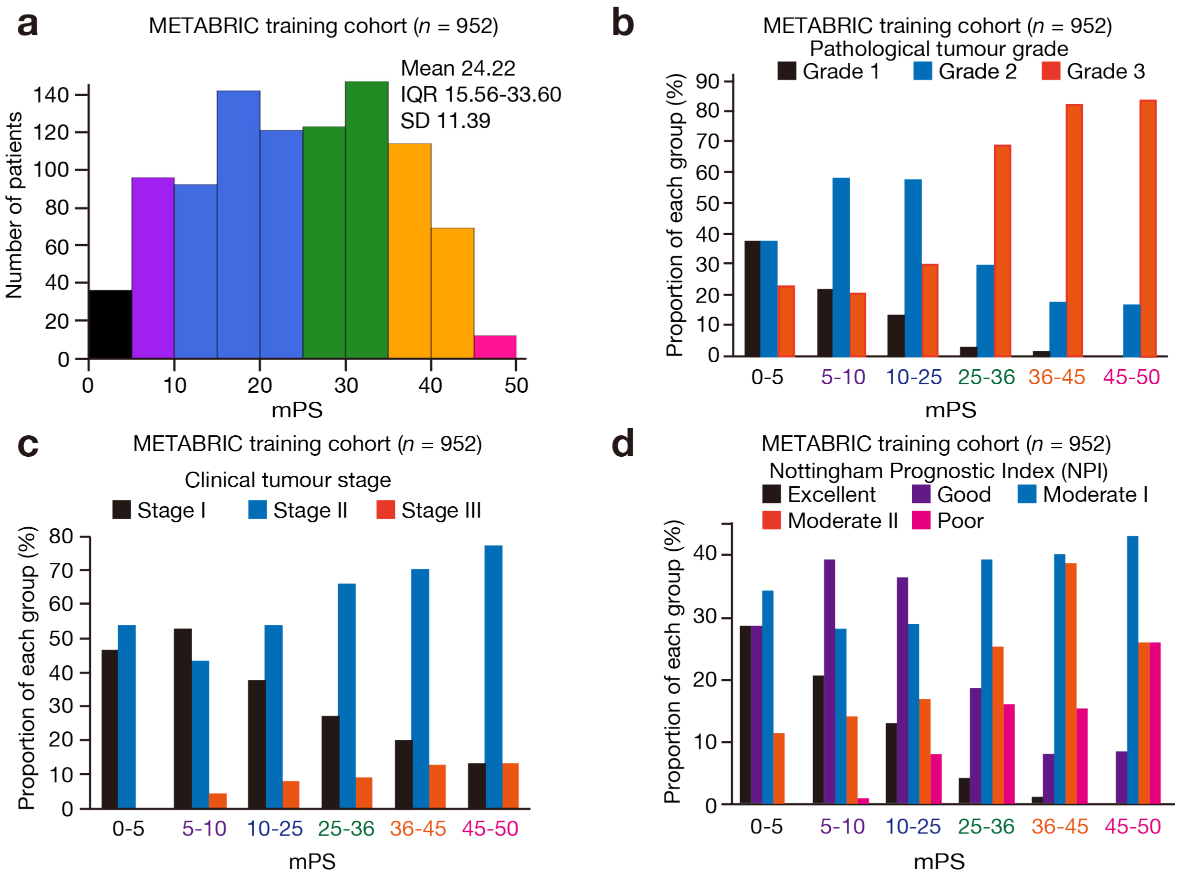
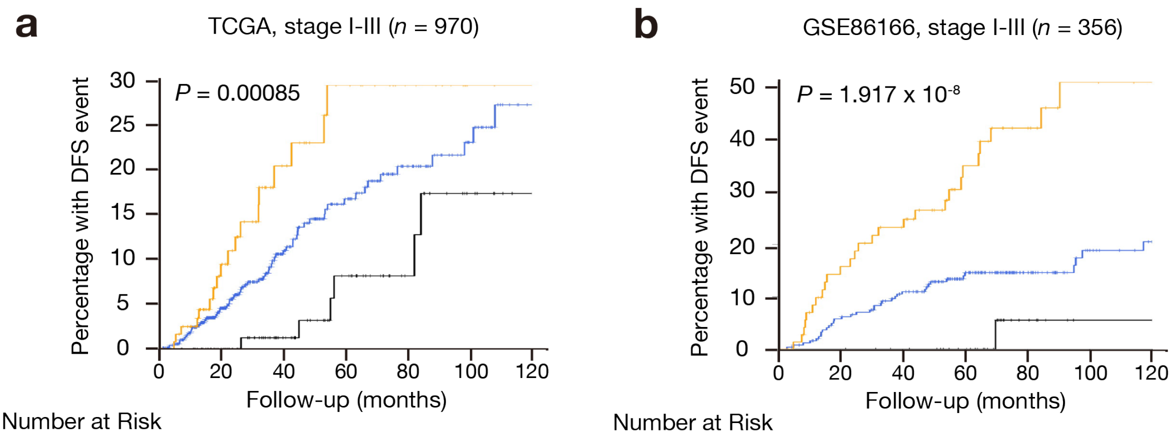


Fig. S4. Characteristics of mPS bins. (**a**) Distribution of mPS (ranging from 0 to 50) for all patients in the METABRIC training cohort. (**b**–**d**) Percentage of patients classified according to pathological grade (**b**), clinical tumour stage (**c**), or NPI cluster (**d**) in each of six mPS bins for the METABRIC training cohort. See also Supplementary Table S4.

Shimizu et al. Figure S5



Number at Risk

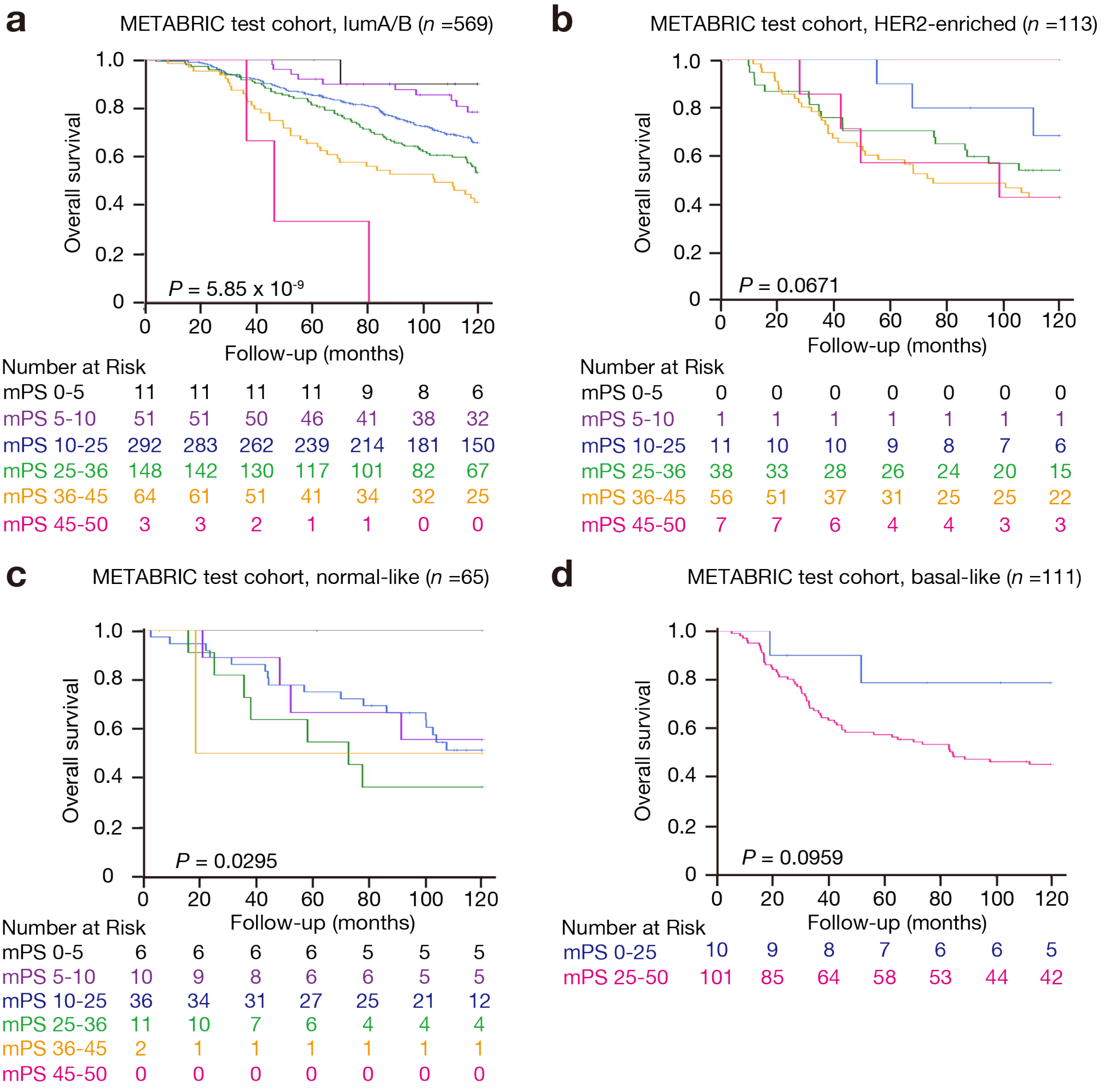
mPS 0-10	118	92	58	39	21	10	2
mPS 10-36	716	435	235	151	92	55	26
mPS 36-50	136	70	36	23	13	2	1

Number at Risk

mPS 0-10	35	35	32	26	8	2	2
mPS 10-36	247	225	199	141	81	55	46
mPS 36-50	74	57	49	28	18	10	8

Fig. S5. Stratification of DFS by mPS. (**a**) Kaplan-Meier curves according to mPS for DFS events in patients at stage I, II, or III in the TCGA cohort. (**b**) Kaplan-Meier curves according to mPS for DFS events in patients at stage I, II, or III in the GSE86166 data set. Only patients with DFS data are shown.

Shimizu et al. Figure S6



a Number at Risk

mPS 0-5	11	11	11	11	9	8	6
mPS 5-10	51	51	50	46	41	38	32
mPS 10-25	292	283	262	239	214	181	150
mPS 25-36	148	142	130	117	101	82	67
mPS 36-45	64	61	51	41	34	32	25
mPS 45-50	3	3	2	1	1	0	0

b Number at Risk

mPS 0-5	0	0	0	0	0	0	0
mPS 5-10	1	1	1	1	1	1	1
mPS 10-25	11	10	10	9	8	7	6
mPS 25-36	38	33	28	26	24	20	15
mPS 36-45	56	51	37	31	25	25	22
mPS 45-50	7	7	6	4	4	3	3

c Number at Risk

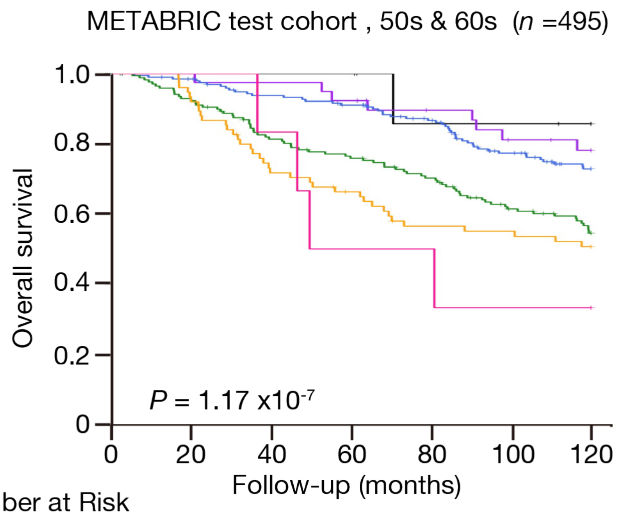
mPS 0-5	6	6	6	6	5	5	5
mPS 5-10	10	9	8	6	6	5	5
mPS 10-25	36	34	31	27	25	21	12
mPS 25-36	11	10	7	6	4	4	4
mPS 36-45	2	1	1	1	1	1	1
mPS 45-50	0	0	0	0	0	0	0

d Number at Risk

mPS 0-25	10	9	8	7	6	6	5
mPS 25-50	101	85	64	58	53	44	42

Fig. S6. Stratification of patients according to mPS for intrinsic subtypes of breast cancer. Kaplan-Meier curves according to mPS were constructed for OS of patients in the METABRIC test cohort with luminal A or B (lumA/B) (**a**), HER2-enriched (**b**), normal-like (**c**), or basal-like (**d**) intrinsic subtypes.

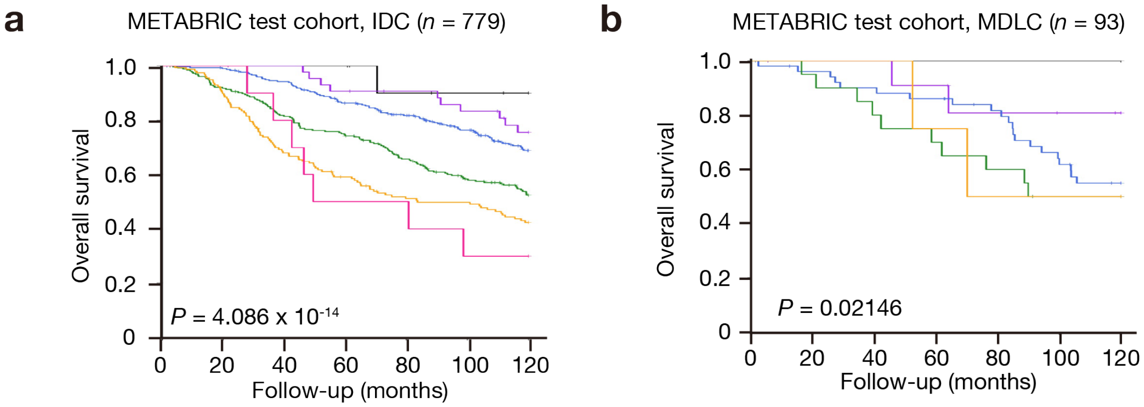
Shimizu et al. Figure S7



Number at Risk							
mPS 0-5	9	9	9	9	6	6	5
mPS 5-10	40	39	38	36	32	28	26
mPS 10-25	195	189	177	170	154	129	110
mPS 25-36	170	155	134	124	113	92	79
mPS 36-45	75	69	53	47	39	38	35
mPS 45-50	6	6	5	3	3	2	2

Fig. S7. Kaplan-Meier curves according to mPS for OS of patients in the METABRIC test cohort in their 50s or 60s.

Shimizu et al. Figure S8



Number at Risk

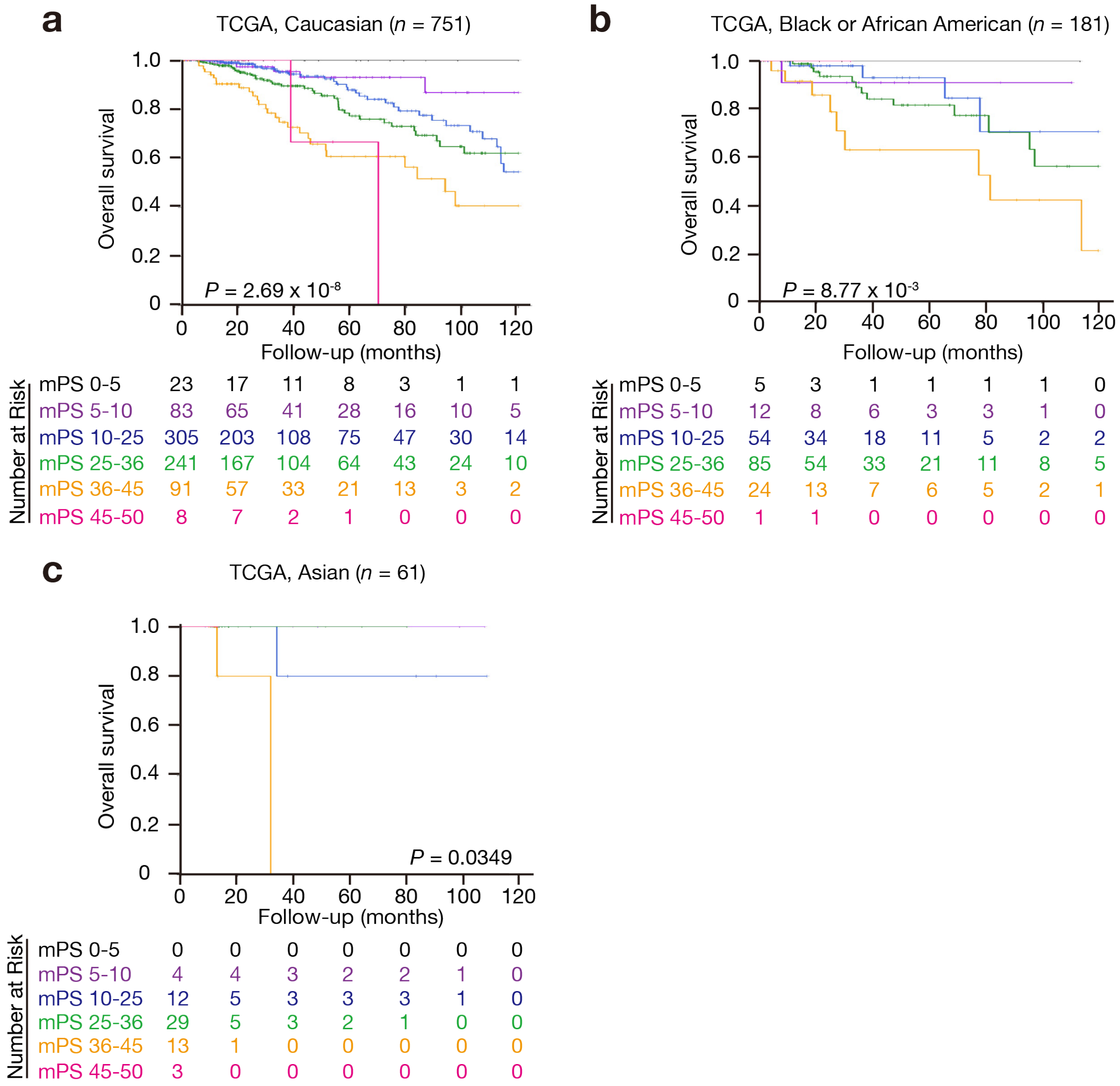
mPS 0-5	12	12	12	12	9	8	7
mPS 5-10	44	44	43	39	37	34	29
mPS 10-25	300	293	276	248	225	195	158
mPS 25-36	256	233	205	187	161	134	113
mPS 36-45	157	141	106	90	76	74	63
mPS 45-50	10	10	8	5	5	3	3

Number at Risk

mPS 0-5	5	5	5	5	5	5	5
mPS 5-10	12	11	11	10	8	7	6
mPS 10-25	52	48	45	43	37	28	23
mPS 25-36	20	19	16	14	12	9	9
mPS 36-45	4	4	4	3	2	2	2
mPS 45-50	0	0	0	0	0	0	0

Fig. S8. Kaplan-Meier curves according to mPS for OS of patients in the METABRIC test cohort with IDC (**a**) or MDLC (**b**).

Shimizu et al. Figure S9



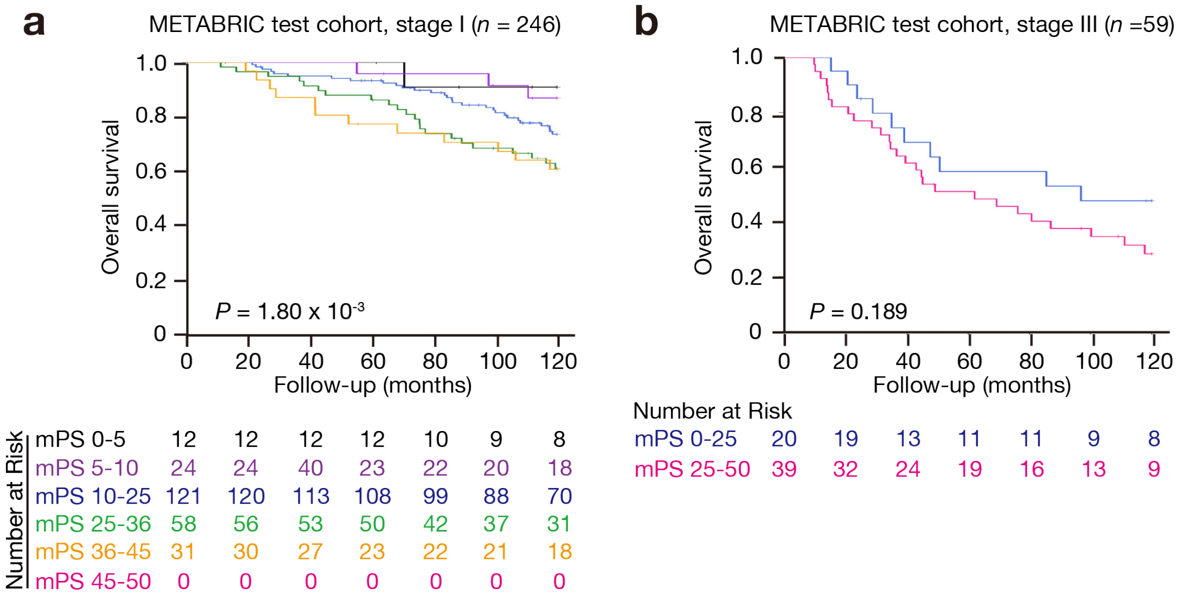
Number at Risk							
mPS 0-5	23	17	11	8	3	1	1
mPS 5-10	83	65	41	28	16	10	5
mPS 10-25	305	203	108	75	47	30	14
mPS 25-36	241	167	104	64	43	24	10
mPS 36-45	91	57	33	21	13	3	2
mPS 45-50	8	7	2	1	0	0	0

Number at Risk							
mPS 0-5	5	3	1	1	1	1	0
mPS 5-10	12	8	6	3	3	1	0
mPS 10-25	54	34	18	11	5	2	2
mPS 25-36	85	54	33	21	11	8	5
mPS 36-45	24	13	7	6	5	2	1
mPS 45-50	1	1	0	0	0	0	0

Number at Risk							
mPS 0-5	0	0	0	0	0	0	0
mPS 5-10	4	4	3	2	2	1	0
mPS 10-25	12	5	3	3	3	1	0
mPS 25-36	29	5	3	2	1	0	0
mPS 36-45	13	1	0	0	0	0	0
mPS 45-50	3	0	0	0	0	0	0

Fig. S9. Stratification of breast cancer patients of different races according to mPS. Kaplan-Meier curves according to mPS were constructed for OS of Caucasian (**a**), black or African-American (**b**), and Asian (**c**) patients in the TCGA breast cancer cohort.

Shimizu et al. Figure S10



Number at Risk

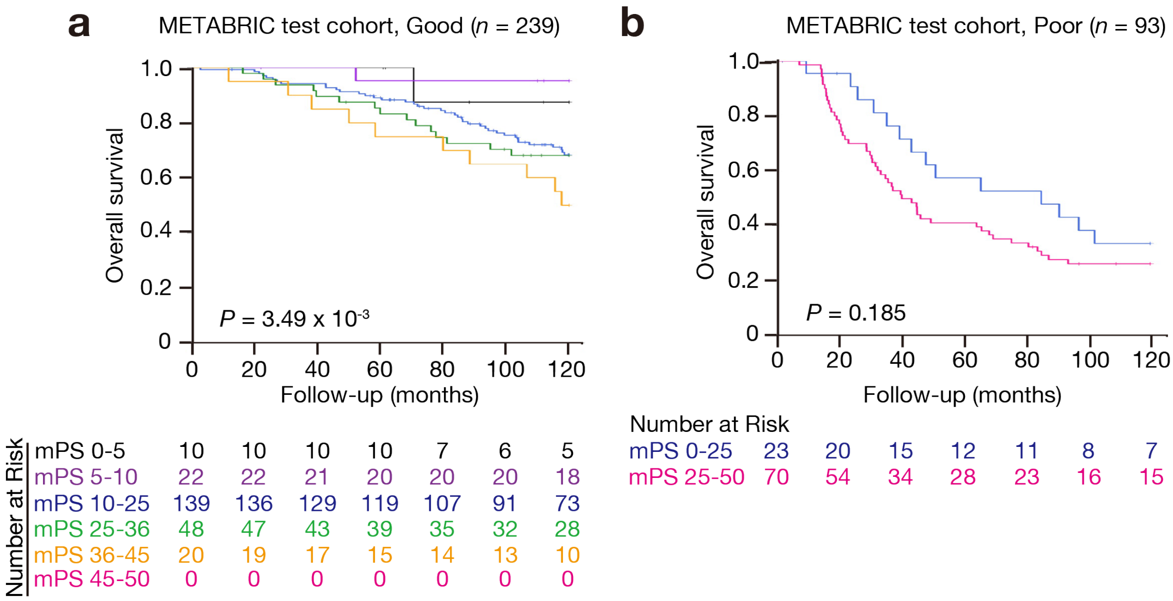
mPS 0-5	12	12	12	12	10	9	8
mPS 5-10	24	24	40	23	22	20	18
mPS 10-25	121	120	113	108	99	88	70
mPS 25-36	58	56	53	50	42	37	31
mPS 36-45	31	30	27	23	22	21	18
mPS 45-50	0	0	0	0	0	0	0

Number at Risk

mPS 0-25	20	19	13	11	11	9	8
mPS 25-50	39	32	24	19	16	13	9

Fig. S10. Kaplan-Meier curves according to mPS for OS of patients in the METABRIC test cohort at clinical TNM stage I (**a**) or III (**b**).

Shimizu et al. Figure S11

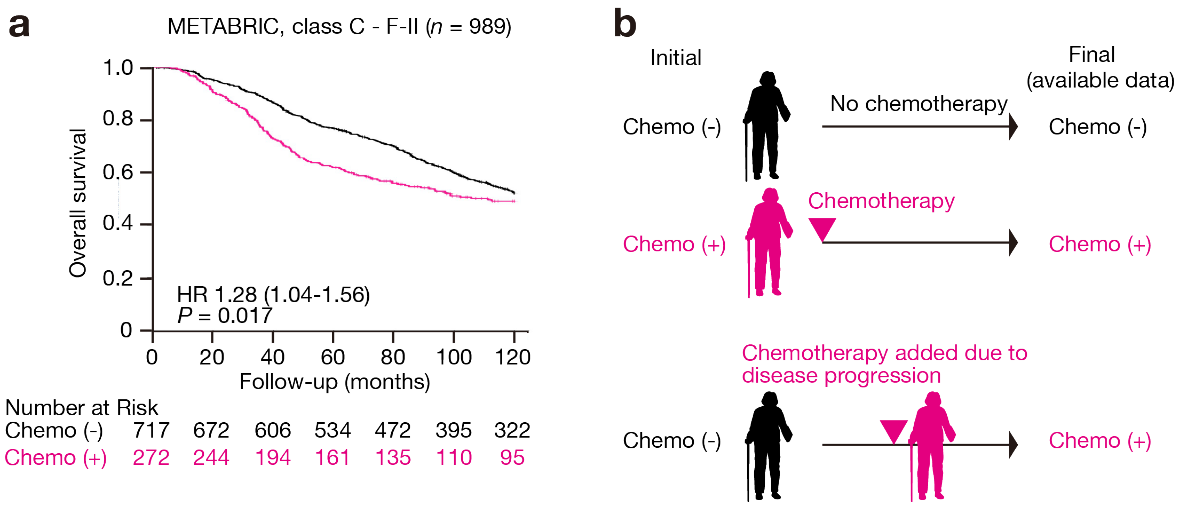


Number at Risk							
mPS 0-5	10	10	10	10	7	6	5
mPS 5-10	22	22	21	20	20	20	18
mPS 10-25	139	136	129	119	107	91	73
mPS 25-36	48	47	43	39	35	32	28
mPS 36-45	20	19	17	15	14	13	10
mPS 45-50	0	0	0	0	0	0	0

Number at Risk							
mPS 0-25	23	20	15	12	11	8	7
mPS 25-50	70	54	34	28	23	16	15

Fig. S11. Stratification of patients according to mPS regardless of NPI. Kaplan-Meier curves were constructed according to mPS for OS of patients in the METABRIC test cohort assigned to the NPI clusters of Good (**a**) or Poor (**b**). Even in the Poor (NPI > 5.40) group, mPS-high patients tend to show a worse prognosis than mPS-low patients.

Shimizu et al. Figure S12



Number at Risk

Chemo (-)	717	672	606	534	472	395	322
Chemo (+)	272	244	194	161	135	110	95

Fig. S12. Relation of chemotherapy to OS in the METABRIC cohort. (**a**) Kaplan-Meier curves for patients in class C, D, E, F-I, or F-II according to whether they received cytotoxic chemotherapy or not during the follow-up time. (**b**) Limited availability of clinical data. Evaluation of potential utility as a predictive score requires information regarding whether the patient received chemotherapy at initial diagnosis. The available data, however, reflect the final status of chemotherapy (performed or not), which means that even if chemotherapy was performed because of disease progression or relapse, the final chemotherapy status is recorded as “Yes” in this data set.