

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. Stumpe¹, Michael E. Salazar¹, Ameen A. Salahudeen¹, Ruth A. Pe Benito¹, Calvin McCarter¹, Benjamin D. Leibowitz¹, Matthew Kase¹, Catherine Igartua¹, Robert Huether¹, Ashraf Hafez¹, Nike Beaubier¹, Michael D. Axelson¹, Mark D. Pegram², Sarah L. Sammons³, Joyce A. O'Shaughnessy⁴, and Gary A. Palmer¹

*co-first authors

¹Tempus Labs, Chicago, IL, U.S.A.

²Stanford University School of Medicine, Stanford, CA, U.S.A.

³Department of Medicine, Duke University Medical Center, Duke University, Durham, NC, U.S.A.

⁴Texas 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.

1 Introduction

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

10 An essential step towards strengthening RWE validity is demonstrating
11 consistency between population statistics derived from observational RWD and those
12 from controlled, experimental data. Despite the overwhelming support for RWE utility in
13 oncology, technical barriers must be addressed for RWD/RWE to reach its full clinical
14 potential. Incorporating administrative data, ancillary data, and unstructured clinical text
15 from a variety of institutions to generate RWE is a complex task. For example, no
16 standardization exists for abstracting and structuring highly heterogeneous data
17 sources, and many natural language processing algorithms cannot account for these
18 incongruencies.^{2,3} Consequently, clinical endpoints may not be accurately captured²²
19 and even when data is properly abstracted and prepared for analysis, extraneous
20 variables in raw RWD can introduce confounding biases.^{7,23} Similarly, the integration of
21 omics data with RWD requires a controlled approach for large-scale data analytics.²⁴

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

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

42

43

Patients and Methods

44

45 **Cohort Selections**

46 Two retrospective breast cancer cohorts were randomly selected from the Tempus
47 clinicogenomic database after applying clinically relevant inclusion criteria. All data were
48 de-identified in accordance with the Health Insurance Portability and Accountability Act
49 (HIPAA). Dates used for analyses were relative to the breast cancer primary diagnosis
50 (pdx) date, and year of pdx was randomly off-set. Pdx within the cohorts spanned from
51 1990-2018. The first group was a clinical abstraction (CA) cohort of 4,000 breast cancer
52 patients selected as a representative sample of RWD structured in the Tempus
53 oncology database. Records were required to have data for a pdx, pdx date, age, race,
54 sex, stage, histological subtype, and estrogen receptor (ER), progesterone receptor
55 (PR), and human epidermal growth factor receptor 2 (HER2) status. The recorded stage
56 and histological subtype were required to fall within 30 days relative to the pdx date,
57 while the receptor statuses may have been recorded within 30 or 50 days, depending
58 on the testing modality (see methods: *Molecular Subtype Determination*). A second
59 cohort was selected, the Tempus molecular sequenced (MLC) cohort, which included
60 400 primary breast cancer patients with pdx dates and whose tumor biopsy underwent
61 RNA-seq and targeted DNA sequencing (DNA-seq) with the Tempus xT (n=344), xO
62 (n=55), or xE (n=1) panels between 2017-2019. While only patients with reported
63 variants were included in the cohort, less than 1% of all breast cancer cases in the
64 Tempus database have no DNA variants reported.

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

68

69 ***Abstracted Molecular Markers***

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

79

80 ***Molecular Subtype Determination***

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

91

92 ***Clinical Data Abstraction***

93 Clinical data were extracted from the Tempus real-world oncology database of
94 longitudinal structured and unstructured data from geographically diverse oncology
95 practices, including integrated delivery networks, academic institutions, and community
96 practices. Many of the records included in this study were obtained in partnership
97 through ASCO CancerLinQ. Structured data from electronic health record systems were
98 integrated with unstructured data collected from patient records via technology-enabled
99 chart abstraction and corresponding molecular data, if applicable. Data were
100 harmonized and normalized to standard terminologies from MedDRA, NCBI, NCI, NCI, NCI, NCI, RxNorm, and SNOMED.
101

102

103 ***Menopausal Status Determination***

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

110

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

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

116

117 Menopausal status circumstances beyond the scope of these criteria were denoted as
118 “Unknown.”

119

120 ***Overall Survival Analysis***

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

127

128 ***Genomic Testing***

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

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

140

141 ***Genomic Test Variant Reporting***

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

149

150 ***Tumor mutational burden (TMB)***

151 TMB was calculated by dividing the number of non-synonymous mutations by the
152 adjusted panel size of the xT, xO, or xE assay (2.4 Mb, 5.86 Mb, and 36 Mb,
153 respectively). All non-silent somatic coding mutations, including missense, indel, and
154 stop-loss variants with coverage greater than 100x and an allelic fraction greater than
155 5% were counted as non-synonymous mutations.

156

157 ***RNA-based Prediction of Molecular Subtypes***

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

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

169

170 ***Gene Expression Collection, Processing, and Normalization***

171 Gene expression was generated through RNA-seq of FFPE tumor samples using an
172 exome capture-based protocol.³⁷ Transcript-level quantification to GRCh37 was
173 performed using Kallisto 0.44. Transcript counts were then corrected for GC content
174 and length using quantile normalization and adjusted for sequencing depth via a size
175 factor method. Normalized counts in protein coding transcripts covered by the exome
176 panel were then summed to obtain gene-level counts. Subsequent expression analyses
177 were performed on log₁₀-transformed counts.

178

179 ***RNA-seq Pathway Analyses***

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

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

198

199 **Results**

200

201 ***Real-world evidence from a clinical abstraction breast cancer cohort***

202

203 *Patient demographics and clinical characteristics in the CA cohort*

204 We first determined whether key demographic and clinical characteristics captured in

205 RWD replicate clinical studies, and found the deidentified data were consistent with

206 previous large-scale breast cancer cohort studies (**Table 1**).⁴⁶⁻⁴⁹ The cohort mostly
207 comprised females (99.3%, n=3,970) with a median age at diagnosis of 61.0 years.
208 Year of diagnosis among the cohort ranged from 1990 to 2018 (**Supplemental Fig. 1**).
209 The self-reported race was 83.3% White (n=3,332), 13.1% Black or African American
210 (n=523), and 3.6% Asian or Pacific Islander (n=145). In 2,042 females with menopausal
211 data, 87.4% (n=1,784) were postmenopausal. Abstracted stage at initial diagnosis
212 primarily consisted of stage I (49.6%, n=1,986) and II (33.3%, n=1,333), followed by III
213 (10.5%, n=420), IV (5.5%, n=219), and 0 (1.1%, 42). Most tumors had a histological
214 classification of invasive ductal carcinoma (77.4%, n=3,095), and 9.5% (n=378) had an
215 invasive ductal component or were NOS. Several rare cancer types were also
216 represented.

217

218 *Molecular subtype determination in the CA cohort*

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

229 HR+/HER2+ patients ($P=0.015$) and TNBC versus HR+/HER2- patients ($P=6.38 \times 10^{-9}$).
230 The largest proportions of high Ki67 IHC test results were in TNBC (82.0%, n=50 of 61)
231 and HR-/HER2+ patients (75.0%, n=15 of 20), while most low Ki67 results were in
232 HR+/HER2- patients (44.0%, n=140 of 318).

233

234 Anti-HER2 therapy analysis in the CA cohort

235 We next examined anti-HER2 therapy treatment patterns from longitudinal RWD.
236 Curated anti-HER2 therapies included trastuzumab, ado-trastuzumab emtansine,
237 neratinib, lapatinib and pertuzumab. Among CA cohort patients, 13.7% (n=546) were
238 HER2+ at initial diagnosis, of whom 74.2% (n=405) received anti-HER2 therapy at
239 some point in their clinical care. Approximately 70.0% of patients who received anti-
240 HER2 therapy did so within 3 months of a positive test result and the majority (73.5%)
241 had early-stage cancer (**Fig. 2A**). These results are consistent with previous breast
242 cancer cohort studies.^{16,53} Moreover, a small portion of HER2- patients exhibited
243 evidence of receiving an anti-HER2 therapy (1.1%, 36 of 3,352 HER2- patients) (**Fig.**
244 **2B**). Of those patients, 33.3% (n=12) had evidence of a discordant result at initial
245 diagnosis, 44.4% (n=16) had only HER2- results, and 22.2% (n=8) had a HER2-
246 equivocal or positive result recorded beyond initial diagnosis. A small portion of patients
247 (n=37) were not assigned a HER2 treatment time frame due to date quality issues.

248

249 HER2 test result analyses in the CA cohort

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

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

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

268

269 Overall survival in the CA cohort

270 OS analyses from longitudinal RWD revealed overall 5-year and 10-year survival rates
271 (92.2% and 85.7%, respectively) relatively consistent with average U.S. percentages
272 (**Fig. 3A**).⁴⁶ Survival rates were expectedly high, varying by stage ($P<0.0001$) and
273 receptor status. Stage IV patients exhibited worse OS than earlier-stage patients (**Fig.**
274 **3B**). The 5-year survival rate was 93.5% in stage I-IV HER2+ patients and 92.0% in

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

282

283 ***Genomic testing insights from the Tempus molecular sequenced cohort***

284

285 *Patient demographics and clinical characteristics in the MLC cohort*

286 Abstracted clinical characteristics and patient demographics from the 400 MLC cohort
287 patients were assessed (**Table 1**), and found to be relatively consistent with the CA
288 cohort and other large-scale breast cancer cohort studies.⁴⁶⁻⁴⁹ The cohort had a slightly
289 younger median age at diagnosis of 55.8 years (45.2-66.4), and higher percentage of
290 Black or African American (14.6%, $n=35$) and Asian or Pacific Islander patients (5.4%,
291 $n=13$) than the CA cohort. Patients with known stage information were mostly stage II at
292 diagnosis (38.4%, $n=83$), followed by stages IV (26.4%, $n=57$), III (21.3%, $n=46$), and I
293 (13.9%, $n=30$), indicating an overall higher risk population compared with the CA cohort.
294 A total of 75.0% ($n=267$) of tumors were invasive ductal carcinoma, with several rare
295 cancer types also represented in the cohort.

296

297 *DNA sequencing analysis of the MLC cohort*

298 The top three genes with reported alterations were *TP53*, *PIK3CA*, and *GATA3*, which
299 were found in 55.0% (n=220), 29.0% (n=116), and 13.8% (n=55) of the MLC cohort,
300 respectively (**Supplemental Fig. 2A**). These findings are consistent with a previous
301 analysis of The Cancer Genome Atlas breast cancer data.⁵⁵ **Supplemental Fig. 2B**
302 shows the distribution of variant types in the 20 most frequently reported genes.
303 Assessment of patients with tumor/normal-matched DNA-seq (n=356) identified 18
304 patients (5.1%) with pathogenic germline variants in 12 NCCN-designated familial high-
305 risk genes (**Supplemental Fig. 2C**). This sub-population may be underrepresented as
306 exon-level duplications or deletions were not included. Among the 18 patients harboring
307 a pathogenic germline variant in any of those 12 genes, most contained variants in
308 *BRCA 1* or *2* (n=11), followed by *CHEK2* (n=6), *ATM* (n=3), and *PALB2* (n=2). Because
309 TMB and MSI status are integrated biomarker measurements in the Tempus platform,
310 we observed a wide range of TMB across the cohort with a median of 1.7 mutations/Mb
311 (**Supplemental Fig. 2D**). Consistent with previous studies,⁵⁶ the majority of patients
312 (84.7%, n=339) were MSI stable, while only 0.3% (n=1) were MSI high and 0.5% (n=2)
313 were MSI low.

314

315 *RNA-based prediction of receptor status for molecular subtypes*

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

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

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

334

335 *RNA-based HER2 and ER pathway analyses*

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

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

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

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

369

370 **Discussion**

371

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

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

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

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

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

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

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

427

428 **Conclusion**

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

436

437 **Clinical Practice Points**

438

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

- 458
- These data provide rational for use of the Tempus clinicogenomic database to
- 459 generate RWE and conduct real-time, hypothesis-driven analyses of large RWD
- 460 cohorts in the future. Clinicians may utilize these large-scale databases to
- 461 circumvent the restrictive exclusion criteria of controlled studies, clarify real-world
- 462 patient needs, and aid the development of clinical trials. Furthermore, our results
- 463 suggest molecular data may bolster deficiencies in standard breast cancer
- 464 diagnostic tests.

465

466 **Acknowledgments**

467 We are thankful to ASCO CancerLinQ for their partnership and the clinical data,

468 operations, data science, engineering, pathology, and lab teams at Tempus Labs. We

469 sincerely thank the entire clinical data abstraction team, Jeff Ottens, and April Manhertz.

470 We thank Kelly McKinnon for proofreading and figure assembly and design. We thank

471 Kevin White and Kimberly Blackwell for scientific review and discussion of the

472 manuscript. We thank Hailey Lefkofsky for initial discussions and Eric Lefkofsky for his

473 support and discussions.

474

475

476

477

478

479

480

481

482 **References**

- 483 1. Nabhan C, Klink A, Prasad V: Real-world Evidence—What Does It Really Mean?
484 JAMA Oncology 5:781, 2019
- 485 2. Administration USFD: Framework for the FDA’s Real-world Evidence Program.
486 2018
- 487 3. Jourquin J, Reffey SB, Jernigan C, et al: Susan G. Komen Big Data for Breast
488 Cancer Initiative: How Patient Advocacy Organizations Can Facilitate Using Big Data to Improve
489 Patient Outcomes. JCO Precision Oncology:1-9, 2019
- 490 4. Administration USFD: Submitting Documents Using Real-World Data and Real-
491 World Evidence to FDA for Drugs and Biologics Guidance for Industry, 2019
- 492 5. Administration USFD: Use of Electronic Health Record Data in Clinical
493 Investigations: Guidance for Industry. 2018
- 494 6. Administration USFD: Use of Real-World Evidence to Support Regulatory
495 Decision-Making for Medical Devices: Guidance for Industry and Food and Drug Administration
496 Staff, 2017
- 497 7. Khozin S, Blumenthal GM, Pazdur R: Real-world Data for Clinical Evidence
498 Generation in Oncology. JNCI: Journal of the National Cancer Institute 109, 2017
- 499 8. Quek RGW, Mardekian J: Clinical Outcomes, Treatment Patterns, and Health
500 Resource Utilization Among Metastatic Breast Cancer Patients with Germline BRCA1/2
501 Mutation: A Real-World Retrospective Study. Advances in Therapy 36:708-720, 2019
- 502 9. Taylor-Stokes G, Mitra D, Waller J, et al: Treatment patterns and clinical
503 outcomes among patients receiving palbociclib in combination with an aromatase inhibitor or
504 fulvestrant for HR+/HER2-negative advanced/metastatic breast cancer in real-world settings in
505 the US: Results from the IRIS study. The Breast 43:22-27, 2019
- 506 10. Khozin S, Carson KR, Zhi J, et al: Real-World Outcomes of Patients with
507 Metastatic Non-Small Cell Lung Cancer Treated with Programmed Cell Death Protein 1
508 Inhibitors in the Year Following U.S. Regulatory Approval. The Oncologist 24:648-656, 2019
- 509 11. Martina R, Jenkins D, Bujkiewicz S, et al: The inclusion of real world evidence in
510 clinical development planning. Trials 19, 2018
- 511 12. Harrell M, Fabbri D, Levy M: Analysis of Adjuvant Endocrine Therapy in Practice
512 From Electronic Health Record Data of Patients With Breast Cancer. JCO Clinical Cancer
513 Informatics:1-8, 2017
- 514 13. El-Galaly TC, Jakobsen LH, Hutchings M, et al: Routine Imaging for Diffuse Large
515 B-Cell Lymphoma in First Complete Remission Does Not Improve Post-Treatment Survival: A
516 Danish–Swedish Population-Based Study. Journal of Clinical Oncology 33:3993-3998, 2015
- 517 14. Przepiorcka D, Ko CW, Deisseroth A, et al: FDA Approval: Blinatumomab. Clinical
518 Cancer Research 21:4035-4039, 2015
- 519 15. Hernandez AF, Fleurence RL, Rothman RL: The ADAPTABLE Trial and PCORnet:
520 Shining Light on a New Research Paradigm. Annals of Internal Medicine 163:635-636, 2015

- 521 16. Dawood S, Broglio K, Buzdar AU, et al: Prognosis of Women With Metastatic
522 Breast Cancer byHER2Status and Trastuzumab Treatment: An Institutional-Based Review.
523 Journal of Clinical Oncology 28:92-98, 2010
- 524 17. Pfizer: U.S. FDA APPROVES IBRANCE® (PALBOCICLIB) FOR THE TREATMENT OF
525 MEN WITH HR+, HER2- METASTATIC BREAST CANCER, 2019
- 526 18. Administration USFD: <IBRANCE® (palbociclib) capsules, for oral use - Label.pdf>.
527 2019
- 528 19. Gierach GL, Curtis RE, Pfeiffer RM, et al: Association of Adjuvant Tamoxifen and
529 Aromatase Inhibitor Therapy With Contralateral Breast Cancer Risk Among US Women With
530 Breast Cancer in a General Community Setting. JAMA Oncology 3:186, 2017
- 531 20. Daniels B, Kiely BE, Lord SJ, et al: Long-term survival in trastuzumab-treated
532 patients with HER2-positive metastatic breast cancer: real-world outcomes and treatment
533 patterns in a whole-of-population Australian cohort (2001–2016). Breast Cancer Research and
534 Treatment 171:151-159, 2018
- 535 21. McNamara DM, Goldberg SL, Latts L, et al: Differential impact of cognitive
536 computing augmented by real world evidence on novice and expert oncologists. Cancer
537 Medicine, 2019
- 538 22. Cowie MR, Blomster JI, Curtis LH, et al: Electronic health records to facilitate
539 clinical research. Clinical Research in Cardiology 106:1-9, 2017
- 540 23. Skovlund E, Leufkens HGM, Smyth JF: The use of real-world data in cancer drug
541 development. European Journal of Cancer 101:69-76, 2018
- 542 24. Warner JL, Jain SK, Levy MA: Integrating cancer genomic data into electronic
543 health records. Genome Medicine 8:113, 2016
- 544 25. Shimelis H, Laduca H, Hu C, et al: Triple-Negative Breast Cancer Risk Genes
545 Identified by Multigene Hereditary Cancer Panel Testing. JNCI: Journal of the National Cancer
546 Institute 110:855-862, 2018
- 547 26. Vallon-Christersson J, Häkkinen J, Hegardt C, et al: Cross comparison and
548 prognostic assessment of breast cancer multigene signatures in a large population-based
549 contemporary clinical series. Scientific Reports 9, 2019
- 550 27. Gyórfy B, Pongor L, Bottai G, et al: An integrative bioinformatics approach
551 reveals coding and non-coding gene variants associated with gene expression profiles and
552 outcome in breast cancer molecular subtypes. British Journal of Cancer 118:1107-1114, 2018
- 553 28. Avazpour N, Hajjari M, Tahmasebi Birgani M: HOTAIR: A Promising Long Non-
554 coding RNA with Potential Role in Breast Invasive Carcinoma. Frontiers in Genetics 8, 2017
- 555 29. Plitas G, Konopacki C, Wu K, et al: Regulatory T Cells Exhibit Distinct Features in
556 Human Breast Cancer. Immunity 45:1122-1134, 2016
- 557 30. Guo W, Wang Q, Zhan Y, et al: Transcriptome sequencing uncovers a three–long
558 noncoding RNA signature in predicting breast cancer survival. Scientific Reports 6:27931, 2016
- 559 31. Craig DW, O'Shaughnessy JA, Kiefer JA, et al: Genome and Transcriptome
560 Sequencing in Prospective Metastatic Triple-Negative Breast Cancer Uncovers Therapeutic
561 Vulnerabilities. Molecular Cancer Therapeutics 12:104-116, 2013
- 562 32. Zoon CK, Starker EQ, Wilson AM, et al: Current molecular diagnostics of breast
563 cancer and the potential incorporation of microRNA. Expert Review of Molecular Diagnostics
564 9:455-466, 2009

- 565 33. BreastCancer.org: MammaPrint Test, 2019
- 566 34. Beaubier N, Bontrager M, Huether R, et al: Integrated genomic profiling expands
567 clinical options for patients with cancer. *Nature Biotechnology*, 2019
- 568 35. Michuda J, Igartua C, Taxter T, et al: Transcriptome-based cancer type prediction
569 for tumors of unknown origin. *Journal of Clinical Oncology* 37:3081-3081, 2019
- 570 36. Grewal JK, Tessier-Cloutier B, Jones M, et al: Application of a Neural Network
571 Whole Transcriptome–Based Pan-Cancer Method for Diagnosis of Primary and Metastatic
572 Cancers. *JAMA Network Open* 2:e192597, 2019
- 573 37. Brueffer C, Vallon-Christersson J, Grabau D, et al: Clinical Value of RNA
574 Sequencing–Based Classifiers for Prediction of the Five Conventional Breast Cancer Biomarkers:
575 A Report From the Population-Based Multicenter Sweden Cancerome Analysis Network—Breast
576 Initiative. *JCO Precision Oncology*:1-18, 2018
- 577 38. Fumagalli D, Blanchet-Cohen A, Brown D, et al: Transfer of clinically relevant
578 gene expression signatures in breast cancer: from Affymetrix microarray to Illumina RNA-
579 Sequencing technology. 15:1008, 2014
- 580 39. Gill J, Prasad V: Improving observational studies in the era of big data. *The Lancet*
581 392:716-717, 2018
- 582 40. Miksad RA, Abernethy AP: Harnessing the Power of Real-World Evidence (RWE):
583 A Checklist to Ensure Regulatory-Grade Data Quality. *Clinical Pharmacology & Therapeutics*
584 103:202-205, 2018
- 585 41. Gyawali B, Parsad S, Feinberg BA, et al: Real-World Evidence and Randomized
586 Studies in the Precision Oncology Era: The Right Balance. *JCO Precision Oncology*:1-5, 2017
- 587 42. Beaubier N, Tell R, Huether R, et al: Clinical validation of the Tempus xO assay.
588 *Oncotarget* 9, 2018
- 589 43. Beaubier N, Tell R, Lau D, et al: Clinical validation of the tempus xT next-
590 generation targeted oncology sequencing assay. *Oncotarget* 10, 2019
- 591 44. Hänzelmann S, Castelo R, Guinney J: GSVA: gene set variation analysis for
592 microarray and RNA-Seq data. *BMC Bioinformatics* 14:7, 2013
- 593 45. Liberzon A, Birger C, Thorvaldsdóttir H, et al: The Molecular Signatures Database
594 Hallmark Gene Set Collection. *Cell Systems* 1:417-425, 2015
- 595 46. Society AC: Breast Cancer Facts & Figures 2019-2020, in Society AC (ed), 2019
- 596 47. Institute NC: SEER Cancer Statistics Review 1975-2016, 2019
- 597 48. Iqbal J, Ginsburg O, Rochon PA, et al: Differences in Breast Cancer Stage at
598 Diagnosis and Cancer-Specific Survival by Race and Ethnicity in the United States. *JAMA*
599 313:165, 2015
- 600 49. Goss PE, Ingle JN, Martino S, et al: Impact of premenopausal status at breast
601 cancer diagnosis in women entered on the placebo-controlled NCIC CTG MA17 trial of extended
602 adjuvant letrozole. *Annals of Oncology* 24:355-361, 2013
- 603 50. Desantis CE, Ma J, Gaudet MM, et al: Breast cancer statistics, 2019. *CA: A Cancer*
604 *Journal for Clinicians*, 2019
- 605 51. Arena V, Pennacchia I, Vecchio FM, et al: ER–/PR+/HER2– breast cancer type
606 shows the highest proliferative activity among all other combined phenotypes and is more
607 common in young patients: Experience with 6643 breast cancer cases. *The Breast Journal*
608 25:381-385, 2019

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

- 652 71. Zare S, Rong J, Daehne S, et al: Implementation of the 2018 American Society of
653 Clinical Oncology/College of American Pathologists Guidelines on HER2/neu Assessment by FISH
654 in breast cancers: predicted impact in a single institutional cohort. *Modern Pathology* 32:1566-
655 1573, 2019
- 656 72. Wolff AC, Hammond MEH, Allison KH, et al: Human Epidermal Growth Factor
657 Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American
658 Pathologists Clinical Practice Guideline Focused Update. *Journal of Clinical Oncology* 36:2105-
659 2122, 2018
- 660 73. von Minckwitz G, Procter M, de Azambuja E, et al: Adjuvant Pertuzumab and
661 Trastuzumab in Early HER2-Positive Breast Cancer. *New England Journal of Medicine* 377:122-
662 131, 2017
- 663 74. Fehrenbacher L, Cecchini R, Geyer C, et al: Abstract GS1-02: NSABP B-47 (NRG
664 oncology): Phase III randomized trial comparing adjuvant chemotherapy with adriamycin (A)
665 and cyclophosphamide (C) → weekly paclitaxel (WP), or docetaxel (T) and C with or without a
666 year of trastuzumab (H) in women with node-positive or high-risk node-negative invasive breast
667 cancer (IBC) expressing HER2 staining intensity of IHC 1+ or 2+ with negative FISH (HER2-Low
668 IBC). *Cancer Research* 78:GS1-02-GS1-02, 2018
- 669 75. Loi S, Giobbie-Hurder A, Gombos A, et al: Pembrolizumab plus trastuzumab in
670 trastuzumab-resistant, advanced, HER2-positive breast cancer (PANACEA): a single-arm,
671 multicentre, phase 1b–2 trial. *The Lancet Oncology* 20:371-382, 2019
- 672 76. Ayoub NM, Al-Shami KM, Yaghan RJ: Immunotherapy for HER2-positive breast
673 cancer: recent advances and combination therapeutic approaches. *Breast Cancer: Targets and
674 Therapy* Volume 11:53-69, 2019
- 675 77. Sharma A, Koldovsky U, Xu S, et al: HER-2 pulsed dendritic cell vaccine can
676 eliminate HER-2 expression and impact ductal carcinoma in situ. 118:4354-4362, 2012

677
678

679 **Figure and Table Legends**

680

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

683

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

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

692

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

696

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

702

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

710 results, while right panels represent transcriptome-based receptor status predictions.

711 HER2 predictions for samples reported as equivocal are plotted as white dots.

712

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

720

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

728

729 **Supplemental Figure and Table Legends**

730

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

733

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

741

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

748

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

756

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

763

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

769

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

773

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

776

Figure 1

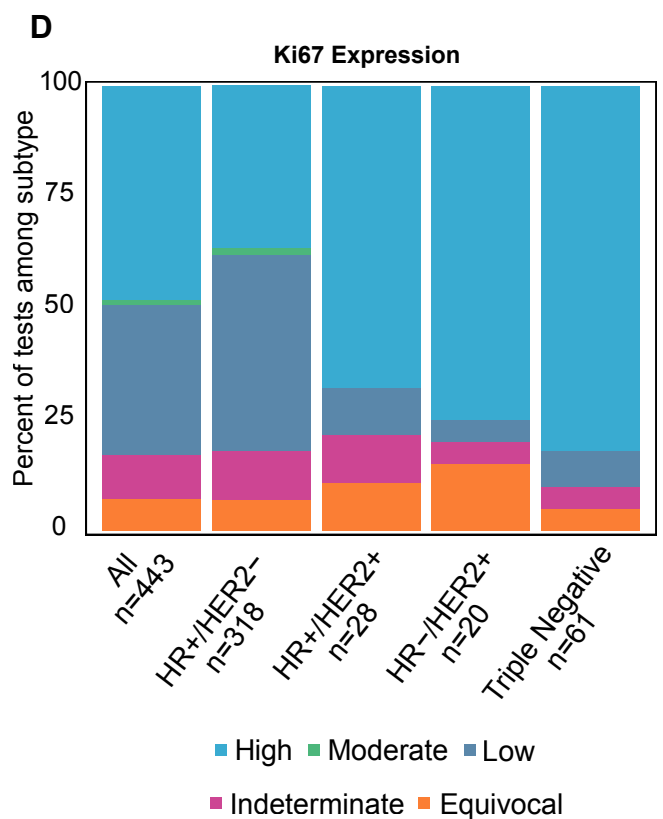
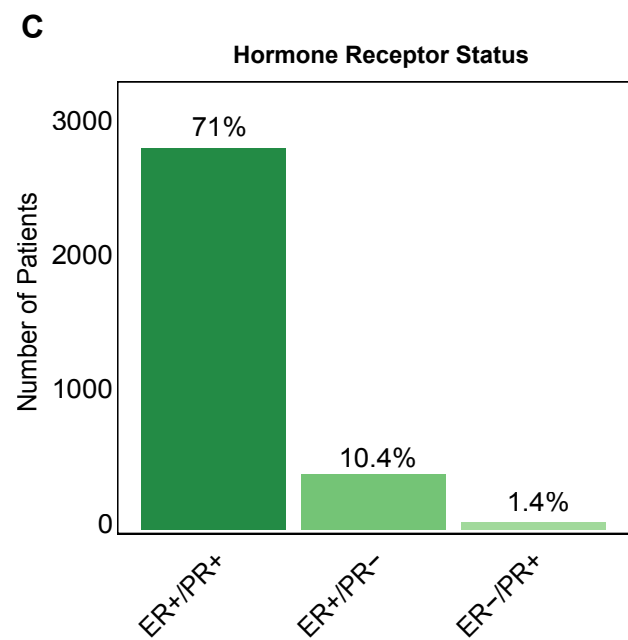
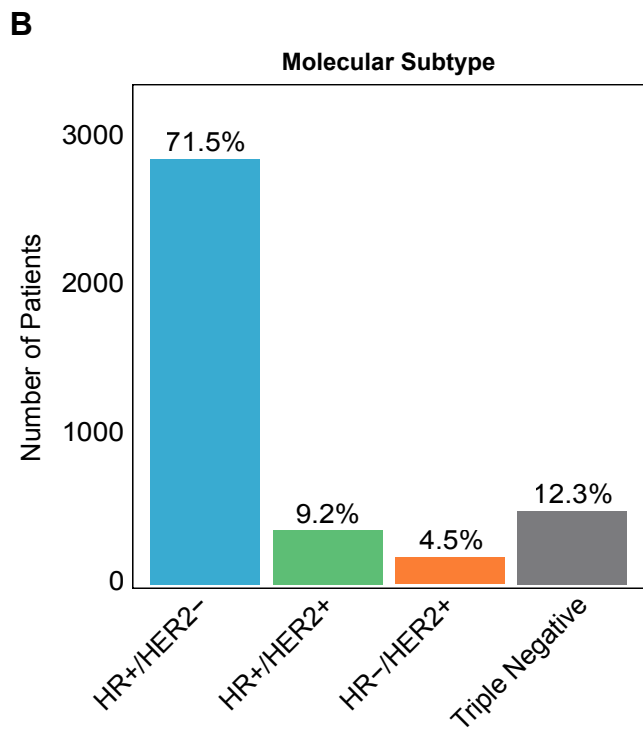
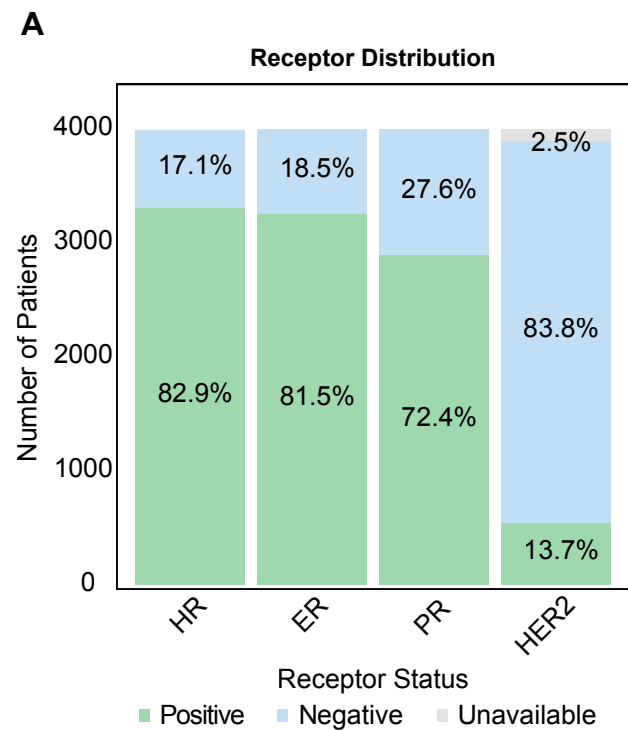


Figure 2

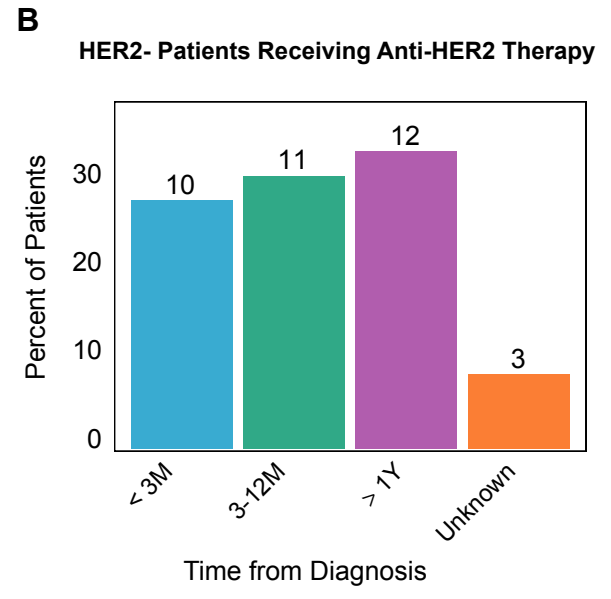
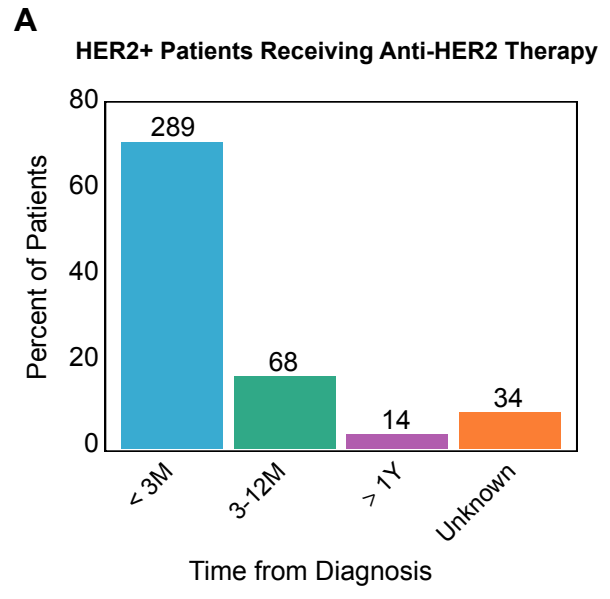


Figure 3

Overall survival from date of primary diagnosis

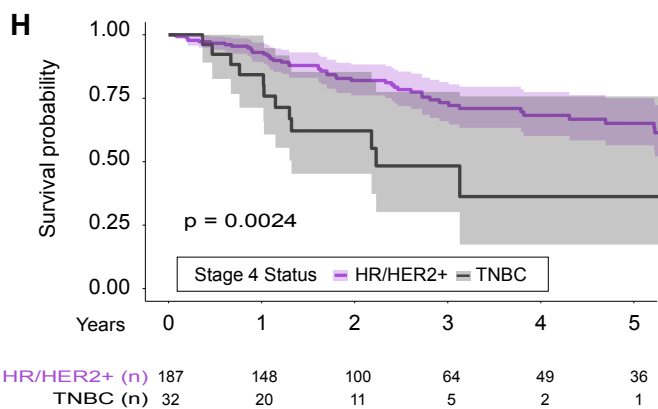
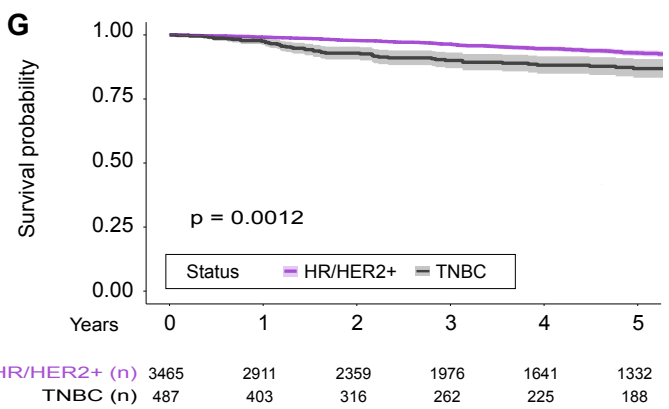
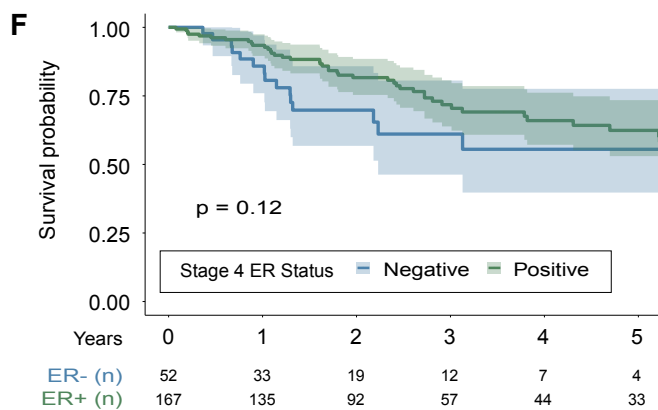
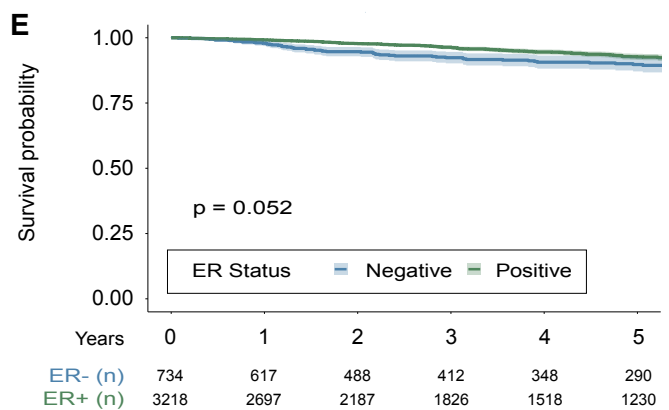
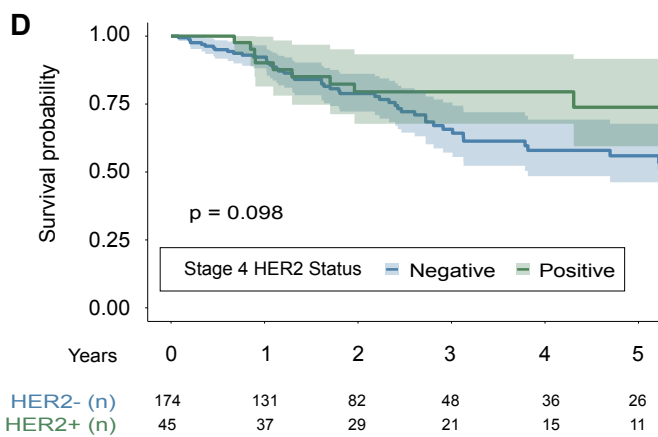
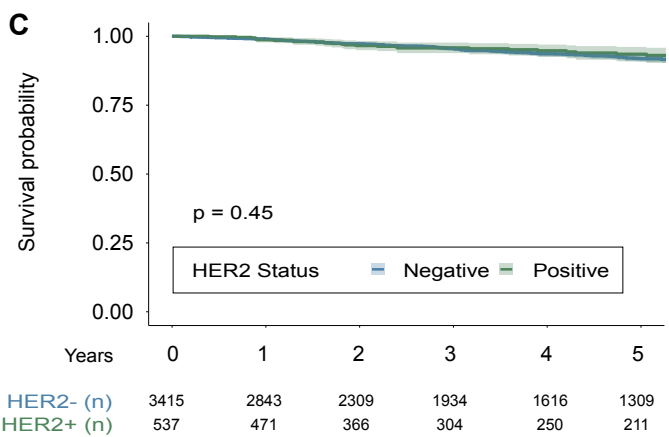
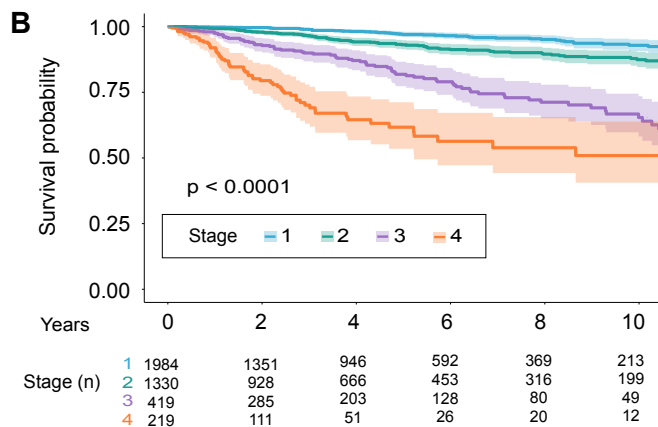
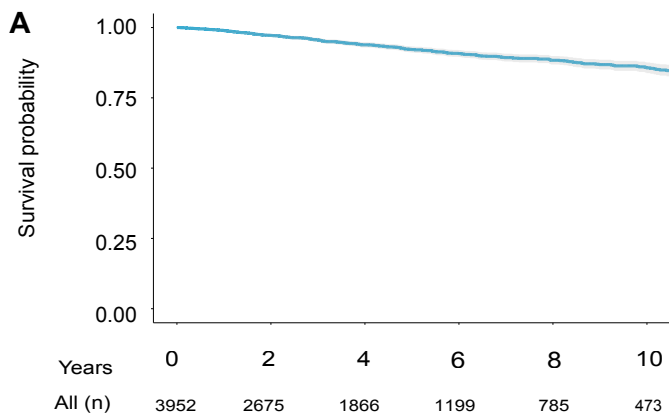


Figure 4

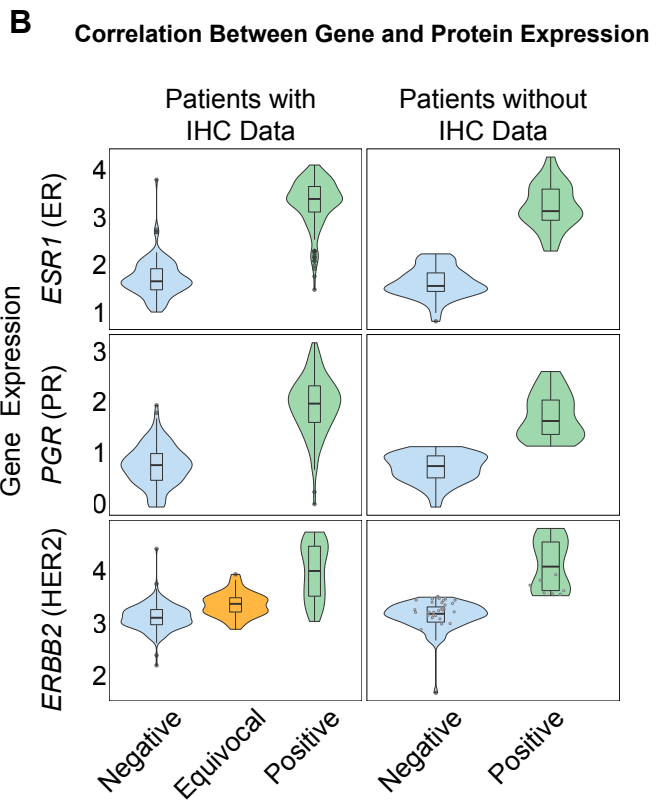
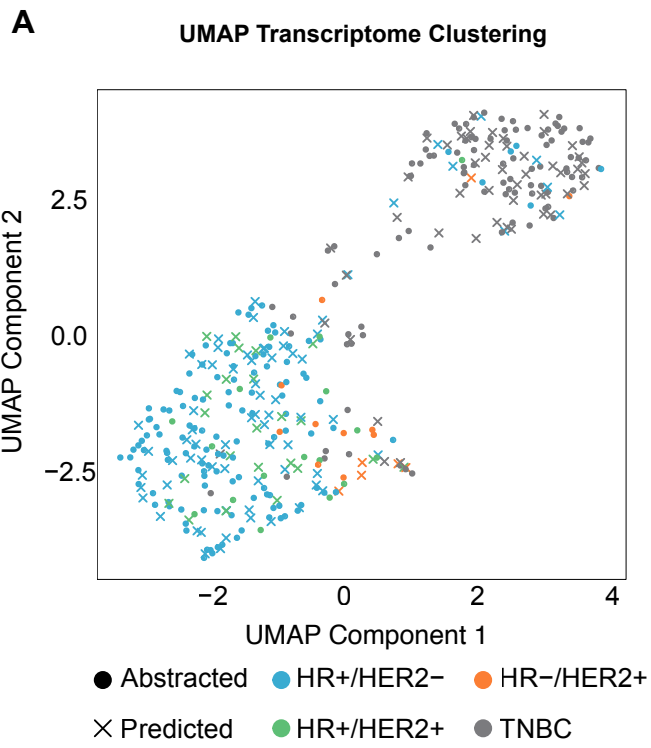
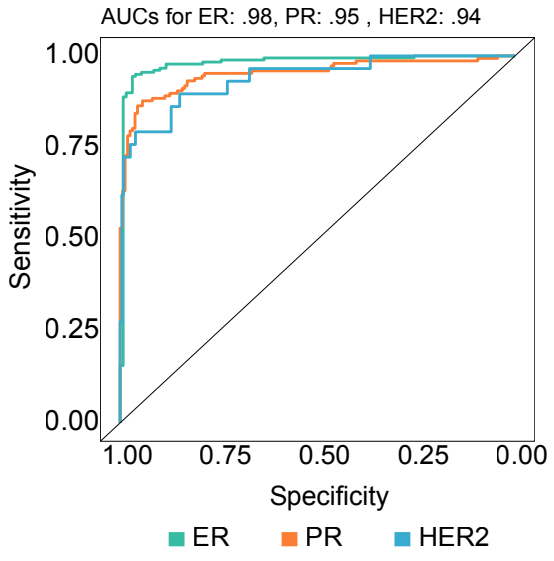
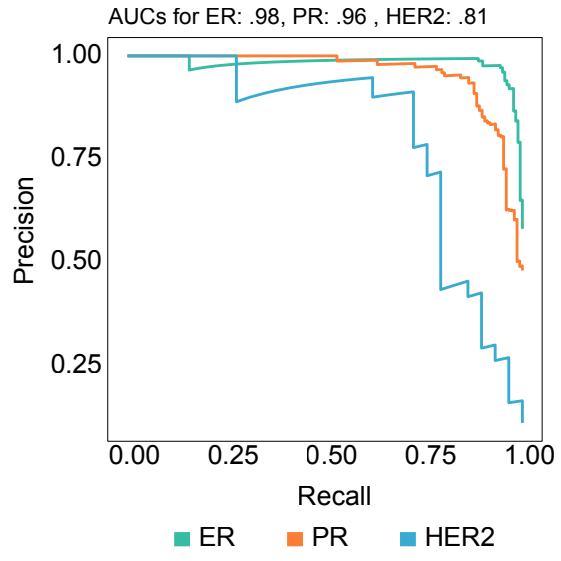


Figure 5

A Specificity and Sensitivity of IHC Status Predictions



B Precision and Recall of IHC Status Predictions



C

	ER		PR		HER2	
Negative Predictions	125	10	132	10	226	8
Positive Predictions	4	169	27	137	6	21
	Negative IHC	Positive IHC	Negative IHC	Positive IHC	Negative IHC	Positive IHC

Figure 6

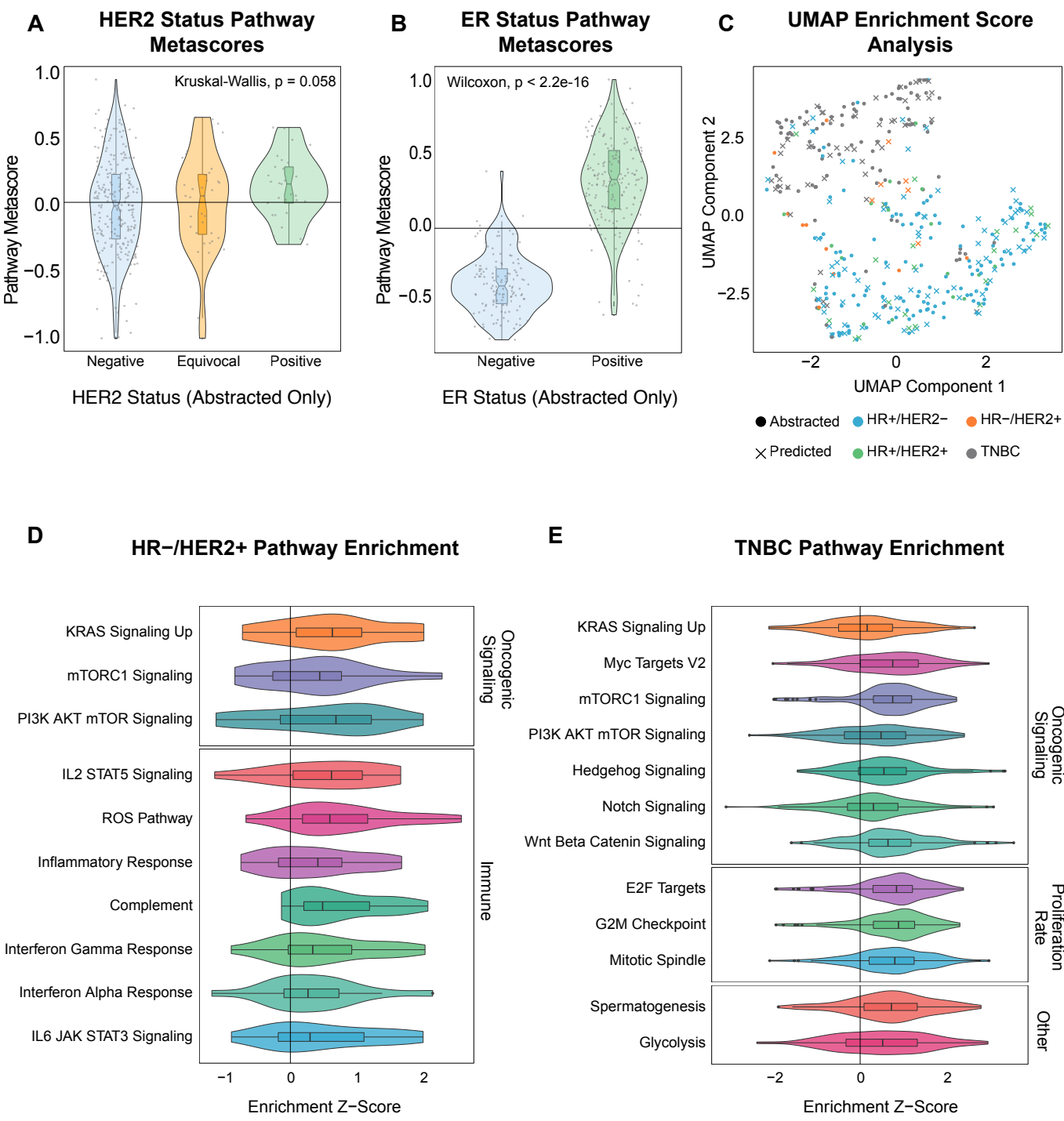


Table 1. Patient demographics and clinical characteristics of the clinical abstraction and Tempus molecular sequenced cohorts at initial diagnosis

		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.