UK NEQAS ICC & ISH Ki-67 Data Reveal Differences in Performance of Primary Antibody Clones

Suzanne Parry, MSc, Mitch Dowsett, PhD, FMedSci,† and Andrew Dodson, MPhil**

Abstract: We examined data from 374 laboratories staining for Ki-67 as part of external quality assessment over 8 runs between 2013 and 2017 (total data sets = 2601). One of 5 primary antibodies was used for 94.8% of submissions, with MIB-1 (Agilent Dako) comprising 58.8% of the total. Examining assessment score as a continuous variable showed the 30-9 (Ventana) and K2 (Leica Biosystems) clones were associated with the highest mean scores (17.0; 95% confidence interval, 16.8-17.2 and 16.3; 95% confidence interval, 15.9-16.6, respectively). Stain quality was not significantly different between them. Both were associated with significantly better staining compared with MIB-1 (Agilent Dako), MM1 (Leica Biosystems), and SP6 from various suppliers ($P < 0.05$). Similarly, categorical assessment of "Good" versus "Not good" staining quality showed that the 30-9 and K2 clones were both significantly associated with "Good" staining (both $P < 0.001$). Other methodological parameters were examined for significant primary antibody-specific effects; none were seen for 30-9, K2, or SP6. The MM1 clone was more likely to be associated with good quality staining when it was used with Leica Biosystems sourced antigen retrieval, detection, and platform, all statistically significant at $P < 0.01$. MIB-1 was more likely to be associated with good quality staining results when it was used with Agilent Dako antigen retrieval, detection, and staining platforms ($P < 0.0001$), and less likely at the same significance level when used with Leica Biosystems reagents and equipment. The data presented here show the importance of not just primary antibody choice but also matching that choice to other methodological factors.

Received for publication June 8, 2020; accepted November 16, 2020.

From the *UK National External Quality Assessment Scheme for Immunocytochemistry and In-Situ Hybridisation; and †Ralph Lauren Centre for Breast Cancer Research, Royal Marsden Hospital, London, UK.

The authors declare no conflict of interest.

Reprints: Andrew Dodson, MPhil, UK National External Quality Assessment Scheme for Immunocytochemistry and In-Situ Hybridisation, Office 127, Finsbury Business Centre, 40 Bowling Green Lane, London EC1R 0NE, UK (e-mail: adodson@ukneqasiccish.org).

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website, www.appliedimmunohist.com.

Key Words: Ki-67, proliferation index, external quality assessment, immunohistochemistry, breast cancer

(*Appl Immunohistochem Mol Morphol* 2021;29:86–94)

Immunohistochemical (IHC) demonstration of Ki-67 protein provides information on proliferation status that aids prognosis and prediction, this being particularly true in breast pathology.[1–3] However, application of Ki-67 to breast cancer clinical management has been hampered by lack of standardization and reproducibility in scoring and staining methodologies.[4,5]

Successful strategies for increasing reproducibility of Ki-67 scoring have included the development and analytical validation of standardized manual scoring methods capable of producing intraobserver results that show very close agreement,[6–8] and use of digital image analysis.[9,10] Clinical validity of manual scoring has been shown in several studies, and most notably in the POETIC clinical trial which involved the assessment of Ki-67 in more than 4500 breast cancer patients.[11] Similarly, the demonstration of clinical validity for automated scoring has being subject to considerable research effort.[9,12]

Nonstandardized IHC staining methodologies contribute to variability of Ki-67 results. For example, a paper by Polley and colleagues found differences in the levels of agreement achieved when observers scored a centrally stained tissue microarray compared with locally stained ones. In the former situation the group achieved an intraclass correlation coefficient (ICC) of 0.71 [95% confidence interval (CI), 0.47-0.78], in the latter the ICC was much lower, at 0.59 (95% CI, 0.37-0.68). The difference between these 2 ICCs indicate ~10% of the variation in Ki-67 results is introduced by differing IHC methods.[13]

We examined the large body of data on Ki-67 IHC accumulated by the UK National External Quality Assessment Scheme for Immunocytochemistry and In-Situ Hybridisation (UK NEQAS ICC & ISH) in the course of its external quality assessments to determine whether any systematic differences in performance could be ascribed to specific components of the IHC methodology. In particular, we looked for difference in staining quality produced by the different commercially available Ki-67 primary antibody clones and also identified the major methodological factors in the IHC staining protocol that impacted significantly on that quality, with the intention of

presenting the results in sufficient detail to allow workers in the field to identify reagents and methodologies yielding optimal staining quality applicable to the reagents and staining platforms used in their own institutions.

MATERIALS AND METHODS

This study presents data gathered from 8 proficiency testing surveys examining the quality of Ki-67 demonstration by IHC conducted between July 2013 and April 2017.

At each assessment participants were provided with formalin-fixed paraffin-embedded sections of both breast cancer tissues and reactive tonsil, which they were required to stain for Ki-67 using their routine method. Stained slides were returned for central assessment by a panel of experts. Feedback on the quality of their own preparations and a comparative assessment of the entire peer-group was provided individually to each participating center to help them monitor, and where necessary, improve their staining.

Methodological data were obtained from each participating center, which included information about:

- nature of primary antibody,
- antigen retrieval method (if used),
- IHC detection system,
- automated staining platform (if any).

A nominal scoring scale, running between 4 and 20 was used to indicate IHC staining quality, with scores of 4 indicating poor staining and scores of 20, excellent staining (see below for full descriptions):

- Scores of 4 to 9 (fail): unreadable staining, which has no utility. Improvement is essential.
- Scores of 10 to 12 (borderline): suboptimal preparation, which is readable but may not be at the expected level of specificity or sensitivity. Improvement is essential.
- Scores of 13 to 15 (pass): adequate staining, which is readable but may not be at the expected level of sensitivity. Improvement is required.
- Scores of 16 to 20 (good): good to excellent demonstration of requested antigen at the expected level of sensitivity. If improvements are required, they are minor.

Data were collated in Microsoft Excel for Office 365 (version 1911; Microsoft, Redmond) and analyzed in GraphPad Prism (version 8.3.0; GraphPad, San Diego).

RESULTS

Descriptive Statistics

Eight assessment runs were carried out, 4 between July 2013 and April 2014, and 4 between July 2016 and April 2017. Median number of participating centers over the 8 runs was 326 (range, 299 to 348). In total, 2601 individual submissions were received, with 1398 (53.7%) coming from UK-based centers and 1203 (46.3%) from centers outside the United Kingdom. Overall, 374 centers made at least 1 submission; of those, 270 (72.2%) made submissions to all 8 runs. See Table 1 for the complete data set relating to primary antibody clones and other methodological parameters, which are discussed in more detail below.

Primary Antibodies

For the vast majority of submissions (>90%), the primary antibody was 1 of 5 clones, 3 were mouse monoclonal antibodies (K2, MM1, and MIB-1), and 2 were rabbit monoclonals (30-9 and SP6). With the exception of SP6, primary antibodies were almost exclusively obtained from a single commercial source; MIB-1 (Agilent Dako, Santa Clara, CA), which was used for 58.8% of submissions, 30-9 (Ventana Medical Systems Inc., Arizona) for 16.0%, MM1 (Leica Biosystems, Buffalo Grove, IL) for 10.1%, K2 (Leica Biosystems) for 5.5%, and SP6 (various commercial suppliers) for 4.4%.

In 1.2% of submissions a variety of other clones were used with each individually being used <100 times in total. For the remaining 4.0% of submissions the clone was not stated (Fig. 1A) illustrates all the usage data.

Each of the 5 major clones showed marked trends in their proportional usage over the course of the study. The proportion of participants using MIB-1 declined from a mean of 60.4% over the first 4 runs to 57.0% for the second 4. Similarly, the use of MM1 declined from 15.1% to 4.5%. In contrast, K2's use increased from 2.6% to 8.7%, becoming the most commonly used Leica clone. 30-9 also saw a large increase in the proportion of participants using it (10.3% in runs 1 to 4, 22.4% in runs 5 to 6). Figure 1B displays this data in more detail.

Other Methodological Parameters

More than 80% of submissions used 1 of 4 heat-mediated antigen retrieval (HMAR) methods. Most (40%) used Cell Conditioning 1 solution (CC1; Ventana Medical Systems Inc.). Epitope Retrieval 1 (ER1) and Epitope Retrieval (ER2) buffers (both Leica Biosystems), were used for 4.4% and 25.0% of submissions, respectively, and the PT-Link ancillary HMAR water-bath (Agilent Dako) was used for 14%.

Together 3 companies supplied almost 90% of the IHC detection system reagents used. Ventana supplied the BenchMark reagents that were used for 42.1% of submissions, a labelled polymer detection system from Leica Biosystems (Bond MAX Refine) accounted for 29.3% and 1 from Agilent Dako (FLEX), 16.1% of the remainder.

And finally, the same 3 companies supplied the 5 automated staining platforms used to stain almost 90% of submitted slides. These being, various versions of the Agilent Dako supplied Autostainer (15.9%), Leica Biosystems supplied BOND MAX (14.9%) and BOND-III (15.0%) machines, and Ventana BenchMark ULTRA (23.3%), and XT (19.1%) platforms.

It can be seen that 3 companies dominated the market. In the cases of Ventana and Leica Biosystems there was a very strong tendency within any 1 center for a single company to be used for the supply of primary antibody, HMAR buffers, detection system reagents, and staining platform. In contrast, the same close associations

TABLE 1. Data Characteristics

	n (%)							
	30-9	**K2**	**MIB-1**	**MM1**	**SP6**	**Not Stated**	**Various**	**Total**
Submissions	415 (16.0)	142 (5.5)	1530 (58.8)	264 (10.1)	115 (4.4)	103 (4.0)	32 (1.2)	2601 (100.0)
Primary antibody supplier								
Agilent Dako	0 (0.0)	0 (0.0)	1528 (99.9)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1528 (58.7)
Leica	0 (0.0)	142 (100.0)	0 (0.0)	257 (97.3)	0 (0.0)	0 (0.0)	0 (0.0)	399 (15.3)
Ventana	414 (99.8)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	414 (15.9)
Not stated	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	103 (100.0)	0 (0.0)	103 (4.0)
Various (use < 100)	1 (0.2)	0 (0.0)	2 (0.1)	7 (2.7)	115 (100.0)	0 (0.0)	32 (100.0)	157 (6.0)
HMAR method (supplier)								
CC1 buffer (Ventana)	399 (96.1)	7 (4.9)	539 (35.2)	20 (7.6)	54 (47.0)	43 (41.7)	6 (18.8)	1068 (41.1)
ER1 buffer (Leica)	0 (0.0)	8 (5.6)	67 (4.4)	37 (14.0)	2 (1.7)	0 (0.0)	0 (0.0)	114 (4.4)
ER2 buffer (Leica)	4 (1.0)	123 (86.6)	336 (22.0)	148 (56.1)	26 (22.6)	4 (3.9)	9 (28.1)	650 (25.0)
PT-Link (Agilent Dako)	0 (0.0)	0 (0.0)	341 (22.3)	13 (4.9)	0 (0.0)	2 (1.9)	8 (25.0)	364 (14.0)
Not stated	0 (0.0)	2 (1.4)	20 (1.3)	7 (2.7)	4 (3.5)	38 (36.9)	0 (0.0)	71 (2.7)
Various (use < 100)	12 (2.9)	2 (1.4)	227 (14.8)	39 (14.8)	29 (25.2)	16 (15.5)	9 (28.1)	334 (12.8)
Detection method (supplier)								
Bond MAX Refine (Leica)	4 (1.0)	133 (93.7)	395 (25.8)	188 (71.2)	28 (24.3)	4 (3.9)	9 (28.1)	761 (29.3)
FLEX (Agilent Dako)	0 (0.0)	0 (0.0)	397 (25.9)	11 (4.2)	0 (0.0)	2 (1.9)	9 (28.1)	419 (16.1)
OptiView (Ventana)	173 (41.7)	4 (2.8)	117 (7.6)	6 (2.3)	6 (5.2)	19 (18.4)	0 (0.0)	325 (12.5)
ultraView (Ventana)	232 (55.9)	0 (0.0)	441 (28.8)	9 (3.4)	48 (41.7)	34 (33.0)	6 (18.8)	770 (29.6)
Not stated	0 (0.0)	0 (0.0)	19 (1.2)	0 (0.0)	4 (3.5)	27 (26.2)	1 (3.1)	51 (2.0)
Various (use < 100)	6 (1.4)	5 (3.5)	161 (10.5)	50 (18.9)	29 (25.2)	17 (16.5)	7 (21.9)	275 (10.6)
Automation (supplier)								
Autostainer (Agilent Dako)	1 (0.2)	0 (0.0)	376 (24.6)	22 (8.3)	4 (3.5)	2 (1.9)	9 (28.1)	414 (15.9)
BenchMark ULTRA (Ventana)	252 (60.7)	7 (4.9)	280 (18.3)	12 (4.5)	31 (27.0)	22 (21.4)	3 (9.4)	607 (23.3)
BenchMark XT (Ventana)	133 (32.0)	0 (0.0)	291 (19.0)	8 (3.0)	30 (26.1)	31 (30.1)	3 (9.4)	496 (19.1)
BOND-III (Leica)	0 (0.0)	72 (50.7)	223 (14.6)	81 (30.7)	12 (10.4)	0 (0.0)	1 (3.1)	389 (15.0)
BOND MAX (Leica)	4 (1.0)	60 (42.3)	184 (12.0)	111 (42.0)	16 (13.9)	4 (3.9)	8 (25.0)	387 (14.9)
Not Stated	0 (0.0)	1 (0.7)	0 (0.0)	0 (0.0)	0 (0.0)	28 (27.2)	0 (0.0)	29 (1.1)
Various (use < 100)	25 (6.0)	2 (1.4)	176 (11.5)	30 (11.4)	22 (19.1)	16 (15.5)	8 (25.0)	279 (10.7)

Data describing primary antibody clone use and suppliers, and the characteristics relating to the principle components of the IHC staining methodologies employed. values in red indicates categories containing >80% of submissions for that antibody.

HMAR indicates heat-mediated antigen retrieval; n (%), count of submissions and proportion (percentage).

were not observed for the SP6 clone or for MIB-1. The MIB-1 clone in particular was used in conjunction with reagents and platforms from all commercial suppliers across the board.

Performance Data

A histogram showing the frequency distribution of assessment scores aggregated across all 8 runs is shown in Figure 2A. In total, 62.4% of submissions achieving a

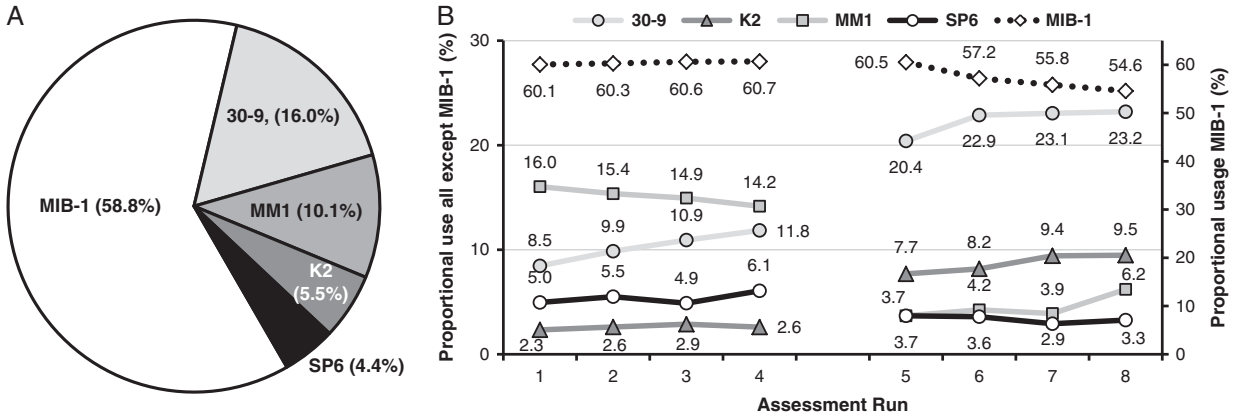

FIGURE 1. Primary antibody clone use. A, Overall primary antibody usage over the course of all 8 external quality assessment runs. B, Primary antibody usage over time, showing relative proportions of submissions occurring at each assessment run. MIB-1 has been displayed against a secondary (right-hand) axis. The gap in the *x*-axis indicates separation in time between runs 1 to 4 (July 2013 to April 2014) and runs 5 to 6 (July 2016 to April 2017). Data for primary antibodies that were individually used <100 times (1.2%) and where the clone was not stated (4.0%) have been omitted from both figures.

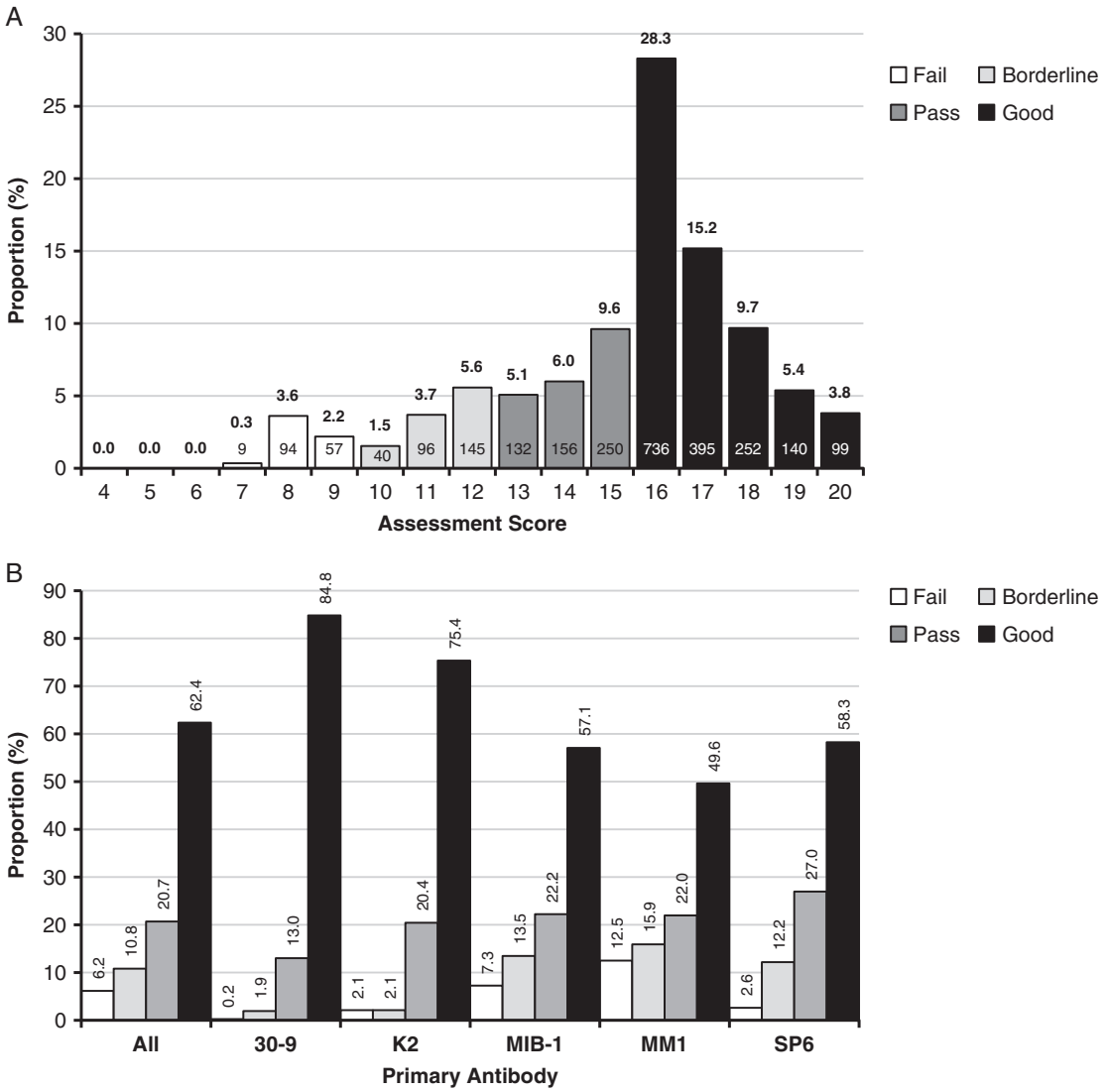

FIGURE 2. Primary antibody clone performance. A, Frequency distribution of assessment scores allocated over the course of all 8 external quality assessment runs. The figure at the top of each column indicates the proportion, the figure at the base, the count for each assessment score. B, Analysis of categorical assessment data. The figure at the top of each column indicates proportion. Data are shown for overall and individual primary antibody clone performances. In both (A) and (B), columns have been shaded to indicate the category: white = fail (4 to 9), light gray = borderline (10 to 12), dark gray = pass (13 to 15), black = good (16 to 20). Data for other and not stated categories have been omitted from both charts to aid legibility.

score at assessment indicative of good staining (score ≥ 16), and only 6.2% of submissions failing (score ≤ 9).

Primary Antibodies

Qualitative differences in performance of the individual primary antibody clones were seen when assessment score data were analyzed (Table 2). Submissions using 30-9 clone had the highest mean score (17.0; 95% CI, 16.8-17.2), whereas those using MM1 had the lowest (14.1; 95% CI, 13.8-14.5). The population of MIB-1 users attained a mean score intermediate between the 2 extremes (15.0; 95% CI, 14.8-15.1).

When it was considered as a continuous variable, the effect of primary antibody clone on assessment score was significant for the grouped comparison of all 5 primary antibodies (ANOVA test, $F_{4,2461} = 63.80$, $P < 0.0001$). Comparisons between clones showed the score means of 30-9 and K2 to be significantly higher than those of the other 3 clones, but not different from each other (Tukey multiple comparisons test, full details are given in Table 3).

Categorical assessment groups were also used to examine performance. The 30-9 clone produced the lowest proportion of fails (0.2%) and highest proportion of submissions assessed as "Good" (84.8%); in comparison, 12.5% of submissions failed at assessment and 49.6% were classified as "Good" when the MM1 clone was used. The remaining 3 antibody clones were associated with performance statistics intermediate between those of 30-9 and MM1. The data are illustrated for all clones in Figure 2B.

TABLE 2. Statistics Describing Performance at Assessment, Overall, and for Individual Primary Antibody Clones

	All	30-9	K2	MIB-1	MM1	SP6
No. values	2601	415	142	1530	264	115
Minimum	7	7	8	7	7	8
25% percentile	14	16	16	13	12	14
Median	16	17	16	16	15	16
75% percentile	17	18	17	17	16	16
Maximum	20	20	20	20	20	19
Range	13	13	12	13	13	11
Mean	15.3	17.0	16.3	15.0	14.1	15.1
SD	2.8	1.9	2.1	2.9	3.0	2.2
SE of mean	0.1	0.1	0.2	0.1	0.2	0.2
Lower 95% CI of mean	15.2	16.8	15.9	14.8	13.8	14.7
Upper 95% CI of mean	15.4	17.2	16.6	15.1	14.5	15.5

Data for other and not stated categories have been omitted.
CI indicates confidence interval.

Contingency table (2×2) were constructed for each primary antibody clone using assessment scores of 4 to 15 (all staining characterized as "Not Good") and scores of 16 to 20 ("Good" staining) to define the 2 categories. The 30-9 and K2 clones were both statistically significantly associated with "Good" staining (both $P < 0.001$); odds ratio for the 30-9 clone was 4.03 (95% CI, 3.06-5.34), and that for the K2 clone was 1.90 (95% CI, 1.30-2.84). In contrast, both MIB-1 and MM1 were significantly less likely to be associated with "Good" staining (both $P < 0.0001$); odds ratio for the MIB-1 clone was 0.57 (95% CI, 0.48-0.67), and that for the MM1 clone was 0.56 (95% CI, 0.43-0.72). The results for the SP6 clone were not statistically significant. The full data set is given in Table 4 and Figure 3A.

TABLE 3. Tukey Multiple Comparisons Test for Performance of Individual Primary Antibody Clones

		95% CI of Dif.			
	Mean Dif.	Lower	Upper	*P* Summary	Adjusted *P*
30-9 vs. K2	0.71	−0.01	1.43	NS	0.05
30-9 vs. MIB-1	2.03	1.62	2.44	****	< 0.0001
30-9 vs. MM1	2.85	2.27	3.43	****	< 0.0001
30-9 vs. SP6	1.88	1.10	2.65	****	< 0.0001
K2 vs. MIB-1	1.32	0.67	1.97	****	< 0.0001
K2 vs. MM1	2.15	1.38	2.91	****	< 0.0001
K2 vs. SP6	1.17	0.24	2.09	**	0.01
MIB-1 vs. MM1	0.82	0.33	1.32	****	< 0.0001
MIB-1 vs. SP6	−0.15	−0.87	0.56	NS	0.98
MM1 vs. SP6	−0.98	−1.80	-0.15	*	0.01

Data for other and not stated categories have been omitted.
CI indicates confidence interval; Dif., difference; NS, not significance.
*$P \leq 0.05$.
**$P \leq 0.01$.
****$P \leq 0.0001$.

TABLE 4. Results of Contingency Table Analyses of Categorical Staining for Each Primary Antibody Clone

Descriptor	30-9	K2	MIB-1	MM1	SP6
Fisher exact test					
P (2-sided)	< 0.0001	0.0009	< 0.0001	< 0.0001	0.3760
P summary	****	***	****	****	NS
Odds ratio					
Value	4.03	1.90	0.57	0.56	0.84
95% CI lower value	3.06	1.30	0.48	0.43	0.57
95% CI upper value	5.34	2.84	0.67	0.72	1.21
Data analyzed (count of submissions)					
All except good (4-15)	63	35	657	133	48
Good (16-20)	352	107	873	131	67
Total	415	142	1530	264	115
Data analyzed (% of submissions)					
All except good (4-15)	15.2	24.6	42.9	50.4	41.7
Good (16-20)	84.8	75.4	57.1	49.6	58.3

Results are for good staining category versus all other categories combined. Data for other and not stated categories have been omitted.
CI indicates confidence interval; NS, not significance.
***$P \leq 0.001$.
****$P \leq 0.0001$.

Other Methodological Parameters

For each primary antibody clone, the data were examined for associations between methodological parameters and performance at assessment. This was done in a similar way to primary antibody performance assessment, with contingency tables. Once again, an assessment score of 16 was used as a cut-point to differentiate good from all other staining quality.

No significant associations were present for clones 30-9, K2, or SP6.

The results for MM1 confirmed a positive association for HMAR using ER2 buffer, Bond MAX Refine detection, and use of the BOND-III staining platform. A negative association with the use of ultraView detection was shown, albeit at a significance of only $P = 0.0473$. The details of results relating to MM1 are given in Table 5 and are illustrated in Figure 3B.

MIB-1-specific contingency table results are given in Table 6. They show a remarkably consistent and highly significant pattern of associations, which is best appreciated if the methodological parameters are arranged according to their commercial supplier as in (Fig. 3C). Submissions made using Leica Biosystems staining reagents and staining automation had odds ratios around 0.3, indicating that they were 3 times less likely to achieve a score at assessment of ≥ 16 than the whole group. In contrast, those made using Ventana and Agilent Dako supplied reagents and equipment had odds ratios substantially higher than 1, indicating an improved chance of obtaining Good staining.

Reanalysis of primary antibody performance was done, considering the effects of methodological parameters and the results are illustrated in Figure 4 (the full set of results are

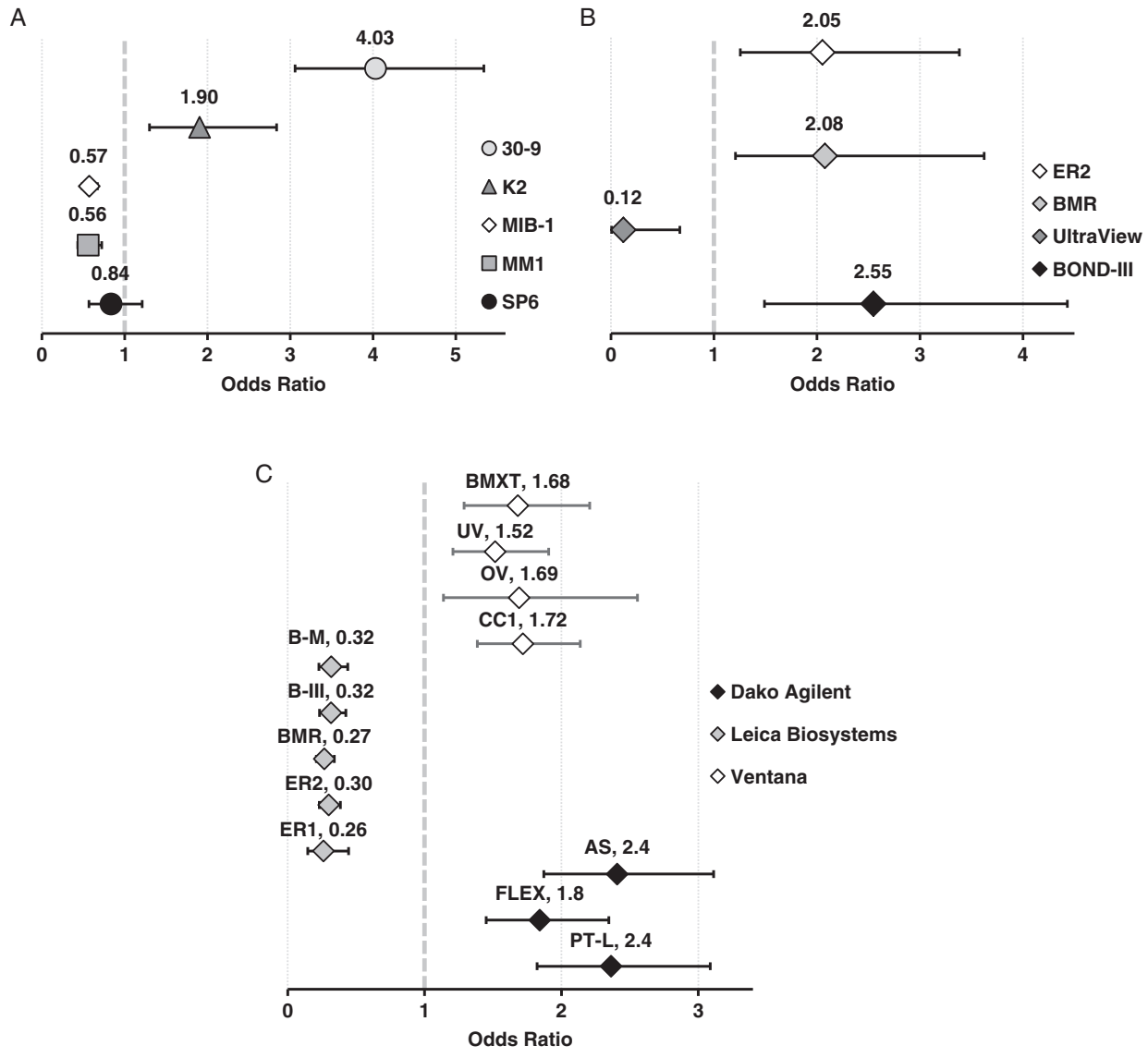

FIGURE 3. Forest plots illustrating primary antibody clone performance and the effects of methodological parameters on MM1 and MIB-1. A, This forest plot shows the odds ratios (ORs) for the staining to be categorized as "Good" for each primary antibody clone. B, ORs for the staining to be categorized as "Good" when the MM1 clone was used in conjunction with various methodological procedures. ER2 = heat-mediated antigen retrieval using ER2 buffer, BMR (Bond MAX Refine) and ultraView = detection systems, BOND-III = automated platform. C, ORs for the staining to be categorized as "Good" when the MIB-1 clone was used in conjunction with various methodological procedures. Reagents and platforms have been grouped and keyed according to their commercial suppliers as shown in the legend. Automated platforms: BMXT = BenchMark-XT, B-M = BOND MAX, B-III = BOND-III, AS = Autostainer; detection systems: UV = ultraView, OV = OptiView, BMR = Bond MAX Refine; heat-mediated antigen retrieval: PT-L = PT-Link. In all plots, the number above markers is the OR for that clone (or clone and procedure), horizontal bars indicate 95% confidence intervals for the OR.

presented in a Supplementary Data Table, Supplemental Digital Content 1, http://links.lww.com/AIMM/A279). For each of the 5 primary antibody clones, the group obtaining the highest mean score was:

- 30-9: all submissions (mean = 17.0; 95% CI, 16.8-17.2)
- K2: all submissions (mean = 16.3; 95% CI, 15.9-16.6)
- MIB-1: with PT-Link HMAR (mean = 16.0; 95% CI, 15.7-16.3)
- MIB-1: stained on an Autostainer platform (mean = 16.0; 95% CI, 15.7-16.2)
- SP6: all submissions (mean = 15.1; 95% CI, 14.7-15.5)
- MM1: stained on a BOND-III platform (mean = 15.0; 95% CI, 14.5-15.6).

DISCUSSION

The Ki-67 methodological data presented here have been gathered from almost 400 health care laboratories evenly distributed between the United Kingdom and the rest of the world, and therefore can be fairly regarded as reflecting current routine practise globally.

TABLE 5. Results of Contingency Table Analyses of Categorical Staining for MM1 Primary Antibody Clone Looking for Associations With Methodological Parameters

Descriptor	ER2 HMAR	Bond MAX Refine Detection	ultraView Detection	BOND-III Automation
Fisher exact test				
P (2-sided)	0.0044	0.0089	0.0473	0.0007
P summary	**	**	*	***
Odds ratio				
Value	2.05	2.08	0.12	2.55
95% CI lower value	1.26	1.21	0.01	0.67
95% CI upper value	3.38	3.62	0.67	4.43
Data analyzed (count of submissions)				
All except good (4-15)	116	76	255	183
Good (16-20)	148	188	9	81
Total	264	264	264	264
Data analyzed (% of submissions)				
All except good (4-15)	43.9	28.8	96.6	69.3
Good (16-20)	56.1	71.2	3.4	30.7

Results are for good staining category versus all other categories combined. Data for other and not stated categories have been omitted.

CI indicates confidence interval; HMAR, heat-mediated antigen retrieval.

*$P \leq 0.05$.

**$P \leq 0.01$.

***$P \leq 0.001$.

The study has shown statistically significant differences in the performance of the 5 most commonly used Ki-67 antibody clones in regard to the quality of staining produced. The 30-9 and the K2 clones were both associated with better quality, whereas the MM1 and MIB-1 clones with less optimal staining quality. However, those broad overview conclusions need to be viewed in context together with the other methodological parameters of IHC staining. Given the strong tendency for IHC reagents and equipment to be sourced from a single supplier in any given laboratory, that is ~98% of the data for 30-9 and 97% for K2 relate to staining using Ventana and Leica supplied reagents and platforms, respectively, the performance of these antibodies is not generalizable to other manufacturer's reagents or systems. MM1 and SP6 did show a tendency to be used with more than 1 supplier's reagents and platforms and here it was clear that performance of the 2 clones was significantly affected by the IHC method used. The SP6 clone performed better when it was used in a Ventana supplied system than it did on a Leica one. The performance of MM1 was suboptimal when used with FLEX detection compared with that with Bond Refine, especially when ER2 HMAR buffer was used with the Refine detection.

The best cross-platform data exist for MIB-1. This clone was the first anti-Ki-67 clone reactive in formalin-fixed paraffin-embedded material to be described,[14,15] and it is still today the most widely used by some margin. There is strong evidence that MIB-1 performs well on both the Agilent Dako and the Ventana systems, but significantly less well on Leica Biosystems platforms with that company's reagents, to an extent that its use should probably be avoided in those circumstances.

TABLE 6. Results of Contingency Table Analyses of Categorical Staining for MIB-1 Primary Antibody Clone Looking for Associations With Methodological Parameters

	HMAR				Detection				Automation			
Descriptor	CC1	ER1	ER2	PT-Link	Bond MAX Refine	FLEX	OptiView	ultraView	Autostainer	BOND-III	BOND MAX	BenchMark XT
Fisher exact test												
P (2-sided)	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.0108	0.0004	<0.0001	<0.0001	<0.0001	0.0002
P summary	****	****	****	****	****	****	**	***	****	****	****	***
Odds ratio												
Value	1.72	0.26	0.30	2.36	0.27	1.84	1.69	1.52	2.41	0.32	0.32	1.68
95% CI lower value	1.39	0.15	0.23	1.82	0.21	1.45	1.14	1.21	1.87	0.23	0.23	1.29
95% CI upper value	2.14	0.44	0.39	3.09	0.34	2.35	2.56	1.91	3.11	0.42	0.44	2.21
Data analyzed (count of submissions)												
All except good (4-15)	991	1463	1194	1189	1135	1133	1413	1089	1154	1307	1346	1239
Good (16-20)	539	67	336	341	395	397	117	441	376	223	184	291
Total	1530	1530	1530	1530	1530	1530	1530	1530	1530	1530	1530	1530
Data analyzed (% of submissions)												
All except good (4-15)	64.8	95.6	78.0	77.7	74.2	74.0	92.4	71.2	75.4	85.4	88.0	81.0
Good (16-20)	35.2	4.4	22.0	22.3	25.8	26.0	7.6	28.8	24.6	14.6	12.0	19.0

Results are for good staining category versus all other categories combined. Data for other and not stated categories have been omitted.

CI indicates confidence interval; HMAR, heat-mediated antigen retrieval.

**$P \leq 0.01$.

***$P \leq 0.001$.

****$P \leq 0.0001$.

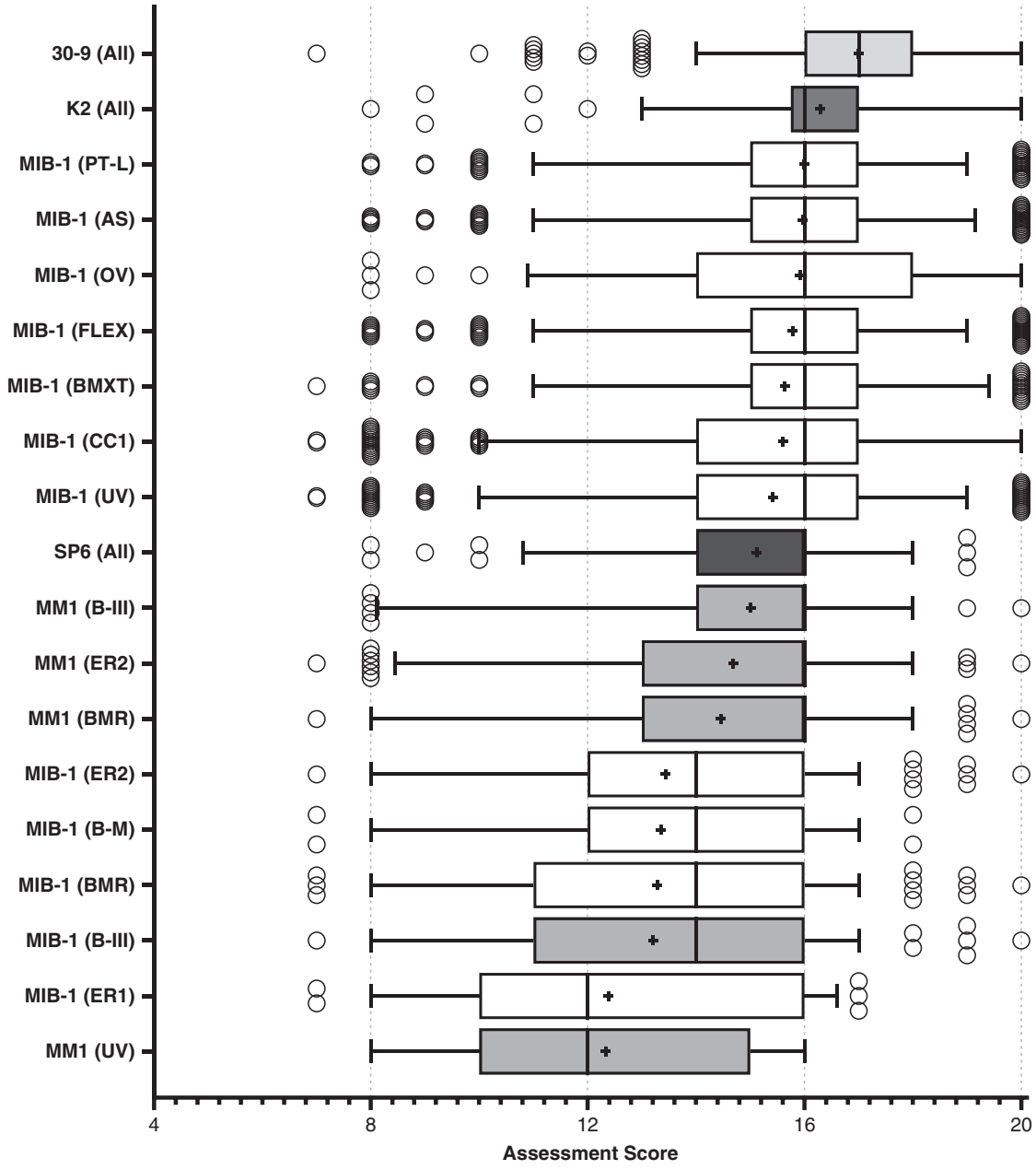

FIGURE 4. Box and whiskers plot. Illustrating subgroup analyses of primary antibody clones and the methodological parameters having a statistically significant effect on their performance. Bounds of box are 25th and 75th quartiles, line within box represents median, the "+" symbol shows the mean, whiskers are 5th and 95th percentile range. Ordered by mean assessment score. Color coding indicates primary antibody clone: black = SP6, dark gray = K2, mid-gray = MM1, light gray = 30-9, white = Agilent Dako. AS = Autostainer (all types), B-III = BOND-III, B-M = BOND MAX, BMXT = BenchMark XT, UV = ultraView, OV = OptiView, BMR = Bond MAX Refine, PT-L = PT-Link.

A very important consideration when viewing the results presented in this paper is the fact that quality of staining was the basis for allocating a score to any given slide. This primarily looked at how well the staining was localized to the nuclear cellular compartment with the absence of nonspecific or inappropriate staining, but not its ability to perform well-producing reproducible quantitative results.

The NordiQC EQA organization have reported on a quantitative study of variability in Ki-67 results seen with different clones and IHC methodologies.[16] They also found significant differences between clones, reagents, and platforms in a similar way to our study. With regard to this UK NEQAS study, an attempt was made to introduce a quantitative aspect to the assessment by using breast tumors showing differing levels of proliferation, but no

formal counting was undertaken to quantify results and assessors instead relied on an "eye-balling" approximation. An approach known to be subjective and to introduce a high level of uncertainty. Therefore, the data are not reported here. The lack of truly quantitative data is a significant limitation of this report. A different approach to the UK NEQAS scheme would be required to evaluate this and, if introduced one which should probably also include an element of scoring by the centers since this was identified as the most significant component of between center variability by the International Ki-67 in Breast Cancer Working Party.[13]

Nevertheless, the study is still important, as clean, "crisp" staining is much more likely to yield reproducible results in both manual and automated scoring than that which is poorly localized and diffuse or shows the presence of nonspecific and/or inappropriate staining. The data it has produced should prove valuable to scientists looking to optimize the staining results they produce for Ki-67 in their own centers.

ACKNOWLEDGMENTS

The authors thank the UK NEQAS ICC & ISH assessors who took part in the original EQA assessment sessions and to all its General Pathology Module participants who submitted slides. M.D. acknowledges the support from the NIHR Biomedical Research Centre at the Royal Marsden Hospital.

REFERENCES

1. Denkert C, Budczies J, von Minckwitz G, et al. Strategies for developing Ki67 as a useful biomarker in breast cancer. *Breast*. 2015;24(suppl 2): S67–S72.
2. Penault-Llorca F, Radosevic-Robin N. Ki67 assessment in breast cancer: an update. *Pathology*. 2017;49:166–171.
3. Dowsett M, Dunbier AK. Emerging biomarkers and new understanding of traditional markers in personalized therapy for breast cancer. *Clin Cancer Res*. 2008;14:8019–8026.
4. Focke CM, Burger H, van Diest PJ, et al. Interlaboratory variability of Ki67 staining in breast cancer. *Eur J Cancer*. 2017;84:219–227.
5. Focke CM, van Diest PJ, Decker T. St Gallen 2015 subtyping of luminal breast cancers: impact of different Ki67-based proliferation assessment methods. *Breast Cancer Res Treat*. 2016;159:257–263.
6. Leung SCY, Nielsen TO, Zabaglo LA, et al. Analytical validation of a standardised scoring protocol for Ki67 immunohistochemistry on breast cancer excision whole sections: an international multicentre collaboration. *Histopathology*. 2019;75:225–235.
7. Leung SCY, 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.
8. Ekholm M, Grabau D, Bendahl PO, et al. Highly reproducible results of breast cancer biomarkers when analysed in accordance with national guidelines—a Swedish survey with central re-assessment. *Acta Oncol*. 2015;54:1040–1048.
9. Acs B, Pelekanou V, Bai Y, et al. Ki67 reproducibility using digital image analysis: an inter-platform and inter-operator study. *Lab Invest*. 2019;99:107–117.
10. Koopman T, Buikema HJ, Hollema H, et al. Digital image analysis of Ki67 proliferation index in breast cancer using virtual dual staining on whole tissue sections: clinical validation and inter-platform agreement. *Breast Cancer Res Treat*. 2018;169:33–42.
11. Smith IE, Robertson J, Kilburn L, et al. Long-term outcome and prognostic value of Ki67 after peri-operative aromatase inhibitor therapy in postmenopausal women with hormone sensitive early breast cancer: The POETIC (peri-operative endocrine therapy–individualising care) trial. *Lancet Oncol*. 2020:1443–1454.
12. Abubakar M, Orr N, Daley F, et al. Prognostic value of automated KI67 scoring in breast cancer: a centralised evaluation of 8088 patients from 10 study groups. *Breast Cancer Res*. 2016;18:104.
13. Polley MY, Leung SC, McShane LM, et al. An international Ki67 reproducibility study. *J Natl Cancer Inst*. 2013;105:1897–1906.
14. Cattoretti G, Becker MH, Key G, et al. Monoclonal antibodies against recombinant parts of the Ki-67 antigen (MIB 1 and MIB 3) detect proliferating cells in microwave-processed formalin-fixed paraffin sections. *J Pathol*. 1992;168:357–363.
15. Gerdes J, Dallenbach F, Lennert K, et al. Growth fractions in malignant non-Hodgkin's lymphomas (NHL) as determined in situ with the monoclonal antibody Ki-67. *Hematol Oncol*. 1984;2: 365–371.
16. Roge R, Nielsen S, Riber-Hansen R, et al. Impact of primary antibody clone, format, and stainer platform on Ki67 proliferation indices in breast carcinomas. *Appl Immunohistochem Mol Morphol*. 2019;27:732–739.