ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

ANNALS OF ONCOLOGY driving innovation in oncology

ORIGINAL ARTICLE

Treatment Efficacy Score—continuous residual cancer burden-based metric to compare neoadjuvant chemotherapy efficacy between randomized trial arms in breast cancer trials

M. Marczyk[1,2], A. Mrukwa[1], C. Yau[3], D. Wolf[3,4], Y.-Y. Chen[5], R. Balassanian[6], R. Nanda[7], B. A. Parker[8], G. Krings[3], H. Sattar[9], J. C. Zeck[10], K. S. Albain[11], J. C. Boughey[12], M. C. Liu[13], A. D. Elias[14], A. S. Clark[15], S. J. Venters[4], S. Shad[3], A. Basu[3], S. M. Asare[16], M. Buxton[3], A. L. Asare[16], H. S. Rugo[17], J. Perlmutter[18], A. M. DeMichele[15], D. Yee[19], D. A. Berry[20], L. van't Veer[4], W. F. Symmans[21], L. Esserman[3] & L. Pusztai[2*], on behalf of the I-SPY Consortium[†]

[1]Department of Data Mining and Engineering, Silesian University of Technology, Gliwice, Poland; [2]Department of Breast Medical Oncology, Yale School of Medicine, New Haven; Departments of [3]Surgery; [4]Laboratory Medicine, [5]Pathology, University of California, San Francisco; [6]Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis; [7]Department of Medicine, Section of Hematology/Oncology, University of Chicago, Chicago; [8]Department of Medicine, Division of Hematology-Oncology, University of California San Diego, La Jolla; [9]Department of Pathology, University of Chicago, Chicago; [10]Department of Pathology, Georgetown University, Washington; [11]Department of Medicine, Division of Hematology-Oncology, Loyola University Chicago Stritch School of Medicine, Maywood; Departments of [12]Surgery; [13]Oncology, Mayo Clinic, Rochester; [14]Department of Medicine, Division of Medical Oncology, University of Colorado Anschutz Medical Center, Aurora; [15]Department of Medicine, Division of Hematology-Oncology, University of Pennsylvania, Philadelphia; [16]Quantum Leap Healthcare Collaborative, San Francisco; [17]Department of Medicine, Division of Hematology-Oncology, University of California, San Francisco; [18]Gemini Group, Ann Arbor; [19]Department of Medicine, Division of Hematology, Oncology, and Transplantation, University of Minnesota, Minneapolis; [20]Berry Consultants, LLC, Houston; [21]Departments of Pathology and Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, USA

Available online XXX

Background: Difference in pathologic complete response (pCR) rate after neoadjuvant chemotherapy does not capture the impact of treatment on downstaging of residual cancer in the experimental arm. We developed a method to compare the entire distribution of residual cancer burden (RCB) values between clinical trial arms to better quantify the differences in cytotoxic efficacy of treatments.

Patients and methods: The Treatment Efficacy Score (TES) reflects the area between the weighted cumulative distribution functions of RCB values from two trial arms. TES is based on a modified Kolmogorov—Smirnov test with added weight function to capture the importance of high RCB values and uses the area under the difference between two distribution functions as a statistical metric. The higher the TES the greater the shift to lower RCB values in the experimental arm. We developed TES from the durvalumab + olaparib arm ($n = 72$) and corresponding controls ($n = 282$) of the I-SPY2 trial. The 11 other experimental arms and control cohorts ($n = 947$) were used as validation sets to assess the performance of TES. We compared TES to Kolmogorov—Smirnov, Mann—Whitney, and Fisher's exact tests to identify trial arms with higher cytotoxic efficacy and assessed associations with trial arm level survival differences. Significance was assessed with a permutation test.

Results: In the validation set, TES identified arms with a higher pCR rate but was more accurate to identify regimens as less effective if treatment did not reduce the frequency of high RCB values, even if the pCR rate improved. The correlation between TES and survival was higher than the correlation between the pCR rate difference and survival.

Conclusions: TES quantifies the difference between the entire distribution of pathologic responses observed in trial arms and could serve as a better early surrogate to predict trial arm level survival differences than pCR rate difference alone.

Key words: breast cancer, neoadjuvant chemotherapy, residual cancer burden

INTRODUCTION

Residual cancer burden (RCB) scores quantify the amount of invasive cancer after neoadjuvant chemotherapy by integrating tumor size, tumor cellularity, and nodal involvement into a single continuous score.[1] The RCB values are grouped into four RCB classes including RCB-0 [RCB value = 0, equivalent to pathologic complete response (pCR) or ypT0/

*Correspondence to: Prof. Lajos Pusztai, Yale Cancer Center, Yale School of Medicine, 300 George St, Suite 120, New Haven, CT 06520, USA. Tel: +1-203-737-8309

E-mail: lajos.pusztai@yale.edu (L. Pusztai).

[†]See Supplementary author list, available at https://doi.org/10.1016/j.annonc.2022.04.072.

is ypN0], RCB-1 (RCB values: 0-1.36), RCB-II (RCB values: 1.37-3.28), and RCB-III (RCB values >3.28) that represent groups with increasingly larger residual cancer and worse recurrence-free survival.[1-4] RCB 0/pCR has been adopted as a clinical efficacy endpoint in neoadjuvant trials because of its strong association with excellent long-term survival at the patient level.[5] Comparing only the pCR rates between trial arms, however, ignores the effect of downstaging of residual cancers in those who fail to achieve a pCR by a more effective treatment. Since the extent of residual cancer correlates strongly with survival, differences in residual disease size distribution can have important effects on trial arm level survival. Moving patients with minimal residual disease to the pCR category will increase the pCR rate, but will have only a small effect on survival, because both minimal residual disease and pCR have a good long-term prognosis. Moving many patients from RCB3 to RCB2, and from RCB2 to RCB1 classes, however, while not altering the pCR rate, could substantially affect trial arm level survival. Several neoadjuvant trials have adopted RCB as a co-primary endpoint, including the I-SPY 2 trial.

Using data from the I-SPY2 trial (NCT01042379), we previously showed that different neoadjuvant therapies cause different types of shifts in RCB score distributions.[6] We noted that immune checkpoint inhibitors or trastuzumab added to chemotherapy causes large shifts towards smaller residual disease values and this may explain the substantial improvement in event-free survival (EFS) observed in the neoadjuvant KEYNOTE-522 (NCT03036488) and GeparNuevo (NCT02685059) immunotherapy trials despite modest increases in pCR rates.[7,8] We hypothesize that efficacy comparisons between two different neoadjuvant cytotoxic regimens could be improved by comparing the entire distribution of RCB values rather than just comparing pCR rates. A score that quantifies this additional treatment benefit in cases with residual cancer could better predict trial arm level survival improvement than increase in pCR rate alone and could aid investigators and regulators to evaluate the impact of new therapies in the neoadjuvant setting.

In this paper, we compare different statistical methods to capture differences in RCB value distributions in trial arms and propose a novel statistical metric, the Treatment Efficacy Score (TES), as a new efficacy measure in neoadjuvant trials. The distribution of RCB values is non-normal and multimodal and therefore changes in mean or median values do not fully capture RCB shifts. We use the non-parametric, two-sample Kolmogorov–Smirnov (KS) test with modifications to compare the entire distributions of RCB values between treatment arms. To develop TES, we used data from the durvalumab + olaparib arm and corresponding controls of the I-SPY 2 trial (i.e. discovery cohort).[9] In subsampling of the discovery set and in simulation experiments we compared TES with KS, Mann–Whitney (MW), and Fisher's exact (F) tests to identify the superior trial arm within biomarker subsets and to assess the robustness of the results under variable sample sizes and sample size imbalances. The remaining 11 experimental arms and 2 additional control arms of the I-SPY2 trial were used as validation sets to examine the association between TES and survival.

METHODS

Discovery and test cohorts

The discovery cohort included 354 human epidermal growth factor receptor-2 (HER2)-negative breast cancer patients treated with paclitaxel ($n = 282$; control arm) or durvalumab + olaparib added to paclitaxel ($n = 72$; experimental arm; Table 1).[9] We selected this arm for discovery because it includes both immune checkpoint therapy and chemotherapy that resulted in an RCB-wide downstaging compared with control. The remaining 11 experimental arms and additional control arms from the I-SPY2 trial[10-12] ($n = 947$ patients) were used as the test cohort (Table 1). Data were analyzed separately within the three breast cancer subtypes including hormone receptor-positive HER2-negative (HR+/HER2−), HER2 positive (HER2+), and triple-negative breast cancer (TNBC). For the HER2+ subtype, paclitaxel + trastuzumab was used as the control treatment; for the other subtypes, paclitaxel alone was used. In all arms, patients also received doxorubicin +

Table 1. Number of patients in different breast cancer subtypes per treatment arm in the discovery and test cohorts. Zero indicates that a given regimen was not tested in that subtype

Cohort	Treatment name	Treatment type	HR+/HER2− subtype	HER2+ subtype	TNBC subtype
Discovery	Paclitaxel	Control	152	0	130
Discovery	Paclitaxel + durvalumab + olaparib	Experimental	52	0	20
Test	Paclitaxel	Control	93	0	78
Test	Paclitaxel + trastuzumab	Control	0	30	0
Test	Regimen 1	Experimental	32	0	37
Test	Regimen 2	Experimental	61	0	52
Test	Regimen 3	Experimental	46	0	42
Test	Regimen 4	Experimental	57	0	45
Test	Regimen 5	Experimental	26	0	30
Test	Regimen 6	Experimental	17	61	32
Test	Regimen 7	Experimental	39	0	28
Test	Regimen 8	Experimental	0	19	0
Test	Regimen 9	Experimental	0	29	0
Test	Regimen 10	Experimental	0	43	0
Test	Regimen 11	Experimental	0	50	0

HER2, human epidermal growth factor receptor-2; HR, hormone receptor; TNBC, triple-negative breast cancer.

cyclophosphamide after completion of the paclitaxel or experimental treatments.

TES

A widely used method to measure distance between two empirical cumulative distribution functions (eCDF) is the KS test. The KS test, however, was designed for continuous data and assumes that the data do not contain repeated values. This assumption is not valid for RCB distributions that contain many zeros (i.e. pCR), and therefore we modified KS to develop a new metric. A weight function was added for calculating eCDF which varies the importance of different RCB scores. The weighted eCDF of continuous variable $x = (x_1, \ldots, x_n)$ is defined as:

$$wF_n(x) = \frac{1}{\sum_{i-1}^{n} w(x_i)} \sum_{i-1}^{n} w(x_i) * 1_{xi \leq x} \quad (1)$$

where x is the RCB score, 1_A is the indicator of event A and $w(x)$ is a weight function. $wF_n(x)$ is a step function which jumps at the unique values of x. The height of the jump at a given point is the total number of tied observations at that value scaled by weight at that value. The following non-negative weight function is used:

$$w(x) = \frac{2}{1 + e^{(xaie^* x)}} \quad (2)$$

where x is the RCB score, and scale is a parameter controlling the shape of weight function $w(x)$. For scale equal to 0, $w(x)$ is equal to 1 for any x and $wF_n(x)$ equal to standard eCDF. As a result, higher positive scale gives higher importance to low RCB score values, whereas lower negative scale gives higher importance to high RCB scores. We define the TES as the area under the difference between wF and wG, not supremum as in KS, that summarizes in a single value the overall benefit, or inferiority, of the experimental treatment. wF and wG are weighted eCDF calculated for RCB scores (x) in the experimental and control arms, respectively. TES is denoted as:

$$\text{TES} = \int_0^{+\infty} (wF(x) - wG(x))dx \quad (3)$$

TES ranges from -1 to $+1$. The larger the positive TES value, the greater the overall shift to smaller RCB values in the experimental arm compared with the control. Statistical significance for TES was calculated using random permutation of control and experimental cohort labels to obtain a null distribution. The observed TES values were compared with the random null distribution and a one-sided P-value was calculated as the proportion of permutations where the calculated statistic was $>$TES.

Subsampling and simulation experiments

A robust efficacy metric gives the same answer over a wide range of population sizes and in the presence of noise. To assess whether our proposed metric is technically robust, four datasets were created, either by subsampling from the discovery cohort, or by creating artificial RCB values using Gaussian mixture models (GMMs).[13] We compared the robustness of TES with KS, MW, and F statistical tests in these datasets. TES, KS, and MW were calculated on continuous RCB values, while the F test was calculated on pCR rates. All tests were one-sided. Spearman (for monotonic relationship) and Pearson (for linear relationship) correlations were used to assess the correlation of TES and differences in pCR rates (ΔpCR) between treatment arms. A one-sided t-test was used to compare correlation coefficients. In all analyses, statistical significance was set to 0.05.

False-positive rate control was evaluated in the discovery set with randomly permuted assignment of patients to control or experimental arms in 10 000 iterations. Robustness to small sample size was evaluated by stratified subsampling from the discovery cohort to create sample sizes between 10% and 90% of the original cohort size, but keeping the proportion of patients in the experimental to control arms the same. Each sampling was repeated 50 times. Robustness to sample size imbalance between the experimental and control arms was evaluated by subsampling the experimental arm from the discovery cohort while keeping the control cohort size the same. Experimental treatment cohort sizes varied between $n = 5$ and $n = 70$. Each subsampling was repeated 50 times. The power and false-positive rate control were evaluated on simulated cohorts with artificially created RCB scores. A two-component GMM was fitted to the distribution of RCB scores in the discovery cohort, excluding cases with pCR, using GaMRed.[1] The estimated parameters of the Gaussians are: (i) $\mu = 1.4648$, $\sigma = 0.44497$; (ii) $\mu = 3.1706$, $\sigma = 0.76035$. Another component was added to simulate the pCR group. The proportions of patients with pCR, RCB drawn from the first component, and RCB drawn from the second component, estimated using GMM weight parameters and pCR rate from the discovery cohort, were equal to 0.3133, 0.33109, and 0.35561, respectively. We created multiple artificial cohorts including 100 patients assigned to control and 100 to experimental treatment using this model. To simulate the difference in RCB scores between treatments, the number of patients with reduced RCB score in the experimental arm compared with control treatment was varied between 0 and 20. For example, reduction of RCB score in one patient in the experimental cohort could mean that there is one additional patient with pCR and one less patient with a non-zero RCB value simulated using second GMM component parameters. Each experiment was repeated 50 times.

Survival analysis

We used the Kaplan–Meier (KM) method to create EFS and distant recurrence-free survival (DRFS) curves and estimated the restricted mean survival time (RMST)[14,15] at 4 years follow-up. RMST corresponds to the area under the

KM curve up to a specific follow-up time. We chose RMST as a survival measure because it is less susceptible to uncertainty than Cox regression hazard ratios when the number of events is small, resulting in narrower confidence intervals. The treatment effect is expressed as the difference in RMST between trial arms (ΔRMST) in months, which can be interpreted as the average difference in survival between the two arms. EFS was defined as the time from registration to first local or distant recurrence, new primary cancer, contralateral breast cancer, or death from any cause, and patients who were alive without an event, or died without an event censored at the date of the last follow-up date. DRFS was defined as the time from registration to the first distant recurrence, and patients who were alive without distant recurrence or died without recurrence were censored at the date of the last follow-up. The proportional hazards assumption for a Cox regression model fit was tested using cox.zph function from R survival package (cran.r-project.org).

RESULTS

Comparison of statistical tests to identify the superior trial arm in the discovery cohort

All three standard statistical tests (KS, MW, F) showed significant benefit from durvalumab + olaparib relative to control therapy in the HER2− ($n = 73$) and HR+/HER2− ($n = 52$) subsets. In TNBC ($n = 20$), significance was borderline (Figure 1A). These results are consistent with the reported efficacy of the regimen using Bayesian modelling as a primary endpoint in I-SPY2, which predicted the probability of success in a hypothetical phase III trial to be 0.944, 0.950, and 0.838 in HER2−, HR+/HER2−, and TNBC, respectively.[8] The MW and F tests yielded significant but smaller *P* values than KS, suggesting that KS has less power to detect differences in response between arms.

TES was developed to improve the discriminatory power of KS by introducing weight function (Figure 1B). We empirically tested scale parameters of TES weight function in a range between −0.5 to +0.5 to minimize the *P* value when comparing the two treatment arms in the discovery cohort. The optimum was observed for a scale of −0.136 (Supplementary Figure S1, available at https://doi.org/10.1016/j.annonc.2022.04.072) that assigns higher importance to high RCB values (i.e. a decrease of RCB score from 4 to 3 is more important than a decrease from 3 to 2, which is more important than a decrease from 2 to 1, etc.). The scaled TES showed significant benefit from durvalumab + olaparib in all HER2− ($P = 0.0015$) and in the HR+/HER2− ($P = 0.00047$) subtypes and a trend for benefit in TNBC ($P = 0.073$) (Figure 1C). We note that the small sample size of the TNBC subset implies lesser statistical power relative to the other subsets.

Random assignment of patients into experimental or control arms was used to estimate false-positive rates (i.e. falsely identify an arm as better using pCR rate

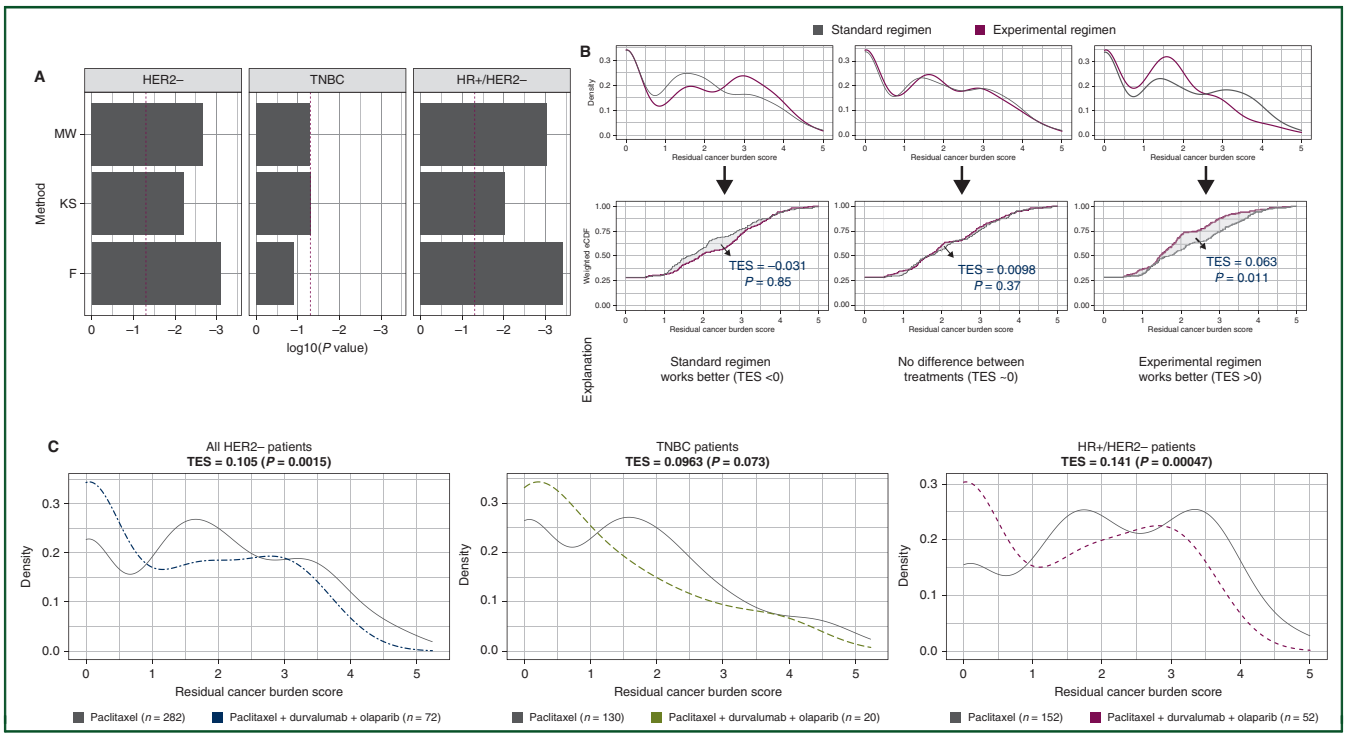

Figure 1. Development of Treatment Efficacy Score in discovery cohort. (A) *P* values from the Kolmogorov–Smirnov (KS), Mann–Whitney *U* (MW), and Fisher's exact (F) tests comparing residual cancer burden (RCB) values between control and experimental arms. Dotted line shows significance level at 0.05. (B). Examples of hypothetical RCB value distributions and their corresponding empirical cumulative distribution functions (eCDF) under different treatment efficacy assumptions. (C) Examples of TES results from I-SPY2 in three breast cancer subtypes. The plots show RCB distributions between two treatment arms with corresponding TES value and significance.

HER2, human epidermal growth factor receptor-2; HR, hormone receptor; TES, Treatment Efficacy Score; TNBC, triple-negative breast cancer.

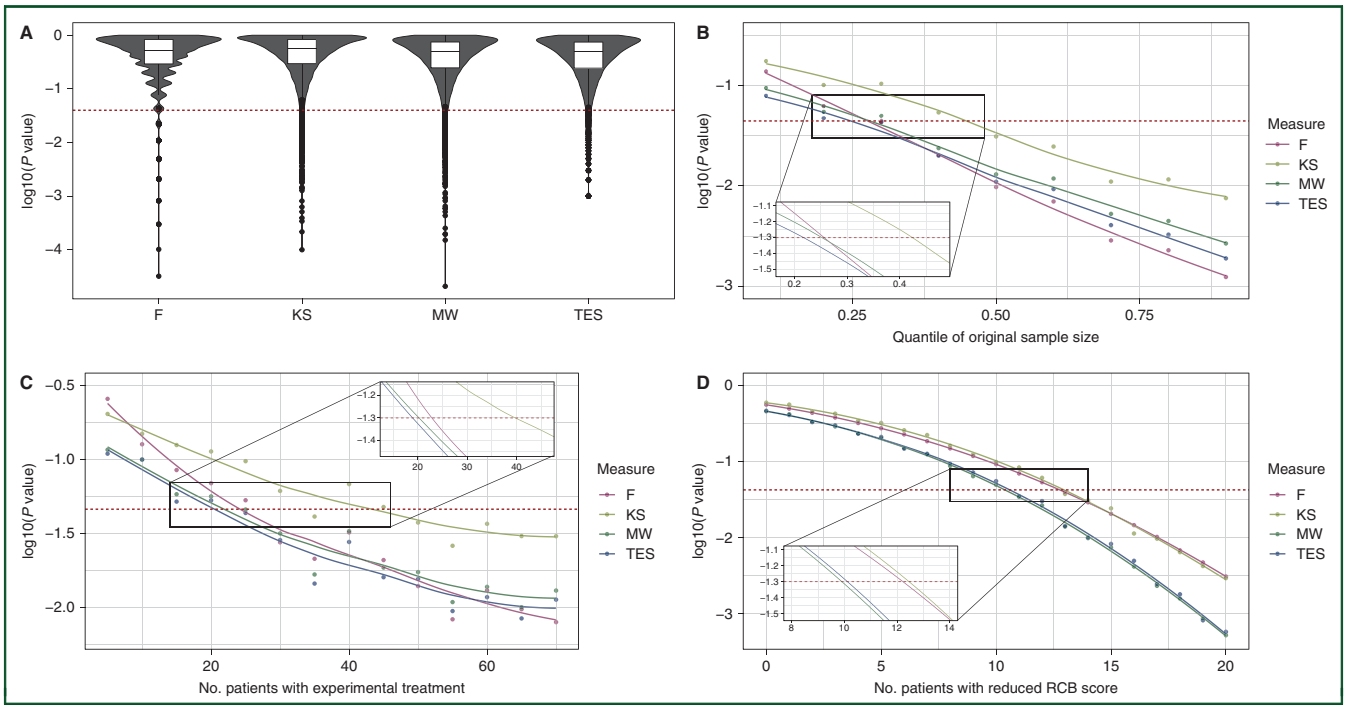

Figure 2. Evaluation of TES performance in the discovery cohort and with simulated data. (A) False-positive rate control. *P* value distributions on randomly permuted patient labels for the four statistical tests. (B) Robustness to small sample size. *P* values for the four tests run with decreasing overall sample sizes. (C) Robustness to treatment arm sample size imbalance. *P* values for tests run with subsampled data for the experimental arm only while keeping the control arm the same ($n = 282$). (D) Power comparison of the test. *P* values for tests run on artificial data with controlled number of high/low RCB score patients. In each plot, the solid lines indicate less smoothing function, and the broken lines show significance level at 0.05.

F, Fisher's exact test; KS, Kolmogorov–Smirnov test; MW, Mann–Whitney test; RCB, residual cancer burden; TES, Treatment Efficacy Score.

difference as the gold standard) (Figure 2A). At a significance level of 0.05, all tests yielded <5% false-positive findings. At small sample sizes, TES reached significance more often than the other methods, indicating the highest power (Figure 2B, Supplementary Figures S1 and S2A, available at https://doi.org/10.1016/j.annonc.2022.04.072).

We also estimated how imbalance in trial arm sizes influences test results by subsampling only from the experimental arm and comparing these to all patients in the control arm (Figure 2C, Supplementary Figure S2B, available at https://doi.org/10.1016/j.annonc.2022.04.072). On average, F, MW, and TES had similar power to detect treatment benefit even with only 25 patients in the experimental arm (8.8% of all samples); KS was the least robust to trial arm sample size imbalance.

Using a mixture model (two normal distributions + constant pCR rate), we generated 200 RCB values to simulate a trial result of $n = 100$ controls and $n = 100$ experimental treatment, to compare the power of the four methods under different RCB distributions (Figure 2D, Supplementary Figure S2C, available at https://doi.org/10.1016/j.annonc.2022.04.072). The performance differences were small, but MW and TES showed the highest power. The TES metric was linearly proportional to the number of patients with lower RCB values in the experimental arm relative to the control ($r = 0.92$, $P < 1 \times 10-16$) demonstrating its utility as an effect size measure (Supplementary Figure S2D, available at https://doi.org/10.1016/j.annonc.2022.04.072).

Assessing TES in independent trial arms

We also assessed the ability of the four tests to identify the more effective experimental therapies in the different molecular subtypes and examined the consistency of the results in the remaining 11 previously reported arms and corresponding controls of I-SPY2 (Figure 3). All tests identified the same experimental treatments as superior, with two exceptions (Figure 4A). In TNBC, experimental regimen 4 was identified as significantly better than the control by F, MW, and TES tests, but not by the KS test. In HER2+ cancers, regimen 11 was borderline significant with TES ($P = 0.059$), but was significant with all other methods. Regimen 11 has graduated in the HER2+ subtype according to efficacy analysis rules of I-SPY2[16]; regimen 4 showed numerically higher pCR rates in TNBC but did not meet the prespecified threshold for graduation of I-SPY.[17] Inspection of the RCB distributions of regimen 11 revealed that the treatment moved many RCB values between 1 and 2 to <1 or to 0, and the proportion of patients with any RCB value >2 was very low in both arms resulting in TES having low power to identify this arm as statistically superior.

The correlation of test statistics (Figure 4B, upper triangle) and *P* values (Figure 4B, lower triangle) between the four statistical methods was high. TES correlated with ΔpCR closely, but captured additional features of response (Figure 4C). Examples of different RCB distributions from arms with significantly increased pCR rates but lower TES values are shown in Figure 4C. In HR+/HER2− cancers,

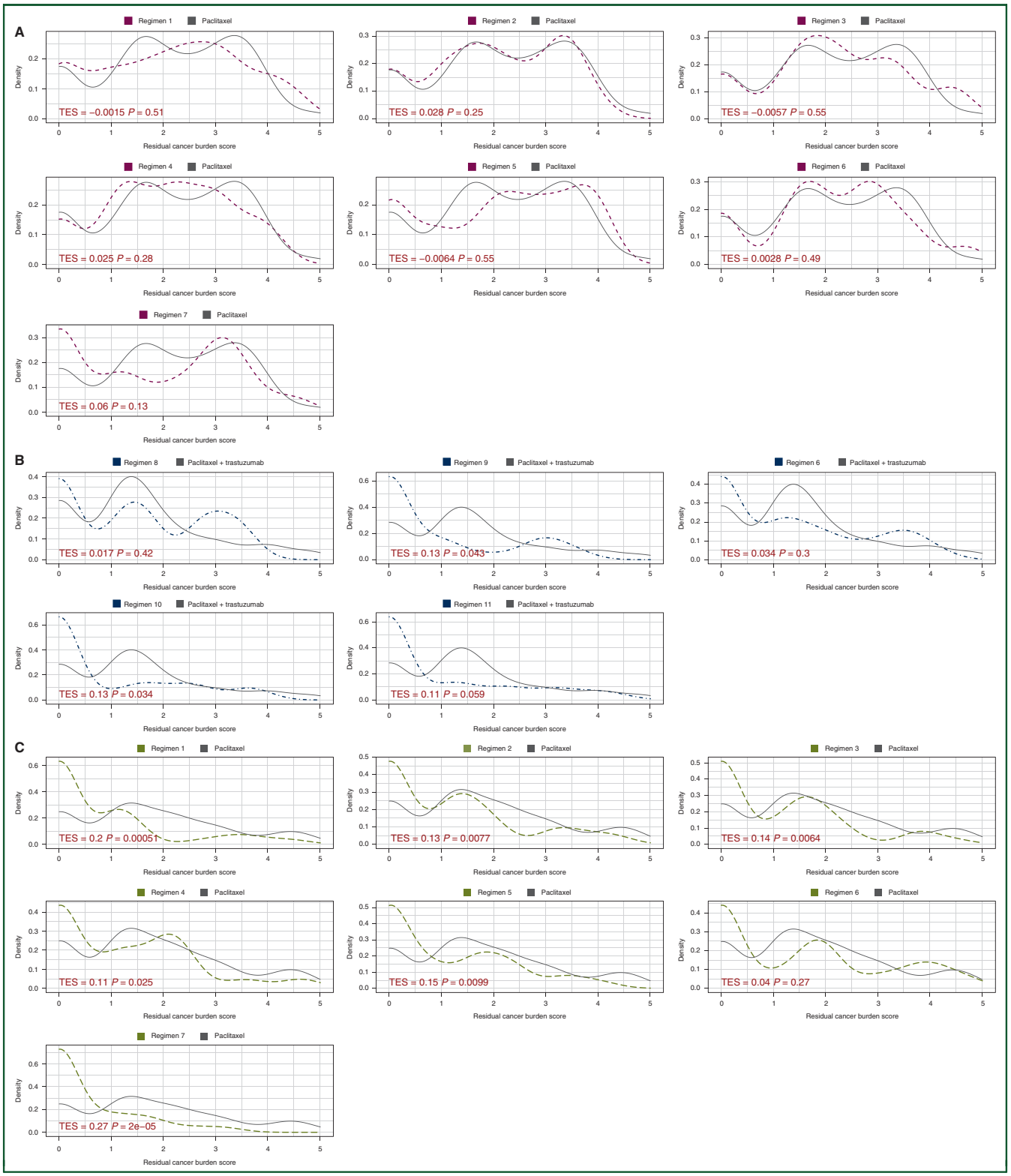

Figure 3. Density plots of residual cancer burden score distributions in (A) HR+/HER2− cancers, (B) HER2+ cancers, (C) TNBC. Density plots of residual cancer burden scores in the experimental arms are represented by the broken line and control arms by the solid line. TES score and its *P* value are shown in the bottom left corner of each plot.

HER2, human epidermal growth factor receptor-2; HR, hormone receptor; TES, Treatment Efficacy Score; TNBC, triple-negative breast cancer.

regimen 7 shows increased pCR and a reduced proportion of patients with RCB values between 1 and 2, but without effecting the proportion of patients with higher RCB values, hence the low TES of 0.059. In TNBC, the same regimen results in a consistent down shift in RCB values across the entire range, leading to a high TES of 0.27 (Figure 3). For regimen 6 in TNBC, increased pCR rate was borderline significant with the F test ($P = 0.054$), but the proportion of

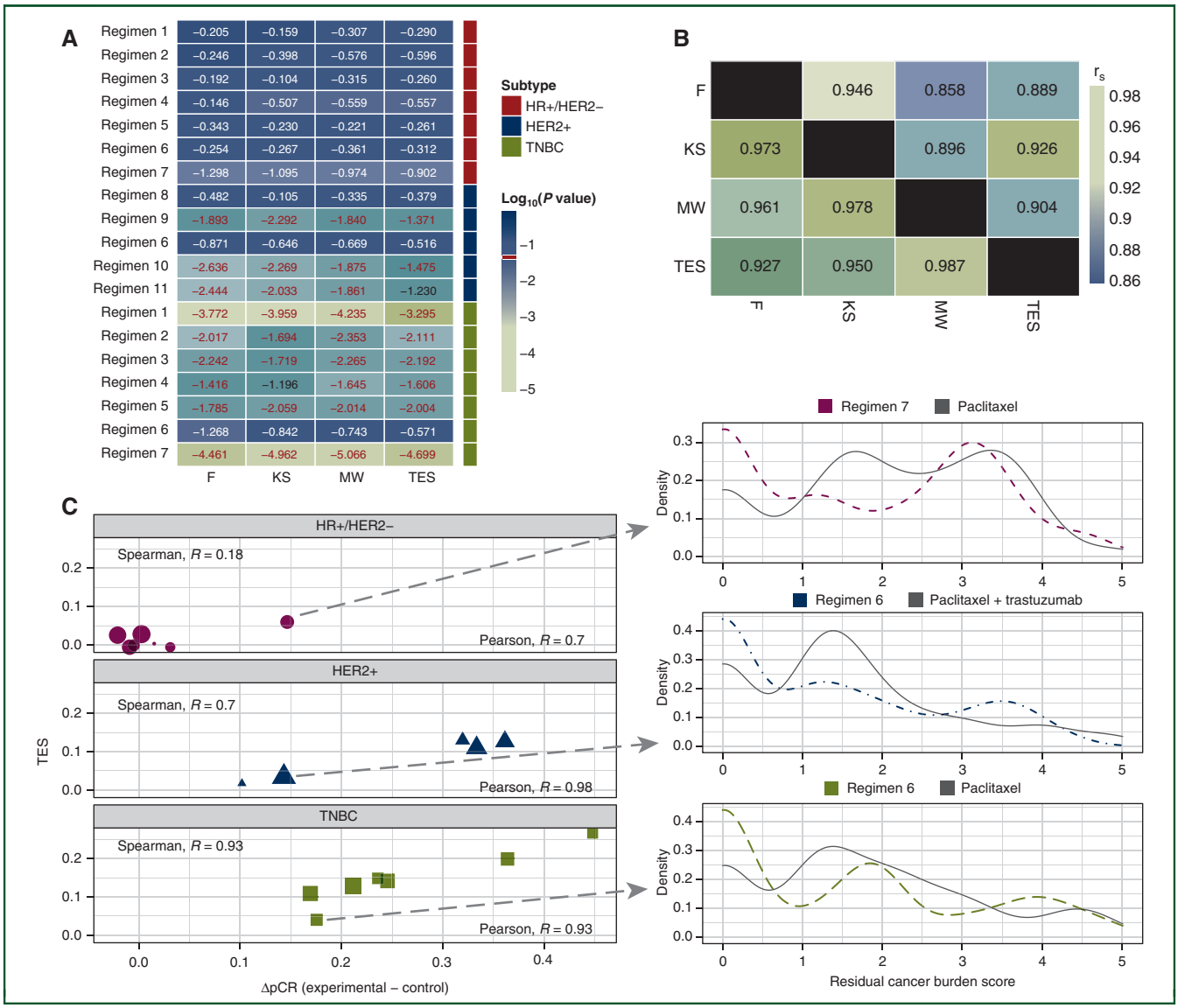

Figure 4. Assessment of TES in independent cohorts. (A) *P*-values from four statistical tests comparing different residual cancer burden (RCB) values between control and experimental treatments in 11 treatment arms (rows) grouped by the three molecular subtypes. The same treatment can be tested in more than one subtype. Red font shows significant result at $P \leq 0.05$. (B) Spearman correlation between test statistics (upper triangle) and *P* values (lower triangle) across all treatments and subtypes combined. (C) Associations between difference in pathologic complete response rates (experimental versus control treatment) and TES for different I-SPY2 trial arms by subtype. On the right side, arrows point to examples of RCB score distributions between treatments and corresponding control from three distinct treatment arms. Circle/dashed line = HR+/HER2−, triangle/dot/dashed line = HER2+, square/long dashed line = TNBC.

F, Fisher's exact test; HER2, human epidermal growth factor receptor-2; HR, hormone receptor; KS, Kolmogorov–Smirnov test; MW, Mann–Whitney test; pCR, pathologic complete response; TES, Treatment Efficacy Score; TNBC, triple-negative breast cancer.

high RCB values was not affected by treatment, leading to low TES (Figure 4C). These illustrate how pCR rate difference can be decoupled from treatment effect on non-zero RCB distribution.

Association of TES with patient survival

The ability to assess the entire distribution of pathologic response makes TES a promising candidate to predict the survival impact of a regimen more accurately than the pCR rate difference alone. Due to the short follow-up of I-SPY2, and violation of proportional hazards in six arms, we used the difference in RMST (ΔRMST) at 4 years, which is the maximum follow-up observed for all 11 treatment arms, as the measure of EFS and DRFS benefit. For both survival endpoints, we observed a stronger correlation between TES and ΔRMST (Spearman R = 0.78 for EFS, R = 0.74 for DRFS), than between ΔpCR and ΔRMST (Spearman R = 0.65 for EFS, R = 0.64 for DRFS) (Figure 5A and B). TES also showed higher correlation with survival than estimates given by F, KS, and MW (Supplementary Figure S3, available at https://doi.org/10.1016/j.annonc.2022.04.072). These differences between correlation coefficients did not reach statistical significance when all treatment arms were considered. In many arms and in several molecular subtypes, the experimental treatment was not superior to control using any of the pathologic response metrics and

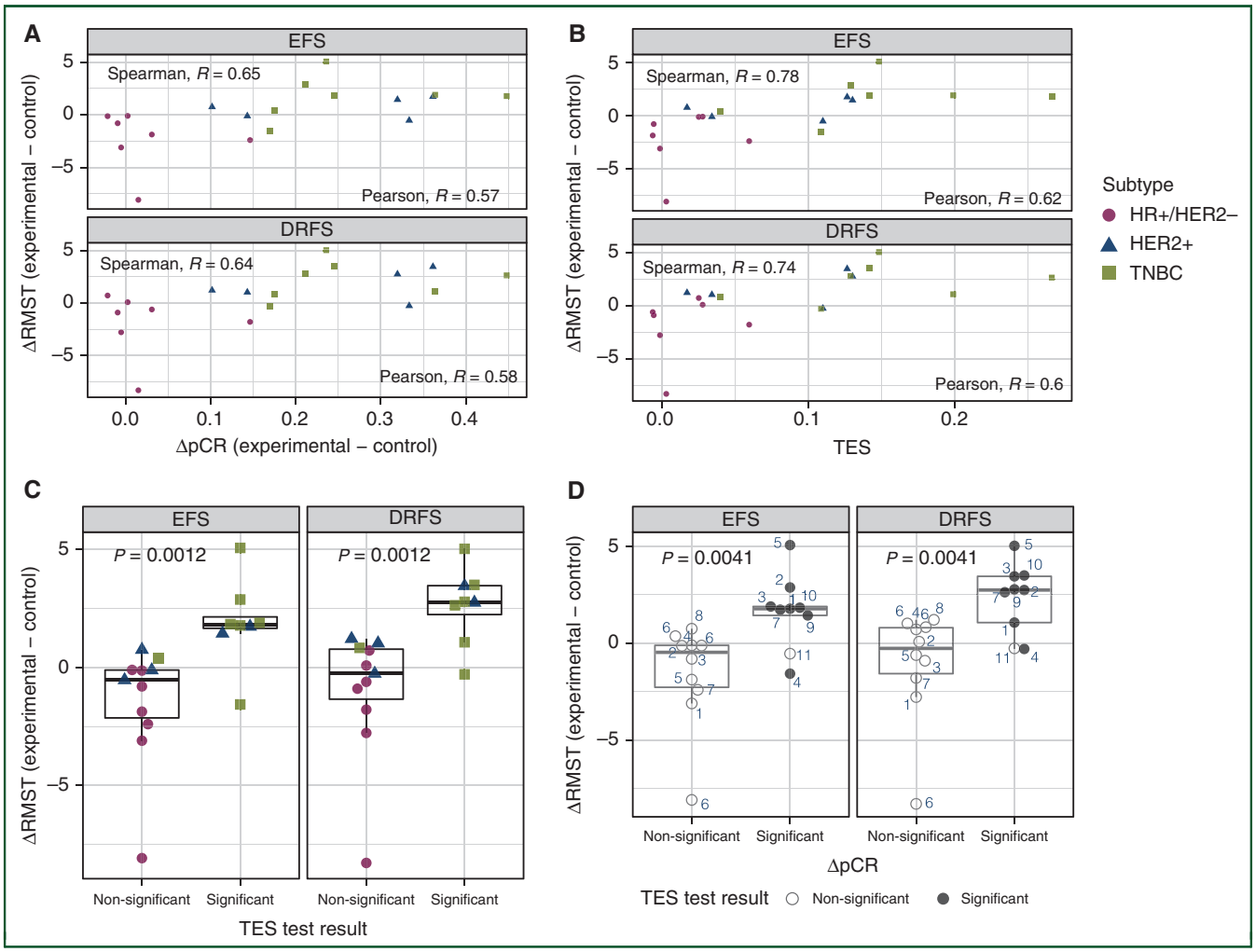

Figure 5. Association between TES and trial arm level survival. (A) Association between pCR rate difference (ΔpCR) and difference in restricted mean survival time (ΔRMST) for event-free survival (EFS) and distant recurrence-free survival (DRFS). (B) Association between TES and ΔRMST for the two survival endpoints. (C) Comparison of ΔRMST between treatment arms that showed significant versus non-significant improvement in TES. (D) Comparison of ΔRMST between treatment arms that showed significant versus non-significant improvement in ΔpCR; each comparison is annotated for being significant for TES (filled dots) or not (open circles). *P* values are from Mann—Whitney test.

HER2, human epidermal growth factor receptor-2; HR, hormone receptor; pCR, pathologic complete response; TES, Treatment Efficacy Score; TNBC, triple-negative breast cancer.

therefore these arms are non-informative in response and survival correlation analysis. When we analyzed only the treatment arms that showed a significant superiority over control using TES, these arms also had significantly higher survival improvement compared with survival improvement seen in arms with non-significant TES (Figure 5C). This supports that a statistically significant TES predicts statistically significant improvement in survival. TES also efficiently captured the traditional pCR rate improvement, because when the pCR rate difference was significant TES was also significant (Figure 5D).

On-line web tools

The R scripts to calculate TES are freely available at GitHub: https://github.com/mmarczyk363/RCB_compare. We also created a free web tool to calculate TES that can be accessed at http://dssoftware.aei.polsl.pl/TES/TES/. The tool requires RCB values from two treatment cohorts in tab-delimited text file format as input and plots RCB distributions, TES, TES *P* value, and provides an estimate of the predicted 4-year ΔRMST (Supplementary Figure S4, available at https://doi.org/10.1016/j.annonc.2022.04.072).

DISCUSSION

Different extents of residual cancer after neoadjuvant chemotherapy have significantly different recurrence-free survivals, and the difference in the proportion of the most favorable prognostic group, pCR/RCB 0, became the dominant metric for efficacy comparisons in neoadjuvant trials. Focusing on pCR ignores the effect of treatment on the non-zero residual cancer distribution. If the treatment-induced improvements were proportional across all RCB values, pCR could serve as a good surrogate of the overall treatment effect. This is, however, rarely observed (Figure 3). Not considering the effect of treatment on non-zero RCB values

may partly explain the difficulty of establishing a clear relationship between pCR rate improvement and improvement in survival at the trial arm level.[18,19]

We developed a statistical tool to compare two different RCB distributions and summarize the difference between them in a single metric, the TES. When we compared the performance of TES with the standard KS test, MW *U* test, or comparing pCR rate differences with the F test under various scenarios including small sample size, unbalanced samples size, and a range of RCB distributions, TES performed equally well, and often better, than these other methods to identify the more cytotoxic regimen. In independent trial arms, all four statistical tests could identify regimens with greater pCR rates. In instances when TES showed only a borderline significant effect despite significant improvement in pCR rate by other methods, inspection of the RCB distributions revealed that TES is more sensitive to a lack of difference in the high end of the RCB value distributions. When an experimental regimen does not result in a downshift in high RCB values, TES may not identify it as more effective, even if the pCR rate is somewhat higher in the experimental arm. This feature could provide an advantage for TES in capturing the impact of treatment on survival.

We found that the correlation between survival and TES was higher than the correlation between survival and pCR rate difference. We also found a significantly greater improvement in EFS and DRFS in the experimental arms compared with controls when TES was statistically significant. Among TNBC, the highest TES, indicating a broad and consistent shift to smaller RCB values across the entire spectrum, were seen in the two immunotherapy arms, pembrolizumab and durvalumab + olaparib, respectively. Both arms graduated in I-SPY2 and KEYNOTE-522 demonstrated improved invasive disease-free survival with the inclusion of pembrolizumab with neoadjuvant chemotherapy, even with the modest final ΔpCR of 7%.[7] This result is consistent with our hypothesis that TES might predict long-term survival differences between trial arms more accurately than ΔpCR.

Our study has limitations. The combined sample size of our test cohort is large, 947 patients; however, the individual treatment arms vary between $n = 19$ to $n = 171$. Because the efficacy of neoadjuvant chemotherapy regimens varies by molecular subtypes, we further subdivided the treatment arms into the three major receptor subtypes resulting in even smaller cohorts for final comparisons. Survival events are still few and the median duration of follow-up is unequal among experimental arms, which further limits the power of survival analyses and comparisons between arms. However, I-SPY2 is the only trial that has RCB data and multiple treatment regimens across multiple molecular subtypes with survival information that allowed us to assess correlation between TES and survival across many arms and molecular subsets. There are also inherent limitations of pCR and primary tumor response as predictors of future survival events. For example, central nervous system (CNS) recurrences are similar across RCB classes (including RCB 0) confirming that the CNS is a treatment sanctuary site.[20] Tumor response also cannot inform about the risk of a future second primary cancer in the breast or elsewhere, or death unrelated to breast cancer. These events weaken associations between RCB metrics and overall survival. As additional larger data sets become available, however, the weight functions could be further optimized and data from other studies will allow independent validation of TES as an early surrogate for trial arm level improvement in recurrence-free survival. Additional validation of our model will be required on the current standard of care regimens by independent trial groups before wide adoption of this proposed new efficacy metric. To facilitate this, we made the R scripts for TES available at GitHub and created a free web tool for clinical trialists to calculate the TES statistic.

In summary, TES is a novel metric to identify a more effective cytotoxic neoadjuvant regimen from the entire distribution of pathologic responses. TES significantly correlates with survival and may serve as a better early surrogate for EFS and DRFS than pCR rate difference.

ACKNOWLEDGEMENTS

The authors sincerely appreciate the ongoing support for the I-SPY2 trial from the Safeway Foundation, the William K. Bowes, Jr. Foundation, and Give Breast Cancer the Boot. Initial support was provided by Quintiles Transnational Corporation, AstraZeneca, Johnson & JohnsonJohnson & Johnson, Genentech, Amgen, the San Francisco Foundation, Eli Lilly, Pfizer, Eisai Co., Ltd, Side Out Foundation, Harlan Family, the Avon Foundation for Women, Alexandria Real Estate Equities, and private individuals and family foundations.

FUNDING

This work was supported by a Breast Cancer Research Foundation Investigator Award (AWDR11559) to LP. The I-SPY2 trial is supported by Quantum Leap Healthcare Collaborative (2013 to present) (no grant number) and the Foundation for the National Institutes of Health (2010 to 2012) (no grantf number), a grant from the Gateway for Cancer Research [grant number G-16-900], and by a grant from the National Cancer Institute Center for Biomedical Informatics and Information Technology [grant number 28XS197].

DISCLOSURE

RN: Consulting from Cardinal Health, Clovis, Fujifilm, G1 Therapeutics, Genentech, Immunomedics/Gilead, iTeos, MacroGenics, Merck, OncoSec, Pfizer, Seattle Genetics; Research Funding to Institution from Arvinas, AstraZeneca, Celgene, Corcept Therapeutics, Genentech/Roche, Immunomedics/Gilead, Merck, OBI Pharma, Odonate Therapeutics, OncoSec, Pfizer, Seattle Genetics, Taiho. BAP: Consulting from BioAtla Inc, Samumed LLC, Dare Biosciences; Stock of Merck; Research Funding to Institution

from Pfizer, GlaxoSmithKline, Novartis, Genentech/Roche, Oncternal. KSA: Grants from Seattle Genetics, Daiichi Sankyo, AstraZeneca. RKM: Participation on a Data Safety Monitoring Board or Advisory Board of Genomic Health/Exact Sciences, Genentech-Roche, Seattle Genetics/Axio; Grants from Seattle Genetics, Daiichi Sankyo, AstraZeneca; Support from Quantum Leap Healthcare Collaborative, Merck, Seattle Genetics, Amgen, Genentech-Roche. JCB: Research Funding from Eli Lilly. MCL: Funding from Eisai, Exact Sciences, Genentech, Genomic Health, GRAIL, Menarini Silicon Biosystems, Merck, Novartis, Seattle Genetics, Tesaro; Support for attending meetings and/or travel from AstraZeneca, Genomic Health, Ionis; Participation on a Data Safety Monitoring Board or Advisory Board of AstraZeneca, Celgene, Roche/Genentech, Genomic Health, GRAIL, Ionis, Merck, Pfizer, Seattle Genetics, Syndax. ASC: Institutional Research Funds from Novartis. HSR: Funding from Pfizer, Merck, Novartis, Lilly, Roche, Daiichi, Seattle Genetics, MacroGenics, Sermonix, Boehringer Ingelheim, AstraZeneca, Ayala, Gilead, and Ayala; Honoraria from Puma, Merck, Samsung, NAPO. JP: Honoraria from Methods in Clinical Research; Support for attending meetings from ASCO, SABCS; Participation on a Data Safety Monitoring Board or Advisory Board of University of Wisconsin Specialized Programs of Research Excellence, VIVLI, Quantum Leap Healthcare Collaborative, Patient Centered Outcomes Institute. DAB: Employee/Leadership, Stock/Ownership, and Consulting/Advisory Board of Berry Consultants; Research Funding from Daiichi Sankyo; Travel/Accommodations/Expenses from Berry Consultants. LV: part-time employee and stockholder of Agendia NV. WFS: Stock owner in Delphi Diagnostics; Patent - "Method of measuring residual cancer and predicting patient survival" (US Patent 7711494B2). LE: Research Funding from Merck; Medical Advisory Panel member of Blue Cross Blue Shield; Website Author of UpToDate. LP: Consulting fees and honoraria from Pfizer, AstraZeneca, Merck, Novartis, Genentech, Eisai, Pieris, Immunomedics, Seattle Genetics, Almac, Biotheranostics, and Natera. All other authors have declared no conflicts of interest.

REFERENCES

1. Symmans WF, Peintinger F, Hatzis C, et al. Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. *J Clin Oncol*. 2007;25(28):4414-4422.
2. Hamy A-S, Darrigues L, Laas E, et al. Prognostic value of the Residual Cancer Burden index according to breast cancer subtype: validation on a cohort of BC patients treated by neoadjuvant chemotherapy. *PLoS ONE*. 2020;15(6):e0234191.
3. Müller HD, Posch F, Suppan C, et al. Validation of residual cancer burden as prognostic factor for breast cancer patients after neoadjuvant therapy. *Ann Surg Oncol*. 2019;26(13):4274-4283.
4. Symmans WF, Wei C, Gould R, et al. Long-term prognostic risk after neoadjuvant chemotherapy associated with residual cancer burden and breast cancer subtype. *J Clin Oncol*. 2017;35(10):1049-1060.
5. Spring LM, Fell G, Arfe A, et al. Pathologic complete response after neoadjuvant chemotherapy and impact on breast cancer recurrence and survival: a comprehensive meta-analysis. *Clin Cancer Res*. 2020;26(12):2838-2848.
6. Symmans WF, Yau C, Chen Y-Y, et al. Assessment of residual cancer burden and event-free survival in the adaptively 1 randomized I-SPY2 trial of neoadjuvant treatment for high-risk breast cancer. *JAMA Oncol*. 2021;7(11):1654-1663.
7. Schmid P, Cortes J, Dent R, et al. Event-free survival with pembrolizumab in early triple-negative breast cancer. *N Engl J Med*. 2022;386:556-567.
8. Loibl S, Untch M, Burchardi N, et al. A randomised phase II study investigating durvalumab in addition to an anthracycline taxane-based neoadjuvant therapy in early triple-negative breast cancer: clinical results and biomarker analysis of GeparNuevo study. *Ann Oncol*. 2019;30(8):1279-1288.
9. Pusztai L, Yau C, Wolf DM, et al. Durvalumab with olaparib and paclitaxel for high-risk HER2-negative stage II/III breast cancer: results from the adaptively randomized I-SPY2 platform trial. *Cancer Cell*. 2021;39(7):989-998.
10. Park JW, Liu MC, Yee D, et al. Adaptive randomization of neratinib in early breast cancer. *N Engl J Med*. 2016;375(1):11-22.
11. Rugo HS, Olopade OI, DeMichele A, et al. Adaptive randomization of veliparib-carboplatin treatment in breast cancer. *N Engl J Med*. 2016;375(1):23-34.
12. Barker AD, Sigman CC, Kelloff GJ, Hylton NM, Berry DA, Esserman LJ. I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clin Pharmacol Ther*. 2009;86(1):97-100.
13. Marczyk M, Jaksik R, Polanski A, Polanska J. GaMRed—adaptive filtering of high-throughput biological data. *IEEE/ACM Trans Comput Biol Bioinform*. 2020;17(1):149-157.
14. Uno H, Claggett B, Tian L, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J clin Oncol*. 2014;32(22):2380.
15. Rahmadian AP, Santos SD, Parshad S, Everest L, Cheung MC, Chan KK. Quantifying the survival benefits of oncology drugs with a focus on immunotherapy using restricted mean survival time. *J Natl Compr Canc Netw*. 2020;18(3):278-285.
16. Clark AS, Yau C, Wolf DM, et al. Neoadjuvant T-DM1/pertuzumab and paclitaxel/trastuzumab/pertuzumab for HER2+ breast cancer in the adaptively randomized I-SPY2 trial. *Nat Commun*. 2021;12:6428.
17. Yee D, Isaacs C, Wolf DM, et al. Ganitumab and metformin plus standard neoadjuvant therapy in stage 2/3 breast cancer. *NPJ Breast Cancer*. 2021;7(1):131.
18. Hatzis C, Symmans WF, Zhang Y, et al. Relationship between complete pathologic response to neoadjuvant chemotherapy and survival in triple-negative breast cancer. *Clin Cancer Res*. 2016;22(1):26-33.
19. Cortazar P, Zhang L, Untch M, et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. *Lancet*. 2014;384(9938):164-172.
20. Yau C, DeMichele A, Symmans WF, et al. Abstract P2-20-02: Site of recurrence after neoadjuvant therapy: clues to biology and impact on endpoints. *Cancer Res*. 2020;80(suppl 4):P2-20-02-P22-20-02.