Validity and Reproducibility of Immunohistochemical Scoring by Trained Non-Pathologists on Tissue Microarrays

Published OnlineFirst July 16, 2021; DOI: 10.1158/1055-9965.EPI-21-0295

Validity and Reproducibility of Immunohistochemical Scoring by Trained Non-Pathologists on Tissue Microarrays

Josien C.A. Jenniskens[1], Kelly Offermans[1], Iryna Samarska[2], Gregorio E. Fazzi[2], Colinda C.J.M. Simons[1], Kim M. Smits[2], Leo J. Schouten[1], Matty P. Weijenberg[1], Piet A. van den Brandt[1,3], and Heike I. Grabsch[2,4]

ABSTRACT

Background: Scoring of immunohistochemistry (IHC) staining is often done by non-pathologists, especially in large-scale tissue microarray (TMA)-based studies. Studies on the validity and reproducibility of scoring results from non-pathologists are limited. Therefore, our main aim was to assess interobserver agreement between trained non-pathologists and an experienced histopathologist for three IHC markers with different subcellular localization (nucleus/membrane/cytoplasm).

Methods: Three non-pathologists were trained in recognizing adenocarcinoma and IHC scoring by a senior histopathologist. Kappa statistics were used to analyze interobserver and intraobserver agreement for 6,249 TMA cores from a colorectal cancer series.

Results: Interobserver agreement between non-pathologists (independently scored) and the histopathologist was "substantial" for nuclear and membranous IHC markers ($\kappa_{range} = 0.67–0.75$ and $\kappa_{range} = 0.61–0.69$, respectively), and "moderate" for the cytoplasmic IHC marker ($\kappa_{range} = 0.43–0.57$). Scores of the three non-pathologists were also combined into a "combination score" (if at least two non-pathologists independently assigned the same score to a core, this was the combination score). This increased agreement with the pathologist ($\kappa_{nuclear} = 0.74$; $\kappa_{membranous} = 0.73$; $\kappa_{cytopasmic} = 0.57$). Interobserver agreement between non-pathologists was "substantial" ($\kappa_{nuclear} = 0.78$; $\kappa_{membranous} = 0.72$; $\kappa_{cytopasmic} = 0.61$). Intraobserver agreement of non-pathologists was "substantial" to "almost perfect" ($\kappa_{nuclear,range} = 0.83–0.87$; $\kappa_{membranous,range} = 0.75–0.82$; $\kappa_{cytopasmic} = 0.69$). Overall, agreement was lowest for the cytoplasmic IHC marker.

Conclusions: This study shows that adequately trained non-pathologists are able to generate reproducible IHC scoring results, that are similar to those of an experienced histopathologist. A combination score of at least two non-pathologists yielded optimal results.

Impact: Non-pathologists can generate reproducible IHC results after appropriate training, making analyses of large-scale molecular pathological epidemiology studies feasible within an acceptable time frame.

Introduction

The introduction of the tissue microarray (TMA) technology by Kononen and colleagues (1) in 1998 has enabled large-scale studies using archival formalin-fixed paraffin-embedded (FFPE) tissue blocks (2, 3). The TMA technology has the advantage that sampling of cores leaves the donor block relatively intact, allowing it to be sampled multiple times (3, 4). Furthermore, immunohistochemistry (IHC) on TMAs is cost effective and less time consuming than performing IHC on full tissue sections (2–6). In addition, a higher level of assay standardization can be achieved, improving reproducibility of results (3, 4, 6–8).

Several studies have shown a high degree of concordance between IHC results obtained from TMA sections and full sections when three 0.6 mm cores per case were used (9–13). Interestingly, a study by Gavrielides and colleagues (14) found slightly higher interobserver agreement for HER2 scoring on TMAs compared with full sections, suggesting a potential benefit of the restricted field of view.

Manual scoring of TMA sections can take a considerable amount of time if individual scores need to be provided for hundreds or thousands of cores (7, 15). Although scoring by automated image analysis has been proposed as a potential alternative to manual scoring, IHC markers present in tumor cells and other cell populations at the same time are challenging to assess automatically (16).

Scoring of IHC stained sections is often done by non-pathologists (17, 18). However, studies on the validity of results from non-pathologists are limited. Jaraj and colleagues (19) suggested that after adequate training, non-pathologists are able to produce valid and reproducible IHC results for a cytoplasmic marker. However, it has been suggested that apart from the expert histopathologist knowledge, the agreement of IHC results between observers might also be affected by the subcellular localization of the marker of interest (nucleus/membrane/cytoplasm) (20). There is a limited number of studies investigating scoring agreement of markers with different subcellular

[1]Department of Epidemiology, GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, the Netherlands. [2]Department of Pathology, GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, the Netherlands. [3]Department of Epidemiology, Care and Public Health Research Institute (CAPHRI), Maastricht University Medical Center+, Maastricht, the Netherlands. [4]Pathology and Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, United Kingdom.

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

J.C.A. Jenniskens and K. Offermans contributed as co-first authors. P.A. van den Brandt and Heike I. Grabsch contributed as co-last authors.

Corresponding Authors: Heike I. Grabsch, Department of Pathology, GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, P. Debyelaan 25, 6229 HX Maastricht, the Netherlands. Phone: 3104-3387-4610; E-mail: h.grabsch@maastrichtuniversity.nl; Piet A. van den Brandt, Department of Epidemiology, GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, P. Debyeplein 1, 6200 MD Maastricht, the Netherlands. E-mail: pa.vandenbrandt@maastrichtuniversity.nl

Cancer Epidemiol Biomarkers Prev 2021;30:1867–74

doi: 10.1158/1055-9965.EPI-21-0295

Published OnlineFirst July 16, 2021; DOI: 10.1158/1055-9965.EPI-21-0295

localizations. One of these studies reported similar overall kappa values for scoring of staining in different subcellular compartments (21, 22), whereas another study reported considerably lower agreement for scoring of cytoplasmic immunostaining (23).

We hypothesized that there is good interobserver agreement between trained non-pathologists and pathologists for IHC scoring on TMAs, and that the interobserver agreement does not depend on the subcellular localization of the staining. Therefore, the aims of the current study were to (i) assess interobserver agreement between trained non-pathologists and an experienced pathologist, and (ii) assess agreement of three IHC markers with different subcellular localization (nucleus/membrane/cytoplasm).

Materials and Methods

Study population, tissue collection, and TMA construction

For TMA construction, tissue blocks from colorectal cancer resections of cases from the Netherlands Cohort Study (NLCS) were collected retrospectively from Dutch hospitals (24–26). Hematoxylin & eosin (H&E)-stained sections were reviewed and the area with the highest tumor density was identified. From this area, three 0.6-mm-diameter cores with tumor and three cores with normal epithelium were sampled per case for TMA construction (TMA-Grandmaster, 3DHISTEC). In total, 78 TMA blocks were constructed containing 7,963 tumor cores.

Ethical approval was obtained from Medical Ethical Committee MUMC, number METC 2019-1085.

Immunohistochemistry

Five μm thick serial sections were cut from all 78 TMA blocks and subjected to IHC using an automated immunostainer (DAKO Autostainer Link 48, Glostrup). TP53, GLUT1, and PTEN were chosen as markers to assess interobserver and intraobserver agreement in scoring nuclear, membranous, and cytoplasmic immunoreactivity, respectively, as these are established IHC markers routinely used in clinical setting. Details of primary antibodies and staining protocols are shown in **Table 1**. Staining protocols for all markers were optimized to eliminate background and nonspecific staining. Sections were counterstained with Mayer's Hematoxylin (VWR International B.V.), dehydrated, and mounted with a glass coverslip and xylene-based mounting medium (DPX, Sigma-Aldrich). All TMA sections were scanned using the Aperio scanner (Leica Microsystems) at 40× magnification at the University of Leeds (Leeds, UK) Scanning Facility.

Quality control

Presence of adenocarcinoma was confirmed for every individual core by reviewing the H&E-stained TMA sections. In case of tumor identification difficulties because of poor tumor differentiation or a large number of inflammatory cells, pan-cytokeratin staining was used to identify tumor cells.

Immunohistochemical scoring

Three non-pathologists (G.E. Fazzi: histology technician; K. Offermans: PhD student; J.C.A. Jenniskens: PhD student) were trained by a senior histopathologist (H.I. Grabsch) in (i) recognizing adenocarcinoma on H&E-stained TMA sections; (ii) recognizing immunoreactivity and distinguishing between immunoreactivity in the nucleus, membrane, and cytoplasm; and (iii) scoring of two TMA sections (∼200 cores) for every immunostaining to ensure that the same criteria were used by all assessors.

After training, the three non-pathologists scored all tumor cores for TP53, GLUT1, and PTEN immunostainings. The scores from the three non-pathologists were combined into a "combination score." If at least two non-pathologists independently assigned the same score to a core, this score became the combination score. If all non-pathologists assigned different scores, the core was categorized as "no agreement." Because not all cores were scored by three non-pathologists for GLUT1 (**Table 2**), the remaining scores of the combination score were based on two non-pathologists. When comparing scores from pairs of trained non-pathologists to the score of the pathologist, non-pathologists' scores were combined as described for the combination score of three non-pathologists.

For evaluation of intraobserver agreement, two non-pathologists (assessor 2 and 3) evaluated 10% randomly selected TMA sections (range: 538–681 cores) per marker for a second time after a period of at least 5 months. These scores were only used to assess intraobserver agreement. To assess interobserver agreement between pathologist and non-pathologists, an experienced pathologist (I. Samarska) evaluated the same 10% randomly selected TMA sections for every marker. The contribution of each assessor to the IHC scoring of the different markers is shown in **Table 2**.

TP53 positivity was defined as unequivocal strong nuclear staining and scored semiquantitatively as published previously (13, 27), with minor adaptations, as: (i) no positive tumor nuclei; (ii) ≤10% positive tumor nuclei; (iii) 11% to 50% positive tumor nuclei; (iv) 51% to 90% positive tumor nuclei; and (v) 91% to 100% positive tumor nuclei (**Fig. 1A**).

GLUT1 positivity was defined as any membranous (complete or incomplete) immunostaining of tumor cells, and scored as published previously (28, 29): (i) no tumor cells with membranous immunostaining; (ii) ≤10% tumor cells with membranous immunostaining; (iii) 11% to 50% tumor cells with membranous immunostaining; (iv) >50% tumor cells with membranous immunostaining (**Fig. 1B**).

Table 1. Overview staining protocols, all performed using the DAKO Autostainer Link 48.

Antibody	Clone	Supplier (catalog number)	Antigen retrieval	Dilution	Incubation time	Visualization system	Chromogen
Pan-CK	AE1/AE3	DAKO (GA05361-2)	PT high[a]	RTU[b]	10 minutes	EnVision FLEX[c]	DAB[e]
TP53	DO-7	DAKO (M700101-2)	PT high[a]	RTU[b]	20 minutes	EnVision FLEX[c]	DAB[e]
GLUT1	—	Thermo Fisher Scientific (RB-9052-P)	PT low[d]	1:200	20 minutes	EnVision FLEX[c]	DAB[e]
PTEN	6H2.1	DAKO (M362729-2)	PT high[a]	1:100	20 minutes	EnVision FLEX[c]	DAB[e]

[a]High pH retrieval (K8004) for 20 minutes on the Dako PT link (Agilent Technologies).
[b]RTU: ready-to-use.
[c]EnVision FLEX Visualization Kit (K8008, DAKO).
[d]Low pH retrieval (K8005) for 20 minutes on the Dako PT link (Agilent Technologies).
[e]DAB: 3,3′-diaminobenzidine.

Published OnlineFirst July 16, 2021; DOI: 10.1158/1055-9965.EPI-21-0295

Table 2. Percentage of slides evaluated per assessor for all IHC markers.

Assessor	Experience	Nuclear (TP53)	Membranous (GLUT1)	Cytoplasmic (PTEN)	Intraobserver[b]
1	NP	100%	25%[a]	100%	X
2	NP	100%	100%	100%	10%
3	NP	100%	100%	100%	10%
4	P	10%	10%	10%	X

Abbreviations: NP, non-pathologist; P, pathologist.
[a]Assessor 1 left the project early because of an unforeseen work relocation.
[b]Percentage of slides rescored per protein.

PTEN scoring was performed as described previously (30), comparing cytoplasmic immunostaining intensity of the tumor cells with that of adjacent stromal cells. PTEN immunostaining was classified as: (i) negative (no PTEN staining in the tumor cells); (ii) weak (staining intensity in the tumor cells weaker than in the stromal cells); (iii) moderate (similar staining intensity in tumor and stromal cells); or (iv) strong (staining intensity in the tumor cells stronger than in the stromal cells), see **Fig. 1C**. In case of heterogeneous immunostaining, the region with the highest staining intensity prevailed.

Uninterpretable (e.g., folded cores) or missing cores were categorized as "uninterpretable" and excluded from analyses for all markers.

Statistical analysis

Interobserver and intraobserver agreement was assessed using all cores that passed quality control. Cohen's kappa was used for assessing interobserver agreement between assessor pairs and for assessing intraobserver agreement within one assessor (31). Fleiss' kappa was used for assessing interobserver agreement between more than two assessors (32). All kappa values were weighted (33), taking into account the magnitude of the disagreement (e.g., ≤10% vs. >50% is worse than ≤10% vs. 11%–50%). A weight of 0.5 was chosen for scoring an adjacent category and a weight of zero for non-adjacent categories. Non-weighted Fleiss' kappa was used for assessing the variation in interobserver agreement between scoring categories. To calculate kappa confidence intervals, the bootstrap method was used with 1,000 repetitions (34–36). The interpretation of kappa values is shown in Supplementary Table S1. Agreement between each pair of assessors was determined, as well as agreement between the combination score of two or three non-pathologists and the pathologist's score (for the latter, cores for which no agreement was reached were excluded from analyses). Data were analyzed using Stata (version 15.1, Statacorp).

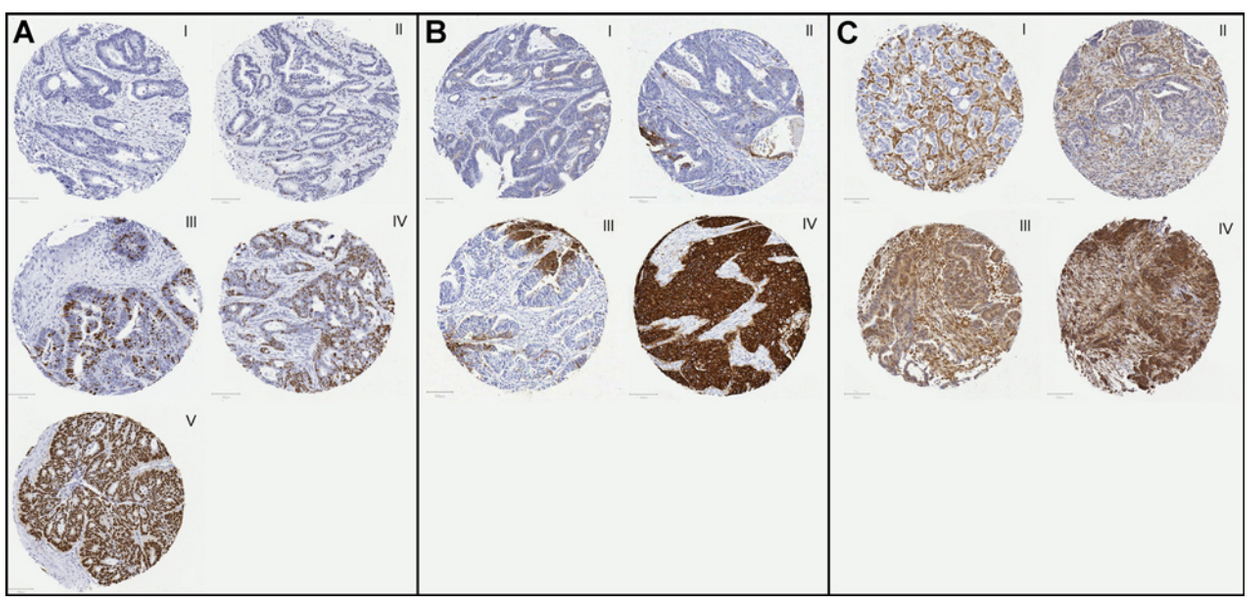

Figure 1.
IHC scoring of nuclear (TP53), membranous (GLUT1), and cytoplasmic (PTEN) protein expression on TMAs. The localization of the antigen–antibody complex is visualized in brown (3,3′-diaminobenzidine) and counterstained in blue (hematoxylin). **A,** Nuclear positivity was defined as unequivocal strong nuclear staining and scored as negative nuclear immunoreactivity (I); ≤10% nuclear immunoreactivity (II); 11%–50% nuclear immunoreactivity (III); 51%–90% nuclear immunoreactivity (IV); 90%–100% nuclear immunoreactivity (V). **B,** Membranous positivity was defined as any membranous (complete or incomplete) immunostaining of tumor cells and scored as negative membranous immunoreactivity (I); ≤10% membranous immunoreactivity (II); 11%–50% membranous immunoreactivity (III); >50% membranous immunoreactivity (IV). **C,** Cytoplasmic protein expression was evaluated by comparing the cytoplasmic immunostaining intensity of the tumor cells to that of adjacent stromal cells, and scored as negative cytoplasmic immunoreactivity (I); weak positive cytoplasmic immunoreactivity (II); moderate positive cytoplasmic immunoreactivity (III); strong positive cytoplasmic immunoreactivity (IV). In case of heterogeneous immunostaining, the region with the highest staining intensity prevailed.

Published OnlineFirst July 16, 2021; DOI: 10.1158/1055-9965.EPI-21-0295

Results

In total, 78 TMA blocks containing 7,963 tumor cores were available. After quality control, 1,714 (21.5%) cores were excluded (464 missing cores; 1,135 cores lacking tumor tissue; 115 uninterpretable tissue cores), leaving 6,249 tumor cores for analyses. All cores were evaluated by at least two assessors (**Table 2**). Frequency distributions of scores assigned by all assessors for nuclear (TP53), membranous (GLUT1), and cytoplasmic (PTEN) immunoreactivity are shown in Supplementary Tables S2–S4.

Interobserver agreement

Non-pathologists versus pathologist

Weighted kappa values of interobserver agreement between non-pathologists and pathologist are shown in **Table 3** (non-weighted kappa values in Supplementary Table S5). Kappa values of each individual non-pathologist with the pathologist showed "substantial" agreement for nuclear ($\kappa_{range} = 0.67–0.75$) and membranous immunostainings ($\kappa_{range} = 0.61–0.69$), and "moderate" for cytoplasmic immunostaining ($\kappa_{range} = 0.43–0.57$). The combination score of the three non-pathologists showed "substantial" agreement with the pathologist's score for nuclear ($\kappa = 0.74$) and membranous immunoreactivity ($\kappa = 0.73$), and "moderate" agreement for cytoplasmic immunoreactivity ($\kappa = 0.57$). The combination score of two non-pathologists showed similar agreement with the pathologist's score as the combination score of three non-pathologists ($\kappa_{nuclear,range} = 0.75–0.81$; $\kappa_{membranous,range} = 0.75–0.79$; $\kappa_{cytoplasmic,range} = 0.54–0.65$). For the majority of scores (range, 90.3%–98.6%), equal or adjacent scoring categories were assigned (**Table 4**) by pathologist and non-pathologists.

In Supplementary Table S6, the agreement per scoring category is shown by non-weighted kappa values. The lowest and highest scoring categories show higher agreement among non-pathologist assessors ($\kappa_{nuclear}$ 0.83 and 0.79; $\kappa_{membranous}$ 0.68 and 0.82; $\kappa_{cytoplasmic}$ 0.61 and 0.51, respectively), than the scoring categories in between ($\kappa_{nuclear,range} = 0.35–0.56$; $\kappa_{membranous,range} = 0.45–0.53$; $\kappa_{cytoplasmic,range} = 0.49–0.53$). Adding the pathologist assessor, this again led to highest agreement in the most extreme categories for nuclear and membranous stainings ($\kappa_{nuclear}$ 0.86 and 0.67; $\kappa_{membranous}$ 0.74 and 0.76, respectively). For cytoplasmic stainings the agreement was highest for the lowest scoring category, and decreased with increasing scoring categories ($\kappa_{category0} = 0.60$; $\kappa_{category1} = 0.53$; $\kappa_{category2} = 0.37$; $\kappa_{category3} = 0.32$).

Non-pathologist versus non-pathologist

Interobserver agreement among non-pathologists is shown in **Table 3** (non-weighted kappa values in Supplementary Table S5). Overall kappa values between all three non-pathologists were similar to those comparing the combination score and the pathologist's score ($\kappa_{nuclear}$ 0.78 vs. 0.74; $\kappa_{membranous}$ 0.72 vs. 0.73; $\kappa_{cytoplasmic}$ 0.61 vs. 0.56, respectively). Scores for nuclear and membranous immunoreactivity showed the highest kappa values among non-pathologists, with an overall weighted kappa of 0.78 ($\kappa_{range} = 0.74–0.80$) and 0.72 ($\kappa_{range} = 0.66–0.81$), respectively. Agreement was lowest for cytoplasmic immunoreactivity, with an overall kappa of 0.61 ($\kappa_{range} = 0.55–0.65$). In the majority of non-pathologists' scores (range, 96.2%–99.8%), equal or adjacent scoring categories were assigned (Supplementary Table S6).

Intraobserver agreement of non-pathologists

Weighted intraobserver kappa values of two non-pathologists are shown in **Table 5** (non-weighted kappa values in Supplementary Table S7). The intraobserver agreement was highest for scoring nuclear and membranous immunoreactivity, showing "almost perfect" agreement ($\kappa_{observer2} = 0.83$; $\kappa_{observer3} = 0.87$), and "substantial" to "almost

Table 3. Interobserver agreement (weighted) between non-pathologists and pathologist.

	Nuclear κ (95% CI)	Membranous κ (95% CI)	Cytoplasmic κ (95% CI)
NP vs. P[a]			
1 vs. 4	0.75 (0.72–0.79)	0.61 (0.55–0.67)[f]	0.57 (0.53–0.61)
2 vs. 4	0.67 (0.63–0.71)	0.69 (0.65–0.73)	0.43 (0.38–0.48)
3 vs. 4	0.70 (0.67–0.74)	0.69 (0.66–0.73)	0.56 (0.52–0.60)
1+2 vs. 4[b,c]	0.80 (0.77–0.84)	0.77 (0.70–0.83)[f]	0.57 (0.51–0.62)
1+3 vs. 4[b,c]	0.81 (0.77–0.84)	0.79 (0.73–0.85)[f]	0.65 (0.60–0.70)
2+3 vs. 4[b,c]	0.75 (0.72–0.79)	0.75 (0.72–0.79)	0.54 (0.50–0.60)
Combination score[c,d] vs. 4	0.74 (0.71–0.78)	0.73 (0.69–0.77)	0.57 (0.52–0.61)
NP vs. NP[e]			
1 vs. 2	0.74 (0.73–0.75)	0.69 (0.67–0.72)[f]	0.55 (0.54–0.57)
1 vs. 3	0.79 (0.79–0.80)	0.66 (0.64–0.69)[f]	0.64 (0.62–0.65)
2 vs. 3	0.80 (0.79–0.81)	0.81 (0.80–0.82)	0.65 (0.64–0.67)
1 vs. 2 vs. 3[g]	0.78	0.72[f]	0.61

Abbreviations: NP, non-pathologist; P, pathologist. Nuclear, TP53; membranous, GLUT1; cytoplasmic, PTEN.
[a]Based on a random 10% of TMA sections (range, 538–681 cores).
[b]Comparison of a combination of two non-pathologists with the pathologist: if the two non-pathologists independently assigned the same score to a core, this was the combined score. If the non-pathologists assigned a different score, the core was categorized as no agreement.
[c]Cores where no agreement was reached between non-pathologists (combination score = no agreement) were excluded for analyses.
[d]The combination score is based on all three non-pathologist's scores: if at least two assessors independently assigned the same score to a core, this was the combination score. If none of the assessors assigned the same score, the core was categorized as no agreement.
[e]Based on all cores ($N = 6,249$).
[f]Assessor 1 left the project early because of an unforeseen work relocation, 1,457 cores were evaluated.
[g]Confidence interval for weighted kappa of multiple assessors (>2) could not be calculated using Stata.

Published OnlineFirst July 16, 2021; DOI: 10.1158/1055-9965.EPI-21-0295

Table 4. Percentage of discrepancies between assessors.

	Nuclear Difference in categories[b]				Membranous[a] Difference in categories[b]				Cytoplasmic Difference in categories[b]			
	0	1	2	3/4	0	1	2	3	0	1	2	3
Interobserver												
NP vs. P[c]												
1 vs. 4	57.9	32.4	8.0	1.8	62.5	29.4	6.0	2.2	54.1	42.1	3.9	0.0
2 vs. 4	51.4	41.0	6.7	0.9	70.5	24.5	4.0	1.0	63.0	33.9	2.7	0.5
3 vs. 4	61.7	32.0	5.1	1.2	70.6	25.2	3.1	1.0	62.1	36.5	1.4	0.0
Combination vs. 4	69.8	27.1	2.9	0.2	73.7	22.7	3.0	0.7	64.6	34.8	0.7	0.0
NP vs. NP[d]												
1 vs. 2	72.0	24.2	3.5	0.3	68.3	29.2	2.1	0.4	65.4	34.0	0.6	0.0
1 vs. 3	76.3	22.2	1.4	0.1	65.5	30.9	3.2	0.4	73.2	26.5	0.3	0.0
2 vs. 3	76.3	22.4	1.2	0.1	81.4	17.2	1.3	0.1	76.0	23.8	0.2	0.0
Intraobserver												
NP vs. NP[c]												
2 vs. 2	82.6	16.3	1.0	0.0	82.4	16.6	1.0	0.0	78.8	21.2	0.0	0.0
3 vs. 3	84.1	15.3	0.6	0.0	74.2	24.8	0.8	0.3	80.0	20.0	0.0	0.0

Note: Uninterpretable cores were excluded.

Abbreviations: NP, non-pathologist; P, pathologist. Nuclear, TP53; membranous, GLUT1; cytoplasmic, PTEN.

[a]Assessor 1 left the project early because of an unforeseen work relocation.

[b]Difference in categories assigned by the two assessors: 0 = same category assigned (no discrepancy); 1 = adjacent categories were assigned (e.g.,<10% positive and 11%–50% positive); 2 = difference between assigned categories was 2 (e.g., <10% positive and >50% positive); 3/4 = difference between assigned categories was 3 or 4 (e.g., negative and >50%).

[c]Based on a random 10% of TMA sections.

[d]Based on all TMA sections.

perfect" agreement ($\kappa_{observer2} = 0.82$; $\kappa_{observer3} = 0.75$), respectively. Scoring of cytoplasmic immunoreactivity showed "substantial" agreement ($\kappa_{observer2} = 0.69$; $\kappa_{observer3} = 0.69$). In the majority of scores (range, 98.9%–100%), equal or adjacent categories were assigned at the first and second timepoint (Supplementary Table S6).

Discussion

TMAs are increasingly used to analyze protein expression by IHC in large-scale studies (2, 3, 5, 37). Scoring is often done by non-pathologists (17, 18); however, only few studies reported validity and reproducibility of scoring results (38, 39). To the best of our knowledge, our study is one of the first to investigate agreement of TMA-based scoring of immunoreactivity in different subcellular localizations by non-pathologists. Our study showed that interobserver agreement between an experienced histopathologist and trained non-pathologists was "moderate" to "substantial." Agreement with the pathologist's score did not further increase when a combination score from three instead of two trained non-pathologists was used.

Table 5. Intraobserver agreement (weighted) of two non-pathologists, based on 10% randomly selected TMA sections.

	Assessor 2 κ (95% CI)	Assessor 3 κ (95% CI)
Nuclear	0.83 (0.80–0.86)	0.87 (0.84–0.90)
Membranous	0.82 (0.79–0.85)	0.75 (0.72–0.78)
Cytoplasmic	0.69 (0.64–0.74)	0.69 (0.64–0.74)

Note: Nuclear, TP53; membranous, GLUT1; cytoplasmic, PTEN. CI, confidence interval.

Interobserver agreement non-pathologists versus pathologist

Our study demonstrates that non-pathologists can generate reproducible results. These results are in line with a previous study by Jaraj and colleagues (19), reporting comparable kappa values for interobserver agreement between pathologists and non-pathologists. Even though it was not their main objective, two other studies reported comparable interobserver agreement between pathologists and non-pathologists (22, 40). However, some of the studies reported weighted kappa values (19, 22), but did not state what weights were assigned to adjacent scoring categories, making a direct comparison of kappa values with our study impossible.

Considering the subjectivity of immunoreactivity scoring, several studies recommended that scoring should be done by multiple assessors to improve interobserver agreement (39, 41, 42). Our study confirmed that combining scores from multiple non-pathologists into a combination score increased interobserver agreement with the pathologist's score. Combining scores of three non-pathologists instead of two did not change interobserver agreement with the pathologist, indicating that IHC scoring by two non-pathologists seems to be sufficient to yield reliable IHC results.

Immunoreactivity scoring in different subcellular localizations

A limited number of studies investigated scoring agreement of immunoreactivity in different subcellular localizations, showing inconsistent results (21–23). We showed that scoring of nuclear and membranous immunoreactivity generally leads to higher interobserver agreement compared with cytoplasmic immunoreactivity, consistent with results of Bolton and colleagues (23). However, this is in contrast to two other studies which did not find a difference in the intraobserver and interobserver agreement when scoring nuclear, membranous and cytoplasmic immunoreactivity (21, 22). These

Published OnlineFirst July 16, 2021; DOI: 10.1158/1055-9965.EPI-21-0295

discrepant results might be explained by the use of different IHC scoring methods between studies.

The IHC markers selected for the current study were chosen to provide a range of subcellular localizations (nucleus/membrane/cytoplasm) for scoring purposes. These markers are generalizable to other IHC stainings considering the subcellular localization.

Interobserver agreement among non-pathologists

Hitherto, few studies reported interobserver agreement of IHC results among non-pathologists. In the current study, we found "substantial" to "almost perfect" agreement among trained non-pathologists, which is in line with previously published results on TMAs and whole tissue sections (17–19).

Intraobserver agreement of non-pathologists

IHC studies often report intraobserver kappa values as a measure of reproducibility. Our study shows that non-pathologists are able to generate reproducible IHC scores after appropriate training, which is in line with previous studies (17–19, 40). Interestingly, intraobserver kappa values of non-pathologists in the current study were similar to those previously reported for pathologists (23, 43). In general, across all three markers, disagreements were limited to one-category discordances (e.g., <10% vs. 11%–50%) for all comparisons.

Limitations

Our study has some limitations. We have no information on intraobserver and interobserver agreement of pathologists, as this was beyond the scope of this article. Furthermore, the current study used TMA cores to assess interobserver and intraobserver agreement. It has been described in the literature that interobserver agreement increases when using TMA cores compared with whole tissue sections (14). Thus, it remains to be clarified whether the agreement among non-pathologists and between non-pathologists and pathologists is similar in full tissue sections. However, the aim of this study was specifically to investigate IHC scoring on TMAs, because non-pathologists will mainly be involved in IHC scoring in large-scale studies using TMAs. Also, we did not directly compare the scoring performance between trained non-pathologists and untrained non-pathologists; thus, we are not able to draw direct conclusions on the necessity of training, and in particular whether similar results would have been obtained without training.

Recommendations

We propose some recommendations which could improve comparability of IHC studies. First, it is important to report what weights were used for analyses of weighted kappa values. In addition, we think it would be of value to report both weighted and non-weighted kappa values. Second, it should be mentioned clearly in the methods what the IHC scoring experience of assessors was. If done by non-pathologists, it is important to report their training. Third, our results showed that disagreements were mostly limited to one-category discordances, suggesting that less refined scoring protocols may potentially improve agreement. This is in line with previous studies (44, 45), in which the authors showed that agreement improved when using scoring protocols with less categories. However, we acknowledge that the number of categories of the scoring protocol depends on the novelty and clinical relevance of the biomarker being studied. Scoring protocols for potential new biomarkers might comprise more categories compared with well-known biomarkers. Finally, we suggest that IHC scoring should be performed by at least two non-pathologists to be able to assess interobserver agreement among assessors. Ideally, these non-pathologists are trained by an expert pathologist and a certain percentage of samples (e.g., 10%) are double-scored by the pathologist to ensure quality of scoring.

Conclusion

In this large study investigating interobserver and intraobserver agreement of TMA-based immunoreactivity scores between pathologists and non-pathologists, we have shown that non-pathologists can generate reproducible IHC scoring results that are similar to those of an experienced pathologist. A combination score of at least two non-pathologists yielded optimal results. Future studies are required to validate our findings and to examine the practical implications and impact of potential misclassification, by comparing effect estimates for established stain-outcome associations when using the pathologist's score versus the non-pathologists' combination score.

Authors' Disclosures

No disclosures were reported.

Authors' Contributions

J.C.A. Jenniskens: Conceptualization, formal analysis, investigation, writing–original draft. **K. Offermans:** Conceptualization, formal analysis, investigation, writing–original draft. **I. Samarska:** Conceptualization, investigation, writing–original draft. **G.E. Fazzi:** Investigation, writing–review and editing. **C.C.J.M. Simons:** Writing–review and editing. **K.M. Smits:** Writing–review and editing. **L.J. Schouten:** Writing–review and editing. **M.P. Weijenberg:** Writing–review and editing. **P.A. van den Brandt:** Conceptualization, resources, supervision, funding acquisition, writing–original draft, project administration. **H.I. Grabsch:** Conceptualization, supervision, writing–original draft.

Acknowledgments

The authors would like to thank the participants of the Netherlands Cohort Study (NLCS), the Netherlands Cancer Registry, and the Dutch Pathology Registry. They are grateful to Ron Alofs and Harry van Montfort for data management and programming assistance; to Jaleesa van der Meer, Edith van den Boezem, and Peter Moerkerk for TMA construction; and the University of Leeds (Leeds, UK) for scanning of all slides.

The Rainbow-TMA consortium was financially supported by BBMRI-NL, a Research Infrastructure financed by the Dutch government (NWO 184.021.007, to P.A. van den Brandt), and Maastricht University Medical Center, University Medical Center Utrecht, and Radboud University Medical Centre, the Netherlands. The authors would like to thank all investigators from the Rainbow-TMA consortium project group [P.A. van den Brandt, A. zur Hausen, H.I. Grabsch, M. van Engeland, L.J. Schouten, J. Beckervordersandforth (Maastricht University Medical Center+, Maastricht, the Netherlands); P.H.M. Peeters, P.J. van Diest, H.B. Bueno de Mesquita (University Medical Center Utrecht, Utrecht, the Netherlands); J. van Krieken, I. Nagtegaal, B. Siebers, B. Kiemeney (Radboud University Medical Center, Nijmegen, the Netherlands); F.J. van Kemenade, C. Steegers, D. Boomsma, G.A. Meijer (Amsterdam University Medical Center, locaties VUmc, the Netherlands); F.J. van Kemenade, B. Stricker (Erasmus University Medical Center, Rotterdam, the Netherlands); L. Overbeek, A. Gijsbers (PALGA, the Nationwide Histopathology and Cytopathology Data Network and Archive, Houten, the Netherlands)] and collaborating pathologists [among others: A. de Bruïne (VieCuri Medical Center, Venlo); J.C. Beckervordersandforth (Maastricht University Medical Center+, Maastricht); J. van Krieken, I. Nagtegaal (Radboud University Medical Center, Nijmegen); W. Timens (University Medical Center Groningen, Groningen); F.J. van Kemenade (Erasmus University Medical Center, Rotterdam); M.C.H. Hogenes (Laboratory for Pathology OostNederland, Hengelo); P.J. van Diest (University Medical Center Utrecht, Utrecht); R.E. Kibbelaar (Pathology Friesland, Leeuwarden); A.F. Hamel (Stichting Samenwerkende Ziekenhuizen Oost-Groningen, Winschoten); A.T.M.G. Tiebosch (Martini Hospital, Groningen); C. Meijers (Reinier de Graaf Gasthuis/S.S.D. Z., Delft); R. Natté (Haga Hospital Leyenburg, The Hague); G.A. Meijer (Amsterdam University Medical Center, locatie VUmc); J.J.T.H. Roelofs (Amsterdam University Medical Center, locatie AMC); R.F. Hoedemaeker (Pathology Laboratory Pathan,

Rotterdam); S. Sastrowijoto (Orbis Medical Center, Sittard); M. Nap (Atrium Medical Center, Heerlen); H.T. Shirango (Deventer Hospital, Deventer); H. Doornewaard (Gelre Hospital, Apeldoorn); J.E. Boers (Isala Hospital, Zwolle); J.C. van der Linden (Jeroen Bosch Hospital, Den Bosch); G. Burger (Symbiant Pathology Center, Alkmaar); R.W. Rouse (Meander Medical Center, Amersfoort); P.C. de Bruin (St. Antonius Hospital, Nieuwegein); P. Drillenburg (Onze Lieve Vrouwe Gasthuis, Amsterdam); C. van Krimpen (Kennemer Gasthuis, Haarlem); J.F. Graadt van Roggen (Diaconessenhuis, Leiden); S.A.J. Loyson (Bronovo Hospital, The Hague); J.D. Rupa (Laurentius Hospital, Roermond); H. Kliffen (Maasstad Hospital, Rotterdam); H.M. Hazelbag (Medical Center Haaglanden, The Hague); K. Schelfout (Stichting Pathologisch en Cytologisch Laboratorium West-Brabant, Bergen op Zoom); J. Stavast (Laboratorium Klinische Pathologie Centraal Brabant, Tilburg); I. van Lijnschoten (PAMM Laboratory for Pathology and Medical Microbiology, Eindhoven); K. Duthoi (Amphia Hospital, Breda)].

This project was funded by The Dutch Cancer Society (KWF 11044, to P.A. van den Brandt).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received March 8, 2021; revised May 4, 2021; accepted July 12, 2021; published first July 16, 2021.

References

1. Kononen J, Bubendorf L, Kallioniemi A, Bärlund M, Schraml P, Leighton S, et al. Tissue microarrays for high-throughput molecular profiling of tumor specimens. Nat Med 1998;4:844–7.
2. Tawfik El-Mansi M, Williams AR. Validation of tissue microarray technology using cervical adenocarcinoma and its precursors as a model system. Int J Gynecol Cancer 2006;16:1225–33.
3. Ilyas M, Grabsch H, Ellis IO, Womack C, Brown R, Berney D, et al. Guidelines and considerations for conducting experiments using tissue microarrays. Histopathology 2013;62:827–39.
4. Camp RL, Neumeister V, Rimm DL. A decade of tissue microarrays: progress in the discovery and validation of cancer biomarkers. J Clin Oncol 2008;26:5630–7.
5. Domeny-Duarte P, Niero L, Domingues MAC. Tissue microarrays of bone marrow aspirate clot allow assessment of multiple samples. Pathology, research and practice. 2020;216:152721.
6. Shergill IS, Shergill NK, Arya M, Patel HRH. Tissue microarrays: a current medical research tool. Curr Med Res Opin 2004;20:707–12.
7. Bubendorf L, Nocito A, Moch H, Sauter G. Tissue microarray (TMA) technology: miniaturized pathology archives for high-throughput in situ studies. J Pathol 2001;195:72–9.
8. Al Kuraya K, Simon R, Sauter G. Tissue microarrays for high-throughput molecular pathology. Ann Saudi Med 2004;24:169–74.
9. Boone J, van Hillegersberg R, van Diest PJ, Offerhaus GJA, Borel Rinkes IHM, Ten Kate FJW. Validation of tissue microarray technology in squamous cell carcinoma of the esophagus. Virchows Arch 2008;452:507–14.
10. Zhang D, Salto-Tellez M, Putti TC, Do E, Koay ESC. Reliability of tissue microarrays in detecting protein expression and gene amplification in breast cancer. Mod Pathol 2003;16:79–84.
11. Hassan S, Ferrario C, Mamo A, Basik M. Tissue microarrays: emerging standard for biomarker validation. Curr Opin Biotechnol 2008;19:19–25.
12. Hoos A, Urist MJ, Stojadinovic A, Mastorides S, Dudas ME, Leung DH, et al. Validation of tissue microarrays for immunohistochemical profiling of cancer specimens using the example of human fibroblastic tumors. Am J Pathol 2001;158:1245–51.
13. Jourdan F, Sebbagh N, Comperat E, Mourra N, Flahault A, Olschwang S, et al. Tissue microarray technology: validation in colorectal carcinoma and analysis of p53, hMLH1, and hMSH2 immunohistochemical expression. Virchows Arch 2003;443:115–21.
14. Gavrielides MA, Conway C, O'Flaherty N, Gallas BD, Hewitt SM. Observer performance in the use of digital and optical microscopy for the interpretation of tissue-based biomarkers. Anal Cell Pathol 2014;2014:157308.
15. Jawhar NM. Tissue microarray: a rapidly evolving diagnostic and research tool. Ann Saudi Med 2009;29:123–7.
16. O'Hurley G, Sjostedt E, Rahman A, Li B, Kampf C, Pontén F, et al. Garbage in, garbage out: a critical evaluation of strategies used for validation of immunohistochemical biomarkers. Mol Oncol 2014;8:783–98.
17. Nielsen JS, Jakobsen E, Holund B, Bertelsen K, Jakobsen A. Prognostic significance of p53, Her-2, and EGFR overexpression in borderline and epithelial ovarian cancer. Int J Gynecol Cancer 2004;14:1086–96.
18. Dijkema IM, Struikmans H, Dullens HF, Kal HB, van der Tweel I, Battermann JJ. Influence of p53 and bcl-2 on proliferative activity and treatment outcome in head and neck cancer patients. Oral Oncol 2000;36:54–60.
19. Jaraj SJ, Camparo P, Boyle H, Germain F, Nilsson B, Petersson F, et al. Intra-and interobserver reproducibility of interpretation of immunohistochemical stains of prostate cancer. Virchows Arch 2009;455:375–81.
20. Lejeune M, Jaen J, Pons L, López C, Salvadó MT, Bosch R, et al. Quantification of diverse subcellular immunohistochemical markers with clinicobiological relevancies: validation of a new computer-assisted image analysis procedure. J Anat 2008;212:868–78.
21. Zlobec I, Terracciano L, Jass JR, Lugli A. Value of staining intensity in the interpretation of immunohistochemistry for tumor markers in colorectal cancer. Virchows Arch 2007;451:763–9.
22. Kirkegaard T, Edwards J, Tovey S, McGlynn LM, Krishna SN, Mukherjee R, et al. Observer variation in immunohistochemical analysis of protein expression, time for a change? Histopathology 2006;48:787–94.
23. Bolton KL, Garcia-Closas M, Pfeiffer RM, Duggan MA, Howat WJ, Hewitt SM, et al. Assessment of automated image analysis of breast cancer tissue microarrays for epidemiologic studies. Cancer Epidemiol Biomarkers Prev 2010;19:992–9.
24. van den Brandt PA, Goldbohm RA, van't Veer P, Volovics A, Hermus RJ, Sturmans F. A large-scale prospective cohort study on diet and cancer in The Netherlands. J Clin Epidemiol 1990;43:285–95.
25. van den Brandt PA, Schouten LJ, Goldbohm RA, Dorant E, Hunen PM. Development of a record linkage protocol for use in the Dutch Cancer Registry for Epidemiological Research. Int J Epidemiol 1990;19:553–8.
26. van den Brandt PA. Molecular pathological epidemiology of lifestyle factors and colorectal and renal cell cancer risk. Maastricht Pathology 2018. 11th Joint Meeting of the British Division of the International Academy of Pathology and the Pathological Society of Great Britain & Ireland, 19–22 June 2018. J Pathol 2018;246:S1–S46.
27. Resnick MB, Routhier J, Konkin T, Sabo E, Pricolo VE. Epidermal growth factor receptor, c-MET, β-catenin, and p53 expression as prognostic indicators in stage II colon cancer: a tissue microarray study. Clin Cancer Res 2004;10:3069–75.
28. Cooper R, Sarioğlu S, Sökmen S, Füzün M, Küpelioğlu A, Valentine H, et al. Glucose transporter-1 (GLUT-1): a potential marker of prognosis in rectal carcinoma? Br J Cancer 2003;89:870–6.
29. Sakashita M, Aoyama N, Minami R, Maekawa S, Kuroda K, Shirasaka D, et al. Glut1 expression in T1 and T2 stage colorectal carcinomas: its relationship to clinicopathological features. Eur J Cancer 2001;37:204–9.
30. Richman SD, Adams R, Quirke P, Butler R, Hemmings G, Chambers P, et al. Pre-trial inter-laboratory analytical validation of the FOCUS4 personalised therapy trial. J Clin Pathol 2016;69:35–41.
31. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.
32. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76:378–82.
33. Cohen J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213–20.
34. Efron B, Tibshirani RJ. An introduction to the bootstrap Boca Raton, FL: Chapman & Hall; 1993.
35. Lee J, Fung KP. Confidence interval of the kappa coefficient by bootstrap resampling. Psychiatry Res 1993;49:97–8.
36. Reichenheim ME. Confidence intervals for the kappa statistic. Stata J 2004;4:421–8.
37. Shiraishi T, Shinto E, Nearchou IP, Tsuda H, Kajiwara Y, Einama T, et al. Prognostic significance of mesothelin expression in colorectal cancer disclosed by area-specific four-point tissue microarrays. Virchows Arch 2020;477:409–20.
38. Cross SS. Observer accuracy in estimating proportions in images: implications for the semiquantitative assessment of staining reactions and a proposal for a new system. J Clin Pathol 2001;54:385–90.

Published OnlineFirst July 16, 2021; DOI: 10.1158/1055-9965.EPI-21-0295

39. Adams EJ, Green JA, Clark AH, Youngson JH. Comparison of different scoring systems for immunohistochemical staining. J Clin Pathol 1999;52:75–7.
40. Gavrielides MA, Gallas BD, Lenz P, Badano A, Hewitt SM. Observer variability in the interpretation of HER2/neu immunohistochemical expression with unaided and computer-aided digital microscopy. Arch Pathol Lab Med 2011;135:233–42.
41. Masmoudi H, Hewitt SM, Petrick N, Myers KJ, Gavrielides MA. Automated quantitative assessment of HER-2/neu immunohistochemical expression in breast cancer. IEEE Trans Med Imaging 2009;28:916–25.
42. Obuchowski NA. How many observers are needed in clinical studies of medical imaging? AJR Am J Roentgenol 2004;182:867–9.
43. Ali A, Bell S, Bilsland A, Slavin J, Lynch V, Elgoweini M, et al. Investigating various thresholds as immunohistochemistry cutoffs for observer agreement. Appl Immunohistochem Mol Morphol 2017;25:599–608.
44. de Jong D, Rosenwald A, Chhanabhai M, Gaulard P, Klapper W, Lee A, et al. Immunohistochemical prognostic markers in diffuse large B-cell lymphoma: validation of tissue microarray as a prerequisite for broad clinical applications–a study from the Lunenburg Lymphoma Biomarker Consortium. J Clin Oncol 2007;25:805–12.
45. Rüschoff J, Dietel M, Baretton G, Arbogast S, Walch A, Monges G, et al. HER2 diagnostics in gastric cancer—guideline validation and development of standardized immunohistochemical testing. Virchows Arch 2010;457:299–307.

Cancer Epidemiology, Biomarkers & Prevention

AACR American Association for Cancer Research

Validity and Reproducibility of Immunohistochemical Scoring by Trained Non-Pathologists on Tissue Microarrays

Josien C.A. Jenniskens, Kelly Offermans, Iryna Samarska, et al.

Cancer Epidemiol Biomarkers Prev 2021;30:1867-1874. Published OnlineFirst July 16, 2021.