Supplementary File 8: Narrative synthesis and additional tables for Chapter 2, Development and analytic validity: IHC4

This supplement is split into four parts:

- S8.1: Development of IHC4
- S8.2: A rapid review of the analytical validity of IHC4
- S8.3: IHC4 methodologies of studies included in the prognostic review
- S8.4: Prognostic review IHC4

S8.1 Development: IHC4

The IHC4 score was derived in a sample of 1,125 patients from the TransATAC trial.[1] Tumour blocks were obtained from patients who had already undergone Oncotype DX testing (patients first reported in Dowsett *et al*. 2010)[2] and for whom sufficient tissue was available for IHC4 testing. Patients were HR+, 90% were HER2-, 26% were LN+ (but the percentage with >3 positive nodes was not reported) and 100% were post-menopausal. As such, the test was developed for a patient spectrum that is wider than the patients defined in the decision problem (which is HR+, HER2-, LN0-3 patients).

A summary of the technical methodology used to conduct the test is given in S8.3. In brief, the process involved constructing tissue microarrays with slides of three representative areas containing tumour cells, which were reviewed by a pathologist and/or experienced lab technician. Three cores were assembled for each patient. The immunohistochemistry and scoring of the slides was conducted as described elsewhere.[3, 4] ER was quantified using the H-score, and ER_{10} obtained by dividing the H-score by 30 (to give a value between 0 and 10). PGR10 was obtained by dividing the percent of cells stained positive for PgR by 10 (to give a value between 0 and 10). HER2 was scored according to manufacturer's recommendations (3+ was positive), with fluorescent *in situ* hybridisation to confirm equivocal (2+) samples. Ki-67 was scored as the percent positively stained cells.

The algorithm was developed in two parts, one using the four IHC components, the other using clinicopathological characteristics of nodal status, tumour size, grade, age and treatment (to account for survival advantages in patients whose endocrine therapy was anastrazole instead of tamoxifen). The most informative combination of the four IHC variables to predict time to distant recurrence (equivalent to DRFI, 100 months median follow-up) was derived using multivariable proportional hazard models and change in likelihood ratio X^2. The model derived was:

IHC4 = $94.7 \times (0.100\ \mathrm{ER}_{10}\ 0.079\ \mathrm{PgR}_{10} + 0.586\ \mathrm{HER2} + 0.240 \ln (1 + 10 \times \mathrm{Ki67}))$.

with likelihood ratio X^2 4 *df*= 39.1; p<0.0001

A further model was developed that incorporated the clinicopathological variables, and the **IHC4+C** score was obtained by summing the scores provided from the two algorithms and multiplying by 100.

Clinical score = $100 \times (0.417N_{1\text{-}3} + 1.566N_4 + 0.930(0.497T_{1\text{-}2} + 0.882T_{2\text{-}3} + 1.838T_{>3} + 0.559Gr_2 + 0.970Gr_3 + 0.130Age_{\geq 65} - 0.149Ana))$

where N_j, T_j, Gr_j, and Age_j denote categories of nodal status, tumor size, grade, and age, respectively, and Ana denotes treatment with anastrozole as opposed to tamoxifen. A shrinkage factor was applied to account for overfitting. The likelihood ratio χ^2 for the clinical variables (9 *df*) was 147, p not reported.

Whilst the score was derived using DRFI, and in a cohort containing some LN+ and some HER2+ patients, the authors state that similar IHC4 scores and models were obtained using the endpoint "all recurrences" and LN0 only patients. In the LN0 group, the likelihood ratio χ^2 was 35.4 for the IHC4 component, but the clinical variables were less informative, with χ^2=40.7 (S8.4) compared to the models in the full cohort.

IHC3 Derivation

A further analysis was conducted in a group of patients who were HER2-, which negated the need for the HER2 component of the IHC4 score. A revised algorithm was developed:

$IHC3 = 93.1 \times (0.086\ ER_{10} - 0.081\ PgR_{10} + 0.281 \ln(1 + 10 \times Ki67))$

which was virtually identical to IHC4 when HER2 was negative and was also highly prognostic with χ^2 22.4, p<0.0001 (S8.4).

Analysis of HR+, HER2-, LN0-3 patient in TransATAC

The TransATAC team conducted analyses for the EAG in a subgroup of the TransATAC data set, specified by the decision problem (see Supplementary File 1). Patients who had been tested for any of IHC4, Oncotype DX, Prosigna or EndoPredict were included. This comprised 829 LN0 patients, and 219 LN+ patients with ER+, HER2- disease and who were treated with endocrine monotherapy in total, but fewer with IHC4 scores (792 and 213 respectively). These data are presented alongside the other prognostic data for IHC4 in S8.4, for ease of comparison, but it should be noted that these patients constitute the derivation cohort, and the prognostic value of IHC4 is likely to be overestimated in TransATAC as a consequence, and that the data reported in S8.4 is from the same patients. The new analyses used a cut off of <10% risk, 10-20% and >20% risk to define low, intermediate and high-risk groups.

S8.2 Analytical validity: IHC4

Background

IHC4 relies on the quantification of the immunohistochemistry (IHC) markers oestrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2) and Ki67 for each patient. Whilst a widely adopted technique, IHC can be criticised for a lack of stringency,[5, 6] which in turn can lead to problems with reproducibility between laboratories. Problems with IHC that can lead to variations in quantitative values produced include:

- Pre-analytical methods (e.g. sample type, fixation, storage)
- Analytical methods (e.g. antibodies, staining techniques and reagents) and
- Interpretation (e.g. manual versus automated scoring, using whole slides versus using hot spots or heterogeneous areas, edge areas versus central areas).

The authors of the IHC4 derivation study[1] note that the use of the IHC4 score in laboratories beyond their own (Royal Marsden Hospital) would raise concerns relating to the reproducibility of the component IHC assays.[1] This summary aims to highlight the main issues relating to the use of IHC4 in laboratories other than the Royal Marsden Hospital laboratory (where the score originated) and the recent work that attempts to address some of these concerns.

Methods

It was not possible, within the time-frame of the review, to conduct a full systematic review of the analytical validity of all components of the IHC4 (namely ER, PR, HER2 and Ki67). Instead, we have conducted a rapid review, using systematic search and snowballing search techniques, to identify the most recent and most relevant literature. We have focussed on studies which consider the analytical validity of the IHC4 test, and on studies which consider the analytical validity of Ki67, as this is the most problematic of the four components.[7]

In order to select the most relevant and recent literature we created a long list of potentially relevant studies and then selected the most relevant literature from this, in three stages:

1) Studies from the following sources:

- The main search (primary or secondary studies, including expert reviews). The search was designed to identify studies relating to the analytical validity of IHC4, but not to the component elements (ER, PR, HER2 and Ki67)
- The reference lists of studies included in the prognostic review of IHC4[1, 8-20]
- The reference lists of studies included or cited in existing systematic or expert reviews[21-24]
- Suggestions from clinical experts

2) Identified key studies and conducted citation searches of these within Google Scholar, and added relevant citations to the long list created in step 1. Where the number of citations for a single study was in excess of 100 studies, these were limited (using the Google Scholar "search within citing articles" facility) to those containing the words "analytical validity". The key studies selected for citation searching were:

- Dowsett 2011[25]: International Ki67 in Breast Cancer Working Group recommendations
- Dodson 2016[7]: IHC4 analytical validity study.
- Engelberger 2015[26]: "Score the Core" development study. This was chosen as it relates directly to attempts to improve IHC4 analytical validity
- Polley 2013; Polley 2015; Leung 2016:[27-29] Ki67 analytical validity studies resulting from the International Ki67 in Breast Cancer Working Group[25]. These were chosen as they are recent developmental studies relating to Ki67.

3) Selected the most relevant studies to include in this summary. These were chosen considering the following factors:

- Inter-laboratory reproducibility of IHC4 or Ki67 compared to the Royal Marsden, as this is the centre where the IHC4 score was generated
- Inter-rater reliability of IHC4 or Ki67

As there were no systematic reviews on the analytical validity of IHC4, recent expert reviews and the discussion points raised in the IHC4 prognostic literature[1, 8-19, 30] were consulted to ensure all points of interest were covered.

Summary of findings

A total of 308 titles were screened for relevance. No systematic review relating to the analytical validity of IHC4 or its components was identified. Eight studies (one Working Group report[25] and 7 primary studies[7, 26-29, 31-34]) were included (**Table 1**). These are broadly split into:

i. Analytical validity of IHC4 between Royal Marsden and external centres

ii. Analytical validity of IHC4 within other centres

iii. Analytical validity of Ki67: Studies related to Ki67 Working Group and Royal Marsden

i. Analytical validity of IHC4 between Royal Marsden and external centres

Dodson et al. 2016[7]

Methods: This study[7] (N=28) originated from the Royal Marsden Hospital (London, UK) and conducted two main assessments (**Table 1**). In the first assessment, sections from ER+, HER2- breast cancer tissue micro-arrays were distributed to three centres, where ER, PR, HER2 and Ki67 were stained according to each centre's own standard procedures, and scored at the Royal Marsden Hospital. Individual IHC scores (ER and PR only) and IHC4+C scores were then compared with those produced from slides stained by the Royal Marsden Hospital. This essentially compares different staining techniques, as all other variables are constant. In the second assessment, tissue microarray sections that had been stained at the Royal Marsden were scored by simplified non-counting methods and compared to results obtained through counting. This essentially compares different scoring methods as all other variables are constant. For ER, two different methods of scoring were used: a "simplified H-Score" where each of the four categories were "eye-balled" (instead of counted) and scored as per the usual protocol where the H-Score = (% cells weakly stained x 1) + (% cells moderately stained x 2) + (% cells strongly stained x 3); and an "estimated H-Score" where the proportion of stained cells was eye-balled and multiplied by the modal intensity score (estimated on a scale of 1-3). For PR and Ki67, the simplified method was an "eye-balled" estimate of the proportion stained cells, regardless of intensity of staining.

Results: Correlations between the external centres and the Royal Marsden were high for ER (r=0.93-0.96) and PR (r=0.91-0.98) but moderate for Ki67 (r=0.80-0.89). Upon calculation of the IHC4 scores, these translated to high correlation for IHC4 (r=0.90-0.93) and IHC4+C (0.98-0.99). For risk of distant recurrence at 10 years the correlation was also high (r=0.97-0.98).

The different scoring methods were also highly correlated for ER (r=0.92-0.93) and PR (r=0.98) but correlations were poorer for Ki67 (r=0.86). Again, correlations for IHC4 (r=0.90 to 0.97) and IHC4+C (r=0.97 to 1.00) were high, as were those for distant recurrence (0.97 to 1.00).

Conclusions: The authors conclude that IHC4+C is tolerant of variation in staining and scoring methods, and that additional confirmatory, comparative studies are required.

Critique: The EAG note that only one variable was altered at a time, namely staining technique and counting technique, and that it is unclear whether similar correlations would be achieved in routine clinical practice, where multiple and potentially different variations could occur. The authors themselves acknowledge this limitation and refer to an ongoing study involving 20 centres which may address some of these concerns. In addition, the authors note that HER2 assessment was not included in this analysis (as all patients were HER2-), and cite the high levels of proficiency in this assay in UK centres reported by UK NEQAS.[35]

The authors also have concerns relating to the Ki67 component, and advise the use of formal counting rather than simplified eye-balling methods. The logarithmic transformation of Ki67 data in the IHC4 algorithm is likely to accentuate differences at the lower end of the scoring scale (ie. 0-20% stained cells), where most patients score, and in could lead to a change in risk category for individual patients.

Engelberg 2015[26]

This study aimed to improve the precision and accuracy of assessing ER, PR, Ki-67, and HER2 (IHC4) through use of the online training tool developed and used in Balassanian 2013[31] & Bishop 2012[32] (see below), now termed "Score the Core" (STC). In Engelberg 2015[26], slides were stained at the Royal Marsden Hospital and scored by two pathologists. The *H* scores had a concordance of 0.90 between the first and second pathologist. Slides were then scanned as whole slide images (WSI) and uploaded to the software and distributed to nine pathologists in the Athena Breast Health Network (University of California), and was opened to pathology residents at the University of California Davis as well. Quantitative image analysis (QIA, an overlay of software-generated image analysis) was not available until after the user had submitted their score. HER2 data were excluded from the analysis as only one tumour was HER2+. As slides were stained at one laboratory, this study tests inter-observer reproducibility in scoring after training.

The training programme resulted in a decrease in error in relation to the reference slides for the Athena pathologists for ER and Ki-67 (ER: from 11.4 to 8.6 on a 100-point scale, p=0.03; Ki-67: from 7.8 to 5.7 percentage points, p=0.03), but not for PR which had reasonable agreement to begin with (6.8 to 4.8 on a 100-point scale, p=0.08). When the residents were included, all improvements were statistically significant.

Kappa scores between the reference slides (Royal Marsden Hospital) and the pathologists (Athena network) after training were ER: 0.73; PR: 0.96; Ki67: 0.87. Kappa scores between pathologists (Athena network) after training were ER: 0.77; PR: 0.87; Ki67: 0.62.

Critique: HER2 was not assessed. These results indicate that training improved scoring agreement, but Kappa values (between Royal Marsden pathologists and Athena pathologists, and between Athena pathologists compared to each other) were not always excellent even after training (range 0.62 to 0.96). Kappas for ER were surprisingly lower than might be expected for an established assay (0.73 and 0.77 respectively). Because slides were pre-stained, this study only provides information about inter-rater reliability and it is unclear whether similar Kappa scores would be achieved in routine clinical practice, where multiple and potentially different variations in pre-analytical, analytical and post-analytical factors could occur.

ii. Analytical validity of IHC4 within other centres

Evidence from the main review

None of the prognostic studies identified by the main review[1, 8-19, 30] reported data relating to analytical validity. If the score had demonstrated prognostic value in multiple analyses, it could be argued that the analytical validity was sufficient for the purpose of prognosis. However, the evidence was somewhat mixed (see Chapter 2, Prognostic performance: IHC4 and IHC4+C, of main report), with some studies reporting statistically significant prognostic value and some not, though this did not seem to be associated with the assay methodologies which sometimes differed from those reported in the derivation study.[1]

Balassanian 2013[31] & Bishop 2012[32]

Two abstracts reported on work conducted by the University of California Athena pathology collaboration, to investigate variance in, and harmonise IHC4 staining and scoring across labs. They report some analytical validity results, but also some attempts to improve standardisation of IHC4 methods. Both are reported here.

The first abstract[31] states that five slides from phenotypically different tumours were sent to 5 University of California laboratories, where IHC4 and HER2 FISH tests were conducted according to the prevailing methodology at each lab. Digital whole slide images (DWSI) were also captured, and analysed using quantitative image analysis (QIA). This study therefore tests staining and scoring variance. The abstracts report that there was variance between technical procedures, and between pathologist's scores, but this was not sufficient to affect the clinical score, and that technical staining variance by different laboratories was observed significantly more often for Ki-67than other IHC tests. Antibody vendor or clone did not explain the variance. Parallel analyses using DWSI with QIA suggests that the main source of variance was technical differences, and that WSI with QIA is a robust method to aid harmonisation of IHC4 scoring.

In a second abstract[32] (assumed to be part of, or an extension of, the same study), a similar (or the same) experiment as reported in Balassanian et al.[31] was described, along with two attempts to improve harmonisation . "Technical variance reduction" was attempted, using a Delphi voting process to identify an "ideal slide". Labs then made technical adjustments to their processes to match the appearance (depth of colour, contrast etc) of the ideal slide, and these slides were then scored by pathologists and by quantitative image analysis. "Scoring variance reduction" was attempted through creation of a digital pathology training tool, later to become "Score the Core".

In addition to some of the results reported by Balassanian et al.[31] mean values and variance were similar between WSI and traditional glass slides, except for HER2. Only early results from the quantitative image analysis relating to the "technical variance reduction" efforts were reported, which suggested that there was reduced variance. No results were reported for the "Scoring variance reduction" efforts.

Critique: the analytical validity data from these abstracts suggest that IHC4 scores conducted according to somewhat heterogeneous technical methods do not vary enough to affect clinical practice. There are more problems with Ki67 than ER, PR and HER2. The study further suggests novel concepts to improve harmonisation across labs, including reference slides to harmonise technical differences, use of WSI with QIA to improve scoring differences, and training through a digital tool.

Borowsky 2016[33]

This study used the "Score the Core" training, as developed and used in Balassanian 2013[31] & Bishop 2012[32] and Engelberg 2015[26] and measured inter-observer variance across four sites and nine pathologists after web-based training. 727 tumour samples were sectioned and stained in one laboratory (not reported which), and scored in a random order by two pathologists, hence testing scoring reproducibility. Kappa values were ER: 0.94; PR: 0.84; Her2: 0.91.

Critique: Excellent agreement was reported after training for ER, PR and HER2. Ki67 was not reported. Because slides were pre-stained, this study only provides information about scoring and it is unclear whether similar Kappa scores would be achieved in routine clinical practice, where multiple and potentially different variations in pre-analytical, analytical and post-analytical factors could occur.

iii. Analytical validity of Ki67: Studies related to Ki67 Working Group and Royal Marsden

Because Ki67 is more problematic than the other components of IHC4 (see Dodson 2016[7] above), we have included some additional literature on this topic. However, the search strategy for the assessment report included search terms for IHC4, but not for Ki67 as this was not included in the scope of the assessment. Therefore, a systematic identification of all studies reporting data relating to Ki67 analytical validity has not been conducted. Instead, we focus on studies stemming from the "International Ki67 in Breast Cancer Working Group" (IKBCWG) and/or studies relating to the Royal Marsden hospital where the IHC4 score was generated, as these have highest relevance to the decision problem. However, it should be noted that there is a much larger body of literature on Ki67 which may address some of the issues not addressed by the selected studies.

The IKBCWG produced a set of recommendations in 2011[25] relating to the pre-analytical and analytical assessment, and interpretation and scoring of Ki67, in an attempt to aid harmonization of methodology. They concluded that, at the time, heterogeneity in pre-analytic and analytical methods were not the major source of variation in Ki67 measurements, and that a lack of standardization in scoring procedures (eg, core-cuts vs whole-tumor sections vs tissue microarrays) was problematic. They also stated that the lack of quality assurance schemes made values produced in different labs non-comparable (though an individual lab may have high reproducibility), making use of the score in clinical decision-making (either on its own or in an algorithm such as IHC4) problematic without labs having their own reference data upon which to standardize values.

From this working group stemmed a series of three studies,[27-29] reported below.

Polley et al. (2013)[29]

This study assessed three questions assessing reproducibility between and within laboratories. The first question was reproducibility for Ki67 between laboratories due to differences in scoring. For this, 100 samples were stained centrally (at the Royal Marsden), then sent to eight laboratories (all having published papers on Ki67 i.e. with expertise in this field) where Ki67 was assessed using local methods of scoring. Reproducibility between local and central laboratories was moderate (intraclass correlation (ICC) 0.71, 95% CI: 0.47 to 0.78), implying that differences in scoring have an impact on Ki67. The second was reproducibility between laboratories due to both staining and scoring; this time, 100 samples were both stained and scored locally. Reproducibility between local and central laboratories was lower than above (ICC 0.59, 95% CI: 0.37 to 0.68), implying that differences in staining also impact on Ki67. The third was within-laboratory reproducibility for Ki67, in which 6 labs locally stained 50 samples each and repeated the scoring on three separate days; reproducibility within laboratories was high (ICC 0.94, 95% CI: 0.93 to 0.97). Factors contributing to between-laboratory discordance included tumour region selection, counting method, and subjective assessment of staining positivity. Formal counting methods gave more consistent results than visual estimation (eye-balling).

Polley et al. (2015)[28]

This study assessed reproducibility for Ki67 between laboratories following web-based training in scoring. For this, 50 samples were stained centrally (at the Royal Marsden) and sent to 16 laboratories in 8 countries. Participants scored Ki67 according to a specific protocol after undertaking training. Reproducibility between laboratories was high (ICC 0.94, 95% credible interval (CrI): 0.90, 0.97) when using central staining and web-based training in scoring.

Leung et al. (2016)[27]

This study compared three methods of Ki67 scoring: global method (assessing four fields of 100 cells each); weighted global method (as global but weighted by estimated percentage of total area); and hot-spot method (assessing a single field of 500 cells). For this, 30 samples were stained centrally (at the Royal Marsden) and sent to 22 laboratories in 11 countries. There was moderate inter-laboratory reproducibility for all three methods: unweighted global (ICC 0.87, 95% CrI 0.81, 0.93); weighted global (ICC 0.87, 95% CrI 80, 0.93) and hot-spot (ICC 0.84, 95% CrI 0.77, 0.92). A few cases still showed large scoring discrepancies. Interestingly, a conference abstract for the same study (Dodson et al., 2016) reported that when these Ki67 assessments were integrated into the IHC4+C score, the correlation for risk of recurrence was very high (ICC 0.99, 95% CI: 0.99 to 1.00), implying that variability in Ki67 had little impact on the combined IHC4+C score.

Discussion

Only two studies reported data relating to the analytical validity of IHC4 in centres external to the Royal Marsden and reported good to moderate correlations for ER, PR and Ki67 when comparing different staining techniques, different scoring methods and different observers. Both studies isolated one analytical or counting variable to alter at a time, and one included additional training and standardisation practices, making it unclear if the same favourable correlations would be achievable when comparing samples prepared in totality at different sites or in isolation of the training programme (Score the Core).

Interestingly, despite moderate Ki67 correlations in Dodson 2016a, the IHC4+C correlations were very high (0.98 to 0.99), suggesting the algorithm is robust to a degree of variation in the scoring of component parts. Similar results were reported in a conference abstract (Dodson 2016b[34]) for the Leung 2016[27] study of Ki67, where incorporation of Ki67 values (by any of three methods of counting) into the IHC4+C score resulted in risk category agreement of 98.6%, and in Balassanian 2013[31] where several labs stained and scored 5 slides, but IHC4 scores were not affected by variance in component scores. Whilst these results are reassuring, they represent only a small number of laboratories, and it seems likely that whilst problems with variance in IHC results persist, clinician confidence in using the score may be affected.

Data relating to the analytical validity of IHC4 within other centres was scarce, though our searches are not comprehensive. One study showed that despite considerable heterogeneity between methods of preparation and interpretation the IHC4 scores did not differ enough to change clinical decisions. Excellent agreement between scoring of ER, PR and Ki67 was achieved after training using "Score the Core" on slides stained at one site.

Notably, across these four studies, only one reported correlation data for HER2 (0.91),[33] meaning this is poorly evidenced. Ki67 was not reported in one study, and identified as more problematic than the other factors in three studies; Dodson 2016,[7] Engleberg 2015[26] (though the kappa for Ki67 was 0.87 between more experienced pathologists, and ER also reported Kappas <0.8, for both experienced and resident pathologists), Balassanian 2013[31]& Bishop 2012.[32]

Attempts to standardise Ki67 appear promising as a result of the IKBCWG programme of work, with high levels of correlation within labs, or when using centrally-stained slides. Web-based training for scoring appears to improve agreement, but has not been used on whole sections and biopsy samples. Problems with variations in staining that were evident in Polley 2013[29] do not appear to have been addressed in the selected literature, probably as the original Working Group[25] findings pointed to problems with scoring being the main source of variance.

It should be noted that there are many examples of attempts to improve IHC measurement in the literature that have not been reviewed here due to time and scope limitations. These include digital imaging (which was used as a reference method in some of the studies included here), double staining, variance in antibodies, use of quantum dots, and even novel ways of measuring the markers themselves, such as use of mRNA, chromogenic in situ hybridization and quantitative immunofluorescence (QIF, e.g AQUA which has been used to validate the IHC4 algorithm).[20]

Conclusions

Excellent levels of agreement appear achievable (with web-based training) when slides are prepared centrally. Standardisation of staining may be achievable with training, but has not yet been fully reported or robustly tested (N=5 tumours). Variance in IHC or Ki67 assays may not affect the IHC4 risk scores in clinically meaningful way, but evidence is extremely limited. Efforts to improve Ki67 appear promising but have not yet addressed all variance issues. External quality assessment schemes may improve inter-laboratory agreement.

Table 1: Study characteristics and results

Reference	Targets	Topic	Samples/setting	Experimental variable	Findings	Conclusions
1. Analytical validity of IHC4 between Royal Marsden and external centres						
Dodson 2016a (full paper)[7]	IHC4+C Ki67 ER PR	1) Inter-laboratory reproducibility for ER, PR & Ki67: slides stained at 3 external centres compared with staining at RMH; RMH scoring of all samples by single assessor (i.e. assessing effect of staining method) 2) Scoring via counting methods vs. simplified non-counting-based methods (all stained & scored at RMH)	N=28 tumour samples, ER+, HER2- 4 centres (all UK)	1)Staining 2) Scoring method	1) External vs RMH staining: High correlation for ER (r=0.93-0.96) and PR (r=0.91-0.98) but moderate for Ki67 (r=0.80-0.89). Translated to high correlation for IHC4 (r=0.90-0.93), IHC4+C (0.98-0.99) and risk of distant recurrence (r=0.97-0.98) 2) Non-counting methods vs counting: high correlation for ER (r=0.92-0.93) and PR (r=0.98) but poorer correlation for Ki67 (r=0.86)	1) External vs RMH staining: high reproducibility for ER and PR, moderate for Ki67. Translated to high correlation for IHC4 and IHC4+C scores and distant recurrence 2) Non-counting vs. counting methods of scoring (same lab): high reproducibility for ER and PR, moderate for Ki67. Recommend formal counting for ki67

Reference	**Targets**	**Topic**	**Samples/setting**	**Experimental variable**	**Findings**	**Conclusions**
Engelberg 2015 (full paper)[26]	IHC4 Ki67 ER PR HER2	Development of "score the core" web-based training 1) 1 RMH pathologist stained and scored reference slides, 2nd pathologist re-scored 2)Athena pathologists scored the RMH reference slides after training 3) Athena pathologists scoring RMH slides after training, compared to each other 4) Pathology Residents scored the RMH reference slides after training	N=32 samples from RMH, 9 pathologists at international centres	1-4) Inter-observer reproducibility in scoring after training	1) Scoring agreement between two RMH pathologists for *H* scores on slide stained at RMH, r=0.90 2) Agreement (kappa) between RMH and Athena pathologists after training on scanned slide stained at RMH: ER: 0.73; PR: 0.96; Ki67: 0.87 3) Agreement (kappa) between Athena pathologists after training on scanned slide stained at RMH: ER: 0.77; PR: 0.87; Ki67: 0.62 4) Agreement between reference slides (RMH) and pathology residents after training: lower correlation for PR (P = .03, pooled 2-sample t test) and no significant difference for ER or Ki-67.	“Score the core” web-based training can improve agreement to reference score and between pathologists. Agreement on IHC4 elements scored by different pathologists were not always good.

2. Analytical validity of IHC4 within other centres						
Balassanian 2013 (CA)[31] Bishop 2012 (CA)[32]	IHC4 ER PR HER2 Ki67	1) IHC4 scoring via traditional techniques versus quantitative image analysis (QIA) with whole slide imaging (WSI); stained and scored at local labs within University of California-Athena pathology collaboration 2) Technical variance reduction through use of "ideal slide" 3) Scoring variance reduction through use of web-based training (Score the Core)	N=5 tumour samples, 5 labs,10 pathologists at University of California	1) Inter-lab variance in staining and scoring 2) intervention to reduce technical (staining) variance 3) intervention to reduce scoring variance	1) Considerable and significant technical and interpretational variances exist between laboratories but IHC4 scores do not differ to a clinically meaningful extent. There are more problems with Ki67 than ER, PR and HER2. 2) Early results suggest reduction in staining variance after intervention 3) Results not reported	See findings
Borowsky 2016 (CA)[33]	IHC4 Ki67 ER PR HER2	Interobserver agreement of IHC4 components after "score the core" web-based training (using tissue microarrays to visually score ER, PR and Ki-67). Sections stained at one lab (not named)	N=727 samples, 4 sites, 9 pathologists (Conf abs)	Inter-observer reproducibility after training	"Experts at multiple sites trained with the Score the Core tool can provide high precision IHC quantitation suitable for clinical decision making." Kappa scores: ER: 0.94; PR: 0.84; HER2: 0.91; Ki67: assessed but no correlation reported	After "score the core" web-based training, agreement between pathologists was good for ER, PR, HER2 (assessed but not reported for Ki67)

3. Analytical valdidty of Ki67: Studies related to Ki67 Working Group and RMH						
Dowsett 2011 (recommendations from Ki67 working group)[25]	Ki67	Summary of issues affecting Ki67 reproducibility and recommendations to mitigate these		NA	Issues include: • Preanalytical (type of biopsy, fixative, storage) • Analytic (antibodies, staining etc) • Interpretation and scoring: determination of percentage positive cells; differences between areas of slide (edge vs central, hot spots), visual vs automated • Data analysis: issues with cutpoints Most problematic is methods of counting and a lack of quality assurance schemes.	
Polley 2013[29] (full paper)	Ki67	1&2) Inter-laboratory reproducibility for Ki67, using central or local staining and own method of scoring 3) Intra-laboratory reproducibility for Ki67, local staining, scored on 3 separate days All used MIB-1 antibody	1&2) 8 labs scored n=100 samples, local and central staining (RMH) 3) 6 labs repeated n=50 slides on 3 days Labs USA & Europe, all had papers on Ki67 i.e. experts	1) Scoring 2) Staining and scoring 3) Intra-lab reproducibility of counting	1&2) Interlab reproducibility was only moderate (central staining: ICC = 0.71, 95% CI = 0.47 to 0.78; local staining: ICC = 0.59, 95% CI = 0.37 to 0.68) “Factors contributing to interlaboratory discordance included tumor region selection, counting method, and subjective assessment of staining positivity. Formal counting methods gave more consistent results than visual estimation.” 3) Intralab reproducibility was high (ICC=0.94, 95% CI;0.93, 0.97)	Reproducibility for Ki67 scoring was high within laboratories but only moderate between laboratories (using central or local staining, and local scoring methods)

Polley 2015[28] (full paper)	Ki67	Inter-laboratory reproducibility for Ki67 after web-based training in scoring. Centrally-stained slides (RMH) sent to external labs for scoring according to specific protocol.	N=50 samples 16 labs, 8 countries	1) inter-Laboratory after training	High inter-laboratory reproducibility following web-based training in scoring (ICC 0.94, 95% CrI 0.90, 0.97) May be possible to standardize scoring of Ki67 among pathology laboratories, but clinically important discrepancies persist. Future research needs to apply this technique to biopsies and whole sections, account for staining variability, and link to outcomes.	Reproducibility for Ki67 scoring was high between laboratories when using central staining AND web-based training in scoring
Leung 2016[27] (full paper) Dodson 2016b (CA)[34]	Ki67	Compares three methods of Ki67 counting: global (4 fields of 100 cells) vs. weighted global (as global but weighted by estimated % of total area) vs. hot-spot method (single field of 500 cells). Centrally-stained slides (RMH)	N=30 samples 22 labs in 11 countries	Counting method	Moderate inter-laboratory reproducibility for all methods: unweighted global (ICC 0.87, 95% CrI 0.81, 0.93); weighted global (ICC 0.87, 95% CrI 80, 0.93) and hot-spot (ICC 0.84, 95% CrI 0.77, 0.92). A few cases still showed large scoring discrepancies. When integrated into IHC4+C, ICC for risk of recurrence was 0.99 (95% CI 0.99, 1.00) and risk category agreement (low/intermediate/high) was 98.6% (Dodson 2016 CA) [34] "Establishment of external quality assessment schemes is likely to improve the agreement between laboratories further."	Moderate reproducibility for Ki67 between laboratories for each of three pre-specified scoring methods (using central staining). Translated to very high correlation for IHC4+C recurrence risk (i.e. variability in Ki67 had little impact on IHC4+C)

RMH, Royal Marsden Hosptial; ER, oestrogen receptor; PR, Progesterone receptor; HER2, human epidermal growth factor receptor 2; IHC, immunohistochemistry; CA conference abstract

S8.3: IHC4 methodologies of studies included in the prognostic review

This table details the IHC4 methods listed in each study included in the prognostic review. Column 4 includes advice received by personal communication with the IHC4 team (Andrew Dodson, National External Quality Assessment Service (UK), September 2017) on how compatible the study methodology was with their own in-house methods.

Table 2: IHC4 methodologies of studies in the prognostic review, with judgement about compatibility with derivation study methodology

Author, year	Lab methods	Algorithm	Advice from IHC4 team

Bartlett 2016[20]	**“DAB (conventional 3,3’-diaminobezidine) method:** Formalin-fixed paraffin-embedded tissue blocks were received at a central laboratory and replicate tissue microarrays constructed. Tissue microarrays were analysed by conventional IHC (DAB)… using the Ariol SL50 image analysis platform previously validated for generation of quantitative H-scores[1] Staining with DAB was performed centrally as previously described.[2] Antibodies used were a single batch of antibody (1:50; ER clone 6F11, Novocastra, Newcastle, United Kingdom; 1:50, PgR clone PgR636; HER2 HerceptTest; and 1:50, Ki-67 clone MIB1; all from Dako, Cambridge, United Kingdom) and reagents were used to perform all assays; incubations were temperature controlled. Replicate tissue microarrays were analyzed for ER (n= 6), PgR (n = 6), HER2/neu (n = 3), and Ki-67 (n = 3) staining by using the average score for HER2/neu across all cores analysed and the summed value for both percentages of positive cells and staining intensity (1þ, 2þ, 3þ) based on individual cell counts for ER/PgR and Ki-67 in the final analysis, as previously described.[1]” Quoted verbatim form methods section. Reproduced, with permission, from Bartlett 2016[20] © 2010 College of American Pathologists	The model[1]1 used a linear combination of ER, PR, HER2/neu, and Ki-67. For DAB scores, ER histoscores were divided by 30; PgR percentage positive cells were divided by 10; Ki-67, as percentage positive cells, was used without modification. HER2/neu was treated as a dichotomous variable on the basis of guidelines current to the time.[36, 37]	DAB: Compatible QIF: incompatible

Cuzick 2011[1]	**See** Cuzick 2011[1] methods section		Compatible
Stephen, 2014[17]	"Immunohistochemical staining for a panel of biomarkers including ER, PgR, HER2, Ki67, HTF9C, CEACAM5,NDRG1, p53 and SLC7A5 and FISH (fluorescence in situ hybridisation) for HER2 was performed using either sextuplet(ER and PgR) or triplicate (all other markers) 0.6mm2 TMA cores. Results were derived from dual scoring by expert observers(as described by Kirkegaard et al (2006)) for the Edinburgh BCScohort for all markers. For TEAM patients, ER, PgR and Ki67scores were derived by quantitative image analysis using the Ariolsystem with algorithms validated against both whole sections andmanual assessment (Faratian et al, 2009; Bartlett et al, 2011a). Data for ER were recorded as a histoscore (Kirkegaard et al, 2006) andfor Ki67 and PgR as a percentage of positive cells (ATAC and Ki67guidelines; Dowsett et al, 2011). Results for HER2 were scoredaccording to the UK guidelines (Walker et al, 2008; Bartlett et al,2011b), with cases regarded as HER2-amplified if any core showed amplification/overexpression. Positivity for p53, HTF9C (recentlyre-named TRIMT2A), CEACAM5, NDRG1 and SLC7A5 wasrecorded as previously described.[7-9]" Reproduced from Stephen, 2014[17]© 2016 The Authors. Published by Springer Nature. This work is licensed under the Creative Commons Attribution-Non-Commercial-Share Alike 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/	The IHC4 model (Cuzick et al, 2011[38]) utilised a linear combination of multiple markers: ER, PgR,HER2 and Ki67. Continuous marker scores were normalised prior to inclusion in the IHC4 model. ER histoscores were divided by 30, and PgR scores as a percentage of cells staining positive were divided by 10 to obtain continuous values between 0 and 10. Ki67 scores were represented as percentage positive cells and HER2 was treated as a dichotomous variable. The IHC4 risk score was generated according to the previously specified algorithm (Cuzick et al, 2011).[38] The IHC4 score is analysed as a continuous risk score, except for Kaplan–Meier analyses, in which the IHC4 score is categorised into three groups using two cutoff points that correspond to a 10-year distant recurrence rate of 10% and 20% from the original study; however, these cutoffs have not been previously validated (Cuzick et al, 2011).[38]	Similar
Gluz, 2016c[9]	See Gluz 2016[39] methods section	IHC4 was computed according to the established formulas.[1, 40]	Broadly compatible, but less granularity

WSG-AGO-Doc[39]		Instead of the H-score used in Cuzick et al , the authors determined a general intensity score value of 0 to 3 and multiplied this by the percentage of ER-positive tumor cells to give a final ER score of 0 to 300.	
Nitz 2017[10, 11, 14] WSG-Plan B	See Nitz 2017[10, 11, 14] methods section.	As Gluz, 2016c[9]; WSG-AGO-Doc[39]	Incompatible: Ki67 assessed in 5% increments, which will alter IHC4 score
Gong 2016[12] N=611	"ER was quantified by using the H-score and was considered positive if greater than 1%. The variable ER10 was obtained by dividing the H-score by 30 to obtain a variable with a range of 0 to 10. PgR was scored as the percentage of cells staining positive with a positive cutoff of 10%. PgR10 was obtained by dividing this percentage by 10 to obtain a variable with a range of 0 to 10. HER2 was scored according to the manufacturer's recommendation: 3+ was positive and equivocal 2+ samples underwent fluorescent in situ hybridization analysis and were considered positive only if the ratio was more than 2. Ki-67 scores were recorded as the percentage of positively staining malignant cells. A histogram of the IHC4 score for all the patients is shown in Fig. S4. The median is 5.86 and the interquartile range (IQR, Q2) is 20.97 to 12.25. The hazard ratio (HR) for a change from the 25th (quartile 1, Q1) to 75th (quartile 3, Q3) percentile of the IHC3 score for all patients was 2.58(95% CI, 1.73 to 3.83) in a univariate analysis in 611 patients. Thus, we stratified the patients into low (Q1)-, intermediate (Q2) - or high (Q3) - risk group for convenient description."	As per Cuzik 2011[1]	Unclear

	Reproduced from Gong 2016[12] © 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).		
Lin, 2015[13]	"Tumours were stained for ER, PgR, and HER2 by using IHC. The ER and PgR statuses were determined using the Ventana Benchmark system (Ventana Medical Systems Inc., Tucson, AZ, USA) and prediluted antibodies (anti-ER clone 6F11 and anti-PgR clone 16). ER and PgR were scored as percentage of tumor cells positively staining nuclei, and tumors with ≥ 10% positively stained cells were considered positive. The HER2 status was determined according to the American Society of Clinical Oncology/College of American Pathologists updated guideline[19]. Briefly, scores of 0 and 1+ by IHC were considered negative and 3 + was considered positive. Cases with a score of 2+ were tested for gene amplification by dual probe fluorescence in situ hybridization. HER2/CEP17 ratio ≥ 2.0 and/ or an average HER2 copy number ≥ 6.0 signals/cell were considered positive. The primary antibody for staining Ki67 was anti-Ki67 (1:200 dilution, clone MIB-1, DakoCytomation, Denmark)[20 21], and tumors with ≥ 13.25% positively stained nuclei were considered as highly expressed.[22] " Copied verbatim from Lin, 2015[13] © 2016 The Authors. Published by Springer Nature. This work is licensed under the Creative Commons Attribution-NonCommercial-Share Alike 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/	As per Cuzick et al.[1] the IHC4 score was calculated as IHC4 = 94.7 × (− 0.100 · ER10 − 0.079 · PgR10 + 0.586 · HER2 + 0.240 ln [1 + 10 · Ki67]). As assay methods differed, study participants were categorised into low, intermediate, and high risk groups according to the IHC4 scores of < 25th, 25th–75th, and > 75th percentiles, respectively.	Unlikely to be compatible – used image analysis for ER+ and PgR, Ki67method unclear
Rohan, 2014[16]	See methods section of Rohan et al. 2014[23]	Cuzik et al. 2011[1]	Unlikely to be compatible – applied re-fitted IHC4+C algorithm to

			the population
Viale 2013[18]	See methods section of Viale 2013[18]	NR	Unclear
Vincente-Salomon 2013[19]	Immunostaining was done according to previously published protocols[26]. The expression of ER (clone 6F11; 1/200; Novocastra), progesterone receptor (PR; clone 1A6; 1/200; Novocastra), ERBB2 (clone CB11; 1/1,000; Novocastra), epidermal growth factor receptor (HER1; clone 31G7; 1/40; Zymed; Clinisciences), cytokeratin 5/6 (clone D5/16B4; 1/50; Dako), and cytokeratin 8/18 (clone DC10; 1/100; Zymed; Clinisciences) were evaluated. For each antibody, internal and external controls were included in the experiments. ER, progesterone receptor, HER2 receptor and KI67 status were assessed by immunohistochemistry on representative formalin-fixed tumor blocks, according to previously published protocols[27]. The semiquantitative KI67 assessment was performed as previously published[28] and as recommended[29]. A cut-off of 14% was used to define tumors with a high KI67 score (according to St Gallen recommendations[30] and cut-off for molecular classification.[13] Internal (normal glands surrounding the carcinoma) and external controls (for ER, PR and HER2: tissue-microarrays composed of tumors with known ER, PR status, and known numbers of HER2 gene copiestogether with normal mammary tissue; for KI67: normal lymph node with germinal centers as positive controls) were included in all immunostaining experiments. Reproduced from Vincente-Salomon 2013[19] © 2013 Vincent-Salomon et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.	Cuzik et al. 2011[1] Used IHC3 algorithm as patients HER2-	Compatible

Prat 2012[15]	See methods section of Prat 2012[15]	IHC4 was computed according to established formulas.[1, 40] Instead of the H-score used in Cuzick et al , the authors determined a general intensity score value of 0 to 3 and multiplied this by the percentage of ER-positive tumor cells to give a final ER score of 0 to 300.	Compatible

S8.4 Prognostic performance: IHC4 and IHC4+C

In addition to the TransATAC derivation cohort[1, 41] (see S8.1 above), IHC4 has been reported in eleven separate cohorts, reported across fourteen publications (see Table 22 of the main report).[1, 8-18, 20, 42] The size of the studies ranged from N=105[42] to 4,598.[8] Data relating to the subgroup of patients relevant to the decision problem (HR+, HER2-, LN0-3) from the derivation cohort (TransATAC) were provided in a personal communication from the transATAC team[41]. One cohort (Tamoxifen vs Exemestane Adjuvant Multinational(TEAM) trial) was reported in two separate analyses,[8, 17, 20] with different aims (validation of IHC4;[8, 20] prognosis of early or late recurrence)[17] and different numbers of patients (n=4598;[8, 20] n=2513)[17] as Stephen *et al.* 2014[17] recruited only those who had received endocrine monotherapy. Laboratory methodologies for conducting IHC4 varied across studies, and is discussed in more detail below (section "IHC4 methodology and cut-offs: IHC4 and IHC4+C prognostic performance").

Study designs: IHC4 and IHC4+C prognostic performance

Five of the validation cohorts[8-11, 14, 15, 17, 18, 20] and the derivation cohort[41] were a reanalysis of prospectively collected RCT data, using archived tissue samples. The remaining six studies[1, 12, 13, 16, 17, 42] were analyses of cohorts of routinely collected patient data; one of these was a case-control study[16] (see Table 22 of the main report).

The derivation RCT was from the UK:

- ATAC[2, 41] – was an international trial, with a translational research continuation (TransATAC) that investigated prognosis of breast cancer recurrence. Only UK samples were included in this analysis. The trial evaluated anastrozole, tamoxifen, or the combination of both treatments. Recruitment ended in 2006. There are numerous TransATAC publications that met the criteria for the review,[1, 2, 43-50] but here we present data provided by the TransATAC team as a personal communication to the EAG, which restricts to HR+, HER2-, LN0-3 patients.[41]

Two RCTs were conducted in the UK and other countries:

- The TEAM trial[51] recruited patients between 2001 and 2006 and randomised them to exemestane alone or following tamoxifen.
- The IES (Intergroup Exemestane Study) trial[52] recruited patients between 1998 and 2003 and randomised them to one of two endocrine therapies: exemestane or tamoxifen.

The remaining three RCTs were conducted in Europe (Spain and Germany):

- WSG (West German Study Group) Plan B trial[53] recruited patients between 2009 to 2011, and randomised them to anthracycline-free or anthracycline-taxane based chemotherapy. In an early protocol amendment, patients with Oncotype DX RS <12 were not given chemotherapy.
- GEICAM 9906 (Grupo Espanol de Investigation en Cancer de Mama)[15, 54] randomised patients with node-positive disease to adjuvant fluorouracil, epirubicin, and cyclophosphamide versus fluorouracil, epirubicin, and cyclophosphamide followed by weekly paclitaxel, and patients with HR-positive disease subsequently received adjuvant endocrine therapy.
- WSG-AGO-Doc (West German Study Group epirubicine and cyclophosphamide-Doc)[39] recruited patients between 2000 and 2005 and randomised them to taxane or non-taxane-based chemotherapy regimens.

There were a total of six retrospective studies. Three studies were from the UK or Europe:

- A cohort from Nottingham, UK[1]
- A cohort from Edinburgh, UK[17]
- A cohort from France (Institut Curie).[42]

One study was from the USA, where clinical advice to the EAG suggests chemotherapy rates are generally higher:

- Patients in the Kaiser Permanente Northwest[16] database

A further two studies were from East Asia:

- A cohort from China[12] from the Sun Yat-sen Memorial Hospital and the Third Hospital of Nanchang City
- A cohort from Taiwan[13] from the National Taiwanese University Hospital.

Clinical advice received by the EAG suggests that these two East Asian studies may be less generalisable to the English context because: (a) patients were treated according to usual clinical practice and this may differ in these countries compared with the UK enough to affect prognostic outcomes, and (b) it is possible that people of different ethnicities have different underlying risk profiles and disease natural history. For this reason, data from these studies should be interpreted with caution and with reference to data from studies where the ethnic profile and clinical practice is similar to the UK.

Patients and treatments: IHC4 and IHC4+C prognostic performance

The studies were highly heterogeneous in terms of the patients recruited and the treatments given. Overall, only the derivation cohort (TransATAC)[41] reported an analysis of 100% ER+, HER2-, LN0-3 patients who had not undergone chemotherapy but had received 5 years of endocrine therapy. Data from this cohort were provided to the EAG as Academic in Confidence, and has limitations in that: (a) it is also the derivation cohort for the IHC4 score, so some overfitting (leading to overestimation of prognostic performance) can be expected, (b) it only recruited post-menopausal women, and (c) it did not recruit PR+ patients.

As such, most of the evidence base has low generalisability to the decision problem, and even the most relevant available evidence has limitations in that TransATAC is the derivation cohort for IHC4 and only recruited ER+ post-menopausal patients. These limitations along with the problems with patient cohorts and treatments given should be borne in mind when interpreting the evidence base.

What follows is a more detailed look at the evidence base from the perspective of each factor of importance to the decision problem:

Lymph node status: The IHC4 test was developed for use amongst LN+ or LN0 patients, though this assessment focusses on those with LN0-3. Amongst the RCT reanalysis studies, TransATAC[1, 41] and WSG Plan B[10, 11, 14] recruited or reported a subgroup of patients with LN0-3, whilst TEAM[8, 17, 20] and IES[18] recruited patients with any lymph node status, and did not report the percentage with more than three positive nodes. GEICAM 9906[15] and WSG-AGO-Doc[9] recruited LN+ patients, with 38% patients having LN>3 in GEICAM 9906 but all patients being LN1-3 in WSG-AGO-Doc.

Amongst the retrospective cohort and case control studies, the Nottingham,[1] the Kaiser Permanente,[16] the Edinburgh (BCS),[17] the Chinese[12] and the Taiwanese[13] data sets all recruited both LN positive and negative patients, but did not report the proportion who were LN>3. The cohort from the Institut Curie[42] were all LN0.

Hormone receptor status: IHC4 was intended for use in HR+ patients. All studies recruited HR+ or ER+ patients except the IES RCT[18] and the study from Taiwan,[13] both of which did not report the percentage of patients who were HR+ (see Table 22 of the main report).

HER2 status: The IHC4 test was developed for both HER2+ and HER2- patients, though this assessment focusses on HER2- patients. Amongst the RCT reanalysis studies (see Table 22 of the main report), TransATAC,[41] WSG Plan B,[10, 11, 14] GEICAM 9906[15] and WSG-AGO-Doc[9] recruited or reported a subgroup of HER2- patients, whilst TEAM[8, 17, 20] and IES[18] did not report the HER2 status of patients. Amongst the retrospective studies (see Table 22 of the main report), the Kaiser Permanente cohort,[16] Institut Curie[42] cohort and the Chinese[12] cohort all recruited 100% HER2- patients whist the Nottingham cohort,[1] Edinburgh (BCS)[17] cohort and the Taiwanese[13] cohort recruited a proportion who were HER2+, or did not report this.

Treatments: IHC4 was intended for use in predicting distant disease recurrence assuming 5 years of endocrine therapy in HER2- patients, and no chemotherapy. As such, failure to treat all HER2- patients with endocrine therapy or treatment of any patients with chemotherapy will affect the survival of patients, and the estimates of prognostic performance may also be affected, especially if the proportion of patients given or not given treatment differs in each risk group; in theory, assuming patients in the higher risk categories get chemotherapy more often (if there is some concordance between clinically-defined risk and tumour profiling test risk), this is likely to reduce the separation in observed risk between IHC4 risk categories reported in these studies. This type of problem is theoretically possible in the retrospective studies of routine practice, where the IHC4 markers alone are likely to have affected treatment decisions, but also in the RCT study WSG Plan B, where patients with Oncotype DX RS<12 were given endocrine monotherapy and those with RS≥12 were given chemotherapy and endocrine therapy, if there is some concordance between Onctoype-DX and IHC4 categorisations.

Only two data sets treated all HER2- patients with endocrine therapy and did not treat any patients with chemotherapy (TransATAC[1, 41] and the analysis of TEAM conducted by Stephen *et al*. 2014 (see Table 22 of the main report).[17] The analysis by Stephen *et al*. is likely to suffer from spectrum bias as patients were excluded if they received chemotherapy, and these patients are likely to be systematically different to those who did not as chemotherapy decisions were based on clinical practice in this trial (only exemestane/tamoxifen treatment was randomised). Five studies treated all HER2- patients with endocrine therapy but also treated some patients with chemotherapy, or were assumed to have treated some patients with chemotherapy as they were treated according to routine practice (WSG Plan B,[10, 11, 39] IES,[18] GEICAM 9906, [15] China[12] cohort and the Bartlett *et al*. 2016[8, 20] analysis of TEAM, (see Table 22 of the main report). The Nottingham IHC4 validation cohort[1] included some HER2- patients who were not treated with endocrine therapy, but applied a correction in the analysis to account for this; however, as the cohort were patients undergoing routine therapy, it is likely that some received chemotherapy and no adjustment for this is reported (see Table 22 of the main report). Three studies (Kaiser Permanente,[16] WSG-AGO-Doc,[9] Taiwan[13] (see Table 22 of the main report) did not treat all patients with endocrine therapy or did not report the proportion who were treated, and one study (Institut Curie[42]) treated some patient with endocrine therapy, but none with chemotherapy.

IHC4 methodology and cut-offs: IHC4 and IHC4+C prognostic performance

The methodology for conducting IHC4 is well known to be problematic. Concerns centre on the performance of Ki-67, and specifically the lack of standardisation of laboratory and analytic methods.[20] [25] We have documented the methods reported in the included studies in S8.3 for reference, but as it was beyond the expertise of the EAG to identify which methods are in accordance with UK practice, and the methods used by the derivation group, [1] we sought advice from the IHC4 team. Their judgement regarding the compatibility of the methods used in the studies to their own methodology (used in their laboratory) is given in S8.3, and in Table 2. Seven datasets were analysed using IHC4 methodologies that were the same or very similar to the IHC4 team's own methodology (referred to from here on in as the standard IHC4 methodology) (TransATAC[41],_TEAM,[8, 17, 20] the Nottingham cohort,[1] the BCS cohort,[17] the Institut Curie[42] cohort, GEICAM 9906[15] and WSG-AGO-Doc)[9] whilst the remaining five datasets were analysed with methodologies that were unclear or dissimilar to the IHC4 team's methods (WSG-Plan B,[10, 11, 14] the Kaiser Permanente cohort,[16] IES,[18] the Chinese cohort [12] and the Taiwanese cohort[13]). Results have not been excluded by IHC4 methodology, as methodologies are not currently standardised and as such all data is of some relevance.

A brief description of methods is given for each study in Table 22 of the main report. Three studies were unclear whether it was the IHC4 score or the IHC4+C score, as they referenced Cuzick *et al*. 2011,[1] but not which score; attempts were made to clarify this point with the authors where contact details were available (IES;[18] Institut Curie cohort;[42] WSG-AGO-Doc).[9] Most other studies used only the IHC4 component of the IHC4 score, without using the clinical component (see Chapter 2, Development and analytic validity: IHC4, of main report) (TEAM analyses by Barlett *et al* 2016[20] and Stephen *et al*. 2014;[17] Edinburgh cohort;[17] WSG Plan B;[10, 11, 14] GEICAM 9906; [15] Kaiser Permanente cohort;[16] China cohort[12]; Taiwan cohort).[13] Data definitely stated to relate to IHC4+C was only available for the Nottingham cohort[1] and TransATAC.[41]

The original IHC4[1] analysis did not report numerical cut-offs for the definition of high, intermediate and low-risk patients, but used quartiles and tertiles, whilst the analysis of TransATAC uses 10%, 10-20% and >20% risk or recurrence as cut offs. Other studies used quartiles and/or tertiles to define the cut-offs, or used the score as a continuous variable in cox proportional hazard models, except the Stephen *et al*. analysis of BCS and TEAM, [17] which stated that the same cut-offs as Cuzick *et al*.[1] were used.

The Insitut Curie trial,[42] which recruited all HER2- patients, stated that they used the IHC3 version of the IHC4 algorithm, where HER2 status is not incorporated. It is unclear whether other studies that recruited only HER2- patients and referenced Cuzick *et al*. 2011[1] as the source of the algorithm also used the IHC3 score, as reported by Cuzick *et al* 2011.[1]

Comparators: IHC4 and IHC4+C prognostic performance

No studies of IHC4 compared the score to a comparator. The TransATAC study reported data with NPI and CTS as comparators. The Nottingham cohort analysis also reported a comparison to the clinical score component of the IHC4+C score.

Quality assessment: IHC4 and IHC4+C prognostic performance

The evidence base was of generally poor quality; no study scored well on all items (Table 4). Of particular concern was the high number of studies that included patients who had received chemotherapy treatment (see section entitled "*Treatments*" above), and the high number of studies that were not able to include all relevant patients due to missing samples or insufficient tissue. This is likely to introduce spectrum bias, as patients with smaller tumours are more likely to have been excluded due to insufficient tissue being available. Very few studies reported that they blinded test assessors, leaving the evidence base at high risk of ascertainment bias. The applicability of the IHC4 tests conducted to the decision problem is acceptable in seven studies (TransATAC[41],_TEAM,[8, 17, 20] the Nottingham cohort,[1] the BCS cohort,[17] the Institut Curie[42] cohort, GEICAM 9906[15] and WSG-AGO-Doc)[9], but unknown or not compatible in five (WSG-Plan B,[10, 11, 14] the Kaiser Permanente cohort,[16] IES,[18] the Chinese cohort [12] and the Taiwanese cohort[13]).

Results: IHC4 prognostic performance: Unadjusted analyses

This section reports unadjusted analyses. Adjusted analyses, which show whether the test has prognostic value over clinicopathological variables, are reported in the section "Additional prognostic value"

DRFS: Three studies[12, 13, 16] reported unadjusted analyses for this outcome and results are reported in Table 23 of the main report. None used methods compatible with the standard IHC4 methodology. Kasier Permanente[16] reported 5-year DRFS for LN0 patients, using tertiles with cut-offs defined as low-risk: ≤-7.81; intermediate-risk: >-7.81 to 88.32; high-risk: >88.32. Not all patients had endocrine therapy and some patients had chemotherapy. An odds ratio analysis of 5-year DRFS for intermediate vs low-risk (1.76 (95% CI 1.10 to 2.84)) and high vs low-risk patients (2.54 (95% CI 0.97 to 6.62)) gave a p value of 0.01. The C-index (AUC) was 0.62

(95% CI NR); values above 0.5 indicate the test is better than chance in placing patients into appropriate risk categories.

The two East Asian studies[12, 13] with uncertain generalisability to the UK context (recruited any lymph node status; variable endocrine and chemotherapy treatments; used methods not compatible with the standard IHC4 methodology) were in general agreement with Kaiser Permanente.[16] They reported statistically significant HRs for high-risk patients (above the 75th percentile) versus low-risk patients (below the 25th percentile) (1.454, (95% CI: 1.133, 1.866, p=0.003) and 2.33 (95% CI: 1.41: 3.85, p NR) respectively). Results for intermediate (between 25th to 75th percentile) vs low were not statistically significant[12] in one study and statistically significant in the other.[13]

DRFI: The Nottingham cohort and the IES study both[1, 18] reported unadjusted analyses for 5 year DRFI, and results are presented in Table 23 of the main report. Only the Nottingham cohort[1] used the standard IHC4 methodology. Both studies reported statistically significant 5 year DRFI HRs for high versus low-risk groups, defined as quartiles (patients above the 75th quartile high-risk; patients below the 25th quartile low-risk) [1] or tertiles (not defined further)[18] but with different 5-year DRFI HRs (4.1 (95% CI: 2.5, 6.8) versus 2.3 (95% CI: 1.1, 4.7) respectively). This may be due to the different categorisation of patients (quartiles versus tertiles) or differences in patients recruited (LN0/+ versus LN0 respectively), or treatments given (not all patients received endocrine therapy in the Nottingham cohort; some patients received chemotherapy in the IES cohort). A comparison of patients between the second and first tertile to those below the first tertile in the IES study[18] was not statistically significant (5-year DRFI HR 1.4 (95% CI: 0.7 2.9)).

RFS: Both Bartlett *et al.*'s analysis of the TEAM trial[8, 20] and the Taiwanese cohort[13] reported 5-year RFS and results are presented in Table 5. Only the TEAM trial[8, 20] analysis used the standard IHC4 methodology. Both studies recruited LN0/+ patients, and both treated some patients with chemotherapy. Both reported statistically significant differences for IHC4 risk categories (HR not reported, p<0.001 in TEAM;[8, 20] HR 2.33 (1.41, 3.85) in the Taiwan cohort)[13], except for an analysis of those below the 25th quartile to those between the 25th and 50th quartile in the TEAM[8, 20] trial (p=0.11).

IDFS: see Table 6. The WSG-Plan B[10, 11, 14] trial (LN0/+), where clinically high-risk patients were recruited, and patients with Oncotype DX <12 received endocrine monotherapy and those with RS ≥12 received endocrine and chemotherapy reported a statistically significant 5 year IDFS HR for those above the 75th versus those below the 25th quartile of 2.04 (95% CI: 1.47, 2.83, p<0.001). Similarly, 5 year IDFS results from the LN+ WSG-AGO-Doc trial,[9] where patients all received chemotherapy and the % receiving endocrine therapy was not reported, were statistically significant for the same analysis (HR 2.12 (95% CI: 1.32, 3.42, p 0.002)). Only the WSG-AGO-Doc trial[9] used the standard IHC4 methodology.

IDFI: See Table 7. The lymph node negative Insitut Curie[42] cohort, where some patients received endocrine therapy and none received chemotherapy, reported a non-statistically significant effect for an analysis of IHC3 as a continuous variable (HR 1.01 (95% CI: 1.00, 1.01, p=0.204)). This study was compatible with the standard IHC4 methodology.

Additional prognostic value: IHC4

This section reports adjusted analyses, which indicate the additional prognostic value of IHC4 over clinicopathological factors. The clinicopathological factors adjusted for vary from study to study, and are detailed in the footnotes to the tables.

None of the seven cohorts that reported data relating to the additional prognostic value of IHC4 over other clinicopathological risk scores or versus clinicopathological factors in multivariable analyses recruited HR+, HER2- LN0-3 patients and treated them with 100% endocrine therapy and 0% chemotherapy (Table 24 of the main report). The closest study to the decision problem was the analysis of TEAM and the Edinburgh cohorts by Stephen *et al*. 2014,[17] though selection of chemotherapy-untreated patients in the Edinburgh cohort and from the TEAM trial may have led to spectrum bias, as patients not treated with chemotherapy in routine practice are likely to be systematically different to those who are treated with chemotherapy. As such, all estimates should be interpreted with caution. Three studies (WSG-Plan B, Kaiser Permanent cohort and the Taiwan cohort)[10, 11, 13, 14, 16] did not use methods compatible with standard IHC4 methodology.

Outcomes included DRFS, DRFI, DFS, IDFS and RFS. Across these outcomes, across the seven cohorts reporting relevant data (Edinburgh cohort, TEAM, WSG Plan B, Kaiser Permanente cohort, WSG-AGO-Doc, GEICAM 9906, Taiwan cohort),[8-11, 13-17, 20] the picture on additional prognostic value was mixed. The analysis conducted by Stephen *et al.*[17] analysed the Edinburgh cohort (median follow-up 12.9 years) and the TEAM cohort (median follow-up 6.2 years) separately, and reported HRs and D-statistics for IHC4 and clinical factors separately, where a difference in D statistics of 0.1 or more indicated improved prognostic separation. HRs (unclear which risk groups compared) were not statistically significant at 0-5 and 5-10 years for DRFI, but the separation in D-statistics between IHC4 and clinicopathological factors were greater at 0-5 year follow-up rather than at full follow-up in both cohorts, and the difference was 0.1 or more in all but the full follow-up analysis of the Edinburgh cohort. The authors interpreted these data as indicating that the additional prognostic value of IHC4 was restricted to the first five years of follow-up. Further to this, multivariable analyses of subgroups of LN0 and LN+ patients showed a statistically significant 0-5 year DRFI HR only for the LN0 subgroup of the Edinburgh cohort (HR 3.16 (95% CI: 1.03, 9.64).

The analysis by Bartlett *et al.*[20] of the TEAM trial (LN0/+, which did not select for endocrine monotherapy and therefore included some patients treated with chemotherapy) also reported a statistically significant HR of 1.006 (95% CI: 1.004, 1.008) when IHC4 was analysed as a continuous variable in a multivariable model including clinicopathological factors, with an increase in likelihood ratio χ^2 over clinicopathological factors of 38.5 (29%). WSG-Plan B,[10, 11, 14] in a mixed cohort of LN0/+, also reported a statistically significant HR of 1.59 (95% CI: 1.15, 2.2), p=0.005) when IHC4 was fractionally ranked by 75th to 25th percentiles in a multivariable model including clinicopathological factors. The Kaiser Permanente[16] LN0/+ cohort reported a statistically significant 5-year DRFS odds ratio of 1.06 (95% CI: 1.00, 1.13) when the score was analysed as a continuous variable in 10 unit increments in a multivariable model including clinicopathological factors, but not when an odds ratio was calculated (1.61 (95% CI: 0.48 5.47) for those above the highest tertile versus those below the lowest tertile). The Taiwanese study also reported a statistically significant HR for those above the 25th percentile versus those below the 25th percentile (1.90 (95% CI: 1.32, 2.73, p<0.001) in a multivariable model including clinicopathological factors.

No studies apart from Stephen *et al.*[17] reported on LN0 patients (see above). Stephen *et al.* [17] reported multivariable DRFI HRs corrected for clinicopathological variables at both 0-5 and 5-10 years in the TEAM and Edinburgh analyses. These were not statistically significant (which was also true for the HRs for the full LN0/+ analysis, where the D-statistic did show an effect), except for 0-5 years in the Edinburgh cohort (HR 3.16 (95% CI: 1.03, 9.64)), but no D-statistics were reported.

WSG-AGO-Doc[9] and GEICAM 9906[15] and the Stephen *et al.*[17] analysis of TEAM and Edinburgh cohorts (see above) reported LN+ cohorts. WSG-AGO-Doc[9] reported a non statistically significant HR in a multivariable analysis corrected for clinicopathological variables, whilst GEICAM 9906[18] reported a statistically significant increase in likelihood ratio χ^2 over clinicopathological variables (13.5, p<0.05). As already stated, the analysis in TEAM and Edinburgh were not statistically significant in multivariable analyses at both 0-5 and 5-10 years for HRs, but no D-statistics were reported..[17]

Broadly speaking, results did not appear to be influenced by the compatibility of the IHC4 methodology with the standard methodology, with both statistically significant and non-significant results being reported in both compatible and non-compatible studies.

Results: IHC4+C prognostic performance: Unadjusted analyses

This section reports unadjusted analyses. Adjusted analyses are reported in the section "Additional prognostic value".

DRFI: Both the Nottingham cohort[1] and the TransATAC[41] derivation cohort re-analysis reported DRFI for IHC4+C, and results are presented in Table 25 of the main report. The TransATAC analysis used the cut-offs of <10% risk, 10-20% and >20% risk to define low, intermediate and high-risk groups and reported data for LN0-3, LN0 and LN1-3. TransATAC analysis reports statistically significant 5 and 10 year DRFI HRs for the LN0-3, the LN0 and LN1-3 analyses (see Table 25 of the main report) for both high versus low and intermediate versus low comparisons. HRs were higher in the LN1-3 subgroup than the LN0 subgroup. For example, the 10 year DRFI high-risk versus low-risk HR was 6.42 (95% CI: 3.37, 12.24) in the LN0 group and 10.34 (95% CI: 2.44, 43.89) in the LN1-3 group. Interestingly, 5 year DRFI HRs were higher than 10 year DRFIs in the LN0 group, but lower in the LN+ group; the high-risk versus low-risk HR at 5 years in the LN0 group was 11.39 (95% CI: 4.05, 32.01) compared with 6.42 (95% CI: 3.37, 12.24) at 10 years, whilst the same analyses were 8.82 (95% CI: 1.14, 68.30) and 10.34 (95% CI: 2.44, 43.89) respectively in the LN1-3 group. A similar trend in the intermediate versus low analyses was reported (see Table 25 of the main report).

The IES study in LN0 patients (100% endocrine therapy, 19% chemotherapy) reported that "*addition of clinical variable to IHC made the effect more profound*" which is ambiguous but could indicate that the addition of the clinical score to the IHC4 score increased the 5 year DRFI HR (those below the 1st tertile versus those above the 3rd tertile), which was 2.3 (95% CI: 1.1, 4.7).

Broadly speaking, results did not appear to be influenced by the compatibility of the IHC4 methodology with the standard methodology.

OS: Only the TransATAC study (derivation cohort) reported OS results for IHC4+C, and only in LN0 and LN1-3 subgroups. The results are presented in Table 8. The HRs were much lower than for DRFI, for example, the 10 year DRFI high-risk versus low-risk HR was 6.42 (95% CI:

3.37, 12.24) in the LN0 group and 10.34 (95% CI: 2.44, 43.89) in the LN1-3 group, whilst the OS were 3.18 (95% CI: 1.52< 6.65) and 2.93 (95% CI: 1.91, 4.50), respectively.

Additional prognostic value: IHC4+C

This section report adjusted analyses, which indicate the additional prognostic value of IHC4+C over clinicopathological factors. The clinicopathological factors adjusted for vary from study to study, and are detailed in the footnotes to the tables.

The additional prognostic value of IHC4+C was analysed in the TransATAC (derivation) cohort[41] and the Nottingham cohort[1] (Table 26 of the main report). Both studies used methodologies compatible with the standard IHC4 methodologies. In the TransATAC analysis, additional prognostic value was assessed via increases in likelihood ratio χ^2 for 5-year and 10-year DRFI, for IHC4+C plus NPI or CTS, over NPI or CTS alone (Table 26 of the main report). Increases in likelihood ratio χ^2 at 5 and 10 years were statistically significant for LN0 patients: 10 year DRFI change in likelihood ratio χ^2 17.14 (p<0.0001) over CTS and 21.91 (p<0.0001) over NPI, but not statistically significant for LN+ patients: 3.08 (p=0.08) over CTS and 2.45 (p=0.10) over NPI (Table 26 of the main report). Similarly, the Nottingham cohort reported an increase in likelihood ratio χ^2 over the clinical score component of the IHC4 total score of 25.89 (p<0.0001) and an HR of 3.9 (95% CI: 2.3, 6.5) in a multivariable analysis adjusted for clinicopathological variables. If the CTS is the same as the clinical component of IHC4+C, then likelihood ratio χ^2 provides the additional prognostic value of IHC4, over CTS.

Table 3: Data relating to the derivation of IHC4 score and IHC3. DRFI (100 months median follow-up). All data from TransATAC

Reference; N	Cohorts	Population	Nodal status	Endo / chemo	Likelihood ratio χ^2	DRFI: HR (95% CI) Unadjusted, 0-25th vs 75-100th percentile:	DRFI: HR (95% CI) Multivariable[a]
Cuzick 2011[1] N=1,125	TransATAC	100% HR+ 90% HER2- Postmeno	LN+/-	100% ET monotherapy	**IHC4:** 39.1, p<0.0001 **Clin:** 147, p NR	**IHC4:** 5.7 (3.4 9.7)	**IHC4:** 3.9 (2.4, 6.7)
N=793			LN0		**IHC4:** 35.4, p NR **Clin:** 40.7, p NR		
N=1,066		100% HER2-	LN+/-		**IHC3:** 22.4, p<0.0001		

IHC4, IHC4 component alone; Clinical, clinical component alone
a multivariable model assumed to include IHC4 score and Clinical score as separate components

Table 4: Quality assessment of prognostic studies: IHC4 and IHC4+C

Reference(s); N	Cohort(s)	Derivation or validation?	Study design appropriate?	All eligible patients included?	Blinding (of test assessors to outcomes)?	Outcome definition standardised or *a priori*?	Applicability: Patient Spectrum	Applicability: Test as per decision problem?
Bartlett 2016[20] Christiansen 2012[8] N=2919[20] N=4598[8]	TEAM	V	N, Some CT	UC	UC	Y	UC (ER2- NR; LN>3 NR)	Y
Cuzick 2011[1] N=786	Nottingham	V	UC, % CT NR	UC	UC	Y	UC, %LN>3 NR, CT NR	Y
Gluz, 2016c[9] N=459	WSG-AGO-Doc	V	N, some CT	N, InsT,	UC	Y	Y	Y

Gong 2016[12] N=611	SYSMH; CCSYU; 3rdHNC	V	N, some CT	N InsT; MD	UC	Y	N, InT, MD, CT,	UC, assay methods unclear
Lin, 2015[13] N=605	National Taiwan University Hospital	V	N, some CT	N, InsT	UC	UC, unclear if DRFS includes deaths	N, InsT, CT, LN>3 NR	UC, assay methods unclear
Nitz 2017[10, 11, 14] N=2642	WSG-Plan B	V	N, some CT	N, MS	y	Y	Y, but high-risk	N, assay methods incompatible
Prat 2013[15]	GEICAM 9906	V	N, all CT	UC	UC	Y	N,	Y
Rohan, 2014[16] N=295 (147 cases; 148 controls)	Kaiser Permanente Northwest	V	N, Case control with some CT	N, InsT, MS, MC	Y	UC, unclear if deaths censored or an event	N, InsT, CT, LN>3 NR	N, some assay methods different
Stephen, 2014[17] a) BCS N=831 b) TEAM N=2513	a) BCS b) TEAM	V	Y, consecutive cohort; reanalysis of RCT	N, MS, InsT, MD	UC	Y	UC, (HER2 NR; LN>3 NR)	Y
TransATAC N=1048	TransATAC	D	Y, reanalysis of RCT	N, InsT, MS	UC	Y	N, InsT, MS	Y
Viale 2013[18]	IES	V	N, some with CT	UC	UC	UC, unclear if deaths censored or an event	N, CT, % LN>3 NR, % HER2- NR	UC, assay methods unclear
Vincent-Salomon, 2013[42] N=105	Institut Curie	V	Y, Cohort	N, InsT, MS	UC	Y	N, InsT, MS	Y

V, validation; N, no, high risk of biase; UC unclear risk of bias; Y, yes, low risk of bias; NR, not reported; MS, missing samples; InsT, insufficient tissue; MS, missing sample; MD, missing data; CT, chemotherapy; MC, no eligible control; BCS, Edinburgh Breast Conservation Series ; SYSMH, Sun Yat-sen Memorial Hospital; CCSYSU, Cancer Centre of Sun Yat-sen University; 3rdHNC, Third Hospital of Nanchang City

Table 5: Prognostic performance of IHC4: RFS

Reference; N	Cohorts	Population	Nodal status	ET/CT	% pts per group			RFS: HR (95% CI) unless stated otherwise
					Low	Inter	High	0-5 yr
LN0/+, 100% ET, some CT								
Bartlett 2016[20] Christiansen 2012[8] N=2919[20] N=4598[8]	TEAM	100% HR+ % HER2- NR	LN0/+, % NR	100% ET Some CT, % NR [30]	Used Quartiles			**8 year (n=2919):** continuous: 1.008 (1.006, 1.009, p<0.001) [20] Quartiles: p<0.001[20] Q1 vs Q2: p=0.11[20] **Yr NR (n=4598):** continuous: 1.008 (1.007, 1.010)[8]
Retrospective studies: Uncertain generalisability to UK context **LN0/LN+, some ET&CT**								
Lin, 2015[13] N=605	National Taiwan University Hospital	HR+ NR 76.2% HER2-	Any LN, % NR	ET NR 74.6% CT	Used Quartiles			**High vs. low:** [a] 2.33 (1.41, 3.85) **Intermediate vs. low:** [a] 1.88 (1.18, 2.99)

Pts per grp; patient per group; HR+, hormone receptor positive; HER2-, human epidermal growth factor receptor negative; NR, not reported; Q1, first quartile (0.-25%); Q2, second quartile (26-50%); RFS, relapse free survival; NR, not reported; LN, lymph node; ET, endocrine therapy; CT, chemotherapy; HR, hazard ratio; CI, confidence; yr, year

[a] High defined as above 75th percentile; low defined as below 25th percentile; intermediate 25th to 75th percentile.

Table 6: Prognostic performance of IHC4: IDFS

Reference; N	Cohorts	Population	Nodal status	ET/CT	Test or comp.	% pts per grp	Other analyses
LN0/+, 100% ET, some CT							
Nitz 2017[10, 11, 14] N=2642	WSG-Plan B	100% HR+ 100% HER2- High clinical risk 100% female	LN0-3 LN0 58.8% LN1-3 41.2%	RS<12 endo only; RS≥12, chemo + endo	**IHC4**	Used quartiles	**0-5 yr: HR 100th-75th to 0-25th percentil**e: 2.04 (95% CI: 1.47, 2.83, p<0.001)
LN+, ET NR, 100% CT							
Gluz, 2016c[9] N=459	WSG-AGO-Doc[39]	100% HR+ 100% HER2-	LN1-3	% ET NR 100% CT	**IHC4**	Used quartiles	**0-5 yr: HR 100th-75th to 0-25th percentil**e: 2.12 (95% CI: 1.32, 3.42, p 0.002)

Pts per grp; patient per group; RS, recurrence score; HR+, hormone receptor positive; HER2, human epidermal growth factor receptor; LN, lymph node; ET, endocrine therapy; CT, chemotherapy; HR, hazard ratio; CI, confidence interval; yr, year

Table 7: Prognostic performance of IHC4: IDFI

Reference; N	Cohorts	Population	Nodal status	ET/CT	Test or comp.	% pts per grp	IDFI: HR (95% CI, p)
LN0, some ET, 0% CT							
Vincent-Salomon, 2013[42] N=105	Institut Curie	100% ER+ 100% HER2- <3cm	LN0 100%	9.5% ET 0% CT	**IHC3**	NR	**HR continuous:** 1.01 (1.00, 1.01, p=0.204)
Pts per grp; patient per group; IDFI, invasive disease free survival; HR+, hormone receptor positive; HER2, human epidermal growth factor receptor; LN, lymph node; ET, endocrine therapy; CT, chemotherapy; HR, hazard ratio; CI, confidence interval; NR not reported							

Table 8: Prognostic performance of IHC4+C: OS

Reference; N	Cohorts	Population	Nodal status	ET/CT	% pts per group			% OS risk: 0-5 yr			% OS risk: 0-10 yr			OS: HR (95% CI)	
					Low	Inter	High	Low	Inter	High	Low	Inter	High	0-5 yr	0-10 yr
LN0; LN+ subgroups, 100%ET, 0% CT															
TransATAC[41] N=1005	TransATAC	100% ER+ HER2-	LN0	100% ET 0% CT	70 [a]	21 [a]	9 [a]	95.6	85.9	86.8	83.3	63.7	59.7	Inter vs. low: 3.40 (1.92, 6.02) [b] High vs. low: 3.18 (1.52, 6.65) [b]	Inter vs. low: 2.41 (1.71, 3.37) [b] High vs. low: 2.93 (1.91, 4.50) [b]
			LN1-3		28 [a]	34 [a]	38 [a]	94.9	86.2	79.0	81.2	65.5	50.5	Inter vs. low: 2.82 (0.78, 10.24) [b] High vs. low: 4.55 (1.33, 15.52) [b]	Inter vs. low: 2.22 (1.06, 4.64) [b] High vs. low: 3.55 (1.77, 7.12) [b]
Pts per grp; patient per group; OS, overall survival; Yr, year; Endo, endocrine therapy; chemo, chemotherapy ET, endocrine therapy; CT, chemotherapy; LN, lymph node; HR, hazard ratio; CI, confidence interval [a] These analyses used a cut off of <10% risk, 10-20% and >20% risk to define low, intermediate and high-risk groups; [b] this data from the reduced data set															

REFERENCES

1. 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. *J Clin Oncol* 2011;29:4273-8.
2. 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;28:1829-34.
3. Dowsett M, Allred C, Knox J, Quinn E, Salter J, Wale C*, et al.* Relationship between quantitative estrogen and progesterone receptor expression and human epidermal growth factor receptor 2 (HER-2) status with recurrence in the Arimidex, Tamoxifen, Alone or in Combination trial. *J Clin Oncol* 2008;26:1059-65.
4. Zabaglo L, Salter J, Anderson H, Quinn E, Hills M, Detre S*, et al.* Comparative validation of the SP6 antibody to Ki67 in breast cancer. *J Clin Pathol* 2010;63:800-4.
5. Elliott K, McQuaid S, Salto-Tellez M, Maxwell P. Immunohistochemistry should undergo robust validation equivalent to that of molecular diagnostics. *J Clin Pathol* 2015:jclinpath-2015-203178.
6. Goldstein NS, Hewitt SM, Taylor CR, Yaziji H, Hicks DG, Standardization MoA-HCOI. Recommendations for improved standardization of immunohistochemistry. *Appl Immunohistochem Mol Morphol* 2007;15:124-33.
7. Dodson A, Zabaglo L, Yeo B, Miller K, Smith I, Dowsett M. Risk of recurrence estimates with IHC4+C are tolerant of variations in staining and scoring: an analytical validity study. *J Clin Pathol* 2016;69:128-35.
8. Christiansen J, Bartlett JMS, Gustavson M, Rimm D, Robson T, Van De Velde CJH*, et al.* Validation of IHC4 algorithms for prediction of risk of recurrence in early breast cancer using both conventional and quantitative IHC approaches. *J Clin Oncol Conf* 2012;30.
9. Gluz O, Liedtke C, Huober J, Peyro-Saint-Paul H, Kates RE, Kreipe HH*, et al.* Comparison of prognostic and predictive impact of genomic or central grade and immunohistochemical subtypes or IHC4 in HR+/HER2- early breast cancer: WSG-AGO EC-Doc Trial. *Ann Oncol* 2016;27:1035-40.
10. Gluz O, Nitz U, Chrlstgen M, Kates RE, Clemens M, Kraemer S*, et al.* Prognostic impact of 21 gene recurrence score, IHC4, and central grade in high-risk HR+/HER2-early breast cancer (EBC): 5-year results of the prospective Phase III WSG PlanB trial. *J Clin Oncol Conf* 2016a;34.
11. Gluz O, Nitz UA, Christgen M, Kates RE, Shak S, Clemens M*, et al.* West German Study Group Phase III PlanB trial: First prospective outcome data for the 21-gene recurrence score assay and concordance of prognostic markers by central and local pathology assessment. *J Clin Oncol* 2016b;34:2341-9.
12. Gong C, Tan W, Chen K, You N, Zhu S, Liang G*, et al.* Prognostic value of a BCSC-associated microRNA signature in hormone receptor-positive HER2-negative breast cancer. *EBioMedicine* 2016;11:199-209.
13. Lin CH, Chen IC, Huang CS, Hu FC, Kuo WH, Kuo KT*, et al.* TP53 mutational analysis enhances the prognostic accuracy of IHC4 and PAM50 assays. *Scientific Reports* 2015;5:17879.
14. Nitz U, Gluz O, Christgen M, Kates RE, Clemens M, Malter W*, et al.* Reducing chemotherapy use in clinically high-risk, genomically low-risk pN0 and pN1 early breast cancer patients: five-year data from the prospective, randomised phase 3 West German Study Group (WSG) PlanB trial. *Breast Cancer Res Treat* 2017; 10.1007/s10549-017-4358-6.
15. Prat A, Cheang MC, Martin M, Parker JS, Carrasco E, Caballero R*, et al.* Prognostic significance of progesterone receptor-positive tumor cells within immunohistochemically defined luminal A breast cancer. *J Clin Oncol* 2013;31:203-9.

16. Rohan TE, Xue X, Lin HM, D'Alfonso TM, Ginter PS, Oktay MH*, et al.* Tumor microenvironment of metastasis and risk of distant metastasis of breast cancer. *J Natl Cancer Inst* 2014;106.

17. Stephen J, Murray G, Cameron DA, Thomas J, Kunkler IH, Jack W*, et al.* Time dependence of biomarkers: non-proportional effects of immunohistochemical panels predicting relapse risk in early breast cancer. *Br J Cancer* 2014;111:2242-7.

18. Viale G, Speirs V, Bartlett JM, Mousa K, Kalaitzaki E, Palmieri C*, et al.* Pr prognostic and predictive value of IHC4 and erb1 in the intergroup exemestane study (IES)-on behalf of the pathies investigators. *Ann Oncol* 2013;24:iii29-iii30.

19. Vincent-Salomon A, Benhamo V, Gravier E, Rigaill G, Gruel N, Robin S*, et al.* Genomic instability: a stronger prognostic marker than proliferation for early stage luminal breast carcinomas. *PLoS ONE* 2013;8:e76496.

20. Bartlett JM, Christiansen J, Gustavson M, Rimm DL, Piper T, van de Velde CJ*, et al.* Validation of the IHC4 breast cancer prognostic algorithm using multiple approaches on the multinational TEAM clinical trial. *Arch Pathol Lab Med* 2016;140:66-74.

21. Ward S, Scope A, Rafia R, Pandor A, Harnan S, Evans P*, et al.* Gene expression profiling and expanded immunohistochemistry tests to guide the use of adjuvant chemotherapy in breast cancer management: a systematic review and cost-effectiveness analysis. *Health Technol Assess* 2013;17:1-302.

22. Harbeck N, Sotlar K, Wuerstlein R, Doisneau-Sixou S. Molecular and protein markers for clinical decision making in breast cancer: today and tomorrow. *Cancer Treatment Reviews* 2014;40:434-44.

23. Hayes DF. Clinical utility of genetic signatures in selecting adjuvant treatment: Risk stratification for early vs. late recurrences. *Breast* 2015;24 Suppl 2:S6-S10.

24. Institut für Qualität und Wirtschaftlichkeit im G. Biomarker-based tests for the decision for or against adjuvant systemic chemotherapy in primary breast cancer. 2016.

25. Dowsett M, Nielsen TO, A'Hern R, Bartlett J, Coombes RC, Cuzick J*, et al.* Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *J National Cancer Institute* 2011;103:1656-64.

26. Engelberg JA, Retallack H, Balassanian R, Dowsett M, Zabaglo L, Ram AA*, et al.* "Score the Core" Web-based pathologist training tool improves the accuracy of breast cancer IHC4 scoring. *Hum Path* 2015;46:1694-704.

27. Leung SC, Nielsen TO, Zabaglo L, Arun I, Badve SS, Bane AL*, et al.* Analytical validation of a standardized scoring protocol for Ki67: phase 3 of an international multicenter collaboration. *NPJ Breast Cancer* 2016;2:16014.

28. Polley M-YC, Leung SC, Gao D, Mastropasqua MG, Zabaglo LA, Bartlett JM*, et al.* An international study to increase concordance in Ki67 scoring. *Mod Pathol* 2015;28:778-86.

29. Polley M-YC, Leung SC, McShane LM, Gao D, Hugh JC, Mastropasqua MG*, et al.* An international Ki67 reproducibility study. *J Natl Cancer Inst* 2013;105:1897-906.

30. Bartlett JM, Brookes CL, Robson T, van de Velde CJ, Billingham LJ, Campbell FM*, et 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. *J Clin Oncol* 2011;29:1531-8.

31. Balassanian R, Engelberg JA, Bishop JW, Borowsky AD, Cardiff RD, Carpenter PM*, et al.* Harmonization of immunohistochemical stains for breast cancer biomarkers-an athena pathology collaboration. *Lab Invest* 2013;93:29A.

32. Bishop JW, Engelberg J, Apple S, Balassanian R, Borowsky AD, Cardiff RD*, et al.* Raising the bar: Breast cancer biomarkers IHC4 harmonization from University of California-Athena pathology collaboration. *J Clin Oncol Conf: ASCO's Quality Care Symposium* 2012;30.

33. Borowsky A, Balassanian R, Yau C, Engelberg JA, Thompson CK, Retallack HEG*, et al.* Interobserver agreement of breast cancer IHC4 after "score the core" training. *Lab Invest* 2016;96:33A.

34. Dodson A, Zabaglo L, Martins V, Yeo B, Hayes D, McShane L*, et al.* Between-lab variability in Ki67 scoring by a standardised method in core-cuts has little impact on risk estimates by the

IHC4+Clinical (IHC4+C) Score. A study presented on behalf of the International Ki67 in Breast Cancer Working Group of the Breast International Group. *Eur J Cancer* 2016;57:S142-S3.

35. Bartlett JM, Ibrahim M, Jasani B, Morgan JM, Ellis I, Kay E*, et al.* External quality assurance of HER2 FISH and ISH testing: three years of the UK national external quality assurance scheme. *Am J Clin Pathol* 2009;131:106-11.

36. Welsh AW, Moeder CB, Kumar S, Gershkovich P, Alarid ET, Harigopal M*, et al.* Standardization of estrogen receptor measurement in breast cancer suggests false-negative results are a function of threshold intensity rather than percentage of positive cells. *J Clin Oncol* 2011;29:2978-84.

37. Bartlett J, Going JJ, Mallon EA, Watters AD, Reeves JR, Stanton P*, et al.* Evaluating HER2 amplification and overexpression in breast cancer. *J Pathol* 2001;195:422-8.

38. Cuzick J, Dowsett M, Wale C, Salter J, Quinn E, Zabaglo L*, et al.* Prognostic value of combined ER, PgR, Ki67, ER2 immunohistochemical score (IHC4) and comparison with the GHI recurrence score in early breast cancer. *J Clin Oncol* 2011.

39. Nitz U, Gluz O, Huober J, Kreipe H, Kates R, Hartmann A*, et al.* Final analysis of the prospective WSG-AGO EC-Doc versus FEC phase III trial in intermediate-risk (pN1) early breast cancer: efficacy and predictive value of Ki67 expression. *Ann Oncol* 2014;25:1551-7.

40. Prat A, Cheang MCU, Martín M, Parker JS, Carrasco E, Caballero R*, et al.* Prognostic significance of progesterone receptor–positive tumor cells within immunohistochemically defined luminal A breast cancer. *J Clin Oncol* 2012;31:203-9.

41. Sestak I, Dowsett M, Cuzick J. NICE request - TransATAC data analysis. In; 2017.

42. Vincent-Salomon A, Benhamo V, Gravier E, Rigaill G, Gruel N, Robin S*, et al.* Genomic instability: a stronger prognostic marker than proliferation for early stage luminal breast carcinomas. *PLoS ONE [Electronic Resource]* 2013;8:e76496.

43. Buus R, Sestak I, Kronenwett R, Denkert C, Dubsky P, Krappmann K*, et al.* Comparison of EndoPredict and EPclin with Oncotype DX recurrence score for prediction of risk of distant recurrence after endocrine therapy. *J Natl Cancer Inst* 2016;108.

44. Dowsett M, Sestak I, Lopez-Knowles E, Sidhu K, Dunbier AK, Cowens JW*, 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;31:2783-90.

45. Sestak I, Buus R, Cuzick J, Dubsky P, Kronenwett R, Ferree S*, et al.* Comprehensive comparison of prognostic signatures for breast cancer recurrence in TransATAC. San Antonio Breast Cancer Symposium, abstract no. 5773.

46. Sestak I, Cuzick J, Dowsett M, Lopez-Knowles E, Filipits M, Dubsky P*, et al.* Prediction of late distant recurrence after 5 years of endocrine treatment: a combined analysis of patients from the Austrian breast and colorectal cancer study group 8 and arimidex, tamoxifen alone or in combination randomized trials using the PAM50 risk of recurrence score. *J Clin Oncol* 2015;33:916-22.

47. Sestak I, Dowsett M, Zabaglo L, Lopez-Knowles E, Ferree S, Cowens JW*, et al.* Factors predicting late recurrence for estrogen receptor-positive breast cancer. *J Natl Cancer Inst* 2013;105:1504-11.

48. Sestak I, Zhang Y, Schroeder BE, Schnabel CA, Dowsett M, Cuzick J*, et al.* Cross-Stratification and Differential Risk by Breast Cancer Index and Recurrence Score in Women with Hormone Receptor-Positive Lymph Node-Negative Early-Stage Breast Cancer. *Clin Cancer Res* 2016b;22:5043-8.

49. 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;14:1067-76.

50. Tang G, Cuzick J, Costantino JP, Dowsett M, Forbes JF, Crager M*, et al.* Risk of recurrence and chemotherapy benefit for patients with node-negative, estrogen receptor-positive breast cancer: recurrence score alone and integrated with pathologic and clinical factors. *J Clin Oncol* 2011b;29:4365-72.

51. Van de Velde CJ, Rea D, Seynaeve C, Putter H, Hasenburg A, Vannetzel J-M*, et al.* Adjuvant tamoxifen and exemestane in early breast cancer (TEAM): a randomised phase 3 trial. *The Lancet* 2011;377:321-31.

52. Coombes RC, Hall E, Gibson LJ, Paridaens R, Jassem J, Delozier T*, et al.* A randomized trial of exemestane after two to three years of tamoxifen therapy in postmenopausal women with primary breast cancer. *N Engl J Med* 2004;350:1081-92.

53. Harbeck N, Gluz O, Clemens MR, Malter W, Reimer T, Nuding B*, et al.* Prospective WSG phase III PlanB trial: Final analysis of adjuvant 4xEC→4x doc vs. 6x docetaxel/cyclophosphamide in patients with high clinical risk and intermediate-to-high genomic risk HER2-negative, early breast cancer. *J Clin Oncol* 2017;35:504-.

54. Martín M, Rodríguez-Lescure Á, Ruiz A, Alba E, Calvo L, Ruiz-Borrego M*, et al.* Randomized phase 3 trial of fluorouracil, epirubicin, and cyclophosphamide alone or followed by Paclitaxel for early breast cancer. *J Natl Cancer Inst* 2008;100:805-14.