ORIGINAL PAPER

# Interlaboratory agreement in assessment of gynaecological cytology in Cervical Cancer Screening Programme in Poland – a pilot evaluation

Kinga Zalewska-Otwinowska[1], Anna Macios[1,2], Katarzyna Komerska[1], Małgorzata Rekosz[1], Andrzej Nowakowski[1]

[1]Department of Cancer Prevention, Maria Sklodowska-Curie National Research Institute of Oncology, Warsaw, Poland
[2]Department of Gastroenterology, Hepatology and Clinical Oncology, Centre of Postgraduate Medical Education, Warsaw, Poland

In our pilot study we have aimed to assess interlaboratory variability of cytological diagnoses in selected laboratories participating in the Polish Cervical Cancer Screening Programme (CCSP) to establish grounds for certification system for cytodiagnosticians and to monitor the quality of services. Set of 50 selected Pap smears, previously reassessed by an expert on the grounds of clinical, colposcopic and histological data was blinded and sent to 15 laboratories in Poland with request for evaluation according to routine practice according to the Bethesda 2001 system. Concordance with expert diagnoses reached a median of 82% (range: 66% to 92%), with median unweighted κ coefficient at κ = 0.67 (range 0.40 to 0.86) depending on laboratory. This indicates substantial agreement among laboratories, however with essential differences in proper evaluation in some outlying laboratories. Agreement was highest in samples with high-grade, lower for low-grade abnormalities. Slides with ASC-US and ASC-H expert diagnoses were most troubling for cytodiagnosticians. Sets of highly selected cytological slides with expert diagnoses may serve as a tool in the process of comprehensive periodic recertification of cytodiagnosticians in the screening programme. A benchmark level of agreement with expert diagnoses should be established to guide corrective actions for cytodiagnosticians with lowest agreement.

**Key words:** Pap test, cervical cancer screening, interlaboratory agreement, cytodiagnosticians.

## Introduction

Cervical cancer (CC) is the fourth most common cancer and cancer-related cause of death in women worldwide [1] and it is still an unsolved epidemiological problem especially in low- and middle-income countries [2]. In 2017 in Poland 2,502 new cases and 1,609 deaths due CC were reported by the National Cancer Registry which translates into world-age standardized incidence and mortality ratios in wom-

en of 7.7 and 4.2 per 100,000, respectively [3]. In Poland, more than half of all CC cases occur among women between 45 and 64 years old. Incidence and mortality have been decreasing for the last few decades in women under the age of 60 but these rates are still stable in older women [4]. Also 5-years relative CC-related survival rates have remained unchanged throughout last decade [5].

Screening for CC is recommended by European Commission since 2003 as a form of secondary

prevention [6] and has the potential to greatly reduce both CC incidence and mortality through early identification of precancerous lesions which can be treated far more effectively than in case of symptomatic invasive disease [7]. In most of the developed countries which run organised/opportunistic screening programmes, exfoliative cytology (PAP test) is still the basic screening test. Even in countries which switched to high risk human papillomavirus (hr-HPV) test as a more sensitive screening method, PAP tests are an integral part of triage algorithms after a positive hr-HPV test [8, 9, 10, 11].

All types of cytological procedures are affected by the problem of ambiguous results that hinder the implementation of proper diagnostic and therapeutic procedures. As a result of a multidisciplinary effort, in 1988 The Bethesda System (TBS) was developed in order to create a standardized framework for cervical cytology reports (including evaluation of specimen adequacy, optional general categorization, descriptive diagnosis) [12, 13].

TBS was accepted internationally. The system was revised in 1991, 2001 and 2014 in reaction to evolution of management and new research on cervical cancer [13]. TBS for assessing abnormalities in the cervix and in the thyroid gland was very readily accepted by pathologists and clinicians, and the presented classifications created the foundation for a unified clinical procedure [12].

In 2006 in Poland, an organised CC screening programme for women at the age between 25 and 59 was initiated. Pap test is offered every 3 years, in accordance with the guidelines of the European Commission and Polish Gynaecological Society [7, 14]. In Poland, by law, cytopathologists (medical doctors with specialty in pathology) and cytodiagnosticians (laboratory diagnosticians with specialty in medical cytomorphology) are entitled to assess cytological samples. In OCCSP all Pap tests with abnormal results must be approved be a senior cytodiagnostician or a cytopathologist. Results are coded according to modified TBS 2001. Information about screening results is stored in an IT System for Prevention Monitoring (pol. System Informatyczny Monitorowania Profilaktyki, SIMP) and National Cancer Registry (NCR) collects reports of cancer diagnosis. Algorithms for abnormal results are set and include repeated testing and colposcopy/biopsy.

Apart from the coverage of target population by the screening programme, quality assurance is a key factor for obtaining full impact of screening on CC epidemiology. Certification of laboratories and personnel is also important to maintain high quality of cytological cervices. Evaluation of agreement between expert cytological diagnosis and laboratories/cytodiagnosticians routinely working in CC screening programmes may be one of the components of certi-

fication process. It may also be important for quality assurance which plays a crucial role in cytology-based CC screening [7]. In Poland quality assurance is led by Central Coordination Centre (pol. COK – Centralny Ośrodek Koordynujący). Although cytological laboratories carrying out CC screening programme in Poland undergo certification, to our knowledge interlaboratory agreement has never been evaluated and is the subject of this pilot project.

## Material and methods

In 2018 COK carried out a pilot study to assess interlaboratory variability of cytological diagnoses in selected laboratories operating in the Cervical Cancer Screening Programme (CCSP). A set of 50 expert-selected conventional cytology slides with clinical, colposcopic and histological confirmation of diagnoses, collected in 2012-2015 was prepared. Slides originally diagnosed by cytodiagnosticians of the Cytological Laboratory of the Maria Sklodowska-Curie National Research Institute of Oncology in Warsaw, Poland, were reviewed again by an expert cytodiagnostician with 20 years of experience from the same lab. The expert confirmed cytological diagnoses taking into account clinical, colposcopic and histopathology reports and selected smears formed a set representative for each category of diagnoses according to the Bethesda 2001 system (expert diagnoses are presented in Table I). The laboratory is high-throughput, with internal quality-assurance and is co-operating with a Cervical Pathology Clinic performing large-scale screening, colposcopic/histological triage and treatment of women with cervical pathology also within the CCSP.

The set with each-time blinded and mixed slides was sent to 15 cytodiagnostic laboratories including 8 laboratories in Mazovian Voivodeship, 4 in Lublin Voivodeship, 2 in Świętokrzyskie Voivodeship and 1 in Łódź Voivodeship, all operating in the CCSP. Standard clinical data accompanied each slide. By the decision of the head in each lab, a cytodiagnostician or cytodiagnosticians were asked to evaluate the set smears over a week along with routine everyday practice. Each slide was to be assessed according to the Bethesda 2001 system. After the evaluation three types of coding were applied in COK to allow for agreement analyses as follows:

- *Coding 1 (general):* unsatisfactory for evaluation *vs* normal (no intraepithelial lesion or malignancy – NILM) vs. abnormal (atypical squamous cells of undetermined significance [ASC-US] or low-grade squamous intraepithelial lesion [LSIL] or atypical squamous cells cannot exclude HSIL [ASC-H] or high-grade squamous intraepithelial lesion [HSIL] or atypical glandular cells [AGC] or squamous cell carcinoma [SCC]);

**Table I.** Experts' and laboratories' evaluations of prepared set of cytological slides with calculated percentage of laboratories coherent with experts in specific slide evaluation

| Blinded slide number | Experts' diagnoses | Lab1 | Lab2 | Lab3 | Lab4 | Lab5 | Lab6 | Lab7 | Lab8 | Lab10 | Lab11 | Lab12 | Lab13 | Lab14 | Lab15 | Percentage of laboratories which evaluated slides coherently with experts' | | | Majority diagnosis |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | | | According to detailed coding | According to aggregated coding | According to general coding | |
| 1 | USFE | USFE | NILM | USFE | USFE | USFE | USFE | USFE | USFE | USFE | USFE | USFE | USFE | USFE | USFE | 85.7% | 85.7% | 85.7% | USFE |
| 2 | USFE | NILM | NILM | NILM | NILM | USFE | NILM | NILM | NILM | NILM | NILM | NILM | NILM | USFE | NILM | 14.3% | 14.3% | 14.3% | NILM |
| 3 | USFE | USFE | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | 7.1% | 7.1% | 7.1% | NILM |
| 4 | USFE | NILM | NILM | NILM | NILM | NILM | USFE | NILM | NILM | USFE | NILM | NILM | NILM | NILM | NILM | 14.3% | 14.3% | 14.3% | NILM |
| 5 | NILM | NILM | NILM | NILM | NILM | | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | 100.0% | 100.0% | 100.0% | NILM |
| 6 | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | ASCUS | NILM | ASCUS | NILM | NILM | NILM | NILM | 85.7% | 85.7% | 85.7% | NILM |
| 7 | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | 100.0% | 100.0% | 100.0% | NILM |
| 8 | NILM | ASCUS | NILM | NILM | NILM | ASCUS | NILM | NILM | NILM | NILM | ASCUS | NILM | NILM | NILM | NILM | 78.6% | 78.6% | 78.6% | NILM |
| 9 | NILM | NILM | NILM | NILM | AGC | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | 92.9% | 92.9% | 92.9% | NILM |
| 10 | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | 100.0% | 100.0% | 100.0% | NILM |
| 11 | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | ASCUS | ASCUS | ASCUS | NILM | NILM | NILM | NILM | 78.6% | 78.6% | 78.6% | NILM |
| 12 | NILM | NILM | NILM | AGC | AGC | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | 85.7% | 85.7% | 85.7% | NILM |
| 13 | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | 100.0% | 100.0% | 100.0% | NILM |
| 14 | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | 100.0% | 100.0% | 100.0% | NILM |
| 15 | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | 100.0% | 100.0% | 100.0% | NILM |
| 16 | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | 100.0% | 100.0% | 100.0% | NILM |
| 17 | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | 100.0% | 100.0% | 100.0% | NILM |
| 18 | NILM | NILM | NILM | NILM | NILM | NILM | AGC | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | 92.9% | 92.9% | 92.9% | NILM |
| 19 | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | 100.0% | 100.0% | 100.0% | NILM |
| 20 | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | 100.0% | 100.0% | 100.0% | NILM |
| 21 | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | 100.0% | 100.0% | 100.0% | NILM |
| 22 | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | 100.0% | 100.0% | 100.0% | NILM |
| 23 | NILM | NILM | NILM | NILM | NILM | USFE | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | 92.9% | 92.9% | 92.9% | NILM |
| 24 | NILM | NILM | NILM | NILM | USFE | NILM | NILM | NILM | NILM | NILM | NILM | NILM | ASCUS | NILM | NILM | 85.7% | 85.7% | 85.7% | NILM |
| 25 | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | 100.0% | 100.0% | 100.0% | NILM |

**Table I.** Cont.

| Blinded slide number | Experts' diagnoses | Lab1 | Lab2 | Lab3 | Lab4 | Lab5 | Lab6 | Lab7 | Lab8 | Lab10 | Lab11 | Lab12 | Lab13 | Lab14 | Lab15 | Percentage of laboratories which evaluated slides coherently with experts' | | | Majority diagnosis |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | | | According to detailed coding | According to aggregated coding | According to general coding | |
| 26 | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | AGC | AGC | NILM | NILM | NILM | NILM | NILM | 92.9% | 92.9% | 92.9% | NILM |
| 27 | ASCUS | ASCUS | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | NILM | 7.1% | 7.1% | 7.1% | NILM |
| 28 | ASCUS | ASCUS | ASCUS | ASCUS | NILM | NILM | NILM | ASCUS | ASCUS | ASCUS | ASCUS | NILM | ASCUS | NILM | NILM | 57.1% | 57.1% | 57.1% | ASCUS |
| 29 | ASCUS | HSIL | HSIL | AGC | NILM | ASCH | ASCH | ASCH | HSIL | HSIL | HSIL | ASCH | HSIL | HSIL | NILM | 0.0% | 0.0% | 85.7% | HSIL |
| 30 | ASCUS | LSIL | ASCUS | NILM | NILM | ASCUS | NILM | NILM | LSIL | NILM | ASCUS | NILM | LSIL | NILM | NILM | 21.4% | 42.9% | 42.9% | NILM |
| 31 | ASCUS | HSIL | ASCH | ASCUS | NILM | ASCUS | NILM | NILM | ASCH | ASCH | HSIL | HSIL | HSIL | HSIL | ASCUS | 21.4% | 21.4% | 78.6% | HSIL |
| 32 | ASCUS | ASCUS | ASCUS | NILM | NILM | USFE | NILM | NILM | NILM | NILM | ASCUS | NILM | LSIL | NILM | NILM | 21.4% | 28.6% | 35.7% | NILM |
| 33 | ASCH | ASCH | LSIL | ASCUS | NILM | NILM | NILM | NILM | ASCH | ASCUS | ASCH | ASCH | ASCH | ASCUS | ASCUS | 35.7% | 35.7% | 71.4% | ASCH |
| 34 | ASCH | ASCH | ASCH | NILM | NILM | NILM | NILM | NILM | ASCH | ASCUS | ASCH | ASCH | ASCH | LSIL | ASCUS | 42.9% | 42.9% | 64.3% | ASCH |
| 35 | ASCH | NILM | NILM | NILM | ASCUS | NILM | NILM | NILM | HSIL | ASCUS | ASCH | NILM | ASCUS | NILM | ASCUS | 7.1% | 7.1% | 42.9% | NILM |
| 36 | LSIL | LSIL | ASCUS | ASCUS | LSIL | LSIL | ASCUS | LSIL | ASCUS | LSIL | ASCH | LSIL | LSIL | LSIL | ASCUS | 57.1% | 100.0% | 100.0% | LSIL |
| 37 | LSIL | LSIL | LSIL | ASCUS | LSIL | LSIL | NILM | ASCUS | ASCUS | LSIL | ASCH | LSIL | LSIL | LSIL | ASCUS | 57.1% | 92.9% | 92.9% | LSIL |
| 38 | LSIL | LSIL | LSIL | LSIL | LSIL | LSIL | NILM | LSIL | ASCH | LSIL | ASCH | LSIL | LSIL | ASCUS | NILM | 64.3% | 71.4% | 85.7% | LSIL |
| 39 | LSIL | HSIL | ASCH | LSIL | LSIL | ASCUS | ASCH | LSIL | ASCH | ASCH | HSIL | LSIL | HSIL | LSIL | ASCUS | 35.7% | 50.0% | 100.0% | LSIL |
| 40 | LSIL | LSIL | LSIL | LSIL | LSIL | LSIL | LSIL | LSIL | LSIL | LSIL | LSIL | LSIL | LSIL | LSIL | LSIL | 100.0% | 100.0% | 100.0% | LSIL |
| 41 | LSIL | ASCUS | ASCUS | NILM | NILM | NILM | NILM | NILM | NILM | ASCUS | NILM | NILM | ASCUS | NILM | NILM | 0.0% | 28.6% | 28.6% | NILM |
| 42 | LSIL | LSIL | LSIL | ASCUS | LSIL | ASCUS | NILM | LSIL | LSIL | LSIL | LSIL | ASCUS | LSIL | LSIL | LSIL | 64.3% | 85.7% | 85.7% | LSIL |
| 43 | HSIL | HSIL | HSIL | ASCH | HSIL | HSIL | ASCH | LSIL | HSIL | HSIL | HSIL | HSIL | HSIL | HSIL | HSIL | 78.6% | 85.7% | 100.0% | HSIL |
| 44 | HSIL | HSIL | HSIL | ASCH | ASCUS | HSIL | HSIL | ASCUS | AGC | HSIL | ASCH | HSIL | HSIL | HSIL | ASCUS | 57.1% | 78.6% | 100.0% | HSIL |
| 45 | HSIL | HSIL | ASCH | AGC | HSIL | HSIL | AGC | HSIL | HSIL | HSIL | HSIL | HSIL | AGC | HSIL | HSIL | 71.4% | 100.0% | 100.0% | HSIL |
| 46 | HSIL | HSIL | HSIL | ASCH | NILM | NILM | HSIL | NILM | HSIL | ASCH | ASCH | ASCH | HSIL | ASCH | ASCUS | 35.7% | 71.4% | 78.6% | HSIL |
| 47 | HSIL | HSIL | HSIL | HSIL | HSIL | HSIL | HSIL | HSIL | HSIL | HSIL | HSIL | HSIL | HSIL | HSIL | HSIL | 100.0% | 100.0% | 100.0% | HSIL |
| 48 | HSIL | LSIL | ASCH | LSIL | LSIL | LSIL | ASCUS | LSIL | LSIL | LSIL | LSIL | LSIL | HSIL | LSIL | LSIL | 7.1% | 7.1% | 100.0% | LSIL |
| 49 | HSIL | HSIL | HSIL | LSIL | HSIL | HSIL | HSIL | LSIL | HSIL | HSIL | HSIL | HSIL | HSIL | HSIL | ASCH | 78.6% | 85.7% | 100.0% | HSIL |
| 50 | SCC | SCC | SCC | ASCH | NILM | ASCH | ASCH | SCC | SCC | SCC | SCC | HSIL | SCC | SCC | NILM | 57.1% | 85.7% | 85.7% | SCC |

**Table II.** Consistency of experts' and laboratories evaluations according to general, aggregated and detailed coding

| Experts' diagnoses | Number of slides prepared for the study | General coding — Cytodiagnosticians' diagnoses, N (%) | | | | | Aggregated coding — Cytodiagnosticians' diagnoses, N (%) | | | | | Detailed coding — Cytodiagnosticians' diagnoses, N (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Coherent with experts' diagnosis | Incoherent with experts' diagnosis — Overall | Under-stated | Over-stated | USFE | Coherent with experts' diagnosis | Incoherent with experts' diagnosis — Overall | Under-stated | Over-stated | USFE | Coherent with experts' diagnosis | Incoherent with experts' diagnosis — Overall | Under-stated | Over-stated | USFE |
| USFE | 4 | 18 (32.1) | 38 (67.9) | . | . | . | 18 (32.1) | 38 (67.9) | . | . | . | 18 (32.1) | 38 (67.9) | . | . | . |
| NILM | 22 | 291 (94.8) | 14 (4.6) | . | 14 (4.6) | 2 (0.7) | 291 (94.8) | 14 (4.6) | . | 14 (4.6) | 2 (0.7) | 291 (94.8) | 14 (4.6) | . | 14 (4.6) | 2 (0.7) |
| ASC-US | 6 | 42 (50.0) | 41 (48.8) | 41 (48.8) | . | 1 (1.2) | 22 (26.2) | 61 (72.6) | 41 (48.8) | 20 (23.8) | 1 (1.2) | 18 (21.4) | 65 (77.4) | 41 (48.8) | 24 (28.6) | 1 (1.2) |
| LSIL | 7 | 83 (84.7) | 15 (15.3) | 15 (15.3) | . | 0 (0.0) | 72 (73.5) | 26 (26.5) | 15 (15.3) | 11 (11.1) | 0 (0.0) | 53 (54.1) | 45 (45.9) | 34 (34.7) | 11 (11.1) | 0 (0.0) |
| ASC-H | 3 | 25 (59.5) | 17 (40.5) | 17 (40.5) | . | 0 (0.0) | 13 (31.0) | 29 (69.0) | 29 (69.0) | . | 0 (0.0) | 12 (28.6) | 30 (71.4) | 29 (69.0) | 1 (2.4) | 0 (0.0) |
| HSIL | 7 | 95 (96.9) | 3 (3.1) | 3 (3.1) | . | 0 (0.0) | 75 (76.5) | 23 (23.5) | 23 (23.5) | . | 0 (0.0) | 60 (61.2) | 38 (38.8) | 38 (38.8) | 0 (0.0) | 0 (0.0) |
| SCC | 1 | 11 (78.6) | 3 (21.4) | 3 (21.4) | . | 0 (0.0) | 11 (78.6) | 3 (21.4) | 3 (21.4) | . | 0 (0.0) | 7 (50.0) | 7 (50.0) | 7 (50.0) | 0 (0.0) | 0 (0.0) |
| Low-grade lesions | 13 | 125 (68.7) | 56 (30.8) | 56 (30.8) | . | 1 (0.5) | 94 (51.6) | 87 (47.8) | 56 (30.8) | 31 (17.0) | 1 (0.5) | 71 (39.0) | 110 (60.4) | 75 (41.2) | 35 (19.2) | 1 (0.5) |
| High-grade lesions | 11 | 131 (85.1) | 23 (14.9) | 23 (14.9) | . | 0 (0.0) | 99 (64.3) | 55 (35.7) | 55 (35.7) | . | 0 (0.0) | 79 (51.3) | 75 (48.7) | 74 (48.1) | 1 (0.6) | 0 (0.0) |
| Abnormal | 24 | 256 (76.2) | 79 (23.5) | 79 (23.5) | . | 1 (0.3) | 193 (57.4) | 142 (42.3) | 111 (33) | 31 (9.2) | 1 (0.3) | 150 (44.6) | 185 (55.1) | 149 (44.3) | 36 (10.7) | 1 (0.3) |

USFE – unsatisfactory for evaluation; NILM – no intraepithelial lesions or malignancy; ASC-US – atypical squamous cells of undetermined significance; LSIL – low grade squamous intraepithelial lesions; ASC-H – atypical squamous cells – cannot exclude HSIL; HSIL – high grade intraepithelial lesions; SCC – squamous cell carcinoma; low-grade lesions include ASC-US and LSIL; high-grade lesions include ASC-H, HSIL and SCC; abnormal results include diagnoses ASC-US and more severe
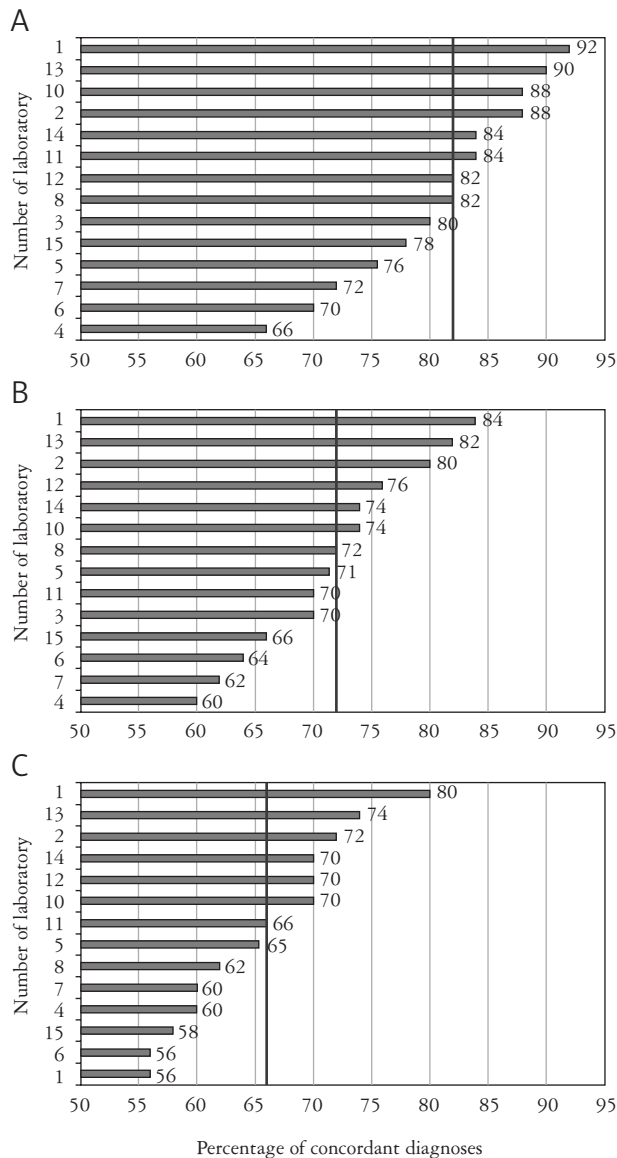
A



B

C

Percentage of concordant diagnoses

**Fig. 1.** Percentage of concordant diagnoses in (A) general, (B) aggregated, (C) detailed coding by laboratory. Vertical line indicates median percentage of correct diagnoses by lab in specific coding

- *Coding 2 (aggregated):* unsatisfactory for evaluation vs. normal (NILM) vs. low-grade lesions (LSIL or ASC-US) vs. high-grade lesions (ASC-H or HSIL or AGC or SCC);
- *Coding 3 (detailed):* unsatisfactory for evaluation vs. normal (NILM) vs. ASC-US vs. LSIL vs. ASC-H vs. AGC vs. HSIL vs. SCC.

The study followed the Declaration of Helsinki and the protocol was approved by the Ministry of Health.

**Statistical analysis**

Percentages of laboratories' diagnoses coherent with expert's evaluations were calculated in each type of coding. Unweighted Cohen's κ statistics were es-

timated for laboratories participating in pilot study. Weighted ($κ_w$) κ were also computed since unweighted κ does not account for severity of the diagnoses. Weights were adjusted so that unsatisfactory for evaluation diagnosis was considered as in total disagreement with any distinct diagnosis and other diagnoses were squarely weighted. Level of <0.05 was established as significant. Landis and Koch's approach was applied to interpret agreement expressed by κ coefficients as: < 0 – poor; 0.0-0.2 – slight; 0.2-0.4 – fair; 0.4-0.6 – moderate; 0.6-0.8 – substantial and 0.8-1.0 – almost perfect. Stata 15 software [15] was used for statistical analyses.

**Results**

Of 15 diagnostic centres involved in study, 13 labs evaluated all of the 50 prepared slides, 1 lab evaluated 49 smears and 1 lab – only 29 slides. Since the distribution of normal and abnormal slides in the set evaluated by last laboratory was not coherent with the distribution of set of all 50 smears, this laboratory was excluded from further analysis (Table I: full source data with expert's and laboratories' diagnoses). Total number of 728 readings was collected and 699 of them were analysed: 21 considered as unsatisfactory for evaluation and 678 as adequate. However, only 32.1% of truly unsatisfactory for evaluation assessments were recognized correctly (Table II). Slides originally assessed as adequate for evaluation were recognized properly in 99.5%.

Cytodiagnosticians managed to evaluate correctly 94.8% of normal smears overall (range from 86.4% to 100% by lab) and recognized correctly 76.2% of abnormal slides (range from 50% to 95.8% by lab). Of 336 slides with expert's diagnosis of "presence of abnormal cells", 79 were apprised as normal (23.5%). On the other hand, 14 of 307 expert's "normal (no intraepithelial lesion or malignancy) readings were identified as abnormal (4.6%). The vast majority of slides assessed originally as "unsatisfactory for evaluation" by the expert (38 of 56) were diagnosed as normal (67.9%).

High-grade lesions were correctly recognized more often than low-grade ones (64.3% vs. 51.6%). Slides with ASC-US or LSIL were considered as normal rather than high-grade (56 readings vs. 31 readings, 30.8% vs. 17.0%).

Of 8 categories in detailed coding, smears with ASC-US and ASC-H expert's diagnoses were evaluated discordant by cytodiagnosticians most frequently (21.4% and 28.6% of coherent diagnoses, respectively). Of 14 squamous cell carcinoma readings in laboratories undergoing assessment, in 3 the cases were apprised as normal (21.4%). Among 408 slides considered as normal by laboratories, 11 were misdi-

agnosed (79 of them (19.4%) was abnormal according to expert's opinion).

The proportion of correct diagnoses, weighted and unweighted κ coefficients were estimated for each laboratory. The median percentage was calculated as 82% (range 66-92%) in general coding, 72% (range 60-84%) in aggregated coding and 66% (range 56-80%) in detailed coding (Fig. 1). Unweighted κ coefficients among labs ranged from 0.40 to 0.86 for general coding, from 0.37 to 0.76 for aggregated coding and from 0.34 to 0.73 for detailed coding and median unweighted κ coefficients correspond with substantial (0.67), moderate (0.59) and moderate agreement (0.53), respectively (Table III).

Weighted κ coefficients were calculated to account for differences in severity of diagnoses. In the detailed coding $\kappa_w$ ranged from 0.40 to 0.76 with median of 0.65, identifying substantial agreement.

Each slide was assigned to percentage of diagnostic centres which diagnosed it correctly by each type of coding (Table II). Of 50 smears, 35 (70%) were recognized correctly by more than 80% of laboratories in general coding, 30 (60%) in aggregated coding and 23 (46%) in detailed coding. However, 4 slides (8%) were properly evaluated by at most 20% of diagnostic centres in general, 7 (14%) in aggregated and 8 (16%) in detailed coding. Two slides were identified incorrectly by all of the laboratories in detailed and one on them also in aggregated coding. Expert diagnosed one of these slides as LSIL and one as ASC-US. Most of smears with more than 80% proper identifications were normal ones [22]. Overall, each slide was recognized properly by average of 81.0% laboratories in general, 71.9% in aggregated and 65.7% in detailed coding (median 92.9%, 85.7% and 78.6%, respectively).

## Discussion

Cytological laboratories participating in the national population-based cervical cancer screening programmes should be certified and subject to both internal and external quality assurance [7]. Internal quality assurance include re-screening of slides, monitoring primary screening detection rates, assessment of cyto-histologic, cyto-clinical, cyto-virological correlations and audit of interval cancers according to EU Guidelines [7]. External quality assurance is less defined, varies between countries and may include proficiency testing and comparison of laboratory and personal reporting rates with set national standards (p. 167 in [7]).

Although cervical cytology is commonly used world-wide in developed countries, its accuracy in detection of cervical neoplasia may vary greatly since it is a very subjective method [14, 16] and diagnoses may differ from lab to lab [17]. Some studies show

**Table III.** Agreement among laboratories. Median and range was calculated for percentage of agreement, weighted and unweighted κ coefficients

| Type of coding | Coefficient of agreement, median (range) | | |
|---|---|---|---|
| | Percentage [%] | κ coefficient | |
| | | Unweighted | Weighted |
| General | 82 (66-92) | 0.67 (0.40-0.86) | 0.67 (0.40-0.86) |
| Aggregated | 72 (60-84) | 0.59 (0.37-0.76) | 0.69 (0.36-0.79) |
| Detailed | 66 (56-80) | 0.53 (0.34-0.73) | 0.65 (0.40-0.76) |

only fair agreement (with κ < 0.3) between cytodiagnosticians evaluating cytology samples [18]. Low grade abnormalities causes most troubles in establishing correct diagnosis [19].

In this study we have aimed to develop grounds for future periodic proficiency testing in assessment of cytological slides and run a pilot evaluation of selected cytodiagnosticians operating in cervical cancer screening programme in Poland. These activities are a part of effort to form a certification/periodic recertification process. We have also aimed to assess interlaboratory agreement in assessment of a set of highly selected cytological slides representing a wide spectrum of normal and abnormal conditions of the uterine cervix. Although the CCSP in Poland has been in place since 2006/2007 such activities have been attempted for the first time.

In general, we have found average to very high agreement between the original expert cytological diagnosis and diagnosis reached by laboratories participating in our pilot project. Interpretation of Pap smears is subjective so achieved concordance with expert diagnoses of each slide by average of 81.0% laboratories in general, 71.9% in aggregated and 65.7% in detailed coding is a fairly good result.

One laboratory (number 1) had repeatedly highest agreement rate in all three types of coding. On the other hand, another laboratory (number 4) had the lowest agreement rate in two types (general, aggregated) coding of diagnoses according the Bethesda system. This may indicate that quality of assessment may differ between cytological laboratories participating in national screening and reasons for high disagreement in outlying laboratories should be sought.

Our results show that cytodiagnosticians found proper classification of unsatisfactory for evaluation slides from normal slides troubling. Only 32.1% of slides with original expert diagnosis of "unsatisfactory for evaluation" were recognized correctly. This should be taken under consideration in future verification of cytological diagnoses and more impact

should be place on the smear quality during the process of training of cytodiagnosticians in Poland.

Percentage of slides correctly diagnosed as abnormal was 76.2%, however, results strongly vary between specific diagnoses from 21.4% slides correctly diagnosed as ASC-US to 61.2% slides correctly diagnosed as HSIL. Some of previous studies show that the greatest source of disagreement in cytology results involved interpretation of low-grade lesions [20, 21, 22] and our study seems to be in agreement with these findings. The number of recognized abnormal slides varied from 50% to 95.8% by laboratory which suggests great differences in reporting between some of them and should trigger corrective measures in labs with low detection rates which directly influence the quality of the whole screening process.

Slides were evaluated additionally to normal workload in laboratories, so it may have influenced the results. Other limitation of our pilot study is the limited number of laboratories which took part in our project. However we have initiated actions to run similar analyses for each cytodiagnostician participating in the CCSP in Poland and make it a future requirement of certification and periodic recertification.

COK as part of Department of Cancer Prevention, The Maria Sklodowska-Curie National Research Institute of Oncology, Warsaw, Poland, on the basis of the agreement made with the Ministry of Health is obliged, among other responsibilities, to monitor the quality of services provided by laboratories participating in the CCSP. We have carried out this study to evaluate consistency of cytological diagnoses in selected laboratories in Poland. The results of the pilot study will lead to the development of methods for evaluation laboratories in the CCSP and therefore to ensure the high quality cytological diagnoses. Periodic skills evaluation system is planned to be introduced to certify cytologists working in the CCSP.

The analysis carried out on the basis of the data collected during the pilot study clearly indicates the necessity of conducting further verification of the consistency in cytological diagnoses. External audits should be run systematically and also include monitoring of colposcopy examinations, histopathological results and effectiveness of treatment and COK initiated actions and is working to implemented these monitoring activities into the CCSP in Poland.

---

## References

1. Bernard W. Stewart and Christopher P. Wild. IARC: World cancer report 2014, Lyon 2014.
2. World Health Organization: Cervical cancer, Geneva 2018. Available at: https://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/ (access 16.07.2019).
3. Wojciechowska U, Czaderny K, Ciuba A, et al. Cancer in Poland in 2016. Polish National Cancer Registry, Department of Epidemiology and Cancer Prevention 2018.
4. Nowakowski A, Wojciechowska U, Wieszczy P, et al. Trends in cervical cancer incidence and mortality in Poland: is there an impact of the introduction of the organised screening? Eur J Epidemiol 2017; 32: 529-532.
5. The Maria Sklodowska-Curie National Research Institute of Oncology, Warsaw: National Cancer Registry. Available at: http://onkologia.org.pl/nowotwory-szyjki-macicy-kobiet/ (access 16.07.2019).
6. Council of the European Union (2003). Council recommendation of 2 December 2003 on cancer screening (2003/878/EC). Off J Eur Union 2003; L 327: 34-38.
7. Arbyn M, Anttila A, Jordan J, et al. (eds.). European guidelines for quality assurance in cervical cancer screening. 2nd edition. International Agency for Research on Cancer 2008.
8. Gultekin M, Karaca MZ, Kucukyildiz I, et al. Initial results of population based cervical cancer screening program using HPV test in one million Turkish women. Int J Cancer 2018; 142: 1952-1958.
9. Rebolj M, Rimmer J, Denton K, et al. Primary cervical screening with high risk human papillomavirus testing: observational study. BMJ 2019; 364: 1240.
10. Polman NJ, Snijders PJF, Kenter GG, et al. HPV-based cervical screening: rationale, expectations and future perspectives of the new Dutch screening programme. Prev Med 2019; 119: 108-117.
11. Bergengren L, Lillsunde-Larsson G, Helenius G, et al. HPV-based screening for cervical cancer among women 55-59 years of age. PLoS One 2019; 14: e0217108.
12. Żelazik M, Michalska A, Radowicz-Chil A, et al. Unified cytological reports determine the clinical management. Medical Studies 2019; 35: 147-152.
13. Nayar R, Wilbur DC. The Bethesda System for Reporting Cervical Cytology: A Historical Perspective. Acta Cytol 2017; 61: 359-372.
14. Polskie Towarzystwo Ginekologiczne: Rekomendacje PTG dotyczące diagnostyki, profilaktyki i wczesnego wykrywania raka szyjki macicy. Ginekol Pol 2006; 77: 655-659.
15. StataCorp. 2017. Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC.
16. Baena A, Guevara E, Almonte M, et al. Factors related to inter-observer reproducibility of conventional Pap smear cytology: a multilevel analysis of smear and laboratory characteristics. Cytopathology 2017; 28: 192-202.
17. Klinkhamer PJ, Vooijs GP, de Haan AF. Intraobserver and interobserver variability in the quality assessment of cervical smears. Acta Cytol 1989; 33: 215-218.
18. da Fonseca AJ, Murari RS, Moraes IS, et al. Validity of cervicovaginal cytology in a Brazilian State with high incidence rate of cervical cancer. Rev Bras Ginecol Obstet 2014; 36: 347-352.
19. Sørbye SW, Suhrke P, Revå BW, et al. Accuracy of cervical cytology: comparison of diagnoses of 100 Pap smears read by four pathologists at three hospitals in Norway. BMC Clin Pathol 2017; 17: 18.
20. Stoler MH, Schiffman M. Interobserver reproducibility of cervical cytologic and histologic interpretations: realistic estimates from the ASCUS-LSIL Triage Study. JAMA 2001; 285: 1500-1505.

21. Türkmen IC, Başsüllü N, Bingöl B, et al. Interobserver variability in cervical smears from patients with a history of abnormal cytology: comparison of conventional Pap smears and liquid-based cytology. Erciyes Med J 2013; 1: 13-17.
22. Wright TC Jr, Stoler MH, Behrens CM, et al. Interlaboratory variation in the performance of liquid-based cytology: insights from the ATHENA trial. Int J Cancer 2014; 134: 1835-1843.

## Address for correspondence

**Kinga Zalewska-Otwinowska**
Maria Sklodowska-Curie National Research Institute of Oncology
Department of Cancer Prevention
Roentgen 5
02-781 Warsaw, Poland
tel./fax: +48 22 546 30 21
e-mail: kinga.zalewska-otwinowska@pib-nio.pl