



# Staining Performance of ALK and ROS1 Immunohistochemistry and Influence on Interpretation in Non—Small-Cell Lung Cancer

Cleo Keppens,<sup>\*</sup> Jan von der Thüsen,<sup>†</sup> Patrick Pauwels,<sup>‡§</sup> Ales Ryska,<sup>¶</sup> Nils 't Hart,<sup>||\*\*</sup> Ed Schuurin,<sup>||</sup> Keith Miller,<sup>††</sup> Erik Thunnissen,<sup>‡‡</sup> Karen Zwaenepoel,<sup>‡§</sup> and Elisabeth M.C. Dequeker<sup>\*</sup>

From the Biomedical Quality Assurance Research Unit,<sup>\*</sup> Department of Public Health and Primary Care, University of Leuven, Leuven, Belgium; the Department of Pathology,<sup>†</sup> Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands; the Department of Pathology,<sup>‡</sup> University Hospital Antwerp, Edegem, Belgium; the Centre for Oncological Research,<sup>§</sup> University of Antwerp, Edegem, Belgium; the Department of Pathology,<sup>¶</sup> Charles University Medical Faculty Hospital, Hradec Kralove, Czech Republic; the Department of Pathology,<sup>||</sup> University of Groningen, University Medical Center Groningen, Groningen, the Netherlands; the Department of Pathology,<sup>\*\*</sup> Isala Klinieken, Zwolle, the Netherlands; the UK National External Quality Assessment Scheme for Immunocytochemistry and in Situ Hybridisation,<sup>††</sup> University College London Cancer Institute, London, United Kingdom; and the Department of Pathology,<sup>‡‡</sup> Vrije Universiteit Amsterdam Medical Center, Amsterdam, the Netherlands

Accepted for publication  
September 16, 2020.

Address correspondence to  
Elisabeth M.C. Dequeker, Ph.D.,  
Department of Public Health and  
Primary Care, Biomedical  
Quality Assurance Research  
Unit, University of Leuven,  
Kapucijnenvoer 35d, Box 7001,  
3000 Leuven, Belgium. E-mail:  
[els.dequeker@kuleuven.be](mailto:els.dequeker@kuleuven.be).

Selection of non—small-cell lung cancer patients for treatment relies on the detection of expression of anaplastic lymphoma kinase (ALK) and ROS proto-oncogene 1 (ROS1) protein by immunohistochemistry (IHC). We evaluated staining performance for different IHC protocols and laboratory characteristics, and their influence on ALK and ROS1 interpretation during external quality assessment schemes between 2015 and 2018. Participants received five formalin-fixed, paraffin-embedded cases for staining by their routine protocol, whereafter at least two pathologists scored them simultaneously under a multihead microscope and awarded a graded expert staining score (ESS) from 1 to 5 points based on staining quality. European Conformity in Vitro Diagnostic kits (such as D5F3) revealed a better ALK ESS compared with laboratory-developed tests. ESS was indifferent to the applied antibody dilution or a recent protocol change. Lower ESSs were observed for higher antibody incubation times and temperatures. ESS for various ROS1 protocols were largely similar. Overall, for both markers, ESS improved over time and for repeated external quality assessment participation but was independent of laboratory setting or experience. Except for ROS1, ESS positively correlated with laboratory accreditation. IHC stains with lower ESS correlated with increased error rates in ALK and ROS1 interpretation and analysis failures. Laboratory characteristics differently affected staining quality and interpretation, and laboratories should assess both aspects, and less common protocols need improvement in staining performance. (*J Mol Diagn* 2020, 22: 1438–1452; <https://doi.org/10.1016/j.jmoldx.2020.09.006>)

Disclosures: J.v.d.T. received fees for participating in advisory board meetings from Roche, Merck, Merck Sharp & Dohme, Bristol-Myers Squibb, and AstraZeneca, and research support from Bristol-Myers Squibb and Roche; P.P. received fees for participating in advisory board meetings from Biocartis, Boehringer Ingelheim, Roche, Novartis, Pfizer, Merck, Merck Sharp & Dohme, Bristol-Myers Squibb, and AstraZeneca, and research support from Roche and AstraZeneca; A.R. received honoraria for participating in advisory board meetings from Merck Sharp & Dohme, Bristol-Myers Squibb, Roche, Pfizer, and AstraZeneca, and unrestricted research support from Roche, Merck Sharp & Dohme, and Bristol-Myers Squibb; N.t.H. received fees for participating in advisory board meetings from Merck, Roche, Pfizer, and AstraZeneca, and unrestricted research support from Roche and Pfizer; E.S. received fees for consultancy/advisory board from AstraZeneca, Roche, Pfizer, Bayer, Novartis, Bristol-Myers Squibb, BioRad, Illumina, Ageno BioSciences, Janssen Cilag

(Johnson&Johnson), and BioCartis; speaker's fee from AstraZeneca, Roche, Pfizer, Novartis, BioRad, Illumina, and BioCartis; and (unrestricted) grants from Boehringer Ingelheim, Bristol-Myers Squibb, Biocartis, BioRad, Ageno BioSciences, and Roche (all outside the submitted work and fees to University Medical Center Groningen); E.T. acted as consultant for Merck Sharp & Dohme, Pfizer, Clovis, Bristol-Myers Squibb, AstraZeneca, Diaceutics, Amgen, Abbvie, Roche Ventana, Bayer, Takeda, and currently for Zelfstandigen Zonder Personeel and Histogenex; and has received honoraria for speaker from AstraZeneca, Pfizer, and Roche trainer SP142 and SP263; the Vrije Universiteit Amsterdam Medical Center received grants from the Investigator-Initiated Research Program from Pfizer and AstraZeneca; E.M.C.D. received an unrestricted research grant from Pfizer Oncology for the coordination of the European Society of Pathology lung external quality assessment schemes, not related to the research performed in this study.

Personalized medicine by appropriate targeting of molecular targets in tumors has improved survival of patients with non–small-cell lung cancer (NSCLC). Besides testing for variants in the epidermal growth factor receptor gene, the identification of rearrangements in anaplastic lymphoma kinase (*ALK*) and ROS proto-oncogene 1 (*ROS1*) genes is required to select patients for treatment with tyrosine kinase inhibitors.<sup>1</sup>

*ALK* and *ROS1* rearrangements are mutually exclusive, and occur in approximately 3% to 7% and 1% to 2% of NSCLC cases, respectively.<sup>2</sup> The most common fusion partner for *ALK* includes the echinoderm microtubule-associated protein-like 4 gene,<sup>3</sup> whereas the CD74-*ROS1* fusion occurs most frequently for *ROS1*. More than 20 fusion partners have been described for both *ALK* and *ROS1*, but the clinical significance of different fusion products requires further investigation.

To date, five *ALK* tyrosine kinase inhibitors (crizotinib, ceritinib, alectinib, brigatinib, and lorlatinib) have received approval by the US Food and Drug Administration (<https://www.fda.gov>, last accessed July 20, 2020) and the European Medicines Agency (<https://www.ema.europa.eu/en/medicines/download-medicine-data>, last accessed July 20, 2020) for treatment of advanced *ALK*-positive NSCLC. In 2016 and 2019, crizotinib and entrectinib were approved for the treatment of advanced *ROS1*-rearranged NSCLC, respectively.<sup>4,5</sup>

Although fluorescence *in situ* hybridization (FISH) testing was originally considered the gold standard, the detection of increased protein expression to identify underlying gene fusions by immunohistochemistry (IHC) is currently widely accepted.<sup>6,7</sup> Namely, IHC is reported to be fast, cheap, and particularly useful in small biopsy specimens with a limited number of neoplastic cells, and showed a better correlation with clinical outcome.<sup>6–8</sup>

Several primary antibody clones are commercially available for *ALK* IHC. In 2016, the D5F3 antibody as part of the *ALK* (D5F3) IHC CDx Assay (Roche, Ventana, Oro Valley, AZ) received approval for the selection of *ALK*-positive NSCLC, making *ALK* IHC a valuable alternative compared with FISH testing.<sup>9</sup> In contrast to antibodies using a binary scoring system (eg, D5F3 IHC CDx), confirmation of an intermediate *ALK* staining pattern is recommended for antibodies using an intensity-based score (eg, 5A4 *ALK* IHC), using FISH or other independent *ALK*-fusion detection assays.<sup>2,10,11</sup> For the *ALK1* antibody, conflicting results are reported, as most studies found lower sensitivity of IHC using *ALK1* compared with D5F3 or 5A4.<sup>12</sup> However, there are some studies reporting higher sensitivity for *ALK1* compared with the other antibodies.<sup>13</sup>

For *ROS1*, a single antibody clone (D4D6; Cell Signaling Technology, Danvers, MA) was for some time the only commercially available primary antibody, until the 1A1 (Origene, Rockville, MD) and the *ROS1* SP384 antibody (Roche, Ventana) were also introduced.<sup>14,15</sup> Detection systems and other conditions of the protocols (eg, temperature

and duration of incubation) reported in combination with D4D6 have been shown to vary considerably,<sup>16</sup> and cross-platform studies are needed. As only part of the *ROS1* IHC-positive cases are confirmed by *ROS1* FISH positivity, the IHC can be used for screening of *ROS1*-rearranged NSCLC, but positive *ROS1* IHC results should be confirmed by a molecular or cytogenetic method.<sup>1</sup> Because of the rarity of *ALK* and *ROS1* rearrangements, multiple testing strategies are used in different countries, including sequential testing of markers based on the clinical needs or parallel testing using next-generation sequencing to enable concurrent detection of as many potentially targetable mutations as possible. Thus, *ALK* or *ROS1* IHC is often performed as a screening tool before FISH confirmation.<sup>1</sup>

Given the high incidence and mortality of lung cancer (11.6% of total cancer cases and 18.4% of cancer-related deaths) worldwide,<sup>17</sup> of which 80% is accounted for by NSCLC, the correct identification of *ALK* and *ROS1* rearrangements is indispensable for appropriate treatment selection. Even though recommendations and guidelines are available, these are mainly general principles and do not provide specific data required to help laboratories to evaluate the technical performance of IHC assays routinely used for diagnostics.<sup>18–20</sup>

European external quality assessment (EQA) schemes have been organized to evaluate the performance of *ALK* and *ROS1* IHC analyses and to assist laboratories in generating accurate test results. EQA results showed room for improvement with regard to the analytical outcome.<sup>21–23</sup> Staining quality needs to be addressed with a focus on antibodies and protocol parameters to allow identification of required elements for appropriate staining and their relation to analytical outcome. Several EQA schemes already indicated a variety in *ALK* IHC protocols and detection methods, both affecting EQA pass rates.<sup>10,20</sup>

This study evaluated staining performances from the European Society of Pathology NSCLC EQA schemes between 2015 and 2018, independently performed by a team of expert pathologists, for various *ALK* and *ROS1* IHC protocols and different laboratory characteristics. These findings were compared with the sample outcomes as scored by the participants, to assess how the scores awarded by the experts translate into the participants' interpretation.

## Materials and Methods

Four external quality assessment schemes were organized for *ALK* IHC between 2015 and 2018, and three schemes were organized for *ROS1* IHC between 2016 and 2018. All schemes were in accordance with ISO17043:2010 (conformity assessment: general requirements for proficiency testing, available from International Organization for Standardization, Geneva, Switzerland) and open to all laboratories worldwide. During every scheme, participating laboratories received two unstained sections (3 μm thick)

from five NSCLC resection specimens or cell lines for staining by their routine IHC protocol (Supplemental Table S1). Participants were able to choose whether to use both unstained slides for staining or use one slide as a spare or negative control for the primary antibody. Samples were validated beforehand by a central reference laboratory for the IHC status and corresponding FISH status.

Participants were given 14 calendar days for staining and interpretation. Participants were requested to return stained slides for a central review. In 2015 and 2016, five cases were requested for review. From 2017, the distribution of the five cases occurred in three separate runs for ALK. In the first run, three cases were sent, and one case was sent in the second and third round. For ROS1, two separate runs were organized in 2017 during which two and three cases were sent for staining, respectively. As these runs were 2 to 4 months apart, only the slides from the run with three cases were sent back for review. This was to ensure that all samples were analyzed by the participants at a comparable moment in time and were treated with an identical protocol, and to reduce the administrative burden for participants by sending multiple packages back to the coordination center. Participants also completed an electronic datasheet with their laboratory characteristics (such as their accreditation status or setting) and details of detection protocols. In that same sheet, participants had to provide their individual scoring of the ALK and ROS1 expression in the EQA samples (IHC status positive or negative, or analysis failure) for each case, according to their routine protocol for interpretation. The reported accreditation statuses and laboratory settings (ie, university hospital, general hospital, or private or industry laboratory) were validated afterwards on the websites of the relevant national accreditation bodies and the laboratory websites, respectively. Accreditation was defined as adhering to ISO15189 (medical laboratories: particular requirements for quality and competence, available from International Organization for Standardization), ISO17025 (general requirements for the competence of testing and calibration laboratories, available from International Organization for Standardization), or relevant national standards (such as CAP15189) (<https://www.cap.org/laboratory-improvement/accreditation/cap-15189-accreditation-program>, last accessed July 20, 2020), and could include both specific accreditation for ALK or ROS1 IHC or general laboratory accreditation for all executed analyses.

Scoring criteria for the technical assessment were discussed beforehand by a team of experts in molecular pathology (all pathologists; J.v.d.T., P.P., A.R., N.t.H., K.M., and E.T.), and were based on a graded scale: 5, excellent staining; 4, pass with minor remark; 3, deficiency without affecting clinical output; 2, incorrect staining with clinical output affected; and 1, failed staining, impossible to interpret. Before scoring, one round of harmonization was performed to ensure equal assessment compared with several reference slides. A

team of two to three pathologists scored the stained slides simultaneously under a multihead microscope. An expert staining score of 1 to 5 points (hereafter referred to as ESS) was awarded based on the staining quality for all the evaluated slides combined, relative to the optimal staining pattern for the specific protocol used by the participants. Control tissue was only taken into account if sent back for review.

At the end of the EQA scheme, images of (sub)optimal stains, their corresponding ESS, and protocols were made available for the participants. They also received a general scheme summary on the performance for the different sample outcomes and ESS, and individual comments on the staining quality were sent to each participant separately.

Statistics on these EQA scheme data were performed using SAS software version 9.4 of the SAS System for Windows (SAS Institute Inc., Cary, NC). Two generalized linear models were used with estimation based on generalized estimating equations to account for clustering in the data (ie, tests performed by the same laboratory). First, proportional odds models were used to evaluate the association of laboratory characteristics (such as setting or accreditation status) or used method (antibody, antigen retrieval, or detection) with the ESS as ordinal outcome. Results are presented by odds ratios (ORs) with 95% CIs. The ESS, as assessed on the slides, is a combination of both the used primary antibody clone and all subsequent protocol steps, and the overall ESS values of the complete protocols are presented. For the less common antibodies, protocols were grouped into one category for statistical analysis, and individual ESS scores are shown separately. The number of EQA participations, samples tested per year, and involved staff members in the complete test process were considered as ordinal variables (instead of categorical) to evaluate the influence of a +1 level increase in these ordinal variables on the ESS.

Second, Poisson models were used to analyze the association of the ESS from the assessors and the laboratory characteristics, with the scoring of the participants, represented as the number of analysis failures, false-positive or false-negative results, as count outcome variables. Only cases for which both an ESS and an outcome scored by the participants were available were taken into account. Results are presented as incidence rate ratios (IRRs) with 95% CIs, and taking into account the log of the total number of samples analyzed during the EQA scheme as an offset variable.

## Results

### Influence of Laboratory Characteristics on Expert Staining Score

In total, data from 174 unique laboratories from 37 countries were analyzed for ALK IHC (352 participations) and 82 unique participants from 26 countries for ROS1 IHC (137

**Table 1** Average ESS for Laboratory Characteristics of the ALK and ROS1 EQA Scheme Participants

Laboratory characteristic	ALK (n = 352)		ROS1 (n = 137)	
	n (%)	Average ESS OR (95% CI)	n (%)	Average ESS OR (95% CI)
Samples tested in last 12 months		0.92 (0.76–1.11) NS		1.10 (0.88–1.38) NS
No clinical testing	10 (2.8)	4.6	9 (6.6)	3.6
<10	8 (2.3)	4.4	13 (9.5)	4.5
10–99	107 (30.4)	4.1	27 (19.7)	4.1
100–249	115 (32.7)	4.1	25 (18.3)	4.4
250–499	70 (19.9)	4.1	32 (23.4)	4.0
>500	38 (10.8)	4.2	19 (13.9)	4.3
Missing data	4 (1.1)	4.0	12 (8.8)	4.3
Staff involved in testing†		1.18 (0.96–1.45) NS		1.24 (0.91–1.70) NS
1–5	146 (41.5)	4.1	45 (32.8)	4.1
6–10	114 (32.4)	4.0	48 (35.1)	4.1
11–20	58 (16.5)	4.1	28 (20.4)	4.2
>20	25 (7.1)	4.6	12 (8.8)	4.4
Missing data	9 (2.6)	4.0	4 (2.9)	4.3
EQA participations		1.11 (0.91–1.36) NS		4.43 (2.62–7.48)***
First participation	175 (49.7)	4.1	82 (59.85)	3.9
Second participation	93 (26.4)	4.2	39 (28.47)	4.4
Third participation	59 (16.8)	4.3	16 (11.68)	4.9
Fourth participation	25 (7.1)	3.9	NA†	NA‡
EQA round		****		***
2015	73 (20.7)	3.7	NA†	NA‡
2016	91 (25.9)	4.2	31 (22.63)	4.0
2017	96 (27.3)	4.3	52 (37.96)	3.9
2018	92 (26.1)	4.1	54 (39.42)	4.5
Laboratory setting§¶		NS		NS
Industry	6 (1.7)	4.5	1 (0.7)	3.0
(Private) laboratories	55 (15.6)	4.2	6 (4.4)	4.0
Hospital laboratories	107 (30.4)	4.1	30 (21.9)	4.2
University and research	184 (52.3)	4.1	99 (72.3)	4.2
Missing data	0 (0.0)	NA	1 (0.7)	4.0
Accreditation status¶		NS		*
No	172 (48.9)	4.1	40 (29.2)	3.9
Yes	175 (49.7)	4.2	87 (63.5)	4.3
Missing data	5 (1.4)	4.0	10 (7.3)	4.0

Proportional odds models with generalized estimating equations for clustering of the data were used to analyze the difference in ESS. The first three characteristics (samples tested, staff involved, and EQA participations) are evaluated on an ordinal level. OR >1 represents a higher ESS for a higher category level. OR <1 represents a lower ESS for a higher category level. Other characteristics are evaluated as a categorical variable: overall significance levels are given. ORs for every pairwise comparison between categories are described in the main text.

\**P* < 0.05, \*\*\**P* < 0.001, and \*\*\*\**P* < 0.0001.

†As the number of staff members involved in the complete test process was related to the number of samples tested annually, this characteristic was used as a measure of the size of the laboratory and available expertise/resources.

‡No EQA scheme was organized to evaluate the ESS of ROS1 in 2015.

§Industry are laboratories involved in the development of diagnostic commercial kits. (Private) laboratories are not within a hospital's infrastructure. Hospital laboratories included private and public hospitals. University and research included education and research hospitals, university hospitals, university laboratories, and anticancer centers.

¶Laboratory setting and accreditation were validated on the websites of the laboratories and national accreditation bodies. Accreditation was defined as adhering to ISO15189 (medical laboratories: particular requirements for quality and competence, available from International Organization for Standardization, Geneva, Switzerland), ISO17025 (general requirements for the competence of testing and calibration laboratories, available from ISO), or relevant national standards (such as CAP15189), specifically for the ALK or ROS1 test or a general laboratory accreditation.

ALK, anaplastic lymphoma kinase; EQA, external quality assessment; ESS, expert staining score; OR, odds ratio; NA, not applicable; NS, not significant; ROS1, ROS proto-oncogene 1.

participations) (Supplemental Table S2). An overview of the different laboratory characteristics and their relation to the ESS is given in Table 1.

For both markers, there was a significant improvement of the average ESS depending on the EQA round. For ALK, a higher ESS was observed in 2016 (OR, 2.34; 95%

**Table 2** Average ESS for ALK and ROS1 IHC General Method Characteristics

Method characteristic	ALK ( <i>n</i> = 352)		ROS1 ( <i>n</i> = 137)	
	<i>n</i> (%)	Average ESS	<i>n</i> (%)	Average ESS
Method type		*		
Approved kit (CDx)	149 (42.3)	4.3	NA†	NA†
LDT	203 (57.7)	4.0	137 (100.0)	4.2
Switched protocol between EQA schemes‡		NS		NS
No	130 (36.9)	4.2	50 (36.5)	4.5
Yes	47 (13.4)	4.1	5 (3.6)	4.8
NA‡	175 (49.7)	4.1	82 (59.9)	3.9
Antibody dilution		NS		NS
<1:50	67 (19.1)	4.3	13 (9.5)	4.5
1:50–1:100	79 (22.4)	4.0	89 (65.0)	4.2
>1:100	41 (11.7)	4.0	35 (25.6)	4.1
RTU	165 (46.9)	4.1	0 (0.0)	
Incubation time, minutes		**		NS
1–30	241 (68.5)	4.2	39 (28.5)	4.1
31–60	91 (25.9)	4.1	83 (60.6)	4.3
>60	19 (5.4)	3.6	15 (11.0)	3.9
Missing data	1 (0.3)	3.0	0 (0.0)	NA
Incubation temperature, °C		**		*
Room temperature	114 (32.4)	4.1	49 (35.8)	4.0
1–40	219 (62.2)	4.2	85 (62.0)	4.3
>40	18 (5.1)	3.6	3 (2.2)	3.3
Missing data	1 (0.3)	3.0	0 (0.0)	NA

Proportional odds models with generalized estimating equations for clustering of the data were used to analyze the difference in ESS as a categorical variable: overall significance levels are given. Odds ratios (95% CIs) for every pairwise comparison between categories are described in the main text.

\* $P < 0.05$ , \*\* $P < 0.01$ .

†No approved kit is currently available for ROS1 immunohistochemistry.

‡A switch included the change in primary antibody, antigen retrieval, or detection kit. NA includes entries from first participations for which no method information from previous rounds was available.

ALK, anaplastic lymphoma kinase; EQA, external quality assessment; ESS, expert staining score; IHC, immunohistochemistry; LDT, laboratory-developed test; NA, not applicable; NS, not significant; ROS1, ROS proto-oncogene 1; RTU, ready to use.

CI, 1.39–3.95;  $P = 0.0014$ ), 2017 (OR, 3.64; 95% CI, 2.11–6.29;  $P < 0.0001$ ), and 2018 (OR, 2.17; 95% CI, 1.26–3.75;  $P = 0.0056$ ) compared with the first round in 2015. The ESS did not differ between any of the subsequent rounds (2016 to 2018). For ROS1 IHC, there was a significant improvement in ESS in the latest 2018 scheme compared with 2016 (OR, 4.48; 95% CI, 1.87–10.75;  $P = 0.0008$ ) and 2017 (OR, 4.85; 95% CI, 2.24–10.53;  $P < 0.0001$ ), but not between 2016 and 2017. An overview of the comments provided to the participants for the awarded ESS is shown in [Supplemental Table S3](#).

Most participants were university and research laboratories for both ALK (52.3%) and ROS1 (72.3%) IHC, followed by (general) hospital laboratories (30.4% and 21.9% for ALK and ROS1, respectively). Private laboratories less frequently performed ROS1 (4.4%) analyses compared with ALK (15.6%) analyses. There was no significant difference in the ESS depending on the laboratory's setting.

Almost half of the laboratories (49.7%) were accredited for ALK IHC testing according to ISO15189 or relevant national standards, compared with 63.5% for ROS1 testing. Only for ROS1 analysis, a higher ESS was observed for

accredited participants (OR, 2.31; 95% CI, 1.15–4.65;  $P = 0.0191$ ).

The number of staff members involved in the complete IHC testing process (from sample receipt until readout) was most frequently between 1 and 5 staff members for ALK analysis (41.5%) and between 6 and 10 staff members for ROS1 detection (35.1%). An increased number of involved staff members resulted in a higher ESS, although the difference was not significant.

With regard to laboratory experience, there was no significant relation between the number of samples tested annually and ESS. For ROS1 analysis, ESS improved significantly if a laboratory participated in more successive EQA rounds (OR, 4.43; 95% CI, 2.62–7.48;  $P < 0.0001$ ). This was not the case for ALK analysis ( $P = 0.3039$ ).

#### Expert Staining Score for the Different Protocols

The relationship between the ESS and general IHC protocol characteristics is described in [Table 2](#). Overall, 57.7% of ALK IHC tests were performed by a laboratory-developed test (LDT), compared with 100.0% for ROS1 (as no

approved kits were available at the time of this study). The use of an approved kit according to the manufacturer's instructions resulted in a better ESS for *ALK* (OR, 1.67; 95% CI, 1.08–2.56;  $P = 0.0201$ ) in comparison to LDTs.

Performance was not affected by a switch in test method (primary antibody, antigen retrieval, or detection platform) between two schemes or the applied antibody dilution. In contrast, lower average ESS values were observed for longer incubation times and higher temperatures of the primary antibody. In more detail, an incubation time of >60 minutes negatively affected the ESS compared with a time between 31 and 60 minutes (OR, 0.45; 95% CI, 0.21–0.97;  $P = 0.0405$ ) or between 1 and 30 minutes (OR, 0.31; 95% CI, 0.17–0.60;  $P = 0.0005$ ) for the *ALK* primary antibody, but not for *ROS1* ( $P = 0.0544$ ). In total, 19 participants used an incubation time of >60 minutes. For these participants, a lower signal/noise ratio was observed (weak antigen detection along with high background). Most of these participants (13/19) used the Ventana UltraView, i-view, or 1-view detection method, and received an individual comment that polymer detection is recommended. Of those 19 participants, 12 used less common antibodies, such as 5A4 (Abcam, Cambridge, UK), which demonstrated lower performances.

Incubation temperatures <40°C resulted in better ESS compared with temperatures >40°C (OR, 3.10; 95% CI, 1.51–6.34;  $P = 0.0020$ ) for *ALK*. For *ROS1*, on the other hand, incubation at room temperature resulted in lower performance compared with usage of a specific temperature <40°C (OR, 0.45; 95% CI, 0.23–0.86;  $P = 0.0160$ ).

The ESS was also evaluated for the different combinations of primary antibodies, antigen retrieval, and detection platforms, as reported by the EQA participants. ORs to obtain a good ESS for the most frequently used combinations relative to other methods are visualized in [Table 3](#).

Most *ALK* IHC participants (39.8%,  $n = 352$ ) used the D5F3 *ALK* IHC CDx kit from Ventana, including the D5F3 antibody clone in combination with the Cc1 kit and Optiview DAB IHC detection kit. There was no significant difference in ESS between the most frequently used protocols. Also, the 1A4 clone (Origene) displayed a higher ESS compared with most other methods. Participants using the 5A4 (Novocastra, Nussloch, Germany) antibody in combination with the Optiview, Bond, or Envision flex detection kits demonstrated a good performance. However, when using this antibody with other detection methods, the 5A4 antibody resulted in a lower performance. A similarly better performance was noticed for other antibodies when using any of these three detection methods compared with participants using the antibodies with other detection methods, such as the ZytoChem method. Hence, the combination of the applied antibody and detection system is important. An example of optimal and suboptimal *ALK* IHC staining patterns for different protocols is shown in [Supplemental Figure S1](#).

There was also a significant difference between the methods used for *ALK* antigen retrieval ( $P = 0.0073$ ).

Laboratory-developed EDTA or TRIS-EDTA based approaches performed significantly worse compared with the commercial methods, such as Cc1 (Ventana) and Omnis Envision FLEX TRS (Dako, Santa Clara, CA) (data not shown). For antigen detection, laboratories using the ZytoChem Plus (HRP) Polymer Kit (Zytomed) performed suboptimal to all other reported detection kits ( $P < 0.0001$ ).

Of 137 tests, 131 (95.6%) for *ROS1* were performed by the D4D6 (Cell Signaling Technology) primary antibody, most frequently (48.2%) in combination with the Cc1 kit and Optiview DAB IHC detection kit from Ventana. ORs were all >1, implying a higher ESS relative to all other methods used. The ESS was significantly higher for the D4D6 clone in combination with Optiview, compared with using the same clone in combination with the UltraView Universal DAB Detection kit or PT module TRS High Envision Flex (Dako) ([Table 4](#)). For *ROS1*, there were no observed differences between any of the other used protocols, or between the individual methods for antigen retrieval and detection. An example of optimal and suboptimal *ROS1* IHC staining patterns for the D4D6 antibody in combination with the different platforms is shown in [Supplemental Figure S2](#). The number of users and average ESS for the other primary antibodies are represented in [Table 5](#). ESS values are highly variable, with scores ranging between 1.0 and 5.0 on a total of 5 points. The number of users for these other primary antibodies is small, with minimum one and maximum four laboratories applying the antibody.

### Participants' Scoring of *ALK* and *ROS1* Expression in the EQA Samples

In total, 1379 *ALK* cases and 470 *ROS1* cases were returned to the EQA provider for assessment of the ESS. For *ALK*, incorrect interpretations (false-positive or false-negative outcomes) and analysis failures (failure to stain or interpret the slides) were observed in 34 (2.5%) and 19 (1.4%) cases, respectively. For *ROS1*, 4 (0.9%) misinterpretations and 6 (1.3%) analysis failures were observed.

A lower staining performance, as determined by the ESS, was significantly correlated to the incidence of analysis misinterpretations or failures for the total number of *ALK* and *ROS1* cases tested ([Figure 1](#)). For *ALK* IHC, the false-negative results are shown in [Figure 1A](#) and analysis failures in *ALK* expression negative cases are shown in [Figure 1B](#). For *ROS1* IHC, lower ESS resulted in more false-positive interpretations and more analysis failures in all evaluated samples ([Figure 1](#)). The IRR for incorrect outcomes in positive cases is not given as only one error was made.

The incidence of misinterpretations and analysis failures for the above-mentioned laboratory characteristics is presented in [Supplemental Table S4](#). Incorrect interpretations for laboratories participating in the *ALK* subscheme increased in later EQA rounds (IRR, 1.39; 95% CI, 1.02–1.90;  $P = 0.0355$ ). Analysis failures diminished

**Table 3** ALK ESS for Different Combinations of Primary Antibodies, Antigen Retrieval, and Detection Kits

Primary antibody	Antigen retrieval method	Detection system	Times used (n = 352) n (%)	Average ESS on 5 points ± SD	Method code	OR (95% CI) relative to method			
						a	b	c	d
D5F3 (Ventana)	Cc1 (Ventana)	OptiView DAB IHC Detection Kit (Ventana)	140 (39.8)	4.3 ± 0.8 a		NA	2.37 (0.73–7.70)	1.23 (0.52–2.90)	1.14 (0.58–2.23)
		Other combination of antigen retrieval and detection	9 (2.6)	3.7 ± 0.9 b	0.42 (0.13–1.38)	NA	0.52 (0.13–2.08)	0.48 (0.14–1.71)	
D5F3 (Cell Signaling Technology)	Cc1 (Ventana)	OptiView DAB IHC Detection Kit (Ventana)	22 (6.3)	4.2 ± 0.8 c	0.82 (0.35–1.93)	1.92 (0.48–7.65)	NA	0.93 (0.37–2.34)	
		Bond Epitope Retrieval 2 (Leica)	13 (3.7)	4.2 ± 0.8 d	0.88 (0.45–1.72)	2.08 (0.59–7.38)	1.08 (0.43–2.73)	NA	
		Other combination of antigen retrieval and detection	31 (8.8)	3.9 ± 0.8 e	0.48 (0.22–1.05)	1.13 (0.30–4.26)	0.59 (0.20–1.72)	0.54 (0.21–1.39)	
5A4 (Novocastra)	Cc1 (Ventana)	OptiView DAB IHC Detection Kit (Ventana)	27 (7.7)	4.3 ± 0.7 f	1.07 (0.51–2.26)	2.53 (0.69–9.32)	1.31 (0.47–3.68)	1.22 (0.50–2.97)	
		Bond Epitope Retrieval 2 (Leica)	15 (4.3)	3.7 ± 1.0 g	0.32 (0.09–1.06)	0.71 (0.14–3.71)	0.39 (0.09–1.59)	0.36 (0.10–1.32)	
		Envision FLEX TRS, High pH (Dako)	14 (4.0)	4.3 ± 1.1 h	1.39 (0.37–5.28)	3.32 (0.59–18.76)	1.70 (0.37–7.93)	1.58 (0.38–6.62)	
		Other combination of antigen retrieval and detection	17 (4.8)	3.6 ± 1.1 i	<b>0.25</b> <b>(0.10–0.66)</b> **	0.60 (0.14–2.52)	0.31 (0.10–1.02)	<b>0.29</b> <b>(0.10–0.81)</b> *	
5A4 (Abcam)	Various combinations of antigen retrieval and detection	10 (2.8)	3.5 ± 0.5 j	<b>0.21</b> <b>(0.12–0.40)</b> ****	0.51 (0.15–1.76)	<b>0.26</b> <b>(0.10–0.68)</b> **	<b>0.24</b> <b>(0.11–0.53)</b> ***		
1A4 (Origene)	Various combinations of antigen retrieval and detection	19 (5.4)	4.5 ± 0.9 k	2.76 (0.99–7.75)	<b>6.54</b> <b>(1.50–28.57)</b> *	3.41 (0.96–12.20)	<b>3.15</b> <b>(1.01–9.80)</b> *		
Other antibodies (11)	Various combinations of antigen retrieval and detection	35 (9.9)	3.2 ± 1.1 l	0.23 (0.09–0.60)	0.54 (0.13–2.16)	<b>0.28</b> <b>(0.08–0.94)</b> *	<b>0.26</b> <b>(0.09–0.79)</b> *		

(table continues)

Proportional odds models with generalized estimating equations for clustering of the data were used to analyze the difference in ESS. Differences in ESS are represented as ORs (95% CIs) for every method (row level) relative to other methods used (column level). OR >1 represents a higher ESS for a given method (column level) relative to the other method (row level). OR <1 represents a lower ESS for a method relative to other methods. Significant results are highlighted in bold.

\* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ , and \*\*\*\* $P < 0.0001$ .

ALK, anaplastic lymphoma kinase; ESS, expert staining score; NA, not applicable; OR, odds ratio.

when a higher number of samples were tested annually (IRR, 0.60; 95% CI, 0.44–0.83;  $P = 0.0020$ ), but increased when laboratories switched from one protocol to another one (IRR, 15.75; 95% CI, 1.65–150.71;  $P = 0.0168$ ).

For *ROS1*, the opposite was observed with fewer analysis failures in later EQA rounds (IRR, 0.29; 95% CI, 0.09–0.91;  $P = 0.0346$ ) and more analysis failures when a laboratory performs more *ROS1* IHC samples annually in

routine practice (IRR, 1.92; 95% CI, 1.25–2.93;  $P = 0.0027$ ). If computed, there was no difference depending on the number of previous participations, staff members involved, or the accreditation status and technique type (LDT versus *in vitro* diagnostic–labeled kit) used.

Because of the wide variety of different protocols and low number of analysis misinterpretations and failures, IRRs for each separate ALK and *ROS1* IHC method were not analyzed.

**Table 3** (continued)

OR (95% CI) relative to method							
e	f	g	h	i	j	k	l
2.11 (0.95–4.64)	0.94 (0.44–1.98)	3.18 (0.94–10.72)	0.72 (0.19–2.73)	<b>3.94</b> <b>(1.52–10.21)**</b>	<b>4.68</b> <b>(2.53–8.66)****</b>	0.36 (0.13–1.01)	<b>4.39</b> <b>(1.66–11.63)**</b>
0.89 (0.24–3.37)	0.40 (0.11–1.46)	1.40 (0.27–7.30)	0.30 (0.05–1.70)	1.67 (0.40–6.99)	1.98 (0.57–6.88)	<b>0.15</b> <b>(0.04–0.67)*</b>	1.86 (0.46–7.46)
1.70 (0.58–5.00)	0.76 (0.27–2.15)	2.59 (0.63–10.65)	0.59 (0.13–2.73)	3.19 (0.98–10.38)	<b>3.78</b> <b>(1.47–9.77)**</b>	0.29 (0.08–1.05)	<b>3.55</b> <b>(1.06–11.88)*</b>
1.85 (0.72–4.75)	0.82 (0.34–2.01)	2.79 (0.76–10.31)	0.63 (0.15–2.65)	<b>3.47</b> <b>(1.23–9.73)*</b>	<b>4.11</b> <b>(1.89–8.94)***</b>	<b>0.32</b> <b>(0.10–0.99)*</b>	<b>3.86</b> <b>(1.27–11.69)*</b>
NA	0.45 (0.17–1.18)	1.58 (0.39–6.33)	0.34 (0.08–1.52)	1.87 (0.59–5.92)	2.22 (0.93–5.30)	<b>0.17</b> <b>(0.05–0.57)**</b>	2.09 (0.67–6.49)
2.25 (0.85–5.96)	NA	3.40 (0.89–13.00)	0.77 (0.18–3.35)	<b>4.22</b> <b>(1.38–12.87)*</b>	<b>5.00</b> <b>(2.17–11.51)***</b>	0.39 (0.12–1.27)	<b>4.69</b> <b>(1.52–14.52)**</b>
0.63 (0.16–2.55)	0.30 (0.08–1.13)	NA	0.23 (0.04–1.20)	1.19 (0.28–5.10)	1.41 (0.39–5.12)	<b>0.11</b> <b>(0.02–0.51)**</b>	1.33 (0.30–5.81)
2.95 (0.66–13.21)	1.30 (0.30–5.65)	4.41 (0.83–23.36)	NA	<b>5.53</b> <b>(1.03–29.58)*</b>	<b>6.56</b> <b>(1.58–27.25)**</b>	0.51 (0.10–2.59)	<b>6.16</b> <b>(1.22–30.99)*</b>
0.53 (0.17–1.68)	<b>0.24</b> <b>(0.08–0.72)*</b>	0.84 (0.20–3.61)	<b>0.18</b> <b>(0.03–0.97)*</b>	NA	1.19 (0.44–3.21)	<b>0.09</b> <b>(0.03–0.34)***</b>	1.11 (0.33–3.80)
0.45 (0.19–1.07)	<b>0.20</b> <b>(0.09–0.46)</b> ***	0.71 (0.20–2.58)	<b>0.15</b> <b>(0.04–0.63)</b> **	0.84 (0.31–2.29)	NA	<b>0.08</b> <b>(0.03–0.23)****</b>	0.94 (0.35–2.54)
<b>5.81</b> <b>(1.76–19.23)</b> **	2.58 (0.79–8.48)	<b>9.17</b> <b>(1.95–47.48)</b> **	1.97 (0.39–10.10)	<b>10.87</b> <b>(2.93–40.00)***</b>	<b>12.93</b> <b>(4.30–38.88)****</b>	NA	<b>12.20</b> <b>(3.09–47.62)***</b>
0.48 (0.15–1.50)	<b>0.21</b> <b>(0.07–0.66)</b> **	0.76 (0.17–3.31)	<b>0.16</b> <b>(0.03–0.82)*</b>	0.90 (0.26–3.06)	1.07 (0.39–2.88)	<b>0.08</b> <b>(0.02–0.32)***</b>	NA

**Discussion**

Immunohistochemistry has evolved into an indispensable diagnostic tool for the selection of NSCLC patients for targeted therapies, because of its low costs and fast turn-around times.<sup>6</sup> However, IHC is reported to lack standardization, causing a risk of suboptimal technical performance, which potentially leads to incorrect interpretations.<sup>24</sup> This study showed an improvement in staining performance over time, with varying performance for the different protocols for ALK IHC but not for ROS1. The staining performance was affected by several laboratory characteristics. The

outcome of the samples was influenced by the staining performance as well as the participant’s interpretation.

**Influence of Laboratory Characteristics on Expert Staining Score**

Our results clearly demonstrate the importance of EQA participation in reaching high-quality staining for ROS1, as laboratories obtained higher ESS when they participated in successive EQA schemes. This suggests the educational value of comparison to peers, the availability of individual feedback on staining quality, and examples of good-quality



**Table 4** ROS1 ESS for Different Combinations of Primary Antibodies, Antigen Retrieval, and Detection Kits

Primary antibody	Antigen retrieval kit	Detection system	Times used (n = 137) n (%)	Average ESS on 5 points ± SD	Method code	OR (95% CI) relative to method a
D4D6 (Cell Signaling Technology)	Cc1 (Ventana)	OptiView DAB IHC Detection Kit (Ventana)	66 (48.2)	4.3 ± 0.7	a	NA
		UltraView Universal DAB Detection kit (Ventana)	13 (9.5)	3.8 ± 0.7	b	<b>0.29 (0.11–0.75)*</b>
	Omnis Envision FLEX TRS, High pH (Dako)	Envision flex (Dako)	13 (9.5)	3.9 ± 1.1	c	0.43 (0.15–1.23)
	PT module TRS High envision Flex (Dako)		10 (7.3)	3.8 ± 0.6	d	<b>0.24 (0.11–0.53)***</b>
	Bond Epitope Retrieval 2 (Leica)	Bond polymer refine detection system (Leica)	8 (5.8–)	4.1 ± 0.6	e	0.50 (0.16–1.59)
	LDT (TRIS-) EDTA (with/without pressure cooker)	Different detection systems	11 (8.0)	4.0 ± 0.9	f	0.70 (0.17–2.92)
	Different combinations antigen retrieval/detection		10 (7.3)	3.9 ± 0.9	g	0.77 (0.22–2.61)
Other antibodies (3)	Different combinations antigen retrieval/detection		6 (4.4)	3.9 ± 0.8	h	0.58 (0.13–2.62)

(table continues)

Proportional odds models with generalized estimating equations for clustering of the data were used to analyze the difference in ESS. Differences in ESS are represented as ORs (95% CIs) for every method (row level) relative to other methods used (column level). OR >1 represents a higher ESS for a given method (column level) relative to the other method (row level). OR <1 represents a lower ESS for a method relative to other methods. Significant results are highlighted in bold.

\* $P < 0.05$ , \*\*\* $P < 0.001$ .

ESS, expert staining score; LDT, laboratory-developed test; NA, not applicable; OR, odds ratio; ROS1, ROS proto-oncogene 1.

stains and protocols. Similar increased performances between rounds have also been reported for estrogen receptor IHC by NordiQC,<sup>24</sup> whereas they also observed a surprising decrease in pass rate for ALK ([https://www.nordiqc.org/downloads/assessments/122\\_14.pdf](https://www.nordiqc.org/downloads/assessments/122_14.pdf), last accessed July 20, 2020). The lack of significant improvement for ALK might be explained by the fact that laboratories have been testing for this marker much longer and scores have stabilized after 2015, coinciding with the introduction of the ALK D5F3 CDx kit.

Recently, it was shown that accredited and research laboratories, in comparison to nonaccredited laboratories, more swiftly achieve a better performance during implementation of novel markers in routine practice.<sup>25</sup> The current study demonstrated that accreditation also positively affects the technical ROS1 staining quality, but no difference was observed with regard to laboratory setting (research setting versus private or community hospitals). Nevertheless, ROS1 seemed to be more frequently performed in research institutes than ALK, which was more frequently represented in this EQA by laboratories based in general hospitals. In addition, more staff members were involved for ROS1 compared with ALK IHC testing, even though this does not mean that more individuals were involved in the interpretation. As the number of staff members involved in the complete test process was related to the number of samples tested annually, this characteristic was used as a measure of the size of the laboratory and available expertise/resources.

The higher number of personnel involved for ROS1 can be explained as the research institutes, which more frequently tested ROS1 compared with other institutes, also reported to have a higher number of staff members involved in the test. This is in contrast to the general institutes, which more frequently reported to only test for ALK and reported fewer personnel. However, it cannot be excluded that the individual completing the survey affected the responses received. These findings may suggest that ROS1 IHC is considered to require more follow-up and practice, although more data are needed to investigate the interrelationship of different laboratory characteristics. Nevertheless, laboratories are performing well for this marker and have improved over time. Staining quality did not differ significantly between the applied protocols.

ROS1 IHC readout has been reported to be more difficult to interpret and operator dependent compared with ALK IHC for several reasons. For instance, ROS1 expression can be seen in a patchy pattern, typically at weak intensity in up to a third of tumors that do not have an underlying rearrangement.<sup>1</sup> Also, benign type 2 pneumocytes may show focal positivity with ROS1, in contrast to ALK IHC, which is negative in normal lung tissue.<sup>26</sup>

#### Expert Staining Score for the Different Protocols

This study confirmed the high variation in detection protocols applied by laboratories worldwide, with 16 different

**Table 4** (continued)

OR (95% CI) relative to method							
b	c	d	e	f	g	h	
<b>3.44 (1.33–8.90)*</b>	2.34 (0.82–6.71)	<b>4.12 (1.89–8.99)***</b>	1.99 (0.63–6.25)	1.43 (0.34–5.94)	1.31 (0.38–4.46)	1.73 (0.38–7.85)	
NA	0.68 (0.19–2.44)	1.20 (0.43–3.34)	0.58 (0.15–2.17)	0.41 (0.09–2.01)	0.38 (0.10–1.51)	0.50 (0.10–2.49)	
1.47 (0.41–5.26)	NA	1.76 (0.56–5.54)	0.85 (0.20–3.51)	0.61 (0.12–3.20)	0.56 (0.13–2.44)	0.74 (0.14–3.95)	
0.83 (0.30–2.33)	0.57 (0.18–1.79)	NA	0.48 (0.15–1.60)	0.35 (0.08–1.52)	0.32 (0.09–1.13)	0.42 (0.09–1.88)	
1.74 (0.46–6.54)	1.18 (0.29–4.90)	2.08 (0.63–6.90)	NA	0.72 (0.13–3.96)	0.66 (0.14–3.04)	0.87 (0.16–4.88)	
2.42 (0.50–11.77)	1.65 (0.31–8.62)	2.90 (0.66–12.82)	1.40 (0.25–7.69)	NA	0.92 (0.16–5.29)	1.21 (0.18–8.33)	
2.64 (0.66–10.53)	1.80 (0.41–7.87)	3.17 (0.89–11.36)	1.52 (0.33–7.04)	1.09 (0.19–6.29)	NA	1.33 (0.23–7.75)	
1.99 (0.40–9.90)	1.36 (0.25–7.25)	2.39 (0.53–10.75)	1.15 (0.21–6.44)	0.82 (0.12–5.64)	0.76 (0.13–4.42)	NA	

commercial ALK antibodies used in combination with several antigen retrieval methods and detection kits. US Food and Drug Administration and European Conformity in Vitro Diagnostic approved kits reached a higher ESS compared with LDTs. This is in line with results reported by NordiQC, where LDT performance for human epidermal growth factor receptor 2 IHC improved over time but remained suboptimal compared with approved tests,<sup>24</sup> and where adaptations by the laboratory to the ready-to-use systems resulted in lower pass rates for ALK ([https://www.nordiqc.org/downloads/assessments/122\\_14.pdf](https://www.nordiqc.org/downloads/assessments/122_14.pdf), last accessed July 20, 2020). An explanation could be that the use of a US Food and Drug Administration–approved test restricts the performing laboratory to a specified reagent set and a closed protocol, and to an internal validation process that provides some assurance of run-to-run stability. For good laboratory practice, any deviation from the standard reagents or protocol requires appropriate validation. Deviations from the protocol have previously been shown to have negative effects on the ability to demonstrate proteins.<sup>27</sup> However, in general, more cases were tested by approved kits in the current study, which may have contributed to the observed difference in performance. Reasons for underperformance of LDTs might include the lack of comparative literature on sensitivity and specificity for less common antibodies, as well as unavailability of training by the respective manufacturers.

The most widely used antibodies included D5F3 and 5A4. D5F3 (Ventana) is part of an *in vitro*

Diagnostic–labeled kit. This antibody demonstrated similar performances compared with the D5F3 clone from another manufacturer (Cell Signaling Technology) and the 5A4 (Novocastra) clone. Results indicated a statistically better performance for the 1A4 (Origene) primary antibody, even though this antibody has previously been reported to have a lower specificity compared with D5F3.<sup>2</sup> In case of equivocal results by 1A4, confirmation by an additional independent method has therefore been advised.<sup>2</sup> However, compared with the percentage of D5F3 users (39.8%), the percentage of 1A4 users is relatively small (5.4%), and more data are needed to confirm this statement. In contrast, it appeared that application of less common antibodies, such as 5A4 (Abcam), ALK1 (Dako), or ALK01 (Ventana), resulted in a lower ESS. For ALK1, this is not surprising, as several studies demonstrated the lower performance for this clone, and this clone is therefore not recommended for use in NSCLC.<sup>1,28</sup> These findings are in line with quality assessment results from NordiQC and UK National External Quality Assessment Scheme, where lower performances were observed for ALK1.<sup>10</sup> It must be taken into consideration that the ESS as assessed on slides is a combination of both the used primary antibody clone and all subsequent protocol steps. This is the reason to present the difference in overall ESS in Table 3.

A difference in antibody performance could therefore also be attributed to the method of antigen retrieval and detection systems. Indeed, this was exemplified by the

**Table 5** Average ESS for Less Common *ALK* and *ROS1* IHC Protocols

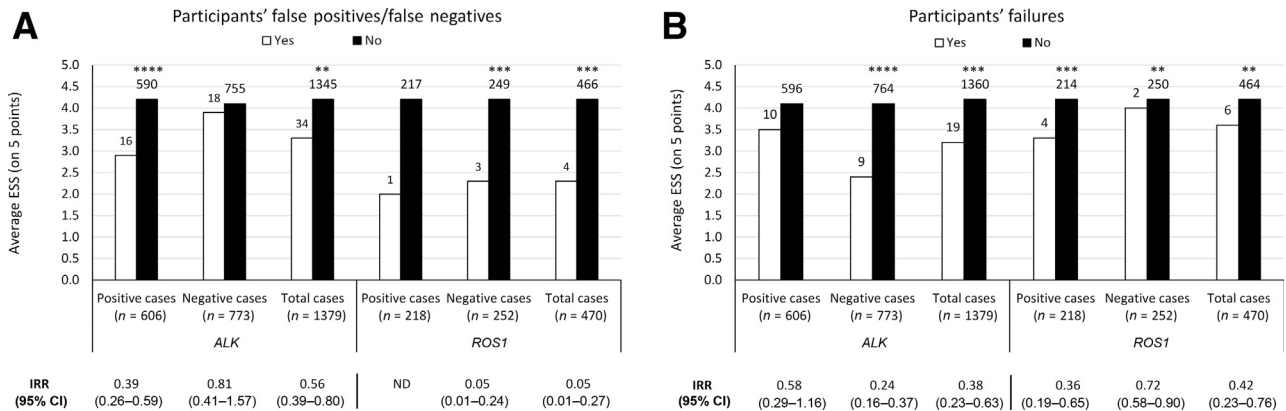
Antibody	Antigen retrieval	Detection platform	Users, <i>n</i>	Users, %	Average ESS (on 5 points)
ALK ( <i>n</i> = 352)					
D5F3 (Genemed)	Bond Epitope Retrieval 2 (Leica)	Bond polymer refine detection system (Leica)	1	0.3	3.0
5A4 (Biocare Medical)	Cc1 (Ventana)	UltraView Universal DAB Detection kit (Ventana)	1	0.3	4.0
	LDT EDTA or TRIS-EDTA (with/without pressure cooker)	Powervision (Immunovision Technologies)	2	0.6	3.5
5A4 (Clinisciences)	Cc1 (Ventana)	OptiView DAB IHC Detection Kit (Ventana)	3	0.9	4.3
		UltraView Universal DAB Detection kit (Ventana)	1	0.3	3.0
	LDT EDTA or TRIS-EDTA (with/without pressure cooker)	Bond polymer refine detection system (Leica)	1	0.3	3.0
5A4 (Diagnostic Biosystems)	Cc1 (Ventana)	UltraView Universal DAB Detection kit (Ventana)	1	0.3	3.0
5A4 (Histofine)	HISTOFINE ALK Detection KIT	HISTOFINE ALK Detection KIT	1	0.3	4.0
5A4 (Medac)	LDT EDTA or TRIS-EDTA (with/without pressure cooker)	ZytoChem Plus (HRP) Polymer Kit (Zytomed)	1	0.3	3.0
5A4 (Monosan)	Cc1 (Ventana)	OptiView DAB IHC Detection Kit (Ventana)	5	1.4	4.4
		DAKO Omnis Envision FLEX TRS, High pH	1	0.3	5.0
5A4 (Zytomed)	Cc1 (Ventana)	OptiView DAB IHC Detection Kit (Ventana)	2	0.6	3.5
		Novolink Polymer Detection System (Leica)	3	0.9	5.0
	LDT EDTA or TRIS-EDTA (with/without pressure cooker)	ZytoChem Plus (HRP) Polymer Kit (Zytomed)	1	0.3	3.0
1A4 (Zytomed)	Cc1 (Ventana)	UltraView Universal Alkaline Phosphatase Red Detection Kit (Ventana)	1	0.3	4.0
	LDT EDTA or TRIS-EDTA (with/without pressure cooker)	ZytoChem Plus (HRP) Polymer Kit (Zytomed)	1	0.3	2.0
ALK01 (Ventana)	Cc1 (Ventana)	OptiView DAB IHC Detection Kit (Ventana)	4	1.1	3.0
		UltraView Universal DAB Detection kit (Ventana)	1	0.3	2.0
ALK1 (Dako)	Bond Epitope Retrieval 1 (Leica)	Bond polymer refine detection system (Leica)	1	0.3	1.0
	Cc1 (Ventana)	OptiView DAB IHC Detection Kit (Ventana)	1	0.3	4.0
		DAKO Omnis Envision FLEX TRS, High pH	1	0.3	2.0
		EnVisionFLEX Target Retrieval Solution, low pH (Dako Omnis)	1	0.3	1.0
ROS1 ( <i>n</i> = 137)					
D4D6 (Bioké)	Cc1 (Ventana)	OptiView DAB IHC Detection Kit (Ventana)	1	2.9	3.0
EP282 (Epitomics)			3	8.6	4.7
D4D6 (Genemed)		UltraView Universal DAB Detection kit (Ventana)	2	5.7	4.0

During the external quality assessment scheme, an ESS of  $\geq 3$  on a total of 5 points was considered acceptable.

*ALK*, anaplastic lymphoma kinase; ESS, expert staining score; IHC, immunohistochemistry; LDT, laboratory-developed test; *ROS1*, ROS proto-oncogene 1.

lower ESS for homebrew antigen retrieval reagents and the ZytoChem Plus (HRP) Polymer Kit (Zytomed, Barga-teheide, Germany) for detection of ALK. Also, the 5A4

(Novocastra) performed equally well as other primary antibodies in combination with the three most common detection kits [OptiView DAB IHC Detection Kit



**Figure 1** Average anaplastic lymphoma kinase (*ALK*) and ROS proto-oncogene 1 (*ROS1*) expert staining score (ESS) related to participants' scoring incidence of false positives/false negatives (A) and analysis failures (B). Bar labels represent the number of cases with or without incorrect interpretations/analysis failures observed. Poisson models with generalized estimating equations were used to analyze the association of the ESS with the number of incorrect interpretations (false-positive and false-negative results; A) and the number of analysis failures (B) by the participants, observed in the external quality assessment (EQA) schemes as count outcome variables. Results are presented as incidence rate ratio (IRR) (95% CI), taking into account the log of the total number of samples analyzed during the EQA scheme as an offset variable. IRRs <1 represent a lower number of incorrect interpretations or analysis failures for higher ESS. The IRR for incorrect interpretations in *ROS1*-positive cases was not computed as only one error occurred.  $^{*}P < 0.01$ ,  $^{***}P < 0.001$ , and  $^{****}P < 0.0001$ . ND, not determined.

(Ventana), Envision Flex (Dako), or Bond polymer refine detection system (Leica, Nussloch, Germany)], but worse when using less common detection methods (Table 3). The participants with an incubation time >60 minutes also used antibody clones that demonstrated a lower performance in this study.

For the incubation temperature, it might be useful to evaluate the difference between 32°C and 36°C to 37°C in future schemes, as they represent frequently used cutoffs in routine practice. This was not performed in this study as the number of laboratories using 32°C was too low to make valid assumptions (9/351 for *ALK* and 4/137 for *ROS1*).

Because not all participants provided their in-house control tissues, these were not evaluated and might be considered a limitation of this study. The use of appropriate positive and negative controls by participants to evaluate the sensitivity and specificity of the tests is advised and has previously been reported to be highly variable and often suboptimal.<sup>10</sup> In the currently ongoing EQA scheme, an additional request was made to the participants to submit control tissues/slides as well. Also, the next schemes will allow an evaluation of the *ROS1* SP384 (Ventana) or 1A1 (Origene)<sup>14,15</sup> antibodies, which were not yet available during the time of the EQA schemes interrogated for this study.

In contrast to *ALK* IHC, no European Conformity in Vitro Diagnostic certified method was available for *ROS1* analysis at the time of these EQA schemes. Most participants (95.6%) used the D4D6 antibody (Cell Signaling Technology). A variety of protocols were reported, although most frequently in combination with OptiView DAB IHC Detection Kit (Ventana) (Table 4). This combination revealed a better performance compared with other test methods, although only significant compared with

UltraView and PT module TRS High envision Flex (Dako). It is notable that a few laboratories reported the use of other clones, including D4D6 (Genemed, Torrance, CA), D4D6 (Bioké, Leiden, the Netherlands), and EP282 (Epitomics, Cambridge, UK), and these did not differ compared with D4D6 (Cell Signaling Technology) concerning the ESS. *ROS1* testing has been more recently introduced than *ALK*, and international testing guidelines were first made available in 2016.<sup>29</sup> As such, less information is currently available compared with *ALK*, and more interlaboratory studies are needed on the performance of detection protocols.

#### Participants' Scoring of *ALK* and *ROS1* Expression in the EQA Samples

As a lower staining performance (as defined by the ESS) resulted in an increased rate of misinterpretations, technical schemes can be useful in guiding participants in further optimizing their detection protocol through individually tailored feedback, ultimately also resulting in an improvement in their scoring. Detection of positive cases (which might be rare in case of *ALK* and *ROS1* rearrangements in NSCLC) is of utmost importance, as a false-negative outcome could result in the loss of chance to benefit from optimal treatment for patients.<sup>4</sup> False-positive results, on the other hand, still have a chance to be corrected in case FISH confirmation is performed, which is not required in case of using an approved *ALK* IHC method.

The overall number of false-positive and false-negative results observed in the EQA schemes was low, with 34 of 1379 (2.5%) for *ALK* and 19 of 470 (4.0%) for *ROS1*. Surprisingly, the number of total incorrect *ALK* IHC interpretations was significantly higher ( $P = 0.0355$ ) for later EQA scheme years, but not when evaluating positive and

negative samples separately. This was explained by the fact that in the later scheme years, two laboratories made multiple interpretation errors on the set of five or three samples, affecting both positive and negative cases. In 2016, these laboratories denoted all five samples as negative or positive, respectively, suggesting a problem with their IHC interpretation criteria and cutoff. In 2017, both laboratories switched the interpretation for a positive and negative case. In other scheme years or for ROS1 IHC, only one error per laboratory was made. As the overall misinterpretations were low, these two laboratories with multiple errors have skewed the data toward higher error rates in later schemes, even though the number of laboratories making an error remained stable.

Because of the low percentage of incorrect test interpretations reported by participants, statistics on the incidence of these errors were not calculated for different protocols. However, for LDTs, 3.5% (25/802 for ALK and 19/470 for ROS1) of misinterpretations occurred compared with 1.6% for CDx methods (9/577, ALK only).

Besides incorrect false-positive or false-negative sample scoring, technical failures (failed run or suboptimal staining leading to the inability to safely interpret the staining pattern) were significantly related to the ESS. Indeed, a change in the test protocol within the last 12 months introduced more analysis failures for the concerned participants ([Supplemental Table S3](#)), and this suggests the need for careful quality control in case of newly introduced methods.

Repeated testing might cause additional costs, a delay in the final diagnosis, loss of tissue, and, in cases where the test cannot be repeated because of lack of material (as biopsy specimens from NSCLC tumors are often minimal in material), patient distress because of the need for an additional intervention.

In this study, a decrease in errors/technical failures was observed for both markers depending on the number of samples tested annually. This is consistent to previously reported findings in which the probability to have a successful EQA score increased when a higher number of samples was tested annually for *KRAS*, *NRAS*, or *EGFR* mutation status.<sup>25</sup> Thus, even though the staining quality was not affected by a laboratory's experience (annual test volume) in this study, the interpretation of the staining intensity is influenced by experience. Even if the IHC test is completely standardized, interpretation is likely to always be subjective, in part because it requires comparison with appropriate controls, and these may differ for the various antibodies and detection kits used.

## Conclusion

To conclude, a good ESS in the EQA scheme only represents one aspect of the total quality assurance system for IHC, and might not fully reflect staining quality and consistency in routine practice. Nevertheless, this study was able to reveal the high variability of IHC methods, where the

use of restricted protocols by approved kits led to improved performance, in contrast to lower performances for less commonly used protocols. In addition, although *ALK* testing seems to be well implemented in general practice, *ROS1* detection is still improving. More interlaboratory studies are needed to assess performance for different ROS1 IHC protocols. The correlation between the experts' scoring and participants' interpretation, and variable effects of laboratory characteristics on both aspects, stress the importance for EQA providers to evaluate both the staining quality and the interpretation of the staining.

## Acknowledgments

This project would not have been possible without the support of the participating laboratories in the 2015 to 2018 external quality assessment schemes, the European Society of Pathology for the organization of the scheme and administrative support, Ivonne Marondel (Pfizer Oncology) for the unrestricted research grant for coordination of the schemes; colleagues of the BQA Research Unit for the coordination and administrative support; Véronique Tack for the validation of laboratory setting and accreditation in 2015 to 2016 and conceptualization of the laboratory setting taxonomy; the scheme experts, assessors, and members of the steering committee; the laboratories responsible for cutting and labeling of the samples between 2015 and 2018; the reference laboratories involved in validation of the immunohistochemistry samples in 2015 to 2018; University Hospital Antwerp and University Hospital Leuven for the use of the multihead microscope; and Annouschka Laenen, Leuven Biostatistics and Statistical Bio-informatics Center, and the Leuven Cancer Institute for performing the statistical analyses

## Author Contributions

C.K. and E.M.C.D. collected data according to ISO17043 and performed statistical analysis; C.K., E.M.C.D., J.v.d.T., and E.S. interpreted the data; J.v.d.T. selected and scanned the (sub)optimal staining examples; J.v.d.T. and N.t.H. provided medical expertise during the external quality assessment (EQA) schemes and were responsible for sample selection; E.S. was the technical expert and validated the samples during the EQA schemes; A.R. and K.M. prepared the samples for shipment (cutting and labeling); P.P., A.R., N.t.H., K.M., and E.T. conceived and designed the technical assessment; P.P., A.R., N.t.H., K.M., E.T., and K.Z. took part as technical assessors in one or more of the schemes; and C.K., J.v.d.T., P.P., A.R., N.t.H., E.S., and K.Z. were involved as assessors for the analysis outcomes of immunohistochemistry and fluorescence *in situ* hybridization. All authors critically revised the manuscript for important intellectual content.

## Supplemental Data

Supplemental material for this article can be found at <https://doi.org/10.1016/j.jmoldx.2020.09.006>.

## References

- Lindeman NI, Cagle PT, Aisner DL, Arcila ME, Beasley MB, Bernicker EH, Colasacco C, Dacic S, Hirsch FR, Kerr K, Kwiatkowski DJ, Ladanyi M, Nowak JA, Sholl L, Temple-Smolkin R, Solomon B, Souter LH, Thunnissen E, Tsao MS, Ventura CB, Wynes MW, Yatabe Y: Updated molecular testing guideline for the selection of lung cancer patients for treatment with targeted tyrosine kinase inhibitors: guideline from the College of American Pathologists, the International Association for the Study of Lung Cancer, and the Association for Molecular Pathology. *Arch Pathol Lab Med* 2018, 142: 321–346
- Tsao MSHF, Yatabe Y: IASLC Atlas of ALK and ROS1 Testing in Lung Cancer. International Association for the Study of Lung Cancer. North Fort Myers, FL: Editorial Rx Press, 2016
- Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, Fujiwara S, Watanabe H, Kurashina K, Hatanaka H, Bando M, Ohno S, Ishikawa Y, Aburatani H, Niki T, Sohara Y, Sugiyama Y, Mano H: Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* 2007, 448:561–566
- Shaw AT, Solomon BJ: Crizotinib in ROS1-rearranged non-small-cell lung cancer. *N Engl J Med* 2015, 372:683–684
- Drilon A, Siena S, Dziadziuszko R, Barlesi F, Krebs MG, Shaw AT, et al; Trial investigators: Entrectinib in ROS1 fusion-positive non-small-cell lung cancer: integrated analysis of three phase 1-2 trials. *Lancet Oncol* 2020, 21:261–270
- Du X, Shao Y, Qin HF, Tai YH, Gao HJ: ALK-rearrangement in non-small-cell lung cancer (NSCLC). *Thorac Cancer* 2018, 9:423–430
- Wynes MW, Sholl LM, Dietel M, Schuurin E, Tsao MS, Yatabe Y, Tubbs RR, Hirsch FR: An international interpretation study using the ALK IHC antibody D5F3 and a sensitive detection kit demonstrates high concordance between ALK IHC and ALK FISH and between evaluators. *J Thorac Oncol* 2014, 9:631–638
- van der Wekken AJ, Pelgrim R, 't Hart N, Werner N, Mastik MF, Hendriks L, van der Heijden EHF, Looijen-Salamon M, de Langen AJ, Staal-van den Brekel J, Riemersma S, van den Borne BE, Speel EJM, Dingemans AC, Hiltermann TJN, van den Berg A, Timens W, Schuurin E, Groen HJM: Dichotomous ALK-IHC is a better predictor for ALK inhibition outcome than traditional ALK-FISH in advanced non-small cell lung cancer. *Clin Cancer Res* 2017, 23:4251–4258
- Thunnissen E, Lissenberg-Witte BI, van den Heuvel MM, Monkhorst K, Skov BG, Sørensen JB, Mellemgaard A, Dingemans AMC, Speel EJM, de Langen AJ, Hashemi SMS, Bahce I, van der Drift MA, Looijen-Salamon MG, Gosney J, Postmus PE, Samii SMS, Duplaquet F, Weynand B, Durando X, Penault-Llorca F, Finn S, Grady AO, Oz B, Akyurek N, Buettner R, Wolf J, Bubendorf L, Duin S, Marondel I, Heukamp LC, Timens W, Schuurin EMD, Pauwels P, Smit EF: ALK immunohistochemistry positive, FISH negative NSCLC is infrequent, but associated with impaired survival following treatment with crizotinib. *Lung Cancer* 2019, 138:13–18
- Ibrahim M, Parry S, Wilkinson D, Bilbe N, Allen D, Forrest S, Maxwell P, O'Grady A, Starczynski J, Tanier P, Gosney J, Kerr K, Miller K, Thunnissen E: ALK immunohistochemistry in NSCLC: discordant staining can impact patient treatment regimen. *J Thorac Oncol* 2016, 11:2241–2247
- Thunnissen E, Bubendorf L, Dietel M, Elmberger G, Kerr K, Lopez-Rios F, Moch H, Olszewski W, Pauwels P, Penault-Llorca F, Rossi G: EML4-ALK testing in non-small cell carcinomas of the lung: a review with recommendations. *Virchows Arch* 2012, 461: 245–257
- Jiang L, Yang H, He P, Liang W, Zhang J, Li J, Liu Y, He J: Improving selection criteria for ALK inhibitor therapy in non-small cell lung cancer: a pooled-data analysis on diagnostic operating characteristics of immunohistochemistry. *Am J Surg Pathol* 2016, 40: 697–703
- Pyo JS, Kang G, Sohn JH: ALK immunohistochemistry for ALK gene rearrangement screening in non-small cell lung cancer: a systematic review and meta-analysis. *Int J Biol Markers* 2016, 31: e413–e421
- Wang W, Cheng G, Zhang G, Song Z: Evaluation of a new diagnostic immunohistochemistry approach for ROS1 rearrangement in non-small cell lung cancer. *Lung Cancer* 2020, 146:224–229
- Conde E, Hernandez S, Martinez R, Angulo B, De Castro J, Collazo-Lorduy A, et al: Assessment of a new ROS1 immunohistochemistry clone (SP384) for the identification of ROS1 rearrangements in patients with non-small cell lung carcinoma: the ROSING study. *J Thorac Oncol* 2019, 14:2120–2132
- Sholl LM, Sun H, Butaney M, Zhang C, Lee C, Jänne PA, Rodig SJ: ROS1 immunohistochemistry for detection of ROS1-rearranged lung adenocarcinomas. *Am J Surg Pathol* 2013, 37:1441–1449
- Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018, 68:394–424
- O'Hurley G, Sjöstedt E, Rahman A, Li B, Kampf C, Pontén F, Gallagher WM, Lindskog C: Garbage in, garbage out: a critical evaluation of strategies used for validation of immunohistochemical biomarkers. *Mol Oncol* 2014, 8:783–798
- Howat WJ, Lewis A, Jones P, Kampf C, Pontén F, van der Loos CM, Gray N, Womack C, Warford A: Antibody validation of immunohistochemistry for biomarker discovery: recommendations of a consortium of academic and pharmaceutical based histopathology researchers. *Methods* 2014, 70:34–38
- Nielsen S: External quality assessment for immunohistochemistry: experiences from NordiQC. *Biotech Histochem* 2015, 90:331–340
- Tembuysen L, Tack V, Zwaenepoel K, Pauwels P, Miller K, Bubendorf L, Kerr K, Schuurin E, Thunnissen E, Dequeker EM: The relevance of external quality assessment for molecular testing for ALK positive non-small cell lung cancer: results from two pilot rounds show room for optimization. *PLoS One* 2014, 9: e112159
- von Laffert M, Penzel R, Schirmacher P, Warth A, Lenze D, Hummel M, Dietel M: Multicenter ALK testing in non-small-cell lung cancer: results of a round robin test. *J Thorac Oncol* 2014, 9: 1464–1469
- Keppens C, Tack V, Hart N, Tembuysen L, Ryska A, Pauwels P, Zwaenepoel K, Schuurin E, Cabillie F, Tornillo L, Warth A, Weichert W, Dequeker E: EQA assessors expert group: a stitch in time saves nine: external quality assessment rounds demonstrate improved quality of biomarker analysis in lung cancer. *Oncotarget* 2018, 9: 20524–20538
- Vyberg M, Nielsen S: Proficiency testing in immunohistochemistry—experiences from Nordic Immunohistochemical Quality Control (NordiQC). *Virchows Arch* 2016, 468:19–29
- Tack V, Schuurin E, Keppens C, 't Hart N, Pauwels P, van Krieken H, Dequeker EMC: Accreditation, setting and experience as indicators to assure quality in oncology biomarker testing laboratories. *Br J Cancer* 2018, 119:605–614
- Lin JJ, Shaw AT: Recent advances in targeting ROS1 in lung cancer. *J Thorac Oncol* 2017, 12:1611–1625
- Taylor CR: Predictive biomarkers and companion diagnostics: the future of immunohistochemistry: “in situ proteomics,” or just a “stain”? *Appl Immunohistochem Mol Morphol* 2014, 22: 555–561

28. Thunnissen E, Allen TC, Adam J, Aisner DL, Beasley MB, Borczuk AC, Cagle PT, Capelozzi VL, Cooper W, Hariri LP, Kern I, Lantuejoul S, Miller R, Mino-Kenudson M, Radonic T, Raparia K, Rekhtman N, Roy-Chowdhuri S, Russell P, Schneider F, Sholl LM, Tsao MS, Vivero M, Yatabe Y: Immunohistochemistry of pulmonary biomarkers: a perspective from members of the Pulmonary Pathology Society. *Arch Pathol Lab Med* 2018, 142: 408–419
29. Bubendorf L, Büttner R, Al-Dayel F, Dietel M, Elmberger G, Kerr K, López-Ríos F, Marchetti A, Öz B, Pauwels P, Penault-Llorca F, Rossi G, Ryska A, Thunnissen E: Testing for ROS1 in non-small cell lung cancer: a review with recommendations. *Virchows Arch* 2016, 469:489–503