Comparative survival analysis of multiparametric tests—when molecular tests disagree—A TEAM Pathology study

John M. S. Bartlett1,2,3,19,25, Jane Bayani1,19, Elizabeth Kornaga1,2,4,19, Keying Xu1, Greg R. Pond2, Tammy Piper3, Elizabeth Mallon6, Cindy Q. Yao7, Paul C. Boutros7,8,9,10, Annette Hasenburg11, J. A. Dunn12, Christos Markopoulos13, Luc Dirix14, Caroline Seynaeve15, Cornelis J. H. van de Velde16, Robert C. Stein17, and Daniel Rea18

Multiparametric assays for risk stratification are widely used in the management of both node negative and node positive hormone receptor positive invasive breast cancer. Recent data from multiple sources suggests that different tests may provide different risk estimates at the individual patient level. The TEAM pathology study consists of 3284 postmenopausal ER+ve breast cancers treated with endocrine therapy Using genes comprising the following multi-parametric tests OncotypeDx, Prosigna™ and MammaPrint® signatures were trained to recapitulate true assay results. Patients were then classified into risk groups and survival assessed. Whilst likelihood \( \chi^2 \) ratios suggested limited value for combining tests, Kaplan–Meier and LogRank tests within risk groups suggested combinations of tests provided statistically significant stratification of potential clinical value. Paradoxically whilst Prosigna-trained results stratified Oncotype-trained subgroups across low and intermediate risk categories, only intermediate risk Prosigna-trained cases were further stratified by Oncotype-trained results. Both Oncotype-trained and Prosigna-trained results further stratified MammaPrint-trained low risk cases, and MammaPrint-trained results also stratified Oncotype-trained low and intermediate risk groups but not Prosigna-trained results. Comparisons between existing multiparametric tests are challenging, and evidence on discordance between tests in risk stratification presents further dilemmas. Detailed analysis of the TEAM pathology study suggests a complex inter-relationship between test results in the same patient cohorts which requires careful evaluation regarding test utility. Further prognostic improvement appears both desirable and achievable.

npj Breast Cancer (2021) 7:90; https://doi.org/10.1038/s41523-021-00297-7

INTRODUCTION

Multi-parametric molecular tests are central to the treatment management of early breast cancer and their use is incorporated into most major guidelines1 as a pre-requisite for the staging of breast cancer patients, to direct prognostication and to select patients for chemotherapy treatment2,3. Two major challenges related to their use need to be addressed. Firstly, reports highlighting disagreements between tests are disquieting for physicians, health care providers, and patients alike4 since they raise the question “have I recommended/received the right test?” Secondly, the lack of consistency at an individual patient level between different tests suggests additional prognostic information may result from novel tests. Recent results from the MINDACT and TAILORx studies validate the utility of tests to direct chemotherapy use in node-negative patients5,6, which may be extended as new evidence emerges from retrospective3 or prospective studies7,8. In this context an error in assigning appropriate risk classifications would have significant impact on patient treatment and outcomes. Additionally, given recent evidence documenting the long-term risk of relapse for ER+ve breast cancer and the increasing use of extended endocrine therapy9 the selection of the appropriate test to detect recurrence risk over extended time periods is also critical.

Reports of disagreements between tests, based on in silico analyses of existing expression array data, were frequently attributed to methodological challenges and incomplete gene coverage10-14. However, recently direct comparisons, where tests were performed exactly to vendor protocols, demonstrate marked disagreement in risk categorization and subtyping of individual tumors between widely used multiparameter assays15. Furthermore, comparisons between tests in clinical trials derived cohorts provide consistent evidence that combining test results generally improves prognostic value15,16. These results may reflect the relatively modest performance of individual multiparametric tests17.

To date, no direct comparison between different multiparameter assays in a large patient cohort with associated follow-up provides robust information on the impact of discrepant test results for patients. We developed a method to compare signatures using a combined quantitative mRNA array covering key molecular signatures17, trained against the results of the same signatures measured by original methodology18. We analyzed >3000 samples from the TEAM pathology cohort19 using “trained”...
signatures to demonstrate the impact of disagreements between tests on patient outcome in the context of a recent clinical trial cohort.

RESULTS
Comparing signature-trained risk scores—Likelihood ratios
We compared the ability of trained signatures to predict DMFS10 using the likelihood ratio \( \chi^2(\text{LR}) \) based on the Cox models as a measure of the overall prognostic information provided by each model. We illustrated the performance of each “trained” test using Kaplan–Meier survival curves and estimated Hazard ratios as described above (see Fig. 1). We calculated the change in LR values (ΔLR) between the reclassified and single signature models to assess prognostic improvement of reclassification with a second signature versus the single signature using existing trinary and binary (Table 1) cut points as outlined above.

In ER+/HER2—cases \((n = 3284)\), the Prosigna-trained signature provided greater prognostic information compared to Oncotype-trained and MammaPrint-trained signatures (\( \chi^2 = 146.9 \) vs. 118.0 and 119.5, respectively; Table 1). In bivariate models (combining 2 tests) the greatest LR was observed with Oncotype-trained and Prosigna-trained results (Table 1). Comparing bivariate and univariate results combining Oncotype-trained and Prosigna-trained results increased the LR to a far greater extent versus Oncotype-trained results (\( \Delta \text{LR}^2 = 60.0 \)) than versus Prosigna-trained (\( \Delta \text{LR}^2 = 31.0 \)) results. Similarly, when combining tests with Mammaprint-trained results adding Prosigna-trained results showed a greater increase in LR (\( \Delta \text{LR}^2 = 49.3 \)) than did combining Mammaprint-trained results with Oncotype-trained results (\( \Delta \text{LR}^2 = 26.3 \)). Adding Mammaprint-trained results to either Oncotype-trained or Prosigna-trained results to, versus either test produced the smallest improvements in the LR (Table 1). Nonetheless, all test combinations outperformed single tests to a highly statistically significant degree \((p < 0.0001; \text{Table 1})\).

When test results for Oncotype-trained and Prosigna-trained results were dichotomized, there were less marked differences in univariate models between these tests and Mammaprint-trained results (Table 1). Again the largest increase in LR was observed when comparing combined Oncotype-trained and Prosigna-trained classification versus Oncotype-trained alone. All other bivariate models outperformed univariate models to a lesser, but still statistically significant, degree \((p < 0.0001; \text{Table 1})\).

In ER+/HER2— Node •ve, no Chemo (Panel d), Prosigna-trained (Panel e), and Mammaprint-trained results (Panel f). Log-Rank P values for each test are in brackets. Within each panel low (green), moderate (blue) and high (red) risk survival curves are plotted with Log-Rank Hazard ratios for high risk and intermediate risk (Oncotype-trained and Prosigna-trained only) calculated against low risk cases in each subgroup. 95% Confidence intervals for LogRank Hazard ratios are in brackets. For each group the number at risk (Low, moderate, high) are presented under the X axis.
significantly reduced DMFS10 (75.4%) relative to cases low risk by both signatures (HR = 3.19; 95%CI 2.12–4.82; p < 0.001; Table 2; Fig. 2). For Oncotype-trained intermediate-risk cases, 174 (14.8%) were classified as Prosigna-trained low risk with a DMFS10 = 91.5% (p < 0.001; Table 2; Fig. 2), and 618 (52.6%) were classified as Prosigna-trained high risk (DMFS10 = 73.3%; Table 2; Fig. 2). Few Oncotype-trained high-risk tumors were low risk by Prosigna-trained scores and no events were observed in these cases.

Oncotype-trained stratified by MammaPrint-trained

124 Oncotype-trained low-risk cases (8%) were high risk by MammaPrint-trained (DMFS10 = 72.1%; Table 2; Fig. 2; p < 0.001). 52 Oncotype-trained high-risk cases (10%) were low risk by MammaPrint-trained (DMFS10 = 70.4%; Table 2; Fig. 2; p = 0.465). Finally 528 (45%) Oncotype-trained intermediate-risk cases were MammaPrint-trained high risk (DMFS10 = 73.2%; Table 2; Fig. 2; p < 0.001).

Prosigna-trained results. Of 3284 ER+ve/HER2−ve cases with results for Prosigna-trained risk available 25.2% were low risk (DMFS10 = 92.1%, 95%CI 89.8–94.0%), 35.2% intermediate risk (DMFS10 = 84.9%, 95%CI 82.3–87.1%) and 39.7% high risk (DMFS10 = 71.4%, 95%CI 68.6–74.1%; Table 3; Figs. 1b, 3).

Prosigna-trained results stratified by Oncotype-trained results

In Prosigna-trained low-risk cases there were no significant differences in outcome across Oncotype-trained risk groups, all Prosigna-trained low-risk cases experienced DMFS10 > 90% (Table 3; Fig. 3a). Similarly all Prosigna-trained high risk cases experienced a DMFS10 ≤ 80%; those that were also Oncotype-DX-trained high risk experienced significantly poorer outcome (DMFS10 = 65.7%, 95%CI 60.4–70.5%, p < 0.001) than low or intermediate risk by Oncotype-trained (Table 3; Fig. 3c). Of 1155 Prosigna-trained intermediate-risk cases, 685 (59%) were classified low risk by the Oncotype-trained test (DMFS10 = 88.5%; p < 0.001), 89 cases (8%) were Oncotype-trained high risk (DMFS10 = 72.6%; p < 0.001, Table 3; Fig. 3b).

Prosigna-trained stratified by MammaPrint-trained

Excluding Prosigna-trained intermediate-risk cases the majority of results (79.7%) remained in the same risk category (Supplementary Table 2). No stratification of Prosigna-trained low-risk cases occurred using MammaPrint-trained results (Table 3; Fig. 3a). All Prosigna-trained high-risk cases had DMFS10 < 80%, 32% were MammaPrint-trained low risk (Table 3; Fig. 3c). For Prosigna-trained intermediate-risk cases 18% were MammaPrint-trained high risk (DMFS10 = 79.4%; p = 0.005; Table 3, Fig. 3b).

MammaPrint-trained

Of 3284 ER+ve/HER2−ve breast cancers with MammaPrint-Trained risk classification, 66.3% were low risk (DMFS10 = 86.9%) and 33.7% high risk (DMFS10 = 70.7%; Table 4, Figs. 1c, 4).

MammaPrint-trained stratified by Oncotype-trained

Of 2180 MammaPrint-trained low-risk cases, 68% were low risk by Oncotype-trained results (DMFS10 = 89.1%; Table 4; Fig. 4a). MammaPrint-trained low risk Oncotype-trained intermediate-risk cases (30%) exhibited DMFS10 = 83.2% (Table 4, p < 0.001) and Oncotype-trained high-risk cases exhibited DMFS10 = 70.4% (Table 4, p < 0.001; Fig. 4a). In MammaPrint-trained high-risk cases DMFS10 ranged from 73.2–67.3 across Oncotype-trained subgroups and there were marked differences in outcome across Oncotype-trained categories (Table 4, Fig. 4b).

### Table 1. Likelihood ratios by trinary and binary classification.

| Test                  | Trinary classification | Binary classification | df | LRχ² | p-value | df | LRχ² | p-value |
|-----------------------|------------------------|------------------------|----|------|---------|----|------|---------|
| Oncotype              | 2                      | 2                      | 2  | 1.180 | <0.0001 | 1  | 109.87 | <0.0001 |
| Prosigna              | 2                      | 2                      | 2  | 1.469 | <0.0001 | 1  | 127.31 | <0.0001 |
| MammaPrint            | 1                      | 1                      | 1  | 1.195 | <0.0001 | 1  | 119.45 | <0.0001 |

ΔLRχ² = likelihood ratio chi-squared value, all models run excluding at 10 years. Likelihood χ² ratios(LRχ²) for univariate(single test) or bivariate(two tests in sequence) derived using 10-year distant metastasis free survival as end point. ER+ve/HER2−ve cases = all ER+ve/HER2−ve cases (irrespective of nodal status and chemotherapy), ΔLRχ² = change in Likelihood χ² ratio when two tests are used sequentially. Trinary classification: results using results from Oncotype-Dx trained and Prosigna-trained tests categorized as low, intermediate, and high risk, binary classification: results using dichotomous results for all tests, see text for cut-points. ΔLRχ² = change in LRχ² for comparison of 2 tests versus a single test.

**Analysis of test performance by outcome in reclassified patients**

We analyzed agreement between tests by investigating the extent to which re-classifying results for individual patients by performing tests in sequence affected predicted outcome. Example, we estimated the effects of performing a Prosigna-trained test on tumors previously classified as intermediate risk by the Oncotype-trained test.

**Entire ER+ve/HER2−ve population**

**Oncotype-trained.** Of 3284 ER+ve/HER2−ve breast cancers with results for the Oncotype-trained risk classification, 48.9% were classified low risk (DMFS10 = 87.9%), 35.8% intermediate risk (DMFS10 = 78.6%) and 15.3% high risk (DMFS10 = 67.5%) (Table 2; Figs. 1a, 2).

**Oncotype-trained stratified by Prosigna-trained**

When Oncotype-trained results were further stratified by Prosigna-trained results a significant proportion (56.5%) of cases changed risk category (Supplementary Table 2). In Oncotype-trained low-risk cases, 279 (17.4%) were re-classified as high risk by Prosigna-trained results and 9 Oncotype-trained high-risk cases (1.8%) were re-classified as low risk by Prosigna-trained results. Oncotype-trained low risk/Prosigna-trained high-risk cases exhibited a
Table 2. Oncotype-trained results stratified by other test results, trinary classification.

| Oncotype-trained first | Oncotype-trained intermediate risk | Oncotype-trained high risk |
|------------------------|-----------------------------------|----------------------------|
| **HR (95% CI)** | **DMFS (95% CI)** | **P** | **HR (95% CI)** | **DMFS (95% CI)** | **P** | **HR (95% CI)** | **DMFS (95% CI)** | **P** |
| All cases REF 87.9 (85.8-89.6) | <0.001 | 2.03 (1.65-2.50) | 78.6 (75.9-81.1) | <0.001 | 3.47 (2.76-4.34) | 67.5 (62.8-71.7) | <0.001 |
| N-Ch REF 92.5 (88.8-95.0) | <0.001 | 2.08 (1.26-3.43) | 86.3 (81.6-89.9) | 0.004 | 4.06 (2.41-6.83) | 76.7 (68.4-83.0) | <0.001 |
| N-Ch REF 86.4 (83.1-89.1) | <0.001 | 2.03 (1.48-2.78) | 77.0 (72.0-82.1) | <0.001 | 4.64 (3.32-6.47) | 55.1 (46.4-63.0) | <0.001 |
| Ch REF 85.4 (81.2-88.8) | <0.001 | 2.03 (1.45-2.82) | 73.8 (68.8-82.2) | <0.001 | 2.48 (1.58-3.65) | 70.8 (63.0-77.2) | <0.001 |

| Prosinga-trained low | Prosinga-trained Int | Prosinga-trained high |
|-----------------------|----------------------|-----------------------|
| **HR DMFS P** | **HR DMFS P** | **HR DMFS P** |
| All cases REF 92.2 (89.5-94.2) | <0.001 | 1.39 (0.94-2.06) | 88.5 (83.3-91.0) | <0.001 | 3.19 (2.12-4.82) | 75.4 (68.3-81.1) | <0.001 |
| N-Ch REF 97.1 (92.5-98.9) | 0.011 | 2.52 (0.80-7.92) | 92.4 (86.3-95.8) | <0.001 | 5.08 (7.01-91.6) | 83.8 (65.0-99.3) | <0.001 |
| N-Ch REF 89.3 (84.4-92.7) | <0.001 | 0.93 (0.53-1.62) | 89.1 (84.1-92.5) | <0.001 | 2.88 (3.00-5.07) | 70.3 (61.6-79.6) | <0.001 |
| Ch REF 91.9 (86.5-95.2) | <0.001 | 1.98 (1.03-3.80) | 83.4 (76.6-93.9) | <0.001 | 3.76 (57.0-82.1) | 71.7 (67.4-83.0) | <0.001 |

| Mammaprint-trained low | Mammaprint-trained high |
|-------------------------|-------------------------|
| **HR DMFS P** | **HR DMFS P** |
| All cases REF 89.1 (87.1-90.8) | <0.001 | 2.79 (1.83-4.25) | 72.1 (60.9-80.6) | <0.001 | 3.67 (2.69-4.94) | 65.1 (53.7-76.2) |
| N-Ch REF 93.8 (90.0-96.2) | 0.002 | 3.57 (1.69-8.55) | 80.8 (62.6-99.7) | 0.004 | 7.08 (9.27-95.6) | 81.1 (65.6-96.5) |
| N-Ch REF 87.7 (84.4-90.3) | <0.001 | 2.94 (1.55-5.59) | 68.8 (48.0-81.2) | 0.001 | 2.30 (1.47-3.60) | 78.2 (58.3-74.1) |
| Ch REF 86.8 (82.6-90.0) | <0.001 | 3.03 (1.43-6.41) | 64.2 (37.6-81.8) | 0.004 | 2.30 (1.43-4.61) | 75.8 (49.9-88.4) |

HR = hazard ratio, 95% CI = 95% confidence interval, P = p value of log-rank test to compare survival distributions, REF = reference group, P = p value of Wald test for comparison versus reference (low risk) group, DMFS = distant metastasis free survival at 10 years (see text), (N) = number of cases in subgroups, All cases = all ER+/HER2- cases, N-Ch = Node negative cases treated without chemotherapy, N-Ch REF = reference group, Ch = cases treated with chemotherapy (node negative and node positive combined), Int = intermediate, P=p-value of log-rank test to compare survival distributions (global statistical significance of the model). P-value of Wald-test to evaluate whether the hazard ratio is 1 (statistical significance of each individual coefficient).
MammaPrint-Trained results stratified by Prosigna-trained results

In MammaPrint-trained low-risk cases 20% were Prosigna-trained high risk (DMFS10 = 78.1%; Table 4, p < 0.001) and 43% intermediate risk (DMFS10 = 86.1%; Table 4; p < 0.001, Fig. 4a). Amongst MammaPrint-trained high-risk cases, only a small (n = 12) subgroup of Mammaprint-trained high, Prosigna trained low results exhibited DMFS10 = 90% (p = 0.006, Fig. 4b).

Sub-group analysis ER+ve/HER2−ve, Node-ve patients not treated with chemotherapy

Oncotype-trained. Of 970 cases in this subgroup, 47.2% were Oncotype-trained low (DMFS10 = 92.5%), 36.0% intermediate (DMFS10 = 86.3%) and 16.8% high risk (DMFS10 = 76.7%, Table 2; Figs. 1d; 5) respectively.

Oncotype-trained results stratified by Prosigna-trained results. When Oncotype-trained results were stratified by Prosigna-trained results, 57.3% changed risk category (Supplementary Table 3). In Oncotype Dx-trained low risk 95 cases (21%) were Prosigna-trained high risk with DMFS10 = 83.8% (p = 0.006, Table 2; Fig. 5). In Oncotype-trained intermediate-risk cases 12% were Prosigna-trained low risk (DMFS10 = 94.1%; Table 2, p = 0.090; Fig. 5). The 57% of Oncotype-trained intermediate-risk cases classified as Prosigna-trained high risk exhibited DMFS10 = 83.7% (Table 2; p = 0.076, Fig. 5). Only three Oncotype-trained high-risk cases were Prosigna-trained low risk no events were observed in these cases.

Oncotype-trained stratified by MammaPrint-trained. 11% of Oncotype-trained low-risk cases were MammaPrint-trained high risk (DMFS10 = 80.8%, p = 0.004; Table 2, Fig. 5a). In Oncotype-trained intermediate-risk patients 50% were MammaPrint-trained low risk (DMFS10 = 92.2%, p = 0.002; Table 2, Fig. 5b). In Oncotype Dx-trained high-risk cases 11% were MammaPrint-trained low risk, no events were observed in these 18 cases (Table 2, Fig. 5c). MammaPrint-trained scores identified 37.5% of Oncotype-trained cases (intermediate or high) as low risk (DMFS10 > 90%).

Prosina-trained stratified by Oncotype-trained. Neither Prosina-trained low nor moderate risk cases showed statistically significant sub-stratification for outcome by Oncotype-trained risk scores (Table 3, Fig. 6a, b). Within Prosina-trained high-risk cases 22% were Oncotype-trained low risk, however, DMFS10 for this group was 83.8% (Table 3, Fig. 6c).

Prosina-trained stratified by MammaPrint-trained. No impact of MammaPrint-trained scores was observed in the Prosina-trained low-risk group (Table 3, Fig. 6a), with only three discordant results. For both moderate and high risk Prosina-trained results a group of MammaPrint-trained low-risk cases were identified (DMFS10 = 93.1% and 89.6%, respectively, Table 3; Fig. 6b, c).

MammaPrint-trained results

No impact of Oncotype-trained on Mammaprint-trained scores was observed (Fig. 7; Table 4). In Mammaprint-trained low-risk
| Oncotype-trained Low                  | Oncotype-trained Int                  | Oncotype-trained High                  |
|-------------------------------------|--------------------------------------|----------------------------------------|
| HR DMFS P*                          | HR DMFS P*                            | HR DMFS P*                             |
| All cases REF 92.2 (98.9–94.2)      | 0.737 (0.36–2.03)                     | 0.835 (6.35–9.25)                     |
| N- Ch REF 97.1 (92.5–98.9)          | 0.955 (0.12–2.96)                     | 0.976 (4.95–9.97)                      |
| Ch+ REF 91.9 (96.5–95.2)            | 0.552 (6.35–3.60)                     | 0.353 (7.6–9.93)                       |

**Table 3.** Prosigna-trained results stratified by other test results, trinary classification.

HR = Hazard ratio. 95% CI = 95% confidence interval. P* = p value of log-rank test to compare survival distributions. REF = reference group. P = p value of Wald test for comparison versus reference (low risk) group. DMFS = distant metastasis free survival at 10 years (see text). (N) = number of cases in subgroups. All cases = all ER+ve/HER2-ve cases. N-Ch = Node negative cases treated without chemotherapy. N+Ch = Node positive cases treated without chemotherapy. Ch+ = cases treated with chemotherapy (node negative and node positive combined). Int = intermediate.
cases 22% were categorized as Prosigna-trained high risk, with a modest reduction in DMFS10 = 89.6% (p = 0.027, Table 4).

DISCUSSION

Our analysis of 3284 ER+ve/HER2−ve cases using trained signatures demonstrates that the Prosigna-trained signature provides potentially more prognostic information than either the Oncotype-trained or MammaPrint-trained signatures (Table 1). This result is consistent with results in the smaller TransATAC cohort using original vendor methodology.

Critical to our study is the close correlation between the computationally derived “signature trained” scores and true results as shown by us previously. For ROR-PT results the correlation coefficient between “trained” and true assay results was 0.93, comparing true to “trained” results showed 90% of cases within the same risk category (low, intermediate, high—see ref. 16). Similarly for “Oncotype-Dx trained” results the correlation coefficient between true and “trained” results was 0.87 with 75% of results giving the same risk category (see ref. 16) and only 1% of cases disagreeing by more than 1 risk category. For Mammaprint trained results, which were calculated only as categorical high versus low risk groups, over 90% of cases were classified in the same risk group by “trained” and true results. Full details of these results are reported elsewhere.

We also show when two trained tests are combined the overall amount of information is always greater than a single test alone. In this study, adding stratification by Prosigna-trained results to Oncotype-trained results provided the greatest $LR_+=2$, and the improvement was greater for this combined model versus Oncotype-trained results alone than for Prosigna-trained results alone. Collectively these results suggest that, in this study, Prosigna-trained results, either alone or combined with other test results, provide potentially greater prognostic information. However, most critically, all test combinations (where two tests were used for patient stratification) outperformed models with only one test to a highly statistically significant degree. This both confirms earlier reports and suggests that differences between tests reflect quantitative and qualitative differences in the degree of prognostic information collected. This conclusion is supported by recent comparisons by the ATAC group, showing the impact of different signaling modules in ER+ve/HER2−ve cases across different signatures. The conclusion from this work is that different tests capture different aspects of prognostic drivers and therefore that future improvements in prognostic testing remain achievable.

Critically, we dissected the effect of applying a second test to risk-stratified subgroups defined by the initial result; e.g. we examined the effect of applying the Prosigna-trained signature to the “intermediate risk” group identified by the Oncotype-trained signature etc. When combining tests, Prosigna-trained results added value to both Oncotype-trained and MammaPrint-trained results (Table 1). The improved prognostic impact of Prosigna-trained results applied across all ER+ve/HER2−ve cases when Oncotype-trained results was reflected by Prosigna-trained results sub-stratifying patients across both low and intermediate risk Oncotype trained groups (Fig. 2a, b). Even within the node negative ER+ve/HER2−ve population not treated with chemotherapy (Table 2; Fig. 5a, b) Oncotype-trained low and intermediate-risk groups were also further stratified by Prosigna-trained results and 20.7% of Oncotype-trained low-risk cases were
### Table 4. Mammaprint-trained results stratified by other test results, trinary classification.

| Mammaprint-trained first | Mammaprint-trained high risk |
|--------------------------|-----------------------------|
| **Mammaprint-trained low risk** | **DMFS (95% CI)** | **P**<sup>*</sup> | **Mammaprint-trained high risk** | **DMFS (95% CI)** | **P**<sup>*</sup> |
| All cases | REF | 86.9 (85.1–88.4) | <0.001 | (2180) | 2.64 (2.22–3.14) | <0.001 | (1104) |
| N-Ch− | REF | 93.5 (90.5–95.6) | <0.001 | (896) | 4.23 (2.72–6.56) | <0.001 | (371) |
| N-Ch+ | REF | 85.9 (83.1–88.3) | <0.001 | (452) | 3.94 (2.56–4.36) | <0.001 | (531) |
| Ch+ | REF | 82.3 (78.8–85.3) | <0.001 | (681) | 1.86 (1.41–2.46) | <0.001 | (379) |

| Onco-type-trained low | HR (95% CI) | DMFS (95% CI) | P<sup>*</sup> | Onco-type-trained high | HR (95% CI) | DMFS (95% CI) | P<sup>*</sup> |
|-----------------------|-------------|---------------|----------|------------------------|-------------|---------------|----------|
| All cases | REF | 89.1 (87.1–90.8) | <0.001 | (1483) | 72.1 (60.9–80.6) | 0.031 | 1.10 (528) | 0.668 | 1.47 (62.2–71.6) | 0.072 |
| N-Ch− | REF | 93.8 (90.0–96.2) | <0.001 | (407) | 80.8 (62.8–90.7) | 0.196 | 1.12 (175) | 0.784 | 1.68 (65.0–81.0) | 0.212 |
| N-Ch+ | REF | 87.7 (84.4–90.3) | <0.001 | (639) | 67.6 (48.0–81.2) | 0.047 | 1.22 (166) | 0.552 | 1.83 (46.0–63.6) | 0.067 |
| Ch+ | REF | 86.8 (82.6–90.0) | <0.001 | (434) | 64.2 (57.6–81.8) | 0.007 | 1.05 (185) | 0.676 | 0.92 (63.8–78.4) | 0.792 |

| Prosigna-trained low | HR (95% CI) | DMFS (95% CI) | P<sup>*</sup> | Prosigna-trained high | HR (95% CI) | DMFS (95% CI) | P<sup>*</sup> |
|----------------------|-------------|---------------|----------|------------------------|-------------|---------------|----------|
| All cases | REF | 92.2 (89.8–94.0) | <0.001 | (814) | 90.0 (87.3–98.5) | 0.006 | 2.17 (211) | 0.466 | 3.57 (64.9–71.7) | 0.024 |
| N-Ch− | REF | 96.6 (91.8–98.6) | 0.069 | (217) | 89.0 (85.7–93.1) | 0.027 | 0.78 (72) | 83.8 | 0.002 | 0.76 (70.6–81.4) | 0.296 |
| N-Ch+ | REF | 90.0 (85.7–93.1) | <0.001 | (315) | 90.0 (85.7–93.1) | 0.002 | 0.78 (72) | 83.8 | 0.002 | 0.76 (70.6–81.4) | 0.296 |
| Ch+ | REF | 91.1 (86.5–94.2) | <0.001 | (281) | 91.1 (86.5–94.2) | 0.002 | 0.78 (72) | 83.8 | 0.002 | 0.76 (70.6–81.4) | 0.296 |

**HR** = hazard ratio, 95%CI = 95% confidence interval. **P** = p value of log-rank test to compare survival distributions. **REF** = reference group. **P** = p value of Wald test for comparison versus reference (low risk) group. DMFS = distant metastasis free survival at 10 years (see text). (N) = number of cases in subgroups. All cases = all ER+ve/HER2–ve cases. **N-Ch−** = node negative cases treated without chemotherapy. **N-Ch+** = node negative cases treated with chemotherapy. **Ch+** = cases treated with chemotherapy (node negative and node positive combined). **Int** = intermediate.
identified as high risk by Prosigna-trained results, with DMFS10 of 83.8%, which is important as results from prospective trials suggest these cases may benefit from chemotherapy. This difference was more striking when Oncotype-trained results were dichotomized using cut-points applied in the Tailor-X trial. In ER+/HER2− ve, node negative patients treated without chemotherapy 17−24% of cases with Oncotype-trained results ≥25 were low risk (DMFS10 > 90%) when stratified by Mammaprint-trained or Prosigna-trained results respectively (Supplementary Table 4; Supplementary Fig. 2). Conversely 18−30% of Oncotype-trained low risk cases (<25) were high risk when stratified by Mammaprint-trained or Prosigna-trained results and exhibited DMFS < 90% (Supplementary Table 4; Supplementary Fig. 2).

Conversely, only in Prosigna-trained intermediate risk cases did Oncotype-trained results provide additional stratification by risk (Fig. 3; Table 3). However this stratification was not observed in the subgroup of node negative cases treated without chemotherapy (Fig. 6). No stratification of Prosigna-trained low or high risk cases was observed using either Oncotype-trained or Mammaprint-trained results (Fig. 3; Supplementary Table 4). When using dichotomized risk scores for Prosigna-trained ER+/ve/HER2−ve node-negative cases treated without chemotherapy no further stratification using dichotomized Oncotype-trained results was seen (Supplementary Table 5; Supplementary Fig. 5) and all Prosigna-high risk cases exhibited DMFS10 < 85% regardless of dichotomized Oncotype-trained results (Supplementary Table 5; Supplementary Fig. 5). These results are illustrative of and highlight the potential clinical impact of disagreements between tests at an individual patient level previously demonstrated in the OPTIMA-prelim cohort.

A number of conclusions that can be drawn from our analyses. Firstly that, as with previous analyses22 there is additional prognostic value to be gained from combining multiple molecular tests in the research setting. The corollary is that no single existing assay captures the sum of prognostic information available at the transcriptomic level. This confirms earlier findings22 that improvements in prognostic assays remain possible. Such improvements may, however, require integration of additional molecular features beyond transcriptomics.23,28 Secondly, there was evidence, albeit from sub-group analyses, that the known interaction between clinical risk, treatment, and molecular risk profiling may differ depending on the test chosen. If taken at face value, this might provide support for the use of different testing strategies in different patient risk strata.

Our analysis has some potentially important limitations. In particular we have used a computational approach to generate test scores for the different tests described herein. At an individual tumor level, the trained score may not be identical to the equivalent generated using original methodology. We trained our signatures in an independent cohort using the same signatures measured using original methodology, achieving extremely high correlations with commercial test results. Additionally, the broad agreement between our analysis with the (more limited) analysis of Sestak et al.20 using original methodology and a slightly different statistical approach is highly reassuring.

Additionally, although our cohort is exclusively postmenopausal ER-positive, 30% of cases were treated with adjuvant chemotherapy. All patients in the TEAM trial were postmenopausal, with a median age of 64 years, results presented here may not be representative of the premenopausal population. We included chemotherapy-treated patients to maximize the power of our main analysis. However, the conclusions of our analysis performed on the node-negative subgroup who were not chemotherapy-treated are broadly similar to those in the analysis of the entire cohort, suggesting that these findings are robust both in this clinically critical node negative sub-group and indeed across all patients in the TEAM cohort.

The goal of our study was to provide robust information on the impact of discordant risk classification by different molecular prognostic signatures in postmenopausal, ER+ve early breast cancer. Existing evidence highlights discordance between tests, which is reiterated here. There is clear evidence that adding clinical information to test results provides additional prognostic information, which is supported by sub-group analyses performed here, and that information provided by any individual assay is relatively modest. To date comparisons between tests have been limited either by relatively small sample sizes or by a lack of evidence that signatures extracted from global expression data reflect actual test performance and can therefore inform patients and clinicians on the impact of discordant test results on outcome in the real-world setting. This study provides data on a large clinical trial cohort (the TEAM trial) using test signatures

### Table 5: DMFS (95%CI) and P-value

| Test         | DMFS (95%CI) | P-value | N   |
|--------------|--------------|---------|-----|
| Mammaprint   |              |         |     |
| Low Risk     | 88.9 (85.1−88.4) | <0.001  | 2180|
| Oncotype     | 89.1 (87.1−90.8) | <0.001  | 1483(68%)|
| Moderate     | 83.2 (79.6−86.1) |        | 645(30%)|
| High         | 70.4 (53.7−82.1) |        | 502(2%)|
| Low          | 92.2 (89.8−94.0) | <0.001  | 814(37%)|
| Moderate     | 86.1 (83.3−88.5) |        | 944(43%)|
| High         | 78.1 (73.0−82.3) |        | 422(20%)|

**Fig. 4 Forest plot of Mammaprint-trained test results re-stratified by other tests, all ER+ve/HER2-ve cases. DMFS10 = distant metastasis free survival at 10 years post diagnosis. (95% CI) = 95% confidence interval, P = p value, N = number of cases in each subgroup, % = percentage of cases within each risk strata, X axis = percent distant metastasis free survival. Open boxes represent primary test DMFS10 by risk group. Solid boxes represent sub-stratification by secondary tests with 95% confidence intervals (bars). Top panel (a) Mammaprint-trained low-risk cases stratified by Oncotype-trained and Prosigna-trained results. Bottom panel (b) Mammaprint-trained high-risk group.**
trained in a second cohort (OPTIMA-prelim) to match actual commercial test performance.

In summary, our study provides novel evidence for the potential clinical impact of discordant molecular test results in a large population. Further improvements in test performance are potentially within reach and would be of benefit to patients.

Evidence presented here suggests the differences in test performance are more nuanced than previously reported and that careful consideration to test selection, in the context of treatment and clinical risk may be appropriate.

METHODS

Study design

Our primary analyses explored the impact of signature-trained prognostic scores, categorized in accordance with published cut-points for each assay, for patients with centrally confirmed estrogen receptor positive (ER+ve) HER2 negative (HER2−ve) disease and HER2 positive (HER2+ve) cases were excluded since during recruitment of the TEAM trial HER2 targeted therapies were not used in this setting. We performed a secondary analysis using dichotomized scores for Oncotype Dx and Prosigna to reflect the results of the TailorX study. We also report a complete cohort analysis, including HER2+ve cases (see Supplementary Information), since no assay used was trained on samples treated with HER2-targeted therapies. Supplementary analyses further sub-divide patient groups into node negative cases treated with endocrine therapy (but not chemotherapy), node positive cases treated with endocrine therapy (but not chemotherapy) and cases treated with chemotherapy and endocrine therapy (both node negative and node positive, supplementary methods, data and figures).

Patient samples

Patient samples were derived from the Tamoxifen Exemestane Adjuvant Multicenter (TEAM) Trial pathology study (Supplementary Table 1; NCT00279448/NCT0032126/NCT0036270, NTR267, UMIN C000000057) and included only hormone receptor positive, post-menopausal cancers. Patients provided informed consent and this study was approved by the University of Toronto REB (protocol number 29021).

RNA profiling using NanoString

Profiling of all samples was performed using mRNA previously extracted and analyzed using a custom NanoString codeset as described previously and included only hormone receptor positive, post-menopausal cancers. Patients provided informed consent and this study was approved by the University of Toronto REB (protocol number 29021).

Signature-trained Risk Stratification Scores from candidate assays

We compared two different approaches to the generation of simulated risk scores and selected a training and validation approach using results obtained from the OPTIMA prelim study to fit risk stratification scores generated for this study to those derived from the relevant commercial assay. For all tests, we used the suffix-trained to discriminate the
Fig. 6 Forest plot of Prosigna-trained test results re-stratiﬁed by other tests, Node-ve ER+ve/HER2-ve cases treated without chemotherapy. DMFS10 = distant metastasis free survival at 10 years post diagnosis, (95% CI) = 95% conﬁdence interval, P = p value, N = number of cases in each subgroup, % = percentage of cases within each risk strata, X axis = percent distant metastasis free survival. Open boxes represent primary test DMFS10 by risk group. Solid boxes represent sub-stratiﬁcation by secondary tests with 95% conﬁdence intervals (bars). Top panel (a) Prosigna-trained low-risk cases stratiﬁed by Oncotype-trained and Mammaprint-trained results. Middle panel (b) Prosigna-trained moderate risk group. Bottom panel (c) Prosigna-trained high risk group.

Fig. 7 Forest plot of Mammaprint-trained test results re-stratiﬁed by other tests, Node-ve ER+ve/HER2-ve cases treated without chemotherapy. DMFS10 = distant metastasis free survival at 10 years post diagnosis, (95% CI) = 95% conﬁdence interval, P = p value, N = number of cases in each subgroup, % = percentage of cases within each risk strata, X axis = percent distant metastasis free survival. Open boxes represent primary test DMFS10 by risk group. Solid boxes represent sub-stratiﬁcation by secondary tests with 95% conﬁdence intervals (bars). Top panel (a) Mammaprint-trained low-risk cases stratiﬁed by Oncotype-trained and Prosigna-trained results. Bottom panel (b) Mammaprint-trained high-risk group.
computationally derived assays scores from the commercially derived scores, e.g. Oncotype-trained vs. Oncotype-DX™.

**Methods for cross comparisons between Tests**

Results were available for 3811 subjects. Cases were grouped into the pre-defined risk categories for each test as follows: Oncotype DX—low risk < 18, intermediate risk 18–31 (supplementary methods), high risk ≥ 31; Prosigna ROR-T—low risk < 41, intermediate risk 41–60, high risk ≥ 61 (31); MammaPrint—low risk and high risk (32). We also performed a dichotomized risk analysis for Oncotype Dx using low/intermediate risk 0–25 and high risk > 25, in line with the TailorX study (33), and for Prosigna RT using low/intermediate risk < 61 and high risk ≥ 61. Grouped analyses were performed as follows: (1) ER+/HER2−; (2) and (2) hormone-receptor positive (HR+) regardless of HER2 status (n = 3811). Subjects were considered HR+ if ER and/or progesterone receptor (PR) was reported as positive (34). Differences in distant metastasis free survival (DMFS; i.e. time to first distant recurrence or death, excluding ipsilateral breast cancer recurrences but including distant metastasis, contralateral breast cancer and death from breast cancer) were evaluated using the Kaplan–Meier method with test equality of survivor functions assessed by log-rank and fidence intervals (95% CI) were calculated as DMFS (35). Hazard ratios (HRs) were calculated using Cox proportional hazards regression models, and endpoints into a phase III, randomized clinical trial (SWOG 51007) evaluating OncotypeDX-guided management for women with breast cancer involving lymph nodes. Contemp. Clin. Trials 34, 1–9 (2013).

5. Pan, H. et al. 20-Year Risks of breast-cancer recurrence after stopping endocrine therapy at 5 years. N. Engl. J. Med. 377, 1836–1846 (2017).
6. Pyo, A. E., Ellis, M. J. & Perou, C. M. Practical implications of gene expression-based assays for breast oncologists. Nat. Rev. Clin. Oncol. 9, 48–57 (2012).
7. Fan, C. et al. Concordance among gene-expression-based predictors for breast cancer. N. Engl. J. Med. 355, 560–569 (2006).
8. Kelly, C. M. et al. Agreement in risk prediction between the 21-gene recurrence score assay (Oncotype DX(R)) and the PAM50 breast cancer intrinsic Classifier in early-stage estrogen receptor-positive breast cancer. Oncologist 17, 492–498 (2012).
9. Mackay, A. et al. Microarray-based class discovery for molecular classification of breast cancer: analysis of interobserver agreement. J. Natl Cancer Inst. 103, 662–673 (2011).
10. Weigelt, B. et al. Breast cancer molecular profiling with single sample predictors: a retrospective analysis. Lancet Oncol. 11, 339–349 (2010).
11. Dowsett, M. et al. Comparison of PAM50 risk of recurrence score with oncopotype DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. J. Clin. Oncol. 31, 2783–2790 (2013).
12. Sgroi, D. C. et al. Prediction of late distant recurrence in patients with oestrogen-receptor-positive breast cancer: a prospective comparison of the breast-cancer index (BCI) assay, 21-gene recurrence score, and IHC4 in the TransATAC study population. Lancet Oncol. 14, 1067–1076 (2013).
13. Bayani, J. et al. Molecular stratification of early breast cancer identifies drug targets to drive stratified medicine. npj Breast Cancer 3, 3 (2017).
14. Bartlett, J. M. S. et al. Computational approaches to support comparative analysis of multiparametric tests: modelling versus Training. PLoS ONE 15, e0238593–e0238593 (2020).
15. van de Velde, C. J. H. et al. Adjuvant tamoxifen and exemestane in early breast cancer (TEAM): a randomised phase 3 trial. Lancet 377, 321–331 (2011).
16. Sestak, I. et al. Comparison of the performance of 6 prognostic signatures for estrogen receptor–positive breast cancer: a secondary analysis of a randomized clinical trial prognostic signatures for estrogen receptor–positive breast cancer–prognostic signatures for estrogen receptor–positive breast cancer. JAMA Oncol. 4, 545–553 (2018).
17. Buus, R. et al. Molecular drivers of oncopotype DX, Prosigna, EndoPredict, and the Breast Cancer Index: a TransATAC study. J. Clin. Oncol. 20, 00853 (2020).
18. Bayani, J. et al. Molecular stratification of early breast cancer identifies drug targets to drive stratified medicine. npj Breast Cancer 3, 3 (2017).
19. Bayani, J. et al. Identification of distinct prognostic groups: implications for patient selection to targeted therapies among anti-estrogen therapy-resistant early breast cancers. JCO Precis. Oncol. 3, 1–13 (2019).
20. Pereira, B. et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nat. Commun. 7, 11479 (2016).
21. Vallen-Christersson, J. et al. Cross comparison and prognostic assessment of breast cancer multigene signatures in a large population-based contemporary clinical series. Sci. Rep. 9, 12184 (2019).
22. Cuzick, J. et al. Prognostic value of a combined ER, PgR, Ki67, HER2 immunohistochemical (IHC) score and comparison with the GH recurrence score - results from TransATAC. Cancer Res. 69, 5035–5039 (2009).
23. Dowsett, M. et al. Prediction of risk of distant recurrence using the 21-gene recurrence score in node-negative and node-positive postmenopausal patients with breast cancer treated with anastrozole or tamoxifen: a TransATAC study. J. Clin. Oncol. 28, 1829–1834 (2010).
24. Cuzick, J. et al. Prognostic value of a combined estrogen receptor, progesterone receptor, Ki-67, and human epidermal growth factor receptor 2 immunohistochemical score and comparison with the genomic health recurrence score in early breast cancer. J. Clin. Oncol. 29, 4273–4278 (2011).
25. Sestak, I. et al. Factors predicting late recurrence for estrogen receptor-positive breast cancer. J. Natl Cancer Inst. 105, 1504–1511 (2013).
26. Bartlett, J. M., Rea, D. & Rimm, D. L. Quantification of hormone receptors to guide adjuvant therapy choice in early breast cancer: better methods required for improved utility. J. Clin. Oncol. 29, 3715–3716 (2011).
27. Bartlett, J. M. et al. Mammostrat as an immunohistochemical multigene assay for prediction of early relapse risk in the tamoxifen versus exemestane adjuvant multicenter trial pathology study. J. Clin. Oncol. 30, 4477–4484 (2012).
ACKNOWLEDGEMENTS
Research at the Ontario Institute for Cancer Research is supported by the Government of Ontario. P.C.B. was supported by Genome Canada, by CIHR New Investigator Award and by a Terry Fox Research Institute New Investigator Award. R.C. S. was supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre. The funder had no role in the analysis or reporting of results.

AUTHOR CONTRIBUTIONS
J.M.S.B., J.B., R.C.S., D.R. contributed to the conception and design of the work, the acquisition, analysis, and interpretation of data, and were in involved in drafting, critical review and approval of the final submitted version. They have agreed to be personally accountable for their contributions. E.K., K.X., G.R.P., C.Q.Y., and P.C.B. contributed the acquisition, analysis, and interpretation of data, and were in involved in critical review and approval of the final submitted version. They have agreed to be personally accountable for their contributions. T.P., E.M., A.H., J.A.D., C.M., L.D., C.S. and C.J.H.v.d.V. contributed the acquisition and interpretation of data, and were in involved in critical review and approval of the final submitted version. They have agreed to be personally accountable for their contributions. All authors contributed to the conception or design of the work or to the acquisition, analysis or interpretation of data. All authors were involved in the drafting and/or critical review.

COMPETING INTERESTS
J.M.S.B. has received consultancy or honoraria from Insight Genetics Inc., BioNTechAG, Biothernostics Inc., RNA Diagnostics Inc., oncoXchange, NanoString Technologies Inc, and research funding from ThermoFisher Scientific, Genoptics, Agenda, NanoString Technologies Inc., Biotheranostics Inc. J.B. has received honoraria from ThermoFisher Scientific. G.P. has received consulting fees from Merck, Astra-Zenea, Profound Medical, outside of submitted work; Honorarium for DSMB membership from Takeda outside of submitted work. A.H. has received honoraria from MedConcept GmbH, Med Update GmbH, Pfizer, Roche Pharma AG, Streamed up GmbH, Tesaro Bio Germany GmbH and serves on Advisory Boards for PharmaMar, Roche Pharma AG, Tesaro Bio Germany GmbH. C.M. has received consultancy from Genomic Heath. All other authors declare no competing interests.

ADDITIONAL INFORMATION
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41523-021-00297-7.

Correspondence and requests for materials should be addressed to J.M.S.B.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021