SPAG5 as a prognostic biomarker and chemotherapy sensitivity predictor in breast cancer: a retrospective, integrated genomic, transcriptomic, and protein analysis

Tarek M A Abdel-Fatah, Devika Agarwal, Dong-Xu Liu, Roslin Russell, Oscar M Rueda, Karen Liu, Bing Xu, Paul M Moseley, Andrew R Green, Alan G Pockley, Robert C Rees, Carlos Caldas, Ian O Ellis, Graham R Ball*, Stephen Y T Chan**

Lancet Oncol 2016; 17: 1004–18

Published Online June 13, 2016 http://dx.doi.org/10.1016/S1470-2045(16)00174-1

See Comment page 863

*Contributed equally

Clinical Oncology Department, Nottingham University Hospitals NHS Trust, Nottingham, UK (T M A Abdel-Fatah PhD, P M Moseley BSc, Prof S Y T Chan DM); **John van Geest Cancer Research Centre, School of Science and Technology, Nottingham Trent University, Clifton Campus, Nottingham, UK** (D Agarwal MSc, Prof A G Pockley PhD, Prof R C Rees PhD, Prof G R Ball PhD); **Liggins Institute, University of Auckland, Auckland, New Zealand** (D-X Liu PhD, K Liu MSc, B Xu PhD); **The Institute of Genetics and Cytology, Northeast Normal University, Changchun, China** (D-X Liu); **Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Cambridge, UK** (R Russell PhD, O M Rueda PhD, Prof C Caldas FMedSci); **and Division of Cancer and Stem Cells, School of Medicine, University of Nottingham and Nottingham University Hospitals NHS Trust, Nottingham, UK** (A R Green PhD, Prof I O Ellis FRCPath)

Correspondence to: Prof Stephen Y T Chan, Clinical Oncology Department, Nottingham University Hospitals NHS Trust, Nottingham NG5 1PB, UK **steve.chan@nuh.nhs.uk**

Summary

Background Proliferation markers and profiles have been recommended for guiding the choice of systemic treatments for breast cancer. However, the best molecular marker or test to use has not yet been identified. We did this study to identify factors that drive proliferation and its associated features in breast cancer and assess their association with clinical outcomes and response to chemotherapy.

Methods We applied an artificial neural network-based integrative data mining approach to data from three cohorts of patients with breast cancer (the Nottingham discovery cohort (n=171), Uppsala cohort (n=249), and Molecular Taxonomy of Breast Cancer International Consortium [METABRIC] cohort; n=1980). We then identified the genes with the most effect on other genes in the resulting interactome map. Sperm-associated antigen 5 (*SPAG5*) featured prominently in our interactome map of proliferation and we chose to take it forward in our analysis on the basis of its fundamental role in the function and dynamic regulation of mitotic spindles, mitotic progression, and chromosome segregation fidelity. We investigated the clinicopathological relevance of *SPAG5* gene copy number aberrations, mRNA transcript expression, and protein expression and analysed the associations of *SPAG5* copy number aberrations, transcript expression, and protein expression with breast cancer-specific survival, disease-free survival, distant relapse-free survival, pathological complete response, and residual cancer burden in the Nottingham discovery cohort, Uppsala cohort, METABRIC cohort, a pooled untreated lymph node-negative cohort (n=684), a multicentre combined cohort (n=5439), the Nottingham historical early stage breast cancer cohort (Nottingham-HES; n=1650), Nottingham early stage oestrogen receptor-negative breast cancer adjuvant chemotherapy cohort (Nottingham-oestrogen receptor-negative-ACT; n=697), the Nottingham anthracycline neoadjuvant chemotherapy cohort (Nottingham-NeoACT; n=200), the MD Anderson taxane plus anthracycline-based neoadjuvant chemotherapy cohort (MD Anderson-NeoACT; n=508), and the multicentre phase 2 neoadjuvant clinical trial cohort (phase 2 NeoACT; NCT00455533; n=253).

Findings In the METABRIC cohort, we detected *SPAG5* gene gain or amplification at the Ch17q11.2 locus in 206 (10%) of 1980 patients overall, 46 (19%) of 237 patients with a PAM50-HER2 phenotype, and 87 (18%) of 488 patients with PAM50-LumB phenotype. Copy number aberration leading to *SPAG5* gain or amplification and high *SPAG5* transcript and SPAG5 protein concentrations were associated with shorter overall breast cancer-specific survival (METABRIC cohort [copy number aberration]: hazard ratio [HR] 1·50, 95% CI 1·18–1·92, p=0·00010; METABRIC cohort [transcript]: 1·68, 1·40–2·01, p<0·0001; and Nottingham-HES-breast cancer cohort [protein]: 1·68, 1·32–2·12, p<0·0001). In multivariable analysis, high *SPAG5* transcript and SPAG5 protein expression were associated with reduced breast cancer-specific survival at 10 years compared with lower concentrations (Uppsala: HR 1·62, 95% CI 1·03–2·53, p=0·036; METABRIC: 1·27, 1·02–1·58, p=0·034; untreated lymph node-negative cohort: 2·34, 1·24–4·42, p=0·0090; and Nottingham-HES: 1·73, 1·23–2·46, p=0·0020). In patients with oestrogen receptor-negative breast cancer with high SPAG5 protein expression, anthracycline-based adjuvant chemotherapy increased breast cancer-specific survival overall compared with that for patients who did not receive chemotherapy (Nottingham-oestrogen receptor-negative-ACT cohort: HR 0·37, 95% CI 0·20–0·60, p=0·0010). Multivariable analysis showed high *SPAG5* transcript concentrations to be independently associated with longer distant relapse-free survival after receiving taxane plus anthracycline neoadjuvant chemotherapy (MD Anderson-NeoACT: HR 0·68, 95% CI 0·48–0·97, p=0·031). In multivariable analysis, both high *SPAG5* transcript and high SPAG5 protein concentrations were independent predictors for a higher proportion of patients achieving a pathological complete response after combination cytotoxic chemotherapy (MD Anderson-NeoACT: OR 1·71, 95% CI, 1·07–2·74, p=0·024; Nottingham-ACT: 8·75, 2·42–31·62, p=0·0010).

Interpretation *SPAG5* is a novel amplified gene on Ch17q11.2 in breast cancer. The transcript and protein products of *SPAG5* are independent prognostic and predictive biomarkers that might have clinical utility as biomarkers for combination cytotoxic chemotherapy sensitivity, especially in oestrogen receptor-negative breast cancer.

Funding Nottingham Hospitals Charity and the John and Lucille van Geest Foundation.

Research in context

Evidence before this study

We searched PubMed with the term "biomarkers of proliferation that predict response to chemotherapy in Breast cancer" on May 1, 2012, to review the state of the literature. Since then, advances in molecular biology have generated a huge amount of data, which have then been used to generate multigene profiles to guide chemotherapy treatment. However, most of these approaches face common issues, such as insufficiently high levels of evidence, overfitting of computational models, high false discovery rates, and absence of potential biological mechanisms to support their use as accurate predictors of therapeutic response. As most of the prognostic power of these assays comes from genes related to cell proliferation, a St Gallen International Expert Consensus has recommended the use of proliferation markers or profiles when choosing appropriate systemic treatments for breast cancer. However, the best molecular marker or test to use continues to be debated.

Added value of this study

We used artificial neural network analysis to identify *SPAG5* as a potential prognostic and predictive biomarker in breast cancer and validated it in ten breast cancer cohorts. Our findings suggest that amplification or gain of the *SPAG5* locus at Ch17q11.2 occurs in 10–19% of all breast cancers; *SPAG5* gene copy number aberrations and its transcript and protein are associated with poor clinical outcome and adverse clinicopathological features; and that high expression of *SPAG5* mRNA transcript and SPAG5 protein are independent predictors for response to anthracycline-based chemotherapy. To the best of our knowledge this is the first multidimensional study (ie, one involving interacting variables and parameters such as DNA copy number aberrations, RNA transcription, and protein expression, as well as the clinical variables, treatment variables, and effect on patient survival) to report on the clinicopathological relevance of *SPAG5* as a predictive marker in breast cancer.

Implications of all the available evidence

The transcript and protein products of *SPAG5* are independent prognostic and predictive biomarkers that might have clinical utility as biomarkers for combination cytotoxic chemotherapy sensitivity, especially in oestrogen receptor-negative breast cancer. Our findings have the potential to introduce an accurate predictive biomarker for chemotherapy response, which would help to tailor treatments to individual patients with breast cancer. This work could lead to the development of novel strategies for more effective management and treatment of breast cancer.

Introduction

Roughly 1·68 million women are diagnosed with breast cancer worldwide each year, with more than 500 000 dying of the disease (about 1400 deaths per day).[1] Despite chemotherapy, either alone or in combination with other targeted therapies, being offered to about 60% of patients with breast cancer,[2] results from a meta-analysis of 123 randomised trials including more than 100 000 patients showed that chemotherapy reduces recurrence and mortality by only 20–33%.[3] The delivery of effective precision medicine for breast cancer requires the discovery of novel therapeutic targets in subgroups of breast cancer and improvements in the efficacy of treatments through identification of stratification biomarkers that predict an individual patient's response to a particular therapy.[4] Although a St Gallen International Expert Consensus[5] recommended the use of proliferation markers and profiles for the selection of appropriate systemic treatments, debate continues about which molecular marker or test is the best to use.

The main aim of our study was to identify a proliferation biomarker that could be used to stratify outcomes for breast cancer patients. We used an artificial neural network algorithm[6] to identify a candidate proliferation-related gene and assessed its association with clinicopathological features and outcomes in patients with breast cancer.

Methods

Study design and cohorts

This study was a retrospective, integrated genomic, transcriptomic, and protein analysis done in ten international cohorts of patients with breast cancer.

All patients provided written informed consent for excess tumour tissue to be used in research. The study was approved by the institutional review boards or independent ethics committees and the hospital research and innovations departments at all participating sites. We followed the Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK) criteria[7] throughout this study.

The Nottingham discovery cohort consisted of a set of 171 patients with stage I and II invasive breast cancer, who we have described previously.[8] Median follow-up in this cohort was 180 months (IQR 143–194). All 171 patients had primary operable early stage (I and II) invasive breast carcinomas with relatively small size and low Nottingham prognostic index treated from 1990 to 1996 at the Nottingham City Hospital, Nottingham, UK. Patients within the good prognosis group (Nottingham prognostic index <3·4) did not receive systemic adjuvant therapy. Postmenopausal, oestrogen receptor-positive patients with Nottingham prognostic index greater than 3·4 were offered hormonal therapy.

The Uppsala cohort included 315 women with early stage (stages I and II) breast cancer, representing 65% of

all breast cancers resected in Uppsala County, Sweden, between 1987 and 1989.[9] Median follow-up was 126 months (IQR 119–134). Gene expression data were available for only 249 patients; only these patients were included in our study.

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort was a set of 1980 patients with early stage (stages I and II) breast cancers. Median follow-up was 109 months (IQR 62–155).[10] In this cohort, patients with oestrogen receptor-negative and lymph node-positive breast cancer received adjuvant chemotherapy whereas those with oestrogen receptor-positive and lymph node-negative breast cancer did not. Additionally, no patients with HER2 overexpression received trastuzumab.

To create a cohort of patients with untreated lymph node-negative breast cancer, we pooled data from three publically available datasets with a total of 684 patients with early stage (stage I) breast cancer (Wang and colleagues,[11] n=286; Desmedt and colleagues,[12] n=198; and Schmidt and colleagues,[13] n=200). These patients did not receive any adjuvant systemic therapy, and the samples represent the natural history of the disease in these patients. The median follow-up of each cohort is shown in the appendix (pp 1–3).

See Online for appendix

We also created a large combined breast cancer dataset, the multicentre combined cohort, which included 5439 patients with early stage breast cancer (IA, IIA, and B) and was sourced from 36 publically available international datasets by use of online bc-GenExMiner version 4.0.[14] A list of all the datasets used to create the multicentre combined cohort is shown in the appendix (pp 4–6).

The Nottingham historical early stage breast cancer cohort (Nottingham-HES) consisted of 1650 patients with breast cancer (age <71 years, stage IA, IIA, or B)[15] whose tissues were suitable for immunohistochemistry. This cohort was comprised of 1285 patients with oestrogen receptor-positive breast cancer and 365 patients with oestrogen receptor-negative breast cancer. These patients were diagnosed and treated according to the same protocol between 1986 and 1999 at the Nottingham City Hospital, Nottingham, UK. All patients underwent surgery. Patients with good prognoses (Nottingham prognostic index <3·4) did not receive systemic adjuvant therapy. Premenopausal patients with moderate or poor prognoses were eligible for chemotherapy with intravenous cyclophosphamide, methotrexate, and fluorouracil (cyclophosphamide 750 mg/m², methotrexate 50 mg/m², and fluorouracil 1 g/m² on day 1 of a 21-day cycle). Postmenopausal oestrogen receptor-positive patients with moderate or poor Nottingham prognostic index scores were offered hormonal therapy (>90% accepted this therapy), whereas oestrogen receptor-negative patients received chemotherapy with cyclophosphamide, methotrexate, and fluorouracil. Clinical data were maintained on a prospective basis with a median follow-up of 143 months (IQR 114–174).[15] The median follow-up of the individual prognostic subgroups is summarised in the appendix (pp 1–3).

We also included four cohorts of patients to specifically assess biomarkers in the adjuvant and neoadjuvant chemotherapy settings. We analysed a consecutive series of 697 patients with early stage (stages IA, IIA, and B) oestrogen receptor-negative breast cancers who had been diagnosed and managed at Nottingham City Hospital between 1986 and 2007 (the Nottingham early stage oestrogen receptor-negative breast cancer adjuvant chemotherapy cohort [Nottingham oestrogen receptor-negative-ACT]). This series included the oestrogen receptor-negative patients from the Nottingham-HES cohort who were managed between 1986 and 2000 and treated with either no chemotherapy or adjuvant chemotherapy (n=332), and patients with oestrogen receptor-negative early stage breast cancer who were managed from 2001 and received either no chemotherapy or intravenous anthracycline-based adjuvant chemotherapy (fluorouracil 500–700 mg/m², epirubicin 75–100 mg/m², and cyclophosphamide 500–700 mg/m²; n=365).[16] The median follow-up of the different treatment subgroups is summarised in the appendix (pp 1–3).

The Nottingham anthracycline-based neoadjuvant chemotherapy (Nottingham-NeoACT) cohort consisted of core biopsies and post-chemotherapy surgical specimens from 200 female patients with locally advanced primary breast cancer (stage IIIA–C) who had been treated with anthracycline-based neoadjuvant chemotherapy[17] at Nottingham City Hospital between 1996 and 2012. There was no overlap of patients with Nottingham-HES. All patients received six cycles of an intravenous anthracycline-based therapy (fluorouracil 500 mg/m², epirubicin 75–100 mg/m², and cyclophosphamide 500 mg/m² on day 1 of a 21-day cycle) and 73 (37%) patients also received a taxane (intravenous paclitaxel 175 mg/m² or docetaxel 75–100 mg/m²). All patients underwent mastectomy or breast-conserving surgery and axillary dissection, followed by adjuvant radiation therapy. Patients with oestrogen receptor-positive breast cancer were offered 5 years of adjuvant endocrine therapy. The median follow-up time was 67 months (IQR 27–81).

We included 508 patients from the University of Texas MD Anderson Cancer Center taxane plus anthracycline-based neoadjuvant chemotherapy (MD Anderson-NeoACT) cohort who we selected for newly diagnosed HER2-negative breast cancer and treatment with sequential taxane-based and anthracycline-based neoadjuvant chemotherapy (followed by endocrine adjuvant therapy if oestrogen receptor positive). The patient characteristics and details of the drugs, doses, and routes of administration used have been previously reported.[18] The median follow-up time was 38 months (IQR 26–53).

The phase 2 NeoACT clinical trial cohort was from a randomised, open-label, multicentre, phase 2 clinical trial (NCT00455533) in which women with early stage breast

cancer (T2–3, N0–3, M0, tumour size ≥2 cm) received anthracycline-based neoadjuvant chemotherapy regimens (cyclophosphamide plus doxorubicin, followed by ixabepilone or paclitaxel). Patients received four cycles of doxorubicin (60 mg/m² intravenously) and cyclophosphamide (600 mg/m² intravenously) every 3 weeks followed by 1:1 randomisation to either ixabepilone (40 mg/m², 3-h infusion) every 3 weeks for four cycles or paclitaxel (80 mg/m², 1-h infusion) weekly for 12 weeks). Full details of the study design and the patient characteristics have been reported previously.[19] Of 295 patients enrolled into the trial, 253 patients had available data about gene expression and pathological complete response and were included in our study.

A control cohort consisted of 85 individuals of European ancestry for whom non-cancerous tissues were genotyped and gene expression values from normal tissue were available from the METABRIC study.[10]

Procedures

To identify factors that could drive proliferation and its associated features in breast cancer, we analysed several factors that are directly and indirectly related to proliferation, which we defined as clinical class questions (histological grade; mitotic index; *Ki67*, *TOP2A*, *KIF2C*, and *BIRC5* expression; and 5-year survival) through application of an artificial neural network modelling-based data mining approach. We applied this approach to three gene expression array transcriptomic datasets: the Nottingham discovery, Uppsala, and METABRIC cohorts. We selected artificial neural networks to data mine the clinical datasets because such networks have previously been shown to be able to identify biomarkers that have high sensitivity and specificity for clinical features and excellent validity.[6] Additionally, artificial neural networks, unlike conventional statistical approaches (such as hierarchical clustering, principal components analysis, or linear regression), are not limited by linear functionality; this provides improved representation of biological features. We compared the ranked orders of genes produced in this way across several proliferation-related clinical class questions within a given dataset. We compared the top 200 ranked genes for the prediction of each clinical class question, based on minimum average route mean squared error and identified commonalities at the probe level. We then made further comparisons for the same clinical class questions in the other datasets to establish a consensus list of gene probes across all of the features and datasets. The strongest 100 integrated interactions were selected for visualisation in Cytoscape version 3.1.1.[20] Further details of the artificial neural network approach are in the appendix (pp 7–10).

Sperm-associated antigen 5 (*SPAG5*) featured prominently (ie, had the most effect on other genes) in the interactome map of proliferation that was produced by our artificial neural network analysis. Given that SPAG5 has a fundamental role in the function and dynamic regulation of mitotic spindles and in mitotic progression and chromosome segregation fidelity,[21] we postulated that *SPAG5* could be a novel measurement of proliferation activity and provide a biomarker for the efficacy of systemic therapies in breast cancer, and therefore took *SPAG5* forward for all subsequent analyses. Because *Ki67* has been used by many investigators as a marker for proliferation when choosing appropriate systemic treatments, we used it as a control in our study.

We retrieved data about copy number aberrations at the *SPAG5* locus on chromosome 17q11.2 from high resolution (<100 kb) oligonucleotide microarrays (Nottingham discovery cohort; n=171), comparative genomic hybridisation (Nottingham discovery cohort), and Genome-Wide Human SNP Array 6.0 platform (Affymetrix, Santa Clara, CA, USA) profiling (METABRIC cohort), which we have described previously.[8,10] The oligonucleotide array data can be accessed online (series accession number GSE8757) and the SNP data are available through the European Genotype Archive (accession number EGAS00000000082). We did an additional copy number aberration analysis in our control set of 85 individuals of European ancestry.[10]

We retrieved and analysed *SPAG5* and *Ki67* mRNA expression data in the Nottingham discovery, Uppsala, and METABRIC cohorts. We used Agilent gene expression arrays (Agilent, Santa Clara, CA, USA) for the Nottingham discovery cohort (accession number E-TABM-576), Affymetrix U133A&B gene chip microarray profiling for the Uppsala cohort (series accession number GSE4922), and HumanHT-12 v3 Expression BeadChip arrays (Illumina, San Diego, CA, USA)[10] for the METABRIC cohort (accession number EGAS00000000082). Additionally, we retrieved *SPAG5* and *Ki67* mRNA expression data from the three publically available datasets of lymph node-negative breast cancer in which patients did not receive any adjuvant systemic therapy: Wang and colleagues[11] (accession number GSE2034), Desmedt and colleagues[12] (accession number GSE7390), and Schmidt and colleagues[13] (accession number GSE11121). For the multicentre combined cohort, details of the gene expression data processing, normalisation, and statistical testing have been described previously.[14] In the multicentre combined cohort, gene expression data were converted to a common scale (median equal to 0 and standard deviation equal to 1) to merge all of the data from all of the included studies and to create combined cohorts (appendix p 8).[22] We also downloaded gene expression data from the MD Anderson-NeoACT cohort (accession number GSE25066) and the phase 2 NeoACT clinical trial (accession number GSE41998). In each cohort, we used the median as the cutoff between low and high expression of the gene (ie, high expression if >median or low expression if <median).

For the **oligonucleotide array data** see http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8757

For the **European Genotype Archive** see http://www.ebi.ac.uk/ega/page.php

For the **mRNA transcription data for the Nottingham discovery cohort** see http://www.ebi.ac.uk/miamexpress/

For the **mRNA transcription data for the Uppsala cohort** see http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4922

For the **mRNA transcription data for the METABRIC cohort** see https://www.ebi.ac.uk/ega/studies/EGAS00000000082

For the **MD Anderson Cancer Center Residual Cancer Burden Calculator** see http://www3.mdanderson.org/app/medcalc/index.cfm?pagename=jsconvert3

We did immunohistochemistry profiling in the Nottingham discovery, Nottingham-HES, Nottingham-oestrogen receptor-negative-ACT, and Nottingham-NeoACT cohorts for SPAG5, Ki67, and other biological parameters (appendix p 12). Tissue microarrays (appendix pp 8–12) were used for immunohistochemistry profiling of SPAG5 in all cohorts except in Nottingham-NeoACT, for which full-face sections of core biopsies were used. In each cohort, we used the median as the cutoff between low and high protein expression.

Outcomes

The outcomes of interest were clinicopathological and biomarker associations with *SPAG5*; breast cancer-specific survival (time from diagnosis to death from breast cancer) at 10 years; disease-free survival (time from diagnosis to recurrence or distant metastasis relapse) at 10 years; distant relapse-free survival (time from diagnosis to the occurrence of distant metastasis relapse) at 10 years; pathological complete response (defined as the absence of any residual invasive carcinoma at both the primary site and in axillary lymph nodes); and residual cancer burden (as calculated with the MD Anderson Cancer Center Residual Cancer Burden Calculator). We studied different outcomes in different cohorts, depending on the data available and patient setting (figure 1).

We investigated the association between clinicopathological parameters and molecular characteristics of *SPAG5* transcript expression in the Uppsala, METABRIC, multicentre combined, and MD Anderson-NeoACT cohorts. We used the METABRIC cohort to analyse molecular and clinicopathological associations with

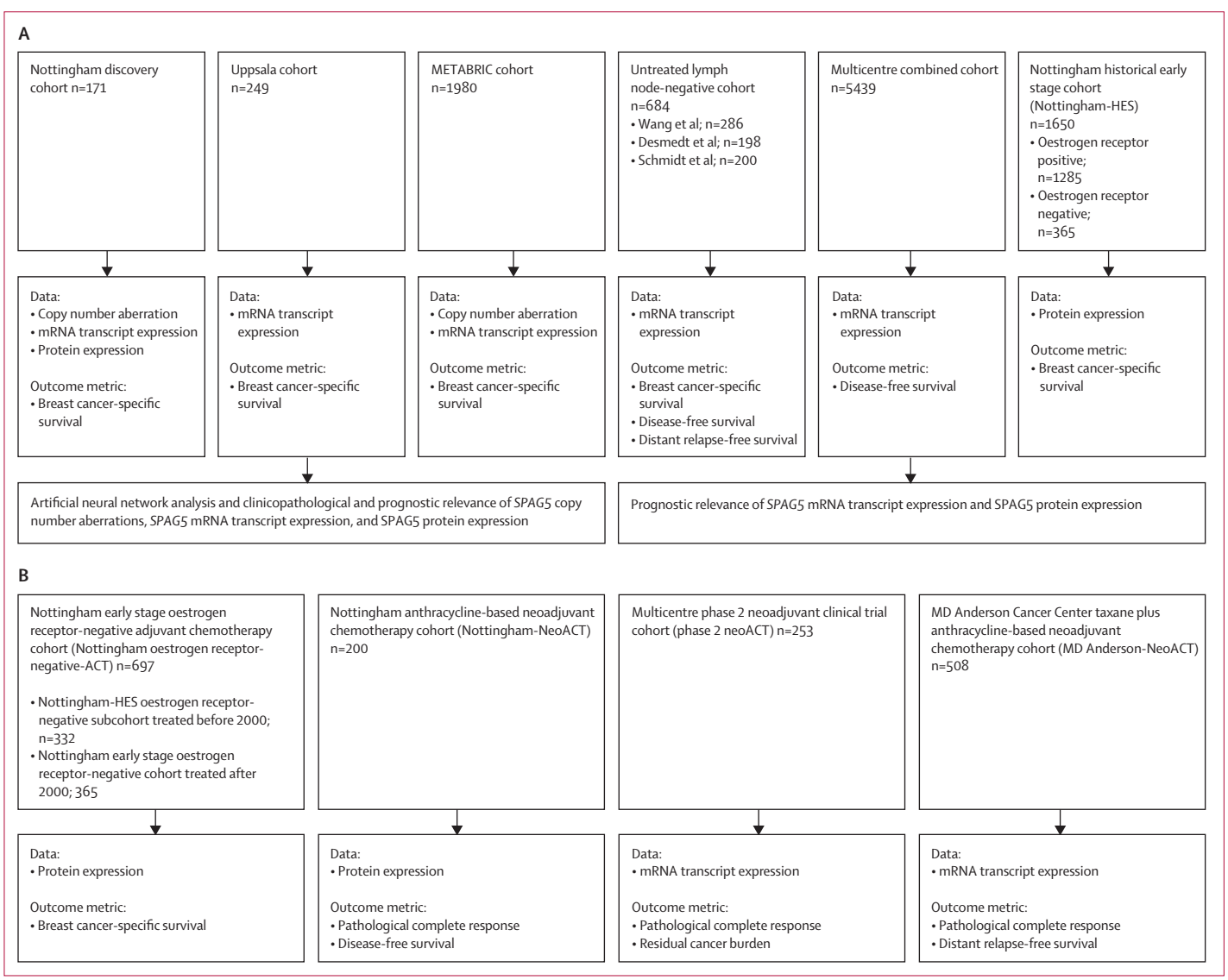

Figure 1: **Study design and patient cohorts to assess the association between *SPAG5*, clinicopathological features, and clinical outcomes (A) and the association between *SPAG5* and clinical outcome after chemotherapy (B)**

SPAG5 copy number aberrations. Associations between SPAG5 protein expression and clinicopathological parameters, as well as prognostic biomarkers and indices, were analysed in the Nottingham-HES, Nottingham-oestrogen receptor-negative-ACT, and Nottingham-NeoACT cohorts.

We explored the association of *SPAG5* transcript expression with breast cancer-specific survival in the Nottingham discovery cohort and validated our findings in the Uppsala cohort, METABRIC cohort, and the untreated lymph node-negative cohort from Desmedt and colleagues. We tested associations between breast cancer-specific survival and *SPAG5* copy number aberrations in the METABRIC cohort and SPAG5 protein expression in the Nottingham discovery, Nottingham-HES, and Nottingham-oestrogen receptor-negative-ACT cohorts.

We examined associations of *SPAG5* transcript expression with disease-free survival in the untreated lymph node-negative cohorts (Wang and Desmedt and cohorts), the multicentre combined cohort, and the Nottingham-NeoACT cohort.

We assessed the effect of *SPAG5* transcript expression on distant relapse-free survival in the untreated lymph node-negative cohorts (Schmidt and Desmedt cohorts). Furthermore, to test *SPAG5* transcript expression as a predictive biomarker for outcome after neoadjuvant combination cytotoxic chemotherapy, we investigated its association with distant relapse-free survival in the MD Anderson-NeoACT cohort.

To assess SPAG5 protein and transcript expression as a predictive biomarker for response to combination cytotoxic chemotherapy, we analysed the association with pathological complete response in the Nottingham-NeoACT, MD Anderson-NeoACT, and phase 2 NeoACT clinical trial cohorts. We analysed the association with residual cancer burden[23] in the phase 2 NeoACT clinical trial cohort. In the control cohort of 85 individuals of European ancestry, we tested for copy number aberrations and mRNA expression. The control cohort was not assessed for any clinical outcomes.

Statistical analysis

TMAA-F and GRB, who were blinded to the clinical data, did the statistical analyses with Dell-STATISTICA version 13 and SPSS version 17. We used the χ^2 test to test associations between categorical variables, and we fitted a multivariable Cox model to all time-to-event data (ie, those with survival as the endpoint). We calculated survival data for the overall study period rather than at a specific timepoint. Odds ratios (OR) and 95% CIs were calculated with SPSS software. The clinicopathological parameters and molecular characteristics included in multivariable analyses were tumour size, age at diagnosis, lymph node status, histological grade, menopause status, *TP53* mutation, PAM50 subclasses (LumA, LumB, HER2, basal-like, and normal),[24] *HER2* amplification or overexpression, progesterone and oestrogen receptor status, *Ki67* expression, integrative clusters (IntClust), *BCL2* status, and *TOP2A* status. The following molecular indices that predict response to chemotherapy were included in some multivariable analysis: residual cancer burden class RCB-0 and RCB-I,[23] the genomic chemosensitivity predictor,[18] the genomic excellent pathological response predictor,[18] the 96-gene genomic grade index,[25] the diagonal linear discrimination analysis of 30-gene signature,[26] the adjuvant-online index, the 76-gene prognostic signature,[12] clinical American Joint Committee on Cancer (AJCC) stage, Nottingham prognostic index, and PAM50 subclasses.[24] Treatment with hormone therapy and chemotherapy, and type of therapy were also included. The interaction terms between *SPAG5* expression and the status of both hormone therapy (*SPAG5**hormone therapy) and chemotherapy (*SPAG5**chemotherapy) were included in the multivariable analysis. The *SPAG5**hormone therapy interaction term was defined as the interaction between hormone therapy (yes or no) and *SPAG5* (high or low). The *SPAG5**chemotherapy interaction term was defined as the interaction between chemotherapy (yes or no) and *SPAG5* (high or low). All tests were two-sided with a 95% CI and we deemed a p value less than 0·05 as statistically significant. We applied multiple testing correction to all p values via the Bonferroni method. The range of corrections was 5–48 803 across the different analyses. We assessed the significance of any correlation between copy number aberrations and aberrant gene expression with the Jonckheere's trend test. We used Pearson correlation coefficients to assess the association between mRNA expression log intensity values and protein expression (H score) to establish whether mRNA expression was correlated with protein concentrations (appendix pp 8–9).

We did retrospective power calculations for the artificial neural network model by use of a logistic regression power model (of which artificial neural networks are an extension with a greater power) with G*Power 3.1.9 software.[27] To establish sample size, we chose an α of 0·050, a power of 0·80, an effect size leading to an odds ratio of 1·72, and two-tailed test for binary questions or classes (eg, low *vs* high expression). Based on the assumptions of the power model, the desired sample size was 88 (44 in each low and high class). We used a Monte Carlo cross-validation strategy to prevent false discovery and overfitting, and to increase the power of the algorithm used (appendix p 7). Overfitting was prevented by repeatedly testing on an unseen dataset and stopping accordingly. False discovery was further reduced in this study through parallel analysis on various questions in several datasets. Each separate analysis reduces the probability that a gene could be discovered by random chance, and yet still be a common result across multiple analyses of separate datasets. The probability of the top 30 genes commonly occurring in the top 200 out of the whole expression

For the **adjuvant-online index** see https://www.adjuvantonline.com/

array for the three cohorts used for artificial neural network analysis for a minimum of four proliferation-related factors was $1{\cdot}43\times10^{-31}$ (appendix p 7).

We did a retrospective power analysis to establish the confidence in the calculated hazard ratio (HR) and associated p value for 10-year survival (all time-to-event analyses) and to ascertain how applicable the result would be to a global population. We calculated the study power with PASS version 13 software.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

Figure 1 and the appendix (pp 1–3) show summaries of the study design and patient cohorts included in this study. Our artificial neural network analysis in the Nottingham discovery, Uppsala, and METABRIC cohorts identified the top 200 ranked genes that predict most proliferation-related features in breast cancer (appendix pp 13–27). We chose to further investigate *SPAG5* because it was among the 30 common gene probes that were predictive across most proliferation features and datasets, and because it had the most effect on other genes in the resulting interactome maps (appendix pp 25–27). Additionally, the *SPAG5* transcript has been reported to be one of a few genes associated with poor prognosis in oestrogen receptor-positive breast cancer.[28]

Gain or amplification at the *SPAG5* locus occurred in 26 (15%) of 171 patients in the Nottingham discovery cohort and 206 (10%) of 1980 patients in the METABRIC cohort. *SPAG5* gain or amplification was more common in histological high grade (grade 3) breast cancer (127 [14%] of 925 patients with high grade breast cancer) than in low or intermediate (grade 1 and 2) grade breast cancer (66 [7%] of 943 patients with grade 1 or 2) and in PAM50-HER2 (46 [19%] of 237 PAM50-HER2 cases in METABRIC) and PAM50-LumB (87 [18%] of 488 PAM50-LumB cases in METABRIC) breast cancer than in other subtypes. *SPAG5* copy number aberrations and *SPAG5* transcript expression were strongly correlated (Nottingham discovery: Spearman correlation r=0·81, Bonferroni-adjusted p=0·010; METABRIC: Spearman correlation r=0·87; Bonferroni-adjusted p<0·0001). Patients with both oestrogen receptor-negative and oestrogen receptor-positive breast cancer had higher *SPAG5* transcript expression than did healthy individuals in our cohort of 85 healthy individuals of European ancestry (METABRIC oestrogen receptor-positive: r=0·19, Bonferroni-adjusted p<0·0001; METABRIC oestrogen receptor-negative: r=0·37, Bonferroni-adjusted p<0·0001). However, in the METABRIC cohort, the amount of *SPAG5* transcript was higher in oestrogen receptor-negative disease than in oestrogen receptor-positive disease (r=0·18; Bonferroni-adjusted p<0·0001). Furthermore, in the METABRIC cohort, the PAM50-LumB, PAM50-Basal, and PAM50-HER2 breast cancer subclasses had more *SPAG5* transcript than did PAM50-normal-like, PAM50-LumA disease, and normal tissue, (all Bonferroni-adjusted p<0·0001 in both oestrogen receptor-positive and oestrogen receptor-negative cancers; appendix p 28).

As a continuous and categorical variable, compared with low *SPAG5* transcript expression, high *SPAG5* transcript expression was associated with high-grade (grade 3), *TP53* mutation, and *HER2* gain or amplification (METABRIC cohort; appendix pp 29–32). In the METABRIC cohort, ten novel prognostic biological subgroups have been previously identified by the joint clustering of copy number aberration and gene expression data (integrative clusters [IntClust]).[10] In our study, *SPAG5* gain or amplification was associated with IntClust 1, 5, and 6 (all p<0·0001), whereas high *SPAG5* transcript expression was associated with IntClust 1, 5, 9, and 10 (appendix pp 29–32). Furthermore, high *SPAG5* transcript expression was associated with values for other molecular parameters and indices that predict higher probability of response to neoadjuvant chemotherapy: residual cancer burden class RCB-0 and RCB-I,[23] genomic chemosensitivity predictor,[18] genomic excellent pathological response predictor,[18] 96-gene genomic grade index,[25] diagonal linear discrimination analysis of 30-gene signature,[26] and PAM50-HER2 and PAM50-Basal[24] (all p<0·0001; appendix p 33).

Additionally, in the Nottingham discovery cohort *SPAG5* transcript expression and SPAG5 protein expression were strongly correlated (Pearson correlation r=0·75, Bonferroni-adjusted p=0·001). In the Nottingham-HES cohort, 272 (20%) of 1368 patients with immunohistochemistry data available showed high SPAG5 protein expression (H score ≥10), which was associated with aggressive phenotypes including HER2 overexpression (p=0·030), an absence of hormone receptor expression, and *TP53* mutation (appendix pp 34–37). In the Nottingham oestrogen receptor-negative-ACT cohort, high SPAG5 protein expression occurred in 355 (51%) of 697 patients and was associated with lymphovascular invasion, high grade, and high Ki67 expression (all p<0·0001; appendix pp 38–41). In the Nottingham-NeoACT cohort, high SPAG5 protein expression was detected in 50 (25%) of 200 prechemotherapy core biopsies and was associated with high grade, oestrogen receptor-negative and HER2-negative phenotypes, and *TP53* mutation (all Bonferonni-adjusted p<0·0001). Neither *SPAG5* transcript nor protein expression was associated with lymph node stage or disease clinical stage in any of the cohorts tested (appendix pp 29–41).

In the METABRIC cohort, *SPAG5* gain or amplification was associated with shorter overall breast cancer-specific survival than that for normal *SPAG5* copy number or

SPAG5 loss in all patients and in the oestrogen receptor-positive subgroup, but not in the oestrogen receptor-negative subgroup (figure 2A–C). As a continuous variable, high *SPAG5* transcript expression was associated with shorter overall breast cancer-specific survival than was low *SPAG5* transcript expression (METABRIC cohort: HR 1·89, 1·55–2·31, p<0·0001; Nottingham discovery cohort: HR 1·50, 95% CI 0·98–2·32, p=0·065; Uppsala cohort: 1·99, 1·44–2·76, p<0·0001). Furthermore, as a categorical variable, high *SPAG5* transcript expression was associated with shorter overall breast cancer-specific survival than was low *SPAG5* transcript expression in the METABRIC cohort (figure 2D) and the Uppsala cohort (HR 1·98, 95% CI 29–3·04; p=0·0020; appendix p 42). In the METABRIC cohort, high *SPAG5* transcript expression was associated with a shorter overall breast cancer-specific survival than was low *SPAG5* transcript in the oestrogen

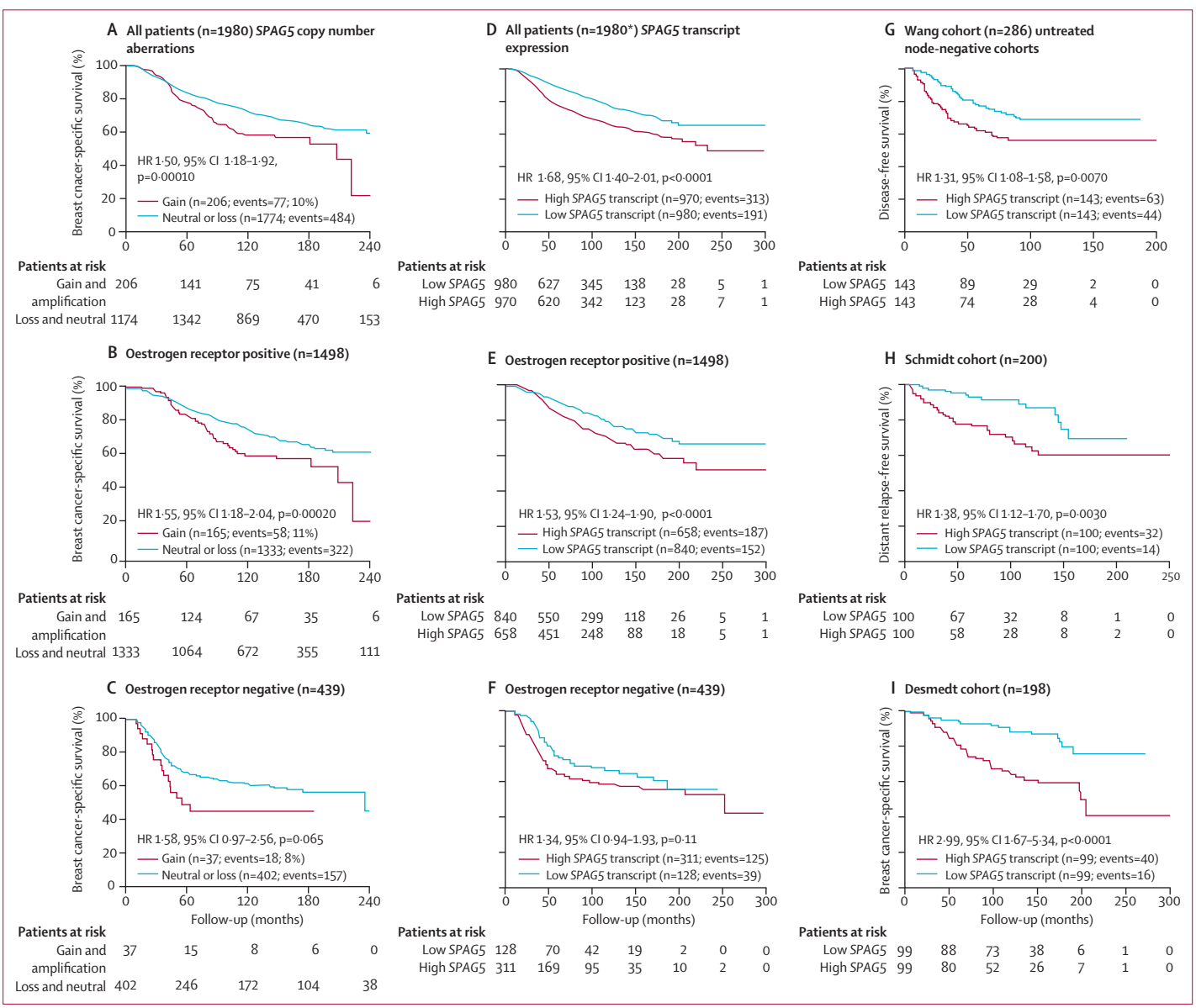

Figure 2: **Clinical outcome of *SPAG5* gain or amplification and transcript expression in the METABRIC and untreated lymph node-negative cohorts**
Kaplan-Meier curves for breast cancer-specific survival in the METABRIC cohort stratified by *SPAG5* gain or amplification in all patients (A), in the oestrogen receptor-positive patients (B), and oestrogen receptor-negative patients (C); and stratified by *SPAG5* transcript expression in all patients (D), in the oestrogen receptor-positive patients (E), and oestrogen receptor-negative patients (F) in the METABRIC cohort (43 patients in METABRIC had unknown oestrogen receptor status). Association between *SPAG5* transcript expression and clinical outcome in untreated lymph node negative cohorts (G–F). Kaplan-Meier curves stratified by *SPAG5* transcript expression for disease-free survival in the Wang cohort[12] (G), distant relapse-free survival in the Schmidt cohort[14] (H), and breast cancer-specific survival in the Desmedt cohort[13] (I). METABRIC=Molecular Taxonomy of Breast Cancer International Consortium. HR=hazard ratio. *Only 1950 patients in METABRIC had available gene expression data.

receptor-positive subgroup, but not in the oestrogen receptor-negative subgroup (figure 2E, F). Additionally, in low-risk breast cancer (n=658; Nottingham prognostic index <3·4), lymph node-negative breast cancer (n=914), and lymph node-positive breast cancer (n=936) samples in the METABRIC cohort, high *SPAG5* transcript expression was associated with a shorter overall breast cancer-specific survival than was low *SPAG5* transcript expression (appendix p 42).

The Uppsala cohort, with 249 cases (including 124 with high *SPAG5* transcript expression), has a retrospective power of 83% to detect an HR of 1·98, when 10-year breast cancer-specific survival is 53% for high *SPAG5* and 71% for low *SPAG5* transcript expression, with a p value less than 0·05. Similarly for the transcript expression analysis in the METABRIC cohort, a power model using a two-sided log-rank test with an overall sample size of 1950 cases with gene expression data available (including 970 with high *SPAG5* transcript expression) achieved a retrospective power of at least 99·9% to detect an HR of 1·68, when breast cancer-specific survival at 10 years was 78% for high *SPAG5* and 66% for low *SPAG5*, with a p value less than 0·0001.

In the untreated lymph node-negative cohorts, high *SPAG5* transcript expression was associated with shorter disease-free survival, distant relapse-free survival, and overall breast cancer-specific survival than was low *SPAG5* transcript expression (figure 2G–I). In the untreated lymph node-negative breast cancer cohorts comparing high to low *SPAG5* transcript expression, the retrospective power was 82% to detect an HR of 1·3 for 10-year disease-free survival in the Wang cohort, 84% to detect an HR of 1·4 for 10-year distant relapse-free survival in the Schmidt cohort, and 98% to detect an HR of 1·99 for 10-year breast cancer-specific survival in the Desmedt cohort, with p values less than 0·050 for all.

In the Uppsala cohort, multivariable Cox regression analysis showed that high *SPAG5* transcript expression (*vs* low transcript expression) and positive lymph node status (*vs* negative) were independently associated with shorter breast cancer-specific survival (table 1). Similarly, in the METABRIC cohort, a multivariable Cox regression model showed that high *SPAG5* transcript expression was independently associated with shorter breast cancer-specific survival than was low *SPAG5* transcript expression (table 1). Lymph node status, histological grade, tumour size, age at diagnosis, HER2 status, and progesterone receptor status were also independently associated with breast cancer-specific survival in the METABRIC cohort (table 1). Furthermore, separate multivariable Cox regression models in the METABRIC cohort showed that high *SPAG5* transcript expression was associated with clinical outcome independently of both PAM50 and IntClust prognostic subclasses, although in these models PAM50 subclasses, hormone therapy, chemotherapy, and IntClust were also independently associated with clinical outcome (table 1).

In the untreated lymph node-negative Desmedt cohort, high *SPAG5* transcript expression was associated with shorter breast cancer-specific survival than was low transcript expression after adjustment for oestrogen receptor status and three prognostic signatures and indices (table 1).

In the multicentre combined cohort, high *SPAG5* transcript expression was associated with shorter overall disease-free survival than was low *SPAG5* transcript expression in all patients and in the lymph node-negative, lymph node-positive, and oestrogen receptor-positive breast cancer subgroups, but not in the oestrogen receptor-negative subgroup (appendix p 43). In a retrospective power calculation, the multicentre combined cohort (5439 patients including 2711 with high *SPAG5* transcript expression) had at least 99·9% power to detect an HR of 1·32 for disease-free survival, with a p value less than 0·0010. In the multicentre combined cohort, multivariable Cox regression models showed that high *SPAG5* transcript expression is an independent factor for shorter disease-free survival after controlling for the Nottingham prognostic index (*vs* low *SPAG5* transcript expression: HR 1·19, 95% CI 1·09–1·30, p=0·00020), the adjuvant-online index (1·18, 1·03–1·35, p=0·017), and the 72-proliferation gene signature[29] (1·18, 1·10–1·27, p<0·0001). Univariate analysis showed that high *Ki67* transcript expression was also associated with a shorter disease-free survival than was low *Ki67* expression in this cohort (HR 1·21, 95% CI 1·15–1·27, p<0·0001). However, multivariable Cox regression models showed that *Ki67* transcript expression was not an independent prognostic factor for disease-free survival after controlling for the Nottingham prognostic index (1·09, 1·00–1·20, p=0·060) and the adjuvant-online index (0·93, 0·83–1·05, p=0·26).

High SPAG5 protein expression was associated with shorter overall breast cancer-specific survival than was low SPAG5 protein expression (Nottingham discovery: HR 1·06, 95% CI 1·02–1·09, p=0·0010; Nottingham-HES: 1·68, 1·32–2·12, p<0·0001; figure 3A). High SPAG5 protein was also associated with shorter overall breast cancer-specific survival than was low SPAG5 protein expression in the oestrogen receptor-positive subgroup, but not in the oestrogen receptor-negative subgroup of the Nottingham-HES cohort (figure 3B, 3C). In the low-risk, lymph node-negative, and lymph node-positive subgroups in the Nottingham-HES cohort, high SPAG5 protein expression was associated with shorter overall breast cancer-specific survival than was low SPAG5 expression (appendix p 44). The Nottingham discovery cohort (128 patients including 24 with high SPAG5 protein) had a retrospective power of 80% to detect an HR of 1·10 with p value less than 0·05, when breast cancer-specific survival in the high SPAG5 subgroup at 10 years is 60%. The Nottingham-HES cohort had 99% power to detect an HR of 1·68 and a 10-year breast cancer-specific survival of 63% in all patients with a p value less than 0·05 for a sample size of

	Hazard ratio (95% CI)	p value
Breast cancer-specific survival (Uppsala cohort; *SPAG5* transcript; n=249)		
SPAG5 mRNA (high *vs* low)	1·62 (1·03–2·53)	0·036
MKI67 mRNA (high *vs* low)	0·991 (0·486–1·71)	0·77
Lymph node status (positive *vs* negative)	1·61 (1·01–2·57)	0·050
96-gene genomic grade index[26]	..	0·34
G1	1	..
G2a	0·94 (0·50–1·79)	..
G2b	1·77 (0·82–3·96)	..
G3	1·73 (0·76–3·97)	..
Age at diagnosis*	1·01 (0·99–1·03)	0·16
Tumour size (continuous; mm)	1·09 (0·95–1·24)	0·21
Oestrogen receptor (positive *vs* negative)	1·43 (0·76–2·71)	0·27
TP53 mutation (mutant *vs* wild-type)	1·07 (0·62–1·86)	0·80
Breast cancer-specific survival (METABRIC cohort; *SPAG5* transcript; n=1980; model 1)		
SPAG5 mRNA (high *vs* low)	1·27 (1·02–1·58)	0·034
Lymph node stage		<0·0001
Negative	1·00	..
1–3 positive lymph nodes	1·68 (1·31–2·16)	..
>3 positive lymph nodes	3·42 (2·59–4·52)	..
Histological grade	..	0·017
Low	1·00	..
Intermediate	1·79 (1·08–2·95)	..
High	2·05 (1·23–3·39)	..
Tumour size (continuous; mm)	1·01 (1·007–1·015)	<0·0001
Age at diagnosis*	1·01 (1·002–1·02)	0·015
HER2 (overexpression *vs* normal expression)	1·50 (1·18–1·91)	0·0010
Progesterone receptor (positive *vs* negative)	0·77 (0·62–0·96)	0·020
Oestrogen receptor (positive *vs* negative)	1·06 (0·78–1·45)	0·70
Hormone therapy (yes *vs* no)	1·23 (0·82–1·02)	0·12
Chemotherapy (yes *vs* no)	1·31 (0·96–1·78)	0·090
Hormone therapy**SPAG5*	0·62 (0·41–0·93)	0·021
Chemotherapy**SPAG5*	0·84 (0·55–1·28)	0·42
Breast cancer-specific survival (METABRIC cohort; *SPAG5* transcript; n=1980; model 2)		
SPAG5 mRNA (high *vs* low)	1·31 (1·04–1·65)	0·020
PAM-50 molecular subclasses[28]	..	<0·0001
PAM50-LumA	1	..
PAM50-LumB	2·13 (1·62–2·80)	..
PAM50-HER2	2·34 (1·72–3·18)	..
PAM50-basal-like	1·89 (1·38–2·59)	..
PAM50-normal	1·45 (1·01–2·08)	..
Hormone therapy (yes *vs* no)	1·31 (1·06–1·60)	0·010
Chemotherapy (yes *vs* no)	1·31 (1·66–2·59)	<0·0001
Hormone therapy**SPAG5*	0·57 (0·38–0·84)	0·0050
Chemotherapy**SPAG5*	1·18 (0·78–1·78)	0·43

(Table 1 continues in next column)

(Continued from previous column)

	Hazard ratio (95% CI)	p value
Breast cancer-specific survival (METABRIC cohort; *SPAG5* transcript; n=1980; model 3)		
SPAG5 mRNA (high *vs* low)	1·33 (1·06–1·67)	0·014
Integrated Clusters (IntClust)[11]	..	<0·0001
IntClust 1	1	..
IntClust 2	1·47 (0·92–2·34)	..
IntClust 3	0·38 (0·24–0·61)	..
IntClust 4	0·69 (0·46–1·03)	..
IntClust 5	1·58 (1·09–2·30)	..
IntClust 6	1·13 (0·70–1·81)	..
IntClust 7	0·58 (0·37–0·93)	..
IntClust 8	0·65 (0·44–0·97)	..
IntClust 9	1·08 (0·72–1·63)	..
IntClust 10	0·75 (0·50–1·13)	..
Hormone therapy (yes *vs* no)	1·23 (1·00–1·50)	0·047
Chemotherapy (yes *vs* no)	2·02 (1·62–2·51)	<0·0001
Hormone therapy**SPAG5*	0·53 (0·36–0·77)	0·020
Chemotherapy**SPAG5*	1·18 (0·78–1·78)	0·66
Breast cancer-specific survival (untreated lymph node-negative patients; Desmedt cohort; *SPAG5* transcript; n=198)		
SPAG5 mRNA (high *vs* low)	2·34 (1·24–4·42)	0·0090
Oestrogen receptor (positive *vs* negative)	0·67 (0·38–1·22)	0·19
Nottingham prognostic index (continuous)	1·74 (0·712–4·23)	0·22
Adjuvant-online index (continuous)	0·76 (0·30–1·94)	0·56
76-gene prognostic signature[13] (continuous)	1·52 (0·75–3·06)	0·24
Breast cancer-specific survival at 20 years follow-up (Nottingham-HES; SPAG5 protein; n=1650)		
SPAG5 protein expression (positive *vs* negative)	1·73 (1·23–2·46)	0·0020
Tumour size (continuous; mm)	1·18 (1·07–1·31)	0·0010
Lymph node status	..	<0·0001
Negative	1	..
Positive	1·95 (1·51–2·52)	..
Histological grade	..	0·00020
Low or intermediate	1	..
High	1·83 (1·33–2·50)	
Oestrogen receptor (positive *vs* negative)	1·20 (0·82–1·74)	0·350
HER2 overexpression (positive *vs* negative)	1·60 (1·16–2·52)	0·0040
Progesterone receptor status (positive *vs* negative)	0·66 (0·47–0·92)	0·015
Ki67 (positive *vs* negative)	1·44 (1·03–2·01)	0·034
Chemotherapy status (CMF; yes *vs* no)	1·55 (1·13–2·17)	0·010
Hormone therapy (yes *vs* no)	1·31 (0·99–1·73)	0·059
Chemotherapy*SPAG5	1·65 (0·85–3·23)	0·14
Hormone therapy *SPAG5	1·95 (1·14–3·35)	0·015

(Table 1 continues in next column)

	Hazard ratio (95% CI)	p value
(Continued from previous column)		
Breast cancer-specific survival at 10 years follow-up (Nottingham oestrogen receptor-negative-ACT cohort; SPAG5 protein expression; n=697; without interaction terms)		
SPAG5 protein expression (positive *vs* negative)	0·68 (0·50–0·92)	0·013
Tumour size (continuous; mm)	1·06 (1·02–1·09)	0·0010
Lymph node status	..	<0·0001
Negative	1	..
Positive	2·60 (1·92–3·50)	..
Histological grade	..	0·059
Low or intermediate	1	..
High	1·67 (0·98–2·86)	..
Menopausal status (postmenopausal *vs* premenopausal)	1·34 (0·99–1·82)	0·060
HER2 overexpression (positive *vs* negative)	0·92 (0·64–1·31)	0·64
BCL2 expression (positive *vs* negative)	0·60 (0·40–0·90)	0·013
Chemotherapy status		
No chemotherapy	1	..
CMF	0·80 (0·54–1·18)	0·260
Anthracycline	0·61 (0·42–0·89)	0·010
Breast cancer-specific survival at 10 years follow-up (Nottingham oestrogen receptor-negative-ACT cohort; SPAG5 protein; n=697; with interaction terms)		
SPAG5 protein expression (positive *vs* negative)	0·48 (0·30–0·76)	0·0020
Tumour size (continuous; mm)	1·05 (1·02–1·09)	0·0030
Lymph node status		<0·0001
Negative	1	..
Positive	2·57 (1·90–3·46)	..
Histological grade		0·066
Low or intermediate	1	..
High	1·65 (0·97–2·82)	..
Menopausal status (postmenopausal *vs* premenopausal)	1·35 (0·99–1·84)	0·056
HER2 overexpression (positive *vs* negative)	0·93 (0·65–1·34)	0·70
BCL2 expression (positive *vs* negative)	0·63 (0·42–0·94)	0·023
Chemotherapy status		
No chemotherapy	1	..
CMF	0·79 (0·54–1·16)	0·23
Anthracycline	0·59 (0·40–0·87)	0·008
SPAG5*CMF	0·70 (0·32–1·50)	0·36
SPAG5*Anthracycline	0·43 (0·20–0·93)	0·032
(Table 1 continues in next column)		

	Hazard ratio (95% CI)	p value
(Continued from previous column)		
Distant relapse-free survival (MD Anderson-NeoACT cohort; SPAG5 transcript; n=508)		
SPAG5 transcript expression (positive *vs* negative)	0·68 (0·48–0·97)	0·031
Chemosensitivity prediction signature[19] (low *vs* high)	0·49 (0·36–0·67)	<0·0001
96-gene genomic grade index[26] (grade 1 or grade 2a *vs* grade 2b or grade 3-like)	0·65 (0·32–1·29)	0·21
30-gene diagonal linear discrimination analysis[27] (low *vs* high)	2·09 (0·95–4·56)	0·065
PAM-50 molecular subclasses[28]	..	0·042
PAM50-LumA *vs* others	0·16 (0·04–0·58)	0·0060
PAM50-LumB *vs* others	0·24 (0·07–0·88)	0·031
PAM50-HER2 *vs* others	0·14 (0·03–0·66)	0·013
PAM50-basal-like *vs* others	0·28 (0·07–1·10)	0·068
Clinical AJCC stage (I or II *vs* III)	2·03 (1·38–2·99)	0·00040
MKI67 transcript (positive *vs* negative)	1·22 (0·71–2·07)	0·47

METABRIC=Molecular Taxonomy of Breast Cancer International Consortium. Hormone therapy*SPAG5=interaction term between hormone therapy status (yes or no) and *SPAG5* expression (high or low). Chemotherapy*SPAG5=interaction term between chemotherapy status (yes or no) and *SPAG5* expression (high or low). PAM50=prediction analysis of microarray 50-gene signature. IntClust=integrative clusters. Nottingham-HES=Nottingham historical early stage breast cancer. CMF=cyclophosphamide, methotrexate, and fluorouracil chemotherapy. Nottingham oestrogen receptor-negative-ACT=Nottingham early stage oestrogen receptor-negative breast cancer adjuvant chemotherapy. SPAG5*CMF=interaction term between CMF chemotherapy status (yes or no) and SPAG5 expression (high or low). SPAG5*anthracycline=interaction term between anthracycline-based chemotherapy status (yes or no) and SPAG5 expression (high or low). MD Anderson-NeoACT=MD Anderson taxane plus anthracycline-based neoadjuvant chemotherapy. AJCC=American Joint Committee on Cancer. *Age was a continuous value with increments of 5 years.

Table 1: **Multivariable Cox regression model analyses by cohort**

1340 patients with both immunohistochemistry and survival data available with 273 high SPAG5 protein samples. Multivariable Cox regression analysis of the Nottingham-HES cohort showed that high SPAG5 protein expression was independently associated with a shorter breast cancer-specific survival at 10 years compared with low SPAG5 expression after adjustment for adjuvant hormone therapy and chemotherapy, grade, size, lymph node stage, HER2, oestrogen receptor, progesterone receptor, age, Ki67, and interaction terms (SPAG5*chemotherapy and SPAG5*hormone therapy; table 1). Tumour size, lymph node status, histological grade, oestrogen receptor and progesterone receptor status, Ki67 status, HER2 status, and chemotherapy status were also independently associated with breast cancer-specific survival (table 1).

In the Nottingham oestrogen receptor-negative-ACT cohort, high SPAG5 protein expression was associated with longer overall breast cancer-specific survival than was low SPAG5 protein expression (HR 0·85, 95% CI 0·78–0·94, p=0·0010; figure 3D). However, a subgroup analysis of adjuvant chemotherapy-naive patients showed that overall breast cancer-specific survival was similar between patients with high and low SPAG5 protein expression (0·90, 0·63–1·27, p=0·54; appendix p 45). In the subgroup who received adjuvant chemotherapy, high SPAG5 protein was associated with longer overall breast cancer-specific survival than was low SPAG5

protein expression (0·41, 0·26–0·64, p<0·0001; appendix p 45). In oestrogen receptor-negative breast cancer with high SPAG5 protein expression, treatment with anthracycline-based adjuvant chemotherapy increased overall breast cancer-specific-survival compared with no chemotherapy (figure 3E). In oestrogen receptor-negative, low SPAG5 protein expression tumours, anthracycline-based adjuvant chemotherapy had no effect on overall breast cancer-specific survival (figure 3F). Multivariable Cox regression models in this cohort showed that SPAG5 protein expression was a predictive marker for breast cancer-specific survival and that the interaction term between SPAG5 protein expression and anthracycline-based adjuvant chemotherapy was a significant predictor for breast cancer-specific survival (table 1). These models also showed that tumour size, lymph node status, *BCL2* status, and anthracycline treatment were independently associated with breast cancer-specific survival (table 1).

In the MD Anderson-NeoACT cohort, after receiving combination cytotoxic chemotherapy, overall distant relapse-free survival was similar in patients with high *SPAG5* and low *SPAG5* transcript expression (HR: 1·3, 95% CI 0·92–1·95, p=0·12; appendix p 43). Multivariable Cox regression analysis showed that high *SPAG5* transcript expression was independently associated with longer disease-free survival than was low SPAG5 expression in these patients who all received neoadjuvant chemotherapy (0·68, 0·48–0·97, p=0·031, table1).

To validate our previous observation of an association between *SPAG5* expression and the clinical outcome of patients with breast cancer after receiving chemotherapy,

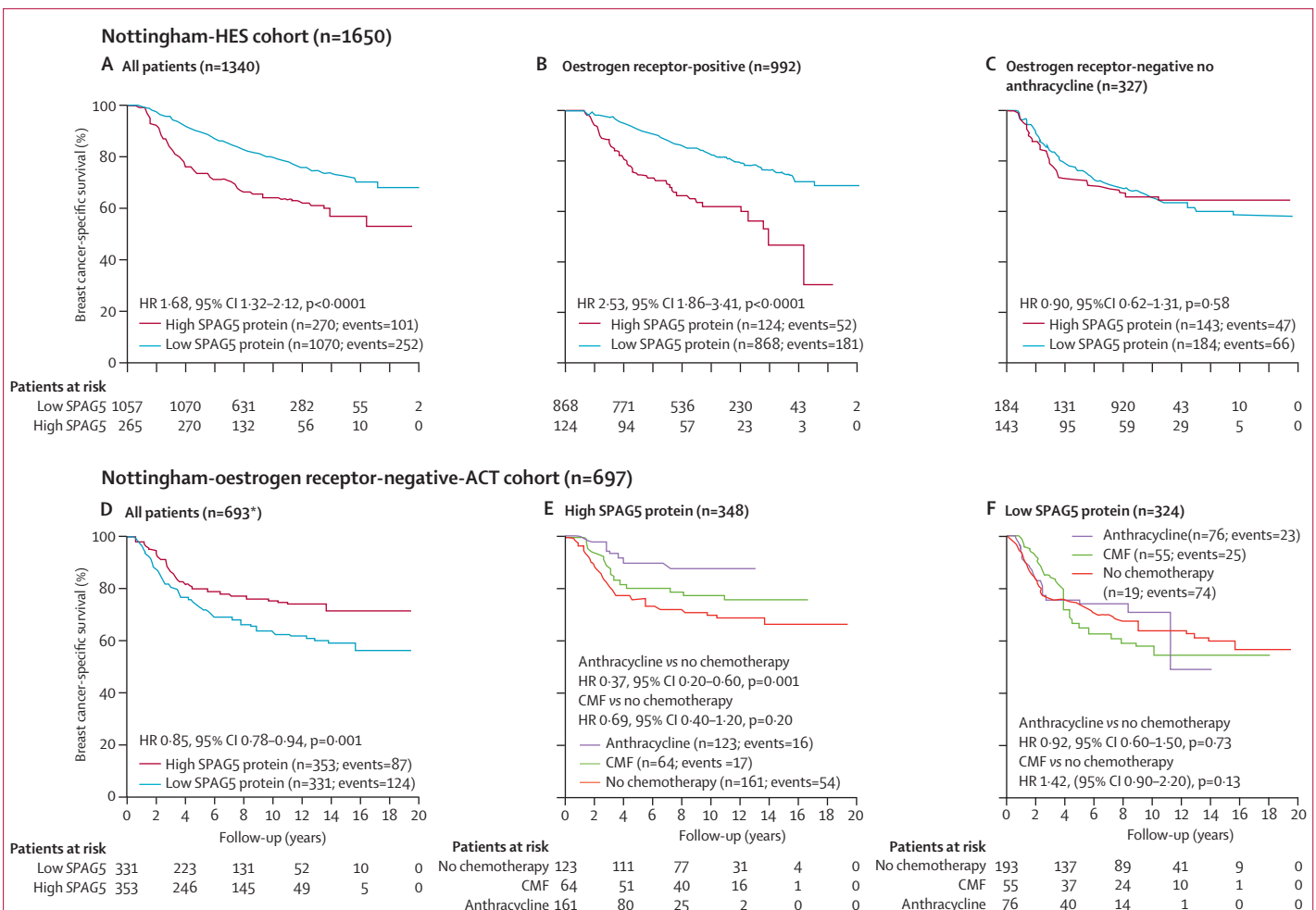

***Figure 3*: Clinical outcome and SPAG5 protein expression in the Nottingham-HES and Nottingham-oestrogen receptor-negative-ACT cohorts**

Both immunohistochemistry and survival data were available for 1340 patients. Kaplan-Meier curves for breast cancer-specific survival in the Nottingham-HES cohort stratified by *SPAG5* protein expression in all patients (A), in the oestrogen receptor-positive patients (B), and oestrogen receptor-negative patients (C). Kaplan-Meier curves for breast cancer-specific survival in the Nottingham-oestrogen receptor-negative-ACT cohort stratified by SPAG5 protein expression in all patients (D) and stratified by SPAG5 protein expression and adjuvant chemotherapy treatment status (E, F). CMF=cyclophosphamide, methotrexate, and fluorouracil chemotherapy. HR=hazard ratio. Nottingham-HES=Nottingham historical early stage breast cancer cohort. Nottingham-oestrogen receptor-negative-ACT cohort=Nottingham early stage oestrogen receptor-negative adjuvant chemotherapy cohort. *672 patients had treatment and survival data available.

	Odds ratio (95% CI)	p value
Pathological complete response (MD Anderson-NeoACT cohort; n=508)		
SPAG5 transcript (high *vs* low)	1·71 (1·07–2·74)	0·024
Pathological complete response prediction signature (low *vs* high)	1·17 (0·44–3·10)	0·75
96-gene genomic grade index[26] (grade 1 or grade 2a *vs* grade 2b or grade 3-like)	0·26 (0·09–0·78)	0·016
30-gene diagonal linear discrimination analysis[27] (low *vs* high)	1·17 (0·44–3·10)	0·75
PAM50 molecular subclasses[28]		0·042
PAM50-LumA *vs* others	0·16 (0·04–0·58)	0·0060
PAM50-LumB *vs* others	0·24 (0·07–0·88)	0·031
PAM50-HER2 *vs* others	0·14 (0·03–0·66)	0·013
PAM50-basal-like *vs* others	0·28 (0·07–1·10)	0·068
Clinical AJCC stage (I or II *vs* III)	0·31 (0·76–0·45)	0·012
Histological grade (G1 or G2 *vs* G3)	2·37 (1·15–4·89)	0·020
Age*	0·99 (0·96–1·01)	0·26
Oestrogen receptor status (positive *vs* negative)	0·46 (0·21–1·04)	0·063
Progesterone receptor (positive *vs* negative)	1·09 (0·86–1·39)	0·47
Residual cancer burden (phase 2 Neo-ACT cohort; n=253)		
SPAG5 transcript (high *vs* low)	1·80 (1·02–3·02)	0·044
Oestrogen receptor status (positive *vs* negative)	0·59 (0·25–1·36)	0·21
Progesterone receptor (positive *vs* negative)	0·41 (0·02–1·02)	0·042
HER2 overexpression (high *vs* low)	0·96 (0·36–2·62)	0·94
Age (≥50 *vs* <50 years)	0·40 (0·22–0·73)	0·0030
Tumour size (≥5 *vs* <5 cm)	0·59 (0·32–1·09)	0·090
Pathological complete response (Nottingham-NeoACT cohort; n=200)		
SPAG5 protein expression (high *vs* low)	8·75 (2·42–31·62)	0·0010
Ki67 protein expression (high *vs* low)	2·8 (0·77–10·24)	0·11
BCL2 protein expression (positive *vs* negative)	0·19 (0·05–0·69)	0·010
TOP2A protein expression (overexpression *vs* normal)	3·81 (0·98–14·73)	0·053
Oestrogen receptor protein expression (positive *vs* negative)	0·77 (0·42–2·84)	0·25
HER2 expression (overexpression *vs* normal)	0·84 (0·23–3·12)	0·79
Taxane (yes *vs* no)	0·67 (0·21–2·21)	0·52
Age*	1·04 (0·98–1·10)	0·25
AJCC stage (I or II *vs* III or IV)	0·35 (0·11–1·52)	0·084
Histological grade (G1 or G2 *vs* G3)	0·42 (0·11–1·54)	0·18

MD Anderson-NeoACT=MD Anderson taxane plus anthracycline-based neoadjuvant chemotherapy. Phase 2 Neo-ACT=multicentre phase 2 neoadjuvant chemotherapy clinical trial. Nottingham-NeoACT=Nottingham anthracycline-based neoadjuvant chemotherapy. PAM50=prediction analysis of microarray 50-gene signature. AJCC=American Joint Committee on Cancer. *Age was a continuous value with increments of 1 year.

Table 2: **Multivariable logistic regression model analysis for treatment outcome after neoadjuvant chemotherapy**

we investigated the association between *SPAG5* transcript expression and tumour response to combination cytotoxic chemotherapy in the MD Anderson-NeoACT cohort, in which 488 of 508 patients had available pathological complete response data. Of these patients, 99 (20%) achieved a pathological complete response. In patients who did not achieve a pathological complete response, high *SPAG5* transcript expression was associated with shorter distant relapse-free survival than was low *SPAG5* transcript expression (HR 1·74, 95% CI 1·17–2·52, p=0·0070; appendix p 46). In patients who achieved a pathological complete response, the association of high and low *SPAG5* transcript expression with distant relapse-free survival was similar (HR 1·80, 0·63–5·14; p=0·270). In the overall MD Anderson-NeoACT patient population, as a continuous variable, high *SPAG5* transcript expression was associated with an increased proportion of patients achieving pathological complete response compared with low *SPAG5* transcript expression (OR 2·6, 95% CI 1·8–3·9, p<0·0001). As a categorical variable, high *SPAG5* transcript expression was associated with an increased proportion of patients achieving a pathological complete response (70 [28%] of 246 patients) compared with low *SPAG5* transcript expression (29 [12%] of 242 patients; OR 2·90, 95% CI 1·80–4·70, p<0·0001). Multivariable logistic regression analysis showed that high *SPAG5* transcript expression was an independent predictor for more patients achieving a pathological complete response (table 2). The 96-gene genomic grade index, PAM50 subclass, AJCC stage, and histological grade were also independent predictors for pathological complete response (table 2).

We validated our results in the phase 2 NeoACT clinical trial cohort,[19] in which 69 (27%) of 253 patients achieved a pathological complete response and 86 (34%) patients achieved a residual cancer burden of RCB-0 or RCB-I. As a continuous variable, high *SPAG5* transcript expression was associated with more patients achieving pathological complete response (OR 1·33, 95% CI 0·98–1·79, p=0·065) and RCB-0 or RCB-I (1·29, 0·98–1·71, p=0·075). High *SPAG5* transcript expression was also associated with a higher proportion of patients achieving pathological complete response (1·99, 1·13–3·45, p=0·016) and RCB-0 or RCB-I (1·97, 1·16–3·34, p=0·010) compared with low *SPAG5* transcript. In a multivariable logistic regression model high *SPAG5* transcript expression was significantly associated with patients achieving RCB-0 or RCB-I (table 2).

In the Nottingham-NeoACT cohort, 29 (15%) of 200 patients achieved pathological complete response and 20 (40%) of 50 patients with high SPAG5 protein expression achieved pathological complete response compared with nine (6%) of 150 patients with low SPAG5 protein expression (OR 10·8, 95% CI 4·5–26·29, p<0·0001; appendix p 43). Furthermore, 18 (37%) of 49 patients with high SPAG5 protein expression at baseline in the pre-chemotherapy tumour biopsy became negative for SPAG5 protein expression after receiving anthracycline-based neoadjuvant chemotherapy (McNemar test p=0·0040 for a comparison of SPAG5 expression in the paired matched pre-chemotherapy biopsy and the residual tumours after chemotherapy). Multivariable logistic regression analysis showed that high SPAG5 protein expression was an independent predictor for pathological complete response, whereas Ki67 was not (table 2).

Patients in the Nottingham-NeoACT cohort with high SPAG5 protein expression prior to chemotherapy who received anthracycline-based neoadjuvant chemotherapy had similar 5-year disease-free survival after surgery compared with patients who had low SPAG5 protein expression (HR 1·1, 95% CI 0·90–1·30, p=0·40;

appendix p 43). Importantly, at the 5-year follow-up, patients with high SPAG5 protein expression in the residual tumour specimen after they had received anthracycline-based neoadjuvant chemotherapy had shorter disease-free survival compared with those who had residual tumours with low SPAG5 protein expression (3·5, 1·8–7·0, p=0·00030; appendix p 46).

In agreement with the results from data mining of the Oncomine microarray database,[28] we noted that breast cancer, similar to most human cancers, has increased *SPAG5* transcript expression compared with normal tissue (appendix pp 47–53). In turn, high *SPAG5* expression is associated with poor clinical outcome (appendix pp 54–56), especially in oestrogen receptor-positive breast cancer.

Discussion

Our findings show that amplification or gain of the *SPAG5* locus at Ch17q11.2 occurred in 10–19% of breast cancers; *SPAG5* gene copy number aberration and expression of its transcript and protein are associated with poor clinical outcome and adverse clinicopathological features, including *TP53* mutation, PAM50-LumB phenotype, and PAM50-HER2 phenotype; and both *SPAG5* transcript and SPAG5 protein expression are independent predictors for response to chemotherapy. To our knowledge, this is the first multidimensional study (ie, one involving interacting variables and parameters such as DNA copy number aberrations, RNA transcription, and protein expression, as well as the clinical variables, treatment variables, and effect on patient survival) to report on the clinicopathological relevance of *SPAG5* as a predictive marker in breast cancer.

Advances in molecular biology have generated a huge amount of data that have been used to generate multigene profiles (eg, Oncotype DX, PAM50 index, proliferation signature, genomic grade index, and Ki67) to guide chemotherapy treatment. Unfortunately, almost all of these molecular approaches share common issues, such as insufficiently high levels of evidence, overfitting of computational models, high false discovery rates,[30] and absence of potential biological mechanisms to support their use as predictors of therapeutic response. Furthermore, these approaches do not offer a substantial improvement in predictive accuracy compared with well established pathological parameters or cheaper conventional immunohistochemistry approaches (eg, oestrogen receptor, progesterone receptor, or HER2 status, histological grade, tumour size, and lymph node metastases), and they might not be available for logistical or financial reasons.[31] Most of the prognostic power of these assays comes from genes that are related to cell proliferation. Our data show that the prognostic and predictive power of *SPAG5* is independent of many of these multigene tests and *Ki67* expression. Furthermore, our integrated network inference analysis showed that *Ki67* had less effect on other proliferation factors than SPAG5 and was not central to the interactome map compared with other probes. Our results, in accordance with those of a previous study,[32] showed substantial amplification of Ch17q11.2 (the locus of *SPAG5*), in *HER2*-overexpressing, oestrogen receptor-positive breast cancer. Duplication of the *CEP17* locus has been proposed as a marker of chromosomal instability and spindle assembly checkpoint deregulation, and has been linked to anthracycline sensitivity in vitro and the clinical outcome of anthracycline-based adjuvant chemotherapy in breast cancer.[33] Similarly, since *SPAG5* has an essential role in the progression of the cell cycle during the mitotic phase, *SPAG5* dysregulation could contribute to chromosome instability and aneuploidy, both of which are hallmarks of malignant cells and could confer vulnerability to chemotherapeutics drugs on the cancer cell. Drugs such as the anthracyclines and taxanes, which interfere with the normal progression of mitosis, are among the most successful chemotherapeutic compounds used for anticancer treatment. Consequently *SPAG5* could represent a molecular target for the development of next-generation antimitotic drugs. Results from studies in cervical cancer[34,35] have shown that *SPAG5* is upregulated in cervical cancer and suggest that downregulation of *SPAG5* inhibits cell proliferation and growth, increases apoptosis, and hinders cell migration and invasion.[35] By inhibiting cell proliferation and growth of cells, anti-*SPAG5* agents could be used as novel therapeutics in breast cancer. Additionally, induction of *SPAG5* expression could sensitise resistant breast cancer cells to existing treatment regimens.

The most clinically useful aspect of our results is the potential ability to identify patients with breast cancer who are likely to benefit from anthracycline-based chemotherapy. Validation of our results in a randomised trial of anthracycline-based neoadjuvant chemotherapy would determine whether patients whose predicted tumour response would be poor could be spared from the unnecessary risk of cardiac toxicity when other more effective agents could be used. Although the mechanism linking *SPAG5* upregulation and anthracycline response is unknown and further investigation is warranted, it could be because of the accumulation of DNA damage, abnormal mitoses, and subsequent mitotic catastrophe.[36]

Our findings potentially introduce an accurate predictive biomarker for chemotherapy response in breast cancer, which would enable the effective tailoring of treatment to the individual patient. The identification of the role of *SPAG5* as a prognostic and predictive biomarker in breast cancer could lead to the development of novel therapeutic strategies to treat breast cancer, thereby increasing the chance of a cure.

Contributors

SYTC, TMAA-F, and GRB provided intellectual input, conceived the conceptual framework, and designed the study. SYTC, TMAA-F, D-XL, DA, RR, OMR, KL, BX, PMM, ARG, AGP, RCR, CC, IOE, and GRB, were involved in the drafting of the manuscript and took part in critically reviewing it for publication. TMAA-F, DA, and GRB did the statistical analysis, gene expression analysis, and artificial neural network

modelling. RR, OMR, and CC provided *SPAG5* gene copy number aberration data, gene expression data, and did the statistical analysis for the METABRIC cohort. SYTC, TMAA-F, DA, GRB, D-XL, and IOE analysed and interpreted the data. PMM did the immunohistochemistry staining. TMAA-F pathologically assessed experimental slides. PMM, TMAA-F, ARG, and RR collected and managed the patient data.

Declaration of interests

TMAA-F, GRB, and SYTC are named inventors on a PCT patent application for the clinical utilities and implication of SPAG5 in human cancers which is jointly held by the NHS Trust and Nottingham Trent University (US patent publication number 14/404,163 published on June 1, 2012). GRB is named on a patent held by Nottingham Trent University (PCT/GB2009/051412, US [Granted], 8788444, EP [pending] 09796034.8) which covers the artificial neural network algorithms used. The other authors declare no competing interests.

Acknowledgments

DA's PhD programme is supported by Nottingham Trent University Vice Chancellor Bursary Scheme. SYTC and his research group were supported by the Nottingham Hospitals Charity. D-XL's research programme is supported by the New Zealand Breast Cancer Foundation and the Breast Cancer Cure charities.

References

1 Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, et al. Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. *Eur J Cancer* 2013; **49:** 1374–403.

2 Schmidt M, Victor A, Bratzel D, et al. Long-term outcome prediction by clinicopathological risk classification algorithms in node-negative breast cancer—comparison between Adjuvant!, St Gallen, and a novel risk algorithm used in the prospective randomized Node-Negative-Breast Cancer-3 (NNBC-3) trial. *Ann Oncol* 2009; **20:** 258–64.

3 Peto R, Davies C, Godwin J, et al. Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100 000 women in 123 randomised trials. *Lancet* 2012; **379:** 432–44.

4 Diamandis M, White NM, Yousef GM. Personalized medicine: marking a new epoch in cancer patient management. *Mol Cancer Res* 2010; **8:** 1175–87.

5 Yerushalmi R, Woods R, Ravdin PM, Hayes MM, Gelmon KA. K n breast cancer: prognostic and predictive potential. *Lancet Oncol* 2010; **11:** 174–83.

6 Lancashire LJ, Powe DG, Reis-Filho JS, et al. A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks. *Breast Cancer Res Treat* 2010; **120:** 83–93.

7 McShane LM, Altman DG, Sauerbrei W, et al. Reporting recommendations for tumor marker prognostic studies (REMARK). *J Natl Cancer Inst* 2005; **97:** 1180–84.

8 Chin SF, Teschendorff AE, Marioni JC, et al. High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol* 2007; **8:** R215.

9 Ivshina AV, George J, Senko O, et al. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res* 2006; **66:** 10292–301.

10 Curtis C, Shah SP, Chin SF, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 2012; **486:** 346–52.

11 Wang Y, Klijn J, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005; **365:** 671–79.

12 Desmedt C, Piette F, Loi S, et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res* 2007; **13:** 3207–14.

13 Schmidt M, Bohm D, von Torne C, et al. The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res* 2008; **68:** 5405–13.

14 Jezequel P, Campone M, Roche H, et al. A 38-gene expression signature to predict metastasis risk in node-positive breast cancer after systemic adjuvant chemotherapy: a genomic substudy of PACS01 clinical trial. *Breast Cancer Res Treat* 2009; **116:** 509–20.

15 Abdel-Fatah TM, Powe DG, Agboola J, et al. The biological, clinical and prognostic implications of p53 transcriptional pathways in breast cancers. *J Pathol* 2010; **220:** 419–34.

16 Abdel-Fatah TM, Perry C, Dickinson P, et al. Bcl2 is an independent prognostic marker of triple negative breast cancer (TNBC) and predicts response to anthracycline combination (ATC) chemotherapy (CT) in adjuvant and neoadjuvant settings. *Ann Oncol* 2013; **24:** 2801–07.

17 Abdel-Fatah TM, McArdle SE, Agarwal D, et al. HAGE in triple negative breast cancer (TNBC) is a novel prognostic, predictive and actionable biomarker: a Transcriptomic and protein expression analysis. *Clin Cancer Res* 2016; **22:** 905–14.

18 Hatzis C, Pusztai L, Valero V, et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA* 2011; **305:** 1873–81.

19 Horak CE, Pusztai L, Xing G, et al. Biomarker analysis of neoadjuvant doxorubicin/cyclophosphamide followed by ixabepilone or Paclitaxel in early-stage breast cancer. *Clin Cancer Res* 2013; **19:** 1587–95.

20 Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003; **13:** 2498–504.

21 Manning AL, Bakhoum SF, Maffini S, Correia-Melo C, Maiato H, Compton DA. CLASP1, astrin and Kif2b form a molecular switch that regulates kinetochore-microtubule dynamics to promote mitotic progression and fidelity. *EMBO J* 2010; **29:** 3531–43.

22 Shabalin AA, Tjelmeland H, Fan C, Perou CM, Nobel AB. Merging two gene-expression studies via cross-platform normalization. *Bioinformatics* 2008; **24:** 1154–60.

23 Symmans WF, Peintinger F, Hatzis C, et al. Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. *J Clin Oncol* 2007; **25:** 4414–22.

24 Parker JS, Mullins M, Cheang MC, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* 2009; **27:** 1160–67.

25 Liedtke C, Hatzis C, Symmans WF, et al. Genomic grade index is associated with response to chemotherapy in patients with breast cancer. *J Clin Oncol* 2009; **27:** 3185–91.

26 Hess KR, Anderson K, Symmans WF, et al. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J Clin Oncol* 2006; **24:** 4236–44.

27 Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods* 2007; **39:** 175–91.

28 Buechler S. Low expression of a few genes indicates good prognosis in estrogen receptor positive breast cancer. *BMC Cancer* 2009; **9:** 243.

29 Dexter TJ, Sims D, Mitsopoulos C, et al. Genomic distance entrained clustering and regression modelling highlights interacting genomic regions contributing to proliferation in breast cancer. *BMC Syst Biol* 2010; **4:** 127.

30 Hayes DF, Khoury MJ, Ransohoff D. Why hasn't genomic testing changed the landscape in clinical oncology? *Am Soc Clinl Oncol Edu Book* 2012; **2012:** e52–55.

31 Coates AS, Winer EP, Goldhirsch A, et al. Tailoring therapies—improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. *Ann Oncol* 2015; **26:** 1533–46.

32 Jonsson G, Staaf J, Vallon-Christersson J, et al. Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics. *Breast Cancer Res* 2010; **12:** R42.

33 Bartlett JM, McConkey CC, Munro AF, et al. Predicting anthracycline benefit: TOP2A and CEP17—not only but also. *J Clin Oncol* 2015; **33:** 1680–87.

34 Huang L, Zheng M, Zhou QM, et al. Identification of a gene-expression signature for predicting lymph node metastasis in patients with early stage cervical carcinoma. *Cancer* 2011; **117:** 3363–73.

35 Yuan LJ, Li JD, Zhang L, et al. SPAG5 upregulation predicts poor prognosis in cervical cancer patients and alters sensitivity to taxol treatment via the mTOR signaling pathway. *Cell Death Dis* 2015; **6:** e1784.

36 Bertheau P, Espie M, Turpin E, et al. TP53 status and response to chemotherapy in breast cancer. *Pathobiology* 2008; **75:** 132–39.