

Original citation:

Bartlett, John M. S., Bayani, Jane, Marshall, Andrea, Dunn, Janet A., Campbell, Amy, Cunningham, Carrie, Sobol, Monika S., Hall, Peter S., Poole, Christopher J., Cameron, David A. et al.. (2016) Comparing breast cancer multiparameter tests in the OPTIMA Prelim Trial : no test is more equal than the others. Journal of the National Cancer Institute, 108 (9). djw050.

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/79314>

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions. Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:

This is a pre-copyedited, author-produced PDF of an article accepted for publication in. Journal of the National Cancer Institute following peer review. The version of record Comparing breast cancer multiparameter tests in the OPTIMA Prelim Trial : no test is more equal than the others. Journal of the National Cancer Institute, 108 (9). djw050. is available online at: <http://dx.doi.org/10.1093/jnci/djw050>

A note on versions:

The version presented here may differ from the published version or, version of record, if you wish to cite this item you are advised to consult the publisher's version. Please see the 'permanent WRAP URL' above for details on accessing the published version and note that access may require a subscription.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

JNCI 15-1086R1

Article

Comparing breast cancer multi-parameter tests in the UK OPTIMA prelim trial: All tests are equal – none are more equal than others

John M.S. Bartlett^{1,2,3*}, Jane Bayani^{1*}, Andrea Marshall⁴, Janet A. Dunn⁴, Amy Campbell⁴; Carrie Cunningham³, Monika S. Sobol³, Peter S. Hall³, Christopher J. Poole⁵, David A. Cameron³, Helena M. Earl⁶, Daniel W. Rea⁷, Iain R. Macpherson⁸, Peter Canney⁸, Adele Francis⁹, Christopher McCabe¹⁰, Sarah E. Pinder¹¹, Luke Hughes-Davies¹², Andreas Makris¹³, Robert C. Stein¹⁴, on behalf of the OPTIMA TMG.

1. Ontario Institute for Cancer Research, Toronto, Ontario, Canada
2. University of Toronto, Toronto, Canada
3. University of Edinburgh, Edinburgh, United Kingdom
4. Warwick Clinical Trials Unit, University of Warwick, Coventry, United Kingdom
5. University Hospitals Coventry and Warwickshire NHS Trust, Coventry, United Kingdom
6. University of Cambridge Department of Oncology and NIHR Cambridge Biomedical Research Centre, Cambridge, United Kingdom.
7. Cancer Research UK Institute for Cancer Studies, University of Birmingham, Birmingham, United Kingdom.
8. University of Glasgow, Beatson West of Scotland Cancer Centre, Glasgow, United Kingdom.
9. University Hospitals Birmingham NHS Foundation Trust, Birmingham, United Kingdom.

10. University of Alberta, Edmonton, AB, Canada.
11. Kings College London, Guy's Hospital, London, United Kingdom.
12. Addenbrooke's Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdom.
13. Mount Vernon Cancer Centre, East and North Hertfordshire NHS Trust, Middlesex, United Kingdom.
14. National Institute for Health Research University College London Hospitals Biomedical Research Centre, London, United Kingdom.

*These authors contributed equally to this work.

Corresponding author:

John M.S. Bartlett, Director of Transformative Pathology, Ontario Institute for Cancer Research, MaRS Centre, 661 University Avenue, Suite 510, Toronto, Ontario, Canada M5G 0A3; Telephone: 647-259-4251; Fax: 416-977-7446; email: John.Bartlett@oicr.on.ca.

ABSTRACT

Background: Previous reports identifying discordance between multi-parameter tests at the individual patient level have been largely attributed to methodological shortcomings of multiple in silico studies. Comparisons between tests, when performed using actual diagnostic assays, have been predicted to demonstrate high degrees of concordance. OPTIMA prelim compared predicted risk stratification and subtype classification of different multi-parameter tests performed directly on the same population.

Methods: 313 women with early breast cancer were randomised to standard (chemotherapy and endocrine therapy) or test-directed (chemotherapy if Oncotype DX™ recurrence score >25) treatment. Risk stratification was also determined with, Prosigna™(PAM50), MammaPrint™, MammaTyper™, NexCourse Breast™ (IHC4-AQUA™) and conventional IHC4 (IHC4). Subtype classification was provided by Blueprint™, MammaTyper™ and Prosigna™.

Results: Oncotype DX™ predicted a higher proportion of tumours as low risk (82.1%; 95%CI 77.8-86.4%) than were predicted low/intermediate risk using Prosigna™ (65.5%; 95%CI 60.1-70.9%), IHC4 (72.0% 95%CI 66.5-77.5), MammaPrint™ (61.4%; 95%CI 55.9-66.9%) or NexCourse Breast™ (61.6%; 95%CI 55.8-67.4%). Strikingly, the five tests showed only modest agreement when dichotomising results between high versus low/intermediate risk. Only 119 (39.4%) tumours were classified uniformly as either low/intermediate risk or high risk, 183 (60.6%) were assigned to different risk categories by different tests, although 31.1% (94) showed agreement between four/five tests. All three subtype tests assigned 59.5-62.4% of tumours to luminal A subtype, but only 121 (40.1%) were classified as luminal A by all three tests and only 58 (19.2%) were uniformly assigned as non-luminal A. Discordant subtyping was observed in 123 (40.7%) cases.

Conclusions: Existing evidence on the comparative prognostic information provided by different tests suggests current multi-parameter tests provide broadly equivalent risk information for the population

of women with ER-positive breast cancers. However, for the individual patient, tests may provide differing risk categorisation and subtype information.

INTRODUCTION

For over 40 years (1-3) the impact of tumour molecular markers on patient outcome and treatment response has been central to breast cancer management. Gene-expression profiling (4;5) to describe the intrinsic subtypes of breast cancer was followed by the independent development, in 2004, of the first multi-parameter molecular diagnostic assay stratifying breast cancer patients with estrogen receptor (ER) positive disease based on risk of relapse following treatment (6). The past decade saw a rapid expansion in the number of such multi-parameter molecular residual risk tests for breast cancer patients (see (7)). These herald an era of more personalised medicine due to their potential to inform rational treatment decisions on a patient-by-patient basis. The initial goal was to identify patients who, despite “favourable” clinico-pathological characteristics, have a poor outcome following conventional endocrine treatment and to advise aggressive therapy, which may reduce relapse risk.

Over time, interest has also grown in the potential for multi-parameter assays to predict chemo-sensitivity (8;9). These tests may also allow an estimate of the intrinsic chemotherapy sensitivity of tumours, reducing the importance of stage information. There are women who gain little from chemotherapy and women who have clinically significant gains. There is therefore a rationale for using stratified medicine to identify patients who may safely avoid toxicities associated with chemotherapy.

The OPTIMA trial (7) is designed as a prospective test of the effectiveness of multi-parameter testing in identifying the subgroup of women with breast cancer (amongst those who would be routinely offered adjuvant chemotherapy, based on conventional criteria), whose tumours are intrinsically insensitive to chemotherapy and for whom such treatment offers only toxicity and delay in starting more effective adjuvant endocrine therapy, and radiotherapy, without any clinically meaningful additional benefit. A key objective of “OPTIMA prelim”, the in-built feasibility phase

of OPTIMA, was to evaluate the performance of alternative multi-parameter tests, to aid selection of a test for the main study that would ensure the results of such a trial be robust and broadly applicable to the patient population, both now and in the future. Critical to this decision was the ability to compare test performance at both the population and individual patient level. Existing data directly comparing individual test performance is limited. A series of studies performing statistical comparisons between tests suggest that, at a population level, four tests (IHC4, PAM50, BCI and Oncotype DX) TM provided broadly equivalent prognostic information on the risk of relapse up to five years post treatment (10-12). Further studies, based largely on in silico reconstruction of existing tests from publically available gene expression data sets suggest a statistically significant degree of discordance between signatures at the individual patient level (13-17). These observations are predominantly attributed to methodological differences due the in silico reconstruction of signatures (15;17). This thesis has not, to date, been robustly tested using actual test methodologies. Limited data shows that concordance between different tests in assigning patients to similar risk groups is low (10). This is consistent with the marked differences in genes measured by different tests (See Supplementary Table 1) and with the relatively modest predictive value, in terms of recurrence, offered by these tests at the individual patient level. Here we report the direct patient level comparison of multiple commercial residual risk profiles in the OPTIMA prelim study, performed to gather information on their performance.

MATERIALS AND METHODS

Recruitment and patient samples

OPTIMA prelim (Optimal Personalised Treatment of early breast cancer using Multi-parameter Analysis preliminary study, ISRCTN42400492) (18) is a multi-centre study that randomised women aged ≥ 40 with ER-positive HER2-negative early breast cancer and either 1-9 involved axillary

nodes or tumour size ≥ 30 mm (if node-negative) between standard treatment (chemotherapy followed by endocrine therapy) and test-directed therapy (7). In the test-directed arm an Oncotype DX™ test was performed; patients with Recurrence Score (RS) >25 (“high” risk) were assigned chemotherapy followed by endocrine therapy, those with RS ≤ 25 (“intermediate/low” risk) received endocrine therapy alone. Chemotherapy, selected from regimens commonly used in the UK NHS, was specified at patient registration. The study was partially blinded so that neither patients nor referring centres were aware of whether chemotherapy was assigned on the basis of Oncotype DX™ RS or by randomisation to the standard treatment arm. Central re-testing of ER and HER2 status was performed on all patients. Following confirmation of eligibility, samples were sent to Genomic Health for Oncotype DX™ assays to be performed with funding from the OPTIMA prelim study. No patient outcome data is available for this analysis. All patients gave written informed consent to participate in the study. The study was approved by the South East Coast - Surrey Research Ethics Committee.

To facilitate the comparison of alternative tests a number of test vendors were approached for support (**Supplementary Table 2**). Ultimately five tests in addition to Oncotype DX™ were included in the OPTIMA prelim study: MammaPrint/BluePrint™, Prosigna™ (PAM50), MammaTyper™, NexCourse Breast by Aqua™ (IHC4-AQUA) and IHC4 by conventional immunohistochemistry. Multi-parameter assays were performed irrespective of patient randomisation. Vendors that did not participate expressed concerns about transposing specific tests into novel applications.

Residual tumour samples from patients were collected at a central good clinical laboratory practice pathology repository (Edinburgh UK). Tissue MicroArrays (TMAs) were constructed as previously described (19) using triplicate 0.6mm cores. TMA sections, tissue sections or extracted mRNA were provided either to the Ontario Institute for Cancer Research (Prosigna™; IHC4: ER, PgR and Ki67 by quantitative image analysis (Ariol) using standard IHC with HER2 testing by ISH at

UCL Advanced Diagnostics) or to Genoptix (IHC4-AQUA), Agendia (MammaPrint™/Blueprint™) and Stratifyer (MammaTyper™). Results from individual tests were collated at the Warwick Clinical Trials Unit (CTU) for analysis.

Statistical Analysis

OPTIMA prelim was designed to recruit 300 patients to enable the kappa value for agreement between tests to be estimated with good accuracy. Assuming 70% of patients would be assigned to no chemotherapy by the test and the true kappa value was 0.8 (14), this would provide a lower 95% confidence limit of 0.73. These numbers were also sufficient to allow for the assumed proportion of patients assigned to no chemotherapy to vary from 55% to 80% (lower confidence limit for kappa varied from 0.74 to 0.72 respectively).

The proportion of tumours assigned to risk groups and/or subtypes was determined. The kappa coefficient and associated 95% confidence interval (95%CI) was used to assess agreement between tests. The predicted benefits of endocrine therapy with or without chemotherapy individualised to patients were estimated using two nomograms, Adjuvant! (20) (version 8, without correction for HER2 status) and PREDICT (21-23). A multivariable logistic regression model using stepwise elimination was performed to determine factors predicting discordant cases. To explore the post hoc hypothesis that individual tests were more likely to agree at the extremes of their ranges, two-by-two scatterplots for the tests that provide risks scores and agreement charts for the categorisation of tumours were constructed (24). Statistical analyses were performed using the SAS statistical package (version 9.3; SAS Institute Inc., Cary, NC, USA) and R version 3.0.3 (25). All statistical tests were two-sided and a p-value of less than 0.05 was considered statistical significant.

RESULTS

Patients

Between October 2012 and June 2014, 313 patients were randomised from 35 UK hospitals (see the **Notes**), of whom 302 had samples available for multi-parameter testing (**Table 1**). Eleven patients were excluded from multi-parameter testing; four withdrew consent, one was ineligible and samples for six patients were insufficient for testing (**Supplementary Figure 1**).

Results from predictive nomograms

The majority of patients recruited were either at intermediate (74.8%) or high (21.2%) risk using the Nottingham Prognostic Index (NPI) (26). All 12 patients with lower risk NPI scores (≤ 3.4) had tumours ≥ 3.0 cm. The median 10 year overall survival estimated by PREDICT (21-23) or Adjuvant! (20) differed by 6.2-8.4% reflecting expected differences between the risk estimate provided by these tools (**Table 2**).

Multi-parameter tests

Results from all tests were available for 236 (78.1%) patients. One patient on the standard arm had insufficient invasive tumour for Oncotype DX™ testing, but sufficient for alternative testing. Test results were unobtainable from Prosigna™ for three patients; from MammaTyper™ for four patients; from MammaPrint™ for four patients and Blueprint™ for seven patients. IHC4 and IHC4-AQUA™ could not be determined for 45 (14.9%) and 31 (10.3%) patients respectively, reflecting use of TMAs for this assessment.

Risk Scores

Five tests provided quantitative or semi-quantitative risk scores and a pre-defined categorised risk assessment (low, intermediate, high). For OPTIMA prelim, Oncotype DX™ RS was dichotomised around 25 separating “low/intermediate” from “high” risk cases as only patients with a high risk of recurrence were allocated chemotherapy (**Table 3**). Using this approach for all tests (**Supplementary Methods**), the proportion of cases classified as low/intermediate risk for Oncotype DX™ was 82.1% (95%CI 77.8-86.4%), 72.0% (95%CI 66.5-77.5%) for IHC4, 65.6% (95%CI 60.1-70.9%) using Prosigna™ risk of recurrence score including proliferation and tumour size, 61.6% (95%CI 55.8-67.4%) for IHC4-AQUA, and 61.4% (95%CI 55.9-66.9%) for MammaPrint™ (**Table 3**).

Agreement between tests when patients were subdivided into combined low/intermediate versus high-risk groups using predefined cut-points was modest; Kappas ranged from 0.33 (95%CI 0.21-0.44) between MammaPrint™ and IHC4 to 0.60 (95%CI 0.50-0.70) between IHC4 and IHC4-AQUA (**Table 4, Supplementary Table 3**). Only 119 (39.4%) tumours were uniformly classified as either low/intermediate or high by all five test; 30.8% ($n=93$) tumours were classified as low/intermediate risk by all tests, a further 8.6% ($n=26$) classified as high risk by all tests. The majority (60.6%; $n=183$) of cases gave no consensus result across all five tests. However for 31.1% of cases ($n=94$) agreement was observed in four of the five tests. There were also no clear differences between tests in terms of the agreement with other tests (**Table 5**). No statistically significant differences in clinico-pathological features between tumours that were concordant or discordant were observed (**Supplementary Table 4**). There is no evidence from the scatterplots of risk scores that individual tests were more likely to agree at the extremes of their ranges (**Supplementary Figure 2**). Disagreement spanning one of three risk categories was common, e.g. low risk to intermediate risk, and disagreement spanning two categories was not infrequent, i.e.

low risk to high risk (**Figure 1; Supplementary Table 5**). An exploratory analysis using a categorisation of low versus intermediate/high risk to more closely reflect current test usage was performed (**Supplementary Tables 6-7**) again modest agreement between tests was observed.

Intrinsic Subtypes

The three tests that provide subtype information categorised similar proportions of patients as having “luminal A” tumours (BluePrint™: 60.7%; 95%CI 55.2%-66.3%, Prosigna™: 59.5%; 95%CI 53.9-65.1% and MammaTyper™ (combined luminal A and low-risk luminal B): 62.4%; 95%CI 56.9-67.9%). Thirteen (4.3%) patients were classified as having HER2 enriched/positive tumours by at least one test. Two (0.7%) patients had basal like tumours using Prosigna™ subtyping; one of whom also had a basal like tumour using BluePrint™ but triple negative breast cancer using MammaTyper™. All these patients were classified as ER-positive and HER2-negative on central review. Agreement between all three tests providing subtype assignment was obtained for 179 (59.3%) patients; 121 (40.1%) tumours classified as luminal A; 58 (19.2%) as all other subtypes. Discordant results across these tests were seen in 123 (40.7%) patients. Moderate agreement between tests was confirmed by Kappa statistics of 0.39 (95%CI 0.29-0.50) between BluePrint™ and MammaTyper™, 0.44 (95%CI 0.34-0.54) between Prosigna™ and MammaTyper™, and 0.55 (95%CI 0.45-0.64) between BluePrint™ and Prosigna™ subtype.

Assessing relationship between the Prosigna™ subtyping and risk of recurrence score

Prosigna™ is unique amongst the multi-parameter assays evaluated in providing both a subtype and a continuous risk of recurrence score (ROR) with pre-defined risk categories derived from an identical set of genes. All 178 tumours classified as luminal A had a ROR score below the predefined high risk cut-point and none of the 113 luminal B tumours were classified as low-risk

(Table 6). Eight tumours, all of which were centrally confirmed as ER-positive/HER2-negative were categorised into either the basal like (n=2) or HER2-like (n=6) subtypes and these were either intermediate or high risk by ROR score respectively.

DISCUSSION

The evaluation of candidate multi-parameter tests within OPTIMA prelim to determine the best assessment of risk stratification for the main OPTIMA study presented a clinically significant challenge given: 1) Evidence that these tests provide broadly similar prognostic information at the population level (26); 2) The use of markedly different gene panels to estimate the same endpoint; 3) The use of different technologies including immunohistochemistry, PCR, quantitative and semi-quantitative array based technologies.

Previous in silico comparisons of multiple gene signatures have identified statistically significant discordance between different “diagnostic tests” (13;15-17). However, to date, this has been attributed to sub-optimal comparisons, since in the majority of studies genomic prediction scores have been estimated from published expression profiles. It has been argued that, in any direct comparison of validated diagnostic genomic assays, a high level of concordance could, and should be obtained (14). In the current study we performed such a direct comparison, each commercial assay was performed as prescribed by the relevant manufacturer (although the AQUA-IHC4 assay used TMAs for convenience). What is striking is that, amongst five tests with robust independent technical and clinical validation as predictors of residual risk (MammaPrint™, Oncotype DX™, Prosigna™, IHC4 and IHC4-AQUA) and three that measure a recognised risk factor (molecular subtype) there is disagreement across *all* tests. Indeed for all tests the level of agreement was “moderate” as defined by Prat et al, reaching only level 3 reproducibility (κ 0.40-0.59) (14) This suggests that agreement for risk classification between different molecular tests

applied to the same patient sample is on the level of agreement for pathological assessment of tumour grade.

The observed disagreement in risk categorisation for 60.6% of cases raises questions as to how patient management may be impacted by the choice of test used for risk stratification. Interestingly there does not seem to be better correlation between tests at the extremes of their ranges (the very low and high risk tumours in our cohort) than in the mid-range. It was less common, although not infrequent, for tumours placed into the lowest risk group by one test to be assigned into the highest risk group by another.

Each test is independently validated and adopted for prediction of risk of recurrence, so what should we do when they disagree? Paradoxically the result of this study can be viewed as either predictable or unexpected, depending on perspective. From a purely biological and technical perspective it is entirely predictable that tests which measure fundamentally different genes using different technologies give dissimilar results even when each individual assay remains technically valid. For example MammaPrint™ and Prosigna™, despite measuring the broadest range of genes (70 and 50 respectively) have only three genes in common and use different technical approaches (27;28). Even those tests measuring the same genes (IHC4, IHC4-AQUA and MammaTyper™) use different technologies (PCR versus IHC) or different antibodies, detection and quantification methods.

From a clinical perspective the disagreement between multiple tests each assessing residual risk is highly perplexing. The disagreement extends to an inability to demonstrate strong agreement on molecular subtyping between tests which again seems counter-intuitive. However, it is less surprising that disagreement between molecular subtyping, in this context predominantly between luminal A and luminal B, should exist in the absence of any clinical or molecular

agreement as to the true boundary between a “luminal A” and “luminal B” cancer (16). Again, the Prosigna™ and BluePrint™ tests for subtyping have minimal gene overlap with only seven genes in common.

What about risk prediction? The prediction of disease recurrence based on clinico-pathological and molecular features of a cancer is notoriously challenging within populations and even more so at the individual patient level. Biologically and clinically aggressive cancers which, if left untreated, are destined to progress may be “cured” by surgery, radiotherapy, chemotherapy or endocrine therapy. Tests predicting risk, therefore face an important challenge in that they seek to measure both the risk of recurrence based on the biology of tumours and must function within a clinical setting where biology may reflect risk that is not realised due to medical intervention. What then can we learn from comparisons between validated assays that seek to stratify patients by risk of recurrence, if indeed we can learn anything? We argue that there is value in such comparisons, even in the absence of outcome data. Each test applied in this study is externally validated and adopted or available for adoption in multiple clinical jurisdictions (6;27-32). However none is, or claims to be, the ultimate discriminator of risk for patients. This study suggests there is more than one way of predicting residual risk.

Our study is not without limitations. Whilst unable to determine subtle nuances in the performance of different tests within this population, we also recognise that existing data, both from the original studies validating individual tests and from comparisons, at a population level, of test performance in a single population (10-12) cannot provide a clear discrimination between them. No outcome data from OPTIMA prelim were available at the time of analysis. As the sample size is comparatively small it is highly unlikely that it will prove possible to compare the ability of the tests studied here to predict patient outcome.

In conclusion, in the widest and most comprehensive patient level direct diagnostic comparisons to date between multi-parametric tests of “residual risk” (after local treatment and endocrine therapy) we present further data that the proportions of patients identified as low, intermediate or high risk are broadly similar irrespective of which test is employed. However, both with respect to risk stratification and molecular sub-typing, marked differences were observed when categorisation of individual patients was considered. Such data, when considered with existing data on efficacy comparisons between different tests, support the conclusion that many current risk stratification tools are broadly equivalent and that further improvements in both prediction of relapse risk and therapeutic targeting would be of clinically significant value for patients at high risk of disease relapse (14).

FUNDING

This work was supported by the National Institute for Health Research Health Technology Assessment programme (grant number 10/34/01) and will be published in full in the Health Technology Assessment Journal. Further information available at: <http://www.nets.nihr.ac.uk/projects/hta/103401>. This publication presents independent research commissioned by the National Institute for Health Research. The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the National Health Service; National Institute for Health Research; Medical Research Council; Central Commissioning Facility; NIHR Evaluation, Trials and Studies Coordinating Centre; Health Technology Assessment programme; or Department of Health. Research at the Ontario Institute for Cancer Research is funded by the Government of Ontario. Agendia Inc., NanoString Technologies, Stratifyer/BioNTech Diagnostics, and Genoptix Medical Laboratories supported testing by provision of reagents and test results (as appropriate) at no financial cost to the current study. RCS was supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre.

NOTES

Acknowledgments

Trial Management Group

John M.S. Bartlett (Program Director & Hon. Professor, Ontario Institute of Cancer Research, Canada); David A. Cameron (Professor of Oncology & Head of Cancer Services, University of Edinburgh, UK); Amy Campbell (Clinical Trial Manager, Warwick Clinical Trials Unit, University of Warwick, UK); Peter Canney (Consultant Oncologist, retired); Jenny Donovan

(Professor of Social Medicine, University of Bristol, UK); Janet A. Dunn (Professor of Clinical Trials, University of Warwick, UK); Helena M. Earl (Reader in Clinical Cancer Medicine, University of Cambridge Department of Oncology and NIHR Cambridge Biomedical Research Centre, UK); Mary Falzon (Consultant Histopathologist, UCL Hospitals, London, UK); Adele Francis (Consultant Breast Surgeon, University Hospital Birmingham, UK); Peter S. Hall (Senior Lecturer and Consultant Medical Oncologist, University of Edinburgh & Visiting Health Economist, AUHE, University of Leeds, UK); Victoria Harmer (Breast Care Nurse, Imperial College NHS Healthcare Trust, London, UK); Helen Higgins (Senior Project Manager, Warwick Clinical Trials Unit, University of Warwick, UK); Luke Hughes-Davies (Consultant Oncologist, Addenbrookes Hospital, Cambridge, UK); Claire Hulme (Director Academic Unit of Health Economics, Leeds Institute of Health Sciences, Leeds, UK); Iain R. Macpherson (Clinical Senior Lecturer in Medical Oncology, University of Glasgow, Beatson West of Scotland Cancer Centre, UK); Andrea Marshall (Principal Research Fellow in Medical Statistics, University of Warwick, UK); Andreas Makris (Consultant Clinical Oncologist, Mount Vernon Cancer Centre, Northwood, UK); Christopher McCabe (Professor of Health Economics, University of Alberta, Canada); Adrienne Morgan (Patient Advocate & Chair of Independent Cancer Patients' Voice Trustees); Sarah E. Pinder (Professor of Breast Pathology, Kings College London, Guy's Hospital, UK); Christopher J. Poole (Professor of Medical Oncology, University Hospitals Coventry & Warwickshire NHS Trust, UK); Daniel W. Rea (Senior Lecturer in Medical Oncology, University of Birmingham, UK); Leila Rooshenas (Research Associate, University of Bristol, UK); Nigel Stallard (Professor of Medical Statistics, University of Warwick, UK); Robert C. Stein (Consultant Medical Oncologist & Hon. Senior Lecturer, UCL Hospitals, London, UK).

Participating Centres

The following centres and Principal Investigators contributed patients to the trial:

Addenbrooke's Hospital, Cambridge, Dr Luke Hughes-Davies; Alexandra Hospital, Redditch, Dr Denise Hrouda; University Hospital Ayr, Ayr, Dr Graeme Lumsden; Barnet Hospital, London, Dr Rob Stein; Beatson West of Scotland Cancer Centre, Glasgow, Dr Iain Macpherson; Bedford Hospital (Primrose Oncology Unit), Bedford, Dr Sarah Smith; Bristol Haematology and Oncology Centre, Bristol, Dr Jeremy Braybrooke; City Hospital, Birmingham, Dr Daniel Rea; Dumfries & Galloway Royal Infirmary, Dumfries, Dr Tamsin Evans; Forth Valley Royal Hospital, Larbet, Dr Judith Fraser; Hairmyres Hospital, Lanarkshire, Dr Grainne Dunn; Inverclyde Royal Hospital, Greenock, Dr Abdulla Alhasso; Luton & Dunstable University Hospital, Luton, Dr Mei-Lin Ah-See; Mount Vernon Hospital, Northwood, Dr Andreas Makris; Musgrove Park Hospital, Taunton, Dr John Graham; Norfolk and Norwich University Hospital, Norwich, Dr Adrian Hartnett; Northwick Park Hospital, Harrow, Dr Andreas Makris; Peterborough City Hospital, Peterborough, Dr Karen McAdam; Queen Elizabeth Hospital, Birmingham, Dr Daniel Rea; Queen Elizabeth Hospital, King's Lynn, Dr Margaret Daly; Royal Alexandra Hospital, Paisley, Dr Abdulla Alhasso; Royal Devon & Exeter Hospital, Exeter, Dr David Hwang; Royal Glamorgan Hospital, Llantrisant, Dr Jacinta Abraham; Royal United Hospital Bath, Bath, Dr Mark Beresford; St Bartholomew's Hospital, London, Dr Rebecca Roylance; The Christie, Manchester, Dr Anne Armstrong; The Woodlands Centre, Hinchingsbrooke, Dr Cheryl Palmer; Torbay Hospital, Torbay, Dr Andrew Goodman; University Hospital Coventry, Coventry, Professor Christopher Poole; University Hospital Crosshouse, Kilmarnock, Dr Graeme Lumsden; Velindre Cancer Centre, Cardiff, Dr Annabel Borley; Western General Hospital, Edinburgh, Dr Angela Bowman; Wishaw

General Hospital, Lanarkshire, Dr Jonathan Hicks; Yeovil District Hospital, Yeovil, Dr Urmila Barthakur; York District Hospital, York, Dr Andrew Proctor.

Author contributions

John M.S. Bartlett* (Program Director & Hon. Professor, Ontario Institute of Cancer Research, Canada) was the translational research lead for the trial. He contributed to study design and managed tissue banking, the establishment of commercial relationships for undertaking multi-parameter assays, the performance of laboratory assays and data analysis. He was responsible for drafting all sections of the paper and had final editorial responsibility.

Jane Bayani* (Research Scientist Ontario Institute of Cancer Research, Canada) was responsible for RNA extraction, Prosigna and IHC4 analysis and contributed to manuscript writing.

Andrea Marshall (Principal Research Fellow in Medical Statistics, University of Warwick, UK) is the trial statistician. She contributed to the statistical analysis plan, conducted the statistical analysis of the data and contributed to manuscript writing.

Janet A. Dunn (Professor of Clinical Trials, University of Warwick, UK) was the CTU lead and senior statistician for the study. She substantially contributed to the trial design, conduct including day-to-day management and monitoring as well as the statistical analysis plan.

Amy Campbell (Trial Manager, Warwick Clinical Trials Unit, University of Warwick, UK) was responsible for the day-to-day management of the trial and monitored data collection, sample collection & analysis.

Carrie Cunningham (Edinburgh Cancer Research Centre, University of Edinburgh, UK) was responsible for all aspects of sample collection, management checking pathology quality, TMA construction and sample shipping to various laboratories.

Monika S. Sobol (Edinburgh Cancer Research Centre, University of Edinburgh, UK) was responsible for all aspects of sample collection, management checking pathology quality, TMA construction and sample shipping to various laboratories.

Peter S. Hall (Senior Lecturer and Consultant Medical Oncologist, University of Edinburgh & Visiting Health Economist, AUHE, University of Leeds, UK) contributed to the health economics aspects of the study design and its overall conduct.

Christopher J. Poole (Professor of Medical Oncology, University Hospitals Coventry and Warwickshire NHS Trust, UK) contributed to the study design and its overall conduct, and advised on the clinical aspects of the trial.

David A. Cameron (Professor of Oncology & Head of Cancer Services, University of Edinburgh, UK) contributed to the study design and its overall conduct, and advised on the clinical aspects of the trial.

Helena M. Earl (Reader in Clinical Cancer Medicine, University of Cambridge Department of Oncology and NIHR Cambridge Biomedical Research Centre, UK) contributed to the study design and its overall conduct, and advised on the clinical aspects of the trial.

Daniel W. Rea (Senior Lecturer in Medical Oncology, University of Birmingham, UK) contributed to the study design and its overall conduct, and advised on the clinical aspects of the trial.

Iain R. Macpherson (Clinical Senior Lecturer in Medical Oncology, Beatson West of Scotland Cancer Centre, University of Glasgow, UK) contributed to the overall conduct of the study, advised on the clinical aspects of the trial and contributed to manuscript writing.

Peter Canney (Consultant Oncologist, Beatson West of Scotland Cancer Centre, Glasgow, UK, retired) contributed to the study concept and design and advised on the clinical aspects of the trial.

Adele Francis (Consultant Breast Surgeon, University Hospital Birmingham, UK) contributed to the study design and its overall conduct and advised on the surgical aspects of the trial.

Christopher McCabe (Professor of Health Economics, University of Alberta, Canada) contributed to the health economics aspects of study design.

Sarah E. Pinder (Professor of Breast Pathology, Kings College London, UK) contributed to the trial design, advised on pathology aspects of trial conduct.

Luke Hughes-Davies (Consultant Oncologist, Addenbrookes Hospital, Cambridge, UK) is co-chief investigator. He contributed to the concept and design of the study, its day-to-day management and overall conduct.

Andreas Makris (Consultant Clinical Oncologist, Mount Vernon Hospital, Northwood, UK) is co-chief investigator. He contributed to the concept and design of the study, its day-to-day management and overall conduct.

Robert C. Stein (Consultant Medical Oncologist & Hon. Senior Lecturer, UCL Hospitals, London, UK) is chief investigator and lead of clinical aspects of the trial. He made substantial contributions to the concept and design of the study, its day-to-day management and overall conduct, data analysis and contributed to manuscript writing.

On behalf of the OPTIMA TMG

*These authors contributed equally to this work.

Role of study sponsor

The sponsors of this study had no role in study design, data collection, analysis, interpretation, writing of the report, or the decision to publish. The authors had full access to the data and are responsible for the content of this manuscript.

REFERENCE LIST

- (1) McGuire WL. Estrogen receptors in human breast cancer. *J Clin Invest* 1973;52(1):73-7.
- (2) McGuire WL, Chamness GC, Costlow ME, Shepherd RE. Hormone dependence in breast cancer. *Metabolism* 1974;23(1):75-100.
- (3) Slamon DJ, Clark GM, Wong SG. Human breast cancer: Correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* 1987;235(4785).
- (4) Perou CM, Sorlie T, Eisen MB, Van de Rijn M, Jeffrey SS, Rees CA et al. Molecular portraits of human breast tumours. *Nature* 2000;406(6797):747-52.
- (5) Perou CM, Jeffrey SS, Van de Rijn M, Rees CA, Eisen MB, Ross DT et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences of the United States of America* 1999;96(16):9212-7.
- (6) Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New Engl J Med* 2004;351(27):2817-26.
- (7) Bartlett J, Canney P, Campbell A, Cameron D, Donovan J, Dunn J et al. Selecting breast cancer patients for chemotherapy: the opening of the UK OPTIMA trial. *Clin Oncol (R Coll Radiol)* 2013;25(2):109-16.

- (8) Paik S, Tang G, Shak S, Kim C, Baker J, Kim W et al. Gene Expression and Benefit of Chemotherapy in Women With Node-Negative, Estrogen Receptor-Positive Breast Cancer. *Journal of Clinical Oncology* 2006;24(23):3726-34.
- (9) Albain KS, Barlow WE, Shak S, Hortobagyi GN, Livingston RB, Yeh IT et al. Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. *The Lancet Oncology* 11(1):55-65.
- (10) Dowsett M, Sestak I, Lopez-Knowles E, Sidhu K, Dunbier AK, Cowens J et al. Comparison of PAM50 risk of recurrence score with oncotype DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. *J Clin Oncol* 2013;31(22):2783-90.
- (11) Sgroi DC, Sestak I, Cuzick J, Zhang Y, Schnabel CA, Schroeder B 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* 2013;14(11):1067-76.
- (12) Cuzick J, Dowsett M, Pineda S, Wale C, Salter J, Quinn E 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. *Journal of Clinical Oncology* 2011;29(32):4273-8.

- (13) Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, Nobel AB et al. Concordance among gene-expression-based predictors for breast cancer. *N Engl J Med* 2006;355(6):560-9.
- (14) Prat A, Ellis MJ, Perou CM. Practical implications of gene-expression-based assays for breast oncologists. *Nat Rev Clin Oncol* 2012;9(1):48-57.
- (15) Kelly CM, Bernard PS, Krishnamurthy S, Wang B, Ebbert MT, Bastien RR 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* 2012;17(4):492-8.
- (16) Mackay A, Weigelt B, Grigoriadis A, Kreike B, Natrajan R, A'Hern R et al. Microarray-based class discovery for molecular classification of breast cancer: analysis of interobserver agreement. *J Natl Cancer Inst* 2011;103(8):662-73.
- (17) Weigelt B, Mackay A, A'Hern R, Natrajan R, Tan DS, Dowsett M et al. Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *Lancet Oncol* 2010;11(4):339-49.

- (18) Stein RC, Dunn JA, Bartlett JMS, Campbell AF, Marshall A, Hall Pet al. OPTIMA: a randomised feasibility study of personalised care in the treatment of women with early breast cancer. Health Technology Assessment (South Hampton, NY) 2015.
- (19) Bartlett JMS, Brookes CL, Robson T, van de Velde CJH, Billingham LJ, Campbell FMet al. Estrogen Receptor and Progesterone Receptor As Predictive Biomarkers of Response to Endocrine Therapy: A Prospectively Powered Pathology Study in the Tamoxifen and Exemestane Adjuvant Multinational Trial. Journal of Clinical Oncology 2011;29(12):1531-8.
- (20) Ravdin PM, Siminoff LA, Davis GJ, Mercer MB, Hewlett J, Gerson Net al. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. J Clin Oncol 2001;19(4):980-91.
- (21) Wishart GC, Bajdik CD, Azzato EM, Dicks E, Greenberg DC, Rashbass Jet al. A population-based validation of the prognostic model PREDICT for early breast cancer. Eur J Surg Oncol 2011;37(5):411-7.
- (22) Wishart GC, Bajdik CD, Dicks E, Provenzano E, Schmidt MK, Sherman Met al. PREDICT Plus: development and validation of a prognostic model for early breast cancer that includes HER2. Br J Cancer 2012;107(5):800-7.

- (23) Wishart GC, Azzato EM, Greenberg DC, Rashbass J, Kearins O, Lawrence Get al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res* 2010;12(1):R1.
- (24) Bangdiwala SJ, Shankar V. The Agreement Chart. *BMC Med.Res.Methodol.* 13, 97. 2013.
- (25) R Core Team. A language and environment for statistical computing. 2014.
- (26) Galea M, Blamey R, Elston C, Ellis I. The Nottingham prognostic index in primary breast cancer. *Breast Cancer Res Tr* 1992;22(3):207-19.
- (27) Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery Tet al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology* 2009;27(8):1160-7.
- (28) Chang JC, Makris A, Gutierrez MC, Hilsenbeck SG, Hackett JR, Jeong Jet al. Gene expression patterns in formalin-fixed, paraffin-embedded core biopsies predict docetaxel chemosensitivity in breast cancer patients. *Breast Cancer Res Treat* 2008;108(2):233-40.
- (29) Cuzick J, Dowsett M, Wale C, Salter J, Quinn E, Zabaglo Let al. Prognostic Value of a Combined ER, PgR, Ki67, HER2 Immunohistochemical (IHC4) Score and Comparison with the GHI Recurrence Score - Results from TransATAC. *Cancer Res* 2009;69(24):503S.

- (30) Dowsett M, Cuzick J, Wale C, Forbes J, Mallon EA, Salter J 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* 2010;28(11):1829-34.
- (31) Nielsen TO, Parker JS, Leung S, Voduc D, Ebbert M, Vickery T et al. A Comparison of PAM50 Intrinsic Subtyping with Immunohistochemistry and Clinical Prognostic Factors in Tamoxifen-Treated Estrogen Receptor-Positive Breast Cancer. *Clin Cancer Res* 2010;16(21):5222-32.
- (32) Chia SK, Bramwell VH, Tu D, Shepherd LE, Jiang S, Vickery T et al. A 50-Gene Intrinsic Subtype Classifier for Prognosis and Prediction of Benefit from Adjuvant Tamoxifen. *Clin Cancer Res* 2012;18(16):4465-72.

Table 1: Characteristics of the 302 patients

Characteristic	Total	
	n	%
Age years, Median(Range)	58 (40-78)	
Menopausal status of participant		
Pre/peri-menopausal	97	32.1
Postmenopausal	205	67.9
Number of involved nodes		
None	57	18.9
1-3	192	63.6
4-9	42	13.9
+ve sentinel node biopsy without clearance surgery	11	3.6
Histological grade		
1	19	6.3
2	201	66.6
3	82	27.1
Largest tumour size in mm, Median(Range)	28 (2-170)	
<=30mm	172	57.0
>30mm	130	43.0
Lymphovascular invasion reported		
No	169	56.0
Yes	122	40.4
Not Known	11	3.6
Tumour type		
Ductal	214	70.9
Lobular	65	21.5
Tubular/Cribriform	2	0.7
Mucinous	4	1.3
Micropapillary	1	0.3
Mixed	16	5.3

Table 2: Clinical risk of patients (n=302)

Risk of Patient	Median (range)
Nottingham Prognostic Index	4.6 (2.8-8.2)
≤3.4, No. (%)	12 (4.0%)
>3.4 - ≤5.4, No. (%)	226 (74.8%)
>5.4, No. (%)	64 (21.2%)
PREDICT 10 year overall survival (%)	
Endocrine therapy only	77.0 (25.1-94.6)
Chemotherapy and endocrine therapy	82.6 (39.8-95.9)
Additional benefit of chemotherapy	5.5 (1.2-25.8)
Adjuvant! 10 year risk overall survival (%)	
Endocrine therapy only	68.6 (25.4-90.4)
Chemotherapy and endocrine therapy	76.4 (31.0-93.6)
Additional benefit of chemotherapy	6.8 (1.2-25.8)
Adjuvant! 10 year relapse free survival (%)	
Endocrine therapy only	60.5 (22.0-82.1)
Chemotherapy and endocrine therapy	72.9 (29.1-89.4)
Additional benefit of chemotherapy	10.5 (2.7-33.3)

Table 3: Risk categorisation by each test

Risk group	Oncotype DX*	MammaPrint†	Prosigna	IHC4	IHC4-AQUA‡
No (%)	301 (99.7%)	298 (98.9%)	299 (99.0%)	257 (85.1%)	271 (89.7%)
Low risk	163 (54.2%)	183 (61.4%)	108 (36.1%)	62 (24.1%)	87 (32.1%)
Intermediate risk	84 (27.9%)	--	88 (29.4%)	123 (47.9%)	80 (29.5%)
Mid risk	--	--	--	--	55 (20.3%)
High risk	54 (17.9%)	115 (38.6%)	103 (34.5%)	72 (28.0%)	49 (18.1%)

*Oncotype DX is divided into three risk groups with intermediate defined as Recurrence Score 18-25 for the current analysis.

†MammaPrint divides tumours into two risk groups only.

‡IHC4-AQUA divides tumours into four risk groups: low, low-mid (here called intermediate), mid and high (combined as high risk).

Table 4: Kappa statistics and 95% confidence interval (CI) for tests providing risk predictions*

Test	MammaPrint (Low), Kappa statistic (95%CI)	Prosigna (Low/ Intermediate) , Kappa statistic (95%CI)	IHC4 (Low/ Intermediate), Kappa statistic (95%CI)	IHC4-AQUA† (Low/Low-Mid), Kappa statistic (95%CI)
Oncotype DX (Recurrence Score ≤ 25)	0.40 (0.30-0.49)	0.44 (0.33-0.54)	0.53 (0.41-0.65)	0.40 (0.30-0.51)
MammaPrint	--	0.53 (0.43-0.63)	0.33 (0.21-0.44)	0.42 (0.30-0.53)
Prosigna (Low/Intermediate)	--	--	0.39 (0.27-0.50)	0.43 (0.31-0.54)
IHC4 (Low/Intermediate)	--	--	--	0.60 (0.50-0.70)

*Kappa statistics are for agreement between categorisation into combined low and intermediate risk versus high risk.

†IHC4-AQUA mid risk and high risk are combined for this analysis.

Table 5: Number of tests agreeing with each test

Number of other tests agreed with test	Oncotype DX, No (%)	Prosigna, No (%)	MammaPrint, No (%)	IHC4, No (%)	IHC4-AQUA, No (%)
4	119 (39.4%)	119 (39.4%)	119 (39.4%)	119 (39.4%)	119 (39.4%)
3	84 (27.8%)	77 (25.5%)	73 (24.2%)	67 (22.2%)	75 (24.8%)
2	54 (17.9%)	52 (17.2%)	47 (15.6%)	36 (11.9%)	33 (10.9%)
1	31 (10.3%)	33 (10.9%)	34 (11.2%)	25 (8.3%)	27 (9.0%)
0	13 (4.3%)	18 (6.0%)	25 (8.3%)	10 (3.3%)	17 (5.6%)
Missing	1 (0.3%)	3 (1.0%)	4 (1.3%)	45 (14.9%)	31 (10.3%)

Table 6: Relationship between Prosigna subtyping and the continuous risk of recurrence (ROR) score

Prosigna test result	Subtype			
	Luminal A No (%)	Luminal B No (%)	Basal like No (%)	HER2 enriched No (%)
No. of patients	178 (59.5%)	113 (37.8%)	2 (0.7%)	6 (2.0%)
Median ROR (Inter-quartile range)	37 (28-44)	70 (63-78)	53 (47-58)	76 (72-78)
Range	5-59	43-96	47-58	64-84
Risk Groups				
Low Risk	108 (60.7%)	0	0	0
Intermediate Risk	70 (39.3%)	16 (14.2%)	2 (100%)	0
High Risk	0	97 (85.8%)	0	6 (100%)

Figure Legend

Figure 1: Agreement charts for two by two comparison of tests according to risk groups. A)

Prosigna against Oncotype DX; B) IHC4 against Oncotype DX; C) IHC4-AQUA against Oncotype DX; D) IHC4 against Prosigna; E) IHC4-AQUA against Prosigna; F) IHC4 against IHC4-AQUA. Only tests that provide three risk categories are included in this analysis. The Oncotype DX intermediate risk group is defined as RS 18-25. The IHC4-AQUA mid risk group was combined with the high risk group. Rectangles are drawn for each level of the test outcomes, i.e. low, intermediate and high risk, based on the row and column cumulative totals. Thus for the low risk rectangle of the test 1 vs test 2 comparison, all tumours categorised as low risk by either test are included. The boundaries of the rectangles along both axes represent the number of tumours that were categorised as that outcome for each test. Black squares within the rectangles represent exact agreement between the levels of the two tests, e.g. both low scores, and are of size based on the cell frequencies and located according to the cumulative totals of the previous levels. Grey rectangles represent partial agreement, where the scores from one test are within one level of those from the other test, i.e. a low score on one test but intermediate on the other test. White areas within the rectangle reflect disagreement by more than level, i.e. low scores on one test and high scores on the other test.