Diagnostics Assessment Report commissioned by the NIHR HTA Programme on behalf of the National Institute for Health and Care Excellence

A rapid evidence review of the analytical validity of IHC4: ADDENDUM to "Tumour profiling tests to guide adjuvant chemotherapy decisions in people with breast cancer (update of DG10)"

Produced by	School of Health and Related Research (ScHARR), The University of Sheffield
Authors	Sue Harnan, Senior Research Fellow, ScHARR, University of Sheffield
	Paul Tappenden, Reader in Health Economic Modelling, ScHARR, University of Sheffield
	Katy Cooper, Senior Research Fellow, ScHARR, University of Sheffield
	John Stevens, Reader in Decision Science, ScHARR, University of Sheffield
	Alice Bessey, Research Associate, ScHARR, University of Sheffield
	Rachid Rafia, Research Fellow, ScHARR, University of Sheffield
	Sue Ward, Senior Research Fellow, ScHARR, University of Sheffield
	Ruth Wong, Information Specialist, ScHARR, University of Sheffield
	Robert C Stein, Consultant Oncologist and Professor of Breast Oncology, NIHR University College Hospitals Biomedical Research Centre, London and Research Department of Oncology, UCL, London
	Janet Brown, Professor of Medical Oncology, Department of Oncology and Metabolism, University of Sheffield
Correspondence Author	Paul Tappenden, Reader in Health Economic Modelling, ScHARR, University of Sheffield, Sheffield, UK
Date completed	Date completed 27th October

Source of funding: This report was commissioned by the NIHR HTA Programme as project number 16/30/03.

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,[1 2] 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[3] 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.[3] 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.[4]

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[3 5-17]
- The reference lists of studies included or cited in existing systematic or expert reviews[18-21]
- 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[22]: International Ki67 in Breast Cancer Working Group recommendations
- Dodson 2016[4]: IHC4 analytical validity study.
- Engelberger 2015[23]: "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:[24-26] Ki67 analytical validity studies resulting from the International Ki67 in Breast Cancer Working Group[22]. 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[3 5-16 27] 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[22] and 7 primary studies[4 23-26 28-31]) were included (Table 1). These are broadly split into:

1. **Analytical validity of IHC4 between Royal Marsden and external centres**
2. **Analytical validity of IHC4 within other centres**
3. **Analytical validity of Ki67: Studies related to Ki67 Working Group and Royal Marsden**

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

Dodson et al. 2016[4]

Methods: This study[4] (N=28) originated from the Royal Marsden Hospital (London, UK) and conducted two main assessments (Table XX). 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.[32]

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[23]

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[28] & Bishop 2012[29] (see below), now termed "Score the Core" (STC). In Engelberg 2015[23], 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.

2. Analytical validity of IHC4 within other centres

Evidence from the main review

None of the prognostic studies identified by the main review[3 5-16 27] 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 section XXX 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.[3] .

Balassanian 2013[28] & Bishop 2012[29]

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[28] 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[29] (assumed to be part of, or an extension of, the same study), a similar (or the same) experiment as reported in Balassanian et al.[28] 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.[28], 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[30]

This study used the "Score the Core" training, as developed and used in Balassanian 2013[28] & Bishop 2012[29] and Engelberg 2015[23] 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.

3. 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[4] 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[22] 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,[24-26] reported below.

Polley et al. (2013)[26]

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)[25]

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)[24]

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[31]) for the Leung 2016[24] 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[28] 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),[30] 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,[4] Engleberg 2015[23] (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[28]& Bishop 2012.[29]

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[26] do not appear to have been addressed in the selected literature, probably as the original Working Group[22] 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).[17]

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 1Study 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)[4]	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)[23]	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						

Reference	Targets	Topic	Samples/setting	Experimental variable	Findings	Conclusions
Balassanian 2013 (CA)[28] Bishop 2012 (CA)[29]	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)[30]	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)

Reference	Targets	Topic	Samples/setting	Experimental variable	Findings	Conclusions
3. Analytical valdidty of Ki67: Studies related to Ki67 Working Group and RMH						
Dowsett 2011 (recommendations from Ki67 working group)[22]	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[26] (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)

Reference	Targets	Topic	Samples/setting	Experimental variable	Findings	Conclusions
Polley 2015[25] (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) "Although these data are potentially encouraging, suggesting that it may be possible to standardize scoring of Ki67 among pathology laboratories, clinically important discrepancies persist. Before this biomarker could be recommended for clinical use, future research will need to extend this approach 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[24] (full paper) Dodson 2016b (CA)[31]	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) [31] "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)

Reference	Targets	Topic	Samples/setting	Experimental variable	Findings	Conclusions
RMH, Royal Marsden Hosptial; ER, oestrogen receptor; PR, Progesterone receptor; HER2, human epidermal growth factor receptor 2; IHC, immunohistochemistry; CA conference abstract						

1. Elliott K, McQuaid S, Salto-Tellez M, et al. Immunohistochemistry should undergo robust validation equivalent to that of molecular diagnostics. *Journal of clinical pathology* 2015:jclinpath-2015-203178.
2. Goldstein NS, Hewitt SM, Taylor CR, et al. Recommendations for improved standardization of immunohistochemistry. *Applied Immunohistochemistry & Molecular Morphology* 2007;15(2):124-33.
3. Cuzick J, Dowsett M, Pineda S, et al. Prognostic value of a combined estrogen receptor, progesterone receptor, Ki-67, and human epidermal growth factor receptor 2 immunohistochemical score and comparison with the Genomic Health recurrence score in early breast cancer. *Journal of Clinical Oncology* 2011;29(32):4273-78. doi: https://dx.doi.org/10.1200/JCO.2010.31.2835
4. Dodson A, Zabaglo L, Yeo B, et al. Risk of recurrence estimates with IHC4+C are tolerant of variations in staining and scoring: an analytical validity study. *Journal of Clinical Pathology* 2016;69(2):128-35. doi: https://dx.doi.org/10.1136/jclinpath-2015-203212
5. Christiansen J, Bartlett JMS, Gustavson M, et al. Validation of IHC4 algorithms for prediction of risk of recurrence in early breast cancer using both conventional and quantitative IHC approaches. *Journal of Clinical Oncology Conference* 2012;30(15 SUPPL. 1)
6. Gluz O, Liedtke C, Huober J, 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. *Annals of Oncology* 2016;27(6):1035-40. doi: https://dx.doi.org/10.1093/annonc/mdw070
7. Gluz O, Nitz U, Chrlstgen M, 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. *Journal of Clinical Oncology Conference* 2016a;34(no pagination)
8. Gluz O, Nitz UA, Christgen 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. *Journal of Clinical Oncology* 2016b;34(20):2341-49. doi: https://dx.doi.org/10.1200/JCO.2015.63.5383
9. Gong C, Tan W, Chen K, et al. Prognostic value of a BCSC-associated microRNA signature in hormone receptor-positive HER2-negative breast cancer. *EBioMedicine* 2016;11:199-209. doi: https://dx.doi.org/10.1016/j.ebiom.2016.08.016
10. Lin CH, Chen IC, Huang CS, et al. TP53 mutational analysis enhances the prognostic accuracy of IHC4 and PAM50 assays. *Scientific Reports* 2015;5:17879. doi: https://dx.doi.org/10.1038/srep17879
11. Nitz U, Gluz O, Christgen M, 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 Research and Treatment* 2017 doi: 10.1007/s10549-017-4358-6
12. Prat A, Cheang MC, Martin M, et al. Prognostic significance of progesterone receptor-positive tumor cells within immunohistochemically defined luminal A breast cancer. *Journal of Clinical Oncology* 2013;31(2):203-09. doi: https://dx.doi.org/10.1200/JCO.2012.43.4134
13. Rohan TE, Xue X, Lin HM, et al. Tumor microenvironment of metastasis and risk of distant metastasis of breast cancer. *Journal of the National Cancer Institute* 2014;106(8) doi: https://dx.doi.org/10.1093/jnci/dju136
14. Stephen J, Murray G, Cameron DA, et al. Time dependence of biomarkers: non-proportional effects of immunohistochemical panels predicting relapse risk in early breast cancer. *British Journal of Cancer* 2014;111(12):2242-47. doi: https://dx.doi.org/10.1038/bjc.2014.530
15. Viale G, Speirs V, Bartlett JM, et al. Pr prognostic and predictive value of IHC4 and erb1 in the intergroup exemestane study (IES)-on behalf of the pathies investigators. *Annals of Oncology* 2013;24:iii29-iii30. doi: http://dx.doi.org/10.1093/annonc/mdt078
16. Vincent-Salomon A, Benhamo V, Gravier E, et al. Genomic instability: a stronger prognostic marker than proliferation for early stage luminal breast carcinomas. *PLoS ONE* 2013;8(10):e76496. doi: https://dx.doi.org/10.1371/journal.pone.0076496

17. Bartlett JM, Christiansen J, Gustavson M, et al. Validation of the IHC4 breast cancer prognostic algorithm using multiple approaches on the multinational TEAM clinical trial. *Archives of Pathology and Laboratory Medicine* 2016;140(1):66-74. doi: 10.5858/arpa.2014-0599-OA [published Online First: 2015/12/31]
18. Ward S, Scope A, Rafia R, 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 Technology Assessment* 2013;17(44):1-302. doi: 10.3310/hta17440 [published Online First: 2013/10/04]
19. Harbeck N, Sotlar K, Wuerstlein R, et al. Molecular and protein markers for clinical decision making in breast cancer: today and tomorrow. *Cancer Treatment Reviews* 2014;40(3):434-44. doi: https://dx.doi.org/10.1016/j.ctrv.2013.09.014
20. 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. doi: https://dx.doi.org/10.1016/j.breast.2015.07.002
21. Institut für Qualität und Wirtschaftlichkeit im G. Biomarker-based tests for the decision for or against adjuvant systemic chemotherapy in primary breast cancer. 2016
22. Dowsett M, Nielsen TO, A'Hern R, et al. Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *Journal of the National Cancer Institute* 2011;103(22):1656-64.
23. Engelberg JA, Retallack H, Balassanian R, et al. "Score the Core" Web-based pathologist training tool improves the accuracy of breast cancer IHC4 scoring. *Human Pathology* 2015;46(11):1694-704. doi: https://dx.doi.org/10.1016/j.humpath.2015.07.008
24. Leung SC, Nielsen TO, Zabaglo L, et al. Analytical validation of a standardized scoring protocol for Ki67: phase 3 of an international multicenter collaboration. *NPJ Breast Cancer* 2016;2:16014.
25. Polley M-YC, Leung SC, Gao D, et al. An international study to increase concordance in Ki67 scoring. *Modern Pathology* 2015;28(6):778-86.
26. Polley M-YC, Leung SC, McShane LM, et al. An international Ki67 reproducibility study. *Journal of the National Cancer Institute* 2013;105(24):1897-906.
27. Bartlett JM, Brookes CL, Robson T, 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. *Journal of Clinical Oncology* 2011;29(12):1531-38.
28. Balassanian R, Engelberg JA, Bishop JW, et al. Harmonization of immunohistochemical stains for breast cancer biomarkers-an athena pathology collaboration. *Laboratory Investigation* 2013;93:29A. doi: http://dx.doi.org/10.1038/labinvest.2013.14
29. Bishop JW, Engelberg J, Apple S, et al. Raising the bar: Breast cancer biomarkers IHC4 harmonization from University of California-Athena pathology collaboration. *Journal of Clinical Oncology Conference: ASCO's Quality Care Symposium* 2012;30(34 SUPPL. 1)
30. Borowsky A, Balassanian R, Yau C, et al. Interobserver agreement of breast cancer IHC4 after "score the core" training. *Laboratory Investigation* 2016;96:33A. doi: http://dx.doi.org/10.1038/labinvest.2016.3
31. Dodson A, Zabaglo L, Martins V, 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. *European Journal of Cancer* 2016;57:S142-S43.
32. Bartlett JM, Ibrahim M, Jasani B, et al. External quality assurance of HER2 FISH and ISH testing: three years of the UK national external quality assurance scheme. *American Journal of Clinical Pathology* 2009;131(1):106-11.