

**UNIVERSIDADE DE BRASÍLIA
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS MÉDICAS**

RAFAEL DE NEGREIROS BOTAN

**CorePAM: ESCORE PROGNÓSTICO DE 24 GENES
DERIVADO DO PAM50 COM VALIDAÇÃO EXTERNA
CROSS-PLATAFORMA PARA CÂNCER DE MAMA**

**Brasília — DF
2026**

RAFAEL DE NEGREIROS BOTAN

**CorePAM: Escore prognóstico de 24 genes derivado do
PAM50 com validação externa *cross-plataforma* para câncer de
mama**

Tese apresentada ao Programa de Pós-Graduação em Ciências Médicas da Faculdade de Medicina da Universidade de Brasília como requisito parcial à obtenção do grau de Doutor em Ciências Médicas.

Área de Concentração: Medicina

Orientador: Prof. Dr. João Batista de Sousa

Brasília — DF

2026

Ficha catalográfica

Espaço reservado para a ficha catalográfica, a ser elaborada pela Biblioteca Central da Universidade de Brasília (BCE/UnB) e inserida no verso da folha de rosto.

FOLHA DE APROVAÇÃO

RAFAEL DE NEGREIROS BOTAN

CorePAM: Escore prognóstico de 24 genes derivado do PAM50 com validação externa *cross-plataforma* para câncer de mama

Tese apresentada ao Programa de Pós-Graduação em Ciências Médicas da Faculdade de Medicina da Universidade de Brasília como requisito parcial à obtenção do grau de Doutor em Ciências Médicas.

Tese defendida e aprovada em 22 de junho de 2026, na Faculdade de Medicina da Universidade de Brasília.

BANCA EXAMINADORA

Prof. Dr. João Batista de Sousa — Presidente da banca e orientador

Faculdade de Medicina, Universidade de Brasília (UnB)

Profa. Dra. Doralina do Amaral Rabello Ramos — Membro titular

Departamento de Patologia, Faculdade de Medicina, Universidade de Brasília (UnB)

Prof. Dr. Hélio Humberto Angotti Carrara — Membro titular externo

Departamento de Ginecologia e Obstetrícia, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo (FMRP-USP)

Prof. Dr. Ruffo de Freitas Junior — Membro titular externo

Professor Titular, Faculdade de Medicina, Universidade Federal de Goiás (UFG)

Dr. João Nunes de Matos Neto — Membro suplente externo

Oncologista clínico do Cettro/Oncoclínicas, Centro de Câncer de Brasília

DEDICATÓRIA

Dedico este trabalho àqueles que, a partir dele, enxergarão algo que ninguém ainda enxergou.

Aos pesquisadores anônimos cujo trabalho silencioso sustenta cada linha do que escrevi. A eles, minha gratidão; a eles, esta tese também pertence.

Mais do que a qualquer outro, dedico-a aos que vierem depois — aos alunos, aos clínicos, aos pesquisadores que um dia abrirem estas páginas procurando não uma resposta, mas um ponto de apoio para erguer um novo ciclo. Melhorem esta tese; rasguem-na, se for preciso. Enxerguem mais longe do que eu consegui.

Porque a ciência, no fim, é isso: uma corrente de gente comum tentando, cada um à sua vez, em silêncio, deixar o mundo um pouco melhor para quem vem depois.

Que esta tese sirva para despertar em outros a paixão que um dia despertaram em mim.

AGRADECIMENTOS

Toda tese é, em segredo, uma obra coletiva. A minha não é exceção. Carrego dentro dela pedaços de muita gente — e é justo nomeá-los aqui, ainda que palavra alguma dê conta do que devo a cada um.

Às minhas filhas, **Alice** e **Bruna**, que desde muito cedo aprenderam a conviver com um pai dividido entre o consultório, o hospital e a tela do computador. Vocês foram, sem saber, o motivo silencioso de cada linha escrita tarde da noite. Espero que, um dia, ao lerem estas páginas, encontrem nelas um convite — à curiosidade, à persistência e à coragem de fazer perguntas difíceis.

À **Carol**, minha esposa, que segurou o mundo enquanto eu tentava compreender uma parte pequena dele. Sua paciência, seu cuidado e seu amor são o chão sobre o qual este trabalho foi construído. Sem você, nada disso seria possível — e menos ainda, valeria a pena.

À minha **mãe**, que me ensinou, antes de qualquer livro, que estudar é uma forma de retribuir ao mundo o que ele nos dá. Este doutorado é, também, seu.

Ao meu orientador, **Prof. Dr. João Batista de Sousa**, pela confiança depositada desde o início, pela liberdade intelectual que me concedeu para percorrer este caminho, e pela paciência com que acompanhou cada etapa deste doutorado. Agradeço a orientação firme quando precisei de direção, e o silêncio generoso quando precisei de espaço para errar e

descobrir por conta própria. Sua leitura crítica e suas sugestões metodológicas estão espalhadas por cada capítulo deste trabalho.

Ao **Sérgio Arruda** — a corda do meu cavaquinho. Amigo querido, de quem aprendi menos sobre ciência do que sobre algo mais raro e mais difícil de ensinar: carinho, generosidade e escuta. Uma das pessoas mais fantásticas que já conheci. Obrigado por estar por perto quando importou.

À **banca de qualificação e de defesa**, pelas críticas rigorosas e pelas sugestões que tornaram este trabalho melhor do que eu teria conseguido sozinho.

À **Universidade de Brasília**, ao **Programa de Pós-Graduação em Ciências Médicas** e ao **Hospital Universitário de Brasília (HUB)**, pelo ambiente institucional que tornou esta pesquisa possível.

Aos **consórcios internacionais SCAN-B, TCGA, METABRIC, GEO e I-SPY**, cuja decisão de compartilhar dados publicamente é, por si só, um ato científico e ético. Esta tese é literalmente feita com dados que outros escolheram não manter para si.

Por fim, às **pacientes** — as verdadeiras protagonistas de cada número e de cada curva de sobrevida destas páginas. Que este trabalho sirva, ainda que de forma modesta, para aproximar um passo a mais o dia em que seus nomes não precisem mais estar em coortes de risco.

EPÍGRAFE

“O cientista não busca verdades absolutas, mas aproximações cada vez mais honestas da realidade.

O que o move não é o desejo de estar certo, mas o compromisso de não enganar — nem a si, nem aos outros.

Cada modelo é uma hipótese em tradução; cada resultado, uma forma de escutar a natureza sem distorções.”

— Karl Popper (livre adaptação)

RESUMO

Objetivo: O classificador PAM50 prediz o prognóstico do câncer de mama, mas requer a medição simultânea de 50 genes e plataformas especializadas, com custo elevado e pouco acesso no sistema público brasileiro. O objetivo deste trabalho foi derivar o CorePAM, o menor subconjunto do PAM50 capaz de manter o desempenho prognóstico original, e validar externamente sua performance em múltiplas coortes e plataformas.

Método: Foi aplicada regressão de Cox penalizada com *elastic-net* e validação cruzada determinística de 10 subamostras na coorte SCAN-B (N = 3 069). A seleção de genes respeitou margem pré-especificada de não inferioridade em relação ao modelo completo de 50 genes. A validação externa utilizou quatro coortes independentes: TCGA-BRCA (N = 1 072), METABRIC (N = 1 978), GSE20685 (N = 327) e GSE1456 (N = 159), totalizando 3 536 pacientes em plataformas distintas (*RNA-seq* e *microarray*). Análises secundárias avaliaram a predição de resposta patológica completa (pCR) em quatro coortes neoadjuvantes (N = 697) e em uma coorte contemporânea exploratória (I-SPY2, N = 986).

Resultados: O CorePAM é composto por 24 genes e reproduziu o desempenho prognóstico do painel completo. O score mostrou-se independentemente associado à sobrevida em todas as coortes externas — TCGA-BRCA (HR = 1,20), METABRIC (HR = 1,41), GSE20685 (HR = 1,40) e GSE1456 (HR = 1,71); todos $p < 0,02$. A meta-análise de efeitos aleatórios produziu *hazard ratio* agrupado de 1,37 (IC 95% 1,24–1,52; $p < 0,001$). O CorePAM manteve significância prognóstica após ajuste para tamanho tumoral e acometimento linfonodal. Para pCR, o *odds ratio* agrupado foi 1,69 (IC 95% 1,39–2,05; $p < 0,001$).

Conclusões: O CorePAM preservou o desempenho prognóstico do PAM50 com menos da metade dos genes originais e mostrou-se robusto entre coortes, tecnologias e populações. A redução pode simplificar o desenvolvimento de ensaios futuros mais acessíveis, especialmente para sistemas de saúde com restrições orçamentárias como o SUS.

Descritores: Câncer de mama. Expressão gênica. Assinatura prognóstica. PAM50. Análise de sobrevida. Resposta patológica completa. Validação *cross-plataforma*.

ABSTRACT

Objective: The PAM50 classifier predicts breast cancer prognosis but requires simultaneous measurement of 50 genes on specialised platforms, a cost-intensive assay with limited access in public health systems. This work aimed to derive CorePAM, the smallest subset of PAM50 able to retain prognostic performance, and to externally validate its transportability across cohorts and platforms.

Methods: Cox regression with *elastic-net* penalisation and deterministic 10-fold cross-validation was applied in the SCAN-B cohort (N = 3,069). Gene selection followed a pre-specified non-inferiority margin relative to the full 50-gene comparator. External validation used four independent cohorts: TCGA-BRCA (N = 1,072), METABRIC (N = 1,978), GSE20685 (N = 327) and GSE1456 (N = 159), totalling 3,536 patients across distinct platforms (*RNA-seq* and *microarray*). Secondary analyses evaluated pathologic complete response (pCR) prediction in four neoadjuvant cohorts (N = 697) and in one contemporary exploratory cohort (I-SPY2, N = 986).

Results: CorePAM comprises 24 genes and reproduced the prognostic performance of the complete panel. The *score* was independently associated with survival in every validation cohort — TCGA-BRCA (HR = 1.20), METABRIC (HR = 1.41), GSE20685 (HR = 1.40) and GSE1456 (HR = 1.71); all $p < 0.02$. Random-effects meta-analysis produced a pooled *hazard ratio* of 1.37 (95% CI 1.24–1.52; $p < 0.001$). CorePAM retained prognostic significance after adjustment for tumour size and nodal involvement. For pCR, the pooled *odds ratio* was 1.69 (95% CI 1.39–2.05; $p < 0.001$).

Conclusions: CorePAM preserved PAM50 prognostic performance with fewer than half the original genes and proved robust across cohorts, technologies and populations. This reduction may simplify the development of more accessible future assays — particularly relevant for health systems with resource constraints such as the Brazilian public system.

Key words: Breast cancer. Gene expression. Prognostic signature. PAM50. Survival analysis. Pathologic complete response. Cross-platform validation.

LISTA DE FIGURAS

Figura	Página
Figura 1 — PAM50, centroides moleculares e a pergunta do CorePAM	23
Figura 2 — Subtipos intrínsecos e correlatos clínicos do câncer de mama	25
Figura 3 — Linha do tempo das assinaturas gênicas em câncer de mama	27
Figura 4 — Desenho do estudo CorePAM. Derivação na coorte SCAN-B (N = 3 069)...	36
Figura 5 — Fluxo das análises estatísticas do estudo	37
Figura 6 — Fluxograma das coortes do estudo	42
Figura 7 — Esquema de derivação do CorePAM	47
Figura 8 — Fronteira de Pareto: C-index out-of-fold em função do número de genes (graus...	56
Figura 9 — Pesos dos 24 genes CorePAM (gráfico lollipop)	58
Figura 10 — Estabilidade bootstrap da seleção de genes (B = 200 reajustes no SCAN-B)	60
Figura 11 — Análise de sensibilidade leave-one-gene-out	62
Figura 12 — Curvas de Kaplan-Meier nas cinco coortes (painel múltiplo)	65
Figura 13 — Curva de Kaplan-Meier na coorte de treinamento SCAN-B (N = 3 069; sobrevida...	67
Figura 14 — Curva de Kaplan-Meier na coorte TCGA-BRCA (N = 1 072; RNA-seq; sobrevida...	68
Figura 15 — Curva de Kaplan-Meier na coorte METABRIC (N = 1 978; microarray Illumina...	69
Figura 16 — Curva de Kaplan-Meier na coorte GSE20685 (N = 327; microarray Affymetrix...	70
Figura 17 — Forest plot da meta-análise de sobrevida (HR por 1-DP, K = 4 coortes...	71
Figura 18 — Delta-C-index incremental do CorePAM sobre o modelo clínico CORE-A	74
Figura 19 — Curvas de calibração (risco predito vs observado) em quatro coortes	76
Figura 20 — Correlação entre escores intra-coorte e z-score congelado (parâmetros SCAN-B)...	79
Figura 21 — Análise de curva de decisão (DCA) — quatro coortes de sobrevida	81
Figura 22 — PCA forense da matriz de expressão METABRIC	83
Figura 23 — Forest plot de subgrupos por status de receptor de estrogênio (RE-positivo...	85
Figura 24 — Forest plot de pCR (OR por 1-DP, K = 4 coortes primárias)	87

Figura	Página
Figura 25 — Curvas ROC para predição de pCR em cada coorte neoadjuvante	88
Figura 26 — Distribuição do escore CorePAM por status de pCR em cada coorte neoadjuvante	89
Figura 27 — Taxa de pCR por quartil do escore CorePAM em cada coorte neoadjuvante	90
Figura 28 — Análise de curva de decisão para predição de pCR nas quatro coortes...	91
Figura 29 — Comparação head-to-head CorePAM (24 genes) vs PAM50-full penalizado (49...	94
Figura 30 — Comparação de C-index: CorePAM vs ROR-S vs OncotypeDX-RS (implementações...	96

LISTA DE TABELAS

Tabela	Página
Tabela 1 — Mapa dos blocos de coortes e das perguntas metodológicas do estudo	30
Tabela 2 — Papel de cada coorte na construção e validação do CorePAM	31
Tabela 3 — Características das coortes de sobrevida	41
Tabela 4 — Características das coortes neoadjuvantes (pCR)	42
Tabela 5 — Parâmetros analíticos congelados antes da análise	43
Tabela 6 — Cobertura de genes CorePAM por coorte de validação	45
Tabela 7 — Lista completa dos 24 genes CorePAM com pesos elastic-net	57
Tabela 8 — Desempenho prognóstico do CorePAM por coorte	63
Tabela 9 — Meta-análise de efeitos aleatórios — sobrevida ($K = 4$)	71
Tabela 10 — C-index incremental do CorePAM sobre o modelo CORE-A	73
Tabela 11 — Modelo clínico expandido com estadiamento anatômico	77
Tabela 12 — Modelo clínico completo (6 variáveis) com e sem CorePAM	77
Tabela 13 — Análise de sensibilidade do z-score congelado	77
Tabela 14 — Estatística C de Uno e AUC tempo-dependente	80
Tabela 15 — Calibração formal por coorte	80
Tabela 16 — Análise de subgrupo por status de RE em SCAN-B ($N = 2\,870$; OS) e METABRIC (N...)	84
Tabela 17 — Predição de pCR por coorte neoadjuvante	86
Tabela 18 — Comparação head-to-head: CorePAM (24 genes) vs comparador PAM50-full...	93
Tabela 19 — Comparação com ROR-S e OncotypeDX-RS (implementações genefu)	95
Tabela 20 — Organização do pipeline computacional por blocos	119
Tabela 21 — Lista completa dos scripts do pipeline	119

LISTA DE QUADROS

Quadro	Página
Quadro 1 — Principais assinaturas gênicas em câncer de mama	29

LISTA DE ABREVIATURAS E SIGLAS

- ABNT** — Associação Brasileira de Normas Técnicas
- AUC** — *area under the curve* (área sob a curva ROC)
- CEP** — Comitê de Ética em Pesquisa
- CNS** — Conselho Nacional de Saúde
- ComBat** — *Combating Batch Effects* (método de correção de *batch*)
- CORE-A** — modelo clínico mínimo utilizado como referência multivariada
- CorePAM** — *Core* PAM50 (escore prognóstico de 24 genes derivado neste estudo)
- CPM** — *counts per million* (contagens por milhão)
- DCA** — *decision curve analysis* (análise de curva de decisão)
- DNA** — ácido desoxirribonucleico
- DOI** — *Digital Object Identifier*
- DP** — desvio-padrão
- DSS** — *disease-specific survival* (sobrevida doença-específica)
- EP** — erro-padrão
- RE / ER** — receptor de estrogênio / *estrogen receptor*
- FFPE** — *formalin-fixed paraffin-embedded* (tecido fixado em formalina e embebido em parafina)
- GEO** — *Gene Expression Omnibus*
- GLOBOCAN** — *Global Cancer Observatory*
- HER2** — *human epidermal growth factor receptor 2* (receptor 2 do fator de crescimento epidérmico humano)
- HR** — *hazard ratio* (razão de riscos)
- HUB** — Hospital Universitário de Brasília
- IC 95%** — intervalo de confiança de 95%

INCA — Instituto Nacional de Câncer

I-SPY — *Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging and Molecular Analysis*

K — número de *folds* da validação cruzada

KM — Kaplan–Meier

L1 / L2 — penalidades tipo *lasso* (L1) e *ridge* (L2)

LLM — *large language model* (modelo de linguagem de grande escala)

METABRIC — *Molecular Taxonomy of Breast Cancer International Consortium*

MINDACT — *Microarray In Node-negative and 1 to 3 positive lymph node Disease may Avoid ChemoTherapy*

NRI / IDI — *net reclassification improvement / integrated discrimination improvement*

OOF — *out-of-fold* (predição fora do *fold*)

OR — *odds ratio* (razão de chances)

OS — *overall survival* (sobrevida global)

PAM50 — *Prediction Analysis of Microarray 50*

pCR — *pathologic complete response* (resposta patológica completa)

PCR — *polymerase chain reaction* (reação em cadeia da polimerase)

QALY — *quality-adjusted life year* (ano de vida ajustado pela qualidade)

REML — *restricted maximum likelihood* (máxima verossimilhança restrita)

RMA — *robust multi-array average*

RNA — ácido ribonucleico

RNA-seq — sequenciamento de RNA

ROC — *receiver operating characteristic*

ROR / ROR-S — *Risk of Recurrence score* / ROR baseado em subtipo

RP — receptor de progesterona

RT-qPCR — *reverse transcription quantitative PCR*

SCAN-B — *Sweden Cancerome Analysis Network — Breast*

SHA-256 — *Secure Hash Algorithm 256 bits*

STAR/GDC — alinhador STAR / *Genomic Data Commons*

SUS — Sistema Único de Saúde

TAILORx — *Trial Assigning Individualized Options for Treatment*

TCGA — *The Cancer Genome Atlas*

TCGA-BRCA — braço de câncer de mama do TCGA

TMM — *trimmed mean of M-values*

TRIPOD — diretriz para relato de modelos prognósticos multivariáveis

SUMÁRIO

Seção	Página
1 INTRODUÇÃO	21
1.1 Epidemiologia e carga global da doença	21
1.2 Heterogeneidade tumoral e justificativa da estratificação molecular	21
1.3 Subtipos moleculares intrínsecos e o PAM50	22
1.4 Assinaturas gênicas: panorama atual e evidência clínica	26
1.5 Assinaturas gênicas, plataformas e transportabilidade	27
1.6 Coortes públicas de câncer de mama	30
1.7 A pergunta central	32
2 OBJETIVOS	34
2.1 Objetivo geral	34
2.2 Objetivos específicos	34
3 MÉTODO	35
3.1 Delineamento do estudo e aspectos éticos	35
3.1.1 Tipo de estudo	35
3.1.2 Aspectos éticos	38
3.1.3 Princípios de validação externa	39
3.1.4 Disponibilidade de dados e código	39
3.1.5 Uso de modelos de linguagem de grande escala	39
3.2 Coortes do estudo	40
3.3 Congelamento analítico (analytical freeze)	43
3.3.1 Margem de não inferioridade (Delta-C-index = 0,010)	44
3.3.2 Atribuição determinística de folds (SHA-256)	44
3.3.3 Estratificação por evento	44
3.4 Pré-processamento de expressão gênica	45
3.4.1 Coortes RNA-seq (SCAN-B, TCGA-BRCA)	45
3.4.2 Coortes microarray (METABRIC, GSE20685, GSE1456, GSE25066, GSE20194,	45
3.5 Derivação do CorePAM	47
3.5.1 Regressão de Cox elastic-net: justificativa e parâmetros	48
3.5.2 Validação cruzada out-of-fold	48
3.5.3 Critério de seleção: não inferioridade e fronteira de Pareto	49
3.5.4 Análises de estabilidade da seleção	49

Seção	Página
3.6 Cálculo do escore	49
3.6.1 Pontuação congelada (single-sample scoring)	50
3.7 Análise de sobrevida	50
3.7.1 Modelos de Cox	50
3.7.2 Métricas de discriminação	51
3.7.3 Curvas de Kaplan-Meier	51
3.7.4 Valor incremental: Delta-C-index	51
3.7.5 Calibração	52
3.7.6 Análise de curva de decisão (DCA)	52
3.7.7 Modelos expandidos com estadiamento	53
3.8 Meta-análise	53
3.9 Comparações head-to-head	54
4 RESULTADOS	55
4.1 Derivação do CorePAM: os 24 genes	55
4.2 Validação prognóstica externa	63
4.3 Meta-análise de sobrevida	71
4.4 Valor incremental sobre preditores clínicos	73
4.5 Independência do estadiamento clínico-patológico	77
4.6 Análises de sensibilidade	77
4.7 Análise de subgrupo por status de RE	84
4.8 Resposta patológica completa (pCR)	86
4.9 Comparação head-to-head com PAM50-full	93
4.10 Comparação com assinaturas estabelecidas	95
5 DISCUSSÃO	97
5.1 O CorePAM no contexto da medicina de precisão em oncologia	97
5.2 Derivação reprodutível: inovação metodológica	98
5.3 Coerência biológica: por que esses 24 genes?	99
5.4 Transportabilidade cross-plataforma	100
5.5 O paradoxo de menos genes, mais informação	101
5.6 A questão do z-score: coorte vs amostra única	102
5.7 O sinal em doença RE-positiva	102
5.8 A ponte entre prognóstico e predição de pCR	103
5.9 Calibração: boa discriminação não garante probabilidades corretas	104
5.10 Reflexões metodológicas para estudos prognósticos futuros	104

Seção	Página
5.11 O caminho até a clínica: o que falta?	105
5.12 Relevância para o contexto brasileiro e o SUS	105
5.13 Limitações	106
5.14 Perspectivas e estudos futuros	106
6 CONCLUSÃO	108
REFERÊNCIAS BIBLIOGRÁFICAS	109
APÊNDICE A — Pipeline computacional	118
APÊNDICE B — Artigo publicado no Breast Cancer Research	129

1. INTRODUÇÃO

1.1 Epidemiologia e carga global da doença

O câncer de mama é a neoplasia maligna mais frequentemente diagnosticada em mulheres no mundo. Em 2020, tornou-se também o câncer mais diagnosticado na população global, ultrapassando o pulmão¹; nas estimativas do GLOBOCAN 2022, o câncer de pulmão retomou a primeira posição global (ambos os sexos), enquanto a mama permaneceu o câncer mais incidente em mulheres. Dados mais recentes do GLOBOCAN 2022 estimam 2,31 milhões de novos diagnósticos e 666 103 mortes por câncer de mama naquele ano, configurando a principal causa de morte por câncer em mulheres em mais de 110 países². A incidência cresce em praticamente todas as regiões do mundo, mas os padrões de mortalidade são drasticamente heterogêneos: enquanto países de alta renda apresentam taxas de mortalidade em declínio há mais de duas décadas — fruto combinado de diagnóstico precoce, terapia endócrina eficaz e avanços em quimioterapia e drogas anti-HER2^{3,4} —, países de média e baixa renda ainda veem mortalidade estável ou crescente.

No Brasil, o Instituto Nacional de Câncer (INCA) projetou cerca de 74 mil novos casos anuais para o triênio 2023-2025⁵, o que posiciona o câncer de mama como a neoplasia maligna mais frequente em mulheres brasileiras em todas as regiões, com exceção do Norte (onde é superada pelo câncer de colo uterino). A mortalidade por câncer de mama no Brasil, diferentemente de países com programas maduros de rastreamento mamográfico, apresenta tendência histórica de estabilidade ou discreto crescimento, com desigualdades regionais marcadas — expressas em idade mais jovem ao diagnóstico, estadiamento clínico mais avançado à apresentação e pior acesso a tratamentos sistêmicos de última geração⁶. Essas desigualdades são uma das motivações de fundo para o presente trabalho: qualquer ferramenta prognóstica que dependa de plataformas de teste inacessíveis ao sistema público perpetua, na prática, a iniquidade de acesso à medicina de precisão.

1.2 Heterogeneidade tumoral e justificativa da estratificação molecular

A heterogeneidade biológica do câncer de mama é um dos fenômenos mais bem caracterizados da oncologia moderna. Dois tumores clinicamente semelhantes — mesmo tamanho, mesmo grau histológico, mesmo perfil imuno-histoquímico — podem ter comportamentos biológicos drasticamente diferentes, com respostas terapêuticas e desfechos

a longo prazo discordantes⁷. Esta heterogeneidade ocorre em pelo menos três níveis hierárquicos: inter-tumoral (diferenças entre pacientes), intra-tumoral espacial (diferentes regiões de um mesmo tumor) e temporal (evolução clonal sob pressão terapêutica). A classificação imuno-histoquímica tradicional (RE, RP, HER2, Ki67), embora operacional e universalmente disponível, oferece uma aproximação grosseira dessa complexidade biológica. A constatação de que expressão gênica em larga escala revela uma estrutura de subtipos biologicamente coerente, consolidada ao longo das duas últimas décadas⁸⁻¹⁰, motivou o desenvolvimento sistemático de ferramentas de classificação molecular capazes de refletir essa estrutura subjacente e orientar decisões terapêuticas individualizadas.

1.3 Subtipos moleculares intrínsecos e o PAM50

Os trabalhos seminais de Perou et al. e Sørlie et al.^{8,9} inauguraram a era da classificação molecular do câncer de mama ao demonstrar, por meio de perfis de expressão gênica obtidos por *microarray*, que os tumores mamários se organizam em subtipos intrínsecos com comportamentos clínicos distintos. Essa taxonomia original descreveu cinco grupos: *Luminal A* (RE-positivo, baixa proliferação), *Luminal B* (RE-positivo, alta proliferação), *HER2-enriched* (subtipo transcricional definido por correlação com o centroide, com expressão elevada do cluster HER2/17q, frequentemente mas não obrigatoriamente coincidente com HER2 clínico), *Basal-like* (fenótipo expressivo semelhante ao epitélio basal/mioepitelial da mama normal) e *Normal-like* (grupo de fronteira, parcialmente artefato de contaminação estromal). Estudos subsequentes incorporaram o subtipo *Claudin-low* e refinamentos como a estratificação em dez subgrupos integrativos do METABRIC, que combinam expressão e alterações numéricas somáticas^{11,12}.

A transição dessa taxonomia para uso clínico exigiu um classificador robusto, treinado em subtipos centroides, e independente da plataforma específica utilizada no treinamento. O PAM50 (*Prediction Analysis of Microarray 50*), desenvolvido por Parker et al.¹³, consolidou esse papel ao definir a classificação com base em exatamente 50 genes e validar sua transferência entre plataformas *microarray* e, posteriormente, RT-qPCR e *RNA-seq*^{14,15}. O PAM50 atribui cada tumor a um dos subtipos intrínsecos por correlação com centroides de expressão e, por meio do *Risk of Recurrence Score* (ROR), fornece um escore prognóstico numérico que integra subtipo, proliferação e tamanho tumoral. Sua implementação comercial — o ensaio *Prosigna*, executado na plataforma *NanoString nCounter* — possui validação analítica rigorosa¹⁶ e evidência de nível 1 de utilidade clínica em pacientes RE-positivas pós-

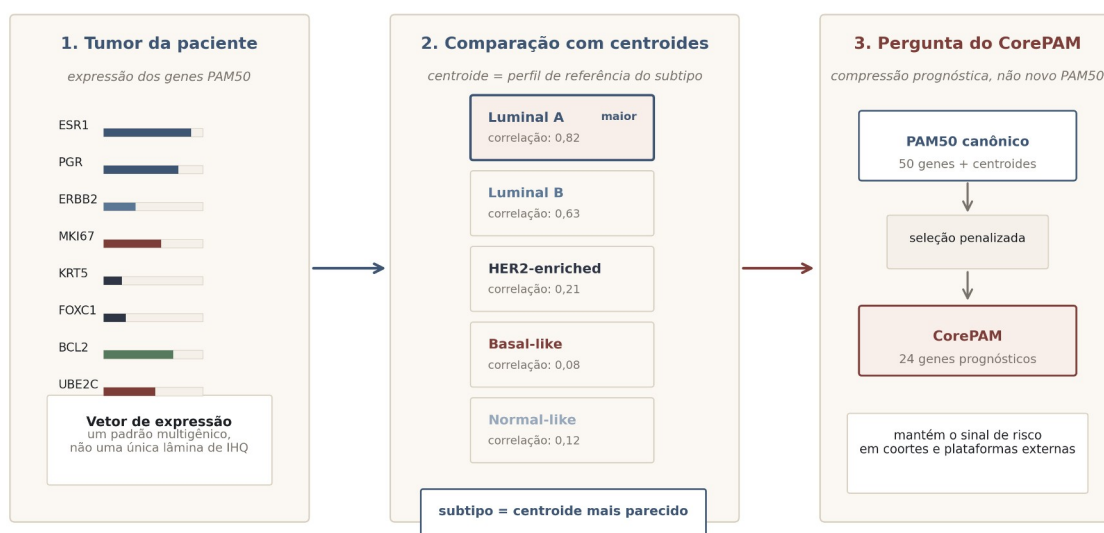
menopausa (estudos TransATAC¹⁷ e ABCSG-8¹⁸), sendo recomendado por diretrizes internacionais^{19–21} para auxílio à decisão de quimioterapia adjuvante em câncer de mama RE-positivo, HER2-negativo, estágio precoce.

Um centroide, nesse contexto, é um perfil de referência: uma espécie de “paciente molecular prototípica” de cada subtipo. O tumor real é convertido em uma lista de expressões gênicas e comparado com esses perfis. Se a expressão do tumor se parece mais com o centroide *Luminal A* do que com os demais, ele recebe a classificação *Luminal A*. Essa comparação é feita no espaço de muitos genes ao mesmo tempo; por isso, não equivale a olhar apenas RE, RP, HER2 ou Ki67. A analogia clínica seria comparar um caso com síndromes bem definidas: não se decide por um único sinal isolado, mas pelo padrão completo.

O CorePAM nasce dessa arquitetura, mas responde a outra pergunta. Ele não redefine os centroides originais do PAM50, não cria uma nova classificação molecular de subtipos e não pretende substituir a leitura clínica do *Prosigna*. A pergunta do CorePAM é mais estreita e mais pragmática: dentro da informação prognóstica contida no painel PAM50, qual subconjunto mínimo de genes preserva a capacidade de ordenar pacientes por risco em coortes externas? Em termos simples, o PAM50 é o atlas molecular completo; o CorePAM tenta conservar as estradas mais úteis para prognóstico, sem redesenhar o mapa inteiro.

PAM50: centroides moleculares e a pergunta do CorePAM

O PAM50 compara a amostra com perfis de referência; o CorePAM pergunta se o sinal prognóstico pode ser comprimido.



Valores de correlação são ilustrativos. IHQ: imuno-histoquímica. O CorePAM não redefine os centroides originais do PAM50; ele reduz o painel para prognóstico.

Figura 1 — PAM50, centroides moleculares e a pergunta do CorePAM. O PAM50 classifica tumores por semelhança com perfis de referência dos subtipos intrínsecos; os valores de correlação na figura são ilustrativos. O CorePAM reduz o painel para uma pergunta prognóstica específica, sem redefinir os centroides originais do PAM50.

Na Figura 1, o primeiro quadro mostra a amostra tumoral como um vetor de expressão: cada gene contribui uma pequena parte da identidade molecular. O segundo quadro mostra a comparação com centroides. O tumor ilustrativo se aproxima mais do centroide *Luminal A*, mas a decisão vem da comparação simultânea com todos os perfis de referência, não de um marcador isolado. O terceiro quadro mostra a diferença conceitual do CorePAM: parte-se do PAM50 canônico de 50 genes, aplica-se uma seleção penalizada voltada a prognóstico e chega-se a 24 genes. A seta não significa que o CorePAM seja “um novo PAM50”; significa que ele é uma compressão prognóstica derivada do espaço PAM50.

É importante separar duas linguagens que se confundiram historicamente. O PAM50 não classifica tumores medindo diretamente RE, RP, HER2 e Ki67; ele compara a expressão de 50 genes com centroides moleculares. Na prática clínica, porém, a imuno-histoquímica passou a ser usada como tradução aproximada desses subtipos porque é barata, padronizada, universalmente disponível e captura os grandes eixos que mais orientam a decisão diária: sinalização estrogênica, ativação da via HER2, proliferação e fenótipo basal. Essa tradução ajudou a levar a taxonomia molecular para a rotina, mas não é equivalente. Tumores com o mesmo painel imuno-histoquímico podem receber subtipos PAM50 diferentes, e tumores classificados como *HER2-enriched* ou *Basal-like* pelo PAM50 nem sempre coincidem perfeitamente com doença HER2-positiva ou triplo-negativa pela definição clínica^{10,14}. Por isso, a Figura 2 deve ser lida como mapa de correlação clínica, não como definição molecular exata dos subtipos.

Subtipos intrínsecos e correlatos clínicos do câncer de mama

PAM50: centroides de expressão gênica; IHQ: aproximação clínica, não equivalência



Leitura correta: PAM50 classifica por centroides de expressão; RE/RP/HER2/Ki67 são correlatos clínicos aproximados.

A categoria Normal-like não é mostrada por não ter substituto IHQ clínico robusto.

Figura 2 — Subtipos intrínsecos e correlatos clínicos do câncer de mama. O PAM50 define subtipos por correlação com centroides de expressão gênica; RE, RP, HER2, Ki67 e marcadores basais mostrados nos quadros representam correlatos imuno-histoquímicos usados historicamente na rotina, não substitutos exatos. A categoria *Normal-like* não é mostrada por não ter correlato imuno-histoquímico clínico robusto.

Na Figura 2, cada quadro deve ser lido em duas camadas. A camada molecular é o nome do subtipo intrínseco, definido pelo PAM50 a partir de centroides de expressão. A camada clínica é o conjunto de marcadores de imuno-histoquímica mostrado abaixo, que funciona como aproximação prática. No quadro *Luminal A*, o leitor vê o padrão de menor proliferação e melhor prognóstico; no *Luminal B*, a presença de Ki67 mais alto indica maior velocidade biológica, mesmo quando o tumor continua sendo RE-positivo. O quadro *HER2-enriquecido* mostra uma armadilha conceitual importante: esse subtipo molecular frequentemente, mas não obrigatoriamente, coincide com HER2 positivo clínico. O quadro *Basal-like* segue a mesma lógica: costuma se aproximar do triplo-negativo, mas não é sinônimo perfeito dele. A figura, portanto, deve ser entendida como se entende um mapa clínico: útil para orientar o trajeto, mas não idêntico ao território molecular.

1.4 Assinaturas gênicas: panorama atual e evidência clínica

O PAM50 não é a única assinatura gênica em uso clínico. O *OncotypeDX Recurrence Score*²² mede 21 genes por RT-qPCR em tecido *FFPE* e foi validado prospectivamente pelo estudo TAILORx²³, que demonstrou não inferioridade de endocrinoterapia isolada em pacientes RE-positivas com escore intermediário (11-25) na faixa etária acima de 50 anos, reescrevendo a conduta para aproximadamente 70% das pacientes com doença RE-positiva linfonodo-negativa. Análises posteriores do TAILORx²⁴ e o ensaio RxPONDER²⁵ estenderam a aplicação do *OncotypeDX* para mulheres com 1-3 linfonodos positivos, consolidando sua utilidade em cenários de decisão clínica específicos. O *MammaPrint*²⁶, assinatura de 70 genes originalmente derivada para risco de recorrência a cinco anos em população RE+/RE-, teve sua utilidade clínica demonstrada pelo estudo MINDACT^{27,28}, que mostrou que pacientes com alto risco clínico mas baixo risco genômico podem evitar quimioterapia com segurança. Outras assinaturas incluem o *EndoPredict*²⁹ (12 genes, foco em recorrência tardia), o Breast Cancer Index³⁰ (genes HOXB13/IL17BR mais módulo de proliferação) e o ROR/*Prosigna*^{13,16}. A comparação *head-to-head* entre essas ferramentas³¹ mostrou que, embora todas ofereçam informação prognóstica significativa, a correlação entre os escores individuais é apenas moderada, indicando que não medem exatamente o mesmo construto biológico.

A distinção fundamental, relevante para o presente trabalho, é que o *OncotypeDX* e o *MammaPrint* não são subconjuntos do PAM50 — foram desenvolvidos independentemente, a partir de pools gênicos distintos e com metodologias e *endpoints* diferentes. Até o presente, nenhuma redução sistemática e orientada por dados do próprio painel PAM50 havia sido tentada usando regressão penalizada moderna com critério pré-especificado de não inferioridade.

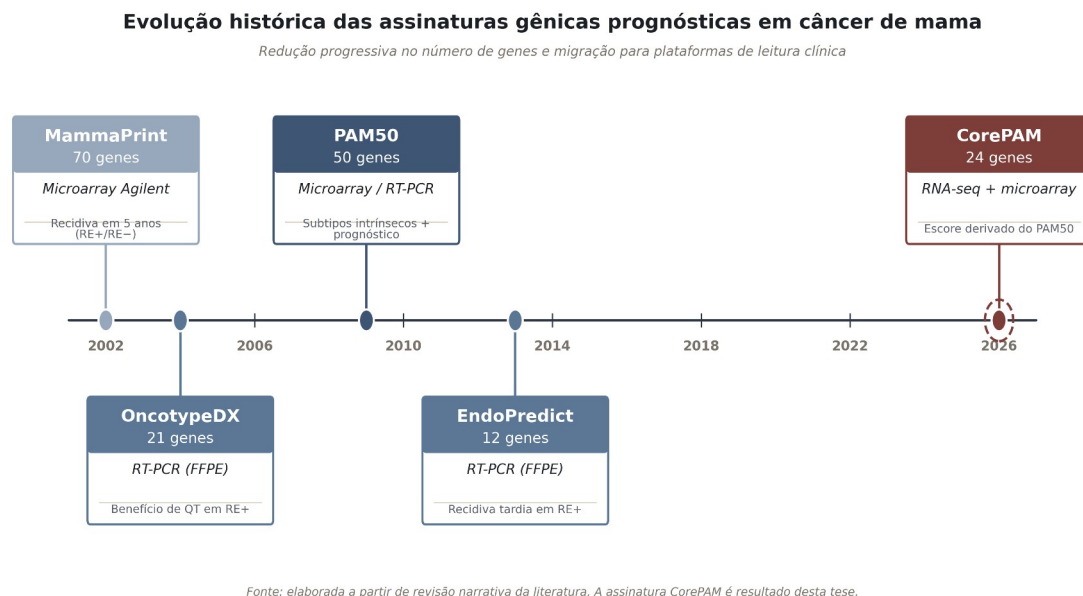


Figura 3 — Linha do tempo das assinaturas gênicas em câncer de mama. Panorama histórico das assinaturas com validação clínica (*OncotypeDX*, *MammaPrint*, *PAM50/Prosigna*, *EndoPredict*, *Breast Cancer Index*) e seus estudos pivô correspondentes, ilustrando a evolução do campo desde os perfis transcricionais seminais até os ensaios prospectivos que consolidaram a utilidade clínica.

Na Figura 3, o eixo horizontal conta a história de uma mudança de paradigma. À esquerda estão as primeiras assinaturas desenvolvidas em *microarray*, quando a principal pergunta era descobrir se padrões de expressão gênica conseguiam antecipar recorrência. O *MammaPrint* aparece como assinatura de 70 genes, nascida de uma estratégia ampla de descoberta. Em seguida, o *OncotypeDX* reduz a pergunta para 21 genes medidos por RT-qPCR, mais compatível com tecido de rotina e com decisão adjuvante em doença RE-positiva. O PAM50 ocupa posição central porque não é apenas um escore de risco: ele conecta prognóstico aos subtipos intrínsecos. O *EndoPredict* e outras ferramentas posteriores acrescentam foco em recorrência tardia e integração com variáveis clínicas. À direita, o CorePAM é colocado como proposta de compressão do PAM50: não reinventa a biologia do câncer de mama, mas pergunta se parte menor do painel preserva o essencial. A leitura importante é que o campo caminhou de painéis grandes e exploratórios para ensaios mais focados, clinicamente interpretáveis e potencialmente mais acessíveis.

1.5 Assinaturas gênicas, plataformas e transportabilidade

Antes de discutir as coortes deste estudo, vale separar três conceitos que costumam ser misturados na prática clínica: **assinatura gênica**, **plataforma de leitura** e **algoritmo de interpretação**. Uma assinatura gênica é uma lista de genes escolhidos porque, em conjunto, carregam informação clínica. A plataforma é a tecnologia de laboratório usada para medir a

atividade desses genes. O algoritmo é a regra matemática que transforma as medidas em um resultado útil, como baixo ou alto risco, subtipo molecular, risco numérico de recorrência ou probabilidade de benefício com quimioterapia. Em termos simples: a assinatura diz “quais genes olhar”; a plataforma diz “como medir”; e o algoritmo diz “como interpretar”.

O funcionamento geral é parecido entre os testes. O patologista seleciona uma área representativa do tumor, o laboratório extrai RNA e mede a expressão de genes relacionados a programas biológicos centrais: sinalização hormonal, proliferação, via HER2, invasão, diferenciação luminal ou fenótipo basal. Depois, os valores medidos são normalizados para reduzir ruído técnico e combinados em um escore. O resultado não substitui estadiamento, grau, linfonodos, idade, comorbidades ou preferência da paciente; ele acrescenta uma camada biológica que não é visível ao microscópio nem pela imuno-histoquímica isolada. Essa é a razão pela qual assinaturas como *OncotypeDX*, *MammaPrint*, *PAM50/Prosigna*, *EndoPredict* e *Breast Cancer Index* entraram nas diretrizes: elas ajudam principalmente na zona cinzenta em que a indicação de quimioterapia adjuvante não é óbvia^{20,21,23,27}.

Um conceito central no PAM50 é o de **centroide**. Centroide é um perfil de referência, uma espécie de “média molecular” de um subtipo. Imagine que cada tumor tenha uma impressão digital composta pela expressão de dezenas de genes. O PAM50 compara a impressão digital da amostra com impressões digitais de referência dos subtipos *Luminal A*, *Luminal B*, *HER2-enriched* e *Basal-like*. O subtipo atribuído é aquele cujo centroide se parece mais com o tumor, geralmente medido por correlação. Portanto, o centroide não é um gene, nem um receptor, nem uma lâmina de imuno-histoquímica; é um padrão multigênico de referência^{13,16}.

Microarray foi a primeira grande tecnologia de expressão em larga escala usada na descoberta dessas assinaturas. Ele funciona como uma lâmina com milhares de sondas fixas: cada sonda é desenhada para capturar um RNA específico. Quanto maior a hibridização, maior o sinal. Foi a tecnologia dos estudos pioneiros de subtipos moleculares e do desenvolvimento inicial do *MammaPrint*^{8,9,26}. Sua vantagem histórica foi permitir medir milhares de genes simultaneamente. Suas limitações são importantes: a plataforma só mede aquilo para o qual há sonda, diferentes fabricantes usam sondas e escalas próprias, e o sinal pode saturar em genes muito expressos ou ficar impreciso em genes pouco expressos.

RT-qPCR é uma tecnologia mais direcionada. Em vez de medir o transcriptoma inteiro, ela mede um conjunto pequeno de genes pré-especificados com alta sensibilidade. É

familiar a muitos laboratórios, funciona bem em tecido fixado em formol e incluído em parafina (*FFPE*) e costuma ser adequada quando a pergunta já está definida. *OncotypeDX* e *EndoPredict* seguem essa lógica: não tentam descobrir novos genes no momento do teste; medem um painel fechado e aplicam uma fórmula previamente validada^{22,29}.

NanoString nCounter também é uma plataforma direcionada, mas usa sondas moleculares com códigos digitais para contar transcritos sem depender do mesmo tipo de amplificação da PCR. O *Prosigna* é a implementação clínica do PAM50 nessa plataforma. Sua força é combinar um painel fixo, compatível com *FFPE*, com fluxo laboratorial padronizado e validação analítica específica¹⁶. Na prática, isso é diferente de “rodar PAM50” em qualquer matriz pública: o teste comercial inclui controles, normalização e algoritmo definidos para aquela plataforma.

RNA-seq é a tecnologia mais ampla e contemporânea. Em vez de perguntar apenas se o RNA se ligou a uma sonda, ela sequencia milhões de fragmentos de RNA e conta quantos fragmentos correspondem a cada gene. Isso oferece grande faixa dinâmica, permite detectar transcritos não previstos por uma sonda antiga e se tornou a base de grandes coortes modernas, como SCAN-B e TCGA-BRCA^{32,33}. Seu custo vem caindo, mas a análise exige bioinformática rigorosa: filtragem, normalização, controle de qualidade, mapeamento de genes e decisão sobre como transformar contagens em uma escala comparável.

Quadro 1 — Principais assinaturas gênicas em câncer de mama

Assinatura	Plataforma ou lógica original	Função clínica principal
<i>OncotypeDX</i>	RT-qPCR; 21 genes em <i>FFPE</i>	Estimar risco de recorrência e benefício de quimioterapia em câncer de mama RE+/HER2-inicial.
<i>MammaPrint</i>	Microarray; 70 genes	Classificar risco genômico alto ou baixo, inclusive em decisões de quimioterapia quando risco clínico e genômico divergem.
PAM50/ <i>Prosigna</i>	PAM50 em NanoString nCounter; 50 genes	Atribuir subtipo intrínseco e estimar risco de recorrência por ROR, sobretudo em doença RE+/HER2- inicial.
<i>EndoPredict</i>	RT-qPCR; 12 genes + tamanho tumoral e linfonodos no EPclin	Estimar risco de recorrência distante, incluindo risco tardio, em doença RE+/HER2-.
Breast Cancer Index	Expressão de HOXB13/IL17BR + módulo de proliferação	Estimar risco tardio e apoiar decisão sobre extensão de hormonioterapia além de cinco anos.

Assinatura	Plataforma ou lógica original	Função clínica principal
CorePAM	Escore investigacional de 24 genes derivados do PAM50	Testar se parte menor do PAM50 preserva informação prognóstica e pode facilitar ensaios futuros mais acessíveis.

A transportabilidade entre plataformas é importante porque um escore clinicamente útil deve medir biologia tumoral, não apenas peculiaridades da máquina que gerou os dados. Se um modelo funciona apenas no equipamento em que foi criado, ele pode estar capturando ruído técnico, tipo de sonda, escala de leitura ou processamento computacional, e não risco real da paciente. Por isso, validar um escore em *RNA-seq*, *microarray Illumina* e *microarray Affymetrix* é um teste mais exigente do que validá-lo apenas em uma coorte parecida com a de treinamento. Neste trabalho, a normalização por *z-score*, descrita em detalhe no capítulo de Metodologia, foi escolhida justamente para converter leituras absolutas diferentes em unidades relativas comparáveis. A escolha tem limites e alternativas, como ComBat ou parâmetros congelados, mas a pergunta central permanece clínica: o sinal prognóstico sobrevive quando mudam laboratório, população, plataforma e escala de medição?^{34–36}.

1.6 Coortes públicas de câncer de mama

Os dados utilizados neste estudo são inteiramente públicos e desidentificados, obtidos de consórcios internacionais e projetos de iniciativa acadêmica que padronizaram e disponibilizaram matrizes de expressão e dados clínicos anotados. O estudo utiliza dez coortes independentes (cinco de sobrevida e cinco neoadjuvantes) organizadas em dois blocos analíticos:

Quando se cria uma assinatura gênica, a pergunta nunca é apenas se ela funciona no conjunto em que nasceu. Isso quase sempre é possível, porque o modelo aprendeu naquele ambiente. A pergunta real é mais exigente: **o sinal continua existindo quando mudam as pacientes, o laboratório, a plataforma, o tempo de seguimento e o desfecho?** Por isso, as coortes foram selecionadas como uma sequência de testes de estresse. Primeiro, uma coorte grande e moderna para derivar o escore; depois, coortes externas para saber se o escore viaja para outras tecnologias e populações; por fim, coortes neoadjuvantes para perguntar se a informação prognóstica também se relaciona com resposta à quimioterapia.

Tabela 1 — Mapa dos blocos de coortes e das perguntas metodológicas do estudo

Bloco analítico	Coortes incluídas	Pergunta que o bloco responde	Por que é necessário
Derivação	SCAN-B	Quais genes e pesos	É o ambiente de

Bloco analítico	Coortes incluídas	Pergunta que o bloco responde	Por que é necessário
		formam o menor escore que preserva desempenho prognóstico?	aprendizado. Deve ser grande, bem anotado e biologicamente rico.
Validação externa de sobrevida	TCGA-BRCA, METABRIC, GSE20685, GSE1456	O escore prediz desfecho fora da coorte em que foi criado?	Testa generalização clínica, populacional e tecnológica.
Validação <i>cross-plataforma</i>	RNA-seq, microarray Illumina, microarray Affymetrix	O escore mede biologia tumoral ou apenas particularidades da plataforma?	Um teste útil precisa sobreviver a diferenças de escala, sonda e processamento.
Análise secundária de pCR	GSE25066, GSE20194, GSE32646, I-SPY1, I-SPY2	Um escore derivado para prognóstico também captura quimiossensibilidade?	Tumores mais proliferativos podem ser mais agressivos e, ao mesmo tempo, mais sensíveis à quimioterapia.

No **bloco de sobrevida**, a coorte de treinamento é a SCAN-B (GSE96058, N = 3 069)^{37,38}, a maior coorte de *RNA-seq* de câncer de mama publicamente disponível, com coleta prospectiva e seguimento populacional. As quatro coortes de validação são: TCGA-BRCA (N = 1 072, *RNA-seq*)³⁹, METABRIC (N = 1 978, *microarray Illumina* HT-12)^{11,12}, GSE20685 (N = 327, *microarray Affymetrix* HG-U133 Plus 2.0; população asiática)⁴⁰ e GSE1456 (N = 159, *microarray Affymetrix* HGU133A/B; coorte de Estocolmo)⁴¹. Juntas, as coortes de validação somam 3 536 pacientes com tecnologias e contextos populacionais intencionalmente heterogêneos.

No **bloco de pCR**, quatro coortes primárias totalizam 697 pacientes (GSE25066⁴², GSE20194³⁴, GSE32646⁴³, I-SPY1⁴⁴) e uma coorte exploratória (I-SPY2, N = 986)⁴⁵ adiciona poder estatístico em uma população neoadjuvante contemporânea.

Tabela 2 — Papel de cada coorte na construção e validação do CorePAM

Coorte	Plataforma e população	Papel no estudo	O que ela testa
SCAN-B	RNA-seq Illumina; coorte sueca prospectiva	Derivação	Escolha dos 24 genes, pesos do escore e limiar de não inferioridade.
TCGA-BRCA	RNA-seq STAR/GDC; consórcio norte-americano	Validação externa	Se o escore se mantém em outro RNA-seq, com outro processamento e seguimento mais curto.
METABRIC	Microarray Illumina; seguimento longo	Validação externa	Se o escore preserva sinal em microarray Illumina e em sobrevida doença-

Coorte	Plataforma e população	Papel no estudo	O que ela testa
			específica.
GSE20685	Microarray Affymetrix; população asiática	Validação externa	Se o escore resiste a população diferente e plataforma Affymetrix.
GSE1456	Microarray Affymetrix antigo; menor N	Validação externa difícil	Se o escore ainda funciona com menor tamanho amostral e cobertura gênica parcial.
GSE25066	Microarray; tratamento neoadjuvante	pCR primária	Relação entre escore e resposta patológica completa à quimioterapia.
GSE20194	Microarray; MAQC-II	pCR primária	Reprodutibilidade em coorte usada para avaliação de modelos preditivos.
GSE32646	Microarray; neoadjuvância	pCR primária	Consistência do sinal de quimiossensibilidade em outra série independente.
I-SPY1	Microarray Agilent; ensaio neoadjuvante	pCR primária	Desempenho em contexto prospectivo de terapia neoadjuvante.
I-SPY2	RNA-seq; ensaio adaptativo contemporâneo	pCR exploratória	Se o achado se mantém em coorte moderna e maior, sem substituir as análises primárias.

1.7 A pergunta central

Apesar da validação clínica robusta do PAM50, sua implementação enfrenta barreiras práticas importantes: custo elevado (US\$ 3 000 a 4 000 por amostra em laboratórios de referência nos EUA; indisponível no Sistema Único de Saúde brasileiro), dependência exclusiva da plataforma *NanoString* para uso comercial e necessidade de medir simultaneamente 50 genes com controles padronizados e infraestrutura específica. Em países como o Brasil, essas barreiras se traduzem em iniquidade concreta de acesso: pacientes atendidas no sistema suplementar podem receber testes de assinatura gênica, mas pacientes do SUS raramente têm essa possibilidade, o que perpetua tanto o sub-tratamento (pacientes de fato de baixo risco expostas a quimioterapia desnecessária) quanto o sobre-tratamento (pacientes de alto risco não identificadas como tal)⁶.

Nesse contexto, é natural perguntar se todo o painel de 50 genes do PAM50 é realmente necessário para capturar a informação prognóstica relevante, ou se uma porção

menor — selecionada de forma sistemática e reprodutível — seria suficiente. A pergunta central deste trabalho pode ser enunciada de forma direta:

Quantos dos 50 genes do PAM50 são realmente necessários para manter seu desempenho prognóstico, dentro de uma margem de não inferioridade pré-especificada?

Este trabalho apresenta o CorePAM: uma redução de 50 para 24 genes derivada por regressão de *Cox elastic-net* na coorte SCAN-B, validada em quatro coortes externas *cross-plataforma*, e avaliada secundariamente como preditor de resposta patológica completa à quimioterapia neoadjuvante. A motivação não é propor mais uma assinatura comercial concorrente, mas demonstrar um princípio: que a informação prognóstica do PAM50 pode ser substancialmente comprimida, e que essa compressão — se metodologicamente rigorosa — abre portas para implementações analíticas mais baratas e mais acessíveis, especialmente em sistemas de saúde com restrições orçamentárias importantes.

2. OBJETIVOS

2.1 Objetivo geral

Derivar e validar externamente o CorePAM — escore prognóstico composto pelo menor subconjunto do painel PAM50 capaz de preservar o desempenho prognóstico do painel completo em câncer de mama.

2.2 Objetivos específicos

1. Derivar, na coorte SCAN-B, o menor subconjunto de genes do PAM50 que preserve o desempenho prognóstico do painel original, respeitando margem pré-especificada de não inferioridade.
2. Caracterizar biologicamente os genes selecionados, verificando sua coerência com os subtipos intrínsecos e com a biologia conhecida do câncer de mama.
3. Validar externamente o escore CorePAM em quatro coortes independentes (TCGA-BRCA, METABRIC, GSE20685 e GSE1456), abrangendo plataformas de *RNA-seq* e *microarray* e populações geograficamente distintas.
4. Avaliar o valor prognóstico incremental do CorePAM em relação a variáveis clínicas tradicionais e ao estadiamento anatômico.
5. Avaliar o CorePAM como preditor de resposta patológica completa (pCR) à quimioterapia neoadjuvante em coortes independentes.

3. MÉTODO

3.1 Delineamento do estudo e aspectos éticos

3.1.1 Tipo de estudo

Trata-se de estudo retrospectivo multicoorte de desenvolvimento e validação externa de modelo prognóstico, utilizando exclusivamente dados de expressão gênica e clínicos publicamente disponíveis e desidentificados. O estudo foi planejado e reportado de acordo com o guideline *TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis)*^{46,47}, que estrutura a descrição de modelos de predição em quatro blocos — desenvolvimento, validação interna, validação externa e modelo final — e recomenda reporte completo de itens relacionados a fontes de dados, variáveis, manejo de valores ausentes, estratégia analítica, desempenho e interpretação clínica.

O delineamento segue uma estrutura clássica de desenvolvimento e validação externa, com os três componentes analíticos mantidos estritamente separados para evitar vazamento de informação:

- **Desenvolvimento (derivação):** realizado inteiramente na coorte SCAN-B (GSE96058). Todo o processo de seleção de genes, ajuste de hiperparâmetros e estabelecimento de pesos ocorre dentro desta coorte, sem acesso aos dados de validação.
- **Validação externa:** realizada em quatro coortes completamente independentes — uma segunda coorte de *RNA-seq* (TCGA-BRCA) e três de *microarray* —, sem reajuste de qualquer parâmetro. Os pesos congelados do SCAN-B são aplicados diretamente aos *z-scores* calculados intra-coorte em cada conjunto de validação.
- **Análise secundária (pCR):** realizada em cinco coortes neoadjuvantes inteiramente distintas das coortes de sobrevida, tanto em população quanto em *endpoint* (resposta patológica completa a quimioterapia neoadjuvante), avaliando se a informação contida no escore transcende o *endpoint* de derivação.

A Figura 4 apresenta o desenho do estudo.

Desenho do estudo

Derivação em coorte populacional, validação externa e análise secundária de resposta à quimioterapia neoadjuvante

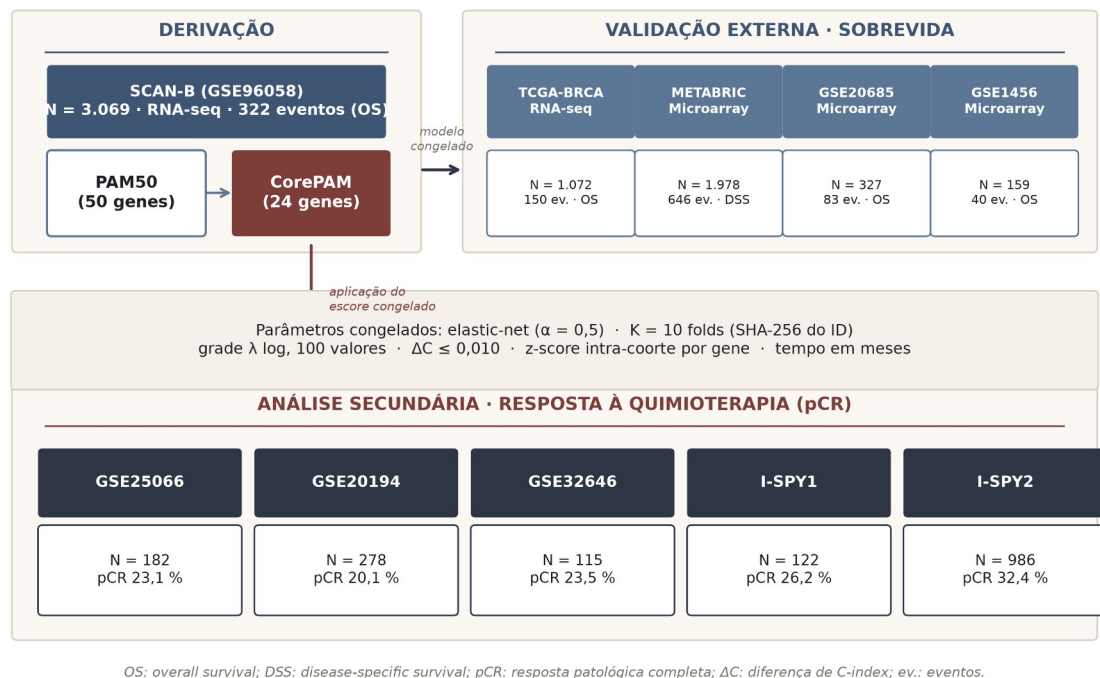


Figura 4 — Desenho do estudo CorePAM. Derivação na coorte SCAN-B (N = 3 069) por regressão de Cox elastic-net com validação cruzada determinística de 10 folds. Validação externa em quatro coortes independentes cross-plataforma (TCGA-BRCA, METABRIC, GSE20685, GSE1456; N total = 3 536). Análise secundária de pCR em cinco coortes neoadjuvantes (N total = 1 683). As setas indicam a direção do fluxo de dados, sem compartilhamento entre etapas.

Na Figura 4, a leitura começa pelo primeiro quadro: a coorte SCAN-B entra como coorte de derivação. Ela é grande, moderna e medida por *RNA-seq*, por isso funciona como um bom “hospital-escola” para o modelo: é ali que o CorePAM aprende quais genes carregar e que peso dar a cada um. O segundo bloco mostra a validação externa de sobrevida. Aqui a pergunta muda: os pesos aprendidos no SCAN-B são congelados e aplicados em TCGA-BRCA, METABRIC, GSE20685 e GSE1456. Essa etapa é análoga a testar uma conduta fora do centro onde ela foi criada. Se o escore só funcionasse no SCAN-B, seria apenas ajuste local; ao funcionar em coortes externas, ganha plausibilidade como sinal biológico. O terceiro bloco mostra as coortes de pCR. Elas não foram usadas para criar o escore, mas para fazer uma pergunta secundária: um escore prognóstico, ligado a agressividade e proliferação, também se associa à chance de resposta completa à quimioterapia? As setas da figura são importantes porque indicam direção única do fluxo: treino primeiro, validação depois, sem ida e volta.

Mapa prático das análises:

- **Selecionar genes:** regressão de Cox com penalização *elastic-net*, porque o desfecho principal é tempo até evento e os genes do PAM50 são correlacionados entre si.
- **Evitar otimismo:** validação cruzada *out-of-fold*, porque cada paciente precisa ser avaliada por um modelo que não usou seus próprios dados no treino.
- **Escolher o tamanho do escore:** margem de não inferioridade no *C-index*, porque a pergunta não é “qual modelo é numericamente máximo?”, mas “qual é o menor modelo que praticamente não perde desempenho?”.
- **Testar se generaliza:** validação externa em coortes de plataformas diferentes, porque um escore clinicamente útil precisa funcionar fora do ambiente em que foi criado.
- **Medir utilidade clínica:** modelos de Cox, curvas de Kaplan-Meier, *C-index*, calibração, curva de decisão e meta-análise, cada um respondendo a uma pergunta distinta: associação, visualização, discriminação, risco absoluto, benefício líquido e consistência entre coortes.

A Figura 5 resume visualmente esse mapa: o escore contínuo alimenta, em sequência, a análise primária (associação e discriminação), a visualização e o valor clínico, e a síntese entre coortes (meta-análise) — com uma camada transversal de análises de sensibilidade que testam a robustez de cada achado.

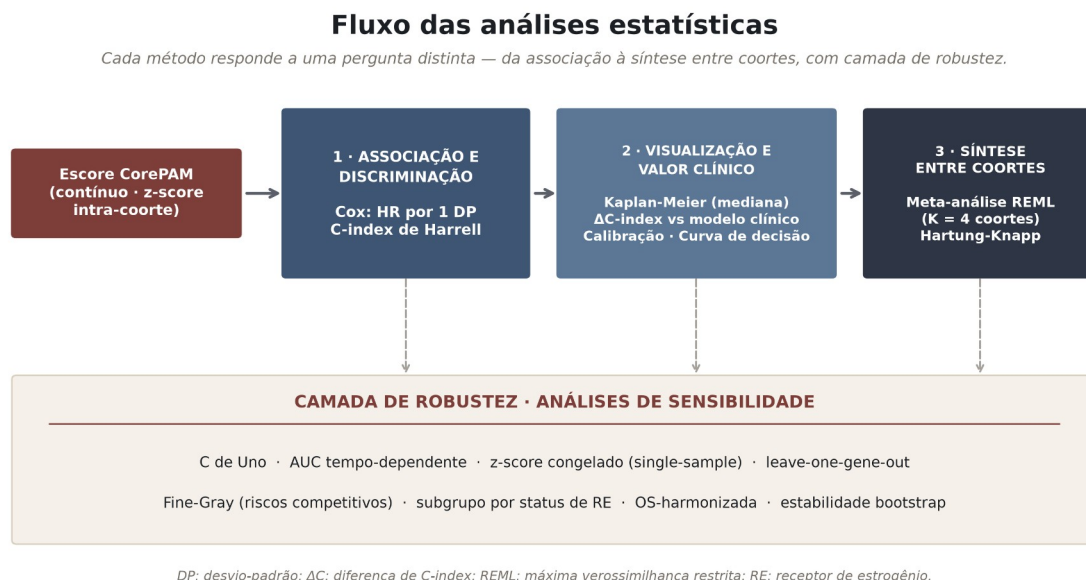


Figura 5 — Fluxo das análises estatísticas do estudo. Cada método responde a uma pergunta distinta: associação (Cox; *hazard ratio* por 1 desvio-padrão), discriminação (*C-index* de Harrell), visualização (Kaplan-Meier), valor incremental sobre o modelo clínico (ΔC -index), risco absoluto (calibração) e utilidade clínica (curva de decisão), culminando na síntese entre coortes por meta-análise. A camada inferior reúne as análises de sensibilidade que testam a robustez de cada achado: C de Uno, AUC tempo-dependente, *z-score* congelado, *leave-one-gene-out*, Fine-Gray (riscos competitivos) e subgrupo por status de RE.

O desenho também exigiu decisões para contornar dificuldades previsíveis em estudos multicoorte. A primeira dificuldade foi evitar vazamento de informação: se as coortes externas influenciassem a escolha dos genes, a validação deixaria de ser externa. Por isso, genes, pesos e parâmetros centrais foram definidos no SCAN-B e congelados antes da aplicação nas demais coortes. A segunda dificuldade foi a heterogeneidade técnica: *RNA-seq*,

Illumina microarray e *Affymetrix microarray* não medem expressão na mesma escala. Para reduzir esse problema sem misturar matrizes, cada coorte foi normalizada separadamente por *z-score*, permitindo comparar a posição relativa de cada tumor dentro da sua própria coorte.

Uma terceira dificuldade foi a cobertura incompleta de genes em algumas plataformas. TCGA-BRCA não continha um dos 24 genes do CorePAM, e GSE1456 continha 21 dos 24. Excluir essas coortes reduziria a validade externa justamente nos cenários mais difíceis; imputar genes ausentes criaria informação artificial. A solução foi pré-especificar um limiar mínimo de cobertura e reescalar o escore pelo somatório dos pesos dos genes efetivamente presentes. A quarta dificuldade foi a diferença de desfechos e tempos de seguimento. METABRIC usa sobrevida doença-específica, enquanto as demais coortes de sobrevida usam sobrevida global; TCGA-BRCA tem seguimento mais curto. Por isso, as análises mantiveram os desfechos identificados, usaram meta-análise de efeitos aleatórios e encurtaram o horizonte de calibração da TCGA-BRCA para 24 meses.

Por fim, havia dificuldade clínica e estatística no tamanho do escore. Escolher arbitrariamente “10”, “20” ou “30” genes facilitaria a narrativa, mas introduziria preferência do pesquisador. O estudo contornou isso definindo previamente a margem de não inferioridade e deixando o número de genes emergir dos dados. Essa estratégia se aproxima da prática clínica: não se escolhe um exame porque ele é bonito ou compacto, mas porque entrega informação suficiente com o menor custo e a menor fragilidade possível.

3.1.2 Aspectos éticos

Houve dispensa de submissão a novo Comitê de Ética em Pesquisa para esta tese porque o estudo utilizou exclusivamente dados secundários, públicos e desidentificados, oriundos de repositórios internacionais (GEO, TCGA/cBioPortal e European Genome-phenome Archive). Não houve recrutamento de participantes, contato com pacientes, intervenção clínica, coleta de material biológico, acesso a prontuários identificáveis ou tentativa de reidentificação. Na prática, a análise partiu de matrizes de expressão e tabelas clínicas já publicadas ou disponibilizadas por consórcios, sem nomes, documentos, datas de nascimento, endereços ou qualquer identificador direto. Esse enquadramento é compatível com a Resolução CNS 510/2016 (Brasil), que prevê que pesquisas com informações de domínio público ou bancos agregados sem possibilidade de identificação individual não sejam registradas nem avaliadas pelo sistema CEP/CONEP, e com a *Common Rule* dos Estados Unidos (45 CFR 46.104(d)(4)), que contempla isenção para pesquisa secundária com

informações publicamente disponíveis ou registradas de modo que a identidade dos participantes não possa ser prontamente determinada. Todas as coortes originais haviam sido aprovadas pelos respectivos comitês de ética das instituições geradoras, conforme documentado nas publicações primárias^{11,37,39–42,44,45}.

3.1.3 Princípios de validação externa

A validação externa seguiu quatro princípios estabelecidos na literatura de modelos prognósticos^{47–49}:

1. **Independência completa** — nenhum registro das coortes de validação foi acessado durante a fase de derivação. Os pesos do modelo foram congelados no SCAN-B e aplicados posteriormente.
2. **Diversidade intencional de plataformas** — a escolha deliberada de coortes em plataformas diferentes da de treinamento (*microarray Illumina*, *Affymetrix* HG-U133+, HGU133A/B, *RNA-seq* STAR/GDC) testa a transportabilidade cross-tecnológica, não apenas populacional.
3. **Não reajuste de parâmetros** — nenhum coeficiente do modelo foi re-estimado nas coortes externas. Qualquer validação externa que reajusta parâmetros é, na prática, uma nova derivação e deve ser reportada como tal.
4. **Pré-processamento independente por coorte** — cada coorte foi normalizada com seus próprios parâmetros de pré-processamento (*z-score* calculado com média e desvio-padrão da própria coorte), sem agrupamento inter-coorte ou correção de *batch* tipo ComBat^{35,36}. Esta escolha é discutida em detalhe adiante.

3.1.4 Disponibilidade de dados e código

Todos os dados utilizados neste estudo são de acesso público. O código-fonte completo, incluindo scripts de *download*, harmonização, derivação, *scoring*, análise de sobrevida, meta-análise, análises de sensibilidade e geração de figuras, encontra-se disponível no repositório público <https://github.com/RafaelBotan/CorePAM-Study>, sob licença permissiva, acompanhado do arquivo de configuração congelado (00_setup.R) que define todos os parâmetros analíticos pré-especificados. A lista completa dos scripts, numerados sequencialmente e organizados em blocos temáticos, encontra-se no APÊNDICE A. A reprodução integral dos resultados depende apenas do *download* dos dados públicos e da execução dos scripts na ordem numérica, não havendo qualquer componente proprietário ou dependência de dados sob embargo.

3.1.5 Uso de modelos de linguagem de grande escala

A redação e a formatação do presente manuscrito receberam assistência de um modelo de linguagem de grande escala (*large language model*, LLM) — Claude (Anthropic, modelo Opus) — empregado como ferramenta de apoio nas seguintes tarefas circunscritas: revisão ortográfica e de estilo em português acadêmico, verificação cruzada de referências bibliográficas e seus respectivos identificadores DOI, pareamento entre descrição textual dos

métodos e os scripts R correspondentes do *pipeline*, geração dos esquemas didáticos incluídos na Introdução e nos Métodos (subtipos moleculares, linha do tempo das assinaturas gênicas, fluxograma de coortes e esquema de derivação) e adequação da formatação do documento às normas da Associação Brasileira de Normas Técnicas (ABNT NBR 14724). O modelo não participou do desenho do estudo, da execução das análises estatísticas, da interpretação dos resultados nem da formulação das conclusões científicas, as quais são de responsabilidade exclusiva do autor. Todas as decisões metodológicas e os achados quantitativos foram verificados diretamente contra a saída dos scripts analíticos versionados no repositório público do estudo.

3.2 Coortes do estudo

O estudo integrou dez coortes públicas (cinco de sobrevida e cinco neoadjuvantes), divididas em dois blocos clínicos: coortes de sobrevida e coortes de quimioterapia neoadjuvante. As coortes não foram escolhidas por serem iguais; foram escolhidas justamente por serem diferentes. Essa heterogeneidade é parte do teste. Um escore que só funciona na coorte onde foi criado tem pouco valor clínico; um escore que preserva sinal em tecnologias, populações e tempos de seguimento distintos passa por um teste mais próximo da prática real.

A **SCAN-B** (GSE96058)^{37,38} foi usada como coorte de derivação. É uma coorte sueca, prospectiva, com tumores mamários primários operáveis medidos por sequenciamento de RNA (*RNA-seq*) *Illumina*. A partir dela foram escolhidos os genes e congelados os pesos do CorePAM. As quatro coortes externas de sobrevida testaram perguntas complementares. **TCGA-BRCA**³⁹ avalia se o escore se mantém em outra fonte de *RNA-seq*. **METABRIC**¹¹ testa uma plataforma *Illumina* de *microarray*, com seguimento muito mais longo e desfecho de sobrevida doença-específica. **GSE20685**⁴⁰ acrescenta população asiática e plataforma *Affymetrix*. **GSE1456**⁴¹, de Estocolmo, testa o cenário mais difícil: menor tamanho amostral, plataforma mais antiga e cobertura parcial dos genes do escore.

As coortes neoadjuvantes tiveram outra finalidade. Elas não foram usadas para treinar o escore de sobrevida; serviram para perguntar se o mesmo sinal biológico também antecipa resposta patológica completa à quimioterapia. **GSE25066**, **GSE20194**, **GSE32646** e **I-SPY1** formaram o conjunto primário de pCR^{34,42-44}. **I-SPY2** foi mantida como coorte exploratória contemporânea, maior, mas clinicamente mais heterogênea⁴⁵.

Mapa de leitura das coortes:

- **SCAN-B:** onde o CorePAM nasce. Serve para selecionar genes e definir pesos.
- **TCGA-BRCA:** pergunta se o escore viaja de um *RNA-seq* para outro *RNA-seq*.
- **METABRIC:** pergunta se o escore resiste a seguimento longo, desfecho doença-específico e plataforma *microarray Illumina*.
- **GSE20685:** pergunta se o escore se mantém em população asiática e *microarray Affymetrix*.
- **GSE1456:** pergunta se o escore ainda funciona quando alguns genes não estão disponíveis.
- **Coortes pCR:** perguntam se o sinal prognóstico também carrega informação sobre quimiossensibilidade.

A Tabela 3 e a Tabela 4 resumem os dados essenciais. Na Tabela 3, “N/eventos” mostra primeiro o número de pacientes analisadas e, depois da barra, o número de óbitos ou eventos do desfecho. Essa forma evita quebrar números na página e facilita a leitura clínica.

Tabela 3 — Características das coortes de sobrevida

Coorte	Papel	Tecnologia e desfecho	N/eventos	Seguimento
SCAN-B	Derivação	<i>RNA-seq Illumina</i> ; OS	3.069/322	54,9 meses
TCGA-BRCA	Validação	<i>RNA-seq</i> GDC; OS	1.072/150	32,0 meses
METABRIC	Validação	<i>Microarray Illumina</i> ; DSS	1.978/646	159,0 meses
GSE20685	Validação	<i>Microarray Affymetrix</i> ; OS	327/83	112,8 meses
GSE1456	Validação	<i>Microarray Affymetrix</i> ; OS	159/40	91,0 meses

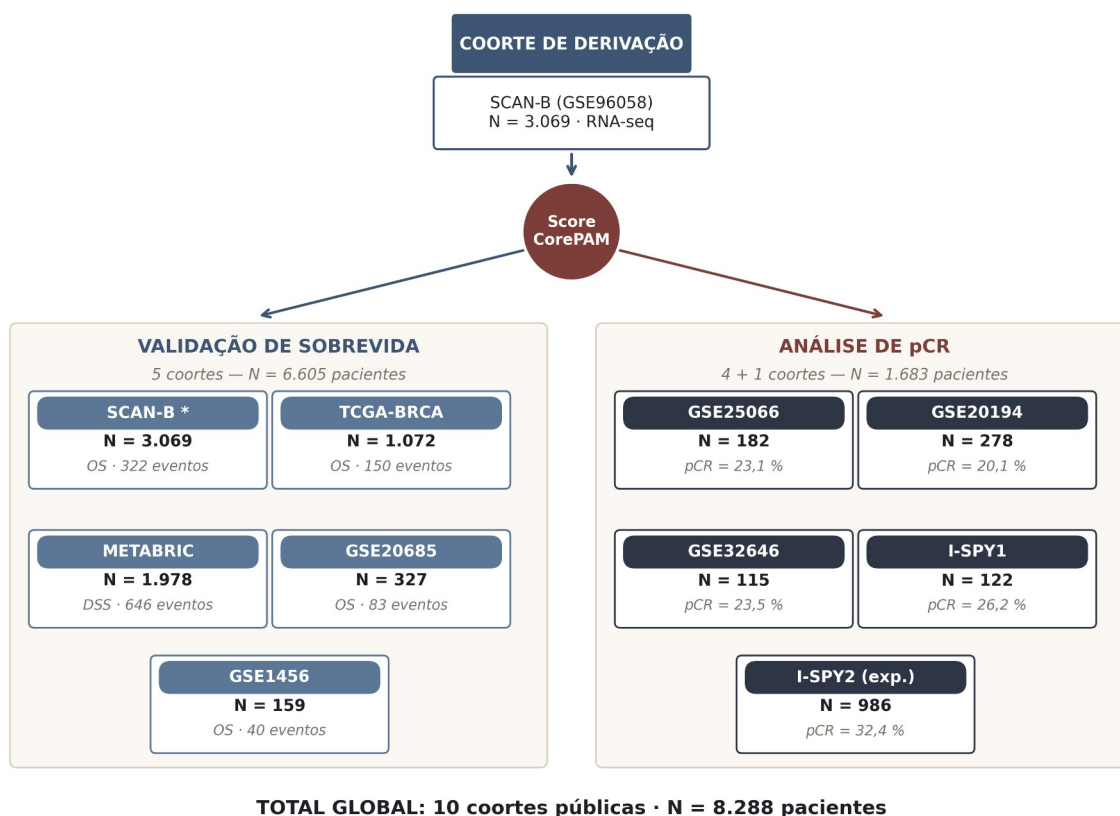
OS indica sobrevida global; DSS indica sobrevida doença-específica. A diferença é importante: OS inclui qualquer causa de morte, enquanto DSS tenta isolar mortes atribuídas ao câncer de mama. Por isso, METABRIC não é uma repetição das demais; ela responde a uma pergunta levemente diferente, porém clinicamente útil.

Tabela 4 — Características das coortes neoadjuvantes (pCR)

Coorte	Plataforma	N/pCR	Papel
GSE25066	HGU133A	182/42 (23,1%)	Primária
GSE20194	HGU133Plus2	278/56 (20,1%)	Primária
GSE32646	HGU133Plus2	115/27 (23,5%)	Primária
I-SPY1	Agilent 44K	122/32 (26,2%)	Primária
I-SPY2	RNA-seq	986/319 (32,4%)	Exploratória

Fluxograma das coortes do estudo

10 coortes públicas — 5 para validação prognóstica (sobrevida) e 4 + 1 para análise secundária (pCR)



* Coorte de treinamento (não usada como validação independente). OS: overall survival; DSS: disease-specific survival; pCR: resposta patológica completa.

Figura 6 — Fluxograma das coortes do estudo. Representação esquemática das dez coortes utilizadas, organizadas por bloco analítico (sobrevida versus pCR), papel (derivação versus validação), plataforma de medida (RNA-seq versus microarray) e tamanho amostral, evidenciando a heterogeneidade deliberada do desenho.

Na Figura 6, o leitor deve observar primeiro a separação entre os dois blocos clínicos. O bloco de sobrevivência responde à pergunta prognóstica: quem tem maior risco de evento ao longo do tempo? O bloco de pCR responde a uma pergunta preditiva: em quais pacientes o tumor tem maior chance de desaparecer por completo — a resposta patológica completa — com a quimioterapia neoadjuvante? Dentro do bloco de sobrevivência, SCAN-B aparece com

papel diferente das demais: é derivação, não validação. TCGA-BRCA testa outro *RNA-seq*; METABRIC testa microarray Illumina e seguimento longo; GSE20685 e GSE1456 testam plataformas Affymetrix e populações ou condições mais difíceis. No bloco de pCR, as coortes primárias avaliam consistência em estudos neoadjuvantes já conhecidos, enquanto I-SPY2 funciona como uma leitura contemporânea e exploratória. A figura é, portanto, o roteiro do estudo: cada caixa existe porque responde a uma pergunta diferente, não porque era apenas mais uma base disponível.

3.3 Congelamento analítico (analytical freeze)

O conceito de *congelamento analítico* (analytical freeze) é um pilar metodológico do presente estudo. Em modelos de predição, a flexibilidade analítica — a possibilidade de reportar diferentes resultados dependendo de escolhas tomadas após ver os dados — é uma fonte documentada de inflação de erro tipo I e de otimismo de desempenho. Para mitigar essa flexibilidade, todos os parâmetros analíticos relevantes foram definidos e documentados em arquivo de configuração versionado (script 00_setup.R do *pipeline*) antes de qualquer análise substantiva. Essa pré-especificação permite que o desempenho observado seja interpretado sem suspeita de seleção retrospectiva de análise favorável.

A Tabela 5 apresenta os parâmetros congelados.

Tabela 5 — Parâmetros analíticos congelados antes da análise

Parâmetro	Valor	Justificativa
Margem de não inferioridade (Delta- <i>C-index</i>)	0,010	< 1/2 EP do <i>C-index</i> OOF
Parâmetro de mistura <i>elastic-net</i> (alpha)	0,5	Balanço entre esparsidade (L1) e estabilidade de grupo (L2)
Folds de validação cruzada (K)	10	Padrão da literatura
Atribuição de <i>folds</i>	Determinística: SHA-256 do ID do paciente	Reprodutibilidade exata
Estratificação de <i>folds</i>	Por status de evento	Representatividade de eventos
Cobertura mínima de genes	80% dos 24 genes CorePAM	Equilíbrio inclusividade/integridade
Número de genes	Não pré-especificado	Resultado da análise, não parâmetro

3.3.1 Margem de não inferioridade (Delta-*C-index* = 0,010)

A escolha de 0,010 como margem de não inferioridade pré-especificada merece justificativa explícita. A ideia central do critério é aceitar um submodelo com menos genes se, e somente se, seu *C-index out-of-fold* (OOF) for no máximo 0,010 unidades inferior ao *C-index* OOF do modelo com todos os 50 genes candidatos. Essa margem foi definida como menor que metade do erro-padrão observado do *C-index* OOF no SCAN-B — uma regra prática que garante que a perda aceitável esteja dentro do ruído estatístico esperado, ao mesmo tempo em que não permite perdas clinicamente relevantes. Margens menores (e.g., 0,005) resultariam em um submodelo praticamente idêntico ao original em tamanho, anulando o ganho de parcimônia; margens maiores (e.g., 0,030) poderiam aceitar submodelos com perda mensurável de informação. A escolha 0,010 representa o compromisso explicitamente pré-especificado.

3.3.2 Atribuição determinística de *folds* (SHA-256)

A atribuição determinística de *folds* por hash SHA-256 do identificador do paciente elimina a arbitrariedade da escolha de semente aleatória, garantindo reprodutibilidade exata de primeiros princípios. O procedimento é o seguinte: para cada paciente, calcula-se o hash SHA-256 do seu identificador, interpretam-se os primeiros bytes como inteiro não-assinado, e atribui-se o *fold* pelo resto da divisão desse inteiro pelo número de *folds* (10). Esta atribuição tem três propriedades desejáveis:

1. **Reprodutibilidade universal:** qualquer pesquisador que baixe os mesmos dados e execute o mesmo código obterá exatamente os mesmos *folds*, sem dependência de semente aleatória ou implementação de RNG específica.
2. **Estabilidade a reordenamento:** a atribuição depende apenas do identificador, não da posição no arquivo. Reordenar linhas da matriz de expressão não altera *folds*.
3. **Independência entre ordem de execução e resultado:** ausência de state-dependent randomness significa que paralelismo e ordenamento não afetam o resultado.

Como o SHA-256 é criptograficamente uniforme, a distribuição dos *folds* é aproximadamente uniforme mesmo com *N* moderado (*N* = 3 069 no SCAN-B produz *folds* de tamanhos entre 295 e 320).

3.3.3 Estratificação por evento

Dentro do esquema determinístico, aplicou-se estratificação subsequente por status de evento (óbito/censura), para garantir representatividade de eventos em cada *fold* de validação. Em coortes com baixa taxa de eventos — como o SCAN-B (10,5%) —, a estratificação evita

que *folds* específicos recebam número desproporcional de eventos, o que degradaria a estimativa OOF do *C-index*.

3.4 Pré-processamento de expressão gênica

Todo o pré-processamento foi realizado independentemente por coorte, sem agrupamento inter-coorte ou correção de *batch* tipo ComBat³⁵ ou metodologias análogas. A decisão de não harmonizar matrizes entre coortes tem motivação técnica e pragmática: correção de *batch* requer acesso simultâneo a todas as matrizes de origem, o que é incompatível com o cenário clínico realístico (aplicar o modelo a uma nova amostra que chega isolada) e pode introduzir vazamento de informação quando covariáveis de interesse são correlacionadas com *batch*³⁶.

3.4.1 Coortes *RNA-seq* (SCAN-B, TCGA-BRCA)

Para coortes *RNA-seq*, o pré-processamento seguiu o *pipeline* padrão do Bioconductor: filtragem de genes de baixa expressão com `filterByExpr` do pacote `edgeR`⁵⁰; normalização por *trimmed mean of M-values* (TMM)⁵¹; transformação log-CPM (log das contagens por milhão) para estabilização de variância e aproximação a distribuição normal. Quando disponíveis, contagens brutas em formato ENSEMBL foram mapeadas a símbolos HUGO usando anotação mais recente do genoma humano (GRCh38).

3.4.2 Coortes *microarray* (METABRIC, GSE20685, GSE1456, GSE25066, GSE20194, GSE32646, I-SPY1)

Para coortes *microarray*, foram utilizadas as intensidades pré-processadas disponibilizadas pelos consórcios originais (METABRIC: escala log2 *Illumina* normalizada; GSE20685/GSE1456: RMA no *Affymetrix*; I-SPY1: loess *Agilent*). Não foi aplicada normalização adicional, de forma a preservar a análise como ela teria sido realizada em reproprocessamento independente. Quando múltiplas sondas mapeavam para o mesmo gene, foi selecionada a sonda com maior variância intra-coorte — escolha conservadora que maximiza o sinal biológico sob variação técnica controlada.

Tabela 6 — Cobertura de genes CorePAM por coorte de validação

Coorte	Plataforma	Genes	Cobertura	Ausentes
SCAN-B	RNA-seq Illumina	24/24	100%	Nenhum
TCGA-BRCA	RNA-seq STAR/GDC	23/24	95,8%	MIA

Coorte	Plataforma	Genes	Cobertura	Ausentes
METABRIC	Illumina HT-12	24/24	100%	Nenhum
GSE20685	Affymetrix HG-U133+	24/24	100%	Nenhum
GSE1456	Affymetrix HGU133A	21/24	87,5%	GPR160; CXXC5; KRT17

Esta tabela evita uma possível leitura equivocada. O CorePAM é, por definição, um escore de **24 genes**. Quando a Tabela 6 mostra 23/24 em TCGA-BRCA ou 21/24 em GSE1456, isso não significa que o modelo mudou de tamanho. Significa apenas que aquela plataforma pública não mediu todos os 24 genes com qualidade suficiente. Nesses casos, o escore foi calculado com os genes disponíveis e reescalado pelo somatório dos pesos presentes, regra definida antes da validação. Essa decisão preserva a coorte no estudo sem imputar artificialmente genes ausentes.

3.5 Derivação do CorePAM

Derivação do CorePAM — pipeline reprodutível

Do painel PAM50 (50 genes candidatos) ao CorePAM (24 genes selecionados)



OOF: out-of-fold (predição fora da dobra) · elastic-net: combinação de regularização L1 (lasso) e L2 (ridge) · ΔC: diferença de C-index.

Fonte: elaborada pelo autor.

Figura 7 — Esquema de derivação do CorePAM. Visão geral do procedimento: partindo dos 50 genes do PAM50 candidatos na coorte SCAN-B, a regressão de Cox *elastic-net* com validação cruzada *out-of-fold* e critério pré-especificado de não inferioridade ($\Delta C \leq 0,010$) seleciona o submodelo de 24 genes — o CorePAM —, subsequentemente avaliado em quatro coortes externas de validação.

Na Figura 7, o fluxo deve ser lido de cima para baixo. No topo entram os 50 genes do PAM50 candidatos no SCAN-B. Em seguida, a regressão de Cox *elastic-net* funciona como uma triagem disciplinada: genes redundantes ou instáveis tendem a perder peso, enquanto genes que contribuem para ordenar o risco permanecem. A validação cruzada *out-of-fold* é o mecanismo de proteção contra autoengano estatístico: cada paciente é avaliada por um modelo que não a viu no treino, como se fosse uma paciente nova no ambulatório. O critério de não inferioridade entra no final da seleção e muda a pergunta de “qual modelo tem o maior número possível?” para “qual é o menor modelo que praticamente não perde desempenho?”. O resultado do diagrama é o CorePAM com 24 genes. Só depois disso o escore vai para

validação externa, já congelado. Essa sequência é fundamental: se a validação pudesse influenciar a escolha dos genes, deixaria de ser validação.

3.5.1 Regressão de Cox *elastic-net*: justificativa e parâmetros

A regressão de Cox foi escolhida porque o desfecho principal é tempo até evento: não basta saber se a paciente morreu ou não morreu; importa também **quando** isso ocorreu e por quanto tempo as demais pacientes foram acompanhadas sem evento. A versão *elastic-net* acrescenta uma penalização aos coeficientes do modelo⁵²⁻⁵⁶. Em linguagem prática, essa penalização funciona como um freio: impede que genes com sinal fraco ou instável recebam peso exagerado apenas porque se ajustaram bem ao acaso da coorte de treinamento.

O *elastic-net* combina duas ideias complementares. A parte chamada *lasso* ajuda a zerar coeficientes pequenos e, portanto, seleciona genes. A parte chamada *ridge* estabiliza os pesos quando vários genes medem aspectos parecidos da biologia tumoral. Essa combinação é particularmente adequada ao PAM50, porque muitos genes pertencem aos mesmos programas biológicos, como proliferação ou diferenciação luminal. Um método que escolhesse apenas um gene representante de cada bloco poderia parecer mais simples, mas seria biologicamente frágil.

O parâmetro de mistura foi congelado em ***alpha* = 0,5**, isto é, metade *lasso* e metade *ridge*. A decisão foi tomada antes das análises principais porque equilibrava três necessidades: reduzir o número de genes, manter estabilidade entre genes correlacionados e evitar um modelo excessivamente adaptado ao SCAN-B. O parâmetro *lambda*, que controla a intensidade da penalização, foi avaliado em 100 valores. Para cada valor, o estudo registrou quantos genes permaneciam no modelo e qual era o desempenho fora da amostra de treino.

3.5.2 Validação cruzada *out-of-fold*

O desempenho durante a derivação foi medido por validação cruzada de 10 partes (*10-fold cross-validation*). A lógica é simples: divide-se a coorte SCAN-B em 10 grupos; em cada rodada, o modelo aprende em nove grupos e prediz o grupo que ficou de fora. Assim, cada paciente recebe uma predição feita por um modelo que não a utilizou no treinamento. Essa predição fora do treino é chamada *out-of-fold*.

Ao final, as predições dos 10 grupos foram reunidas e o *C-index* foi calculado uma única vez sobre todas as pacientes. Essa escolha evita uma média instável de resultados

pequenos por grupo e produz uma estimativa mais limpa do desempenho esperado em pacientes novas.

3.5.3 Critério de seleção: não inferioridade e fronteira de Pareto

Para cada intensidade de penalização, o estudo observou duas coisas: o número de genes ainda presentes no modelo e o *C-index out-of-fold*. A fronteira de Pareto organiza essa troca entre simplicidade e desempenho. Um ponto da fronteira é interessante quando não existe outro modelo ao mesmo tempo menor e melhor.

O CorePAM foi definido como o menor modelo cujo *C-index* estivesse no máximo 0,010 abaixo do melhor *C-index* observado. Essa é a margem de não inferioridade. A pergunta, portanto, não foi “qual ponto tem o maior número na terceira casa decimal?”, mas “qual é o menor conjunto de genes que preserva praticamente todo o desempenho?”. O resultado foi **24 genes**, com *C-index out-of-fold* = **0,670**. O máximo observado foi **0,679**, diferença de **0,009**, dentro da margem definida antes da análise.

3.5.4 Análises de estabilidade da seleção

Depois de congelar os 24 genes, foram feitas três verificações de robustez. O objetivo não era escolher genes novamente, mas entender se o conjunto final dependia de uma coincidência estatística.

1. **Bootstrap (B = 200):** a SCAN-B foi reamostrada 200 vezes, e o procedimento de seleção foi repetido. Genes que reaparecem com frequência alta são mais estáveis.
2. **Leave-one-gene-out:** cada gene do CorePAM foi removido uma vez, e o modelo foi recalculado com os 23 genes restantes. Se a retirada de um único gene derrubasse muito o desempenho, o escore dependeria demais dele.
3. **Direção dos pesos:** os sinais dos coeficientes foram comparados com a biologia esperada. Genes de proliferação deveriam, em geral, apontar para maior risco; genes de diferenciação luminal deveriam apontar para comportamento mais favorável.

3.6 Cálculo do escore

O cálculo do escore CorePAM procede em três etapas explícitas, cada uma com justificativa própria:

Etapas 1 — Z-score intra-coorte: para cada gene, a expressão de cada amostra é convertida em desvios em relação à média da coorte, dividida pelo desvio-padrão da própria coorte. A média e o desvio-padrão são calculados exclusivamente com amostras da mesma coorte. Esta padronização remove as diferenças de escala e de unidade entre plataformas, mantendo apenas a posição relativa da amostra em relação ao restante da coorte. A

consequência pragmática é importante: o escore é invariante a transformações lineares da expressão.

Etapa 2 — Escore bruto ponderado: soma-se, para cada amostra, os *z-scores* dos genes CorePAM disponíveis na coorte, cada um multiplicado pelo peso (coeficiente Cox) derivado no SCAN-B. O resultado é dividido pelo somatório de valores absolutos dos pesos dos genes presentes, garantindo que o escore esteja em escala comparável mesmo quando a coorte não apresenta cobertura integral dos 24 genes — caso do GSE1456 (21/24 genes, 87,5%).

Etapa 3 — Padronização final: o escore bruto ponderado é novamente normalizado para média zero e desvio-padrão 1 em cada coorte, facilitando a interpretação de *hazard ratio* por 1 desvio-padrão como aumento de uma unidade.

Direção: em todas as coortes, verificou-se empiricamente que escores mais altos correspondem a pior prognóstico (todos HRs univariados > 1), confirmando a coerência biológica do modelo.

3.6.1 Pontuação congelada (*single-sample scoring*)

Uma análise de sensibilidade relevante é a aplicação dos parâmetros de normalização do SCAN-B (média e desvio-padrão congelados) a cada amostra de validação isoladamente, em lugar de recalculados intra-coorte. Este cenário simula o uso clínico futuro — em que cada paciente chega isolada — e quantifica a perda de informação ao abrir mão da referência intra-coorte. A concordância entre os dois modos (intra-coorte vs congelado) foi avaliada por coeficiente de correlação de Pearson e por *C-index* na mesma coorte.

3.7 Análise de sobrevida

3.7.1 Modelos de Cox

Foram estimados modelos de Cox univariados (simples) e multivariados (ajustados). No modelo univariado, o único preditor foi o escore CorePAM. Nos modelos ajustados, o escore foi combinado com variáveis clínicas disponíveis, principalmente idade e status do receptor de estrogênio (RE). O resultado principal foi o *hazard ratio* por aumento de 1 desvio-padrão do escore. Essa escala foi escolhida porque permite comparar coortes medidas em plataformas diferentes: em vez de depender de uma unidade laboratorial específica, pergunta-se o que acontece quando a paciente está um desvio-padrão acima da média daquela coorte.

A suposição de riscos proporcionais foi avaliada pelo teste de Grambsch-Therneau^{57,58}. Quando havia sinal de violação, a análise era repetida de forma estratificada pela covariável problemática. Essa checagem é importante porque o modelo de Cox pressupõe que a razão de risco entre grupos seja aproximadamente estável ao longo do tempo.

3.7.2 Métricas de discriminação

O *C-index* de Harrell foi a métrica principal de discriminação⁵⁹. Ele pode ser lido como a probabilidade de o modelo ordenar corretamente duas pacientes comparáveis: a paciente que tem evento mais cedo deveria receber risco maior. Um *C-index* de 0,5 equivale a sorte; valores acima disso indicam ordenação prognóstica progressivamente melhor. Os intervalos de confiança foram estimados por *bootstrap* com 1.000 reamostras⁶⁰.

Como análise de sensibilidade, foi calculada a estatística C de Uno⁶¹, menos sensível a padrões de censura, e a AUC tempo-dependente em horizontes clinicamente relevantes pelo método de Blanche et al.⁶². Essas métricas foram usadas para confirmar que o achado não dependia de uma única forma de medir discriminação.

3.7.3 Curvas de Kaplan-Meier

As curvas de Kaplan-Meier foram usadas para visualização. O escore foi dividido pela mediana de cada coorte, criando grupos de maior e menor risco. Essa divisão facilita a leitura visual, mas não foi usada como análise principal, porque transformar um escore contínuo em dois grupos desperdiça informação. As estimativas principais permaneceram baseadas no escore contínuo.

Na METABRIC, o desfecho principal foi sobrevida doença-específica. Como mortes por outras causas competem com morte por câncer de mama, foi realizada análise de sensibilidade pelo modelo de Fine-Gray⁶³, que estimou *hazard ratio* de subdistribuição de 1,42 ($p < 0,001$) — praticamente idêntico ao *hazard ratio* do modelo de Cox principal (1,41) —, indicando que a competição por outras causas de morte não altera a conclusão prognóstica.

3.7.4 Valor incremental: Delta-*C-index*

O valor incremental pergunta se o CorePAM acrescenta algo ao que o médico já sabe por variáveis clínicas simples. Para isso, comparou-se o *C-index* do modelo clínico basal (CORE-A) com o *C-index* do mesmo modelo após acrescentar o CorePAM. A diferença entre

os dois é o *Delta-C-index*. Valores positivos indicam que o escore molecular melhora a ordenação prognóstica. O intervalo de confiança foi obtido por *bootstrap* com 1.000 reamostras.

Métricas de reclassificação como NRI e IDI não foram adotadas como medidas de valor incremental, por serem instáveis e frequentemente superinterpretadas em estudos prognósticos⁶⁴; o ganho incremental foi quantificado exclusivamente pelo *Delta-C-index*.

3.7.5 Calibração

Calibração é diferente de discriminação. A discriminação avalia se o modelo ordena corretamente as pacientes; a calibração avalia se a probabilidade prevista corresponde à frequência observada de eventos. Um modelo pode ordenar bem e, ainda assim, errar a probabilidade absoluta. Por isso foram construídas curvas de risco predito versus observado, seguindo a hierarquia proposta por Van Calster et al.^{64,65}.

O *slope* de calibração foi usado como resumo numérico. Valor próximo de 1 indica calibração ideal. *Slope* maior que 1 indica subconfiança: o modelo separa as pacientes corretamente, mas comprime demais as diferenças de risco. *Slope* menor que 1 indica o oposto, isto é, previsões exageradamente separadas.

3.7.6 Análise de curva de decisão (DCA)

A análise de curva de decisão foi incluída para aproximar a estatística da decisão clínica⁶⁶. Ela pergunta: se um médico adotar determinado limiar de risco para indicar tratamento, o modelo ajuda mais do que tratar todas as pacientes ou não tratar nenhuma? O resultado é o benefício líquido, que pondera acertos e falsos positivos de acordo com o limiar escolhido.

O limiar de decisão é a menor probabilidade de evento que justificaria uma ação clínica. Em sobrevida, por exemplo, um limiar de 10% em 5 anos significa: “eu aceitaria indicar uma intervenção adicional se o risco estimado de evento fosse pelo menos 10%”. Em pCR, um limiar de 25% significa: “eu consideraria a estratégia neoadjuvante, intensificação ou entrada em protocolo se a chance de resposta completa fosse pelo menos 25%”. A DCA transforma esse raciocínio em curva: para cada limiar, calcula quantos verdadeiros benefícios o modelo entrega, descontando o custo dos falsos positivos segundo a gravidade implícita naquele limiar.

Três referências ajudam a ler a figura. A estratégia **tratar nenhum** é a linha horizontal em zero: se nada é feito, não há falsos positivos, mas também não há ganho por identificar pacientes de maior risco. A estratégia **tratar todos** costuma começar alta quando o limiar é muito baixo, porque quase qualquer risco justificaria intervenção, mas cai e pode ficar negativa conforme o limiar aumenta, pois falsos positivos passam a pesar mais. A curva do modelo só tem utilidade clínica quando fica acima das duas referências. Se a curva do CorePAM estiver abaixo de “tratar todos” ou de “tratar nenhum”, o escore pode até ter associação estatística, mas não melhora a decisão naquele intervalo de limiares.

A análise foi feita principalmente no horizonte de 60 meses. Na TCGA-BRCA, o horizonte foi encurtado para 24 meses nas análises de calibração e sensibilidade porque o seguimento mediano é curto; usar 60 meses ali produziria estimativas instáveis e artificialmente dependentes de poucas pacientes ainda sob risco.

3.7.7 Modelos expandidos com estadiamento

Para testar se o CorePAM traz informação independente do estadiamento anatômico, foram ajustados modelos com variáveis como T, N, grau histológico, HER2 e tamanho tumoral, conforme disponibilidade em cada coorte. Essa etapa responde a uma pergunta clínica direta: o escore molecular ainda informa risco depois que se conhece o tamanho do tumor, o acometimento linfonodal e outros dados de rotina?

3.8 Meta-análise

A meta-análise foi usada para resumir os resultados das coortes sem fingir que elas são idênticas. Para sobrevida, foi adotado modelo de efeitos aleatórios com estimador REML⁶⁷, apropriado quando se espera alguma variação real entre estudos. Essa era a situação aqui: as coortes diferem em plataforma, país, tamanho, seguimento e desfecho.

A heterogeneidade foi quantificada pelo I-quadrado⁶⁸. Em termos simples, ele estima que proporção da variação entre resultados parece decorrer de diferenças reais entre coortes, e não apenas de erro amostral. Como havia poucas coortes, também foi aplicada a correção de Hartung-Knapp⁶⁹, que torna o intervalo de confiança mais conservador. Para pCR, foi usado DerSimonian-Laird⁷⁰, escolha tradicional em meta-análises de desfechos binários e mantida por comparabilidade com a literatura.

3.9 Comparações *head-to-head*

O desempenho do CorePAM (24 genes) foi comparado diretamente com um modelo PAM50-full ajustado no mesmo *pipeline*. O PAM50 canônico continua tendo **50 genes**. O número 49 aparece apenas porque, ao ajustar no SCAN-B um modelo Cox *elastic-net* usando os 50 genes PAM50 como candidatos, o maior modelo efetivamente obtido na trajetória penalizada manteve **49 coeficientes não-zero**; um gene recebeu coeficiente zero pela própria penalização. Portanto, “PAM50-full (49)” nesta tese não significa PAM50 incompleto nem altera a definição original do PAM50; significa o maior comparador penalizado estimável dentro do nosso procedimento. Esta comparação garante que ambos os modelos tenham sido derivados na mesma coorte e com o mesmo método, diferindo apenas no grau de parcimônia. Adicionalmente, foi realizada comparação exploratória contra implementações de pesquisa do *ROR-S* (50 genes) e do *OncotypeDX Recurrence Score* (21 genes), disponibilizadas pelo pacote *genefu*⁷¹. Esta comparação é limitada pelo fato de as implementações *genefu* não reproduzirem exatamente os ensaios comerciais (*Prosigna* e *OncotypeDX*), mas serve como referência aproximada para posicionamento do CorePAM em relação a ferramentas estabelecidas.

4. RESULTADOS

4.1 Derivação do CorePAM: os 24 genes

A regressão de Cox *elastic-net out-of-fold* no SCAN-B identificou 24 genes como o conjunto mínimo satisfazendo o critério de não inferioridade. $C\text{-index}$ OOF = **0,670**, $C\text{-index}$ OOF máximo = **0,679**, diferença = **0,009**. $C\text{-index}$ in-sample = **0,698**.

A Figura 8 demonstra retornos decrescentes a partir de aproximadamente 20 genes. No eixo horizontal está o número de genes retidos ao longo do caminho de regularização *elastic-net*, e no vertical o $C\text{-index out-of-fold}$ — estimativa honesta de desempenho, pois cada paciente é avaliado por um modelo treinado sem ele. A linha tracejada marca o limiar de não inferioridade ($C\text{-index}$ máximo menos a margem pré-especificada de 0,010): qualquer configuração acima dessa linha foi considerada não inferior pelo critério operacional pré-especificado, isto é, com $C\text{-index out-of-fold}$ no máximo 0,010 abaixo do valor máximo observado. O primeiro ponto em que o modelo mais parcimonioso atinge esse limiar é 24 genes, que portanto satisfaz simultaneamente desempenho e compacidade. O platô à direita indica que adicionar genes extras traz apenas ruído redundante — reforçando a hipótese biológica de que há um núcleo finito de programas transcricionais informativos para prognóstico em câncer de mama.

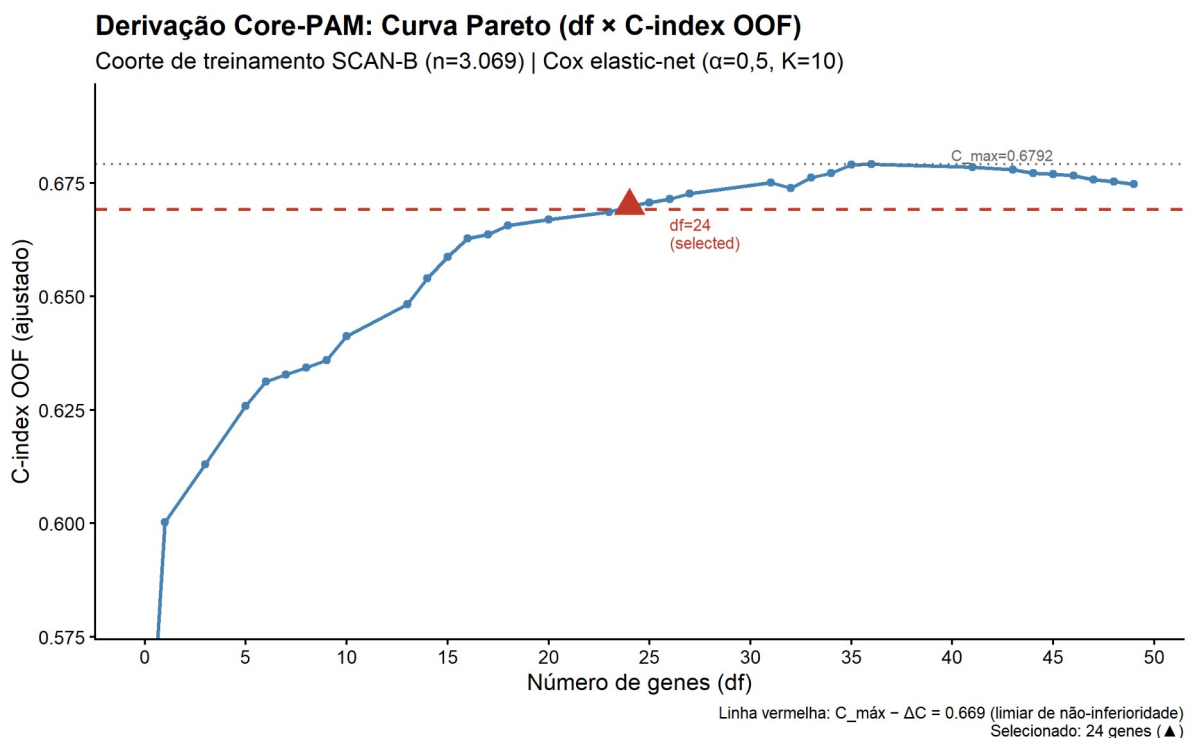


Figura 8 — Fronteira de Pareto: C-index out-of-fold em função do número de genes (graus de liberdade) ao longo do caminho de regularização elastic-net. A linha tracejada indica o critério de não inferioridade (C-index máximo - 0,010). O ponto de intersecção ocorre em $df = 24$. Observa-se um platô de desempenho a partir de aproximadamente 20 genes, indicando redundância funcional nos genes excedentes.

A leitura clínica da Figura 8 é análoga à escolha de um esquema terapêutico: adicionar mais componentes só compensa se houver ganho real. À direita do platô, os pontos têm desempenho numericamente parecido, mas carregam mais genes, custo e fragilidade técnica sem ganho proporcional — razão pela qual o ponto de 24 genes, o primeiro a cruzar o limiar de não inferioridade, é o escolhido.

Tabela 7 — Lista completa dos 24 genes CorePAM com pesos *elastic-net*

Rank	Gene	Peso (w_i)	Direção	Via biológica principal
1	EXO1	+0,2174	Risco (+)	Proliferação / reparo de DNA
2	NAT1	-0,2052	Protetor (-)	Diferenciação luminal
3	BLVRA	-0,1848	Protetor (-)	Metabolismo / estresse oxidativo
4	ACTR3B	-0,1405	Protetor (-)	Citoesqueleto / motilidade celular
5	MIA	-0,1193	Protetor (-)	Marcador <i>HER2-enriched</i>
6	MYBL2	-0,1183	Protetor (-)	Proliferação / ciclo celular
7	PTTG1	+0,1140	Risco (+)	Proliferação / instabilidade cromossômica
8	MDM2	-0,1074	Protetor (-)	Regulação de p53
9	SFRP1	-0,0907	Protetor (-)	Via Wnt / supressor tumoral
10	GPR160	-0,0841	Protetor (-)	Receptor acoplado a proteína G
11	FOXC1	+0,0717	Risco (+)	Fenótipo basal
12	PHGDH	+0,0620	Risco (+)	Metabolismo de serina
13	MYC	-0,0424	Protetor (-)	Proliferação / oncogene
14	ESR1	+0,0396	Risco (+)	Receptor de estrogênio
15	KRT5	-0,0277	Protetor (-)	Fenótipo basal
16	CXXC5	+0,0266	Risco (+)	Regulação epigenética
17	PGR	-0,0262	Protetor (-)	Receptor de progesterona
18	KRT17	-0,0163	Protetor (-)	Fenótipo basal
19	CENPF	+0,0150	Risco (+)	Proliferação / centrômero
20	FGFR4	+0,0110	Risco (+)	Receptor de FGF
21	BCL2	-0,0105	Protetor (-)	Anti-apoptose / diferenciação luminal
22	MLPH	-0,0099	Protetor (-)	Diferenciação luminal
23	ERBB2	-0,0082	Protetor (-)	Receptor HER2
24	GRB7	-0,0053	Protetor (-)	Transdução de sinal HER2

Painel Core-PAM — pesos dos genes

24 genes | não-inferioridade OOF ($\Delta C=0,010$) | treino SCAN-B | pesos do refit no conjunto completo

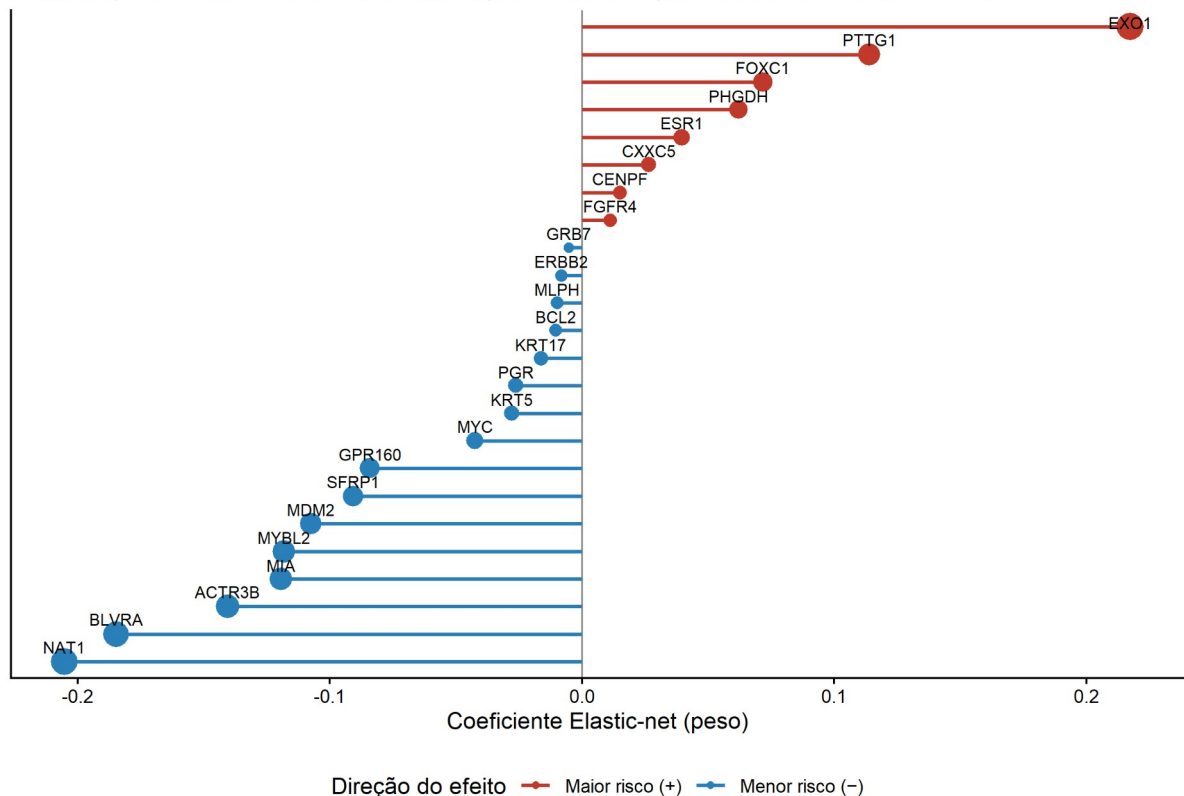


Figura 9 — Pesos dos 24 genes CorePAM (gráfico lollipop). Barras para a direita indicam associação com risco (peso positivo); barras para a esquerda indicam associação protetora (peso negativo). EXO1 e PTTG1 apresentam os maiores pesos positivos (proliferação), e NAT1 e BLVRA, os maiores negativos (diferenciação luminal).

Na Figura 9, a direção das barras é tão importante quanto o tamanho. Barras à direita indicam genes cujo aumento empurra o escore para maior risco; barras à esquerda indicam genes associados a comportamento mais favorável. EXO1 e PTTG1 aparecem como sinais de proliferação e instabilidade genômica, coerentes com maior risco. NAT1, BLVRA, BCL2, MLPH e PGR pertencem ao eixo de diferenciação luminal, que em geral acompanha tumores mais dependentes de hormônio e de melhor comportamento clínico. A figura também mostra que o CorePAM não é uma lista aleatória de marcadores: ele reúne eixos biológicos conhecidos do câncer de mama. Alguns sinais podem parecer contraintuitivos quando vistos gene a gene, como ESR1 com peso positivo ou ERBB2 com peso negativo. Isso ocorre porque o peso de cada gene é estimado dentro de um conjunto correlacionado; ele representa contribuição residual depois que os demais genes já estão no modelo, não uma leitura isolada como uma lâmina de imuno-histoquímica.

Dos 24 genes retidos, 17 mapeiam diretamente nos quatro eixos biológicos fundamentais do PAM50: **diferenciação luminal** (ESR1, PGR, BCL2, NAT1, MLPH, BLVRA), **proliferação** (MYBL2, PTTG1, EXO1, MYC, CENPF), **fenótipo basal** (KRT5, KRT17, FOXC1) e **marcadores HER2** (ERBB2, GRB7, MIA); os demais sete (ACTR3B, MDM2, SFRP1, GPR160, PHGDH, CXXC5, FGFR4) contribuem com funções correlatas — citoesqueleto, regulação de p53, via Wnt, metabolismo e sinalização — que refinam o sinal dentro desses eixos.

A estabilidade da seleção foi avaliada por *bootstrap* ($B = 200$): 15 dos 24 genes foram selecionados em $\geq 70\%$ das reamostras (Figura 10). A análise *leave-one-gene-out* confirmou resiliência estrutural (Figura 11): nenhum gene individual carrega o desempenho do modelo, com perda máxima de apenas 0,044 pontos de *C-index* ao remover o gene mais influente (EXO1). Esse padrão indica que a informação prognóstica está distribuída de forma relativamente homogênea entre os 24 genes, e não concentrada em poucos marcadores dominantes — propriedade desejável para robustez frente a falhas técnicas pontuais de medição.

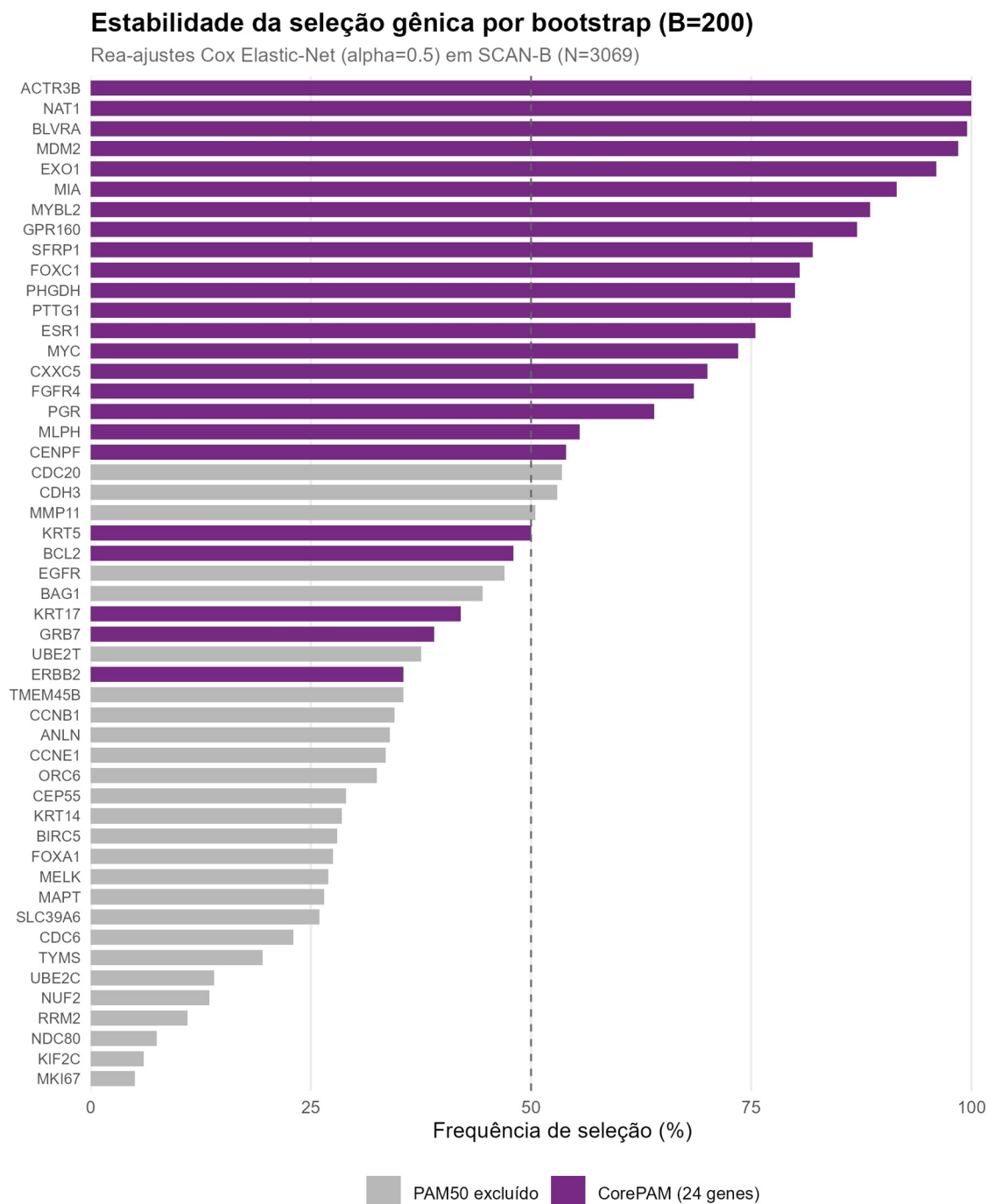


Figura 10 — Estabilidade bootstrap da seleção de genes (B = 200 reajustes no SCAN-B). Cada barra representa a frequência com que um gene foi selecionado nas reamostras. 15 dos 24 genes CorePAM foram selecionados em pelo menos 70% das iterações, com ACTR3B e NAT1 a 100% e frequência mínima de 35,5%.

Na Figura 10, a pergunta é estabilidade. Cada barra mostra quantas vezes o gene reaparece quando a coorte SCAN-B é reamostrada repetidamente. Uma analogia clínica útil é a reprodutibilidade de um achado em consultas diferentes: se um sinal aparece apenas uma vez, talvez seja acaso; se aparece repetidamente, merece confiança maior. ACTR3B e NAT1 surgem em 100% das reamostras, indicando seleção extremamente estável. Genes com

frequência intermediária não são necessariamente irrelevantes; eles podem representar programas biológicos parcialmente redundantes, em que genes vizinhos competem entre si. O ponto principal é que a maioria dos genes do CorePAM reaparece com frequência alta, sustentando que o painel de 24 genes não depende de uma partição fortuita da coorte.

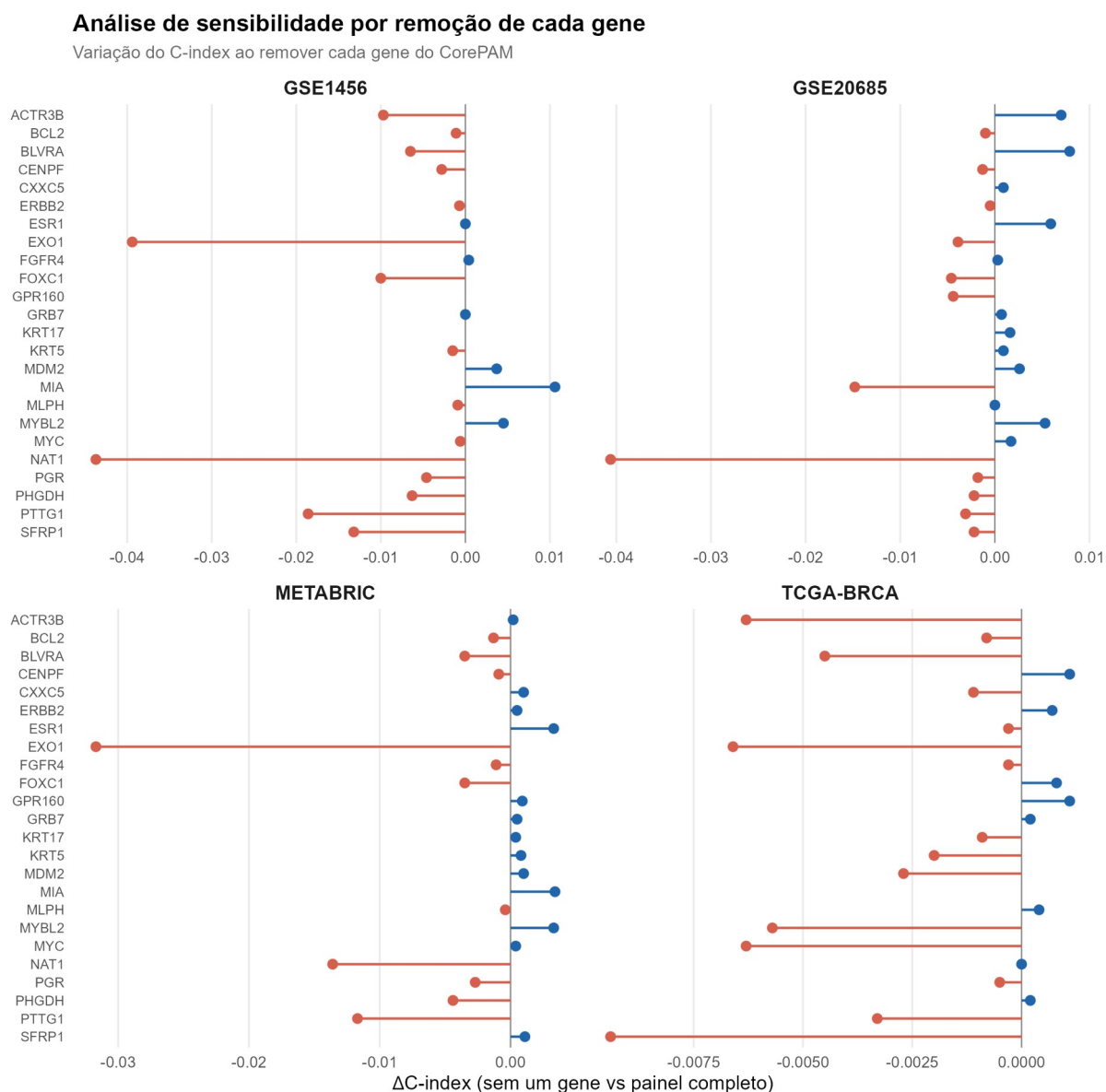


Figura 11 — Análise de sensibilidade leave-one-gene-out. Cada barra mostra a mudança no C-index OOF após remoção individual de cada gene CorePAM. Nenhuma remoção causa perda catastrófica (máximo $|\Delta\text{C}| = 0,044$), demonstrando distribuição equilibrada da informação prognóstica.

Na Figura 11, a pergunta é dependência excessiva. Cada barra mostra o que acontece com o *C-index* quando um gene é retirado do escore. Se uma única retirada derrubasse muito o desempenho, o CorePAM seria frágil, como uma conduta clínica que depende de um único exame sujeito a erro. O padrão observado é diferente: nenhuma exclusão causa colapso do desempenho. EXO1 é o gene cuja remoção mais impacta o modelo, o que combina com seu peso alto e sua ligação com proliferação, mas mesmo essa perda não destrói o escore. A interpretação é que o CorePAM distribui informação entre vários genes, propriedade desejável para um ensaio futuro, especialmente quando se pensa em tecido FFPE, variação pré-analítica e plataformas com cobertura incompleta.

4.2 Validação prognóstica externa

O escore CorePAM foi associado a pior sobrevida em todas as coortes externas de validação (Tabela 8). A leitura da tabela deve ser feita em três camadas. Primeiro, o *hazard ratio* (HR) mostra a força da associação: valores acima de 1 indicam que maior CorePAM corresponde a maior risco de evento. Segundo, o HR ajustado pergunta se essa associação permanece após incluir variáveis clínicas simples. Terceiro, o *C-index* mede a capacidade de ordenar pacientes por risco.

Tabela 8 — Desempenho prognóstico do CorePAM por coorte. SCAN-B é apresentada apenas como referência interna; seu *C-index* na tabela (0,698) é aparente (*in-sample*) e não deve ser comparado diretamente ao *C-index out-of-fold* da derivação (0,670) nem às coortes externas. Nas coortes externas, os pesos já estavam congelados. Em GSE1456, o modelo ajustado CORE-A não foi estimado por ausência de covariáveis clínicas suficientes.

Coorte (desfecho)	N/eventos	HR bruto	HR ajustado	<i>C-index</i>
SCAN-B (OS; treinamento)	3.069/322	1,92 (1,74-2,13); p<0,001	1,86 (1,64-2,12)	0,698 (0,668-0,729)
TCGA-BRCA (OS)	1.072/150	1,20 (1,04-1,40); p=0,016	1,27 (1,09-1,48)	0,624 (0,572-0,678)
METABRIC (DSS)	1.978/646	1,41 (1,31-1,52); p<0,001	1,36 (1,24-1,48)	0,638 (0,615-0,659)
GSE20685 (OS)	327/83	1,40 (1,12-1,75); p=0,003	1,40 (1,12-1,75)	0,626 (0,565-0,684)
GSE1456 (OS; sem ajuste clínico)	159/40	1,71 (1,23-2,38); p=0,001	-	0,661 (0,582-0,741)

Em termos clínicos, a Tabela 8 mostra que um aumento de 1 desvio-padrão no CorePAM multiplicou o risco por 1,20 em TCGA-BRCA, 1,41 em METABRIC, 1,40 em GSE20685 e 1,71 em GSE1456. O efeito menor em TCGA-BRCA deve ser interpretado com cautela: essa coorte tem seguimento mediano curto (32 meses) e menos eventos tardios, o que reduz o poder de capturar diferenças prognósticas em tumores RE-positivos. Já METABRIC tem seguimento muito longo e mais eventos, mas usa DSS, não OS; por isso, ela enxerga melhor mortes relacionadas ao câncer de mama, mas não é diretamente intercambiável com as coortes de OS. GSE1456, apesar de pequena e com apenas 21 dos 24 genes medidos, manteve o maior HR pontual, o que sugere que o núcleo do sinal é robusto mesmo com cobertura parcial.

As curvas de Kaplan-Meier (Figura 12) mostraram separação visual clara entre os grupos de alto e baixo risco em todas as cinco coortes. Cada painel representa uma coorte

independente. O eixo horizontal mostra o tempo em meses, e o eixo vertical mostra a fração de pacientes ainda sem o evento de interesse. A curva vermelha corresponde ao grupo de **alto risco** e, como esperado, fica abaixo da curva verde de **baixo risco** quando há pior sobrevida. A distância vertical entre as curvas representa a diferença absoluta de sobrevida naquele momento; o HR resume essa diferença como risco relativo ao longo do seguimento. A consistência da separação em coortes tão diferentes - Suécia, Estados Unidos/TCGA, METABRIC, Taiwan e Estocolmo - sustenta a transportabilidade do escore.

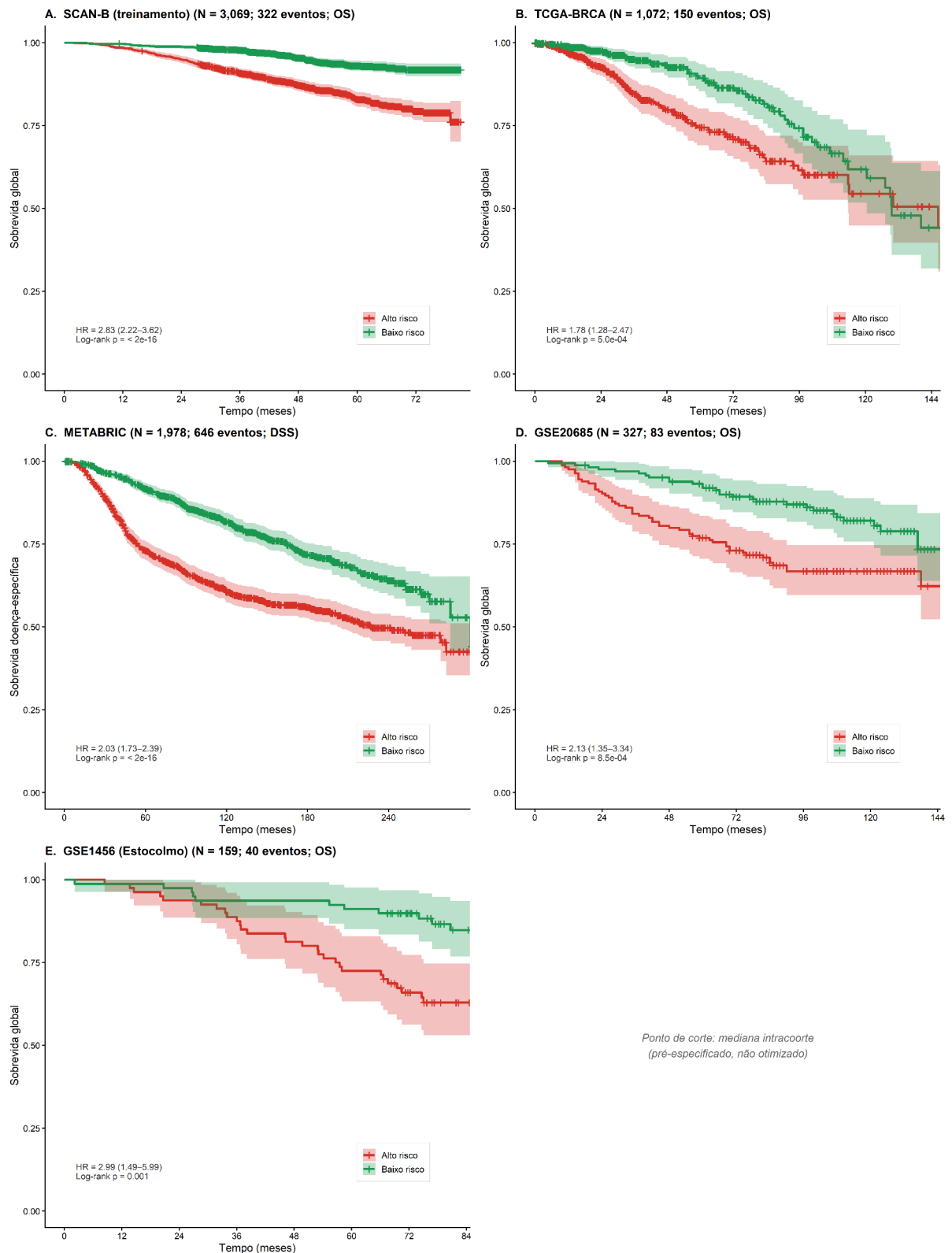


Figura 12 — Curvas de Kaplan-Meier nas cinco coortes (painel múltiplo). Pacientes dicotomizados na mediana intra-coorte do escore CorePAM: grupo alto risco (vermelho) vs baixo risco (verde). Sombreamento: IC 95%. Os HRs dicotomizados variaram de 1,78 (TCGA-BRCA) a 2,99 (GSE1456), com separação significativa em todas as coortes.

Na Figura 12, cada painel conta a mesma história em uma coorte diferente: o SCAN-B mostra a separação mais limpa, por ser a maior e a coorte de derivação; a TCGA-BRCA, mais

discreta, por ter seguimento mais curto e menos eventos; a METABRIC mantém a separação por muitos anos, adequada à doença de história natural longa; e GSE20685 e GSE1456, menores, têm intervalos mais largos, mas preservam a direção do efeito. A figura funciona como a “anamnese visual” do escore: antes de qualquer métrica, mostra se os grupos realmente se comportam de maneira diferente no tempo.

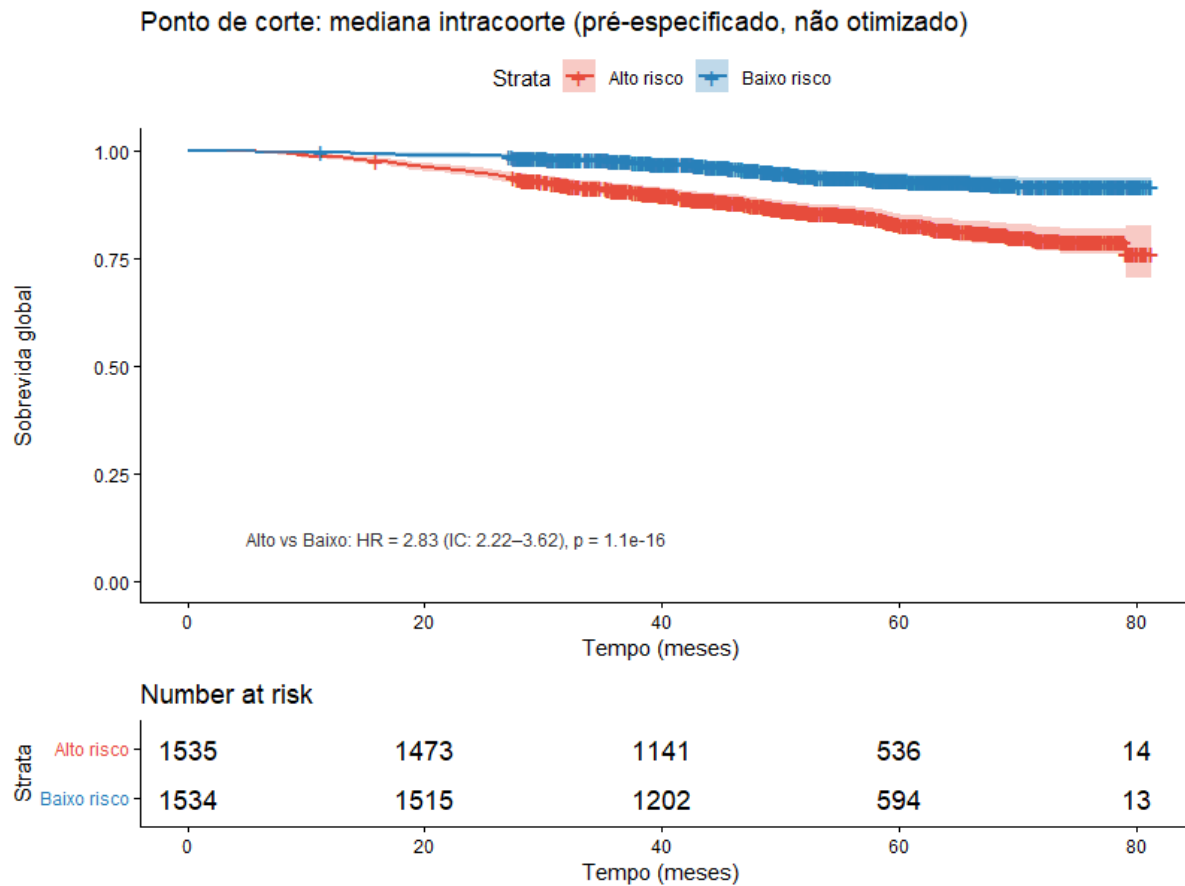


Figura 13 — Curva de Kaplan-Meier na coorte de treinamento SCAN-B (N = 3 069; sobrevida global). Dicotomização na mediana do escore CorePAM. HR = 2,83 (IC 95% 2,22-3,62; $p < 0,001$).

Na Figura 13, a separação é a mais ampla porque o SCAN-B é a coorte de derivação e a maior coorte individual do estudo: o HR dicotomizado de 2,83 indica risco de morte quase três vezes maior acima da mediana do CorePAM. As curvas se separam cedo e permanecem afastadas, sugerindo uma diferença biológica presente desde o início da história da doença. Como a mesma coorte foi usada para derivar o modelo, porém, este resultado deve ser lido como demonstração interna de coerência — não como validação externa nem prova de generalização.

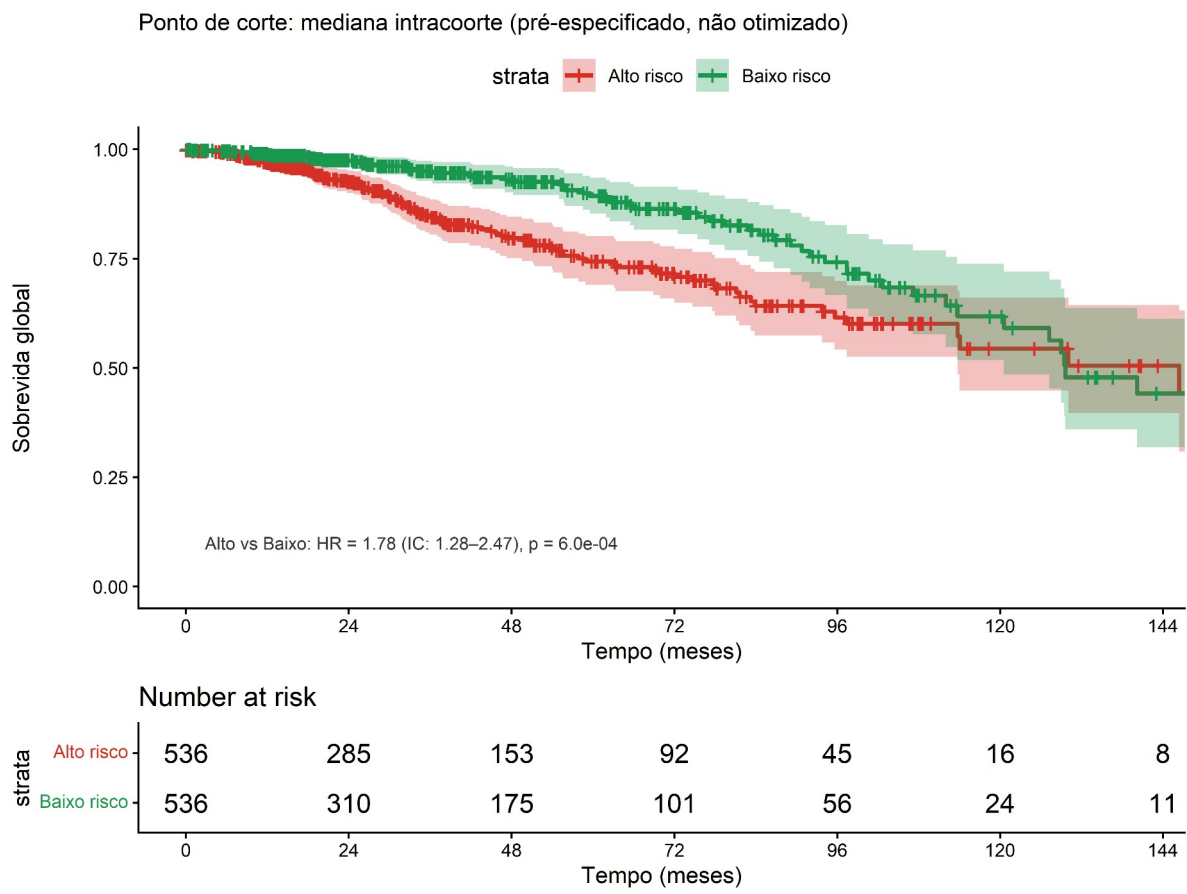


Figura 14 — Curva de Kaplan-Meier na coorte TCGA-BRCA (N = 1 072; RNA-seq; sobrevida global). Dicotomização na mediana do escore CorePAM.

Na Figura 14, a separação é mais discreta — esperado, não contraditório. A TCGA-BRCA tem seguimento mediano curto, menos eventos e composição clínica menos voltada ao seguimento longitudinal; é como uma consulta com pouco tempo de observação, em que a ausência de grande diferença visual reflete menor oportunidade de observar eventos, não ausência de biologia. Ainda assim, a direção do efeito permanece correta e significativa na análise contínua, mostrando que o escore calculado em outro *pipeline* de *RNA-seq* preserva o sinal, ainda que com menor amplitude.

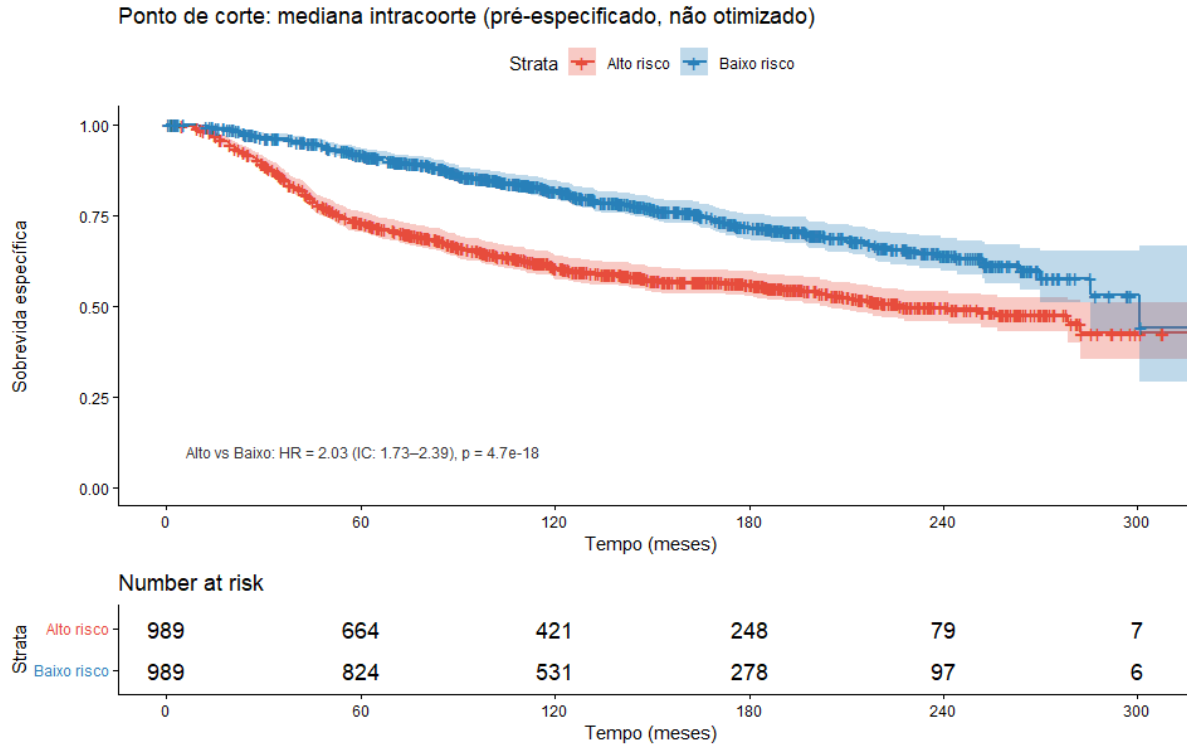


Figura 15 — Curva de Kaplan-Meier na coorte METABRIC (N = 1 978; microarray Illumina; sobrevida doença-específica). Separação ampla mantida ao longo de mais de 200 meses de seguimento.

Na Figura 15, a separação se mantém por mais de 200 meses — leitura particularmente relevante para a doença RE-positiva, em que eventos tardios são frequentes e decisões como a extensão da hormonioterapia dependem de risco prolongado. A METABRIC usa sobrevida doença-específica, o que reduz o ruído de mortes por outras causas (por isso a meta-análise incluiu análise de sensibilidade harmonizada). A persistência da separação por mais de uma década indica que o CorePAM não captura apenas agressividade inicial, mas uma dimensão sustentada da história natural da doença.

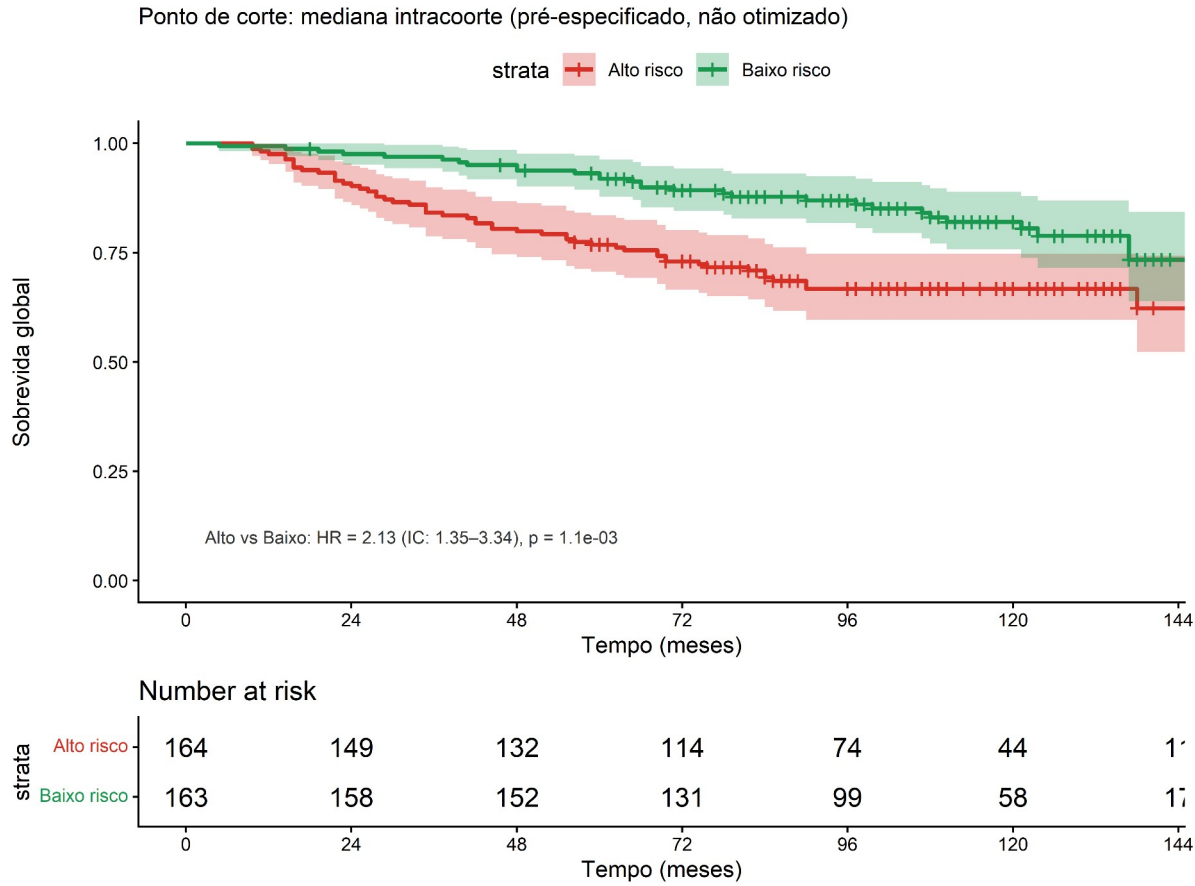


Figura 16 — Curva de Kaplan-Meier na coorte GSE20685 (N = 327; microarray Affymetrix; sobrevivência global). População asiática (Taiwan), testando generalizabilidade populacional e transportabilidade para plataforma Affymetrix.

Na Figura 16, o achado é relevante justamente por ser um teste de adversidade metodológica: coorte menor, asiática e medida por *Affymetrix*. O intervalo de confiança é naturalmente mais amplo, mas a separação das curvas (HR dicotomizado de 2,13 mostrado na figura) e o HR contínuo de 1,40 mostram que o sinal persiste fora das populações predominantemente europeias e da plataforma Illumina. Por isso, mesmo uma separação menos “bonita” tem valor: indica que a direção do risco não depende exclusivamente da coorte sueca.

4.3 Meta-análise de sobrevida

Para combinar formalmente a evidência das quatro coortes de validação em uma única estimativa agrupada, foi realizada meta-análise de efeitos aleatórios por REML (Tabela 9, Figura 17). Cada quadrado no *forest plot* representa a estimativa pontual de *hazard ratio* por 1 desvio-padrão do escore em uma coorte individual, com o tamanho do quadrado proporcional ao peso estatístico dessa coorte (determinado pela variância inversa e pelo tamanho amostral). Os bigodes horizontais são os intervalos de confiança 95%. O diamante no topo representa a estimativa agrupada. O valor I-quadrado quantifica heterogeneidade entre coortes: 38,2% indica heterogeneidade moderada, esperada dada a variação de plataforma, população e desfecho. A robustez da estimativa foi confirmada por dois métodos alternativos (Hartung–Knapp para IC ajustado ao tamanho pequeno de estudos e OS-harmonizada excluindo a METABRIC para sensibilidade ao *endpoint*), ambos produzindo conclusões qualitativamente idênticas.

Tabela 9 — Meta-análise de efeitos aleatórios — sobrevida (K = 4)

Análise	HR (IC 95%)	p	I-quadrado	tau-quadrado
Primária (REML)	1,37 (1,24–1,52)	$1,6 \times 10^{-9}$	38,2%	0,0067
Hartung–Knapp	1,37 (1,15–1,64)	0,011	38,2%	0,0067
OS-harmonizada	1,36 (1,13–1,64)	<0,001	50,5%	0,0151

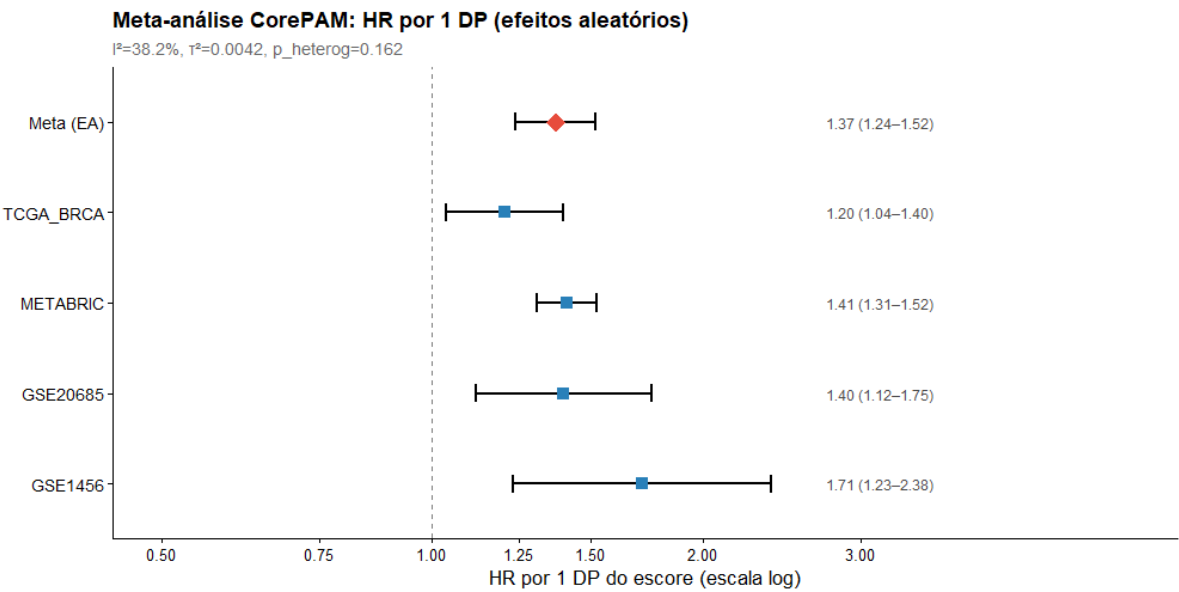


Figura 17 — Forest plot da meta-análise de sobrevida (HR por 1-DP, K = 4 coortes de validação). Modelo de efeitos aleatórios (REML). Diamante: estimativa agrupada (HR = 1,37; IC 95% 1,24-1,52; p = $1,6 \times 10^{-9}$). I-quadrado = 38,2%.

A linha vertical em $HR = 1$ marca a ausência de associação: na Figura 17, todos os quadrados — e o diamante agrupado no topo — ficam à direita dela, indicando maior risco com maior escore. A leitura é análoga à de subgrupos de um ensaio clínico: as coortes têm pesos diferentes, mas o que importa é a consistência de direção.

4.4 Valor incremental sobre preditores clínicos

A Tabela 10 e a Figura 18 respondem à pergunta clinicamente central: o CorePAM agrega informação prognóstica *além* do que já está disponível por variáveis clínicas rotineiras? O modelo de referência CORE-A combina idade ao diagnóstico e status de receptor de estrogênio — as duas variáveis clínico-patológicas mais informativas para prognóstico em câncer de mama. A coluna Delta-C quantifica o ganho de discriminação (*C-index*) ao adicionar o escore CorePAM a esse modelo. Intervalos de confiança 95% foram obtidos por *bootstrap* com 1 000 reamostras, garantindo honestidade estatística em coortes menores. Em todas as quatro coortes o ganho foi positivo, estatisticamente significativo (intervalos de confiança excluem zero) e clinicamente apreciável, com Delta-C variando de 0,030 (SCAN-B) a 0,094 (GSE20685). Notavelmente, o maior ganho ocorre na coorte onde o modelo clínico tem desempenho mais pobre — comportamento consistente com a ideia de que a expressão gênica é particularmente útil quando a informação clínica é limitada.

Tabela 10 — *C-index* incremental do CorePAM sobre o modelo CORE-A

Coorte	CORE-A	N	C sem/com CorePAM	Ganho Delta-C (IC 95%)
SCAN-B	idade + RE	3069	0,750 / 0,780	+0,030 (0,014-0,049)
TCGA-BRCA	idade	1072	0,638 / 0,676	+0,038 (0,011-0,086)
METABRIC	idade + RE	1978	0,603 / 0,646	+0,043 (0,025-0,066)
GSE20685	idade	327	0,536 / 0,630	+0,094 (0,021-0,175)

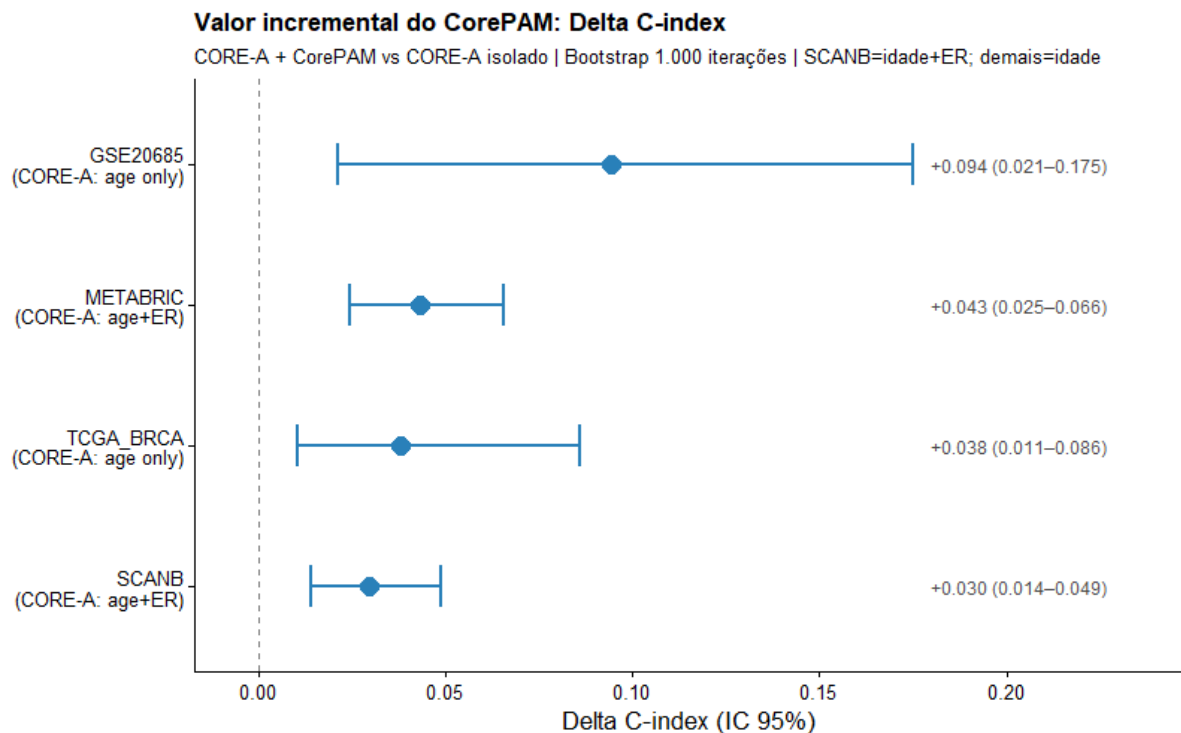


Figura 18 — Delta-C-index incremental do CorePAM sobre o modelo clínico CORE-A. Barras representam o incremento de C-index ao adicionar o escore CorePAM ao modelo clínico basal; bigodes indicam IC 95% bootstrap (B = 1 000). Todos os intervalos excluem zero, indicando melhoria estatisticamente significativa em todas as coortes.

Na Figura 18, cada barra responde à pergunta: quanto melhora a ordenação de risco quando o CorePAM é adicionado ao modelo clínico basal? O zero é a linha de nenhuma melhora. Todas as barras estão acima de zero e seus intervalos de confiança não cruzam zero, o que indica ganho estatisticamente consistente. A leitura clínica é direta: idade e RE ajudam, mas não esgotam a biologia tumoral. Em GSE20685, o ganho é maior porque o modelo clínico basal é mais fraco; é justamente nesse tipo de cenário que uma assinatura molecular pode ter maior utilidade incremental, como um exame adicional que esclarece uma decisão quando a avaliação inicial é insuficiente.

A Figura 19 complementa a análise de desempenho com a dimensão de calibração. Discriminação e calibração respondem a perguntas diferentes. A discriminação, capturada pelo *C-index*, pergunta se o modelo ordena corretamente as pacientes de menor para maior risco. A calibração pergunta se o risco absoluto previsto pelo modelo corresponde ao que de fato se observa.

Em cada painel da Figura 19, o eixo horizontal mostra a probabilidade prevista de estar sem evento no horizonte analisado, e o eixo vertical mostra a probabilidade observada. A linha diagonal representa calibração perfeita. Pontos próximos da diagonal significam que previsão e realidade estão próximas. Curvas acima ou abaixo da diagonal indicam que o modelo precisa de recalibração local antes de produzir probabilidades absolutas para decisão individual.

Os horizontes não são iguais em todos os painéis por um motivo metodológico, não por inconsistência. SCAN-B, METABRIC e GSE20685 foram avaliadas em 60 meses porque tinham seguimento suficiente para esse horizonte. TCGA-BRCA foi avaliada em 24 meses porque seu seguimento mediano é de apenas 32 meses; insistir em 60 meses nessa coorte deixaria a estimativa dependente de poucas pacientes ainda acompanhadas. Por isso, a curva da TCGA-BRCA não deve ser comparada visualmente como se representasse o mesmo tempo clínico das demais.

A TCGA-BRCA também parece diferente porque combina seguimento curto, menor número de eventos e processamento *RNA-seq* próprio do GDC. O painel concentra pontos em faixa estreita de risco e mostra *slope* maior (2,74), isto é, subconfiança: o modelo ainda ordena pacientes, mas comprime demais as diferenças absolutas de risco. O SCAN-B, por ser a coorte de desenvolvimento, apresenta *slope* próximo do ideal (0,96). METABRIC e GSE20685 ficam em posição intermediária, com *slopes* de 1,77 e 1,57. Clinicamente, a mensagem é: o CorePAM é útil para estratificar risco relativo entre pacientes, mas probabilidades absolutas em uma nova população exigem recalibração.

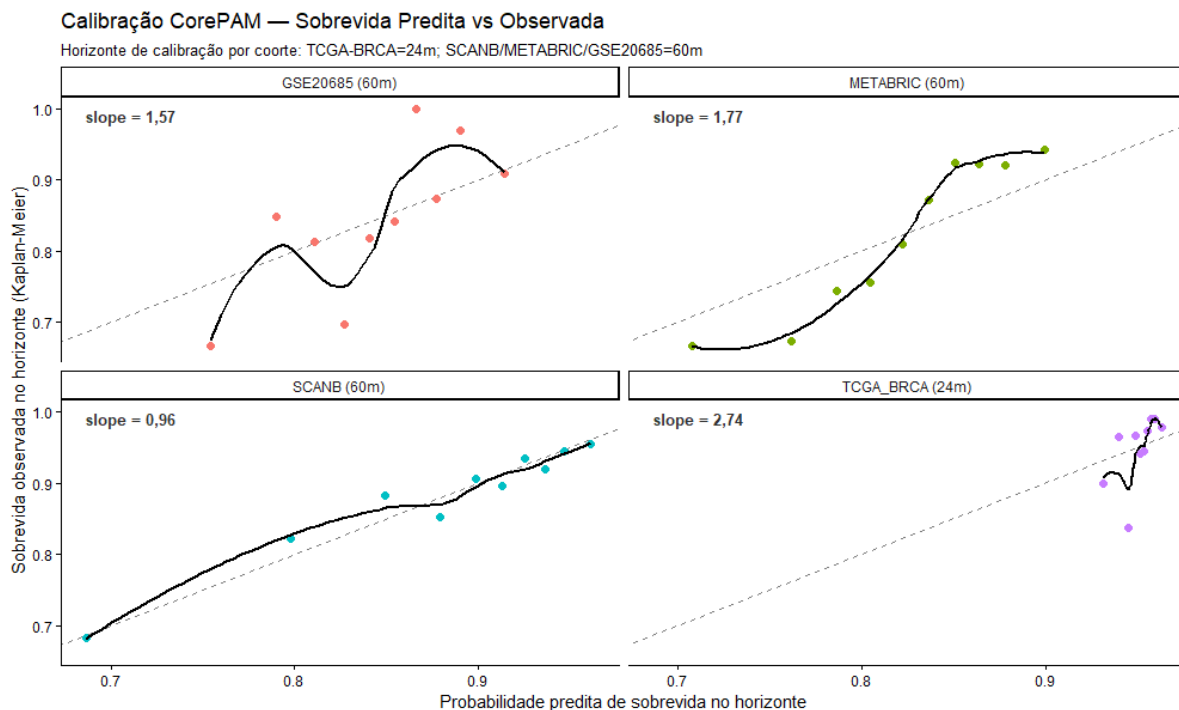


Figura 19 — Curvas de calibração (risco predito vs observado) em quatro coortes. SCAN-B, METABRIC e GSE20685 são avaliadas em 60 meses; TCGA-BRCA é avaliada em 24 meses por ter seguimento mais curto. Diagonal: calibração perfeita. O SCAN-B apresenta slope quase ideal (0,96); coortes externas mostram subconfiança progressiva (slopes 1,57-2,74), indicando que o modelo comprime a separação absoluta de risco.

Na Figura 19, o comportamento das coortes externas é o de um termômetro calibrado para outra escala: ainda distingue quente de frio (ordena as pacientes), mas não fornece a temperatura exata sem ajuste local. A conclusão prática é que o CorePAM, nesta forma, é mais maduro como ferramenta de estratificação de risco relativo do que como calculadora individual definitiva de risco absoluto.

4.5 Independência do estadiamento clínico-patológico

As Tabela 11 e Tabela 12 verificam se o valor prognóstico do CorePAM persiste após ajuste pelo estadiamento anatômico — tamanho tumoral (T) e acometimento linfonodal (N) —, além de idade e status de RE. Em todas as coortes, o escore permaneceu independentemente associado ao risco (HR ajustado entre 1,34 e 1,79; todos $p \leq 0,002$) e acrescentou discriminação ao modelo clínico completo (ganho de *C-index* de +0,022 a +0,027). Em outras palavras, a informação do CorePAM não é mero reflexo do estadiamento: agrega valor prognóstico independente das variáveis clínico-patológicas de rotina.

Tabela 11 — Modelo clínico expandido com estadiamento anatômico

Coorte	N	Ajuste clínico	HR CorePAM ajustado	C sem/com CorePAM	Delta-C
SCAN-B	2948	idade + RE + T + N	1,79 (1,57-2,05); $p < 0,001$	0,766 / 0,788	+0,022
TCGA-BRCA	1054	idade + T + N	1,35 (1,15-1,58); $p < 0,001$	0,719 / 0,746	+0,027
METABRIC	1883	idade + RE + T + N	1,34 (1,23-1,46); $p < 0,001$	0,663 / 0,688	+0,026
GSE20685	327	idade + T + N	1,45 (1,15-1,84); $p = 0,002$	0,697 / 0,724	+0,026

Tabela 12 — Modelo clínico completo (6 variáveis) com e sem CorePAM

Coorte/desfecho	Tipo	N	HR ajustado CorePAM	C sem/com CorePAM	Delta-C; LRT p
SCAN-B OS	Todos	2649	1,75 (1,51-2,03)	0,770 / 0,787	+0,017; $4,9 \times 10^{-13}$
SCAN-B OS	RE+	2450	1,68 (1,43-1,97)	0,770 / 0,784	+0,014; $1,1 \times 10^{-10}$
METABRIC DSS	Todos	1792	1,25 (1,14-1,38)	0,678 / 0,693	+0,015; $6,1 \times 10^{-6}$
METABRIC DSS	RE+	1377	1,47 (1,31-1,65)	0,674 / 0,707	+0,032; $1,0 \times 10^{-10}$

4.6 Análises de sensibilidade

Tabela 13 — Análise de sensibilidade do *z-score* congelado

Coorte	N	Plataforma	C congelado	C local	Delta-C	r
SCAN-B	3069	RNA-seq	0,698	0,698	0,0000	1,000
TCGA-BRCA	1072	RNA-seq	0,624	0,624	+0,0002	0,999

Coorte	N	Plataforma	C congelado	C local	Delta-C	r
METABRIC	1978	Illumina array	0,638	0,638	+0,0007	0,958
GSE20685	327	Affymetrix array	0,591	0,626	-0,036	0,923
GSE1456	159	Affymetrix array	0,622	0,661	-0,040	0,916

A Tabela 13 e a Figura 20 abordam uma questão prática decisiva para aplicação real do CorePAM: o modelo foi calibrado com parâmetros de normalização derivados da coorte sueca (SCAN-B, *RNA-seq*). Para uso em qualquer outra coorte, é preciso decidir se recalcular os parâmetros de padronização localmente (*z-score* intra-coorte) ou congelar os parâmetros SCAN-B. A coluna Delta-C mostra a perda de discriminação ao usar parâmetros congelados em vez dos intra-coorte, e a coluna *r* mostra a correlação entre os dois tipos de escore. Em plataformas de *RNA-seq* e *microarray Illumina*, a correlação é quase perfeita ($r \geq 0,958$) e a perda de *C-index* é praticamente nula, validando o congelamento como estratégia canônica. Nas plataformas *Affymetrix* a atenuação é mensurável ($r = 0,916$ – $0,923$; perda de *C-index* de até 0,04), mas o sinal prognóstico principal permanece preservado. Na prática, isso significa que um laboratório clínico pode aplicar o CorePAM a um paciente individual — sem precisar de uma coorte de referência para normalizar — usando os parâmetros SCAN-B publicados.

Concordância: escore intra-coorte vs z-score congelado (parâmetros SCAN-B)

Cada ponto = uma paciente (escores padronizados); linha = concordância perfeita ($y=x$)

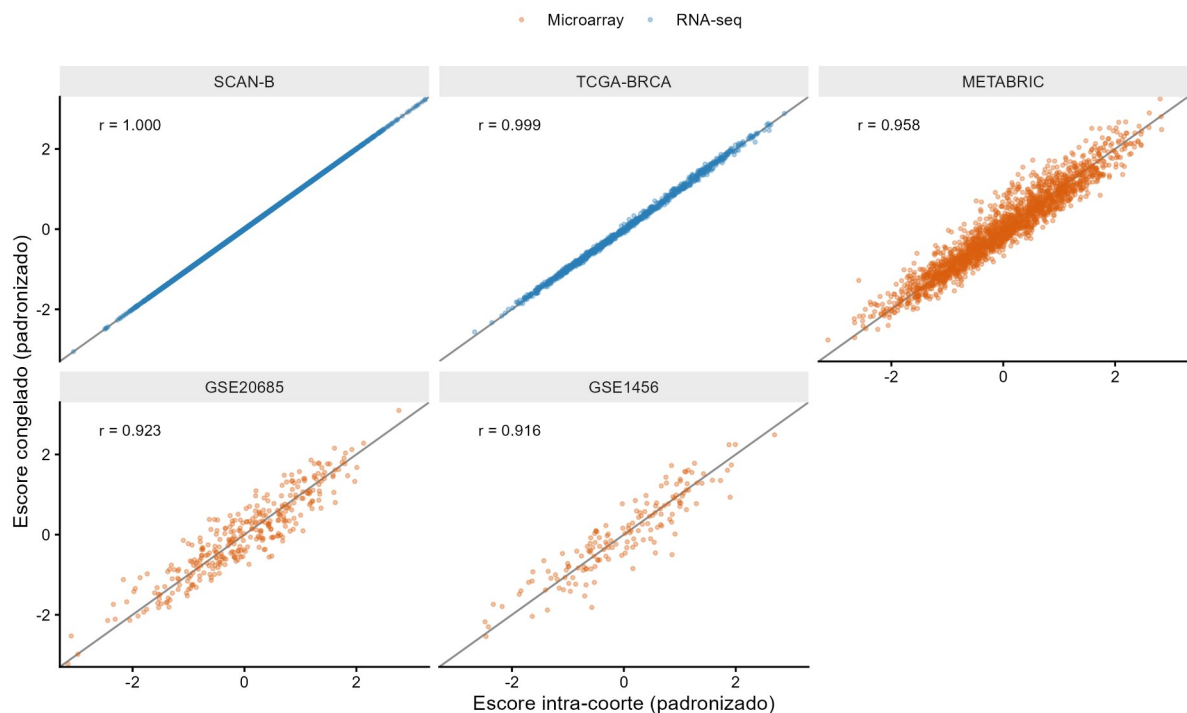


Figura 20 — Correlação entre escores intra-coorte e z-score congelado (parâmetros SCAN-B) por coorte. Diagonal: concordância perfeita. Plataformas RNA-seq e Illumina microarray mostram correlação quase perfeita ($r \geq 0,958$). Plataformas Affymetrix mostram atenuação mensurável ($r = 0,916-0,923$).

Na Figura 20, cada ponto é uma paciente pontuada de duas maneiras — com normalização local e com parâmetros congelados do SCAN-B — e a diagonal representa concordância perfeita. A ponte clínica é o ponto central: no laudo de uma paciente isolada não haverá uma coorte local para normalizar, e a figura mostra que a estratégia de parâmetros congelados é plausível, ao mesmo tempo em que aponta — nas plataformas *Affymetrix* — onde a validação analítica específica será mais necessária.

Duas análises de sensibilidade adicionais reforçam a discriminação e detalham a calibração, apresentadas a seguir (Tabela 14 e Tabela 15).

Tabela 14 — Estatística C de Uno e AUC tempo-dependente

Coorte	N	Horizonte	C de Uno	tdAUC (IC 95%)	C de Harrell
SCAN-B	3 069	60m	0,674	0,679 (0,643-0,715)	0,698
TCGA-BRCA	1 072	24m	0,669	0,673 (0,598-0,748)	0,624
METABRIC	1 978	60m	0,681	0,695 (0,665-0,725)	0,638
GSE20685	327	60m	0,668	0,675 (0,596-0,754)	0,626

Tabela 15 — Calibração formal por coorte

Coorte	N	Horizonte	alpha (EP)	beta (EP)	E/O	Brier
SCAN-B	3 069	60m	0,56 (0,18)	0,96 (0,09)	0,61	0,158
TCGA-BRCA	1 072	24m	5,40 (2,29)	2,74 (0,79)	0,74	0,061
METABRIC	1 978	60m	1,15 (0,25)	1,77 (0,17)	0,97	0,143
GSE20685	327	60m	0,93 (0,75)	1,57 (0,46)	0,98	0,130

A Tabela 14 complementa o *C-index* de Harrell com métricas menos sensíveis ao padrão de censura — o *C* de Uno e a AUC tempo-dependente —, confirmando discriminação consistente em horizontes clínicos (0,67–0,70). A Tabela 15 resume a calibração formal por coorte (intercepto, *slope*, razão esperado/observado e escore de Brier): o *slope* próximo de 1 no SCAN-B (0,96) indica calibração quase ideal na coorte de desenvolvimento, enquanto *slopes* acima de 1 nas coortes externas (1,57–2,74) sinalizam subconfiança — o modelo ordena bem as pacientes, mas comprime as diferenças absolutas de risco.

A Figura 21 apresenta a análise de curva de decisão (*decision curve analysis*) para desfechos de sobrevida. Essa figura deve ser lida como uma simulação de decisão clínica em vários limiares de risco. O eixo horizontal é o limiar: 5%, 10%, 20% ou 30% de risco previsto no horizonte analisado. O eixo vertical é o benefício líquido, isto é, o saldo entre identificar corretamente pacientes de maior risco e causar falsos alarmes. A linha “tratar nenhum” é o zero: nenhuma paciente é selecionada, logo não há benefício nem dano por seleção. A linha “tratar todos” representa a conduta de oferecer a intervenção a todas as pacientes, estratégia que só faz sentido quando o limiar aceito é muito baixo. As curvas do CorePAM, do modelo

por idade e do modelo combinado mostram se uma decisão guiada por modelo melhora esse balanço.

Análise de Curva de Decisão — OS/DSS

Benefício clínico líquido do CorePAM vs tratar-todos e não-tratar

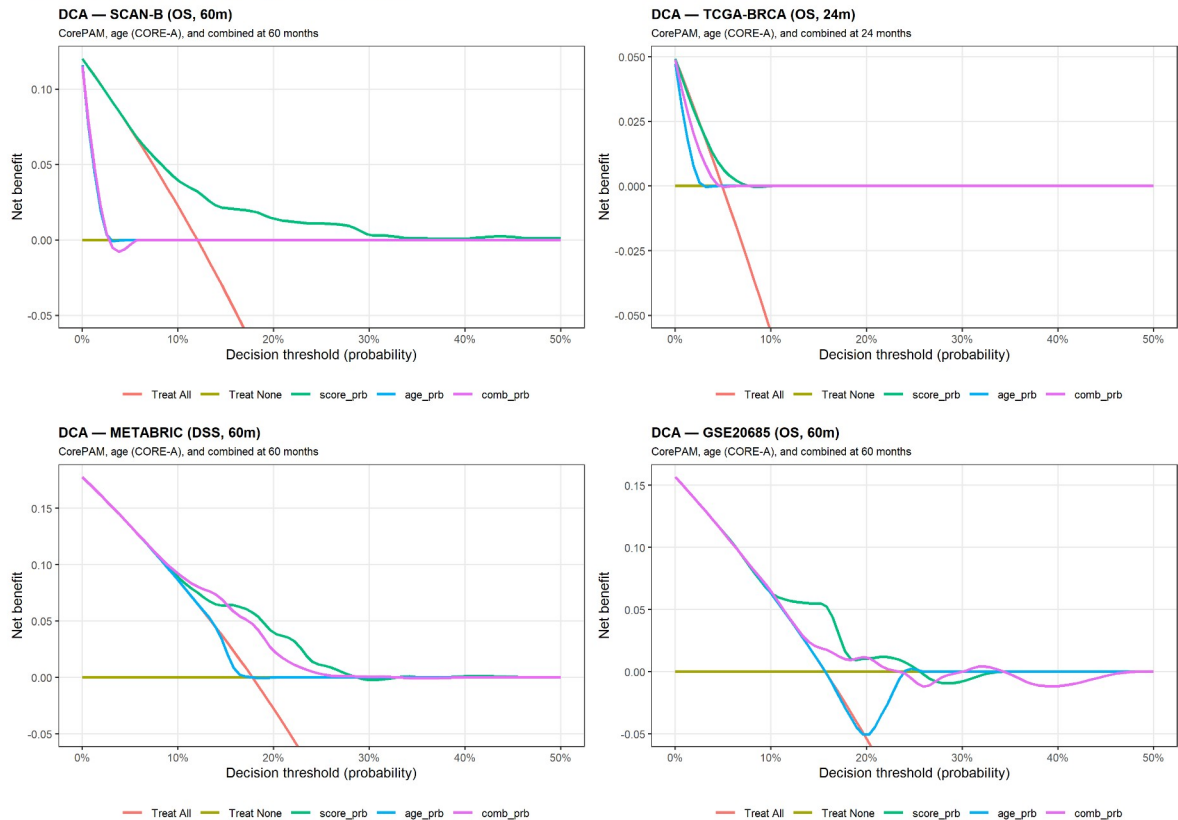


Figura 21 — Análise de curva de decisão (DCA) — quatro coortes de sobrevida. Benefício líquido do CorePAM, do modelo clínico por idade e do modelo combinado comparado com estratégias de tratar todos e tratar nenhum. SCAN-B, METABRIC e GSE20685 foram avaliadas em 60 meses; TCGA-BRCA foi avaliada em 24 meses por seguimento curto.

No painel SCAN-B da Figura 21, o CorePAM isolado mantém benefício líquido acima das estratégias de referência em uma faixa ampla de limiares, sobretudo entre aproximadamente 5% e 30%. Isso significa que, no ambiente de treinamento, o escore consegue separar risco de modo clinicamente útil: melhor do que tratar todas as pacientes e melhor do que não selecionar ninguém. A curva “tratar todos” cai rapidamente e cruza a linha de zero à medida que o limiar aumenta, ilustrando o custo de oferecer intervenção a muitas pacientes que não teriam evento. A idade isolada quase não acrescenta benefício depois dos limiares baixos; essa leitura conversa com a prática clínica, na qual idade ajuda, mas não substitui a biologia tumoral.

No painel TCGA-BRCA, a janela útil é estreita. As curvas têm benefício apenas em limiares muito baixos e se aproximam de zero rapidamente. Esse comportamento não deve ser lido como contradição do efeito prognóstico, mas como consequência do seguimento curto e

da baixa incidência de eventos em 24 meses. Quando poucos eventos ocorreram, há pouco espaço para uma curva de decisão demonstrar ganho líquido sustentado. Em linguagem clínica, é como tentar julgar recorrência tardia depois de uma consulta de seguimento ainda curta: a direção pode estar correta, mas o tempo observado é insuficiente para decisões de longo prazo.

No painel METABRIC, a figura é particularmente informativa. O modelo combinado e o CorePAM mantêm benefício líquido positivo por uma faixa de limiares mais ampla, compatível com o seguimento longo e com o desfecho de sobrevida doença-específica. Esse é o cenário em que a DCA mais se aproxima do problema clínico de tumores luminais: decidir se uma diferença biológica deve modificar intensidade de tratamento ou seguimento quando a história natural se estende por muitos anos. No painel GSE20685, o CorePAM e o modelo combinado também superam as referências em limiares baixos a intermediários, mas as curvas oscilam mais e cruzam zero em limiares altos. Isso é esperado em uma coorte menor, com menos eventos e plataforma *Affymetrix*: a mensagem confiável está na direção e na faixa inicial de benefício, não em pequenas ondulações da curva.

Assim, a Figura 21 não diz que o CorePAM deve ser usado isoladamente para indicar tratamento hoje. Ela mostra algo mais específico: nas faixas de risco em que um médico realmente pondera tratar ou não tratar, o escore frequentemente acrescenta benefício líquido sobre estratégias ingênuas. Esse é o mesmo tipo de raciocínio usado para qualquer biomarcador prognóstico em oncologia: associação estatística é necessária, mas utilidade clínica exige que o marcador melhore uma decisão concreta.

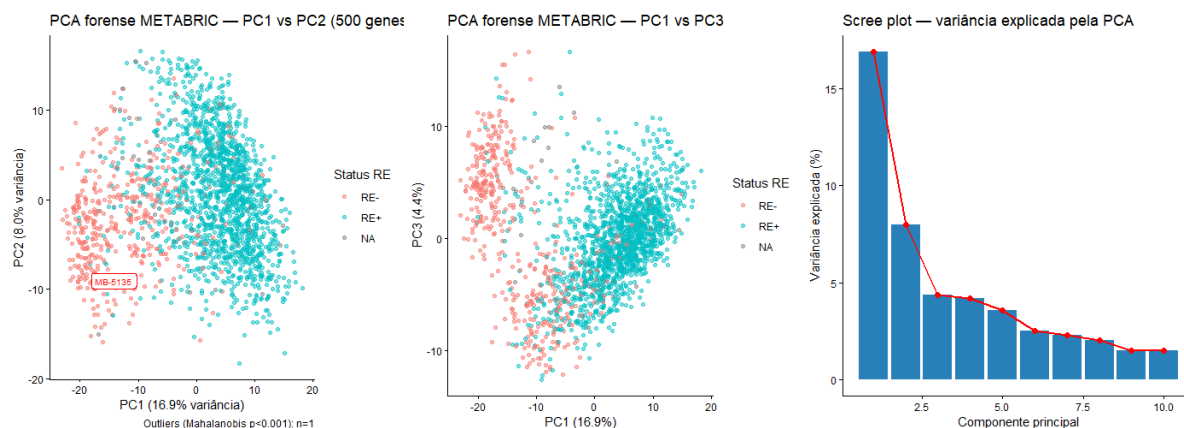


Figura 22 — PCA forense da matriz de expressão METABRIC. Componentes principais coloridos por status de RE. PC1 captura a principal dicotomia biológica (RE+/RE-), sem evidência de efeitos de batch dominantes, confirmando integridade dos dados.

A Figura 22 é uma figura de controle de qualidade, não uma análise prognóstica principal. Ela pergunta se a matriz METABRIC se comporta biologicamente como uma coorte de câncer de mama deveria se comportar. No primeiro painel, cada ponto é uma paciente e os eixos PC1 e PC2 são as duas maiores fontes de variação global na expressão gênica. A separação por status de RE ao longo do primeiro componente é esperada, pois a sinalização estrogênica é um dos eixos mais fortes da biologia transcricional do câncer de mama desde os estudos de Perou e Sørli. Quando uma matriz separa RE+ e RE- de forma coerente, isso sugere que o sinal biológico principal foi preservado.

O segundo painel troca PC2 por PC3 para verificar se a separação observada depende apenas de uma projeção específica. Se houvesse um lote técnico dominante, como uma placa, centro ou data de processamento, poderíamos observar agrupamentos artificiais sem relação clínica clara. Aqui, a leitura principal é que a estrutura global permanece plausível e não há evidência visual de um artefato grosseiro que explique os resultados do CorePAM. O terceiro painel, o *scree plot*, mostra quanta variância cada componente principal captura. Em termos simples, ele informa se a matriz é dominada por um único eixo ou distribuída em múltiplas fontes de variação. Essa verificação é equivalente a checar a qualidade de uma lâmina antes de emitir um laudo: não responde à pergunta clínica final, mas aumenta a confiança de que a amostra técnica é interpretável.

4.7 Análise de subgrupo por status de RE

A heterogeneidade biológica do câncer de mama se reflete, sobretudo, na dicotomia por status de receptor de estrogênio (RE) — doenças RE-positivas e RE-negativas têm histórias naturais, tratamentos e benefícios esperados de ferramentas de estratificação distintos. A Tabela 16 e a Figura 23 avaliam o desempenho do CorePAM separadamente nesses dois subgrupos, em SCAN-B e METABRIC. No SCAN-B, o escore é informativo em ambos os subgrupos com *hazard ratios* comparáveis e teste de interação não significativo ($p = 0,25$), ou seja, não há evidência estatística de que o CorePAM se comporte diferente entre RE+ e RE-. Na METABRIC, ao contrário, o sinal é marcadamente concentrado em RE+ (HR = 1,59; $p < 0,001$) e essencialmente ausente em RE- (HR = 0,97; $p = 0,75$), com teste de interação fortemente significativo ($p = 3,2 \times 10^{-7}$). Essa dissociação é relevante para interpretação clínica: em cenários neoadjuvantes e adjuvantes de câncer RE+/HER2-, onde a decisão quimioterapia-sim/quimioterapia-não é mais frequentemente ambígua, o CorePAM agrega valor prognóstico consistente; em doença RE-negativa (triplo-negativa e HER2+), a assinatura deve ser interpretada com cautela, especialmente em coortes de tratamento adjuvante longo.

Tabela 16 — Análise de subgrupo por status de RE em SCAN-B (N = 2 870; OS) e METABRIC (N = 1 936; DSS)

Coorte	Subgrupo	N	Eventos	HR (IC 95%)	p	C-index
SCAN-B (OS)	RE-positivo	2 646	287	1,92 (1,69-2,17)	<0,001	0,700
SCAN-B (OS)	RE-negativo	224	35	2,53 (1,51-4,25)	<0,001	0,697
SCAN-B (OS)	Interação	—	—	—	$p = 0,25$	—
METABRIC (DSS)	RE-positivo	1 497	460	1,59 (1,44-1,76)	<0,001	0,638
METABRIC (DSS)	RE-negativo	439	186	0,97 (0,82-1,15)	0,75	0,502
METABRIC (DSS)	Interação	—	—	—	$p = 3,2 \times 10^{-7}$	—

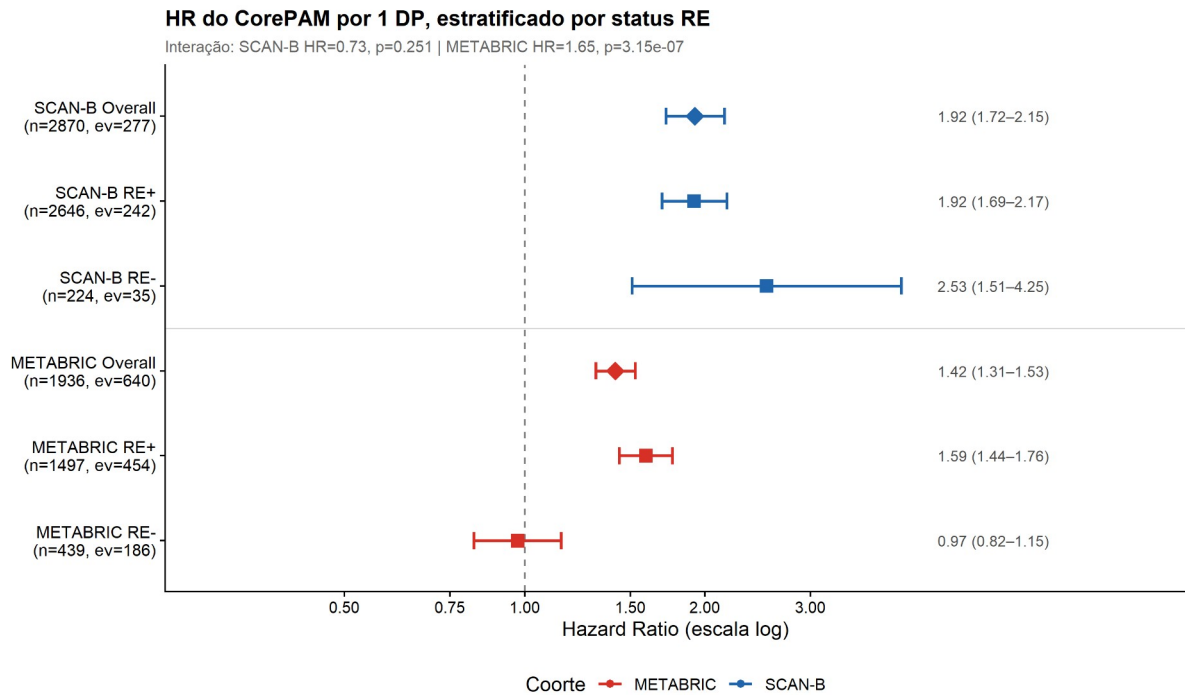


Figura 23 — Forest plot de subgrupos por status de receptor de estrogênio (RE-positivo e RE-negativo). No METABRIC (análise com maior poder), o sinal prognóstico está fortemente concentrado em RE+ (HR = 1,59) com ausência de sinal em RE- (HR = 0,97).

Na Figura 23, a linha vertical em $HR = 1$ marca ausência de associação e pontos à direita indicam que maior CorePAM se associa a maior risco. Em SCAN-B, RE+ e RE- ficam ambos à direita (embora o subgrupo RE- seja pequeno e mais incerto); em METABRIC, o ponto de RE+ está nitidamente à direita, enquanto o de RE- fica próximo de 1 — a tradução visual da interação significativa.

Em síntese, a análise por status de RE posiciona o CorePAM no mesmo nicho clínico das assinaturas estabelecidas: sua informação prognóstica é robusta e clinicamente acionável na doença RE-positiva — justamente o cenário em que a decisão sobre quimioterapia adjuvante é mais frequentemente ambígua — e deve ser interpretada com cautela na doença RE-negativa, cuja biologia de risco é governada por dimensões que o escore capta apenas de forma indireta.

4.8 Resposta patológica completa (pCR)

A Tabela 17 e a Figura 24 testam se o escore CorePAM — derivado para prever *sobrevida* — também se associa, de forma exploratória, a um desfecho clínico distinto: a *resposta patológica completa* após quimioterapia neoadjuvante (pCR), definida como ausência de doença residual invasiva na mama e axila na cirurgia. A medida apresentada, *OR por 1 DP*, representa a multiplicação da chance de pCR quando o escore aumenta em um desvio-padrão; valores acima de 1 indicam que pacientes com escores mais altos têm *maior* probabilidade de resposta completa. O padrão observado é consistente em quatro coortes neoadjuvantes independentes: OR agrupado = 1,69 (IC 95 % 1,39–2,05; $p = 1,9 \times 10^{-7}$), com $I^2 = 0\%$, compatível com homogeneidade entre as quatro coortes — embora o pequeno número de coortes ($K = 4$) limite a precisão dessa estimativa de heterogeneidade —, e consistente entre plataformas de *microarray* diferentes (Affymetrix e Agilent). A Figura 25, a Figura 26 e a Figura 27 ilustram os três aspectos complementares desse achado: capacidade discriminativa (AUC 0,58–0,72), separação das distribuições do escore entre pacientes que atingiram ou não pCR, e gradiente dose-resposta por quartis. Uma interpretação biológica plausível, ainda que não causal, é que tumores com maior assinatura de proliferação (característica dominante do escore CorePAM) tendem a ser simultaneamente mais agressivos e mais quimiossensíveis: o componente proliferativo que aumenta o risco prognóstico pode também associar-se a maior vulnerabilidade à quimioterapia citotóxica, embora a resposta dependa de subtipo, tratamento, imunidade tumoral e mecanismos de resistência.

Tabela 17 — Predição de pCR por coorte neoadjuvante. I-quadrado = 0%.

Coorte	N	pCR n (%)	OR por 1 DP (IC 95%)	p	AUC (IC 95%)
GSE25066	182	42 (23,1%)	1,44 (0,98-2,09)	0,060	0,576 (0,477-0,676)
GSE20194	278	56 (20,1%)	1,73 (1,25-2,38)	<0,001	0,660 (0,586-0,734)
GSE32646	115	27 (23,5%)	2,10 (1,31-3,39)	0,002	0,716 (0,611-0,821)
I-SPY1	122	32 (26,2%)	1,66 (1,05-2,63)	0,030	0,622 (0,517-0,726)
Agrupado — efeitos aleatórios (K=4)	697		1,69 (1,39-2,05)	$1,9 \times 10^{-7}$	

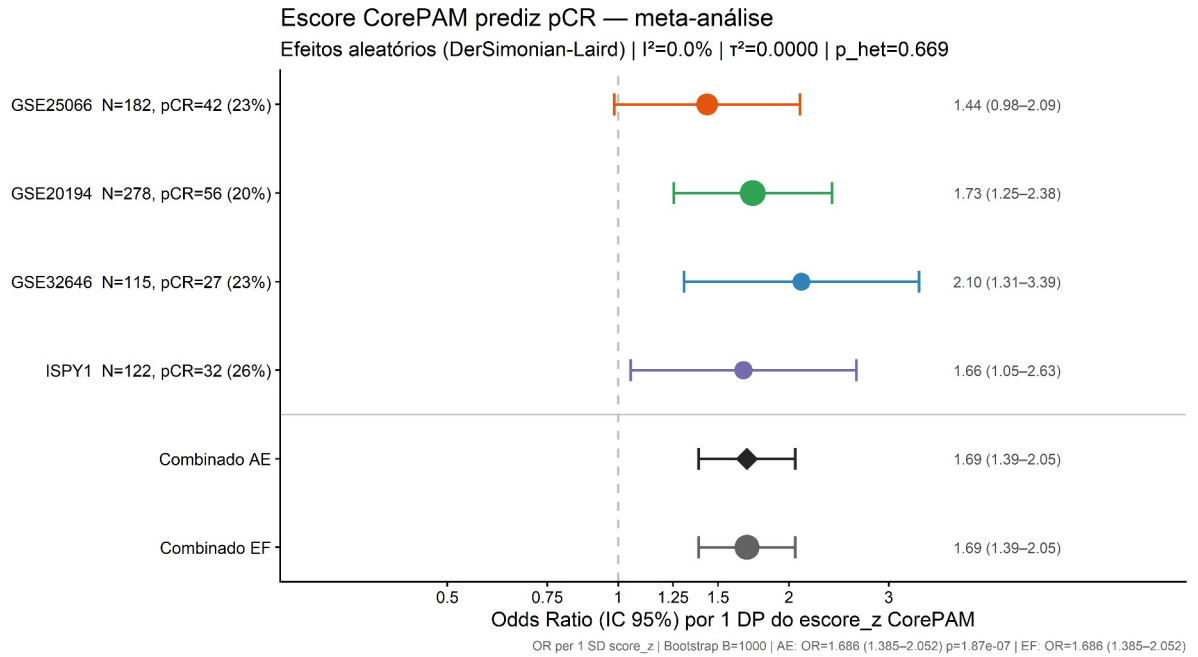


Figura 24 — Forest plot de pCR (OR por 1-DP, K = 4 coortes primárias). Modelo de efeitos aleatórios. OR agrupado = 1,69 (IC 95% 1,39-2,05; $p = 1,9 \times 10^{-7}$). I-quadrado = 0%, compatível com consistência entre coortes de diferentes plataformas (com a ressalva de que K = 4 limita a precisão da estimativa de heterogeneidade).

Na Figura 24, cada linha é uma coorte neoadjuvante e o ponto mostra o aumento da chance de pCR por 1 desvio-padrão do escore; a linha vertical em OR = 1 marca ausência de associação. Todas as estimativas ficam à direita de 1, e o diamante agrupado resume OR = 1,69 com I-quadrado = 0% — consistência notável entre fenótipos e plataformas neoadjuvantes diferentes.

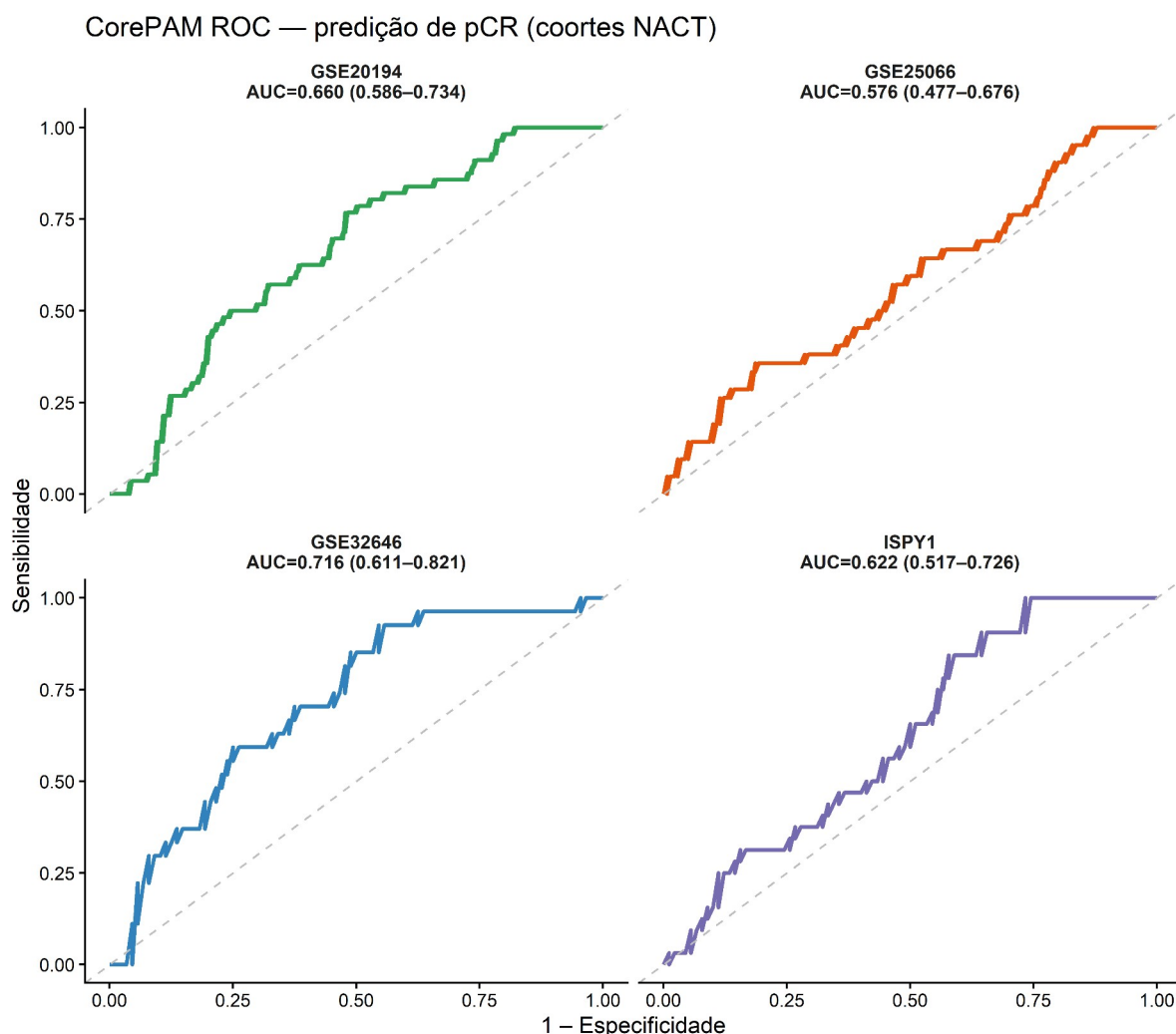


Figura 25 — Curvas ROC para predição de pCR em cada coorte neoadjuvante. AUC variou de 0,576 (GSE25066) a 0,716 (GSE32646), indicando capacidade discriminativa baixa a moderada — ainda assim notável, considerando que o escore foi derivado para sobrevida, não para pCR.

Na Figura 25, cada curva mostra a capacidade do escore de separar pacientes que atingiram pCR daquelas que não atingiram. A diagonal de referência representa desempenho ao acaso; curvas acima dela indicam discriminação. As AUCs entre 0,58 e 0,72 são baixas a moderadas, não definitivas, mas relevantes porque o CorePAM não foi criado para pCR. Em termos clínicos, isso significa que o escore sozinho não deve decidir neoadjuvância, mas carrega informação compatível com quimiossensibilidade. Essa leitura é parecida com a de Ki67: útil para compor o quadro, insuficiente como juiz único.

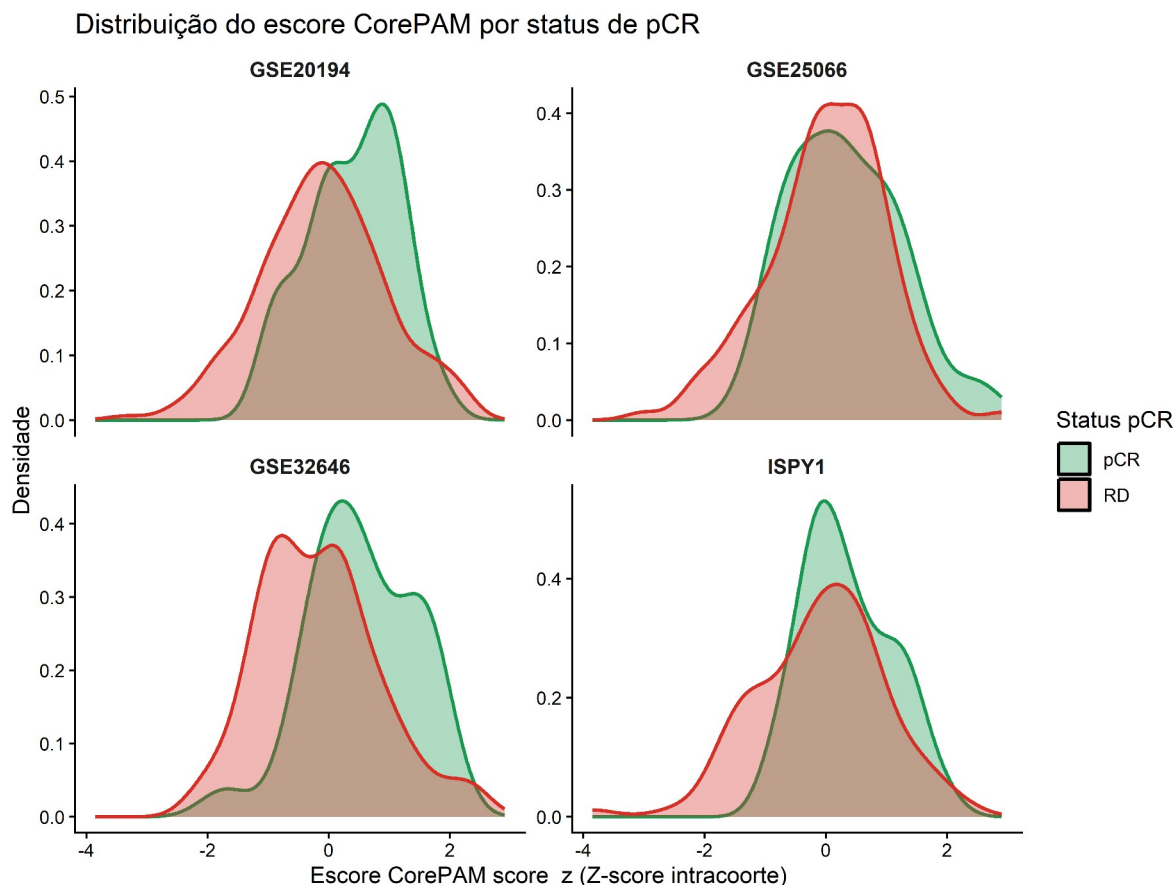


Figura 26 — Distribuição do escore CorePAM por status de pCR em cada coorte neoadjuvante. Pacientes com pCR (vermelho) apresentam escores sistematicamente mais altos, consistente com a hipótese de que tumores mais proliferativos são simultaneamente mais agressivos e mais quimiossensíveis.

Na Figura 26, o eixo horizontal é o escore CorePAM e as distribuições comparam pacientes com e sem pCR. O deslocamento das pacientes com pCR para escores mais altos mostra que o achado não depende apenas de um ponto de corte arbitrário; há mudança global da distribuição. Quando as curvas se sobrepõem, isso indica incerteza individual: muitas pacientes sem pCR também têm escore alto, e algumas com pCR têm escore baixo. Essa sobreposição é esperada em oncologia, porque resposta à quimioterapia depende de múltiplas dimensões além de proliferação, como subtipo, imunidade tumoral, reparo de DNA, esquema terapêutico e resistência farmacológica.

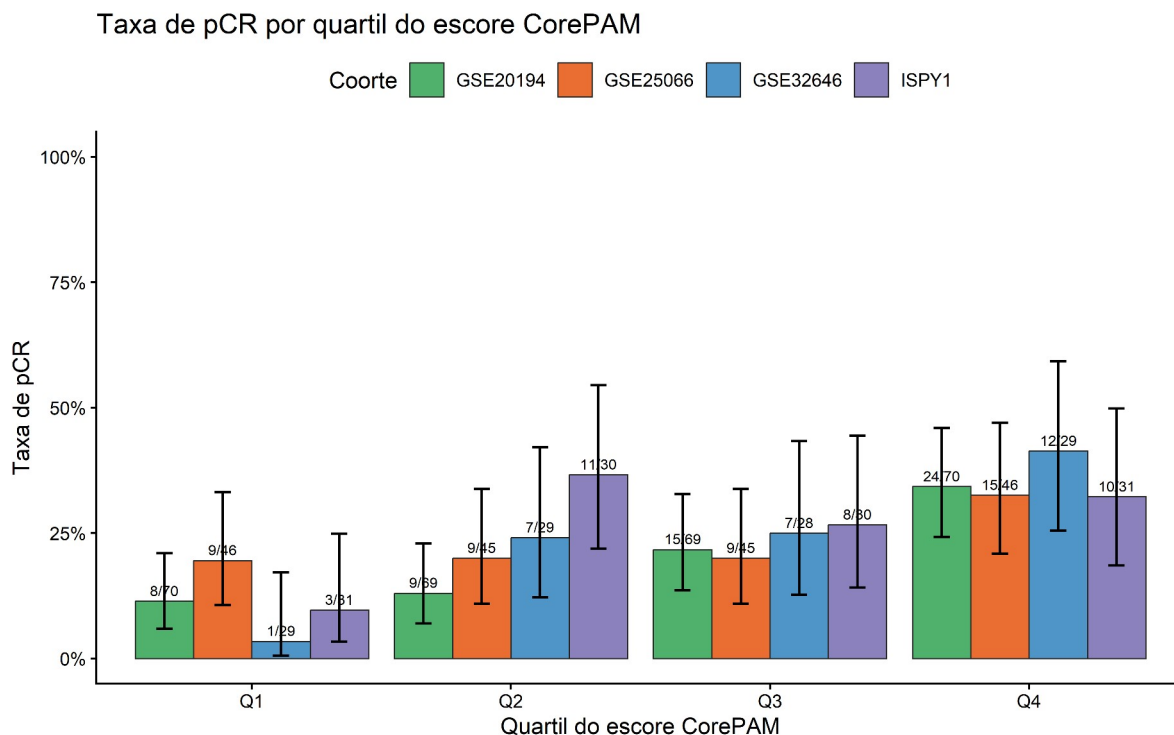


Figura 27 — Taxa de pCR por quartil do escore CorePAM em cada coorte neoadjuvante. Gradiente dose-resposta consistente: quartis mais altos associados a maiores taxas de pCR.

Na Figura 27, o escore é dividido em quatro grupos ordenados. A pergunta é se existe gradiente dose-resposta: quanto maior o CorePAM, maior a taxa de pCR? Esse tipo de leitura é familiar na clínica, como comparar resposta por quartis de Ki67 ou por categorias de carga tumoral. O padrão ascendente reforça coerência biológica, porque uma associação verdadeira tende a aparecer como gradiente, não apenas como diferença entre dois grupos escolhidos depois. Ainda assim, quartis são ferramenta de visualização; a análise principal permanece contínua, usando o escore completo.

A Figura 28 traduz esse achado para decisão clínica exploratória em pCR. Aqui, o limiar não é risco de morte, mas probabilidade mínima de resposta completa que justificaria uma conduta, como priorizar neoadjuvância, discutir intensificação ou considerar entrada em protocolo. A linha “tratar todos” equivale a assumir que todas as pacientes seriam candidatas à estratégia; a linha “tratar nenhum” equivale a não usar o escore para selecionar ninguém. A curva do CorePAM mostra em que limiares a seleção pelo escore melhora o balanço entre identificar respondedoras e selecionar falsos positivos.

Análise de Curva de Decisão — pCR

Benefício clínico líquido do CorePAM na predição de pCR

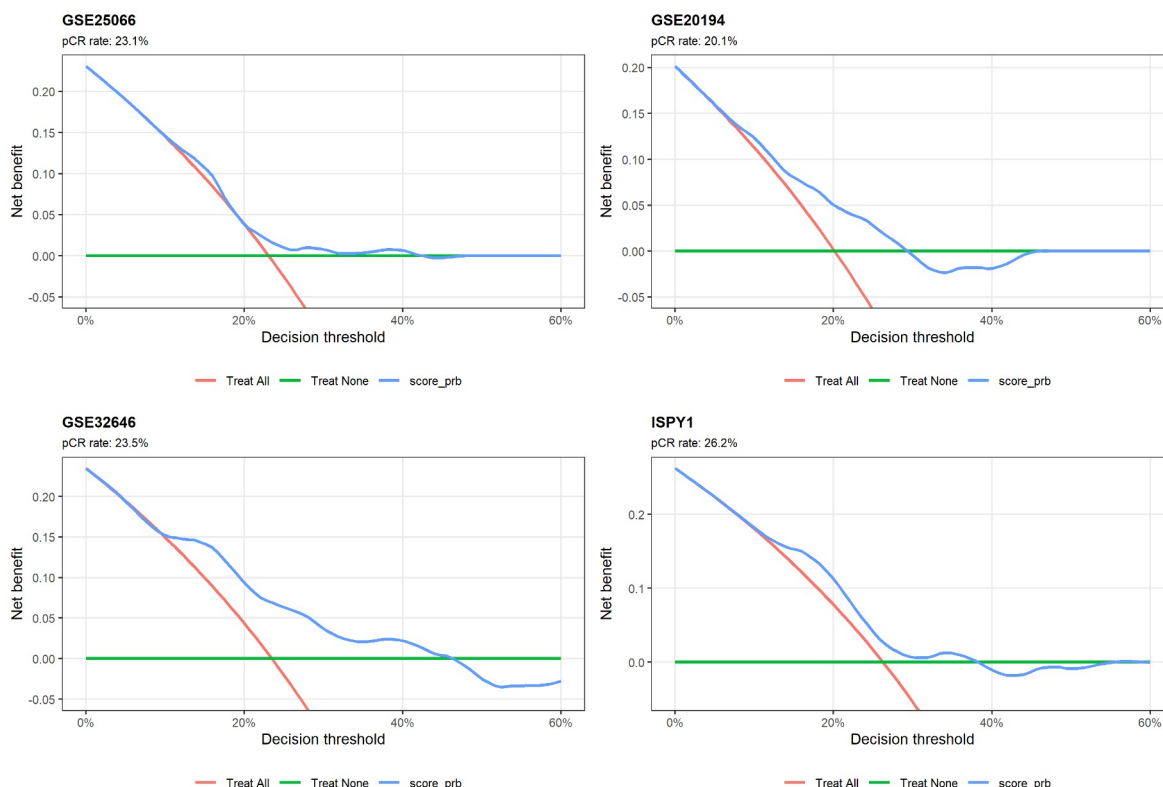


Figura 28 — Análise de curva de decisão para predição de pCR nas quatro coortes neoadjuvantes primárias. A curva do CorePAM é comparada com as estratégias de tratar todos e tratar nenhum em limiares de decisão de 0% a 60%.

Na Figura 28, o painel GSE25066 mostra benefício líquido modesto em limiares baixos a intermediários, com perda de utilidade quando o limiar ultrapassa a faixa em que o escore separa respondedoras de não respondedoras. Em GSE20194, o benefício se estende por faixa um pouco maior, compatível com a AUC mais alta dessa coorte. Em GSE32646 e I-SPY1, a curva permanece acima de “tratar todos” e “tratar nenhum” em parte relevante dos limiares, mas oscila em limiares altos, onde o número de pacientes selecionadas cai e a incerteza aumenta. A mensagem clínica é deliberadamente conservadora: o CorePAM sugere utilidade para enriquecer populações com maior chance de pCR, mas não deve ser interpretado como teste autônomo para escolher quimioterapia neoadjuvante.

Em conjunto, os resultados de pCR mostram que um escore derivado exclusivamente para prognóstico de sobrevida carrega informação reproduzível sobre quimiossensibilidade (OR agrupado = 1,69; $I^2 = 0\%$), coerente com o módulo de proliferação que compõe o CorePAM. Trata-se, porém, de achado secundário e exploratório: a discriminação é apenas moderada e o escore não substitui, isoladamente, a decisão de neoadjuvância — seu papel

plausível é enriquecer populações com maior probabilidade de resposta, hipótese que exige validação prospectiva dedicada.

4.9 Comparação *head-to-head* com PAM50-full

A Tabela 18 e a Figura 29 confrontam diretamente o CorePAM (24 genes) contra o comparador PAM50-full penalizado no mesmo *pipeline* de derivação. Para evitar ambiguidade, vale repetir: o PAM50 original tem 50 genes; o comparador desta tabela tem 49 porque o ajuste *elastic-net* completo, feito no SCAN-B com os 50 genes candidatos, manteve 49 coeficientes não-zero no maior ponto da trajetória. Isso é comportamento esperado de um modelo penalizado e não indica gene ausente nem erro de cobertura.

Cada coorte é avaliada em dois modos de pontuação: *intra* (parâmetros de normalização recalculados localmente) e *congelado* (parâmetros fixados do SCAN-B). A coluna C_{24} é o *C-index* do CorePAM e C_{49} é o *C-index* do comparador PAM50-full penalizado. Delta-C positivo significa vantagem do modelo de 24 genes. O padrão observado é consistente: em TCGA-BRCA (*RNA-seq*, mais próximo da plataforma de treino) os modelos são equivalentes, mas nas três coortes *microarray* (METABRIC, GSE20685, GSE1456) o modelo de 24 genes supera significativamente o de 49. A vantagem cresce à medida que a distância de plataforma em relação ao SCAN-B aumenta, sugerindo que genes adicionais podem carregar ruído técnico que se transporta mal. A coluna r mostra a correlação entre os dois escores paciente-a-paciente; apesar da correlação alta, pequenas divergências em regiões decisivas do ranking explicam a diferença de desempenho.

Tabela 18 — Comparação *head-to-head*: CorePAM (24 genes) vs comparador PAM50-full *elastic-net* com 49 coeficientes não-zero. O PAM50 canônico permanece definido por 50 genes; o número 49 refere-se apenas ao maior modelo penalizado efetivamente obtido neste *pipeline*.

Coorte	Modo	N	C_{24}	C_{49}	Delta-C	IC 95%	r
TCGA-BRCA	<i>intra</i>	1 072	0,624	0,631	-0,008	(-0,038, +0,022)	0,853
TCGA-BRCA	<i>congelado</i>	1 072	0,624	0,634	-0,010	(-0,039, +0,020)	0,864
METABRIC	<i>intra</i>	1 978	0,638	0,616	+0,022	(+0,011, +0,032)	0,902
METABRIC	<i>congelado</i>	1 978	0,638	0,622	+0,017	(+0,007, +0,027)	0,914
GSE20685	<i>intra</i>	327	0,626	0,571	+0,055	(+0,010, +0,101)	0,823
GSE20685	<i>congelado</i>	327	0,591	0,544	+0,047	(+0,007, +0,085)	0,843
GSE1456	<i>intra</i>	159	0,661	0,564	+0,097	(+0,032, +0,166)	0,697
GSE1456	<i>congelado</i>	159	0,622	0,510	+0,111	(+0,045, +0,181)	0,617

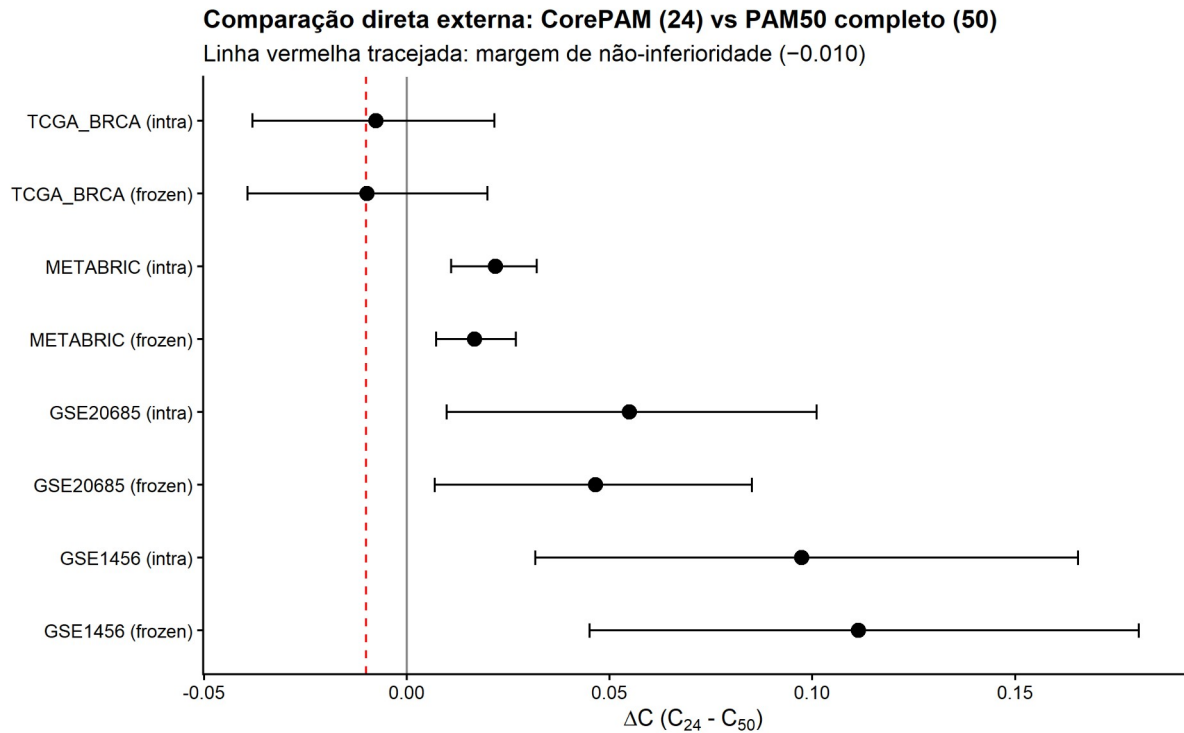


Figura 29 — Comparação head-to-head CorePAM (24 genes) vs PAM50-full penalizado (49 coeficientes não-zero) por coorte e modo de pontuação. Em TCGA-BRCA, desempenho equivalente. Nas três coortes restantes, o CorePAM superou significativamente o comparador penalizado, com vantagem crescente com a distância de plataforma em relação ao SCAN-B.

Na Figura 29, a ponte com a prática clínica é direta: um exame maior nem sempre é melhor se parte da informação adicional é instável entre laboratórios. Onde o Delta-C é positivo (as três coortes *microarray*), menos genes significam não apenas menor custo, mas menor exposição a ruído de plataforma — vantagem que cresce com a distância tecnológica em relação à coorte de treino.

4.10 Comparação com assinaturas estabelecidas

A Tabela 19 e a Figura 30 posicionam o CorePAM em relação a duas assinaturas prognósticas consolidadas: *ROR-S* (componente de subtipo do Prosigna, o teste comercial baseado no PAM50) e *OncotypeDX-RS* (*Recurrence Score*, 21 genes, teste clínico padrão-ouro em câncer de mama RE+). Ambas foram computadas nas mesmas coortes pela biblioteca *genefu*, sob especificação idêntica, garantindo comparação justa. As colunas *C* mostram a discriminação de cada assinatura em cada coorte e as colunas *HR* a razão de risco por 1 DP do escore. O padrão é bem definido: no SCAN-B (coorte de desenvolvimento, *RNA-seq*) o CorePAM supera claramente *ROR-S* e *OncotypeDX-RS*. No TCGA-BRCA (também *RNA-seq*) o CorePAM mantém leve vantagem de discriminação. Nas coortes *microarray* (METABRIC, GSE20685), a ordem se inverte — *OncotypeDX-RS* assume a vantagem. Essa inversão plataforma-dependente é esperada e informativa: cada assinatura carrega a assinatura da plataforma em que foi desenvolvida, e nenhuma ferramenta é universalmente ótima. No entanto, o CorePAM é o único dos três que foi derivado com foco explícito em parcimônia e estabilidade *cross-plataforma*, e permanece competitivo: tem vantagem nas coortes *RNA-seq* (SCAN-B e TCGA-BRCA), onde foi desenvolvido, e desvantagem nas coortes *microarray* (METABRIC e GSE20685), em que *ROR-S* e *OncotypeDX-RS* discriminam ligeiramente melhor. Nenhuma das três ferramentas é universalmente superior; o desempenho de cada uma reflete a plataforma em que foi originalmente construída.

Tabela 19 — Comparação com *ROR-S* e *OncotypeDX-RS* (implementações *genefu*)

Coorte	N	CorePAM C	<i>ROR-S</i> C	ODX-RS C	CorePAM HR	<i>ROR-S</i> HR	ODX-RS HR
SCAN-B	3 069	0,698	0,641	0,607	1,92	1,56	1,40
TCGA-BRCA	1 072	0,624	0,602	0,586	1,20	1,25	1,13
METABRIC	1 978	0,638	0,643	0,665	1,41	1,53	1,51
GSE20685	327	0,626	0,646	0,651	1,40	1,55	1,61

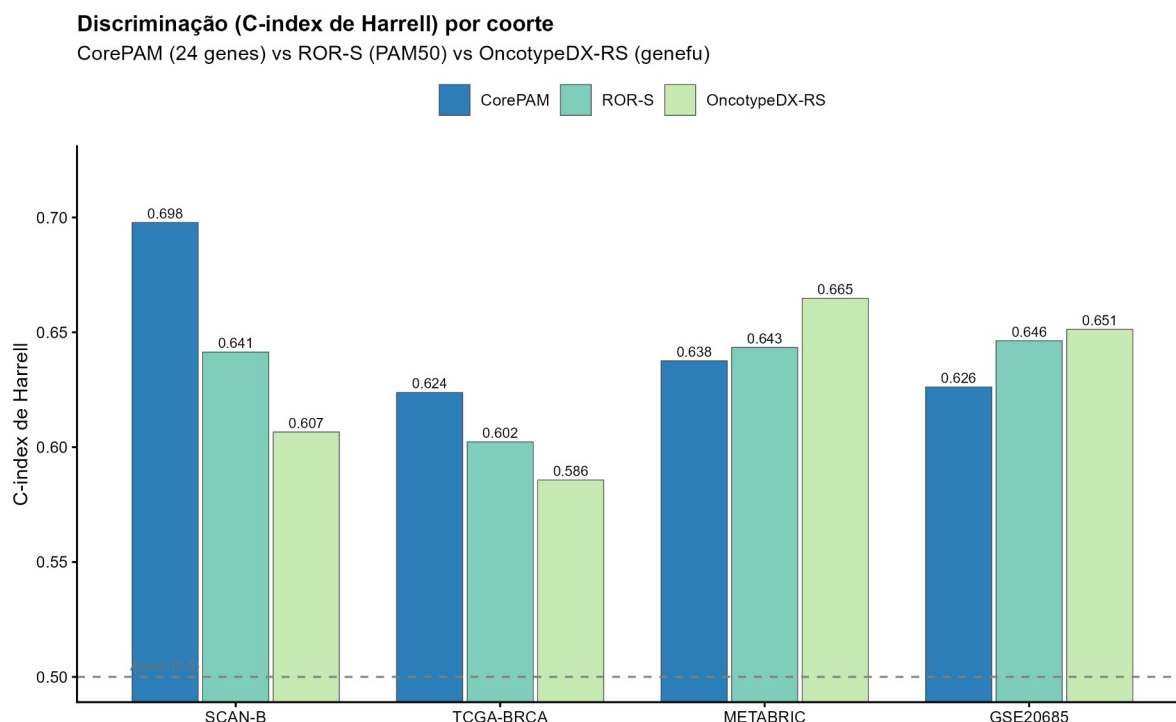


Figura 30 — Comparação de C-index: CorePAM vs ROR-S vs OncotypeDX-RS (implementações genefu) por coorte. O CorePAM demonstra desempenho competitivo, com vantagem nas coortes RNA-seq e déficits nas coortes microarray — refletindo o domínio de treinamento.

Na Figura 30, a mensagem importante não é declarar um vencedor universal — a alternância de desempenho por plataforma é própria da literatura de assinaturas gênicas, em que testes diferentes captam dimensões parcialmente sobrepostas. Para a prática clínica: o CorePAM é competitivo como prova de conceito, mas ainda não substitui ensaios comerciais validados prospectivamente — aponta, sim, um caminho de ensaio mais simples e potencialmente mais acessível.

5. DISCUSSÃO

5.1 O CorePAM no contexto da medicina de precisão em oncologia

A medicina de precisão em oncologia busca identificar, para cada paciente individual, o tratamento com maior probabilidade de benefício e menor probabilidade de toxicidade evitável. No câncer de mama, as assinaturas gênicas multigenes — *OncotypeDX*²²⁻²⁵, *MammaPrint*²⁶⁻²⁸, *PAM50/Prosigna*^{13,17,18}, *EndoPredict*²⁹ e *Breast Cancer Index*³⁰ — já pouparam milhares de mulheres no mundo de quimioterapia adjuvante desnecessária, especialmente em doença RE-positiva, HER2-negativa, linfonodo-negativa ou com 1-3 linfonodos positivos^{20,21}. Em termos práticos, estimativas conservadoras indicam que um terço a metade das pacientes com câncer de mama RE+/HER2- estágio precoce elegíveis para quimioterapia adjuvante no modelo clínico tradicional pode, com segurança, evitá-la quando estratificada por essas ferramentas — reduzindo a exposição a toxicidades agudas e tardias da quimioterapia, como infertilidade induzida, cardiotoxicidade, neoplasias secundárias e, mais raramente, mortalidade relacionada ao tratamento, cujo impacto populacional é substancial^{4,7}.

No entanto, a realidade global é marcada por profundas desigualdades de acesso a essa tecnologia. O custo individual dos testes (US\$ 3 000-4 000 nos EUA, preços análogos em mercado de saúde suplementar no Brasil), a dependência de laboratórios de referência centralizados e a logística de envio internacional são barreiras reais. No Brasil, as assinaturas genômicas prognósticas (*OncotypeDX*, *Prosigna*, *MammaPrint*, *EndoPredict*) não estão incorporadas ao Sistema Único de Saúde (SUS) para uso rotineiro na decisão de quimioterapia adjuvante; sua adoção em oncologia permanece em avaliação pela Comissão Nacional de Incorporação de Tecnologias no SUS (Conitec), inclusive por meio de consultas públicas sobre testagem genômica. A consequência prática é que, mesmo em centros acadêmicos públicos brasileiros de ponta, a indicação de quimioterapia adjuvante em doença RE+/HER2- continua a se apoiar majoritariamente em critérios clínico-patológicos clássicos — idade, T, N, grau, Ki67 por IHQ —, produzindo inevitavelmente um número de sub e sobre-tratamentos que a literatura internacional já demonstrou ser evitável.

O CorePAM se insere neste contexto como **demonstração de princípio**: a informação prognóstica contida no painel PAM50 pode ser capturada por um subconjunto substancialmente menor de 24 genes, com desempenho equivalente ou superior ao maior comparador PAM50-full penalizado obtido no mesmo procedimento. Vinte e quatro genes

representam complexidade analítica reduzida — tanto em termos de número de sondas em RT-qPCR multiplex, quanto em tamanho de painel de *RNA-seq* direcionado —, o que se traduz potencialmente em menor custo unitário por ensaio, menor exigência de infraestrutura e maior plausibilidade de execução em laboratórios de hospitais universitários brasileiros, sem dependência de envio internacional. O presente trabalho não se propõe a substituir um ensaio comercial por outro equivalente, mas sim a mostrar que a estrutura subjacente de informação prognóstica admite compressão considerável, o que é uma descoberta com implicações econômicas e de equidade não triviais.

5.2 Derivação reprodutível: inovação metodológica

Dois elementos da metodologia do presente estudo merecem ênfase por inovarem, ou ao menos sistematizarem de forma mais rigorosa, práticas comuns em estudos de desenvolvimento de assinaturas gênicas.

O primeiro é o **critério de não inferioridade orientado por dados** (*data-driven non-inferiority criterion*). A prática convencional em estudos de redução de painel consiste em pré-especificar um número-alvo de genes — 10, 20, 50 — e ajustar o procedimento de seleção para entregar exatamente esse número. O problema é que a escolha do número-alvo é frequentemente arbitrária ou retroativamente justificada. A presente abordagem inverte a lógica: fixamos a margem de perda de desempenho aceitável ($\Delta C\text{-index} = 0,010$) e deixamos a fronteira de Pareto determinar o número de genes mínimo. A quantidade 24 emerge como consequência da geometria dos dados, não como premissa do analista. Essa inversão tem uma consequência estatística importante: o número de genes deixa de ser um parâmetro que multiplica o número de comparações múltiplas implícitas, passando a ser um resultado da análise única pré-especificada.

O segundo é a **atribuição determinística de *folds* por hash SHA-256**. Em *10-fold cross-validation*, a escolha da semente aleatória é uma fonte silenciosa e subestimada de variação: diferentes seeds podem alterar tanto o conjunto de genes selecionados quanto a estimativa OOF do *C-index* em margens não desprezíveis. O que geralmente se reporta é o resultado obtido com *uma* semente particular, sem justificativa de por que essa, e não outra. A atribuição SHA-256 elimina a arbitrariedade: o *fold* de cada paciente é função determinística do seu próprio identificador, e qualquer análise que seja rodada no mesmo conjunto de dados, em qualquer máquina, em qualquer linguagem que implemente SHA-256 adequadamente, produzirá exatamente os mesmos *folds*. É uma propriedade de *reprodutibilidade universal*

mais forte do que a de uma seed fixa, pois não depende do estado interno do gerador pseudo-aleatório da linguagem.

A combinação desses dois elementos — critério pré-especificado não-circular e atribuição determinística universal — torna a derivação do CorePAM reproduzível de primeiros princípios. Qualquer pesquisador no mundo com acesso aos dados da SCAN-B (GSE96058) e ao código público pode replicar a derivação bit a bit, sem discricionariedade adicional. Esta propriedade de *replicabilidade forte* é, em nosso entendimento, um padrão que estudos futuros de derivação de assinaturas gênicas deveriam aspirar.

5.3 Coerência biológica: por que esses 24 genes?

Os 24 genes do CorePAM não são um conjunto arbitrário; demonstram coerência biológica notável, cobrindo os quatro eixos funcionais reconhecidos do PAM50 e da biologia intrínseca do câncer de mama, além de reguladores correlatos⁷⁻¹⁰:

- **Diferenciação luminal / dependência hormonal:** ESR1, PGR, BCL2, NAT1, MLPH, BLVRA. ESR1 (receptor de estrogênio) é o regulador mestre do programa luminal; sua expressão é alta em tumores RE-positivos que dependem da via estrogênica e, portanto, respondem à terapia endócrina. PGR (receptor de progesterona) é um alvo direto da sinalização estrogênica funcional. BCL2 é regulador antiapoptótico classicamente associado a pior resposta à quimioterapia mas melhor prognóstico global em RE+. NAT1 e MLPH são genes menos conhecidos mas consistentemente altos em luminais bem diferenciados. BLVRA (biliverdina redutase A) integra o metabolismo de heme e a resposta a estresse oxidativo, acompanhando o fenótipo luminal bem diferenciado.
- **Proliferação celular:** MYBL2, PTTG1, EXO1, MYC, CENPF. Este cluster captura, em nível transcricional, o que o Ki67 tenta representar isoladamente na IHQ. Estes genes relacionam-se à progressão do ciclo celular (MYBL2), à segregação cromossômica e à instabilidade mitótica (PTTG1, CENPF), ao reparo de DNA (EXO1) e à sinalização *downstream* de oncogenes (MYC). A alta expressão conjunta deste módulo define tumores biologicamente agressivos, independentemente do subtipo. Na prática clínica, é este módulo que o Ki67 tenta capturar — imperfeita e inconsistentemente, por problemas conhecidos de variação inter-observador⁷².
- **Fenótipo basal e mioepitelial:** KRT5, KRT17, FOXC1. As citoqueratinas de alto peso molecular (KRT5, KRT17) e o fator de transcrição FOXC1 marcam o fenótipo basal/mioepitelial, dominante em tumores triplo-negativos do tipo basal-like.
- **Eixo HER2:** ERBB2, GRB7, MIA. ERBB2 é o gene-alvo clássico (amplificado em ~15-20% dos cânceres de mama); GRB7 co-localiza no amplicon 17q12; MIA integra a resposta *downstream*.
- **Demais reguladores selecionados:** ACTR3B (citoesqueleto/motilidade celular), MDM2 (regulação de p53), SFRP1 (via Wnt/supressão tumoral), GPR160 (receptor acoplado a proteína G), PHGDH (metabolismo de serina), CXXC5 (regulação epigenética) e FGFR4 (sinalização de FGF) completam os 24 genes, refinando o sinal prognóstico dentro e ao redor dos quatro eixos principais.

A preservação dos quatro eixos no CorePAM é biologicamente não trivial: um algoritmo de redução de dimensionalidade puramente maximizador de desempenho poderia, em princípio, concentrar-se em um ou dois eixos (tipicamente proliferação) e descartar os outros, perdendo heterogeneidade biológica em troca de performance bruta. O fato de a *elastic-net* com $\alpha = 0,5$ preservar os quatro eixos é reflexo de sua propriedade de

grouping effect sob colinearidade: genes correlacionados dentro de um mesmo módulo tendem a ser selecionados em conjunto, produzindo assinatura biologicamente equilibrada.

Uma implicação prática importante: o CorePAM, embora selecionado pelo algoritmo de otimização, poderia ter sido racionalmente proposto por um biólogo do câncer de mama que se perguntasse “qual é o subconjunto mínimo de genes que captura proliferação, diferenciação luminal, basal e HER2?”. A convergência entre escolha orientada por dados e escolha racional biológica é um indicador forte de que o resultado não é um artefato de overfitting.

5.4 Transportabilidade *cross-plataforma*

Um dos resultados mais robustos do estudo é a transportabilidade do CorePAM para plataformas intencionalmente diversas da de treinamento. Os pesos, derivados em *RNA-seq* *Illumina* (SCAN-B), preservaram sinal prognóstico significativo quando aplicados a *RNA-seq* STAR/GDC (TCGA-BRCA), *microarray* *Illumina* HT-12 (METABRIC) e dois painéis *Affymetrix* distintos (HG-U133+ em GSE20685 e HGU133A/B em GSE1456). A normalização intra-coorte por *z-score* é o viabilizador técnico dessa transportabilidade: ao operar em unidades relativas — o quanto a expressão de um gene em uma amostra desvia da média da coorte — o escore abstrai as diferenças de calibração entre plataformas. Um *z-score* de +1,5 em METABRIC significa “1,5 desvios acima da média dos tumores de mama daquela coorte”, independentemente de o valor absoluto de expressão estar em escala $\log_2$ *Illumina* ou $\log$ -CPM *RNA-seq*.

A atenuação mensurável em plataformas *Affymetrix* (correlação $r = 0,916 - 0,923$ com *z-score* congelado, versus $r \geq 0,958$ em *RNA-seq* e *Illumina*) é um achado honesto que não deve ser escondido. A hipótese plausível é que a arquitetura das sondas *Affymetrix* HG-U133 — baseada em *perfect match / mismatch pairs* com algoritmos de sumarização (RMA) que diferem substancialmente dos *pipelines* *Illumina* — introduz distorções específicas em determinados transcritos do PAM50 (particularmente genes de baixa expressão basal ou com alta heterogeneidade de isoformas). Esta observação informa uma recomendação clínica clara: para implantação futura, calibração plataforma-específica — ou seja, estudos menores para re-estimativa da escala do escore em cada nova plataforma — seria prudente.

Uma observação conceitual: transportabilidade para novas *plataformas* é distinta de transportabilidade para novas *populações*. O CorePAM foi testado em coortes predominantemente europeias ou de descendência europeia (SCAN-B, TCGA-BRCA,

METABRIC, GSE1456), com uma coorte asiática (GSE20685). Não há dados sobre desempenho em populações africanas, latino-americanas ou brasileiras especificamente. Como demonstrado por estudos de heterogeneidade genômica populacional, sub-coortes com prevalência distinta de subtipos moleculares (e.g., maior frequência de basal-like em populações africanas) podem alterar a calibração, embora a discriminação tenda a ser mais preservada.

5.5 O paradoxo de menos genes, mais informação

Em três das quatro coortes externas — METABRIC, GSE20685 e GSE1456 —, o modelo CorePAM de 24 genes superou significativamente o comparador PAM50-full penalizado, que manteve 49 coeficientes não-zero, em discriminação (*C-index*). O PAM50 canônico continua sendo um painel de 50 genes; o número 49 aqui pertence apenas ao comparador estatístico ajustado no nosso *pipeline*. Apenas em TCGA-BRCA os dois modelos foram equivalentes. A princípio, este resultado é contraintuitivo: como um modelo com menos informação pode discriminar melhor do que um com mais informação?

A explicação reside classicamente no **trade-off viés-variância** da teoria estatística. Um modelo com até 49 coeficientes não-zero tem maior capacidade de se adaptar ao ruído idiossincrático da coorte de treinamento — ou seja, apresenta maior variância de estimativa. Quando este modelo é aplicado a uma coorte externa com distribuição ligeiramente deslocada (plataforma diferente, população diferente, protocolo de processamento diferente), o excesso de parâmetros capturado como sinal no SCAN-B passa a atuar como ruído residual, degradando o desempenho out-of-sample. O modelo de 24 genes, por construção, descarta coeficientes pequenos que refletem majoritariamente ruído e mantém apenas os sinais robustos — sendo, portanto, menos sujeito a esta degradação.

A magnitude do ganho do modelo mais parcimonioso aumenta com a distância da plataforma de teste em relação à plataforma de treinamento (maior em *Affymetrix* GSE1456, menor em TCGA-BRCA). Este gradiente é consistente com a interpretação: quanto mais distante a plataforma, mais o ruído específico do SCAN-B captado pelo comparador penalizado maior deixa de ser útil e passa a atrapalhar.

Um corolário metodológico desta observação: em desenvolvimento de modelos prognósticos destinados a uso *cross-plataforma*, a regularização agressiva deve ser considerada não como concessão cosmética (reduzir tamanho), mas como ferramenta ativa de

melhoria de generalização. O CorePAM não é apenas mais barato de implementar; é, provavelmente, mais transportável.

5.6 A questão do *z-score*: coorte vs amostra única

A decisão de utilizar *z-score* intra-coorte como normalização é a pedra angular da estratégia *cross-plataforma*, mas introduz uma limitação pragmática: o escore de uma paciente individual depende não apenas de sua própria expressão gênica, mas também da distribuição de expressão da coorte da qual faz parte. No uso clínico futuro — em que cada paciente chega isolada —, seria necessário um procedimento alternativo.

Três estratégias são conceitualmente possíveis:

1. **Z-score congelado (parâmetros SCAN-B):** fixar média e desvio-padrão de cada gene nos valores da coorte de treinamento, aplicando-os a cada nova amostra. A análise de sensibilidade mostrou que esta estratégia preserva discriminação quase perfeitamente em *RNA-seq* e *microarray Illumina* (correlação entre escore congelado e escore intra-coorte $r \geq 0,958$), mas sofre atenuação mensurável em *Affymetrix* ($\Delta C\text{-index}$ de $-0,036$ a $-0,040$).
2. **Z-score com coorte de referência local:** calibração inicial em coorte local pequena (e.g., 100-200 pacientes do próprio laboratório de implantação), e subsequente uso destes parâmetros para novas amostras. Esta estratégia requer investimento inicial, mas seria provavelmente a mais robusta.
3. **Ranking-based scoring:** re-expressão do escore em termos de ranks relativos em vez de *z-scores*, que são mais estáveis a variações de plataforma. Esta alternativa não foi avaliada no presente estudo e merece investigação futura.

A atenuação em *Affymetrix* sob parâmetros congelados não invalida a estratégia de *z-score* congelado: mesmo com $\Delta C\text{-index}$ de $-0,04$, o CorePAM mantém discriminação significativa. Mas sinaliza que, para plataformas suficientemente distintas da de treinamento, recalibração local é recomendável.

5.7 O sinal em doença RE-positiva

No METABRIC — a coorte com maior poder estatístico para análise de subgrupo ($N = 1\,936$, 646 eventos DSS) —, observou-se separação expressiva do sinal prognóstico: HR = 1,59 no subgrupo RE-positivo vs HR = 0,97 no subgrupo RE-negativo, com teste de interação altamente significativo ($p < 0,001$). Este achado tem implicação clínica direta, pois é precisamente na doença RE-positiva, HER2-negativa, estágio precoce que os ensaios multigenes são mais utilizados para decisão de quimioterapia adjuvante.

A interpretação biológica é coerente com a composição do CorePAM. Em tumores RE-positivos, a heterogeneidade biológica mais clinicamente relevante é a variação dentro do espectro luminal — de *Luminal A* (baixa proliferação, excelente prognóstico com endocrinoterapia isolada) a *Luminal B* (alta proliferação, risco residual após endocrinoterapia,

candidato a quimioterapia adicional). O CorePAM captura precisamente este espectro, com genes de proliferação e de diferenciação luminal ponderados em direções opostas. Em tumores RE-negativos — dominados por fenótipos basal-like ou *HER2-enriched* —, a heterogeneidade biológica relevante para prognóstico está em dimensões diferentes (carga de infiltrado linfoide⁷³, instabilidade genômica, alvos terapêuticos específicos como HER2), que o CorePAM captura parcial e indiretamente.

A implicação clínica é que o CorePAM — como o *OncotypeDX*, o *MammaPrint* e o *Prosigna* antes dele — tem nicho de utilidade mais claro em doença RE-positiva HER2-negativa. Em RE-negativa, sua aplicação como ferramenta de decisão quimioterápica é menos justificada pela biologia.

5.8 A ponte entre prognóstico e predição de pCR

Um dos resultados mais surpreendentes foi a demonstração de que o CorePAM — derivado exclusivamente para sobrevida global — está significativamente associado à resposta patológica completa (pCR) à quimioterapia neoadjuvante em coortes independentes, com OR agrupado = 1,69 (IC 95% 1,39-2,05; $p < 0,001$) e heterogeneidade I-quadrado = 0%. Esta consistência entre coortes de diferentes plataformas e protocolos neoadjuvantes é notável e sugere que a informação biológica capturada pelo escore transcende o *endpoint* para o qual foi treinado.

A ponte conceitual entre prognóstico e predição de pCR reside no módulo de proliferação dos 24 genes. Proliferação celular alta é, simultaneamente, marcador de pior prognóstico de longo prazo (tumores mais agressivos metastatizam mais) e de melhor resposta à quimioterapia neoadjuvante (drogas citotóxicas atuam preferencialmente em células em divisão). Este *paradoxo* — que tumores biologicamente ruins são, em parte, mais quimiossensíveis — é clássico na oncologia mamária^{74,75} e está na base da utilidade dos ensaios de pCR como biomarcador de resposta. O CorePAM, ao capturar o módulo de proliferação, herda automaticamente essa propriedade dual.

Uma implicação é que o escore pode ter utilidade bivalente em cenários clínicos específicos: em doença RE-positiva, para decidir se é necessário adicionar quimioterapia à endocrinoterapia (como *OncotypeDX*); em doença de alto risco que irá receber quimioterapia neoadjuvante, para estratificar probabilidade de pCR e orientar protocolos intensificados ou desescalados. Esta dupla utilidade justifica investigação clínica prospectiva dedicada.

5.9 Calibração: boa discriminação não garante probabilidades corretas

A calibração — a concordância entre risco predito e risco observado em cada faixa de predição — é, nas palavras de Van Calster et al.⁶⁵, o *calcanhar de Aquiles* da analítica preditiva. Discriminação mede se pacientes com desfecho pior têm escores mais altos que pacientes com desfecho melhor (uma propriedade de ranking), mas não garante que as probabilidades preditas estejam em escala correta.

O CorePAM apresentou discriminação boa (*C-index* de 0,62 a 0,70 nas coortes externas, agrupado em torno de 0,67), mas calibração com limitações claras em coortes externas (slopes de 1,57 a 2,74, indicando subconfiança — o modelo comprime demais a separação de risco em relação à população-alvo). Este padrão não é peculiar do CorePAM; é comum em modelos derivados em uma coorte e aplicados em outra com prevalência de eventos, seguimento e composição clínica distintas.

A implicação prática é que o CorePAM, como derivado no SCAN-B, seria inadequado para fornecer probabilidades calibradas de eventos a 5 ou 10 anos em população brasileira sem recalibração. Uma implementação clínica real exigiria recalibração — tipicamente por regressão de recalibração (*intercept and slope recalibration*) em uma coorte local de tamanho modesto ($N = 200-500$) — antes de uso para decisão individual. A discriminação, em contraste, tende a ser mais preservada e pode ser mais facilmente transportada.

5.10 Reflexões metodológicas para estudos prognósticos futuros

Alguns aprendizados gerais emergem do trabalho que podem ser úteis para estudos futuros de derivação de assinaturas ou modelos prognósticos em oncologia:

1. **Pré-especificar a margem, não o tamanho.** Pré-especificar um critério quantitativo de não inferioridade (por exemplo, $\Delta C\text{-index} \leq 0,010$) e deixar o tamanho do modelo emergir do dado é metodologicamente superior a pré-especificar um número-alvo de variáveis.
2. **Adotar atribuição determinística de folds.** Isto elimina a variabilidade silenciosa de seed e garante replicabilidade forte.
3. **Testar transportabilidade em coortes intencionalmente heterogêneas.** Validar o modelo em coortes idênticas em plataforma e população da de treinamento subestima a degradação que ocorrerá na clínica real.
4. **Reportar calibração com a mesma ênfase que discriminação.** A prática comum de reportar *C-index* ou AUC isoladamente induz ao erro quando decisões clínicas dependem de probabilidades calibradas.
5. **Quantificar o valor incremental sobre preditores baratos.** Antes de propor uma assinatura gênica como útil, demonstrar que ela acrescenta informação sobre variáveis clínicas já disponíveis — idade, T, N, grau, RE, HER2 — é, hoje, requisito mínimo.

5.11 O caminho até a clínica: o que falta?

Para que o CorePAM possa ser utilizado como ferramenta de decisão clínica, uma série de passos adicionais é necessária:

1. **Validação analítica plataforma-específica.** Desenvolvimento de ensaio de RT-qPCR multiplex ou painel de *RNA-seq* direcionado, com genes housekeeping para normalização robusta, e caracterização analítica rigorosa (precisão, exatidão, limite de detecção, robustez a variações pré-analíticas). É o passo técnico obrigatório antes de qualquer uso clínico.
2. **Validação analítica em tecido FFPE.** Todos os estudos atuais de CorePAM utilizam tecido congelado ou matrizes públicas processadas. A transição para *FFPE* — material quase universalmente disponível na rotina clínica — é não trivial, pois a degradação de RNA em *FFPE* afeta transcritos de formas heterogêneas. O *OncotypeDX* e o *Prosigna* foram desenvolvidos diretamente em *FFPE*; o CorePAM não.
3. **Validação clínica prospectiva.** Idealmente um estudo randomizado de decisão terapêutica (como TAILORx e MINDACT), mas minimamente um estudo observacional prospectivo em população brasileira com desfechos sob seguimento.
4. **Avaliação regulatória.** Tanto ANVISA (Brasil) quanto FDA (EUA) requerem documentação específica para aprovação de *in vitro diagnostics*.
5. **Incorporação em diretrizes.** Avaliação pela Conitec para incorporação ao SUS, e por diretrizes internacionais (ASCO, ESMO, St. Gallen).

Este é um caminho de anos, não de meses, e requer investimento substancial. Mas nenhum desses passos é impossível, e todos seguem modelo já trilhado por assinaturas concorrentes.

5.12 Relevância para o contexto brasileiro e o SUS

O Brasil apresenta contexto epidemiológico e de sistema de saúde peculiar para a medicina de precisão em câncer de mama. O volume de novos casos (~74 mil/ano⁵) e o perfil de estadiamento ao diagnóstico (maior frequência de estágios II-III em relação a países com rastreamento mamográfico universalizado⁶) ampliam a demanda potencial por ferramentas que orientem decisão terapêutica, em particular a decisão de quimioterapia adjuvante em doença RE+/HER2- estágio precoce.

Um ensaio baseado em 24 genes, executável em plataformas já presentes em laboratórios brasileiros (RT-qPCR multiplex em tempo real com sondas TaqMan ou painel de *RNA-seq* direcionado de baixo throughput), representaria alternativa com custo potencialmente substancialmente menor — provavelmente abaixo de R\$ 1 500-2 000 por amostra em cenários de escala — e, crucialmente, com execução local, sem dependência de logística internacional. Entretanto, a viabilidade econômica real só pode ser demonstrada após validação analítica e desenvolvimento do ensaio, que exigem investimento inicial.

Em termos de política pública, a hipótese testável é que um ensaio derivado como o CorePAM, com validação local em hospital universitário público (HUB/UnB ou rede

análoga), poderia servir como instrumento de calibração local de assinatura gênica com custo compatível com a realidade do SUS. Esta hipótese é a motivação de longo prazo por trás do presente trabalho e sua validação requer estudo subsequente específico de custo-efetividade no contexto brasileiro.

5.13 Limitações

O estudo tem limitações que devem ser consideradas na interpretação dos achados:

1. **Estudo retrospectivo.** Todas as análises foram conduzidas em coortes já coletadas, com dados clínicos e moleculares existentes. A validação prospectiva em população contemporaneamente coletada não foi realizada.
2. **Populações predominantemente europeias.** SCAN-B (sueca), TCGA-BRCA (maioria americana, predominantemente de descendência europeia), METABRIC (britânica, canadense), GSE1456 (sueca). GSE20685 (Taiwan) é a única coorte com predominância asiática. Nenhuma coorte brasileira ou latino-americana foi testada. Desempenho em populações com composição genética e prevalência de subtipos distintas é desconhecido.
3. **Seguimento curto em TCGA-BRCA** (mediana 32 meses), que limita o poder para detectar eventos tardios em doença RE-positiva (cuja recorrência predomina após 5-10 anos).
4. **Endpoint heterogêneo entre coortes.** OS em SCAN-B, TCGA-BRCA, GSE20685 e GSE1456; DSS em METABRIC. A análise de sensibilidade OS-harmonizada em METABRIC produziu resultados convergentes, mas a heterogeneidade é uma limitação conceitual.
5. **Covariáveis clínicas incompletas ou heterogêneas.** Nem todas as coortes têm informação de grau histológico, T, N, HER2 em formatos harmonizáveis. Os modelos multivariados foram ajustados com os subconjuntos disponíveis.
6. **Comparações com ROR-S e OncotypeDX via genefu.** As implementações do pacote genefu são aproximações de pesquisa, não os ensaios clínicos comerciais. Comparação formal contra *Prosigna* e *OncotypeDX* em plataformas comerciais seria necessária para reivindicar superioridade.
7. **Nenhuma validação analítica em plataforma de implantação.** RT-qPCR multiplex em tempo real, ou painel direcionado de *RNA-seq*, e especialmente *FFPE* — não foram testados. O desempenho em cenário analítico real é desconhecido.
8. **Calibração externa imperfeita.** Como discutido, slopes de calibração externa sugerem que recalibração seria necessária para produção de probabilidades de evento.
9. **Avaliação exploratória de pCR.** Embora o resultado seja estatisticamente sólido, o *endpoint* de pCR não era o foco de derivação e os tamanhos amostrais individuais em algumas coortes neoadjuvantes são modestos.

5.14 Perspectivas e estudos futuros

As perspectivas de trabalho subsequente estruturam-se em seis frentes principais:

1. **Validação analítica em RT-qPCR com 24 genes + housekeeping em FFPE.** Etapa obrigatória de qualquer caminho para implantação clínica.
2. **Validação em coortes brasileiras.** Idealmente a partir de hospitais universitários públicos (Hospital Universitário de Brasília/UnB, Hospital das Clínicas/USP, A.C. Camargo, INCA/RJ, Barretos), com tecido biobanked e desfechos prospectivamente coletados. Esta é, talvez, a mais urgente lacuna a ser preenchida.
3. **Comparação formal com OncotypeDX em coortes pareadas.** Pacientes com ambos os ensaios — *OncotypeDX* (comercial) e CorePAM (investigacional) — sobre mesma amostra *FFPE*, com desfechos prospectivos. Permitiria estabelecer concordância ou discordância de decisão terapêutica entre as duas ferramentas.

4. **Nomograma integrando CorePAM + variáveis clínico-patológicas.** A maioria das diretrizes atuais integra assinatura gênica com variáveis clínicas; o CorePAM isolado é provavelmente subótimo para decisão, e um nomograma seria ferramenta natural.
5. **Estudo de custo-efetividade no contexto do SUS.** Análise formal comparando: (a) estratégia clínica clássica, (b) *OncotypeDX* via SUS (quando disponível), (c) CorePAM hipotético localmente validado. Resultados em custo incremental por QALY seriam fundamentais para pleitear incorporação pela Conitec.
6. **Validação prospectiva em estudo de decisão terapêutica.** Analogamente a TAILORx e MINDACT, mas em escala brasileira e com desenho adequado à realidade do SUS.

Horizontes mais distantes incluem a adaptação do CorePAM para plataformas de terceira geração (sequenciamento de RNA de célula única e sequenciamento Nanopore direto de RNA) e a integração com inteligência artificial multimodal combinando patologia digital (imagens histológicas processadas por deep learning) com assinatura gênica reduzida — cenário que pode produzir ferramentas prognósticas ainda mais robustas e acessíveis.

6. CONCLUSÃO

O CorePAM é uma redução reprodutível e orientada por dados de 50 para 24 genes do painel PAM50, selecionada dentro de um *framework* de não inferioridade pré-especificado.

Os principais achados são:

1. A regressão de Cox *elastic-net* com validação cruzada determinística identificou 24 genes como o menor subconjunto satisfazendo $\Delta C\text{-index} \leq 0,010$ em relação ao modelo de 50 genes.
2. O escore foi significativamente associado a sobrevida em todas as quatro coortes de validação independentes, com HR agrupado de 1,37 (IC 95% 1,24-1,52; $p = 1,6 \times 10^{-9}$).
3. O valor incremental sobre preditores clínicos foi consistente, e o escore permaneceu significativo após ajuste para seis variáveis clínico-patológicas.
4. A associação com pCR em coortes neoadjuvantes independentes (OR = 1,69; I-quadrado = 0%) confirma que a assinatura captura informação biológica que transcende o *endpoint* de derivação.
5. Em comparações externas pareadas, o modelo de 24 genes igualou ou superou o comparador PAM50-full penalizado com 49 coeficientes não-zero em discriminação.

A redução de 50 para 24 genes preserva os quatro eixos biológicos do PAM50 e pode simplificar o desenvolvimento futuro de ensaios prognósticos. A validação analítica plataforma-específica e a validação clínica prospectiva são os próximos passos necessários para qualquer aplicação clínica.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, Freddie Bray. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide. *CA Cancer J Clin.* 2021;71(3):209–49. <https://doi.org/10.3322/caac.21660>
2. Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Isabelle Soerjomataram, Ahmedin Jemal. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2024;74(3):229–63. <https://doi.org/10.3322/caac.21834>
3. Richard Peto, Christina Davies, Jon Godwin, others. Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100,000 women in 123 randomised trials. *Lancet.* 2012;379(9814):432–44. [https://doi.org/10.1016/S0140-6736\(11\)61625-5](https://doi.org/10.1016/S0140-6736(11)61625-5)
4. Early Breast Cancer Trialists' Collaborative Group. Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. *Lancet.* 2011;378(9793):771–84. [https://doi.org/10.1016/S0140-6736\(11\)60993-8](https://doi.org/10.1016/S0140-6736(11)60993-8)
5. Instituto Nacional de Câncer. Estimativa 2023: incidência de câncer no Brasil. INCA. 2022.
6. Arn Migowski, Gulnar Azevedo e Silva, Mario Borges Rosa Dias, Maria do Desterro Paiva Diz, Denise Rangel Sant'Ana, Paulo Nadanovsky. Diretrizes para detecção precoce do câncer de mama no Brasil. *Cad Saúde Pública.* 2018;34(6):e00074817. <https://doi.org/10.1590/0102-311X00074817>
7. Nadia Harbeck, Frederique Penault-Llorca, Javier Cortes, Michael Gnant, Nehmat Houssami, Philip Poortmans, Kathryn Ruddy, Janice Tsang, Fatima Cardoso. Breast cancer. *Nat Rev Dis Primers.* 2019;5(1):66. <https://doi.org/10.1038/s41572-019-0111-2>
8. Charles M Perou, Therese Sorlie, Michael B Eisen, others. Molecular portraits of human breast tumours. *Nature.* 2000;406(6797):747–52. <https://doi.org/10.1038/35021093>

9. Therese Sorlie, Charles M Perou, Robert Tibshirani, others. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA*. 2001;98(19):10869–74. <https://doi.org/10.1073/pnas.191367098>
10. Aleix Prat, Charles M Perou. Deconstructing the molecular portraits of breast cancer. *Mol Oncol*. 2011;5(1):5–23. <https://doi.org/10.1016/j.molonc.2010.11.003>
11. Christina Curtis, Sohrab P Shah, Suet-Feung Chin, others. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. 2012;486(7403):346–52. <https://doi.org/10.1038/nature10983>
12. Bernard Pereira, Suet-Feung Chin, Oscar M Rueda, others. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat Commun*. 2016;7:11479. <https://doi.org/10.1038/ncomms11479>
13. Joel S Parker, Michael Mullins, Maggie C Cheang, others. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009;27(8):1160–7. <https://doi.org/10.1200/JCO.2008.18.1370>
14. Torsten O Nielsen, Joel S Parker, Samuel Leung, David Voduc, Mark Ebbert, Tammi Vickery, Sherri R Davies, Jacqueline Snider, Inge J Stijleman, Jerry Reed, Maggie C U Cheang, Elaine R Mardis, Charles M Perou, Philip S Bernard, Matthew J Ellis. A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clin Cancer Res*. 2010;16(21):5222–32. <https://doi.org/10.1158/1078-0432.CCR-10-1282>
15. Stephen K Chia, Vivien H Bramwell, Dongsheng Tu, others. A 50-gene intrinsic subtype classifier for prognosis and prediction of benefit from adjuvant tamoxifen. *Clin Cancer Res*. 2012;18(16):4465–72. <https://doi.org/10.1158/1078-0432.CCR-12-0286>
16. Brett Wallden, James Storhoff, Torsten Nielsen, others. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Med Genomics*. 2015;8:54. <https://doi.org/10.1186/s12920-015-0129-6>
17. Mitch Dowsett, Jack Cuzick, Christopher Wale, John Forbes, Elizabeth A Mallon, Janine Salter, Emma Quinn, Anita Dunbier, Michael Baum, Aman Buzdar, Anthony Howell, Roberto Bugarini, Frederick L Baehner, Steven Shak. Prediction of risk of distant recurrence using the 21-gene recurrence score in node-negative and node-positive postmenopausal patients with breast cancer treated with anastrozole or tamoxifen: a

TransATAC study. *J Clin Oncol.* 2010;28(11):1829–34. <https://doi.org/10.1200/JCO.2009.24.4798>

18. Michael Gnant, Martin Filipits, Richard Greil, others. Predicting distant recurrence in receptor-positive breast cancer patients with limited clinicopathological risk: using the PAM50 risk of recurrence score in 1478 postmenopausal patients of the ABCSG-8 trial. *Ann Oncol.* 2014;25(2):339–45. <https://doi.org/10.1093/annonc/mdt494>

19. Alan S Coates, Eric P Winer, Aron Goldhirsch, others. Tailoring therapies—improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. *Ann Oncol.* 2015;26(8):1533–46. <https://doi.org/10.1093/annonc/mdv221>

20. Harold J Burstein, Giuseppe Curigliano, Sibylle Loibl, others. Estimating the benefits of therapy for early-stage breast cancer: the St Gallen International Consensus Guidelines for the primary therapy of early breast cancer 2019. *Ann Oncol.* 2019;30(10):1541–57. <https://doi.org/10.1093/annonc/mdz235>

21. Fabrice Andre, Nofisat Ismaila, Kimberly H Allison, others. Biomarkers for adjuvant endocrine and chemotherapy in early-stage breast cancer: ASCO guideline update. *J Clin Oncol.* 2022;40(16):1816–37. <https://doi.org/10.1200/JCO.22.00069>

22. Soonmyung Paik, Steven Shak, Gong Tang, others. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med.* 2004;351(27):2817–26. <https://doi.org/10.1056/NEJMoa041588>

23. Joseph A Sparano, Robert J Gray, Della F Makower, others. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *N Engl J Med.* 2018;379(2):111–21. <https://doi.org/10.1056/NEJMoa1804710>

24. Joseph A Sparano, Robert J Gray, Peter M Ravdin, others. Clinical and genomic risk to guide the use of adjuvant therapy for breast cancer. *N Engl J Med.* 2019;380(25):2395–405. <https://doi.org/10.1056/NEJMoa1904819>

25. Kevin Kalinsky, William E Barlow, Julie R Gralow, others. 21-gene assay to inform chemotherapy benefit in node-positive breast cancer. *N Engl J Med.* 2021;385(25):2336–47. <https://doi.org/10.1056/NEJMoa2108873>

26. Laura J van't Veer, Hongyue Dai, Marc J van de Vijver, others. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415(6871):530–6. <https://doi.org/10.1038/415530aX>
27. Fatima Cardoso, Laura J van't Veer, Jan Bogaerts, others. 70-gene signature as an aid to treatment decisions in early-stage breast cancer. *N Engl J Med*. 2016;375(8):717–29. <https://doi.org/10.1056/NEJMoa1602253>
28. Martine Piccart, Laura J van't Veer, Coralie Poncet, others. 70-gene signature as an aid for treatment decisions in early breast cancer: updated results of the phase 3 randomised MINDACT trial with an exploratory analysis by age. *Lancet Oncol*. 2021;22(4):476–88. [https://doi.org/10.1016/S1470-2045\(21\)00007-3](https://doi.org/10.1016/S1470-2045(21)00007-3)
29. Martin Filipits, Margaretha Rudas, Raimund Jakesz, others. 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(18):6012–20. <https://doi.org/10.1158/1078-0432.CCR-11-0926>
30. Yi Zhang, Catherine A Schnabel, Brock E Schroeder, others. Breast Cancer Index identifies early-stage estrogen receptor-positive breast cancer patients at risk for early- and late-distant recurrence. *Clin Cancer Res*. 2013;19(15):4196–205. <https://doi.org/10.1158/1078-0432.CCR-13-0804>
31. Ivana Sestak, Richard Buus, Jack Cuzick, others. Comparison of the performance of 6 prognostic signatures for estrogen receptor-positive breast cancer. *JAMA Oncol*. 2018;4(4):545–53. <https://doi.org/10.1001/jamaoncol.2017.5524>
32. Michael I Love, Wolfgang Huber, Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. <https://doi.org/10.1186/s13059-014-0550-8>
33. Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47. <https://doi.org/10.1093/nar/gkv007>
34. MAQC Consortium. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol*. 2010;28(8):827–38. <https://doi.org/10.1038/nbt.1665>

35. W Evan Johnson, Cheng Li, Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118–27. <https://doi.org/10.1093/biostatistics/kxj037>
36. Jeffrey T Leek, Robert B Scharpf, Hector Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*. 2010;11(10):733–9. <https://doi.org/10.1038/nrg2825>
37. Lao H Saal, Johan Vallon-Christersson, Jari Hakkinen, others. The Sweden Cancerome Analysis Network—Breast (SCAN-B) Initiative. *Genome Med*. 2015;7(1):20. <https://doi.org/10.1186/s13073-015-0131-9>
38. Christian Brueffer, Johan Vallon-Christersson, Dorte Grabau, others. Clinical value of RNA sequencing-based classifiers for prediction of the five conventional breast cancer biomarkers. *JCO Precis Oncol*. 2018;2:PO.17.00135. <https://doi.org/10.1200/PO.17.00135>
39. TCGA Network. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;490(7418):61–70. <https://doi.org/10.1038/nature11412>
40. Kuo-Jang Kao, Kai-Ming Chang, Hsueh-Chih Hsu, Ai Tao Huang. Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes. *BMC Cancer*. 2011;11:143. <https://doi.org/10.1186/1471-2407-11-143>
41. Yudi Pawitan, Jonas Bjohle, Lukas Amler, others. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res*. 2005;7(6):R953–64. <https://doi.org/10.1186/bcr1325>
42. Christos Hatzis, Lajos Pusztai, Vicente Valero, others. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA*. 2011;305(18):1873–81. <https://doi.org/10.1001/jama.2011.593>
43. Tetsuya Miyake, Takahiro Nakayama, Yasuto Naoi, others. GSTP1 expression predicts poor pathological complete response to neoadjuvant chemotherapy in ER-negative breast cancer. *Cancer Sci*. 2012;103(5):913–20. <https://doi.org/10.1111/j.1349-7006.2012.02241.x>

44. Laura J Esserman, Donald A Berry, Maggie C Cheang, others. Chemotherapy response and recurrence-free survival in neoadjuvant breast cancer depends on biomarker profiles: results from the I-SPY 1 TRIAL (CALGB 150007/150012; ACRIN 6657). *Breast Cancer Res Treat.* 2012;132(3):1049–62. <https://doi.org/10.1007/s10549-011-1895-2>
45. Denise M Wolf, Christina Yau, Julia Wulfschlegel, others. Redefining breast cancer subtypes to guide treatment prioritization and maximize response: predictive biomarkers across 10 cancer therapies. *Cancer Cell.* 2022;40(6):609–623.e6. <https://doi.org/10.1016/j.ccell.2022.05.005>
46. Karel G Moons, Douglas G Altman, Johannes B Reitsma, others. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). *Ann Intern Med.* 2015;162(1):W1–73. <https://doi.org/10.7326/M14-0698>
47. Gary S Collins, Johannes B Reitsma, Douglas G Altman, Karel G M Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ.* 2015;350:g7594. <https://doi.org/10.1136/bmj.g7594>
48. Patrick Royston, Douglas G Altman. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol.* 2013;13:33. <https://doi.org/10.1186/1471-2288-13-33>
49. Ewout W Steyerberg, Frank E Jr Harrell. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol.* 2016;69:245–7. <https://doi.org/10.1016/j.jclinepi.2015.04.005>
50. Mark D Robinson, Davis J McCarthy, Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010;26(1):139–40. <https://doi.org/10.1093/bioinformatics/btp616>
51. Mark D Robinson, Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 2010;11(3):R25. <https://doi.org/10.1186/gb-2010-11-3-r25>
52. Hui Zou, Trevor Hastie. Regularization and variable selection via the elastic net. *J R Stat Soc B.* 2005;67(2):301–20. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>
53. Robert Tibshirani. Regression shrinkage and selection via the lasso. *J R Stat Soc B.* 1996;58(1):267–88. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>

54. Robert Tibshirani. The lasso method for variable selection in the Cox model. *Stat Med.* 1997;16(4):385–95. [https://doi.org/10.1002/\(SICI\)1097-0258\(19970228\)16:4<385::AID-SIM380>3.0.CO;2-3](https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3)
55. Jerome Friedman, Trevor Hastie, Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw.* 2010;33(1):1–22. <https://doi.org/10.18637/jss.v033.i01>
56. Noah Simon, Jerome Friedman, Trevor Hastie, Robert Tibshirani. Regularization paths for Cox’s proportional hazards model via coordinate descent. *J Stat Softw.* 2011;39(5):1–13. <https://doi.org/10.18637/jss.v039.i05>
57. Patricia M Grambsch, Terry M Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika.* 1994;81(3):515–26. <https://doi.org/10.1093/biomet/81.3.515>
58. Terry M Therneau, Patricia M Grambsch. *Modeling Survival Data: Extending the Cox Model.* New York: Springer; 2000. <https://doi.org/10.1007/978-1-4757-3294-8>
59. Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, Robert A Rosati. Evaluating the yield of medical tests. *JAMA.* 1982;247(18):2543–6. <https://doi.org/10.1001/jama.1982.03320430047030>
60. Bradley Efron. Bootstrap methods: another look at the jackknife. *Ann Stat.* 1979;7(1):1–26. <https://doi.org/10.1214/aos/1176344552>
61. Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D’Agostino, Lee-Jen Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med.* 2011;30(10):1105–17. <https://doi.org/10.1002/sim.4154>
62. Paul Blanche, Jean-Francois Dartigues, Helene Jacqmin-Gadda. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat Med.* 2013;32(30):5381–97. <https://doi.org/10.1002/sim.5958>
63. Jason P Fine, Robert J Gray. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc.* 1999;94(446):496–509. <https://doi.org/10.1080/01621459.1999.10474144>
64. Ben Van Calster, Daan Nieboer, Yvonne Vergouwe, Bavo De Cock, Michael J Pencina, Ewout W Steyerberg. A calibration hierarchy for risk models was defined: from

utopia to empirical data. *J Clin Epidemiol.* 2016;74:167–76.
<https://doi.org/10.1016/j.jclinepi.2015.12.005>

65. Ben Van Calster, David J McLernon, Maarten van Smeden, Laure Wynants, Ewout W Steyerberg. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 2019;17(1):230. <https://doi.org/10.1186/s12916-019-1466-7>

66. Andrew J Vickers, Elena B Elkin. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making.* 2006;26(6):565–74.
<https://doi.org/10.1177/0272989X06295361>

67. Wolfgang Viechtbauer. Conducting meta-analyses in R with the metafor package. *J Stat Softw.* 2010;36(3):1–48. <https://doi.org/10.18637/jss.v036.i03>

68. Julian P Higgins, Simon G Thompson, Jonathan J Deeks, Douglas G Altman. Measuring inconsistency in meta-analyses. *BMJ.* 2003;327(7414):557–60.
<https://doi.org/10.1136/bmj.327.7414.557>

69. Joachim Hartung, Guido Knapp. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Stat Med.* 2001;20(24):3875–89.
<https://doi.org/10.1002/sim.1009>

70. Rebecca DerSimonian, Nan Laird. Meta-analysis in clinical trials. *Control Clin Trials.* 1986;7(3):177–88. [https://doi.org/10.1016/0197-2456\(86\)90046-2](https://doi.org/10.1016/0197-2456(86)90046-2)

71. Deena M Gendoo, Natchar Ratanasirigulchai, Markus S Schroder, others. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics.* 2016;32(7):1097–9.
<https://doi.org/10.1093/bioinformatics/btv693>

72. Maggie C U Cheang, Stephen K Chia, David Voduc, others. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *J Natl Cancer Inst.* 2009;101(10):736–50. <https://doi.org/10.1093/jnci/djp082>

73. Carsten Denkert, Gunter von Minckwitz, Silvia Darb-Esfahani, others. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. *Lancet Oncol.* 2018;19(1):40–50. [https://doi.org/10.1016/S1470-2045\(17\)30904-XX](https://doi.org/10.1016/S1470-2045(17)30904-XX)

74. Patricia Cortazar, Liang Zhang, Michael Untch, others. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. *Lancet*. 2014;384(9938):164–72. [https://doi.org/10.1016/S0140-6736\(13\)62422-8](https://doi.org/10.1016/S0140-6736(13)62422-8)
75. Aleix Prat, Cheng Fan, Aranzazu Fernandez, others. Response and survival of breast cancer intrinsic subtypes following multi-agent neoadjuvant chemotherapy. *BMC Med*. 2015;13:303. <https://doi.org/10.1186/s12916-015-0540-zX>
76. Botan, RN, Sousa, JB. CorePAM: a 24-gene PAM50-derived expression score with cross-platform external validation for breast cancer prognosis. *Breast Cancer Res*. 2026. <https://doi.org/10.1186/s13058-026-02298-5>

APÊNDICE A — Pipeline computacional

O *pipeline* analítico completo está disponível em <https://github.com/RafaelBotan/CorePAM-Study> e compreende 69 scripts R numerados sequencialmente. A numeração é parte da rastreabilidade: scripts com prefixo baixo preparam dados e congelam parâmetros; scripts intermediários executam derivação, validação e figuras; scripts finais documentam análises de sensibilidade e respostas às revisões. A organização por blocos temáticos é apresentada na Tabela 20; a lista integral, com descrição objetiva de cada script, segue na Tabela 21.

Para o leitor médico, um **script** pode ser entendido como uma prescrição executável da análise. Em vez de dizer “fazer Cox, calcular *C-index* e gerar figura”, o script contém as instruções exatas que o computador executa para fazer isso sempre do mesmo modo. Ele é semelhante a um protocolo assistencial escrito em detalhes: define entrada, sequência, critérios de exclusão, cálculos, checagens e produto final. A diferença é que, no caso do script, a “equipe” que executa é o computador, sem improviso.

Um **pipeline** é a sequência organizada desses scripts. A lógica é parecida com uma linha de cuidado: primeiro entram os dados brutos, depois vêm limpeza e harmonização, depois cálculo do score, análise estatística, figuras, tabelas e verificações finais. Cada script deixa um produto rastreável para o próximo. Se uma tabela da tese mostra um HR, um *C-index* ou uma curva de Kaplan-Meier, deve existir um script que gerou aquele número ou aquela figura. Essa rastreabilidade é importante porque transforma o resultado em algo auditável: outro pesquisador pode baixar os dados públicos, executar os scripts e verificar se chega ao mesmo resultado.

O **pseudocódigo** apresentado ao final deste apêndice tem outra função. Ele não é para ser executado; é uma tradução do raciocínio dos scripts para linguagem humana, independente de R, Python ou qualquer software. Serve como mapa para entender a análise sem precisar ler código. Em termos clínicos, o pseudocódigo está para o script assim como a descrição de uma técnica cirúrgica está para o ato operatório: não substitui a execução real, mas explica as etapas, a ordem e o motivo de cada passo.

Para facilitar a leitura impressa, os nomes dos arquivos são mantidos exatamente como no repositório, mas a interpretação deve se concentrar no prefixo numérico e na função

descrita na segunda coluna. O prefixo indica a ordem de execução; a descrição indica o papel científico do script.

Tabela 20 — Organização do *pipeline* computacional por blocos

Bloco	Scripts	Função
Setup	00	Configuração do ambiente, paleta de cores, validação
Dados	01-04	<i>Download</i> , harmonização clínica, pré-processamento
Derivação	05	Derivação do CorePAM (CONGELADO)
Scoring	06	<i>Z-score</i> e cálculo do escore por coorte
Sobrevida	07-08	Análise de sobrevida + meta-análise
Valor incremental	11	Delta- <i>C-index</i> e DCA
QC	13-16	Correlações, PCA, <i>schema</i> , anti-hardcoding
Render	17-18	Renderização do manuscrito e <i>bundle</i> de submissão
pCR	19-23b	<i>Download</i> , preparação, análise e figuras de pCR
4ª coorte	24	Integração GSE1456 (Stockholm)
Sensibilidade	25-29	CORE-A expandido, <i>genefu</i> , <i>bootstrap</i> , <i>drop-gene</i>
Revisão major	30-36	<i>Z-score</i> congelado, Uno C, calibração, ER, <i>baseline</i>
R2 (revisão)	37-38	KM <i>multipanel</i> e <i>head-to-head</i> 24 vs 50

Tabela 21 — Lista completa dos scripts do *pipeline*

Script	Descrição
00_colors.R	Paleta de cores reprodutível para figuras
00_setup.R	Parâmetros congelados (alpha, K, margem, lambda grid)
00_validate_all.R	Validação estrutural de entradas e saídas
01_download_raw_data.R	<i>Download</i> das matrizes de SCAN-B, TCGA, METABRIC, GEO
02_harmonize_clinical_GSE20685.R	Harmonização clínica GSE20685
02_harmonize_clinical_METABRIC.R	Harmonização clínica METABRIC
02_harmonize_clinical_SCANB.R	Harmonização clínica SCAN-B
02_harmonize_clinical_TCGA_BRCA.R	Harmonização clínica TCGA-BRCA

Script	Descrição
03_expression_preprocess_GSE20685.R	Pré-processamento de expressão GSE20685
03_expression_preprocess_METABRIC.R	Pré-processamento de expressão METABRIC
03_expression_preprocess_SCANB.R	Pré-processamento de expressão SCAN-B
03_expression_preprocess_TCGA_BRCA.R	Pré-processamento de expressão TCGA-BRCA
03_utils_gene_mapping.R	Utilitários de mapeamento ENSEMBL → HUGO
04_gene_audit_freeze.R	Auditoria e congelamento dos 50 genes candidatos
05_reduce_pam50_to_corepam_FINAL.R	Derivação final do CorePAM (<i>elastic-net</i> CV)
06_zscore_and_score_GSE20685.R	<i>Z-score</i> e escore GSE20685
06_zscore_and_score_METABRIC.R	<i>Z-score</i> e escore METABRIC
06_zscore_and_score_SCANB.R	<i>Z-score</i> e escore SCAN-B
06_zscore_and_score_TCGA_BRCA.R	<i>Z-score</i> e escore TCGA-BRCA
07A_preflight_files_strict.R	<i>Preflight</i> de arquivos obrigatórios
07D_validate_one_cohort_corepam.R	Validação unitária por coorte
07_survival_analysis_GSE20685.R	Sobrevida GSE20685 (Cox + KM + C-index)
07_survival_analysis_METABRIC.R	Sobrevida METABRIC (DSS + Fine-Gray)
07_survival_analysis_SCANB.R	Sobrevida SCAN-B (coorte de treinamento)
07_survival_analysis_TCGA_BRCA.R	Sobrevida TCGA-BRCA
07x_extra_figures.R	Figuras suplementares: Pareto, <i>lollipop</i> , <i>C-index</i> por coorte
07x_pcr_os_extra_figures.R	Figuras suplementares pCR × OS
07y_pam50full_comparison.R	Comparação PAM50-full vs CorePAM
07z_fig1_study_design.R	Figura 1 — desenho do estudo
07z_table_figures_survival.R	Tabelas e figuras consolidadas de sobrevida
08_meta_survival.R	Meta-análise de sobrevida (REML + Hartung-Knapp)
11_incremental_value_and_dca.R	Delta- <i>C-index</i> e DCA (curva de decisão)
11b_dca_corepam.R	DCA auxiliar do CorePAM
13_qc_correlations_offdiag.R	QC — correlações inter-coortes
14_qc_metablic_pca_forensics.R	QC — PCA forensics METABRIC

Script	Descrição
15_qc_schema_range_checks.R	QC — consistência de <i>schema</i> e faixas
16_qc_text_vs_results_assert.R	QC anti-hardcoding (números do texto vs saída)
17_render_manuscript_quarto.R	Renderização do manuscrito (Quarto)
18_make_submission_bundle.R	<i>Bundle</i> de submissão (arquivos + hashes)
19_download_pCR_raw_data.R	<i>Download</i> das coortes neoadjuvantes (pCR)
20_prepare_pCR_GSE20194.R	Preparação GSE20194
20_prepare_pCR_GSE25066.R	Preparação GSE25066
20_prepare_pCR_GSE32646.R	Preparação GSE32646
20_prepare_pCR_ISPY1.R	Preparação I-SPY1
20_prepare_pCR_ISPY2.R	Preparação I-SPY2
20_utils_pcr_extract.R	Utilitários de extração de pCR
21_ismy2_analysis.R	Análise exploratória I-SPY2
21_pCR_logistic_analysis.R	Regressão logística de pCR por coorte
21b_sensitivity_intersection.R	Sensibilidade sobre interseção de genes
21c_pcr_er_interaction.R	Interação pCR × ER
22_meta_pCR.R	Meta-análise de pCR (DerSimonian-Laird)
22b_meta_pCR_with_ismy2.R	Meta-análise de pCR incluindo I-SPY2
23_pCR_figures.R	Figuras de pCR (forest, painéis)
23b_pCR_ismy2_figures.R	Figuras de I-SPY2
24_integrate_GSE1456_stockholm.R	Integração da 4ª coorte (Stockholm)
25_sensitivity_corea_expanded.R	CORE-A expandido (+ T, N, grau, HER2)
26_sensitivity_genefu_comparison.R	Comparação com ROR-S e OncotypeDX (<i>genefu</i>)
27_bootstrap_gene_stability.R	<i>Bootstrap</i> B = 200 (estabilidade de seleção)
28_dropgene_sensitivity.R	<i>Leave-one-gene-out</i> (
29_corea_sensitivity_forest.R	<i>Forest plot</i> de sensibilidade CORE-A
30_frozen_zscore_sensitivity.R	<i>Single-sample scoring</i> com parâmetros congelados
30b_add_GSE1456_frozen.R	Extensão congelada para GSE1456
31_uno_c_and_tdauc.R	C de Uno e AUC tempo-dependente (Blanche)
32_calibration_formal.R	Calibração formal (<i>slope/intercept</i> 60 m)
34_subgroup_er_stratified.R	Subgrupos estratificados por ER
35_er_neg_bootstrap_stability.R	Estabilidade <i>bootstrap</i> em RE-negativo

Script	Descrição
36_full_clinical_baseline.R	<i>Baseline</i> clínico completo (36 variáveis)
37_R2_km_multipanel.R	KM <i>multipanel</i> (resposta à R2)
38_R2_headtohead_24vs50.R	<i>Head-to-head</i> 24 vs 50 genes (resposta à R2)

B.1 Pseudocódigo do procedimento de derivação

O pseudocódigo a seguir descreve, em linguagem algorítmica independente de implementação, o procedimento central de derivação do CorePAM. A correspondência com os scripts R é indicada entre parênteses ao lado de cada bloco. A leitura recomendada não é tentar memorizar símbolos, mas seguir a pergunta clínica de cada algoritmo: como o escore foi criado, como ele é aplicado a uma nova coorte, como se garantiu reprodutibilidade e como os resultados foram combinados.

Algoritmo 1 — Derivação do CorePAM *(script 05_reduce_pam50_to_corepam_FINAL.R)*

Este algoritmo é a etapa em que o CorePAM nasce. Ele parte da SCAN-B, que é a coorte de treinamento, e pergunta quais genes do PAM50 realmente ajudam a ordenar risco de sobrevida. A validação cruzada funciona como um ensaio interno de honestidade: cada paciente é predita por um modelo que não a viu. A seleção final não escolhe o modelo “mais bonito”, mas o menor conjunto que fica dentro da margem de desempenho aceitável.

Entrada:

X : matriz de expressão SCAN-B (N pacientes x 50 genes candidatos)

T : vetor de tempos de seguimento (N)

E : vetor de status de evento 0/1 (N)

ids : vetor de identificadores únicos dos pacientes (N)

alpha : parâmetro elastic-net (0,5)

K : número de folds (10)

lambda_grid : grade de 100 valores em escala log

delta_tol : margem de não inferioridade (0,010)

1. Pré-processamento

Para cada gene $j = 1..50$:

$$z[i,j] = (X[i,j] - \text{media_coorte}(j)) / \text{dp_coorte}(j) \text{ para } i = 1..N$$

2. Atribuição determinística de folds

Para cada paciente $i = 1..N$:

$$h = \text{SHA256}(\text{ids}[i])$$

$$H = \text{inteiro_primeiros_bytes}(h)$$

$$\text{fold}[i] = H \bmod K$$

Aplicar estratificação por E dentro de cada fold.

3. Caminho de regularização com validação cruzada

Para cada lambda em lambda_grid:

Para cada $k = 1..K$:

treino = índices com fold $\neq k$

teste = índices com fold $= k$

```

beta_k = ajuste_cox_elastic_net(z[treino], T[treino], E[treino],
                                alpha, lambda)
eta[teste] = z[teste] %*% beta_k # predição OOF
df[lambda] = número de coeficientes não-zero em beta médio
c_oof[lambda] = C_index_Harrell(eta, T, E)

```

4. Fronteira de Pareto e seleção

```

pareto = remove_pontos_dominados((df[lambda], c_oof[lambda]))
c_max = max(c_oof[lambda])
candidatos = { df em pareto tais que c_oof(df) >= c_max - delta_tol }
df_final = min(candidatos)
lambda_final = arg min lambda tal que df(lambda) = df_final

```

5. Ajuste final e congelamento

```

beta_final = ajuste_cox_elastic_net(z, T, E, alpha, lambda_final)
genes_CorePAM = { j : beta_final[j] != 0 } # |genes_CorePAM| = 24
salvar beta_final, media_coorte, dp_coorte # congelados

```

Saída:

```

genes_CorePAM (24 genes)
pesos beta_final
parâmetros de normalização (media_coorte, dp_coorte)

```

Algoritmo 2 — Cálculo do escore em coorte de validação (*script 06_zscore_and_score.R*)

Este algoritmo descreve como o CorePAM é aplicado depois de pronto. A coorte externa não escolhe novos genes e não muda os pesos; ela apenas recebe o escore congelado. O ponto mais importante é o *z-score*: cada gene é colocado em escala relativa dentro da própria coorte, permitindo comparar plataformas que medem expressão em unidades diferentes. Quando algum gene não está presente, o denominador é recalculado com os pesos disponíveis, evitando imputação artificial.

Entrada:

X_v : matriz de expressão da coorte de validação ($M \times 24$ genes)

genes_CorePAM : subconjunto de 24 genes

beta_final : pesos congelados do SCAN-B

1. Z-score intra-coorte

Para cada gene j em genes_CorePAM:

Se j presente em X_v :

$$z_v[i,j] = (X_v[i,j] - media_v(j)) / dp_v(j) \quad \text{para } i = 1..M$$

2. Escore bruto ponderado

G = genes presentes em X_v

Para cada amostra $i = 1..M$:

$$s[i] = \text{SUM}_{\{j \text{ em } G\}} \text{beta_final}[j] * z_v[i,j] / \text{SUM}_{\{j \text{ em } G\}} |\text{beta_final}[j]|$$

3. Padronização final

$$s_tilde[i] = (s[i] - média(s)) / dp(s)$$

Saida:

s_tilde : escore CorePAM padronizado ($M \times 1$)

Algoritmo 3 — Validação cruzada determinística com SHA-256

Este algoritmo explica como as pacientes foram divididas nos 10 grupos da validação cruzada. Em muitos estudos, essa divisão depende de uma semente aleatória; mudar a semente pode mudar discretamente o resultado. Aqui, a divisão depende de uma transformação fixa do identificador da paciente. Isso torna a análise reprodutível: a mesma paciente sempre cai no mesmo grupo, em qualquer computador, desde que os mesmos dados sejam usados.

Função atribuir_folds(ids, K):

Para cada i :

$h = \text{SHA256}(\text{string}(\text{ids}[i])) \quad \# 32 \text{ bytes}$

$H = \text{inteiro_sem_sinal}(\text{primeiros_4_bytes}(h))$

```
fold[i] = H mod K
```

```
Retornar fold
```

A propriedade crítica deste procedimento é que $\text{fold}[i]$ depende apenas de $\text{ids}[i]$, não do estado interno do gerador pseudo-aleatório, não da ordem de apresentação das amostras e não da versão do software. Portanto, qualquer implementação correta de SHA-256, em qualquer linguagem, produzirá os mesmos *folds* para a mesma coorte.

Algoritmo 4 — Meta-análise de efeitos aleatórios (REML) (*script 08_meta_survival.R*)

Este algoritmo explica como os resultados das coortes externas foram combinados. Ele não assume que todas as coortes são iguais. Pelo contrário: reconhece que TCGA-BRCA, METABRIC, GSE20685 e GSE1456 diferem em população, plataforma, seguimento e desfecho. O modelo de efeitos aleatórios calcula uma estimativa média levando em conta tanto a incerteza dentro de cada coorte quanto a variação real entre coortes. Essa é a forma estatística de dizer: “o efeito é consistente o bastante para ser resumido, mas as coortes não são cópias umas das outras”.

Entrada:

hr_k : hazard ratio por coorte $k = 1..K$

se_k : erro-padrão de $\log(\text{hr_k})$

1. Tau-quadrado por REML

Maximizar a log-verossimilhança restrita de τ^2 dado $(\text{hr_k}, \text{se_k})$.

2. Pesos de efeitos aleatórios

$w_k = 1 / (\text{se}_k^2 + \tau^2)$

$W = \text{SUM}(w_k)$

3. Efeito agrupado

$\log_hr_pool = \text{SUM}(w_k * \log(\text{hr_k})) / W$

$\text{se_pool} = \sqrt{1 / W}$

$\text{hr_pool} = \exp(\log_hr_pool)$

$\text{IC95} = \exp(\log_hr_pool \pm 1,96 * \text{se_pool})$

4. Heterogeneidade

$$Q = \text{SUM}(w_k * (\log(hr_k) - \log_hr_pool)^2)$$

$$I2 = \max(0, (Q - (K - 1)) / Q)$$

APÊNDICE B — Artigo publicado no *Breast Cancer Research*

O artigo completo “CorePAM: a 24-gene PAM50-derived expression *score* with *cross-platform* external validation for breast cancer prognosis”⁷⁶ é reproduzido integralmente nas páginas seguintes, em versão *Article in Press* disponibilizada pela revista.

Corepam: a 24-gene PAM50-derived expression score with cross-platform external validation for breast cancer prognosis

Received: 6 November 2025

Accepted: 27 April 2026

Published online: 21 May 2026

Cite this article as: Negreiros Botan R. & Sousa J.B. Corepam: a 24-gene PAM50-derived expression score with cross-platform external validation for breast cancer prognosis. *Breast Cancer Res* (2026). <https://doi.org/10.1186/s13058-026-02298-5>

Rafael Negreiros Botan & João Batista Sousa

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

CorePAM: a 24-gene PAM50-derived expression score with cross-platform external validation for breast cancer prognosis

Rafael de Negreiros Botan^{1*} and João Batista de Sousa²

^{1*}Department of Oncology, Universidade de Brasília, Brasília, Brazil.

²Department of Proctology, Universidade de Brasília, Brasília, Brazil.

*Corresponding author(s). E-mail(s): oncologista@gmail.com;

Abstract

Background: The PAM50 classifier predicts breast cancer prognosis but requires 50 genes and specialised platforms. We derived CorePAM, the smallest data-driven PAM50 subset maintaining non-inferior prognostic performance relative to a full 50-gene Cox elastic-net model, without pre-specifying gene count.

Methods: Cox elastic-net regression ($\alpha = 0.5$) with deterministic 10-fold cross-validation was applied in the SCAN-B cohort (N = 3,069; GSE96058). Gene selection followed a pre-specified non-inferiority margin (Δ C-index = 0.010). External validation used four independent cohorts: TCGA-BRCA (N = 1,072; RNA-seq), METABRIC (N = 1,978; microarray; disease-specific survival), GSE20685 (N = 327; microarray), and GSE1456 (N = 159; microarray). Incremental value over a clinical model (CORE-A: age and ER status) was assessed by bootstrap Δ C-index. Secondary analyses evaluated pathologic complete response (pCR) prediction in four neoadjuvant cohorts (N = 697) plus I-SPY2 (N = 986).

Results: CorePAM comprises 24 genes with an out-of-fold C-index of 0.670 (gap vs 50-gene maximum: 0.009). The score was independently associated with survival in all validation cohorts: TCGA-BRCA (HR = 1.20), METABRIC DSS (HR = 1.41), GSE20685 (HR = 1.40), GSE1456 (HR = 1.71); all $p < 0.02$. Random-effects meta-analysis (K = 4) yielded pooled HR = 1.37 (95% CI 1.24–1.52; $I^2 = 38.2\%$; $p = 1.6 \times 10^{-9}$). CorePAM provided incremental value beyond CORE-A and remained significant after adjustment for T-stage and nodal status. For pCR, pooled OR = 1.69 (95% CI 1.39–2.05; $I^2 = 0\%$; $p = 1.9 \times 10^{-7}$).

Conclusions: CorePAM—a 24-gene PAM50-derived score—met the pre-specified Δ C-index non-inferiority criterion relative to a 50-gene Cox elastic-net comparator in derivation and retained prognostic discrimination across four external RNA-seq and microarray cohorts, remained significant after anatomical staging adjustment, and was associated with pCR in secondary neoadjuvant analyses. This reduction from 50 to 24 genes may simplify future assay development, although platform-specific analytical validation is required before clinical deployment.

Keywords: Breast cancer, Gene expression, Prognostic signature, PAM50, Elastic-net regression, Survival analysis, Pathologic complete response, Cross-platform validation

Background

The PAM50 intrinsic subtype gene expression test classifies breast tumours into five molecular subtypes—Luminal A, Luminal B, HER2-enriched, Basal-like, and Normal-like [1, 2]—and underpins the Prosigna prognostic assay [3]. Its prognostic utility is well established across multiple independent cohorts and it is widely used for adjuvant chemotherapy decision-making in hormone receptor-positive, HER2-negative early breast cancer [4, 5].

Despite its clinical validation, PAM50 requires simultaneous measurement of all 50 genes with calibrated reference samples. Its commercial implementation (Prosigna assay) is analytically validated exclusively on the NanoString nCounter platform [6], with substantial per-sample costs. These requirements create barriers in settings where RNA quality is limited (e.g., formalin-fixed paraffin-embedded tissue), platforms are unavailable, or per-sample costs are prohibitive. A reduced panel measurable by targeted RT-PCR or small targeted RNA-seq assays could extend the reach of molecular stratification to resource-limited settings. Existing reduced-gene assays such as the OncotypeDX Recurrence Score (21 genes) [7] and the MammaPrint 70-gene signature [8] were developed with different gene pools and endpoints; a PAM50-specific reduction has not been systematically attempted with modern penalised regression.

Previous attempts to reduce gene panels have typically pre-specified the target gene count before analysis, selected genes by threshold on univariate statistics, or used cohorts that mixed RNA-seq and microarray expression without platform-specific preprocessing. These approaches either introduce analytical circularity (optimising for the assumed answer) or ignore the multi-platform generalizability that clinical deployment demands.

We present CorePAM, a data-driven reduction of the PAM50 gene set derived from the complete SCAN-B RNA-seq cohort ($N = 3,069$) using elastic-net Cox regression with deterministic cross-validation and a pre-specified non-inferiority criterion relative to a full 50-gene Cox elastic-net model. Critically, the gene count was not pre-specified: it was derived entirely from the data and the pre-specified (a priori) and frozen $\Delta C\text{-index} = 0.010$ non-inferiority boundary. Validation was performed across four independent cohorts spanning one RNA-seq and three microarray platforms, including the largest publicly available microarray breast cancer cohort (METABRIC, $N = 1,978$). We further assessed whether CorePAM predicts pathologic complete response (pCR) to neoadjuvant chemotherapy—a clinical endpoint entirely distinct from the derivation analysis—in four independent neoadjuvant cohorts and the large I-SPY2 platform trial.

Methods

Study design

This is a retrospective multicohort study exclusively using publicly available de-identified data, reported following the TRIPOD guideline [9]. No new data were collected. All analytical parameters were pre-specified and frozen before any analysis (see Analytical Freeze). External validation followed established principles for prognostic model assessment [10].

Cohorts

Discovery cohort: SCAN-B (GSE96058; N = 3,069 primary operable breast cancers) [11, 12] provided RNA-seq data (Illumina paired-end sequencing) with matched clinical follow-up. All SCAN-B samples with eligible data were used for model development; no predefined training/validation split was provided in the public metadata.

Validation cohorts (OS/DSS block):

- *TCGA-BRCA* (N = 1,072) [13]: RNA-seq (STAR aligner; GDC release); primary endpoint overall survival (OS).
- *METABRIC* (N = 1,978; cBioPortal) [14]: Illumina HT-12 microarray; primary endpoint disease-specific survival (DSS). Deaths from other causes were censored at time of death. OS was assessed as a sensitivity analysis.
- *GSE20685* (N = 327) [15]: Affymetrix HG-U133 Plus 2.0 microarray (GPL570); primary endpoint OS.
- *GSE1456* (N = 159; Stockholm cohort) [16]: Affymetrix HGU133A/B microarray; primary endpoint OS. Clinical covariates (age, ER status) were not available in the public annotation; CORE-A adjustment was not performed for this cohort.

pCR block cohorts (secondary analysis, independent of OS block):

- GSE25066 (N = 508) [17], GSE20194 (N = 278) [18], GSE32646 (N = 154) [19], I-SPY1/GSE22226 (N = 149) [20]: GEO microarray cohorts with pCR data from neoadjuvant chemotherapy trials. After applying eligibility criteria (available pCR outcome and expression data for CorePAM genes), 697 samples were included in the primary pCR analysis.
- I-SPY2/GSE194040 (N = 986) [21]: exploratory external cohort; pre-treatment RNA-seq biopsies; multi-arm neoadjuvant trial; batch-corrected by the original consortium using ComBat.

All cohorts underwent independent harmonisation; no expression or clinical data were pooled across cohorts.

Analytical freeze

The following parameters were pre-specified and frozen before any analysis was performed, archived in the study repository (Table 1).

Table 1 CorePAM derivation parameters and frozen analytical settings.

Parameter	Value
Non-inferiority margin (ΔC -index)	0.010
Elastic-net mixing parameter (α)	0.5
Cross-validation folds (K)	10
Fold assignment	Deterministic (SHA-256 of patient ID)
Minimum gene coverage (validation)	80%
Time unit	months (days / 30.4375)
Gene count	Not pre-specified (derived from data)

Expression preprocessing

All preprocessing was performed independently per cohort with no inter-cohort pooling or batch correction.

RNA-seq cohorts (SCAN-B, TCGA-BRCA): Raw counts were filtered with `edgeR::filterByExpr` [22], normalised by TMM, and transformed to log-CPM (`prior.count = 1`). Ensembl identifiers were mapped to official HGNC symbols via `org.Hs.eg.db` (offline annotation). In the case of multiple Ensembl IDs mapping to the same HGNC symbol, the identifier with the highest intra-cohort variance was retained.

Microarray cohorts (METABRIC, GSE20685, GSE1456): Log2-scaled intensities from GEO Series Matrix files (pre-processed by the original data submitters) were used without additional normalisation. METABRIC identifiers were mapped via `org.Hs.eg.db`; GSE20685 (Affymetrix HG-U133 Plus 2.0, GPL570) via `hgu133plus2.db`;

GSE1456 (Affymetrix HGU133A/B, GPL96) via hgu133a.db. GSE20685 achieves full CorePAM coverage ($24/24 = 100\%$). For GSE1456, the HGU133A platform ceiling is 44/50 PAM50 genes (88%, above the 80% minimum), with three CorePAM genes absent (GPR160, CXXC5, KRT17; $21/24 = 87.5\%$). METABRIC ER status was obtained from the cBioPortal clinical file (ER_IHC column).

Score completeness: When fewer than 24 CorePAM genes are available (e.g., GSE1456: 21/24 genes, 87.5%), the score is computed over available genes only, with the denominator renormalised accordingly. Gene coverage is reported per cohort (Table 2); the pre-specified minimum of 80% was met in all cases. Sensitivity to missing genes is further assessed in a leave-one-gene-out analysis (Additional file 1: Figure S1).

Table 2 CorePAM gene coverage by validation cohort.

Cohort	Platform	Genes present	Coverage (%)	Missing genes
SCAN-B	RNA-seq (Illumina)	24/24	100	—
TCGA-BRCA	RNA-seq (STAR/GDC)	23/24	95.8	MIA
METABRIC	Illumina HT-12 microarray	24/24	100	—
GSE20685	Affymetrix HG-U133 Plus 2.0	24/24	100	—
GSE1456	Affymetrix HGU133A/B	21/24	87.5	GPR160, CXXC5, KRT17

CorePAM derivation

Cox elastic-net regression ($\alpha = 0.5$ [23, 24]; glmnet package [25, 26]) was fitted in SCAN-B using a 10-fold cross-validation scheme. Fold assignment was *deterministic*, based on the first eight hexadecimal digits of the SHA-256 hash of each patient identifier, with event-stratification. This deterministic fold assignment ensures exact reproducibility from first principles without dependence on a random seed as driver of gene selection.

Out-of-fold (OOF) C-index was computed for each degree-of-freedom level (df = number of non-zero genes) along the regularisation path. CorePAM was defined as

the *smallest* df satisfying: $\text{OOF C-index}(df) \geq \text{OOF C-index}(\max) - 0.010$. The non-inferiority margin of 0.010 was chosen because it is less than half of one approximate standard error of the OOF C-index in the training cohort ($\text{SE} \approx 0.026$ for $N = 3,069$ with 322 events), representing a clinically negligible loss in discrimination. Sensitivity analysis confirmed robustness: margins of 0.005, 0.015, and 0.020 would have selected 31, 18, and 16 genes respectively, demonstrating that the selected panel is not brittle to the choice of margin. Final model coefficients were extracted from the full-data path at the selected λ .

Score calculation

For each cohort, the CorePAM score was computed as:

$$\text{score} = \frac{\sum_i w_i \cdot z_i}{\sum_i |w_i|} \quad (1)$$

where the summation runs over the set of CorePAM genes present in the cohort (G_{present}), z_i is the intra-cohort z-score of gene i , and w_i is its elastic-net Cox regression weight from the frozen model. The score was subsequently standardised intra-cohort ($\text{score}_z = \text{scale}(\text{score})$) for proportional hazards regression, yielding hazard ratios per 1 SD. Score direction (positive = higher risk) was established in the SCAN-B training cohort (training HR = 1.92, confirming positive orientation); the same directional convention was confirmed to hold in all four external validation cohorts (all univariate HRs > 1) without requiring sign inversion in any cohort.

Survival analysis

Univariate Cox: score_z as the sole predictor; HR per 1 SD with Wald confidence intervals.

Multivariate Cox (CORE-A model): score_z adjusted for age (continuous) and ER status (binary), when ER was available in harmonised clinical data. CORE-A covariates by cohort: SCAN-B—age + ER status; METABRIC—age + ER status; TCGA-BRCA—age alone (ER status was not recoverable from the GDC-harmonised clinical file); GSE20685—age alone (ER status absent from the GSE Series Matrix clinical annotation). CORE-A denotes the clinical-only model (age $\pm$ ER).

C-index: Harrell's C-index [27] with 95% bootstrap CI ($B = 1,000$). Uno's C-statistic (IPCW) [28] was computed as a sensitivity analysis to address heterogeneous censoring; time-dependent AUC at fixed horizons (60 months; 24 months for TCGA-BRCA) was estimated via the timeROC package [29].

Kaplan–Meier: dichotomised at the intra-cohort median. Cox HR for the dichotomised groups (High vs Low) with 95% CI and Wald p -value are annotated on each figure. Quartile sensitivity analysis.

Proportional hazards assumption: The proportional hazards assumption was assessed graphically via scaled Schoenfeld residual plots and formally tested using the Grambsch–Therneau global test [30] for each cohort. No significant departures were detected in SCAN-B, GSE20685, or GSE1456 (all global $p > 0.05$). METABRIC DSS showed marginal non-proportionality ($p = 0.04$), expected given its 159-month follow-up; the Fine–Gray competing-risks analysis and the concordance-based C-index (which does not assume proportional hazards) mitigate this concern (Additional file 2: Figure S2).

Fine–Gray competing risks [31]: applied to METABRIC DSS (deaths from other causes as competing event).

Incremental value: $\Delta C\text{-index} = C(\text{CORE-A} + \text{score}_z) - C(\text{CORE-A})$; bootstrap 95% CI ($B = 1,000$).

Meta-analysis: Random-effects model using REML estimator for the OS/DSS block; DerSimonian–Laird [32] for the pCR block (metafor package [33]). Heterogeneity

was quantified using I^2 [34]. Four independent validation cohorts were included ($K = 4$: TCGA-BRCA, METABRIC, GSE20685, GSE1456); SCAN-B excluded as training cohort. Hartung–Knapp adjustment [35] applied as sensitivity analysis.

Calibration: Formal calibration was assessed via logistic recalibration at the cohort-specific time horizon. Calibration intercept (α ; ideal = 0) and slope (β ; ideal = 1) were obtained by regressing the binary outcome indicator Y_H (event before horizon H) on the logit of the Cox-predicted risk at H . The expected-to-observed (E/O) ratio and Brier score were also computed.

Frozen z-score sensitivity: As a sensitivity analysis addressing single-sample scoring feasibility, the mean and standard deviation per gene were frozen from the SCAN-B training cohort and applied to all external validation cohorts instead of computing intra-cohort z-scores. No cohort-level re-standardisation (i.e. no `scale()`) was applied to the composite score, ensuring that each patient’s score depends only on their own expression values and the frozen SCAN-B reference—as would occur in clinical deployment with a single specimen. Cox HR and C-index were compared between frozen and intra-cohort approaches on the same raw-score scale.

Subgroup analysis: ER-stratified Cox HR, C-index, and formal score $\times$ ER interaction tests were performed in SCAN-B (OS; ER available in 94%) and METABRIC (DSS; ER available in 98%). These analyses were added post hoc in response to reviewer feedback; all other modelling choices remained frozen.

Full clinical baseline: As a secondary analysis, CorePAM was tested for incremental value beyond a comprehensive 6-variable clinical model (age + ER status + T-stage + N-status + grade + HER2 status) in SCAN-B and METABRIC—the two cohorts where all covariates were available. Complete-case analysis was used; reported sample sizes reflect patients with all six covariates available (SCAN-B: 2,649/3,069; METABRIC: 1,792/1,978). Incremental value was quantified by Δ C-index (bootstrap

$B = 1,000$) and a likelihood ratio test (LRT) comparing the clinical-only versus clinical + CorePAM models. This analysis was also performed restricted to ER-positive patients.

Decision curve analysis (DCA) [36]: Net benefit at clinically relevant probability thresholds using `dcurves::dca` (OS: 60-month horizon for SCAN-B, METABRIC, GSE20685; 24 months for TCGA-BRCA).

Clinical baseline model scope: CORE-A was pre-specified as the minimum clinical baseline (age and ER status) for consistency across cohorts, avoiding post-hoc covariate selection decisions. T-stage and nodal status were reserved for exploratory sensitivity analyses (Additional file 3: Table S1), as their harmonisation quality varied across public data sources and their inclusion in the primary model would have introduced cohort-specific modelling decisions incompatible with a uniform validation framework.

pCR analysis

Logistic regression: $\text{pCR} \sim \text{score}_z$ (univariate); adjusted for age and ER/HER2 status where available. OR per 1 SD with Wald CI. AUC by DeLong method. Bootstrap CI ($B = 1,000$, seed = 42). I-SPY2 sensitivity analysis restricted to the paclitaxel control arm ($N = 179$).

Quality control

Comprehensive QC comprised: (i) off-diagonal score correlation between cohorts via percentile-based Pearson r (Additional file 4: Figure S3); (ii) forensic PCA of METABRIC expression for batch effects and outliers (Additional file 5: Figure S4); (iii) schema/range hard-checks on all analysis-ready files; (iv) automated anti-hardcoding assertions verifying all reported values against source CSV/JSON files.

Data and code availability

All raw data are publicly available (GEO, GDC, cBioPortal). Analysis code is available at <https://github.com/RafaelBotan/CorePAM-Study> [37]. All results are reproducible from the repository.

Computational assistance

A large language model (Claude, Anthropic) was used as a coding assistant for directory organisation, script development, and manuscript formatting. All code was reviewed and validated by the authors before use.

Statistical software

R version 4.5.2. Key packages: glmnet [25] (elastic-net), survival [38]/survminer (Cox, KM), metafor [33] (meta-analysis), dcurves (DCA), edgeR [22] (RNA-seq normalisation), genefu [39] (signature comparison), survC1 (Uno's C-statistic), timeROC [29] (time-dependent AUC).

Results

CorePAM derivation

Out-of-fold Cox elastic-net regression in SCAN-B identified 24 genes as the minimum set satisfying the pre-specified non-inferiority criterion ($\Delta C = 0.010$). The selected model achieved an OOF C-index of 0.670, compared to the maximum OOF C-index achievable with all 50 PAM50 candidates (0.679), with a gap of 0.009—within the 0.010 boundary. The selected 24 genes and their elastic-net Cox weights are illustrated in Additional file 6: Figure S5 and listed in Additional file 7: Table S2.

The 24 genes span luminal differentiation (ESR1, PGR, NAT1), proliferation (MYBL2, PTTG1, EXO1, MYC), basal phenotype (KRT5, KRT17, FOXC1), and

HER2-enriched markers (ERBB2, GRB7, MIA), with full gene list in Additional file 7: Table S2. The Pareto frontier (Additional file 8: Figure S6) shows diminishing returns beyond approximately 15 genes, with the non-inferiority boundary crossed at $df = 24$. Bootstrap resampling ($B = 200$) confirmed that 15 of 24 CorePAM genes were selected in at least 70% of resamples, with a minimum selection frequency of 35.5% (Additional file 16: Figure S10). To assess whether gene reduction compromised external discrimination, we compared CorePAM directly against the PAM50-full elastic-net model trained on all 50 candidate genes (49 non-zero coefficients at the selected λ) in all four validation cohorts under both cohort-standardised and frozen-reference scoring. In paired bootstrap analyses, differences were small in TCGA-BRCA and favoured CorePAM in the remaining three cohorts (Additional file 18: Table S8).

Prognostic validation

The CorePAM score was independently associated with survival in all four external validation cohorts (Table 3). In TCGA-BRCA (RNA-seq), HR per 1 SD = 1.20 (95% CI 1.04–1.40; $p = 0.016$). In METABRIC (DSS endpoint; Illumina microarray), HR = 1.41 (95% CI 1.31–1.52; $p < 0.001$). In GSE20685 (Affymetrix microarray), HR = 1.40 (95% CI 1.12–1.75; $p = 0.003$). In GSE1456 (Stockholm; Affymetrix microarray), HR = 1.71 (95% CI 1.23–2.33; $p = 0.001$). Kaplan–Meier curves stratified at the intra-cohort median score are shown in Figure 2.

TCGA-BRCA has shorter follow-up (median 32 months), attenuating the HR; a 24-month sensitivity analysis yielded HR = 1.64 (95% CI 1.24–2.16; $p = 4.7 \times 10^{-4}$; Additional file 9: Figure S7).

Multivariate Cox regression including the CORE-A clinical model (age and ER status where available) yielded adjusted HRs of 1.27 (TCGA-BRCA), 1.36 (METABRIC), and 1.40 (GSE20685), indicating that CorePAM provides prognostic information independent of clinical covariates. GSE1456 lacked clinical covariates for CORE-A

Table 3 CorePAM prognostic performance across cohorts.

Cohort	Endpoint	N	Events	Univariate HR	Adjusted HR ¹	C-index
SCAN-B ²	OS	3,069	322	1.92 (1.74–2.13)	1.86 (1.64–2.12)	0.698 (0.668–0.729)
TCGA-BRCA	OS	1,072	150	1.20 (1.04–1.40)	1.27 (1.09–1.48)	0.624 (0.572–0.678)
METABRIC	DSS	1,978	646	1.41 (1.31–1.52)	1.36 (1.24–1.48)	0.638 (0.615–0.659)
GSE20685	OS	327	83	1.40 (1.12–1.75)	1.40 (1.12–1.75)	0.626 (0.565–0.684)
GSE1456 ³	OS	159	40	1.71 (1.23–2.38)	—	0.661 (0.582–0.741)

HR: hazard ratio per 1 SD (95% CI); all $p < 0.02$.

¹CORE-A: age + ER (SCAN-B, METABRIC); age only (TCGA-BRCA, GSE20685).

²Training cohort. C-index is in-sample (0.698); OOF C-index = 0.670.

³Clinical covariates unavailable; CORE-A adjustment not performed.

adjustment; the univariate HR of 1.71 nonetheless confirms cross-platform generalizability on an independent Affymetrix dataset. The training cohort (SCAN-B; panel A of Figure 2) showed an in-sample C-index of 0.698 (OOF C-index = 0.670).

Meta-analysis

Random-effects meta-analysis across four validation cohorts yielded a pooled HR of 1.37 (95% CI 1.24–1.52; $p = 1.6 \times 10^{-9}$) with moderate heterogeneity ($I^2 = 38.2\%$; Figure 3). The Hartung–Knapp sensitivity analysis confirmed significance under conservative inference (Table 4). OS-harmonised and leave-one-out sensitivity analyses maintained consistent positive direction.

Table 4 Random-effects meta-analysis of CorePAM prognostic association ($K = 4$ validation cohorts).

Analysis	HR (95% CI)	p	I^2
Primary RE (REML)	1.37 (1.24–1.52)	1.6×10^{-9}	38.2%
Hartung–Knapp	1.37 (1.15–1.64)	0.011	38.2%
OS-harmonised	1.36 (1.13–1.64)	<0.001	50.5%

Incremental value beyond clinical predictors

The CorePAM score improved the C-index over the CORE-A clinical model in all four cohorts (Table 5). CORE-A covariates were: age and ER status in SCAN-B and METABRIC; age alone in TCGA-BRCA and GSE20685 (where ER status was absent from harmonised clinical files). The largest increment was observed in GSE20685 ($\Delta C = +0.093$), consistent with the long follow-up and high event rate in that cohort (Figure 4). Calibration at clinically relevant horizons showed adequate concordance between predicted and observed event rates in the training cohort, with expected under-confidence in external validation cohorts (Figure 5). Decision curve analysis confirmed net clinical benefit across a range of probability thresholds in all cohorts (Additional file 10: Figure S8). Sensitivity analysis comparing age-only versus age + ER baselines confirmed that the CorePAM HR remains significant under both definitions (Additional file 11: Figure S9).

Table 5 Incremental C-index of CorePAM score over the CORE-A clinical model.

Cohort	CORE-A covariates	N	C (CORE-A)	C (CORE-A + CorePAM)	ΔC (95% CI)
SCAN-B	age + ER	3,069	0.750	0.780	+0.030 (0.014–0.049)
TCGA-BRCA	age	1,072	0.638	0.676	+0.038 (0.011–0.086)
METABRIC	age + ER	1,978	0.603	0.646	+0.043 (0.025–0.066)
GSE20685	age	327	0.536	0.630	+0.094 (0.021–0.175)

Sensitivity analyses

Frozen z-score (single-sample scoring feasibility). When the SCAN-B training cohort mean and standard deviation per gene were frozen and applied to external validation cohorts—without any cohort-level re-standardisation—the C-index was preserved in RNA-seq and array cohorts sharing a similar dynamic range: TCGA-BRCA $\Delta C = +0.0002$ (Pearson $r = 0.999$); METABRIC $\Delta C = +0.0007$ ($r = 0.958$). In the

Affymetrix microarray cohorts, which have a markedly different intensity distribution from the RNA-seq training reference, discrimination was attenuated: GSE20685 $\Delta C = -0.036$ ($r = 0.923$); GSE1456 $\Delta C = -0.040$ ($r = 0.916$). These results suggest that frozen single-sample scoring is feasible when the assay platform is compatible with the training reference, but that platform-specific recalibration would be needed for Affymetrix-based deployment (Additional file 12: Table S3).

Uno's C-statistic and time-dependent AUC. IPCW-based discrimination metrics confirmed the Harrell's C-index findings: Uno's C ranged from 0.668 (GSE20685) to 0.681 (METABRIC); time-dependent AUC at the cohort-specific horizon ranged from 0.675 (GSE20685, 60m) to 0.695 (METABRIC, 60m). These concordance estimates are robust to heterogeneous censoring (Additional file 13: Table S4).

Formal calibration. Logistic recalibration at the cohort-specific horizon yielded calibration slopes ranging from 0.96 (SCAN-B, near-ideal) to 2.74 (TCGA-BRCA), with corresponding intercepts of 0.56 (SCAN-B) to 5.40 (TCGA-BRCA). The training cohort showed a near-ideal calibration slope ($\beta = 0.96$), though the non-zero intercept ($\alpha = 0.56$) indicates moderate systematic underprediction of absolute risk. External validation cohorts showed miscalibration that was mild in METABRIC ($\beta = 1.77$) and GSE20685 ($\beta = 1.57$) but substantial in TCGA-BRCA ($\beta = 2.74$), the latter consistent with its short follow-up and low event rate at the 24-month horizon (Additional file 14: Table S5).

ER-stratified subgroup analysis. In SCAN-B ($N = 2,870$; OS), CorePAM was significantly associated with survival in both ER-positive (HR = 1.92; 95% CI 1.69–2.17; $p < 0.001$; $N = 2,646$) and ER-negative (HR = 2.53; 95% CI 1.51–4.25; $p < 0.001$; $N = 224$) subgroups; the interaction test was not significant ($p = 0.25$). In METABRIC ($N = 1,936$; DSS), CorePAM was strongly prognostic in ER-positive patients (HR = 1.59; 95% CI 1.44–1.76; $p < 0.001$; $N = 1,497$) but showed no signal

in ER-negative patients (HR = 0.97; 95% CI 0.82–1.15; $p = 0.75$; $N = 439$); the interaction was significant ($p = 3.2 \times 10^{-7}$). The discrepancy in ER-negative results reflects statistical power: SCAN-B ER– had only 35 events in 224 patients (bootstrap 95% CI for HR: 1.45–4.89; width = 3.44), whereas METABRIC ER– had 186 events in 439 patients (bootstrap 95% CI: 0.83–1.15; width = 0.32; 10.8× narrower). The well-powered METABRIC estimate indicates that the clinically meaningful signal is concentrated in ER-positive disease, consistent with the luminal biology of the retained CorePAM genes.

Independence from clinical-pathological staging

To address whether CorePAM retains prognostic value after adjustment for standard clinical-pathological variables, we fitted expanded multivariable Cox models adjusting for age, ER status (where available), pathological T-stage, and nodal status. T-stage and nodal status were available in 96–100% of samples across the four cohorts with clinical covariates (SCAN-B, TCGA-BRCA, METABRIC, GSE20685). CorePAM remained independently significant in all four (all $p < 0.002$; Additional file 3: Table S1), with HR change of less than 10% compared to the primary CORE-A baseline. Adding CorePAM to the expanded clinical model increased the C-index by +0.022 to +0.027.

In a comprehensive secondary analysis including all six standard clinical-pathological variables (age, ER status, T-stage, nodal status, histological grade, and HER2 status), CorePAM retained independent significance in both SCAN-B (adjusted HR = 1.75; 95% CI 1.51–2.03; $p = 1.6 \times 10^{-13}$; $\Delta C = +0.017$) and METABRIC (adjusted HR = 1.25; 95% CI 1.14–1.38; $p = 5.5 \times 10^{-6}$; $\Delta C = +0.015$; Additional file 15: Table S6). Likelihood ratio tests confirmed that CorePAM adds significant information beyond the full clinical model (both $p < 10^{-5}$). In ER-positive patients,

the incremental value was most pronounced in METABRIC ($\Delta C = +0.032$; 95% CI 0.018–0.050; LRT $p = 10^{-10}$).

Pathologic complete response (secondary analysis)

In four independent neoadjuvant cohorts (N = 697 total), the CorePAM score showed a positive association with pCR in all four cohorts, reaching statistical significance in three of four (Table 6). GSE25066 showed a positive but non-significant trend (OR = 1.44; $p = 0.060$). ORs ranged from 1.44 to 2.10 per 1 SD. The AUC ranged from 0.576 to 0.716, indicating moderate to good discriminative ability for pCR prediction.

Table 6 CorePAM score predicts pCR in neoadjuvant chemotherapy cohorts.

Cohort	N	pCR, n (%)	OR per 1 SD (95% CI)	<i>p</i>	AUC (95% CI)
GSE25066	182	42 (23.1%)	1.44 (0.98–2.09)	0.060	0.576 (0.477–0.676)
GSE20194	278	56 (20.1%)	1.73 (1.25–2.38)	<0.001	0.660 (0.586–0.734)
GSE32646	115	27 (23.5%)	2.10 (1.31–3.39)	0.002	0.716 (0.611–0.821)
I-SPY1	122	32 (26.2%)	1.66 (1.05–2.63)	0.030	0.622 (0.517–0.726)
RE pooled	697		1.69 (1.39–2.05)	<0.001	

OR: odds ratio per 1 SD CorePAM score; AUC: area under the receiver operating characteristic curve; RE: random effects (DerSimonian–Laird). $I^2 = 0\%$.

Random-effects meta-analysis yielded a pooled OR of 1.69 (95% CI 1.39–2.05; $I^2 = 0\%$; $p = 1.9 \times 10^{-7}$; Figure 6). The absence of heterogeneity across four cohorts from different platforms and trials reinforces the cross-platform consistency of the CorePAM pCR association.

I-SPY2 exploratory analysis. In the I-SPY2 external exploratory cohort (GSE194040; N = 986; pCR rate = 32.4%), the CorePAM score predicted pCR univariately (OR = 1.68; 95% CI 1.45–1.95; AUC = 0.648). Inclusion of I-SPY2 in the meta-analysis ($k = 5$) yielded OR = 1.69 (95% CI 1.50–1.90; $I^2 = 0\%$), virtually identical to the primary analysis, confirming external generalizability.

Discussion

CorePAM is a data-driven 24-gene reduction of PAM50 selected within a pre-specified ΔC -index framework relative to a full 50-gene elastic-net comparator. In newly added direct paired external comparisons, discrimination was similar in TCGA-BRCA and favoured the reduced model in METABRIC, GSE20685, and GSE1456 under both cohort-standardised and frozen-reference scoring (Additional file 18: Table S8). These findings support the robustness of the reduced model within the present elastic-net framework. However, external calibration and frozen-reference single-sample performance remained platform-dependent, so these results should not be interpreted as evidence of platform-agnostic deployment or equivalence with commercial PAM50-based assays (Prosigna). No claims of therapeutic equivalence or clinical decision-making impact are made.

Reproducible derivation. Fold assignment was deterministic (SHA-256 of patient identifiers), eliminating random seeds as a driver of gene selection and ensuring exact reproducibility from first principles. The gene count ($N = 24$) was not pre-specified; it was derived from the Pareto frontier of model complexity versus OOF C-index (Additional file 8: Figure S6), where performance plateaus well before 50 genes.

Gene selection stability. Bootstrap resampling ($B = 200$ refits on SCAN-B) confirmed that 15 of 24 CorePAM genes were selected in at least 70% of resamples (Additional file 16: Figure S10), with 2 genes (ACTR3B, NAT1) selected in 100%. The remaining genes show lower but non-negligible selection frequencies (minimum: 35.5%), consistent with the elastic-net grouping property that distributes weight among correlated genes. Leave-one-gene-out analysis showed that no single gene removal causes catastrophic performance loss (Additional file 1: Figure S1; maximum $|\Delta C| = 0.044$).

Heterogeneity is expected and informative. The moderate I^2 in the OS/DSS meta-analysis reflects two structural sources: endpoint divergence (DSS in METABRIC vs. OS in the other cohorts) and differential follow-up duration (32 months in TCGA-BRCA vs. 91–159 months elsewhere). The OS-harmonised sensitivity analysis (substituting METABRIC OS for DSS) reduced I^2 substantially, confirming the pooled estimate is not driven by endpoint choice. The key finding is the consistency of *direction*—a positive HR in all four validation cohorts—rather than effect magnitude. The Hartung–Knapp sensitivity analysis confirmed significance under conservative inference.

Biological coherence. The retained genes recapitulate core PAM50 biology: luminal differentiation (ESR1, PGR, BCL2, NAT1), proliferation (MYBL2, PTTG1, EXO1, MYC), basal phenotype (KRT5, KRT17, FOXC1), and HER2-enriched markers (ERBB2, GRB7). The proliferation genes also explain the pCR signal, as these pathways are targeted by anthracycline-taxane regimens. The 26 excluded genes received zero weight at the selected λ , consistent with functional redundancy under the elastic-net grouping property.

Cross-platform consistency in pCR prediction. The pCR block showed no heterogeneity ($I^2 = 0\%$ across four cohorts), consistent with the biological basis of the signal: the retained proliferation genes are directly targeted by standard neoadjuvant regimens [40].

Calibration and transportability. Calibration at clinically relevant horizons (60 months for SCAN-B, METABRIC, GSE20685; 24 months for TCGA-BRCA) showed adequate concordance between predicted and observed event rates in the training cohort (Figure 5). Formal logistic recalibration confirmed a near-ideal calibration slope in the training cohort ($\beta = 0.96$; intercept $\alpha = 0.56$, indicating moderate underprediction of absolute risk) with further miscalibration in external cohorts (slopes 1.57–2.74), consistent with model under-confidence when applying

scores cross-platform without recalibration—a structural property of intra-cohort z-scoring. Exploratory decision curve analysis suggested positive net benefit across a range of probability thresholds (Additional file 10: Figure S8).

Frozen z-score sensitivity. A key reviewer concern was whether the intra-cohort z-score normalisation—appropriate for retrospective multi-cohort validation—could support single-sample scoring. Our frozen z-score sensitivity analysis applied SCAN-B reference mean/SD per gene to each external cohort without any cohort-level re-standardisation, ensuring that each patient’s score depends only on their own expression values and the frozen reference. Discrimination was preserved in TCGA-BRCA ($\Delta C = +0.0002$; $r = 0.999$) and METABRIC ($\Delta C = +0.0007$; $r = 0.958$). In the Affymetrix cohorts, frozen scoring showed measurable attenuation (GSE20685: $\Delta C = -0.036$, $r = 0.923$; GSE1456: $\Delta C = -0.040$, $r = 0.916$), consistent with the larger distributional gap between RNA-seq and these microarray platforms. These findings support platform-conditional rather than platform-agnostic fixed-reference scoring and indicate that platform-specific recalibration would be required before deployment.

Uno’s C-statistic confirms robustness to censoring. IPCW-based concordance metrics (Uno’s C, time-dependent AUC) yielded estimates consistent with Harrell’s C-index, confirming that the observed discrimination is not an artefact of heterogeneous censoring distributions across cohorts.

ER subgroup analysis. CorePAM is prognostic in unselected cohorts overall, with the clinically meaningful signal concentrated in ER-positive disease. In both SCAN-B and METABRIC, the ER-positive HR was robust (1.92 and 1.59, respectively; both $p < 0.001$). In ER-negative patients, SCAN-B showed an apparent effect (HR = 2.53) that was not replicated in the better-powered METABRIC analysis (HR = 0.97; N = 439, 186 events). Bootstrap analysis confirmed that the SCAN-B ER-negative estimate is highly unstable (95% CI width = 3.44 vs 0.32 in METABRIC),

consistent with its small sample (35 events). This behaviour is consistent with luminal biology and with the clinical use of multigene prognostic assays primarily in ER-positive disease [5, 41].

Independence from clinical-pathological staging. CorePAM retained independent prognostic value after adjustment for all six standard clinical-pathological variables (age, ER, T-stage, N-status, grade, HER2) in both SCAN-B (LRT $p = 4.9 \times 10^{-13}$) and METABRIC (LRT $p = 6.1 \times 10^{-6}$). The ΔC of +0.015 to +0.017 over this comprehensive baseline—and +0.032 in ER-positive METABRIC—confirms that CorePAM captures molecular heterogeneity not accessible through routine clinical-pathological assessment.

Comparison with established signatures. An exploratory head-to-head comparison against research-based implementations (genefu R package [39]) of the PAM50-derived Risk of Recurrence score (ROR-S, 50 genes) and the OncotypeDX Recurrence Score (ODX-RS, 21 genes) showed that CorePAM (24 genes) performs competitively (Additional file 17: Table S7). CorePAM showed advantages in the RNA-seq cohorts (its training domain) and small deficits ($\Delta C < 0.03$) in the microarray cohorts (the development domain of those signatures). This comparison involves research-based score implementations—not commercial assays—and should be interpreted as evidence of competitive performance, not clinical equivalence.

Clinical implications. A smaller gene set may simplify future assay development, but analytical validation on specific deployment platforms and prospective clinical validation were beyond the scope of this study.

Limitations. *Study design and data.* This is a retrospective study using public data, and no prospective validation was performed. TCGA-BRCA has structurally short follow-up (median 32 months), which may attenuate the detectable HR; the 24-month sensitivity analysis (Additional file 9: Figure S7) partially mitigates this concern. METABRIC uses DSS as the primary endpoint, following established precedent

for that dataset. Endpoint heterogeneity across cohorts is reflected in the moderate meta-analytic I^2 and should be considered when interpreting pooled estimates. HER2 status was not available in sufficient cohorts to permit HER2-stratified subgroup analysis.

Analytical choices. The primary CORE-A clinical baseline was restricted to age and ER status—the two variables consistently available in harmonised form—although expanded models including T-stage and nodal status confirmed independent prognostic value (Additional file 3: Table S1). ER status was unavailable in TCGA-BRCA and GSE20685, restricting CORE-A to age alone in those cohorts. The primary score formulation relied on cohort-wise z-score standardisation, which limits direct transportability of absolute risk estimates. In frozen-reference analyses, discrimination remained stable in TCGA-BRCA and METABRIC ($\Delta C < 0.001$) but was attenuated in the Affymetrix cohorts (GSE20685: $\Delta C = -0.036$; GSE1456: $\Delta C = -0.040$), indicating that fixed-reference single-sample scoring is currently platform-dependent rather than platform-agnostic. Gene coverage remained high across validation cohorts (24/24 in METABRIC and GSE20685; 23/24 in TCGA-BRCA; 21/24 in GSE1456), and leave-one-gene-out sensitivity analysis confirmed that the score is robust to the absence of any single gene (Additional file 1: Figure S1). Comparisons with ROR-S and ODX-RS used research-based implementations, not commercial assays.

Clinical translation. A smaller gene set may simplify future assay development, but analytical validation on specific deployment platforms and platform-specific housekeeping-gene calibration were beyond the scope of this study and are the subject of planned technical validation work.

Conclusions

CorePAM is a reproducible 24-gene reduction of the PAM50 panel selected within a pre-specified ΔC -index framework relative to a full 50-gene elastic-net comparator.

Across four external RNA-seq and microarray cohorts, the reduced model retained prognostic discrimination. In cohorts with sufficient clinicopathological annotation, it also provided incremental value beyond a comprehensive six-variable clinical baseline (age, ER, T-stage, N-status, grade, HER2; LRT $p < 10^{-5}$), with the clinically meaningful signal concentrated in ER-positive disease. In direct paired external comparisons, discrimination was similar to or higher than the fuller comparator (Additional file 18). CorePAM was also associated with pathologic complete response in secondary neoadjuvant analyses ($I^2 = 0\%$). External calibration and frozen-reference performance remained platform-dependent, indicating that platform-specific recalibration and prospective analytical validation are required before single-sample deployment. The analytical pipeline is fully reproducible from publicly available data.

List of abbreviations

AUC: Area under the curve; CI: Confidence interval; CORE-A: Clinical baseline model (age $\pm$ ER status); CPM: Counts per million; CV: Cross-validation; DCA: Decision curve analysis; DL: DerSimonian–Laird; DSS: Disease-specific survival; ER: Estrogen receptor; FFPE: Formalin-fixed paraffin-embedded; GDC: Genomic Data Commons; GEO: Gene Expression Omnibus; HER2: Human epidermal growth factor receptor 2; HGNC: HUGO Gene Nomenclature Committee; HR: Hazard ratio; KM: Kaplan–Meier; LLM: Large language model; OOF: Out-of-fold; OR: Odds ratio; OS: Overall survival; PAM50: Prediction Analysis of Microarray 50; PCA: Principal component analysis; pCR: Pathologic complete response; RE: Random effects; REML: Restricted maximum likelihood; ROR-S: Risk of Recurrence score (subtype-based); SD: Standard deviation; SE: Standard error; TMM: Trimmed mean of M-values.

Supplementary information. Additional files 1–18 are available as supplementary material. See the Additional files section at the end of this article for descriptions.

Acknowledgements. The authors thank the SCAN-B, TCGA, METABRIC, I-SPY, and Stockholm (GSE1456) consortia for making their data publicly available.

Declarations

Ethics approval and consent to participate Not applicable. This study uses exclusively publicly available de-identified data (GEO, GDC, cBioPortal). No new human subjects research was conducted.

Consent for publication Not applicable.

Availability of data and materials All raw data are available through GEO (GSE96058, GSE25066, GSE20194, GSE32646, GSE22226, GSE194040, GSE1456), GDC (TCGA-BRCA), and cBioPortal (METABRIC). Analysis code, frozen analytical parameters, and a single-command reproducibility script are available at <https://github.com/RafaelBotan/CorePAM-Study> [37] (release tagged at submission). All results can be reproduced from raw public data using the provided pipeline.

Competing interests The authors declare no competing interests.

Funding No external funding was received. This study was conducted independently by the authors.

Authors' contributions RB designed the study, performed all analyses, and wrote the manuscript. JBS supervised the study and reviewed the manuscript. Both authors read and approved the final manuscript.

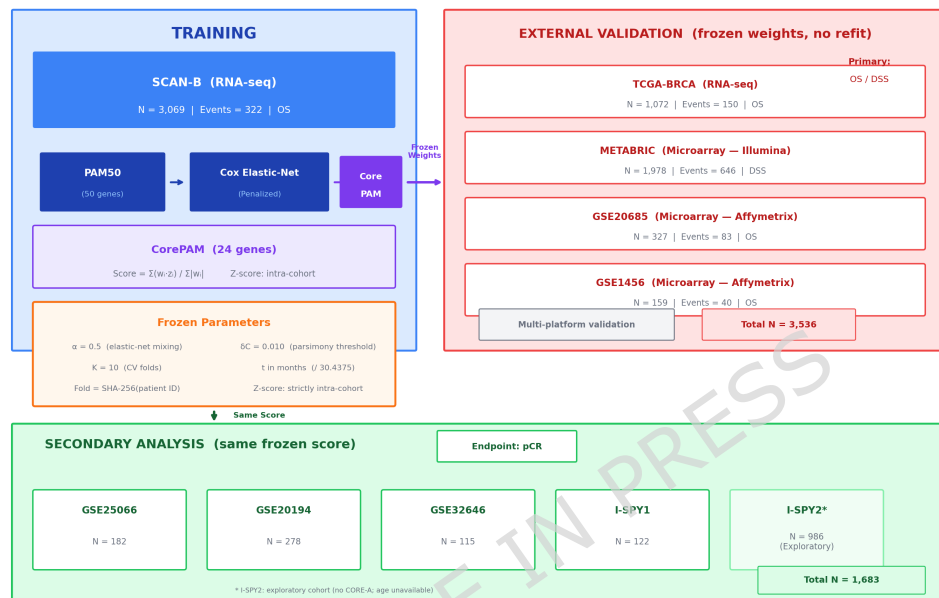


Fig. 1 Study design overview. CorePAM: a 24-gene prognostic signature derived from the PAM50 gene set via elastic-net Cox regression in SCAN-B (N = 3,069; RNA-seq) and validated in four independent cohorts (TCGA-BRCA, METABRIC, GSE20685, GSE1456; total N = 3,536) spanning RNA-seq and microarray platforms, plus five neoadjuvant chemotherapy cohorts for pCR prediction. All analytical parameters were pre-specified and frozen before analysis.

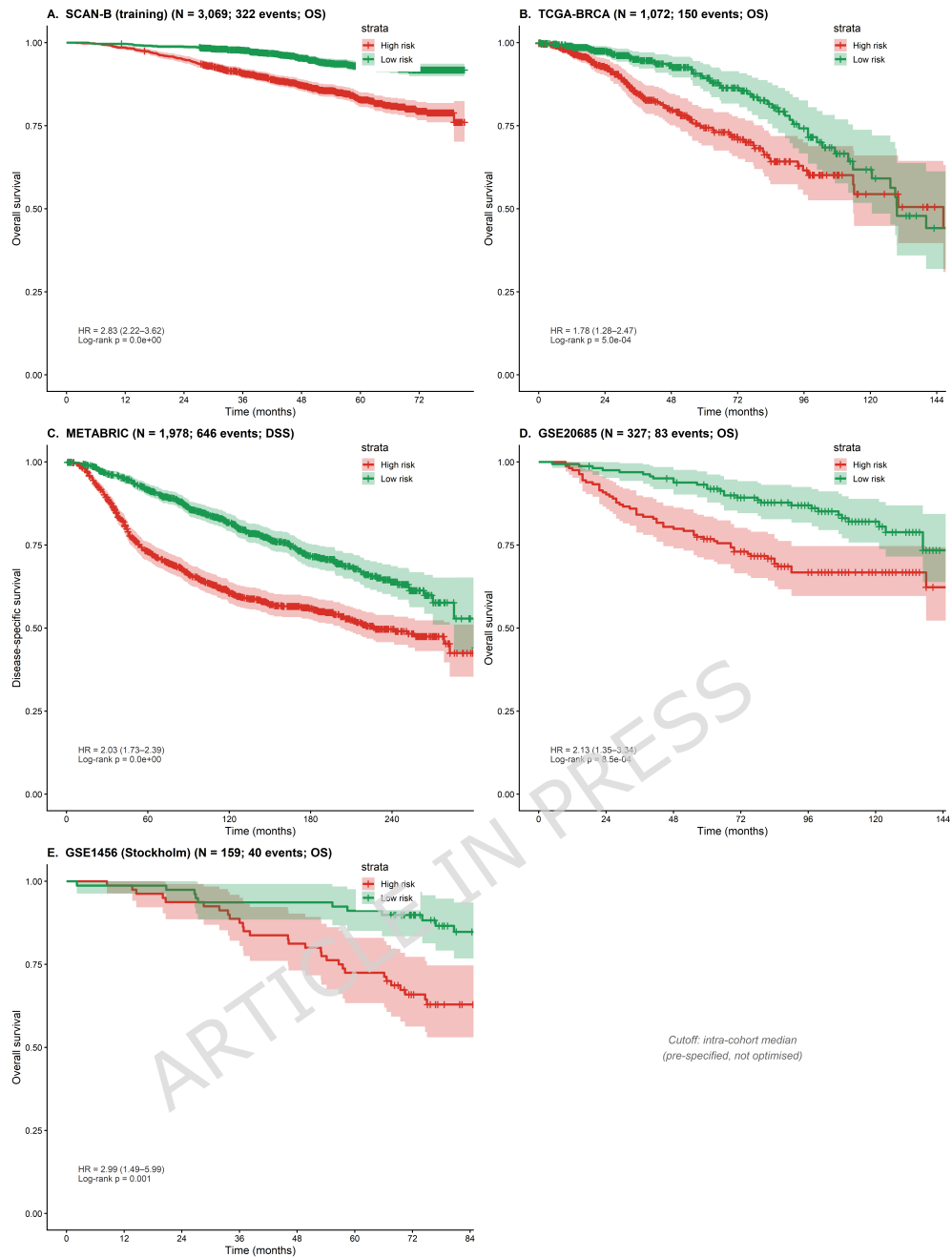


Fig. 2 Kaplan–Meier survival curves across all five cohorts, stratified at the intra-cohort median CorePAM score (High vs Low risk). **A.** SCAN-B training cohort (OS; N = 3,069). **B.** TCGA-BRCA (OS; N = 1,072). **C.** METABRIC (DSS; N = 1,978). **D.** GSE20685 (OS; N = 327). **E.** GSE1456 Stockholm (OS; N = 159). Each panel shows the dichotomised Cox HR with 95% CI and log-rank p -value. Shading: 95% CI. Cutoff: intra-cohort median (pre-specified, not optimised).

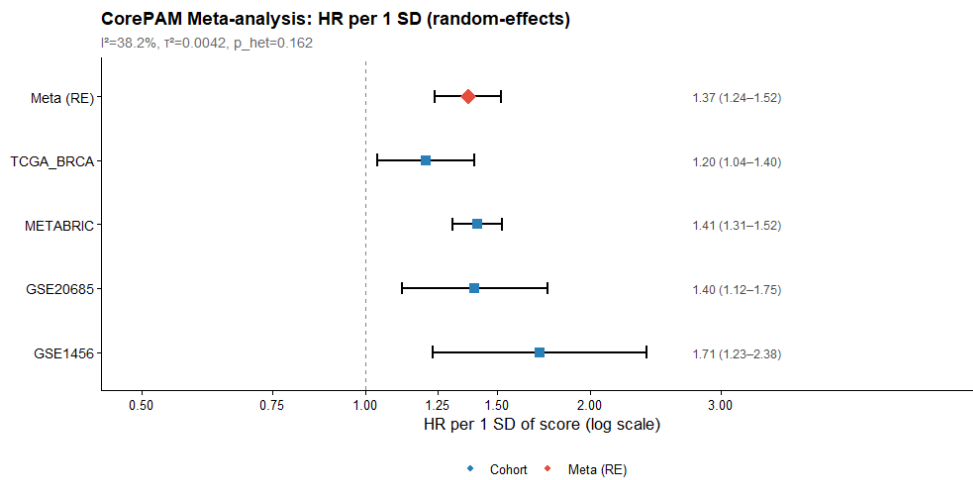


Fig. 3 Forest plot of random-effects meta-analysis. HR per 1 SD CorePAM score across four validation cohorts plus pooled RE estimate ($K = 4$; REML estimator).

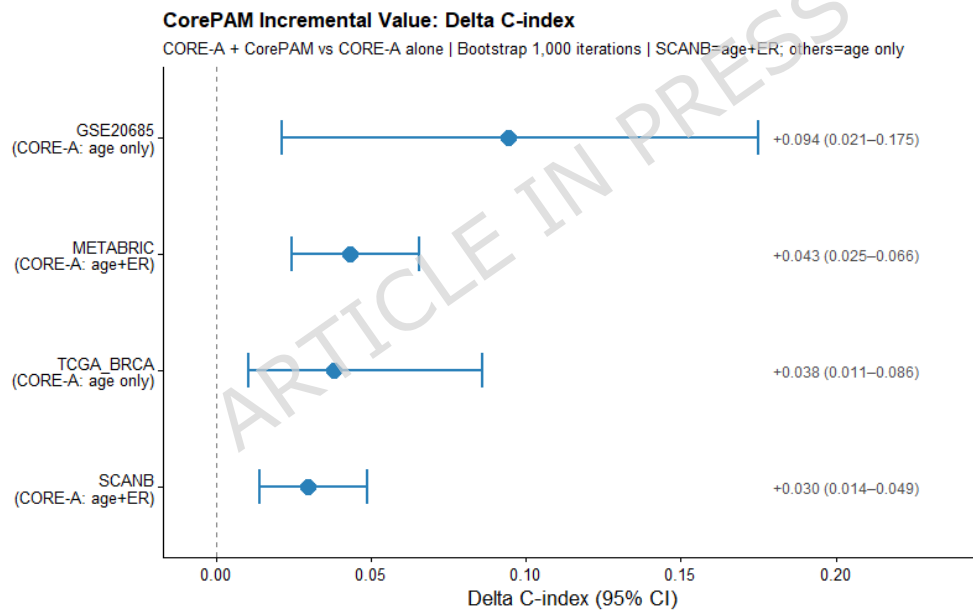


Fig. 4 Incremental C-index (ΔC) of CorePAM score added to the CORE-A clinical model (age $\pm$ ER status) across five cohorts. Error bars: 95% bootstrap CI ($B = 1,000$). CORE-A covariates: age + ER (SCAN-B, METABRIC); age alone (TCGA-BRCA, GSE20685); not available (GSE1456).

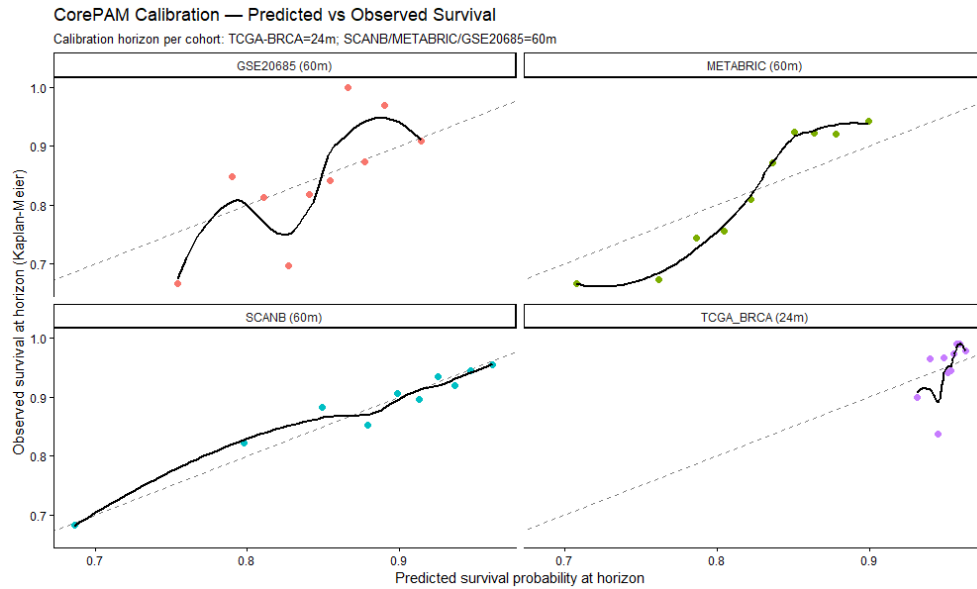


Fig. 5 Calibration curves: predicted versus observed event rates at 60 months (SCAN-B, METABRIC, GSE20685) and 24 months (TCGA-BRCA). Diagonal: ideal calibration. Loess smoother with 95% CI shading.

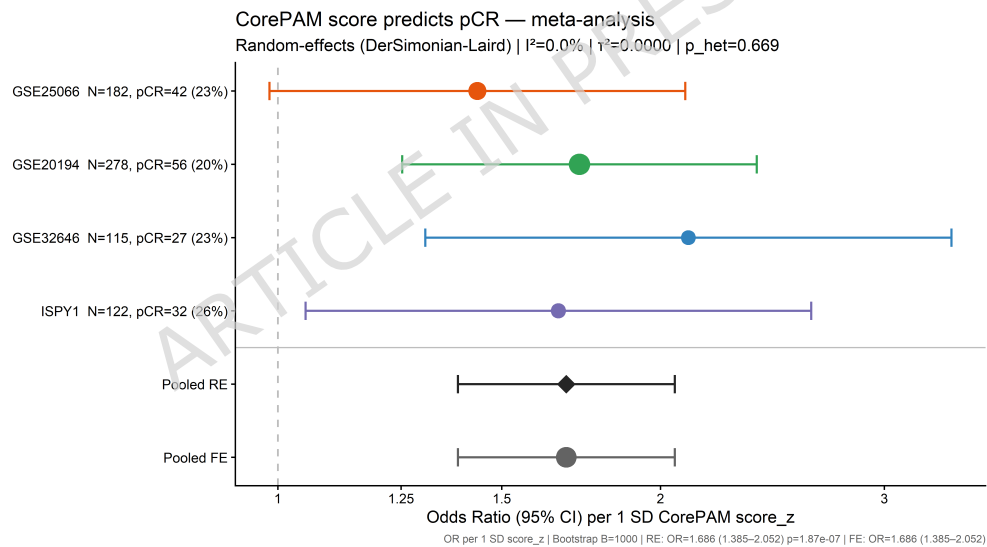


Fig. 6 Forest plot of CorePAM and pathologic complete response (pCR). OR per 1 SD CorePAM score (univariate logistic regression) in four independent neoadjuvant chemotherapy cohorts (GSE25066, GSE20194, GSE32646, I-SPY1; total N = 697), with random-effects pooled estimate (DerSimonian-Laird; $I^2 = 0\%$).

Additional files

Table 7 Additional file 3: Table S1. Expanded CORE-A analysis: CorePAM independence from anatomical staging (T-stage, nodal status). HR per 1 SD CorePAM score from multivariable Cox regression.

Cohort	N	Covariates	HR (95% CI)	<i>p</i>	C clin	C clin+score	ΔC
SCAN-B	2,948	age+ER+T+N	1.79 (1.57–2.05)	<0.001	0.766	0.788	+0.022
TCGA-BRCA	1,054	age+T+N	1.35 (1.15–1.58)	<0.001	0.719	0.746	+0.027
METABRIC	1,883	age+ER+T+N	1.34 (1.23–1.46)	<0.001	0.663	0.688	+0.026
GSE20685	327	age+T+N	1.45 (1.15–1.84)	0.002	0.697	0.724	+0.026

Table 8 Additional file 7:

Table S2. CorePAM gene list with elastic-net Cox regression weights (24 genes). Positive weights: higher expression associated with worse survival; negative weights: protective association.

Gene	Weight	Direction
EXO1	0.2174	Risk (+)
NAT1	-0.2052	Protective (-)
BLVRA	-0.1848	Protective (-)
ACTR3B	-0.1405	Protective (-)
MIA	-0.1193	Protective (-)
MYBL2	-0.1183	Protective (-)
PTTG1	0.1140	Risk (+)
MDM2	-0.1074	Protective (-)
SFRP1	-0.0907	Protective (-)
GPR160	-0.0841	Protective (-)
FOXC1	0.0717	Risk (+)
PHGDH	0.0620	Risk (+)
MYC	-0.0424	Protective (-)
ESR1	0.0396	Risk (+)
KRT5	-0.0277	Protective (-)
CXXC5	0.0266	Risk (+)
PGR	-0.0262	Protective (-)
KRT17	-0.0163	Protective (-)
CENPF	0.0150	Risk (+)
FGFR4	0.0110	Risk (+)
BCL2	-0.0105	Protective (-)
MLPH	-0.0099	Protective (-)
ERBB2	-0.0082	Protective (-)
GRB7	-0.0053	Protective (-)

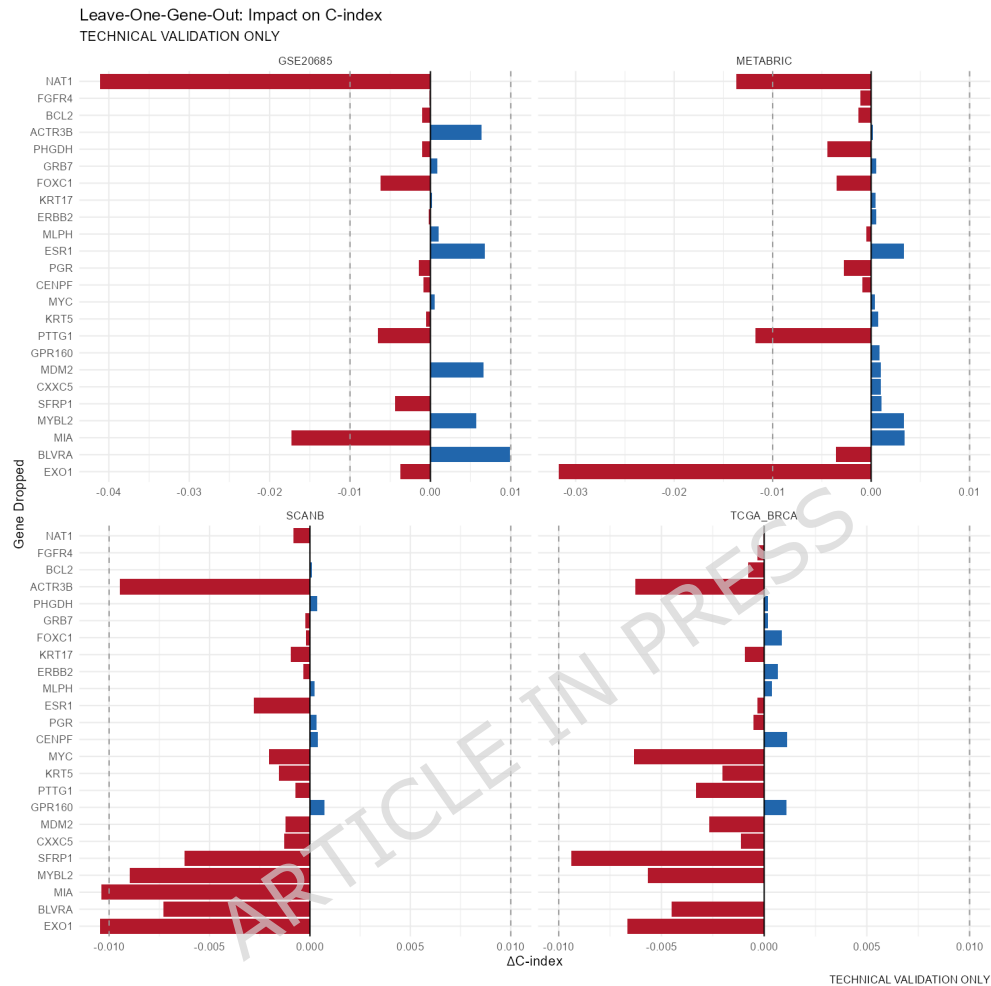


Fig. 7 Additional file 1: Figure S1. Leave-one-gene-out sensitivity analysis. Change in C-index when each CorePAM gene is individually removed.

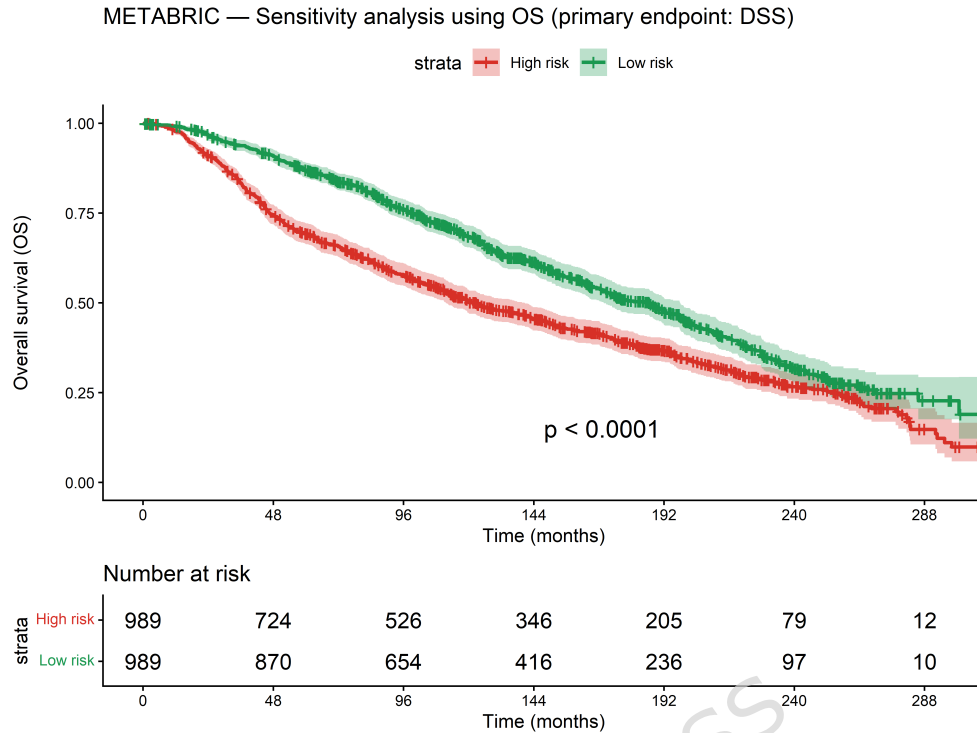


Fig. 8 Additional file 2: Figure S2. METABRIC sensitivity analyses. A: OS endpoint. B: Fine-Gray competing risks model. C: Quartile Kaplan-Meier (DSS).

Table 9 Additional file 12: Table S3. Frozen z-score sensitivity analysis.

C-index and Pearson correlation (r) between frozen and intra-cohort scoring approaches. **Note:** HRs in this table are per unit of raw weighted score (no cohort-level re-standardisation) and are therefore **not directly comparable** to the per-1-SD HRs in Table 3. The relevant comparison metric is the C-index, which is rank-invariant across scoring scales.

Cohort	N	Platform	C frozen	C intra	ΔC	r
SCAN-B	3,069	RNA-seq	0.698	0.698	0.0000	1.000
TCGA-BRCA	1,072	RNA-seq	0.624	0.624	+0.0002	0.999
METABRIC	1,978	Microarray	0.638	0.638	+0.0007	0.958
GSE20685	327	Microarray	0.591	0.626	−0.036	0.923
GSE1456	159	Microarray	0.622	0.661	−0.040	0.916

C frozen/intra: Harrell's C-index under frozen (SCAN-B reference) vs intra-cohort z-score normalisation. $\Delta C = C(\text{frozen}) - C(\text{intra})$. r : Pearson correlation between frozen and intra-cohort scores. C-index is rank-invariant and directly comparable across scoring scales.

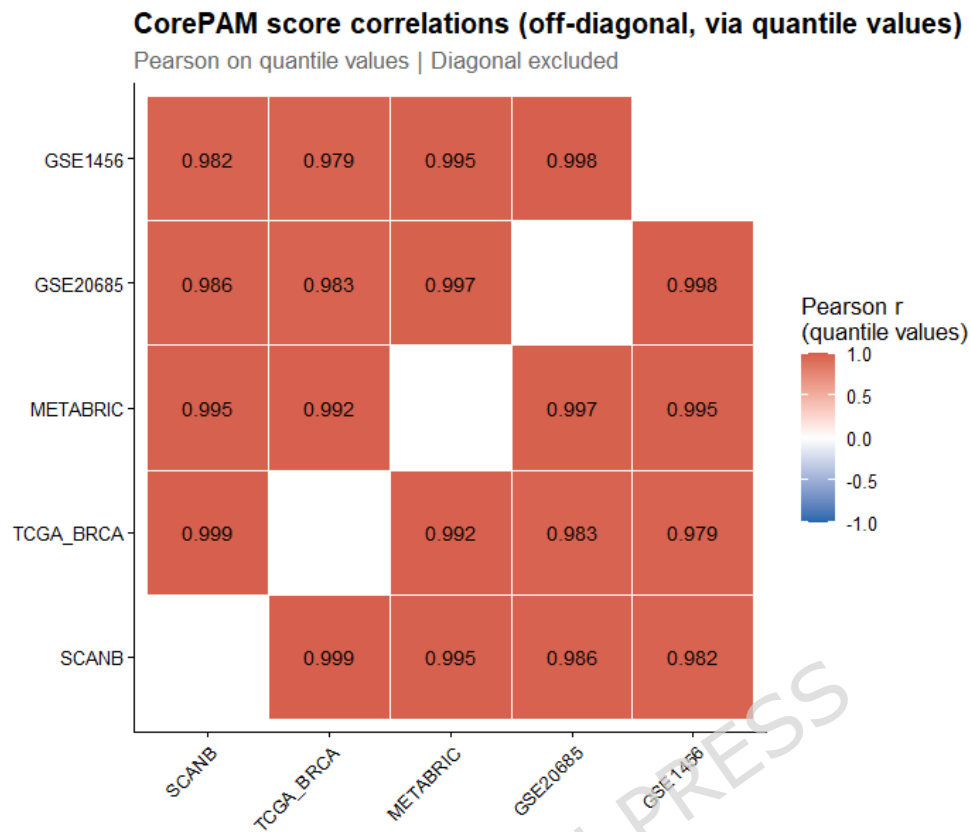


Fig. 9 Additional file 4: Figure S3. Off-diagonal CorePAM score correlations between cohorts. Pearson r computed on paired quantile values.

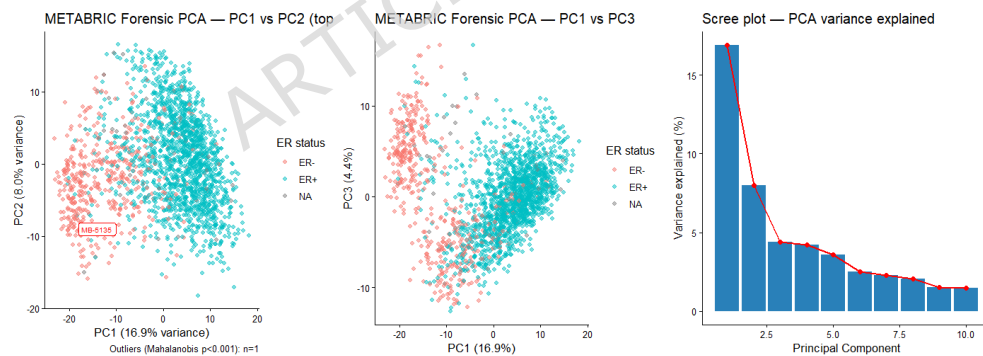


Fig. 10 Additional file 5: Figure S4. Forensic PCA of METABRIC expression matrix. Principal component analysis coloured by ER status.

Core-PAM Panel — Gene Weights

24 genes | OOF non-inferiority ($\Delta C=0.010$) | SCAN-B training | Weights from refit on full training set

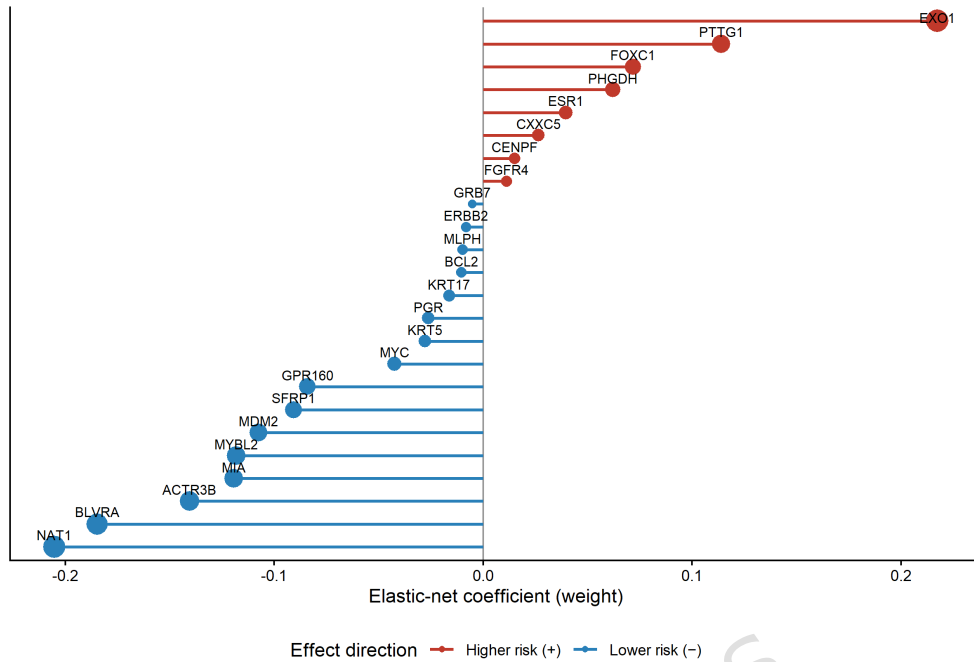


Fig. 11 Additional file 6: Figure S5. CorePAM gene weights (lollipop plot). Elastic-net Cox regression coefficients. Positive weights (red): higher expression associated with worse survival.

Table 10 Additional file 13: Table S4. Uno's C-statistic (IPCW) and time-dependent AUC at cohort-specific horizons.

Cohort	N	Horizon	Uno's C	tdAUC (95% CI)	Harrell's C
SCAN-B	3,069	60m	0.674	0.679 (0.643–0.715)	0.698
TCGA-BRCA	1,072	24m	0.669	0.673 (0.598–0.748)	0.624
METABRIC	1,978	60m	0.681	0.695 (0.665–0.725)	0.638
GSE20685	327	60m	0.668	0.675 (0.596–0.754)	0.626

Table 11 Additional file 14: Table S5. Formal calibration: logistic recalibration intercept (α) and slope (β) at cohort-specific horizons. Ideal: $\alpha = 0$, $\beta = 1$.

Cohort	N	Horizon	α (SE)	β (SE)	E/O	Brier
SCAN-B	3,069	60m	0.56 (0.18)	0.96 (0.09)	0.61	0.158
TCGA-BRCA	1,072	24m	5.40 (2.29)	2.74 (0.79)	0.74	0.061
METABRIC	1,978	60m	1.15 (0.25)	1.77 (0.17)	0.97	0.143
GSE20685	327	60m	0.93 (0.75)	1.57 (0.46)	0.98	0.130

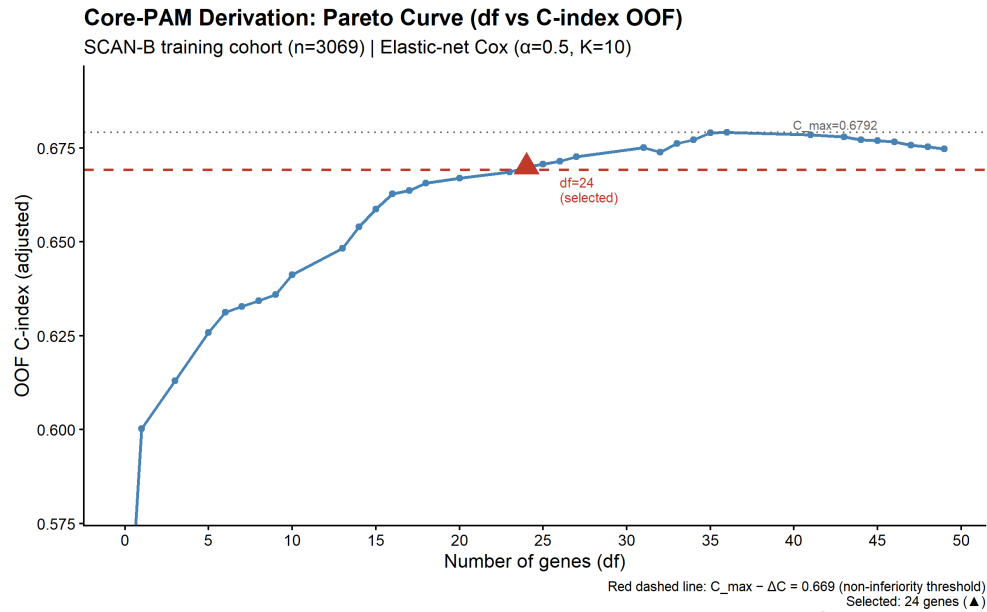


Fig. 12 Additional file 8: Figure S6. Pareto frontier: gene count vs. OOF C-index. Each point represents a unique df level along the glmnet regularisation path. Dashed red line: non-inferiority boundary ($C_{\max} - \Delta C$; $\Delta C = 0.010$).

Table 12 Additional file 15: Table S6. Full clinical baseline incremental value analysis. Cox regression with six clinical-pathological covariates (age, ER, T-stage, N-status, grade, HER2) $\pm$ CorePAM.

Cohort	Population	N	C clinical	C clin+CorePAM	ΔC	HR _{adj}	LRT <i>p</i>
SCAN-B (OS)	All	2,649	0.770	0.787	+0.017	1.75	4.9×10^{-13}
SCAN-B (OS)	ER+	2,450	0.770	0.784	+0.014	1.68	1.1×10^{-10}
METABRIC (DSS)	All	1,792	0.678	0.693	+0.015	1.25	6.1×10^{-6}
METABRIC (DSS)	ER+	1,377	0.674	0.707	+0.032	1.47	1.0×10^{-10}

HR_{adj}: CorePAM HR per 1 SD adjusted for age, ER, T-stage, N-status, grade, HER2. ΔC : bootstrap (B=1,000). LRT: likelihood ratio test for CorePAM addition.

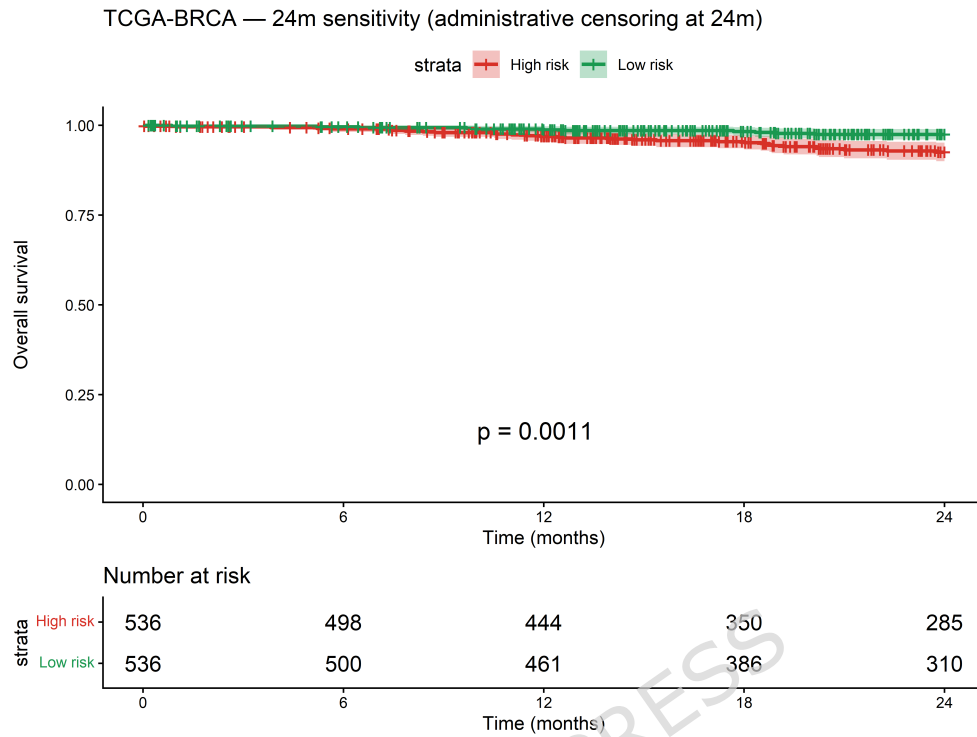


Fig. 13 Additional file 9: Figure S7. TCGA-BRCA 24-month sensitivity analysis.

Table 13 Additional file 17: Table S7. Head-to-head C-index comparison: CorePAM (24 genes) vs research-based ROR-S (50 genes) and OncotypeDX-RS (21 genes). All scores per 1 SD.

Cohort	N	CorePAM C	ROR-S C	ODX-RS C	CorePAM HR	ROR-S HR	ODX-RS HR
SCAN-B	3,069	0.698	0.641	0.607	1.92	1.56	1.40
TCGA-BRCA	1,072	0.624	0.602	0.586	1.20	1.25	1.13
METABRIC	1,978	0.638	0.643	0.665	1.41	1.53	1.51
GSE20685	327	0.626	0.646	0.651	1.40	1.55	1.61

Decision Curve Analysis — OS/DSS

Net clinical benefit of CorePAM vs treat-all and treat-none

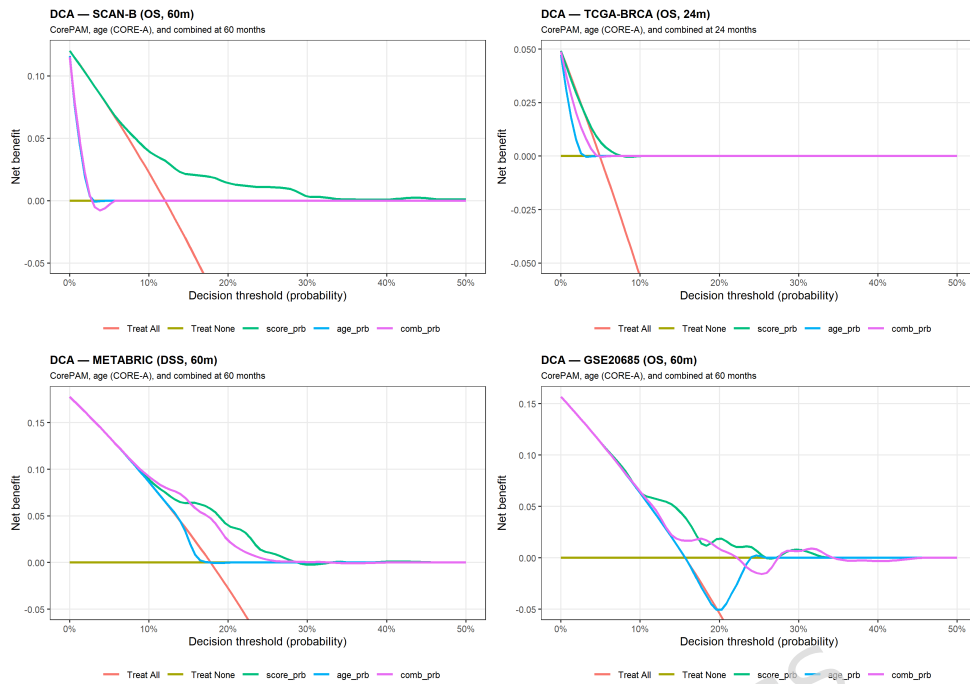


Fig. 14 Additional file 10: Figure S8. Decision curve analysis (OS). Net benefit versus treat-all and treat-none strategies. Horizons: 60 months (SCAN-B, METABRIC, GSE20685) and 24 months (TCGA-BRCA).

Table 14 Additional file 18: Table S8. Head-to-head external comparison: CorePAM (24 genes) vs PAM50-full elastic-net (49 genes). Both models trained in SCAN-B under identical conditions (elastic-net Cox, $\alpha = 0.5$); CorePAM selected at the Pareto-optimal regularisation level, PAM50-full at the maximum-df λ . Scoring modes: intra-cohort z-score (framework-matched) and frozen z-score (deployment-matched, SCAN-B reference).

Cohort	Mode	N	C_{24}	C_{49}	ΔC	95% CI	r
TCGA-BRCA	intra	1,072	0.624	0.631	-0.008	(-0.038, +0.022)	0.853
TCGA-BRCA	frozen	1,072	0.624	0.634	-0.010	(-0.039, +0.020)	0.864
METABRIC	intra	1,978	0.638	0.616	+0.022	(+0.011, +0.032)	0.902
METABRIC	frozen	1,978	0.638	0.622	+0.017	(+0.007, +0.027)	0.914
GSE20685	intra	327	0.626	0.571	+0.055	(+0.010, +0.101)	0.823
GSE20685	frozen	327	0.591	0.544	+0.047	(+0.007, +0.085)	0.843
GSE1456	intra	159	0.661	0.564	+0.097	(+0.032, +0.166)	0.697
GSE1456	frozen	159	0.622	0.510	+0.111	(+0.045, +0.181)	0.617

$\Delta C = C_{24} - C_{49}$; positive values favour the 24-gene model. 95% CI: bootstrap percentile ($B = 2,000$). r : Pearson correlation between the two linear predictors.

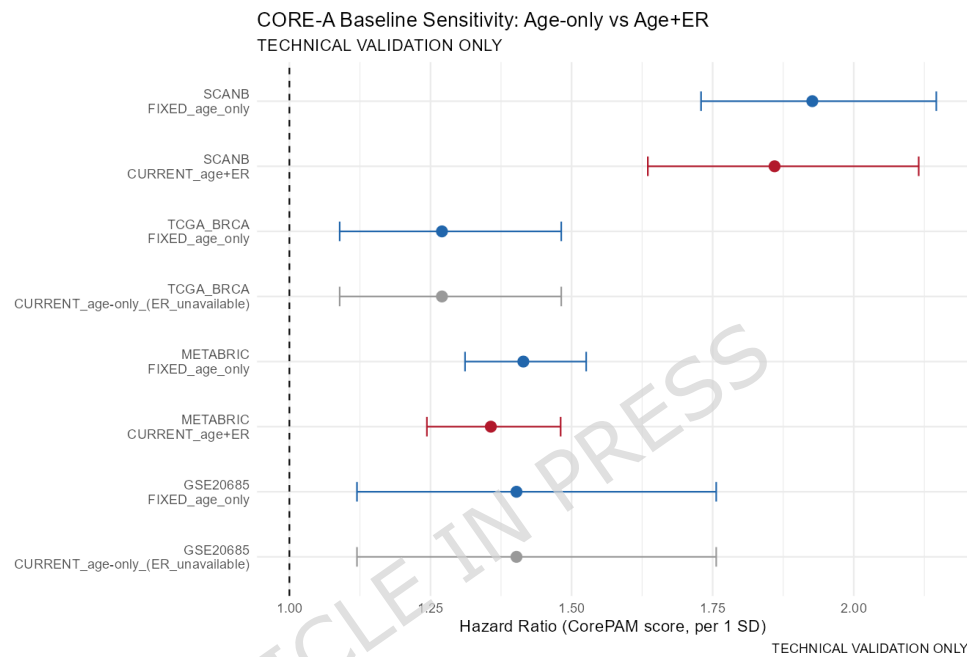


Fig. 15 Additional file 11: Figure S9. CORE-A baseline sensitivity analysis. Forest plot comparing CorePAM adjusted HR under different clinical baselines.

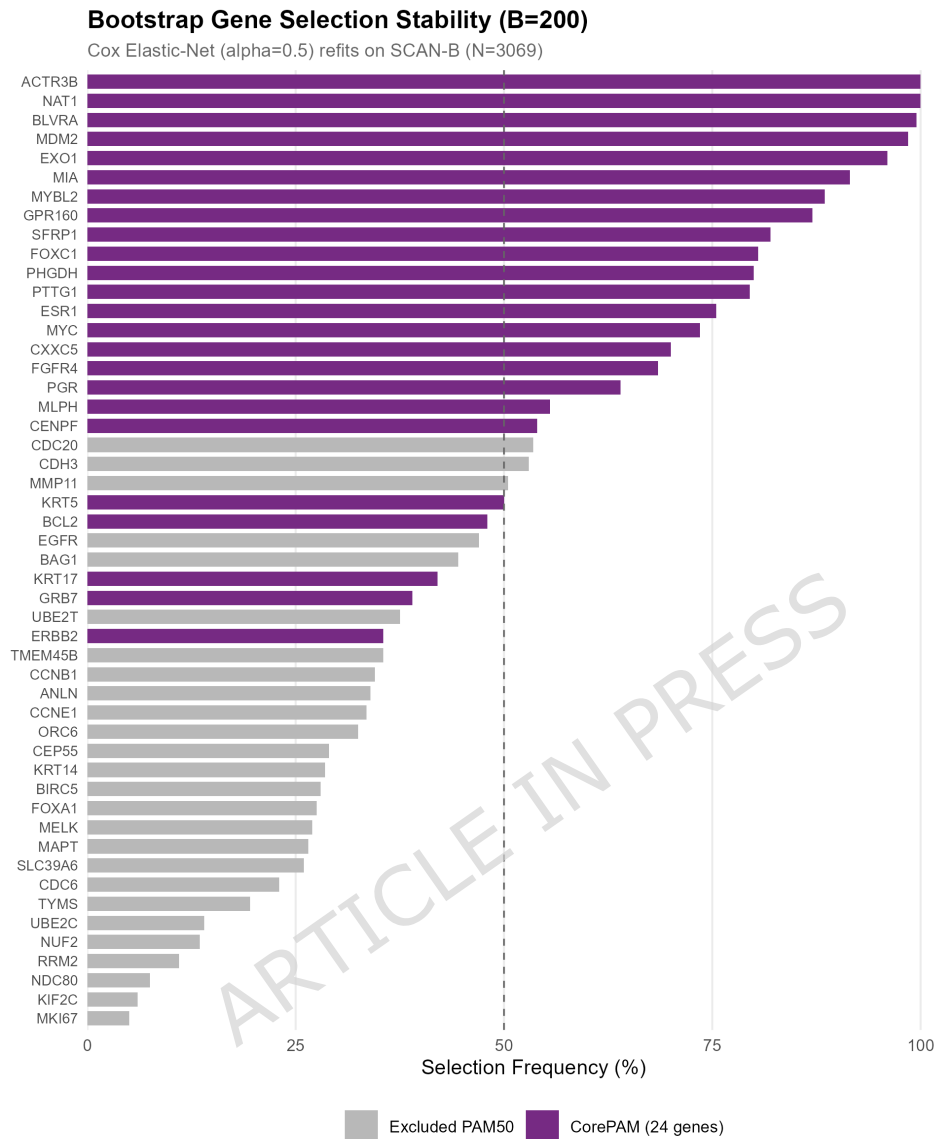


Fig. 16 Additional file 16: Figure S10. Bootstrap stability of CorePAM gene selection ($B = 200$). Frequency of each PAM50 gene across 200 bootstrap refits.

References

- [1] Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406(6797):747–752. <https://doi.org/10.1038/35021093>.
- [2] Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*. 2001;98(19):10869–10874. <https://doi.org/10.1073/pnas.191367098>.
- [3] Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009;27(8):1160–1167. <https://doi.org/10.1200/jco.2008.18.1370>.
- [4] 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(8):717–729. <https://doi.org/10.1056/NEJMoa1602253>.
- [5] Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, et al. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *N Engl J Med*. 2018;379(2):111–121. <https://doi.org/10.1056/NEJMoa1804710>.
- [6] Nielsen TO, Balslev E, Müller S, Krogh M, Shak S, Cardoso F, et al. Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens. *BMC Cancer*. 2014;14:177. <https://doi.org/10.1186/1471-2407-14-177>.
- [7] 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(27):2817–2826. <https://doi.org/10.1056/nejmoa041588>.
- [8] van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415(6871):530–536. <https://doi.org/10.1038/415530a>.
- [9] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1–W73. <https://doi.org/10.7326/M14-0698>.
- [10] Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol*. 2013;13:33. <https://doi.org/10.1186/1471-2288-13-33>.
- [11] Brueffer C, Vallon-Christersson J, Grabau D, Ehinger A, Häkkinen J, Hegardt C, et al. Clinical value of RNA sequencing-based classifiers for prediction of the five conventional breast cancer biomarkers: a report from the population-based multicenter Sweden Cancerome Analysis Network – Breast initiative. *JCO Precis Oncol*. 2018;2:1–18. <https://doi.org/10.1200/po.17.00135>.
- [12] Saal LH, Vallon-Christersson J, Häkkinen J, Hegardt C, Grabau D, Winter C, et al. The Sweden Cancerome Analysis Network – Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. *Genome Med*. 2015;7:20. <https://doi.org/10.1186/s13073-015-0131-9>.
- [13] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;490(7418):61–70. <https://doi.org/10.1038/nature11412>.

- [14] Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. 2012;486(7403):346–352. <https://doi.org/10.1038/nature10983>.
- [15] Kao KJ, Chang KM, Hsu HC, Huang AT. Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. *BMC Cancer*. 2011;11:143. <https://doi.org/10.1186/1471-2407-11-143>.
- [16] Pawitan Y, Bjöhle J, Amler L, Borg ÅL, Egyhazi S, Hall P, et al. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res*. 2005;7(6):R953–R964. <https://doi.org/10.1186/bcr1325>.
- [17] Hatzis C, Pusztai L, Valero V, Booser DJ, Esserman L, Lluch A, et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA*. 2011;305(18):1873–1881. <https://doi.org/10.1001/jama.2011.593>.
- [18] MAQC Consortium. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol*. 2010;28(8):827–838. <https://doi.org/10.1038/nbt.1665>.
- [19] Miyake T, Nakayama T, Naoi Y, Yamamoto N, Otani Y, Kim SJ, et al. GSTP1 expression predicts poor pathological complete response to neoadjuvant chemotherapy in ER-negative breast cancer. *Cancer Sci*. 2012;103(5):913–920. <https://doi.org/10.1111/j.1349-7006.2012.02231.x>.

- [20] Esserman LJ, Berry DA, Cheang MCU, Yau C, Perou CM, Carey L, et al. Chemotherapy response and recurrence-free survival in neoadjuvant breast cancer depends on biomarker profiles: results from the I-SPY 1 TRIAL (CALGB 150007/150012; ACRIN 6657). *Breast Cancer Res Treat.* 2012;132(3):1049–1062. <https://doi.org/10.1007/s10549-011-1895-2>.
- [21] Wolf DM, Yau C, Wulfkühle J, Brown-Swigart L, Gallaher R, Hirst GL, et al. Redefining breast cancer subtypes to guide treatment prioritization and maximize response: predictive biomarkers across 10 cancer therapies. *Cancer Cell.* 2022;40(6):609–623.e6. <https://doi.org/10.1016/j.ccell.2022.05.005>.
- [22] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010;26(1):139–140. <https://doi.org/10.1093/bioinformatics/btp616>.
- [23] Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B.* 2005;67(2):301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>.
- [24] Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med.* 1997;16(4):385–395. [https://doi.org/10.1002/\(sici\)1097-0258\(19970228\)16:4<385::aid-sim380>3.0.co;2-3](https://doi.org/10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3).
- [25] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw.* 2010;33(1):1–22. <https://doi.org/10.18637/jss.v033.i01>.
- [26] Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox’s proportional hazards model via coordinate descent. *J Stat Softw.* 2011;39(5):1–13. <https://doi.org/10.18637/jss.v039.i05>.

- [27] Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA*. 1982;247(18):2543–2546. <https://doi.org/10.1001/jama.1982.03320430047030>.
- [28] Uno H, Cai T, Pencina MJ, D’Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med*. 2011;30(10):1105–1117. <https://doi.org/10.1002/sim.4154>.
- [29] Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat Med*. 2013;32(30):5381–5397. <https://doi.org/10.1002/sim.5958>.
- [30] Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*. 1994;81(3):515–526. <https://doi.org/10.1093/biomet/81.3.515>.
- [31] Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc*. 1999;94(446):496–509. <https://doi.org/10.1080/01621459.1999.10474144>.
- [32] DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177–188. [https://doi.org/10.1016/0197-2456\(86\)90046-2](https://doi.org/10.1016/0197-2456(86)90046-2).
- [33] Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3):1–48. <https://doi.org/10.18637/jss.v036.i03>.
- [34] Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327(7414):557–560. <https://doi.org/10.1136/bmj.327.7414.557>.

- [35] Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Stat Med*. 2001;20(24):3875–3889. <https://doi.org/10.1002/sim.1009>.
- [36] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565–574. <https://doi.org/10.1177/0272989x06295361>.
- [37] Botan RdN.: CorePAM-Study: CorePAM analytical pipeline. Accessed 2026. <https://github.com/RafaelBotan/CorePAM-Study>.
- [38] Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. New York: Springer; 2000.
- [39] Gendoo DMA, Ratanasirigulchai N, Schröder MS, Paré L, Parker JS, Prat A, et al. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics*. 2016;32(7):1097–1099. <https://doi.org/10.1093/bioinformatics/btv693>.
- [40] Prat A, Fan C, Fernández A, Hoadley KA, Martinello R, Vber M, et al. Response and survival of breast cancer intrinsic subtypes following multi-agent neoadjuvant chemotherapy. *BMC Med*. 2015;13:303. <https://doi.org/10.1186/s12916-015-0540-z>.
- [41] Dowsett M, Sestak I, Lopez-Knowles E, Sidhu K, Durbier AK, Cowens JW, et al. Comparison of PAM50 risk of recurrence score with Oncotype DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. *J Clin Oncol*. 2013;31(22):2783–2790. <https://doi.org/10.1200/jco.2012.46.1558>.