Real-world Evidence of Diagnostic Testing and Treatment Patterns in U.S. Breast Cancer Patients with Implications for Treatment Biomarkers from RNA-sequencing Data

Louis E. Fernandes[1]*, Caroline G. Epstein[1]*, Alexandria M. Bobe[1]*, Joshua S.K. Bell[1]*, Martin C. Stumpe[1], Michael E. Salazar[1], Ameen A. Salahudeen[1], Ruth A. Pe Benito[1], Calvin McCarter[1], Benjamin D. Leibowitz[1], Matthew Kase[1], Catherine Igartua[1], Robert Huether[1], Ashraf Hafez[1], Nike Beaubier[1], Michael D. Axelson[1], Mark D. Pegram[2], Sarah L. Sammons[3], Joyce A. O'Shaughnessy[4], and Gary A. Palmer[1]

*co-first authors
[1]Tempus Labs, Chicago, IL, U.S.A.
[2]Stanford University School of Medicine, Stanford, CA, U.S.A.
[3]Department of Medicine, Duke University Medical Center, Duke University, Durham, NC, U.S.A.
[4]Texas Oncology and US Oncology, Dallas, TX, U.S.A.

Corresponding Author
Gary A. Palmer, gary.palmer@tempus.com

Abstract

INTRODUCTION We performed a retrospective analysis of longitudinal real-world data (RWD) from breast cancer patients to replicate results from clinical studies and demonstrate the feasibility of generating real-world evidence. We also assessed the value of transcriptome profiling as a complementary tool for determining molecular subtypes.

PATIENTS AND METHODS De-identified, longitudinal data were analyzed after abstraction from U.S. breast cancer patient records structured and stored in the Tempus database. Demographics, clinical characteristics, molecular subtype, treatment history, and survival outcomes were assessed according to strict qualitative criteria. RNA sequencing and clinical data were used to predict molecular subtypes and signaling pathway enrichment.

RESULTS The clinical abstraction cohort (n=4,000) mirrored U.S. breast cancer demographics and clinical characteristics indicating feasibility for RWE generation. Among HER2+ patients, 74.2% received anti-HER2 therapy, with ~70% starting within 3 months of a positive test result. Most non-treated patients were early stage. In this RWD set, 31.7% of patients with HER2+ IHC had discordant FISH results recorded. Among patients with multiple HER2 IHC results at diagnosis, 18.6% exhibited intra-test discordance. Through development of a whole-transcriptome model to predict IHC receptor status in the molecular sequenced cohort (n=400), molecular subtypes were resolved for all patients (n=36) with equivocal HER2 statuses from abstracted test results. Receptor-related signaling pathways were differentially enriched between clinical molecular subtypes.

CONCLUSION RWD in the Tempus database mirrors the overall U.S. breast cancer population. These results suggest real-time, RWD analyses are feasible in a large, highly heterogeneous database. Furthermore, molecular data may aid deficiencies and discrepancies observed from breast cancer RWD.

Introduction

A growing number of studies have explored real-world data (RWD) and subsequent real-world evidence (RWE) to accelerate treatments for cancer patients. RWD relates to patient information procured during routine care, while RWE is the clinical evidence derived from RWD.[1-2] The feasibility of this approach has increased alongside technological advances and regulatory support to continuously capture and integrate healthcare data sources.[3-7] Several studies demonstrate the ability for RWE to guide clinical development strategies, expand product labels, and address knowledge gaps by examining clinical aspects not captured in clinical trials.[8-21]

An essential step towards strengthening RWE validity is demonstrating consistency between population statistics derived from observational RWD and those from controlled, experimental data. Despite the overwhelming support for RWE utility in oncology, technical barriers must be addressed for RWD/RWE to reach its full clinical potential. Incorporating administrative data, ancillary data, and unstructured clinical text from a variety of institutions to generate RWE is a complex task. For example, no standardization exists for abstracting and structuring highly heterogeneous data sources, and many natural language processing algorithms cannot account for these incongruencies.[2,3] Consequently, clinical endpoints may not be accurately captured[22] and even when data is properly abstracted and prepared for analysis, extraneous variables in raw RWD can introduce confounding biases.[7,23] Similarly, the integration of omics data with RWD requires a controlled approach for large-scale data analytics.[24]

RWE and integrated omics data have the power to impact patient care.[3,25-27] Various studies show the additive value of molecular tumor profiling with RWD for clinically relevant breast cancer insights,[8,28] but further advancements in the field require the integration of genetic and clinical data from a variety of institutions, along with omics-focused capabilities and data analytics. One potential avenue to augment the value of breast cancer RWD is transcriptomics, as RNA-based gene expression analyses have shown prognostic, predictive, and treatment-directing value beyond DNA-sequencing insights.[29-36] Whole-transcriptome RNA sequencing (RNA-seq) can help classify cancer types and breast cancer biomarkers,[37-39] overcoming inconclusive pathology assessments, insufficient tissue quantity, and inter-observer variability of immunohistochemical or in-situ hybridization assays.[39-41]

Here, we address some of the complexities of RWD structuring and analyses. We demonstrate the feasibility of retrospective RWD analysis and test whether results from clinical studies can be replicated using longitudinal RWD from a large, representative breast cancer cohort. Our analyses present key clinical information, such as patient demographics, clinical characteristics, molecular markers, treatment patterns, and overall survival (OS) outcomes; and uncover discrepancies in real-world HER2 testing records. We also provide evidence supporting the integration of RWD with transcriptomic profiling for clinically relevant insights through analyses of RWD and molecular data from breast cancer patients sequenced by Tempus Labs.

Patients and Methods

Cohort Selections

Two retrospective breast cancer cohorts were randomly selected from the Tempus clinicogenomic database after applying clinically relevant inclusion criteria. All data were de-identified in accordance with the Health Insurance Portability and Accountability Act (HIPAA). Dates used for analyses were relative to the breast cancer primary diagnosis (pdx) date, and year of pdx was randomly off-set. Pdx within the cohorts spanned from 1990-2018. The first group was a clinical abstraction (CA) cohort of 4,000 breast cancer patients selected as a representative sample of RWD structured in the Tempus oncology database. Records were required to have data for a pdx, pdx date, age, race, sex, stage, histological subtype, and estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) status. The recorded stage and histological subtype were required to fall within 30 days relative to the pdx date, while the receptor statuses may have been recorded within 30 or 50 days, depending on the testing modality (see methods: *Molecular Subtype Determination*). A second cohort was selected, the Tempus molecular sequenced (MLC) cohort, which included 400 primary breast cancer patients with pdx dates and whose tumor biopsy underwent RNA-seq and targeted DNA sequencing (DNA-seq) with the Tempus xT (n=344), xO (n=55), or xE (n=1) panels between 2017-2019. While only patients with reported variants were included in the cohort, less than 1% of all breast cancer cases in the Tempus database have no DNA variants reported.

The study protocol was submitted to the Advarra Institutional Review Board. The IRB determined the research was exempt from IRB oversight and approved a waiver of HIPAA authorization for this study.

Abstracted Molecular Markers

Protein expression from immunohistochemistry (IHC) results for ER and PR, as well as IHC and fluorescence in-situ hybridization (FISH) results for HER2 were curated during clinical data abstraction. Receptor results included abstracted equivocal, positive, or negative statuses. Hormone receptor (HR) status was classified by combinations of ER and PR statuses. When available, normalized Ki67 results included indeterminant, low, equivocal, moderate, or high statuses. A chi-squared test assessed the significance of Ki67 test result distribution differences. Fisher's exact tests were performed for post-hoc analyses, and *P*-values were adjusted for multiple hypothesis testing using Bonferroni correction.

Molecular Subtype Determination

The molecular subtype of each CA patient was classified as HR+/HER2-, HR+/HER2+, HR-/HER2+, or triple-negative breast cancer (TNBC) based on their receptor statuses at diagnosis. HR statuses were determined from the most recent IHC results or physician notes recorded within 30 days of the pdx date. HR+ status included ER+/PR+, ER+/PR-, and ER-/PR+. HER2 status was determined from the most recent FISH results recorded within 50 days of the pdx date. In the absence of HER2 FISH data, the most recent IHC result or physician note within 30 days of the pdx date was utilized. References to results at "initial diagnosis" imply these 30- and 50-day time frames. Molecular subtypes in the MLC cohort were determined from IHC or FISH results associated with the patient pathology report.

Clinical Data Abstraction

Clinical data were extracted from the Tempus real-world oncology database of longitudinal structured and unstructured data from geographically diverse oncology practices, including integrated delivery networks, academic institutions, and community practices. Many of the records included in this study were obtained in partnership through ASCO CancerLinQ. Structured data from electronic health record systems were integrated with unstructured data collected from patient records via technology-enabled chart abstraction and corresponding molecular data, if applicable. Data were harmonized and normalized to standard terminologies from MedDRA, NCBI, NCIt, NCIm, RxNorm, and SNOMED.

Menopausal Status Determination

Menopausal status was determined using relevant abstracted text fields when available. A patient was considered premenopausal if a single, undated menopause-negative (perimenopausal, premenopausal, or menstruating) status was recorded on or prior to the pdx date and no menopause-positive (menopausal or postmenopausal) status was indicated before diagnosis. Patients were also considered premenopausal at pdx if a menopausal event was recorded after a year from the pdx date.

Likewise, patients with an undated menopause-positive status, and patients with a menopausal or postmenopausal status recorded on or prior to the pdx date, were considered postmenopausal. A patient was also considered postmenopausal if no

menopausal information was available on or prior to the pdx date, but a menopausal or postmenopausal status was indicated within one year after.

Menopausal status circumstances beyond the scope of these criteria were denoted as "Unknown."

Overall Survival Analysis

Overall survival (OS) was calculated for all stage I-IV CA cohort patients with invasive breast cancer (n=3,952). Patients without known relative death dates were right censored at their most recent relative clinical interaction date. Survival curves were generated in R using the survival (v2.43-4) and survminer (v0.4.3) packages with *P*-values calculated by log-rank tests. Results depict the percentage of surviving patients per year, and are stratified based on stage and HER2, ER, and triple-negative status.

Genomic Testing

MLC cohort reported variants were generated from targeted DNA-seq of formalin-fixed, paraffin-embedded (FFPE) slides of primary breast tumor biopsies and, when possible, matched saliva or blood samples. Whole-transcriptome RNA-seq was performed on samples from the same tissue block. Most samples were sequenced with the Tempus xT or xO targeted DNA-seq assays, which detect oncologic targets in solid tumors and hematological malignancies as previously described.[37,42] Two patient samples were sequenced with an updated and refined version of the xT panel targeting clinically relevant exons in 596 genes, and their reported variants were merged for analyses.

Additionally, one sample in the MLC cohort was sequenced with the Tempus xE assay, a whole-exome panel targeting 19,396 genes over a 39 megabase (Mb) genomic region.

Genomic Test Variant Reporting

Because each Tempus assay targets different gene sets, MLC cohort variant analyses only included genes tested across all 400 samples.[42,43] Variants were classified and reported according to previously established clinical guidelines.[37] Reported variants were categorized as alterations, fusions, or copy number variation amplifications or deletions. Alterations include variants of unknown significance (VUS), biologically relevant or potentially actionable alterations, and both germline VUS and pathogenic variants.

Tumor mutational burden (TMB)

TMB was calculated by dividing the number of non-synonymous mutations by the adjusted panel size of the xT, xO, or xE assay (2.4 Mb, 5.86 Mb, and 36 Mb, respectively). All non-silent somatic coding mutations, including missense, indel, and stop-loss variants with coverage greater than 100x and an allelic fraction greater than 5% were counted as non-synonymous mutations.

RNA-based Prediction of Molecular Subtypes

Transcriptome models were used to predict receptor statuses for the MLC cohort, including patients lacking IHC or FISH data. Briefly, single-gene logistic models were

trained on an independent set of Tempus RNA-sequenced breast cancer samples according to the normalized gene expression of *ESR1*, *PGR*, or *ERBB2* using the R glm package v2.0-16. Model performances were assessed separately for primary samples, metastatic samples, and a combined set using 10-fold cross-validation (**Supplemental Table 2**). Performance was evaluated on a testing set comprised of RNA-sequenced samples in the MLC cohort with abstracted IHC or FISH results in the Tempus database (ER n=308, PR n=306, HER2 n=261). These samples were withheld from the training set. Positivity thresholds for IHC prediction models were selected using Youden's J statistic to optimize sensitivity and specificity.

Gene Expression Collection, Processing, and Normalization

Gene expression was generated through RNA-seq of FFPE tumor samples using an exome capture-based protocol.[37] Transcript-level quantification to GRCh37 was performed using Kallisto 0.44. Transcript counts were then corrected for GC content and length using quantile normalization and adjusted for sequencing depth via a size factor method. Normalized counts in protein coding transcripts covered by the exome panel were then summed to obtain gene-level counts. Subsequent expression analyses were performed on $\log_{10}$-transformed counts.

RNA-seq Pathway Analyses

Gene sets were downloaded from the MSigDB website (http://software.broadinstitute.org/gsea/msigdb/index.jsp), and pathway enrichment scores were calculated from normalized gene expression using the ssGSEA function in

Gene Set Variation Analysis (GSVA) R Bioconductor package v1.0.6.[37,44] ER- and HER2-related pathways were identified as those containing the terms "*ESR1*" or "Estrogen" and "*ERBB2*" or "HER2," respectively. Z-scores were calculated for each set of enrichment scores and the sign was reversed for any pathway containing "DN" (down) or "repressed." For select analyses, the mean of the z-score across pathways was calculated to produce a patient pathway metascore. With the exception of the HER2 and ER signaling pathway metascore analyses, receptor status was derived from both abstracted and predicted protein expression. Significance was determined by a Wilcoxon test for any comparison between two groups, and a Kruskal-Wallis test for comparisons between three or more groups, with $P<0.05$ considered significant. A separate gene set analysis was conducted to test the difference in enrichment among the four molecular subtypes relative to the 50 Hallmark pathways, a highly curated list from the MSigDB database.[45] To determine how patients clustered by pathway scores, we performed a second UMAP analysis with enrichment scores for each Hallmark pathway as features.

Results

Real-world evidence from a clinical abstraction breast cancer cohort

Patient demographics and clinical characteristics in the CA cohort

We first determined whether key demographic and clinical characteristics captured in RWD replicate clinical studies, and found the deidentified data were consistent with

previous large-scale breast cancer cohort studies (**Table 1**).[46-49] The cohort mostly comprised females (99.3%, n=3,970) with a median age at diagnosis of 61.0 years. Year of diagnosis among the cohort ranged from 1990 to 2018 (**Supplemental Fig. 1**). The self-reported race was 83.3% White (n=3,332), 13.1% Black or African American (n=523), and 3.6% Asian or Pacific Islander (n=145). In 2,042 females with menopausal data, 87.4% (n=1,784) were postmenopausal. Abstracted stage at initial diagnosis primarily consisted of stage I (49.6%, n=1,986) and II (33.3%, n=1,333), followed by III (10.5%, n=420), IV (5.5%, n=219), and 0 (1.1%, 42). Most tumors had a histological classification of invasive ductal carcinoma (77.4%, n=3,095), and 9.5% (n=378) had an invasive ductal component or were NOS. Several rare cancer types were also represented.

Molecular subtype determination in the CA cohort

We assessed the extent to which RWD captures molecular marker information from clinical testing results. The distributions of all abstracted receptor testing results at initial diagnosis are shown in **Fig. 1A**. Consistent with previous U.S. breast cancer statistics,[50] the most prevalent molecular subtype was HR+/HER2- (71.5%, n=2,859), followed by TNBC (12.3%, n=491) (**Fig. 1B**). Among HR+ patients with non-equivocal statuses, most were ER+/PR+ (71.0%, 2,839 of 3,996) followed by ER+/PR- (10.4%, n=417) and ER-/PR+ (1.4%, n=57) (**Fig. 1C**). Lastly, abstracted Ki67 IHC test results were consistent with the Ki67 expression levels typically indicative of specific breast cancer subtypes (**Fig. 1D**).[51-52] The distribution of Ki67 results differed significantly among molecular subtypes (chi-squared, $P=1.75\text{x}10^{-9}$), particularly between HR+/HER2- versus

HR+/HER2+ patients (P=0.015) and TNBC versus HR+/HER2- patients (P=6.38x10^{-9}). The largest proportions of high Ki67 IHC test results were in TNBC (82.0%, n=50 of 61) and HR-/HER2+ patients (75.0%, n=15 of 20), while most low Ki67 results were in HR+/HER2- patients (44.0%, n=140 of 318).

Anti-HER2 therapy analysis in the CA cohort

We next examined anti-HER2 therapy treatment patterns from longitudinal RWD. Curated anti-HER2 therapies included trastuzumab, ado-trastuzumab emtansine, neratinib, lapatinib and pertuzumab. Among CA cohort patients, 13.7% (n=546) were HER2+ at initial diagnosis, of whom 74.2% (n=405) received anti-HER2 therapy at some point in their clinical care. Approximately 70.0% of patients who received anti-HER2 therapy did so within 3 months of a positive test result and the majority (73.5%) had early-stage cancer (**Fig. 2A**). These results are consistent with previous breast cancer cohort studies.[16,53] Moreover, a small portion of HER2- patients exhibited evidence of receiving an anti-HER2 therapy (1.1%, 36 of 3,352 HER2- patients) (**Fig. 2B**). Of those patients, 33.3% (n=12) had evidence of a discordant result at initial diagnosis, 44.4% (n=16) had only HER2- results, and 22.2% (n=8) had a HER2-equivocal or positive result recorded beyond initial diagnosis. A small portion of patients (n=37) were not assigned a HER2 treatment time frame due to date quality issues.

HER2 test result analyses in the CA cohort

To evaluate inter- and intra-test concordance, we compared HER2 IHC and FISH results among patients with both tests conducted near initial diagnosis (17.7%, n=709).

Among patients with HER2+ IHC results and subsequent FISH testing, 62.2% (n=51 of 82) were inter-test concordant (**Supplemental Table 1**), however, 31.7% with HER2+ IHC were HER2- by FISH (n=26 of 82). This discordance is larger than a previously reported meta-analysis of IHC and FISH HER2 testing worldwide.[54] Four of those 26 patients had received an anti-HER2 therapy in their clinical timeline. Among patients with HER2- IHC results, 3.9% (n=7 of 182) were HER2+ by FISH, similar to historical reports.[54] The majority of these patients (n=6 of 7) received anti-HER2 therapy. HER2-equivocal IHC results (HER2 IHC 2+) were observed in 62.8% (n=445 of 709) of the cohort. Among these patients with equivocal results, 7.8% (n=35 of 445) were later confirmed equivocal by FISH testing. However, 80.7% (n=359 of 445) had subsequent HER2- and 11.5% (n=51 of 445) HER2+ FISH results.

Additionally, intra-test discordance was analyzed in patients with multiple HER2 results at initial diagnosis. Among patients with multiple HER2 IHC results at diagnosis (7.1%, n=253 of 3,561 with HER2 IHC), 18.6% (n=47) exhibited intra-test discordance. Of patients with multiple HER2 FISH results (4.5%, n=52 of 1,157), 21.2% (n=11) exhibited intra-test discordance.

Overall survival in the CA cohort

OS analyses from longitudinal RWD revealed overall 5-year and 10-year survival rates (92.2% and 85.7%, respectively) relatively consistent with average U.S. percentages (**Fig. 3A**).[46] Survival rates were expectedly high, varying by stage ($P<0.0001$) and receptor status. Stage IV patients exhibited worse OS than earlier-stage patients (**Fig. 3B**). The 5-year survival rate was 93.5% in stage I-IV HER2+ patients and 92.0% in

HER2- patients (*P*=0.45), with rates of 74.3% and 57.1%, respectively, among stage IV patients (*P*=0.098) (**Fig. 3C, 3D**). The 5-year survival rate was 92.7% among stage I-IV ER+ patients and 89.8% in ER- patients (*P*=0.052), with rates of 63.7% and 55.5%, respectively, among stage IV patients (*P*=0.12) (**Fig. 3E, 3F**). TNBC patients had significantly worse OS compared to other subtypes, with a 36.3% 5-year survival rate in stage IV TNBC patients compared with 65.1% among stage IV non-TNBC patients (*P*=0.0024) (**Fig. 3G, 3H**).

Genomic testing insights from the Tempus molecular sequenced cohort

Patient demographics and clinical characteristics in the MLC cohort

Abstracted clinical characteristics and patient demographics from the 400 MLC cohort patients were assessed (**Table 1**), and found to be relatively consistent with the CA cohort and other large-scale breast cancer cohort studies.[46-49] The cohort had a slightly younger median age at diagnosis of 55.8 years (45.2-66.4), and higher percentage of Black or African American (14.6%, n=35) and Asian or Pacific Islander patients (5.4%, n=13) than the CA cohort. Patients with known stage information were mostly stage II at diagnosis (38.4%, n=83), followed by stages IV (26.4%, n=57), III (21.3%, n=46), and I (13.9%, n=30), indicating an overall higher risk population compared with the CA cohort. A total of 75.0% (n=267) of tumors were invasive ductal carcinoma, with several rare cancer types also represented in the cohort.

DNA sequencing analysis of the MLC cohort

The top three genes with reported alterations were *TP53*, *PIK3CA*, and *GATA3*, which were found in 55.0% (n=220), 29.0% (n=116), and 13.8% (n=55) of the MLC cohort, respectively (**Supplemental Fig. 2A**). These findings are consistent with a previous analysis of The Cancer Genome Atlas breast cancer data.[55] **Supplemental Fig. 2B** shows the distribution of variant types in the 20 most frequently reported genes. Assessment of patients with tumor/normal-matched DNA-seq (n=356) identified 18 patients (5.1%) with pathogenic germline variants in 12 NCCN-designated familial high-risk genes (**Supplemental Fig. 2C**). This sub-population may be underrepresented as exon-level duplications or deletions were not included. Among the 18 patients harboring a pathogenic germline variant in any of those 12 genes, most contained variants in *BRCA 1* or *2* (n=11), followed by *CHEK2* (n=6), *ATM* (n=3), and *PALB2* (n=2). Because TMB and MSI status are integrated biomarker measurements in the Tempus platform, we observed a wide range of TMB across the cohort with a median of 1.7 mutations/Mb (**Supplemental Fig. 2D**). Consistent with previous studies,[56] the majority of patients (84.7%, n=339) were MSI stable, while only 0.3% (n=1) were MSI high and 0.5% (n=2) were MSI low.

RNA-based prediction of receptor status for molecular subtypes

We developed a whole-transcriptome model based on 19,147 genes to predict IHC receptor status and resolve molecular subtypes in the MLC cohort. Predicted RNA-based subtypes largely aligned with abstracted IHC-based subtypes (**Fig. 4A**). Similar to the literature,[57-58] transcriptome signatures differed between molecular subtypes with TNBC clustering separately. Seventeen samples clustered with TNBC but were

predicted or abstracted as another subtype, suggesting samples that cluster outside of their groups may benefit from further testing or analysis. *ESR1*, *PGR*, and *ERBB2* gene expression correlated with their respective abstracted and predicted receptor statuses (**Fig. 4B**).

RNA-based receptor status predictions were highly accurate for ER (95.5%, AUROC 98.1%) and HER2 (94.6%, AUROC 93.8%) relative to abstracted status, while PR status was predicted with slightly lower accuracy (87.9%, AUROC 95.2%) (**Fig. 5**). Prediction accuracy for all receptors was 92.7%. A detailed overview of the validation data and model performance are available in **Supplemental Table 2.** Patients with incompletely abstracted molecular subtypes (n=150) were classified by predicted receptor statuses from the transcriptomic model. Importantly, patients with equivocal HER2 statuses abstracted from IHC and/or FISH results (n=36) were predicted HER2+ (n=7) or HER2- (n=29) by the model.

RNA-based HER2 and ER pathway analyses

To further evaluate the potential for RNA-seq to enhance breast cancer clinical data, a gene set enrichment analysis was conducted using the MSigDB database. First, we assessed whether measuring the activity of signaling pathways may resolve ambiguous or equivocal IHC and FISH test results. Multiple gene sets that putatively measure such pathway activity were identified by searching for "*ERBB2*," "HER2," "*ESR1*," or "Estrogen" in the MSigDB database (**Supplemental Fig. 3A and 3B**). Results of the pathway analyses were expressed as metascores to avoid the bias introduced when selecting a single pathway. HER2 IHC-positive and FISH-positive samples were

enriched for HER2 activity metascores as expected, but the HER2 signaling results contained substantial variability in pathway activity (**Fig. 6A**). Notably, the GO_ERBB2_SIGNALING_PATHWAY, which directly measures HER2 activity,[59] exhibited a robust correlation with HER2 expression (r=0.453) (**Supplemental Fig. 3A)** and significantly different enrichments between HER2 statuses (*P*=0.00031) **(Supplemental Fig. 5**). While ER enrichment scores were more distinct between IHC-positive and IHC-negative patients, consistent with the relatively higher reliability of ER IHC compared with HER2 tests,[60-62] variability was also observed in the ER signaling results (**Fig. 6B, Supplemental Fig. 6**).

Next, RNA-seq data were analyzed in relation to the highly curated Hallmark pathway gene sets to determine the differential activation of biological pathways between breast cancer subtypes.[45] Most Hallmark pathways (32 of 50) exhibited significantly different enrichment scores between molecular subtypes (**Supplemental Fig. 4A**). A UMAP using only scores from these 50 pathways recapitulated the TNBC clustering observed in the full-transcriptome UMAP (**Fig. 6C**). As expected, HR+ samples, but not HR- or TNBC samples, were highly enriched for two pathways related to estrogen signaling (**Supplemental Fig. 4B**). Among HR-/HER2+ cancers, we observed enrichment for pathways known to be downstream of HER2, RAS, and mTOR (**Fig. 6D**).[63] HER2-driven tumors also showed enrichment for all immune-related Hallmark pathways, a finding consistent with the literature.[64] Many oncogenic signaling pathways were enriched in TNBC (**Fig. 6E**), including Wnt, mTOR, PI3K, Hedgehog, and Notch, consistent with TNBC tumors' reliance on ER-, PR-, and HER2-independent pathways.[65] TNBC samples were also enriched for pathways related to mitotic index, as

expected due to their relatively high growth rate,[66] glycolysis, which is consistent with their elevated Warburg effect and potentially targetable,[67] and cancer/testis antigens.[68]

Discussion

The expanding utility of RWE is evident with the growing number of related studies and regulatory considerations.[2,3,7,69] Compared with randomized controlled trials, however, RWD analyses are complicated by a lack of standardization between records and the introduction of extraneous factors, such as natural language processing errors and uncontrolled confounding variables.[2,3,7,22-24] We aimed to address these concerns by 1) increasing the statistical power of analyses with a relatively large cohort size, 2) incorporating a variety of data sources beyond electronic health records to benefit downstream analyses,[1,3,22,27] and 3) demonstrating consistency between characteristics of the real-world cohort and results from previous clinical studies.

Using only a portion of breast cancer patient records from the extensive Tempus clinicogenomic database, our retrospective analysis provides further evidence for the feasibility and value of generating clinically relevant RWE. We first demonstrated that longitudinal RWD can capture key information regarding patient clinical history, treatment journey, and outcomes. Our RWD analyses generated valid RWE that replicated previously published clinical results and was generally consistent with established databases, indicating feasibility. Although the majority of cohort characteristics were aligned with previous clinical studies, the analyses also highlighted the complexities in breast cancer RWD. For instance, the proportion of pre- and post-

menopausal patients was similar to previous clinical trial data,[49] but menopausal status was only confidently abstracted in approximately 51% of the cohort. Upon further review, many RWD breast cancer studies have either applied simplified definitions of menopause, such as an age cutoff,[19] reported missing statuses in electronic records,[8,70] or did not include menopausal status at all. Simplifying rules for abstraction may fill these gaps in RWD, such as in defining real-world progression-free survival, but can also affect the validity of conclusions.[71,72]

To strengthen the validity of RWE presented here, rules were established and applied to perform relevant analyses and derive statistics from the cohort. For example, rules described in the methods facilitated the definition of molecular subtypes from multiple abstracted test results. Our HER2 test result analyses confirm the existing conflict in standard testing interpretations, an issue evident by recent American Society of Clinical Oncology (ASCO) guidelines, previous clinical studies, and meta-analyses.[54,73-76] Specifically, our findings of IHC intra-test discordance illustrate the subjectivity of IHC testing, prompting standard testing improvements and biomarker discovery.

Upon observation of discrepancies in abstracted HER2 testing results, a separate cohort with complete biopsy data was selected to test the efficacy of a whole-transcriptome model in predicting molecular subtypes. By combining clinical and molecular data, we demonstrate transcriptome profiling is complementary to RWD and can illuminate fundamental biological differences between patients. RNA-seq may supplement standard testing interpretations by providing clinically relevant insights when biopsy test data is inconclusive, exemplified here in the resolution of molecular

subtypes for patients with equivocal statuses. Addressing these cases is critical in the evolving treatment landscape, as molecular subtypes are key criteria for breast cancer treatment decisions.

Furthermore, our signaling pathway investigation uncovered potential pathway-related therapeutic targets, such as oncogenic signaling via the mTOR pathway, for subtypes like TNBC with limited pharmacotherapies available. RNA pathway analyses can also elucidate treatment-related tumor characteristics not captured by standard diagnostic and prognostic tests, such as additional biomarkers or amplifications that may be targetable in HER2+ breast cancer patients.[77-79] Expression-based immune signatures can also predict response to neoadjuvant treatment with several experimental agents/combinations added to standard chemotherapy, including the addition of pembrolizumab in early-stage TNBC.[80] Biomarker selection of immunotherapy in early-stage TNBC will become imperative to therapeutic strategies given its substantial toxicity.

Conclusion

The Tempus data pipeline integrates longitudinal RWD and comprehensive molecular sequencing data into a structured clinicogenomic database capable of generating valid clinical evidence in real-time. While RWD are inherently complex, cancer cohort selection and data insights are feasible using structured data sources and strictly defined analysis criteria. Finally, integrating RNA-seq data with RWD can improve clinically actionable evidence related to clinical markers, potential therapeutic targets, and optimal therapy selection in breast cancer.

Clinical Practice Points

- The feasibility of real-world data (RWD) analysis has increased alongside technological advances and regulatory support to continuously capture and integrate healthcare data sources. Several studies demonstrate the ability for real-world evidence (RWE) to guide clinical development strategies, expand product labels, and address knowledge gaps by examining clinical aspects not captured in clinical trials.
- Despite recent advances and growing regulatory support, RWD from heterogenous structured and unstructured sources is often challenged by various technical barriers. Lack of standardization between electronic records, underpowered natural language processing tools, and uncontrolled extraneous variables threaten the validity of well-sourced RWE.
- Our RWD analyses followed strict qualitative criteria to produce RWE of demographics, clinical characteristics, molecular subtype, treatment history, and survival outcomes from a large, heterogeneous database. Importantly, the results were mostly consistent with data from previous clinical studies, suggesting feasibility of generating valid RWE. We also demonstrate the value of integrating omics data with RWD through the use of whole-transcriptome analyses in relevant breast cancer signaling pathways and a predictive model for receptor statuses.

- These data provide rational for use of the Tempus clinicogenomic database to generate RWE and conduct real-time, hypothesis-driven analyses of large RWD cohorts in the future. Clinicians may utilize these large-scale databases to circumvent the restrictive exclusion criteria of controlled studies, clarify real-world patient needs, and aid the development of clinical trials. Furthermore, our results suggest molecular data may bolster deficiencies in standard breast cancer diagnostic tests.

Acknowledgments

We are thankful to ASCO CancerLinQ for their partnership and the clinical data, operations, data science, engineering, pathology, and lab teams at Tempus Labs. We sincerely thank the entire clinical data abstraction team, Jeff Ottens, and April Manhertz. We thank Kelly McKinnon for proofreading and figure assembly and design. We thank Kevin White and Kimberly Blackwell for scientific review and discussion of the manuscript. We thank Hailey Lefkofsky for initial discussions and Eric Lefkofsky for his support and discussions.

References

1. Nabhan C, Klink A, Prasad V: Real-world Evidence—What Does It Really Mean? JAMA Oncology 5:781, 2019
2. Administration USFD: Framework for the FDA's Real-world Evidence Program. 2018
3. Jourquin J, Reffey SB, Jernigan C, et al: Susan G. Komen Big Data for Breast Cancer Initiative: How Patient Advocacy Organizations Can Facilitate Using Big Data to Improve Patient Outcomes. JCO Precision Oncology:1-9, 2019
4. Administration USFD: Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drugs and Biologics Guidance for Industry, 2019
5. Administration USFD: Use of Electronic Health Record Data in Clinical Investigations: Guidance for Industry. 2018
6. Administration USFD: Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices: Guidance for Industry and Food and Drug Administration Staff, 2017
7. Khozin S, Blumenthal GM, Pazdur R: Real-world Data for Clinical Evidence Generation in Oncology. JNCI: Journal of the National Cancer Institute 109, 2017
8. Quek RGW, Mardekian J: Clinical Outcomes, Treatment Patterns, and Health Resource Utilization Among Metastatic Breast Cancer Patients with Germline BRCA1/2 Mutation: A Real-World Retrospective Study. Advances in Therapy 36:708-720, 2019
9. Taylor-Stokes G, Mitra D, Waller J, et al: Treatment patterns and clinical outcomes among patients receiving palbociclib in combination with an aromatase inhibitor or fulvestrant for HR+/HER2-negative advanced/metastatic breast cancer in real-world settings in the US: Results from the IRIS study. The Breast 43:22-27, 2019
10. Khozin S, Carson KR, Zhi J, et al: Real-World Outcomes of Patients with Metastatic Non-Small Cell Lung Cancer Treated with Programmed Cell Death Protein 1 Inhibitors in the Year Following U.S. Regulatory Approval. The Oncologist 24:648-656, 2019
11. Martina R, Jenkins D, Bujkiewicz S, et al: The inclusion of real world evidence in clinical development planning. Trials 19, 2018
12. Harrell M, Fabbri D, Levy M: Analysis of Adjuvant Endocrine Therapy in Practice From Electronic Health Record Data of Patients With Breast Cancer. JCO Clinical Cancer Informatics:1-8, 2017
13. El-Galaly TC, Jakobsen LH, Hutchings M, et al: Routine Imaging for Diffuse Large B-Cell Lymphoma in First Complete Remission Does Not Improve Post-Treatment Survival: A Danish–Swedish Population-Based Study. Journal of Clinical Oncology 33:3993-3998, 2015
14. Przepiorka D, Ko CW, Deisseroth A, et al: FDA Approval: Blinatumomab. Clinical Cancer Research 21:4035-4039, 2015
15. Hernandez AF, Fleurence RL, Rothman RL: The ADAPTABLE Trial and PCORnet: Shining Light on a New Research Paradigm. Annals of Internal Medicine 163:635-636, 2015

16. Dawood S, Broglio K, Buzdar AU, et al: Prognosis of Women With Metastatic Breast Cancer byHER2Status and Trastuzumab Treatment: An Institutional-Based Review. Journal of Clinical Oncology 28:92-98, 2010
17. Pfizer: U.S. FDA APPROVES IBRANCE® (PALBOCICLIB) FOR THE TREATMENT OF MEN WITH HR+, HER2- METASTATIC BREAST CANCER, 2019
18. Administration USFD: <IBRANCE® (palbociclib) capsules, for oral use - Label.pdf>. 2019
19. Gierach GL, Curtis RE, Pfeiffer RM, et al: Association of Adjuvant Tamoxifen and Aromatase Inhibitor Therapy With Contralateral Breast Cancer Risk Among US Women With Breast Cancer in a General Community Setting. JAMA Oncology 3:186, 2017
20. Daniels B, Kiely BE, Lord SJ, et al: Long-term survival in trastuzumab-treated patients with HER2-positive metastatic breast cancer: real-world outcomes and treatment patterns in a whole-of-population Australian cohort (2001–2016). Breast Cancer Research and Treatment 171:151-159, 2018
21. McNamara DM, Goldberg SL, Latts L, et al: Differential impact of cognitive computing augmented by real world evidence on novice and expert oncologists. Cancer Medicine, 2019
22. Cowie MR, Blomster JI, Curtis LH, et al: Electronic health records to facilitate clinical research. Clinical Research in Cardiology 106:1-9, 2017
23. Skovlund E, Leufkens HGM, Smyth JF: The use of real-world data in cancer drug development. European Journal of Cancer 101:69-76, 2018
24. Warner JL, Jain SK, Levy MA: Integrating cancer genomic data into electronic health records. Genome Medicine 8:113, 2016
25. Shimelis H, Laduca H, Hu C, et al: Triple-Negative Breast Cancer Risk Genes Identified by Multigene Hereditary Cancer Panel Testing. JNCI: Journal of the National Cancer Institute 110:855-862, 2018
26. Vallon-Christersson J, Häkkinen J, Hegardt C, et al: Cross comparison and prognostic assessment of breast cancer multigene signatures in a large population-based contemporary clinical series. Scientific Reports 9, 2019
27. Győrffy B, Pongor L, Bottai G, et al: An integrative bioinformatics approach reveals coding and non-coding gene variants associated with gene expression profiles and outcome in breast cancer molecular subtypes. British Journal of Cancer 118:1107-1114, 2018
28. Avazpour N, Hajjari M, Tahmasebi Birgani M: HOTAIR: A Promising Long Non-coding RNA with Potential Role in Breast Invasive Carcinoma. Frontiers in Genetics 8, 2017
29. Plitas G, Konopacki C, Wu K, et al: Regulatory T Cells Exhibit Distinct Features in Human Breast Cancer. Immunity 45:1122-1134, 2016
30. Guo W, Wang Q, Zhan Y, et al: Transcriptome sequencing uncovers a three–long noncoding RNA signature in predicting breast cancer survival. Scientific Reports 6:27931, 2016
31. Craig DW, O'Shaughnessy JA, Kiefer JA, et al: Genome and Transcriptome Sequencing in Prospective Metastatic Triple-Negative Breast Cancer Uncovers Therapeutic Vulnerabilities. Molecular Cancer Therapeutics 12:104-116, 2013
32. Zoon CK, Starker EQ, Wilson AM, et al: Current molecular diagnostics of breast cancer and the potential incorporation of microRNA. Expert Review of Molecular Diagnostics 9:455-466, 2009

33. BreastCancer.org: MammaPrint Test, 2019
34. Beaubier N, Bontrager M, Huether R, et al: Integrated genomic profiling expands clinical options for patients with cancer. Nature Biotechnology, 2019
35. Michuda J, Igartua C, Taxter T, et al: Transcriptome-based cancer type prediction for tumors of unknown origin. Journal of Clinical Oncology 37:3081-3081, 2019
36. Grewal JK, Tessier-Cloutier B, Jones M, et al: Application of a Neural Network Whole Transcriptome–Based Pan-Cancer Method for Diagnosis of Primary and Metastatic Cancers. JAMA Network Open 2:e192597, 2019
37. Brueffer C, Vallon-Christersson J, Grabau D, 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 Precision Oncology:1-18, 2018
38. Fumagalli D, Blanchet-Cohen A, Brown D, et al: Transfer of clinically relevant gene expression signatures in breast cancer: from Affymetrix microarray to Illumina RNA-Sequencing technology. 15:1008, 2014
39. Gill J, Prasad V: Improving observational studies in the era of big data. The Lancet 392:716-717, 2018
40. Miksad RA, Abernethy AP: Harnessing the Power of Real-World Evidence (RWE): A Checklist to Ensure Regulatory-Grade Data Quality. Clinical Pharmacology & Therapeutics 103:202-205, 2018
41. Gyawali B, Parsad S, Feinberg BA, et al: Real-World Evidence and Randomized Studies in the Precision Oncology Era: The Right Balance. JCO Precision Oncology:1-5, 2017
42. Beaubier N, Tell R, Huether R, et al: Clinical validation of the Tempus xO assay. Oncotarget 9, 2018
43. Beaubier N, Tell R, Lau D, et al: Clinical validation of the tempus xT next-generation targeted oncology sequencing assay. Oncotarget 10, 2019
44. Hänzelmann S, Castelo R, Guinney J: GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics 14:7, 2013
45. Liberzon A, Birger C, Thorvaldsdóttir H, et al: The Molecular Signatures Database Hallmark Gene Set Collection. Cell Systems 1:417-425, 2015
46. Society AC: Breast Cancer Facts & Figures 2019-2020, in Society AC (ed), 2019
47. Institute NC: SEER Cancer Statistics Review 1975-2016, 2019
48. Iqbal J, Ginsburg O, Rochon PA, et al: Differences in Breast Cancer Stage at Diagnosis and Cancer-Specific Survival by Race and Ethnicity in the United States. JAMA 313:165, 2015
49. Goss PE, Ingle JN, Martino S, et al: Impact of premenopausal status at breast cancer diagnosis in women entered on the placebo-controlled NCIC CTG MA17 trial of extended adjuvant letrozole. Annals of Oncology 24:355-361, 2013
50. Desantis CE, Ma J, Gaudet MM, et al: Breast cancer statistics, 2019. CA: A Cancer Journal for Clinicians, 2019
51. Arena V, Pennacchia I, Vecchio FM, et al: ER−/PR+/HER2− breast cancer type shows the highest proliferative activity among all other combined phenotypes and is more common in young patients: Experience with 6643 breast cancer cases. The Breast Journal 25:381-385, 2019

52. Nahed AS, Shaimaa MY: Ki-67 as a prognostic marker according to breast cancer molecular subtype. Cancer Biology & Medicine 13:496, 2016
53. Gullo G, Walsh N, Fennelly D, et al: Impact of timing of trastuzumab initiation on long-term outcome of patients with early-stage HER2-positive breast cancer: the "one thousand HER2 patients" project. British Journal of Cancer 119:374-380, 2018
54. Bahreini F, Soltanian AR, Mehdipour P: A meta-analysis on concordance between immunohistochemistry (IHC) and fluorescence in situ hybridization (FISH) to detect HER2 gene overexpression in breast cancer. Breast Cancer 22:615-25, 2015
55. Network TCGA: Comprehensive molecular portraits of human breast tumours. Nature 490:61-70, 2012
56. Cortes-Ciriano I, Lee S, Park W-Y, et al: A molecular portrait of microsatellite instability across multiple cancers. Nature Communications 8:15180, 2017
57. Bradford JR, Cox A, Bernard P, et al: Consensus Analysis of Whole Transcriptome Profiles from Two Breast Cancer Patient Cohorts Reveals Long Non-Coding RNAs Associated with Intrinsic Subtype and the Tumour Microenvironment. PLOS ONE 11:e0163238, 2016
58. Eswaran J, Cyanam D, Mudvari P, et al: Transcriptomic landscape of breast cancers through mRNA sequencing. Scientific Reports 2, 2012
59. Smid M, Wang Y, Zhang Y, et al: Subtypes of Breast Cancer Show Preferential Site of Relapse. Cancer Research 68:3108-3114, 2008
60. Allott EH, Geradts J, Sun X, et al: Intratumoral heterogeneity as a source of discordance in breast cancer biomarker classification. Breast Cancer Research 18, 2016
61. Robertson S, Rönnlund C, De Boniface J, et al: Re-testing of predictive biomarkers on surgical breast cancer specimens is clinically relevant. Breast Cancer Research and Treatment 174:795-805, 2019
62. Dekker TJA, Smit VTHBM, Hooijer GKJ, et al: Reliability of core needle biopsy for determining ER and HER2 status in breast cancer. 24:931-937, 2013
63. Hare SH, Harvey AJ: mTOR function and therapeutic targeting in breast cancer. American journal of cancer research 7:383-404, 2017
64. Holgado E, Perez-Garcia J, Gion M, et al: Is there a role for immunotherapy in HER2-positive breast cancer? npj Breast Cancer 4, 2018
65. Wu N, Zhang J, Zhao J, et al: Precision medicine based on tumorigenic signaling pathways for triple-negative breast cancer. Oncology letters 16:4984-4996, 2018
66. Elsawaf Z, Sinn H-P: Triple-Negative Breast Cancer: Clinical and Histological Correlations. Breast Care 6:273-278, 2011
67. O'Neill S, Porter RK, McNamee N, et al: 2-Deoxy-D-Glucose inhibits aggressive triple-negative breast cancer cells by targeting glycolysis and the cancer stem cell phenotype. Scientific Reports 9, 2019
68. Thomas R, Al-Khadairi G, Roelands J, et al: NY-ESO-1 Based Immunotherapy of Cancer: Current Perspectives. Frontiers in Immunology 9, 2018
69. Research FoC: Blueprint for Breakthrough: Exploring the Utility of Real World Evidence (RWE), 2016
70. Pobiruchin M, Bochum S, Martens UM, et al: A method for using real world data in breast cancer modeling. Journal of Biomedical Informatics 60:385-394, 2016

71. Zare S, Rong J, Daehne S, et al: Implementation of the 2018 American Society of Clinical Oncology/College of American Pathologists Guidelines on HER2/neu Assessment by FISH in breast cancers: predicted impact in a single institutional cohort. Modern Pathology 32:1566-1573, 2019

72. Wolff AC, Hammond MEH, Allison KH, et al: Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. Journal of Clinical Oncology 36:2105-2122, 2018

73. von Minckwitz G, Procter M, de Azambuja E, et al: Adjuvant Pertuzumab and Trastuzumab in Early HER2-Positive Breast Cancer. New England Journal of Medicine 377:122-131, 2017

74. Fehrenbacher L, Cecchini R, Geyer C, et al: Abstract GS1-02: NSABP B-47 (NRG oncology): Phase III randomized trial comparing adjuvant chemotherapy with adriamycin (A) and cyclophosphamide (C) → weekly paclitaxel (WP), or docetaxel (T) and C with or without a year of trastuzumab (H) in women with node-positive or high-risk node-negative invasive breast cancer (IBC) expressing HER2 staining intensity of IHC 1+ or 2+ with negative FISH (HER2-Low IBC). Cancer Research 78:GS1-02-GS1-02, 2018

75. Loi S, Giobbie-Hurder A, Gombos A, et al: Pembrolizumab plus trastuzumab in trastuzumab-resistant, advanced, HER2-positive breast cancer (PANACEA): a single-arm, multicentre, phase 1b–2 trial. The Lancet Oncology 20:371-382, 2019

76. Ayoub NM, Al-Shami KM, Yaghan RJ: Immunotherapy for HER2-positive breast cancer: recent advances and combination therapeutic approaches. Breast Cancer: Targets and Therapy Volume 11:53-69, 2019

77. Sharma A, Koldovsky U, Xu S, et al: HER-2 pulsed dendritic cell vaccine can eliminate HER-2 expression and impact ductal carcinoma in situ. 118:4354-4362, 2012

Figure and Table Legends

Table 1. Demographics and clinical characteristics of the clinical abstraction and Tempus molecular sequenced cohort.

Figure 1. Breast cancer molecular biomarkers and subtypes in the clinical abstraction cohort. (A) The number of patients with positive, negative, or equivocal IHC or FISH test results for ER, PR, HR, and HER2 status at initial diagnosis. (B) The distributions of breast cancer molecular subtypes as determined by abstracted ER, PR, and HER2 test

results at initial diagnosis, and (C) the distribution of ER and PR status combinations across the cohort. (D) The number of patients with high, moderate, low, indeterminate, or equivocal Ki67 IHC test results or status-indicating physician notes at initial diagnosis, separated by molecular subtype.

Figure 2. Anti-HER2 treatment by HER2 status in the clinical abstraction cohort. Anti-HER2 treatment initiation patterns among (A) HER2+ and (B) HER2- patients who received anti-HER2 therapy at some point in their clinical care. M, month; Y, year.

Figure 3. Overall survival from primary diagnosis dates in the clinical abstraction cohort. Ten-year survival probability in (A) all patients and (B) stage I-IV patients stratified by stage. Five-year survival probabilities stratified by HER2 status in (C) all patients and (D) stage IV patients, ER status in (E) all patients and (F) stage IV patients, and TNBC status in (G) all patients and (H) stage IV patients.

Figure 4. RNA-based receptor status prediction analysis of the Tempus molecular sequenced cohort. (A) UMAP transcriptome clustering of 19,147 genes in the cohort color-coded by molecular subtype. Circles correspond to samples with available IHC or FISH test results for all proteins and X symbols correspond to patients with predicted status for at least one protein. (B) Relationship between ER, PR, and HER2 receptor status and $\log_{10}$-transformed, normalized gene expression of *ESR1*, *PGR*, and *ERBB2*. Left panels represent samples with available receptor status from abstracted test

results, while right panels represent transcriptome-based receptor status predictions. HER2 predictions for samples reported as equivocal are plotted as white dots.

Figure 5. Single-gene logistic model performance for ER, PR, and HER2 status prediction in the Tempus molecular sequenced cohort. The (A) specificity and sensitivity, and (B) precision and recall of transcriptome-based receptor status predictions were evaluated on a testing set comprised of cohort RNA-sequenced samples with abstracted receptor status results in the Tempus database. (C) Confusion matrices depicting transcriptome-based ER, PR, and HER2 status prediction performance.

Figure 6: RNA-seq breast cancer pathway analyses of the Tempus molecular sequenced cohort. (A) HER2 and (B) ER pathway metascores for patients with abstracted HER2 IHC or FISH test results. (C) UMAP of 50 Hallmark enrichment scores. Patients with molecular subtypes based on at least one abstracted receptor status are depicted by circles, while patients with molecular subtypes determined exclusively from RNA-predicted statuses are depicted by X symbols. Distribution of enrichment Z-scores for (D) HR-/HER2+ and (E) TNBC relevant pathways.

Supplemental Figure and Table Legends

Supplemental Figure 1. Patients grouped by year of initial diagnosis. The distribution of patients by year of initial diagnosis across the clinical abstraction cohort.

Supplemental Figure 2. Molecular characteristics of the Tempus molecular sequenced cohort. (A) The distribution of patients with variants in the most frequently reported genes across the cohort. The number of patients harboring mutations in each gene are shown above the bars. (B) The number of variants classified as alterations, amplifications, or deletions within each of the most frequently reported genes in the cohort. (C) The distribution of patients with pathogenic germline alterations in NCCN-designated familial high-risk genes and (D) TMB across the cohort.

Supplemental Figure 3. Breast cancer pathway analyses from RNA-seq data of the Tempus molecular sequenced cohort according to MSigDB and Hallmark pathways. (A) Pearson correlation between *ERBB2* expression and enrichment scores (GSVA) for each HER2-related pathway in MSigDB among the cohort. (B) Correlation between *ESR1* expression and enrichment scores for each ER-related pathway in MSigDB among the cohort.

Supplemental Figure 4. (A) For each Hallmark pathway, the significance of differential enrichment between molecular subtypes was determined by a Kruskal-Wallis test of the enrichment scores. The vertical line indicates P=0.001 and any value to the right of the line was considered significant. (B) Distributions of z-scores among HR+/HER2- (blue), HR+/HER2+ (green), HR-/HER2+ (orange), and TNBC (grey) patients for the two estrogen response Hallmark pathways with the most significant differential enrichments between molecular subtypes.

Supplemental Figure 5. Distribution of enrichment z-scores for each HER2-related pathway in MSigDB among patients in the Tempus molecular sequenced cohort. Patients with negative (blue), equivocal (orange), or positive (green) abstracted or predicted HER2 test results are shown. The *P*-values listed for each pathway represent the results of a Kruskal-Wallis test for the difference between enrichment scores from HER2-, HER2-equivocal, and HER2+ patients.

Supplemental Figure 6. Distribution of enrichment z-scores for each ER-related pathway in MSigDB among patients in the Tempus molecular sequenced cohort. Patients with negative (blue) or positive (green) abstracted or predicted ER test results are shown. The *P*-values listed for each pathway represent the results of a Wilcox rank sum test for the difference between enrichment z-scores from ER+ and ER- patients.

Supplemental Table 1. Inter-test comparison of HER2 status from IHC and FISH results among patients in the clinical abstraction cohort with both tests conducted at initial diagnosis.

Supplemental Table 2. Single-gene logistic model performance results for RNA-based predictions of ER, PR, and HER2 status in the Tempus molecular sequenced cohort.

Figure 1

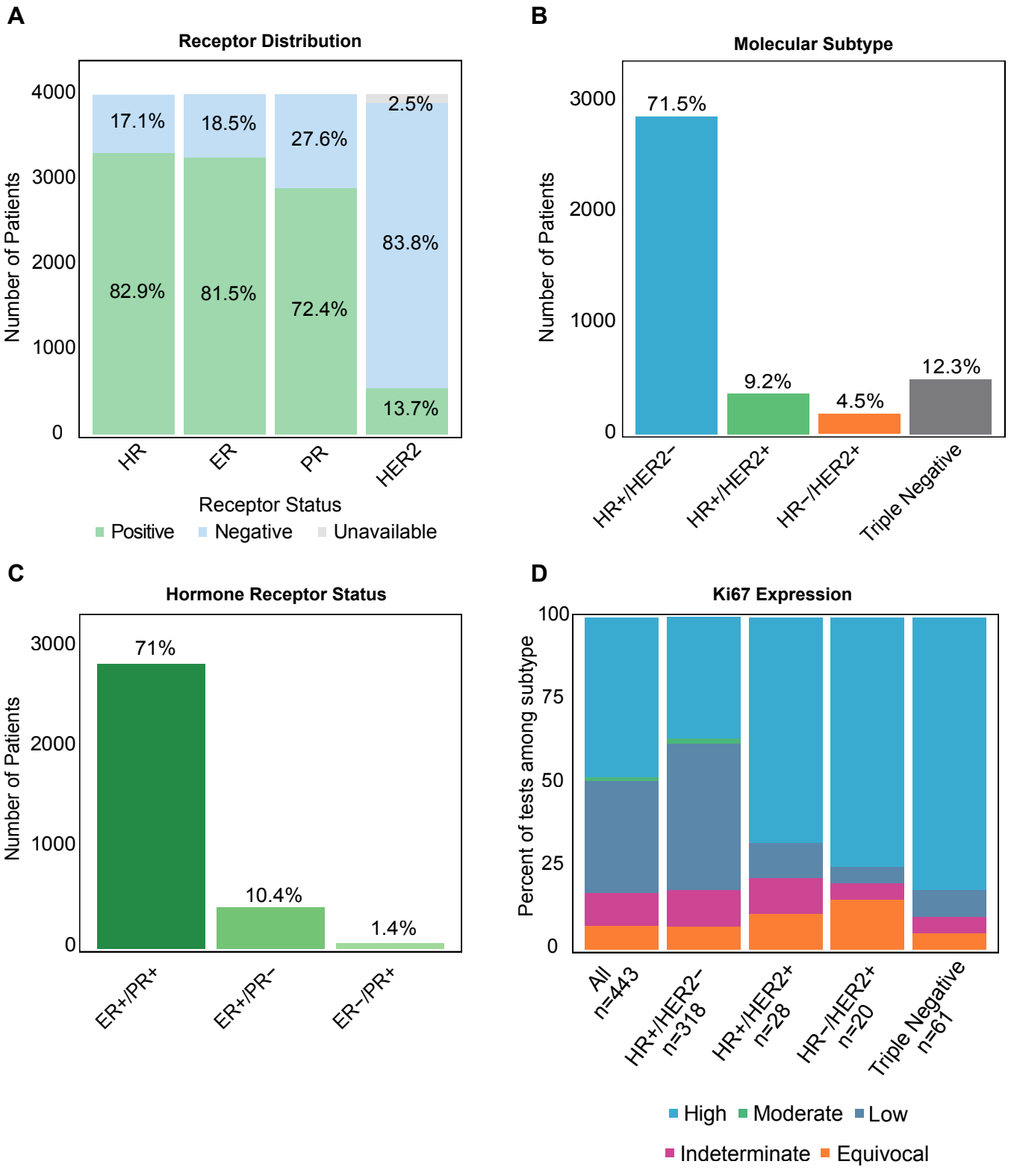
A
Receptor Distribution
Number of Patients
4000
3000
2000
1000
0
17.1%
18.5%
27.6%
2.5%
83.8%
82.9%
81.5%
72.4%
13.7%
HR
ER
PR
HER2
Receptor Status
Positive
Negative
Unavailable
B
Molecular Subtype
Number of Patients
3000
2000
1000
0
71.5%
9.2%
4.5%
12.3%
HR+/HER2−
HR+/HER2+
HR−/HER2+
Triple Negative
C
Hormone Receptor Status
Number of Patients
3000
2000
1000
0
71%
10.4%
1.4%
ER+/PR+
ER+/PR−
ER−/PR+
D
Ki67 Expression
Percent of tests among subtype
100
75
50
25
0
All n=443
HR+/HER2− n=318
HR+/HER2+ n=28
HR−/HER2+ n=20
Triple Negative n=61
High
Moderate
Low
Indeterminate
Equivocal

Figure 2

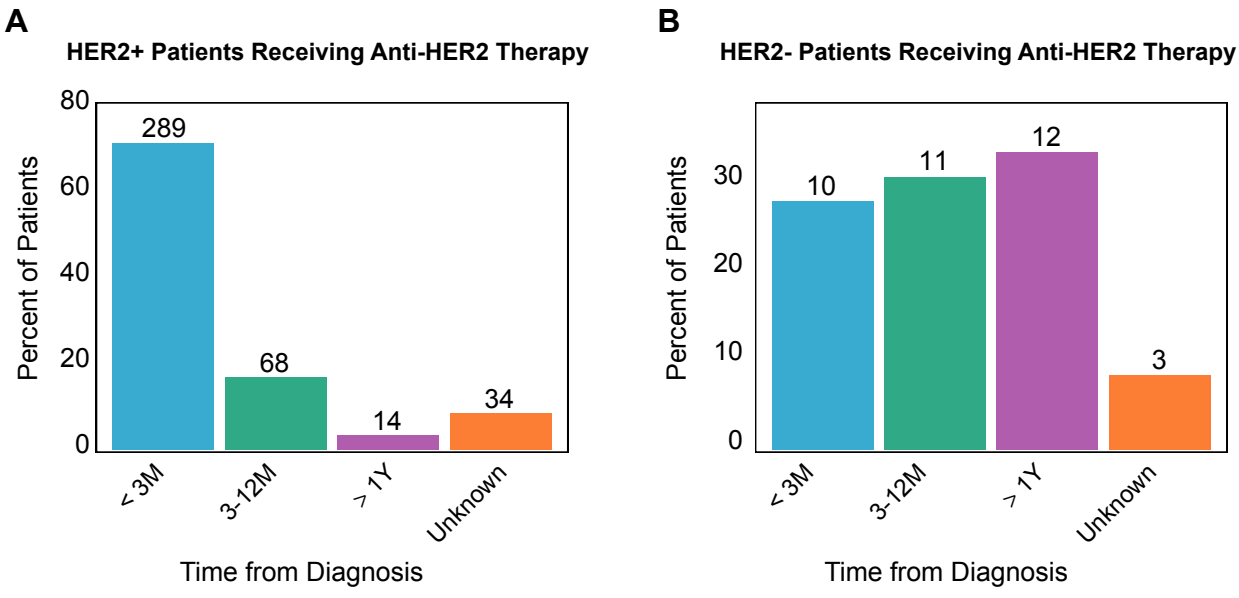
A
HER2+ Patients Receiving Anti-HER2 Therapy
Percent of Patients
80
60
40
20
0
289
68
14
34
< 3M
3-12M
> 1Y
Unknown
Time from Diagnosis
B
HER2- Patients Receiving Anti-HER2 Therapy
Percent of Patients
30
20
10
0
10
11
12
3
< 3M
3-12M
> 1Y
Unknown
Time from Diagnosis

Figure 3

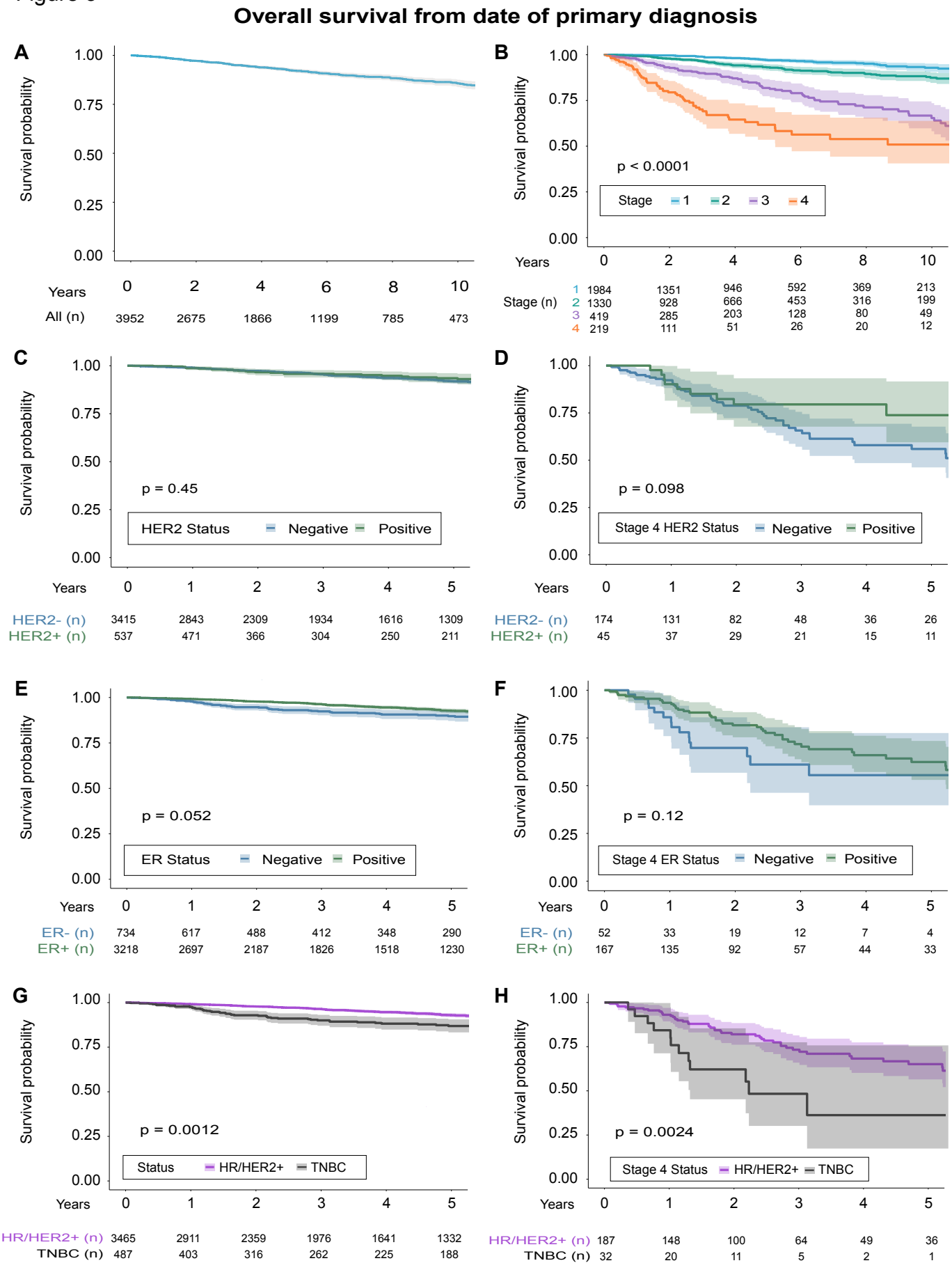
Overall survival from date of primary diagnosis
A
Survival probability
1.00
0.75
0.50
0.25
0.00
Years
0
2
4
6
8
10
All (n)
3952
2675
1866
1199
785
473
B
Survival probability
p < 0.0001
Stage
1
2
3
4
Years
Stage (n)
1 1984 1351 946 592 369 213
2 1330 928 666 453 316 199
3 419 285 203 128 80 49
4 219 111 51 26 20 12
C
Survival probability
p = 0.45
HER2 Status
Negative
Positive
Years
HER2- (n) 3415 2843 2309 1934 1616 1309
HER2+ (n) 537 471 366 304 250 211
D
Survival probability
p = 0.098
Stage 4 HER2 Status
Negative
Positive
Years
HER2- (n) 174 131 82 48 36 26
HER2+ (n) 45 37 29 21 15 11
E
Survival probability
p = 0.052
ER Status
Negative
Positive
Years
ER- (n) 734 617 488 412 348 290
ER+ (n) 3218 2697 2187 1826 1518 1230
F
Survival probability
p = 0.12
Stage 4 ER Status
Negative
Positive
Years
ER- (n) 52 33 19 12 7 4
ER+ (n) 167 135 92 57 44 33
G
Survival probability
p = 0.0012
Status
HR/HER2+
TNBC
Years
HR/HER2+ (n) 3465 2911 2359 1976 1641 1332
TNBC (n) 487 403 316 262 225 188
H
Survival probability
p = 0.0024
Stage 4 Status
HR/HER2+
TNBC
Years
HR/HER2+ (n) 187 148 100 64 49 36
TNBC (n) 32 20 11 5 2 1

Figure 4

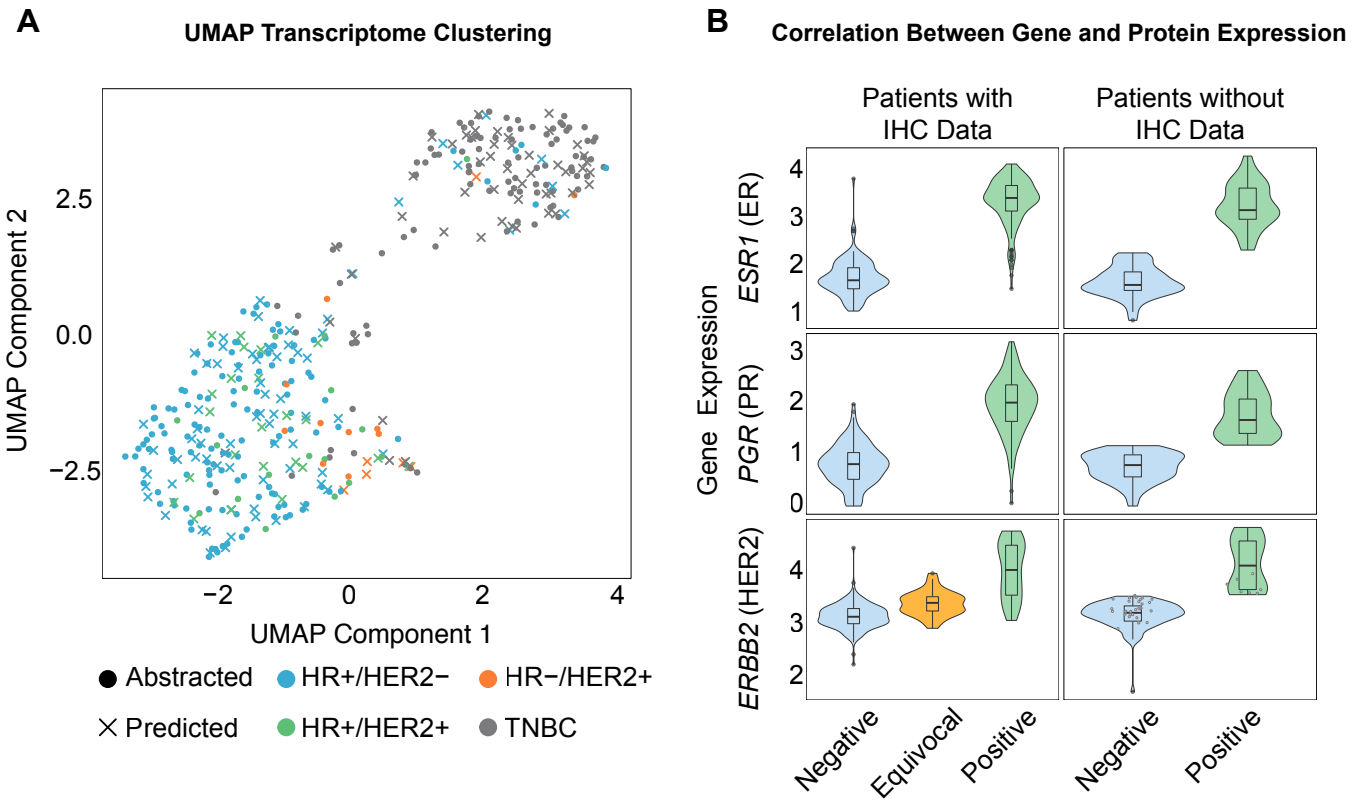
A
UMAP Transcriptome Clustering
UMAP Component 2
2.5
0.0
−2.5
−2
0
2
4
UMAP Component 1
Abstracted
HR+/HER2−
HR−/HER2+
Predicted
HR+/HER2+
TNBC
B
Correlation Between Gene and Protein Expression
Patients with IHC Data
Patients without IHC Data
Gene Expression
ESR1 (ER)
PGR (PR)
ERBB2 (HER2)
Negative
Equivocal
Positive
Negative
Positive

Figure 5

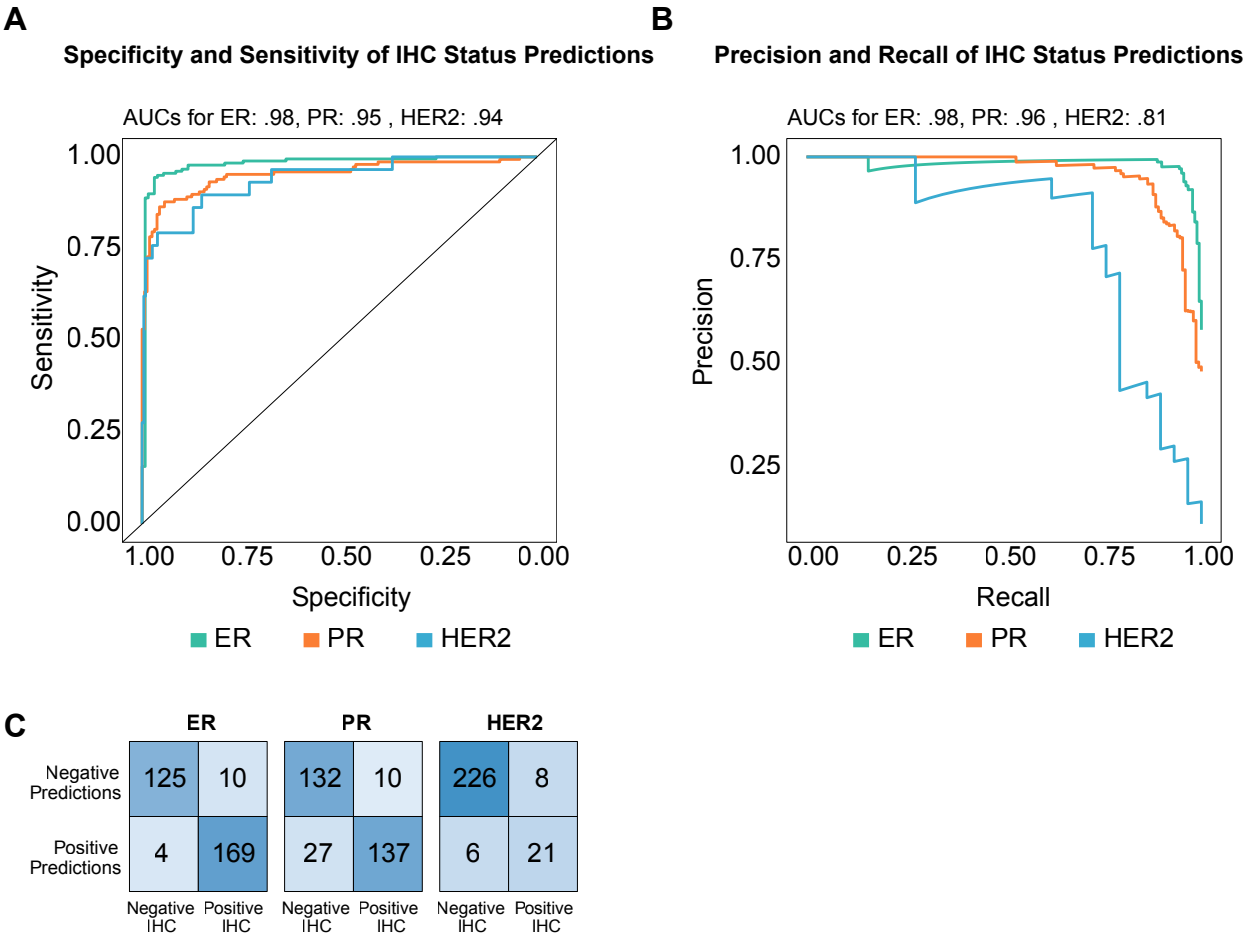
A
Specificity and Sensitivity of IHC Status Predictions
AUCs for ER: .98, PR: .95 , HER2: .94
Sensitivity
1.00
0.75
0.50
0.25
0.00
1.00 0.75 0.50 0.25 0.00
Specificity
ER PR HER2
B
Precision and Recall of IHC Status Predictions
AUCs for ER: .98, PR: .96 , HER2: .81
Precision
1.00
0.75
0.50
0.25
0.00 0.25 0.50 0.75 1.00
Recall
ER PR HER2
C
ER
PR
HER2
Negative Predictions
Positive Predictions
125 10
4 169
132 10
27 137
226 8
6 21
Negative IHC
Positive IHC

Figure 6

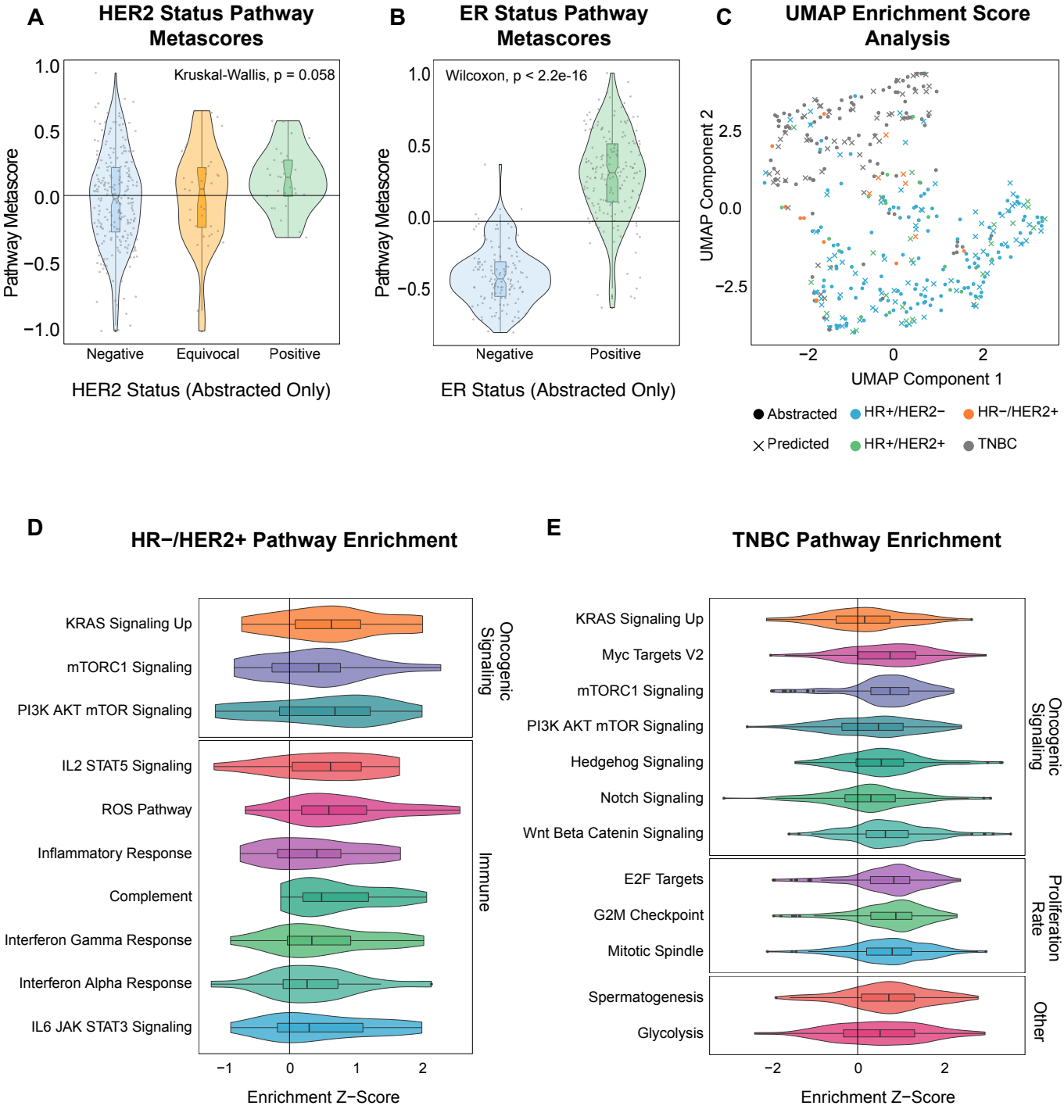
A
HER2 Status Pathway Metascores
Kruskal-Wallis, p = 0.058
Pathway Metascore
1.0
0.5
0.0
−0.5
−1.0
Negative
Equivocal
Positive
HER2 Status (Abstracted Only)
B
ER Status Pathway Metascores
Wilcoxon, p < 2.2e-16
Pathway Metascore
1.0
0.5
0.0
−0.5
Negative
Positive
ER Status (Abstracted Only)
C
UMAP Enrichment Score Analysis
UMAP Component 2
2.5
0.0
−2.5
−2
0
2
UMAP Component 1
Abstracted
Predicted
HR+/HER2−
HR+/HER2+
HR−/HER2+
TNBC
D
HR−/HER2+ Pathway Enrichment
KRAS Signaling Up
mTORC1 Signaling
PI3K AKT mTOR Signaling
IL2 STAT5 Signaling
ROS Pathway
Inflammatory Response
Complement
Interferon Gamma Response
Interferon Alpha Response
IL6 JAK STAT3 Signaling
Oncogenic Signaling
Immune
−1
0
1
2
Enrichment Z−Score
E
TNBC Pathway Enrichment
KRAS Signaling Up
Myc Targets V2
mTORC1 Signaling
PI3K AKT mTOR Signaling
Hedgehog Signaling
Notch Signaling
Wnt Beta Catenin Signaling
E2F Targets
G2M Checkpoint
Mitotic Spindle
Spermatogenesis
Glycolysis
Oncogenic Signaling
Proliferation Rate
Other
−2
0
2
Enrichment Z−Score

Table 1. Patient demographics and clinical characteristics of the clinical abstraction and Tempus molecular sequenced cohorts at initial diagnosis

		Clinical Abstraction Cohort (N=4,000)	Molecular Sequenced Cohort (N=400)
Sex, n (%)	Female	3,970 (99.3%)	396 (99.0%)
	Male	30 (0.7%)	4 (1.0%)
Race, n (%)*	White	3,332 (83.3%)	185 (77.1%)
	Black/AA	523 (13.1%)	35 (14.6%)
	Asian or PI	145 (3.6%)	13 (5.4%)
	Other	0	7 (2.9%)
	Unknown	0	160
Median age (IQR)		61 (51.8-70.2)	55.8 (45.2-66.4)
Stage, n (%)*	0	42 (1.1%)	0
	I	1,986 (49.6%)	30 (13.9%)
	II	1,333 (33.3%)	83 (38.4%)
	III	420 (10.5%)	46 (21.3%)
	IV	219 (5.5%)	57 (26.4%)
	Unknown	0	184
Histological subtype, n (%)*	Invasive ductal	3,095 (77.4%)	267 (75.0%)
	Invasive lobular	345 (8.6%)	23 (6.5%)
	Invasive carcinoma NOS	214 (5.4%)	20 (5.6%)
	Invasive ductal/lobular	167 (4.2%)	20 (5.6%)
	Mucinous (colloid)	61 (1.5%)	0
	Ductal in situ	45 (1.1%)	4 (1.1%)
	Tubular	31 (0.8%)	1 (0.3%)
	Papillary	15 (0.4%)	1 (0.3%)
	Inflammatory	8 (0.2%)	3 (0.8%)
	Metaplastic	6 (0.1%)	12 (3.4%)
	Other	6 (0.1%)	3 (0.8%)
	Medullary	4 (0.1%)	0
	Lobular in situ	1 (0.03%)	1 (0.3%)
	Unmapped malignancy	1 (0.03%)	0
	Phyllodes	1 (0.03%)	1 (0.3%)
	Unknown	0	44
Menopausal status, n (%)*	Postmenopausal	1,867 (86.8%)	67 (91.8%)
	Premenopausal	285 (13.2%)	6 (8.2%)
	Unknown	1,818	313
	Not applicable†	30	4

IQR, interquartile range; AA, African American; PI, Pacific Islander; NOS, not otherwise specified
*Patients with unknown, unreported, or not applicable characteristic/demographic data were not included in population percentage comparisons.
†Represents male patients in the cohort.