Retrospective Diagnostic Study of Ummon AutoReader Performance: A Tool for Prescreening Cytology of Cervical Lesions

Nathan Vinçon, Georges Tarris, Joëlle Depardon, Henri-Philippe Morel, Louis-Oscar Morel

Synopsis of your article in 2-3 sentences for both printing in 'Inside this month's Cytopathology' and tweeting (~280 characters) through Cytopathology twitter handle.

Synopsis : A diagnostic validation of the Ummon AutoReader software as a prescreening tool for p16/Ki-67 dual staining in cervical cytology. Ummon AutoReader demonstrates a higher diagnostic sensitivity for a similar specificity and removes the hurdle of manual prescreening.

Abstract

Background : Cervical cancer is a preventable disease, yet its persistent incidence highlights the need for more effective screening strategies. Among existing methods, p16/Ki-67 dual stain cytology is noted for its cost-effectiveness but is limited by time-consuming processes and the risk of overlooking precancerous cells. To address these issues, we developed Ummon AutoReader, a software that automates the prescreening of cytology slides using a hybrid artificial intelligence approach to minimize the omission of precancerous cells and ease cytology reading.

Objectives : This study aims to evaluate the diagnostic accuracy of Ummon AutoReader by comparing the sensitivity and specificity of manual versus software-assisted readings of p16/Ki-67 dual stain cytology. Additionally, we evaluate the analytical performance of the Ummon AutoReader.

Methods : We analyzed a representative population of 110 cases gathered from the routine workflow of a pathology laboratory. Manual diagnosis derived from clinical records was compared with assisted diagnosis conducted by three cytopathologists using Ummon AutoReader. All discordances were resolved by a consensus committee.

Results : The use of Ummon AutoReader for assisted diagnosis resulted in a higher sensitivity compared to manual reading (100% vs. 81.9%), while maintaining equivalent specificity (100% for both methods). The software achieved a cell detection rate of 98% in a sample of 200 cells, with an area under the curve (AUC) of 0.893 for differentiating between dually and non-dually stained cells. Inter-scanner consistency was confirmed in all 10 cases examined, and inter-scanner correlation of cell scores improved from 0.749 (n=37) without calibration to 0.813 (n=39) with calibration.

Conclusion : Ummon AutoReader enhances the sensitivity of p16/Ki-67 cytology readings without affecting specificity. It also demonstrates robust inter-scanner reliability and improves the ease of cytological evaluations, making it a valuable tool for the early detection of cervical cancer.

Introduction

Cervical cancer represents a significant global health concern, ranking among the top three cancers affecting women younger than 45 years worldwide (Arbyn et al., 2020). In 2022, an estimated 565,541 new cases of cervical cancer were reported and 280,479 new deaths occurred due to cervical cancer worldwide (Momenimovahed et al., 2023). The incidence rate of this disease continues to underscore the importance of effective screening strategies for its prevention and early detection (Bhatla & Singhal, 2020; Jansen et al., 2020; Perkins et al., 2023; Sawaya et al., 2019). These strategies often involve primary HPV testing, because of its high sensitivity to detect precancerous lesions. A negative HPV test indicates a very low cervical cancer risk over the next decade (Dillner et al., 2008; Gage et al., 2014; Katki et al., 2011). However, the moderate specificity of HPV testing, due to its inability to discriminate between transient and persistent infections (Catarino et al., 2015), necessitates additional triage for colposcopy referral (Cuschieri et al., 2018; Sawaya et al., 2019; Wentzensen et al., 2016). This often includes cytology (Papanicolaou tests), but the limited reproducibility of cytology requires frequent retesting (Stoler et al., 2001; Wright Jr et al., 2014).

Another promising triage strategy is concomitant detection of p16 and Ki-67, respectively a HPV-activated protein and a cell proliferation marker, in the same cell. The p16/Ki-67 dual staining has demonstrated higher accuracy in detecting cervical precancerous lesions compared to cytology (Carozzi et al., 2013; Clarke et al., 2019; Ouh et al., 2024; Schmidt et al., 2011; Wentzensen et al., 2012, 2015, 2019; Wright et al., 2017). The p16/Ki-67 dual staining is commercialized by Roche as the CINTec Plus™ technology (Bergeron et al., 2015; Schmidt et al., 2011), for which medico-economic studies showed improved cost-effectiveness (Barré et al., 2017; Petry et al., 2017). Recently, new ASCCP cervical cancer management guidelines included dual-stain triage testing to manage early diagnosis of HPV- positive cervical precancer and cancer (Clarke et al., 2024). These guidelines highlight that, compared to cytology, dual stain requires fewer colposcopies and detects cervical intraepithelial neoplasia grade 3 (CIN3) or worse earlier.

Nevertheless, the manual interpretation of CINTec Plus™ smear tests presents several challenges. Each smear requires meticulous double examination by a cytotechnologist and a cytopathologist, often involving the scrutiny of tens to hundreds of thousands of cells. Despite this thoroughness, the possibility of missing positive cells, even when present among the vast cell population, remains a concern. To mitigate these challenges, automation through algorithmic software has emerged as a promising solution. In 2021, Wentzensen *et al.* (2021) introduced a deep learning algorithm to automatize the p16/Ki-67 dual stain reading and demonstrated the clinical relevance of such tools by showing that automated reading of dual stain provided better risk stratification compared with Pap cytology and manual reading of dual stain. In the conducted study, manual reading was undertaken within a study environment, thereby potentially leading to the Hawthorne effect (McCambridge et al., 2014), a phenomenon that could inflate test sensitivity as compared to routine conditions. Furthermore, external validation using alternative scanners for whole slide image acquisition was not conducted.

To address these limitations, we present Ummon AutoReader (Ummon HealthTech, Dijon, France), a CE-marked software solution designed to streamline the analysis of dual stain p16/Ki-67 cervical smears. Notably, Ummon AutoReader automates only slides pre-screening, identifying diseased cells and facilitating rapid diagnosis by cytologists. This removes the need for exhaustive manual slide exploration while leaving the diagnostic decision to the medical expert. Additionally, Ummon AutoReader integrates a calibration algorithm designed to ensure compatibility across diverse scanner platforms.

In this study, we conduct a comprehensive evaluation of Ummon AutoReader. Firstly, we assess its diagnostic accuracy by comparing the sensitivity and specificity of the software-assisted approach to a fully manual diagnosis across a dataset of 110 slides, focusing specifically on the detection of slides containing at least one dual-stained cell. We then scrutinize individual components of the automated process, including cell detection, scoring, and ranking, to detail the performance at each stage. Lastly, we validate the robustness of the algorithm and examine the impact of the calibration process through an inter-scanner assessment involving 10 different slides. This multifaceted analysis aims to establish a thorough understanding of Ummon AutoReader's capabilities and its potential utility in clinical settings.

Methods

Evaluation Protocol

Study population and data collection

Cervical smears were collected from 110 patients with an age between 25 and 65 years old for which a Pap smear was previously performed and revealed a low-grade squamous intraepithelial lesion (LSIL) or atypical squamous cells of undetermined significance (ASC-US). The CINtec PLUS™ Cytology Kit (Roche mtm Laboratories AG, Mannheim, Germany) was used for the dual immunostaining of cervical smears according to the manufacturer instructions (https://diagnostics.roche.com/content/dam/diagnostics/us/en/products/c/cintec-plus/CINtec-PLUS-Cytology-Package-Insert-US.pdf). Briefly, slides were incubated in ≥95% reagent grade ethanol, then air-dried flat before loading onto the VENTANA BenchMark XT automated slide stainer. After preparation, slides were digitized at ×40 magnification (resolution 0.28 μm/pixel) using a Pannoramic 250 slide scanner

(3DHISTECH, Budapest, Hungary), generating a slide file in the MRXS format. Data was pseudonymised before subsequent use. Among the 110 slides, 2 slides were excluded due to poor scanning quality, and 10 were also scanned with a Hologic Genius™ Digital Imager to perform an inter-scanner evaluation.

Study objectives

In this study, we compare two methods for reading p16/Ki-67 dual stain smears:

1. **Manual diagnosis**: Manual reading of p16/Ki-67 dual stain smears, first by a cytotechnologist and then by a pathologist.
2. **Assisted diagnosis**: Assisted reading of p16/Ki-67 dual stain smears using the Ummon AutoReader software as a diagnostic aid for the pathologist (without a cytotechnologist).

The primary comparison criterion is the sensitivity of each method in detecting p16/Ki-67 dually stained cells.

Study design

Determining the required sample size for clinical analysis involved considering previous estimates. From our experience, manual diagnosis was estimated to have a sensitivity of 0.75 (*p1*), while assisted diagnosis' sensitivity was estimated to be greater than 0.95 (*p2*). These values guided the calculation with a 95% confidence level (α risk of 5%, two-sided) and 80% statistical power (β risk of 20%). Using the sample size estimation formula for proportions (Wang & Chow, 2007) :

$$n = \frac{(Z_{\alpha/2}+Z_{\beta})^2 \times (p_1(1-p_1)+p_2(1-p_2))}{(p_1-p_2)^2}$$

The initial estimate yielded n=47. However, since sensitivity is the proportion of detected samples (true positives) among samples carrying the disease, not all samples will be positive. Based on typical distributions, it is expected that approximately half of the CINTec PLUS samples will carry the disease, thus requiring a doubling of the study sample size, leading to n=94. To ensure robustness, we aimed to have at least 47 cases carrying the disease and added a small margin, resulting in the selection of 110 cases (see Data Acquisition). Note that the cases were not enriched for positive or negative outcomes, ensuring that the population is representative of routine distribution.

Manual diagnostics were obtained from laboratory records, while for assisted diagnosis, three investigators underwent a 30-minutes training session with the Ummon AutoReader software before participating in the study. Investigators made diagnoses within the software, and all diagnoses were exported for subsequent statistical analyses. In cases of discordance between manual and assisted diagnoses, a consensus diagnosis was performed by at least two pathologists using all available information (i.e. the raw slide and the Ummon AutoReader analysis). The diagnoses were established as positive when at least one dual stained cell was found, or negative otherwise (Wentzensen et al., 2015).

Analytical metrics

Three metrics were used to characterize the performance of the algorithm itself.

- Firstly, we used the cell detection sensitivity, corresponding to the number of positive cells that have been detected as a region of interest (ROI) by the software, regardless of their subsequent scoring. A total of 200 ROI (600 x 600 pixels) containing a positive cell were manually extracted from 25 randomly selected slides (18 positive and 7 negative) from the previous set of 110 slides. A positive cell was considered to be found if the distance between the centroid of the cell and the centroid of to the closer ROI found by the Ummon AutoReader software was lower than 300 pixels.

- Secondly, we used the area under the curve (AUC) of the detected cell ranking, which corresponds to the ability of the scoring algorithm to separate positive cells (i.e. dual stained cells) from negative cells. We used 287 positive cells and 5057 negative cells (being 20% of the data described in *Training and validation dataset for cell ranking* in the next section)

- Thirdly, to more precisely account for the user interface design that shows ROI page by page with a default value of 20 ROI per page, we used the proportion of positive slides that have at least one positive cell located inside the first page, and that we refer to as first page sensitivity. This metric evaluates the cell scoring by imitating the user’s behavior. It was performed on the full set of 110 slides.

Statistical Analyses

The sensitivity and negative predictive values were compared using a Chi-squared Proportion Test. Confidence intervals for the cell detection rate were calculated using the Clopper-Pearson exact method. All statistical analyses were conducted in Python 3.10, utilizing the scipy.stats library.

Overview of the Ummon AutoReader software

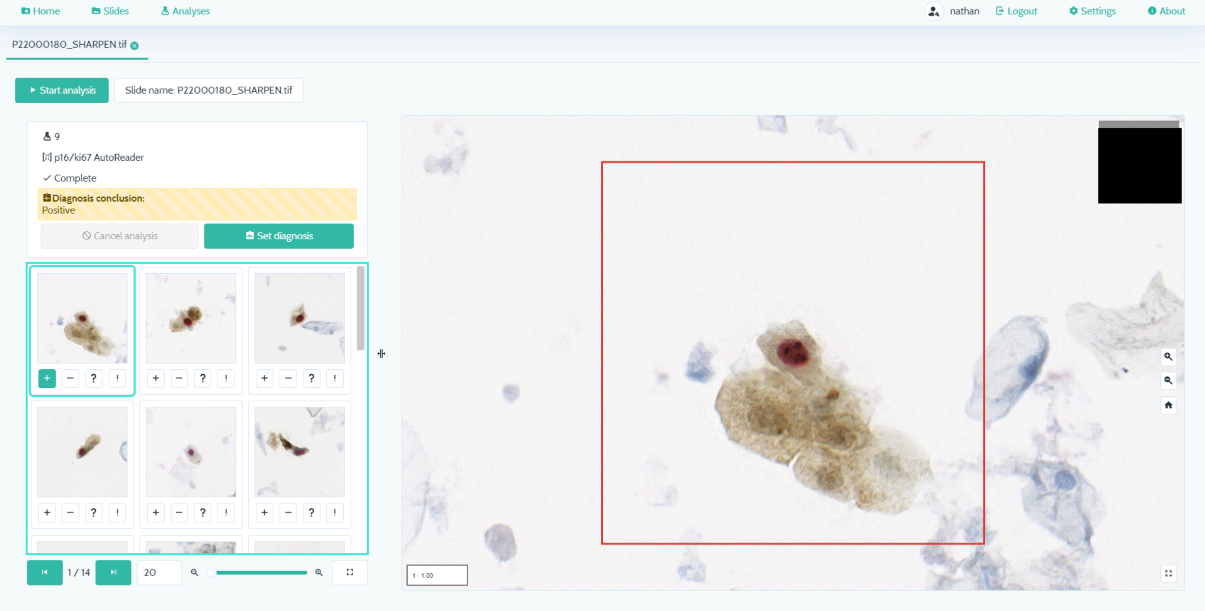
Ummon AutoReader has been certified with a CE marking under Medical device Directive 93/42/CEE since March 2022. Its p16/Ki-67 automated reading algorithm comprises a slide exploration component and a cell scoring component described hereafter.

Slide exploration and cell detection

The slide undergoes processing by an exploration algorithm that traverses the entire slide patch by patch at a ×40 magnification. Each patch is sized at 600 pixels, with overlapping patches by 100 pixels to prevent cells from being missed by being cut between two patches. A pixel classifier, previously trained on pixels extracted from both positive and negative cells, identifies pixels that are Ki-67-positive based on image color. Positive pixels are grouped into connected components, and if the size of a connected component exceeds the threshold size, it is considered as cells. In case a connected component overflows the patch (i.e. touches the border), a new patch is created to capture the full connected component. When a cell is detected, it is centered, and the scoring model is applied (see next subsection).

Model for cell scoring and ranking

Cell scoring utilizes a hybrid model that combines traditional algorithms and deep learning. This model includes a staining score based on the surface area and the intensity of positive staining for both p16 staining and Ki-67 staining, as well as a deep learning score based on classification into classes as defined in the *Training and validation datasets for cell ranking* section. The deep learning model used is a NASNetMobile neural network trained for 20 epochs (Zoph et al., 2018). Subsequently, scores generated by the color-based methods and the deep learning algorithms are combined to achieve optimal scoring, maximizing the discrimination between positive and negative cells represented by the AUC. Detected cells are then added to the comprehensive list and ranked accordingly, with cells showing the highest probability of being positive positioned at the top of the list. The ranked list of detected cells is displayed in pages of 20 cells, as illustrated in Figure 1.



Slide name: P22000180_SHARPEN.tif

9

p16/ki67 AutoReader

✓ Complete

Diagnosis conclusion:
Positive

Figure 1 : *User interface of the Ummon AutoReader software showing a case after automatic analysis and pathologist diagnosis.*

Calibration module

To address inter-scanner variability, the Ummon AutoReader software incorporates a calibration module, which is individually calibrated for each laboratory. This calibration process relies on a stain inference method derived from the Vahadane approach (Vahadane et al., 2015). It adjusts both the cell detection and cell scoring algorithms to accommodate the specific color vector characteristic of each laboratory protocol (staining and scanning). The calibration directly impacts the algorithm in the analysis and is not an independent preprocessing step. In this way it does not increase the total time of the analysis.

Results

Dataset description

The dataset employed in this study initially comprised 110 CINTec Plus™ smear slides obtained from 110 patients between 2018 and 2020. These slides were collected from routine laboratory submissions over a three-month period, ensuring that the data reflects real-life conditions. However, two slides were excluded due to poor scanning quality, leaving 108 slides for analysis. Among these, 68 slides were linked to a positive diagnosis, while 40 slides were associated with a negative diagnosis based on routine diagnostic procedures. The age distribution of the patients spanned from 22 to 65 years and peaked between 25 and 35 years (Table 1), aligning with the french guidelines for utilizing the p16/Ki-67 dual staining method (https://www.e-cancer.fr/content/download/307096/4383798/file/Outil-Pratique-Uterus-2021-@%20DEF%2012032021.pdf).

Age intervals	Number of patients
(20, 25)	9
(25, 35)	68
(35, 45)	19
45+	14

Table 1 : Distribution of ages in the patient population

Diagnostic Performance Comparison of Manual vs. Automated (Ummon AutoReader) p16/Ki-67 Interpretation

The manual diagnosis (i.e. without the Ummon AutoReader) and the assisted diagnosis using Ummon AutoReader showed concordance in 93 out of 108 cases (86%). Following resolution of discrepancies in diagnosis through consensus decision-making by at least two pathologists (see Methods), 83 cases were diagnosed as positive and 25 as negative (Table 2).

The primary comparison criterion was sensitivity. The Ummon AutoReader assisted diagnosis demonstrated superior sensitivity compared to manual diagnosis (100% vs 81.9%, p-value < 0.0001), while maintaining a similar specificity (100% vs 100%). The exceptionally high specificity observed aligns with the robust objectivity of CINTec dual stain positivity, thereby minimizing the likelihood of false positives. Similarly, the positive predictive value (PPV) remained consistent for both methods (100% vs 100%). However, a notable discrepancy was observed in the negative predictive value (NPV), with the Ummon AutoReader assisted diagnosis exhibiting superior performance compared to manual diagnosis (100% vs 62.5%, p-value < 0.0001). Results were confirmed by two expert pathologists.

	Ground Truth		
	Positive	Negative	Total
Manual diagnosis (without Ummon AutoReader)			
Positive	68	0	68
Negative	15	25	40
Total			108
Assisted diagnosis (with Ummon AutoReader)			
Positive	83	0	83
Negative	0	25	25
Total			108

Table 2 : Confusion matrices comparing the manual diagnosis and the assisted diagnosis using the Ummon AutoReader software

Comprehensive Performance Analysis of Ummon AutoReader

Although the diagnostic validation showcased the superiority of Ummon AutoReader assisted diagnosis compared to manual diagnosis, it does not assess the software's reliability comprehensively. Therefore, we aim to expand upon this assessment by leveraging transparent analytical metrics. This approach enhances our understanding of Ummon AutoReader's performance, providing confidence in its output and supporting clinical decision-making.

We define metrics that represent key steps of the complete diagnosis process when using the software, focusing specifically on cell detection and ranking efficacy. To evaluate these steps, we measured the sensitivity of the detected cells as regions of interest (ROIs) and the area under the curve (AUC) of positive cell binary classification.

Cell detection rate

To construct a representative dataset, we manually identified 200 cells as positive by examining 25 slides randomly selected from the 108 slides used in the diagnostic validation. Among these 25 slides, 18 slides actually provided cells for the dataset. This selection process was conducted independently of the established diagnoses to prevent any bias in positive cell sampling. We extracted cell coordinates manually and compared them to the ROIs identified by the Ummon AutoReader analysis conducted during the previous diagnostic validation. Out of the 200 cells examined, the analysis successfully detected 196 cells, resulting in a cell detection sensitivity of 98% (95% CI = [0.9496, 0.9945], calculated using the Clopper-Pearson exact method).

Cell ranking performance

A total of 26 725 cells from 69 slides (distinct from the slide set of the diagnostic validation) were extracted and annotated by experts into 5 categories : double positive (p16+/Ki-67+), single p16 positive (p16+/Ki-67-), single Ki-67 positive (p16-/Ki-67+), double negative (p16-/Ki-67-) or as artifacts (i.e. not a cell or a heavily damaged cell). The resulting class distribution is presented in Table 3. Annotated cells were divided into 80% training and 20% testing sets with a homogeneous class distribution.

Type	Number
p16+/Ki-67+	1439
p16+/Ki-67-	10311
p16-/Ki-67+	8346
p16-/Ki-67-	96
artifact	6533

Table 3 : Distribution of classes in our cell dataset

Prior to integration into the Ummon AutoReader software, the cell scoring model (refer to Methods) underwent training using the training set, which constituted 80% of the dataset. The cell scoring model was then evaluated for its ability to discriminate between positive and negative cells (i.e., p16+/Ki-67+ versus all other categories), achieving an AUC of 0.893. We chose to focus on the discrimination between positive and negative cells rather than the complete classification into all five categories because only the double positive class is relevant for diagnosing CINTec dual immunostaining.

User Interface Efficiency

We also introduce a third metric that is specific to the Ummon AutoReader workflow through its user interface. Detected cells are organized into pages of 20 ROI, impacting the exploration of detected cells by the pathologist, which is likely to explore exhaustively the first page but less likely to explore the following pages. We calculated the proportion of positive analysis having a positive cell in the first page (top 20 ROI, Figure 2) of 100% (83 / 83). Note that most of the time, the positive cell is found in the top 3 cells (78 / 83 cases), corresponding to the first line of detected cells in the user interface with default parameters.

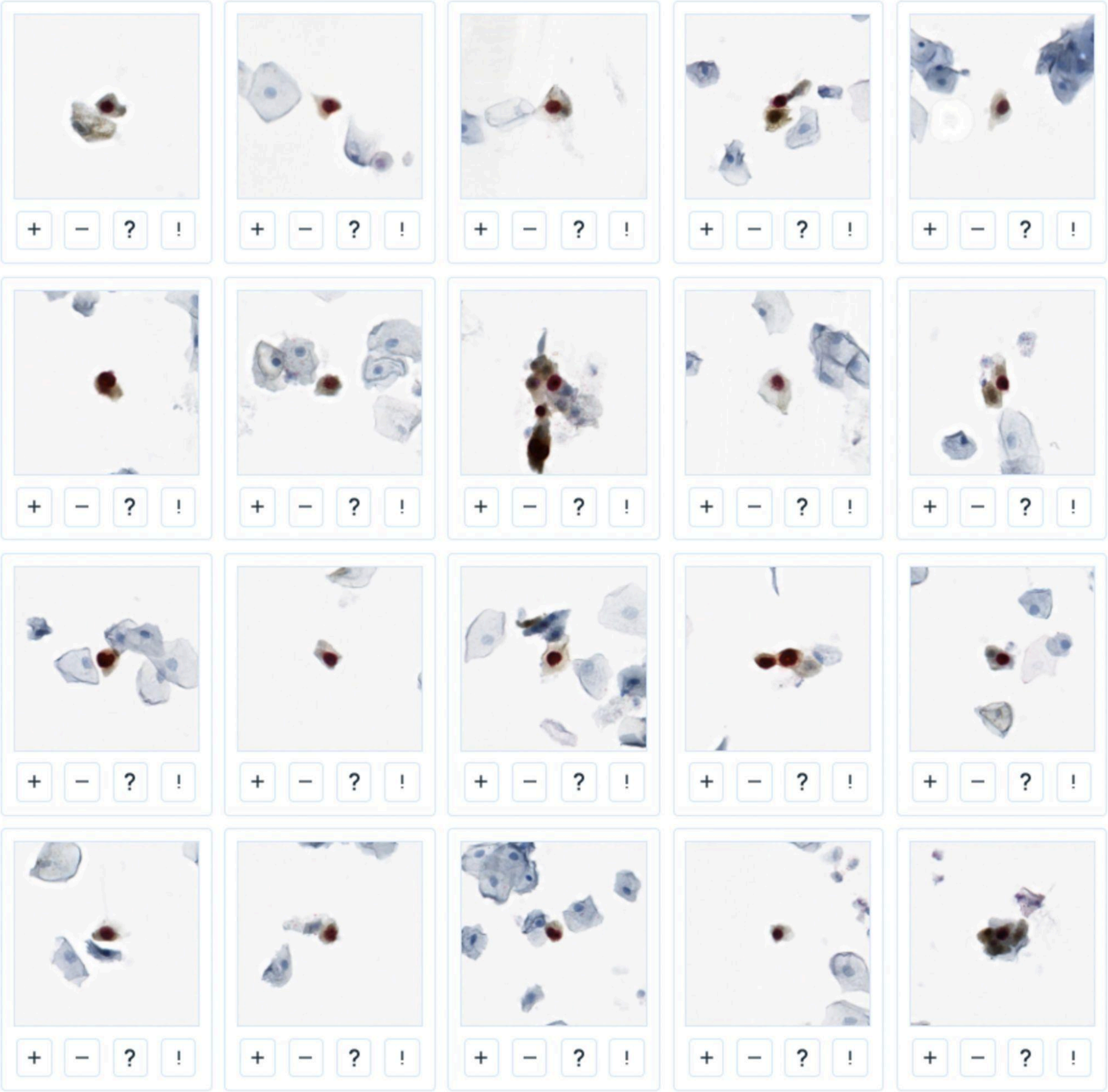


Figure 2 : An example of a page showing 20 detected cells sorted from most likely positive to less likely positive.

Inter-scanner consistency of Ummon AutoReader analyses

A total of 10 slides, including 4 negative slides and 6 positive slides, from the dataset were scanned with a Hologic Genius scanner to evaluate the inter-scanner consistency of the automated analysis. The 10 slides were analyzed by the Ummon AutoReader both with and without a calibration before. A separate diagnosis was then performed, which showed complete consistency (i.e. 10/10 slides).

We then randomly selected cells spanning the full range of cell scores from 3 analyses : the analysis of slides scanned with 3DHistech Panoramic 250 scanner (referred to as the reference), the analysis of slides with Hologic Genius scanner **without** the calibration (referred to as uncalibrated), and the analysis of slides with Hologic Genius scanner **with** the calibration (referred to as calibrated). Correlation between the reference and the uncalibrated slides is 0.749 (n=37), while the correlation between the reference and the calibrated slides reached 0.813 (n=39), showcasing the effectiveness of calibration to ensure consistent results across scanners. Note that there was a little discrepancy in magnification between both scanners, thus on the one hand the correlation is probably underestimated, and on the second hand the scoring algorithm is robust to scaling variations too. Moreover, correlation between calibrated and uncalibrated slides reached 0.887 (n=33).

Discussion

Ummon AutoReader is a software designed to automatically pre-analyze cervical smears dually stained with p16 and Ki-67. It assists diagnosis by generating a cell ranking, prioritizing the most likely positive cells, thus significantly facilitating the diagnostic task for pathologists. The software leverages computer vision and a hybrid algorithm that combines traditional methods with deep learning. Additionally, Ummon AutoReader integrates a calibration algorithm that reduces inter-scanner variability in analyses.

In this study, we have demonstrated that the assisted diagnosis using Ummon AutoReader outperforms the manual diagnosis used routinely in terms of sensitivity of finding a dually stained cell within a smear, while maintaining similar specificity. Using 108 samples, we found that Ummon AutoReader achieved a sensitivity of 100%, compared to 81.9% for manual diagnosis, with both methods showing a specificity of 100%. Achieving 100% sensitivity is not particularly surprising as the primary challenge in p16/Ki-67 dual stain interpretation is identifying potentially positive cells among hundreds of thousands cells, because there is low ambiguity in cell classification. Importantly, the cell detection sensitivity was 98%, indicating a rare possibility of missing a positive cell at the detection stage. Our evaluation of scoring performance showed a ranking performance with an AUC of 0.893 and demonstrated user interface efficiency, with the first positive cell, when present, systematically appearing on the first page of results.

A previous algorithm was proposed by Wentzensen *et al.* (2021) to automatize the p16/Ki-67 dual stain reading, which demonstrated the relevance of such tools by showing that automated reading of dual stain provided better risk stratification compared with Pap cytology and manual reading of dual stain. However, their study involved adapting and retraining the model for each test cohort, thus leading to a center-specific solution. In contrast, we present a ready-to-use software evaluated on new data across different scanners, demonstrating its applicability with various workflows after a simple calibration process.

A major limitation of our study is the absence of clinical outcome data, particularly regarding the occurrence of precancerous lesions for the screened patients, as we lack long-term follow-up and biopsy data for patients with discrepant diagnoses. It is possible that increased sensitivity could lead to a reduced specificity in the detection of these precancerous lesions, and further studies should be made to investigate this question. Nevertheless, our analysis is likely more reliable since the software has a detection probability exceeding 95%. Strengths of our study include the use of real routine data for comparison and multiple scanners to assess robustness. Although in this study we did not precisely measure the time spent by the cytotechnologists and by the cytopathologists, we observed that the cytopathologist typically required less than 2 minutes to make a diagnosis. This gain in time should be considered within the context of slide digitization, which introduces additional human time into the workflow. Further studies should investigate the full workflow to better assess the time-saving benefits of our software. We advocate for the use of Ummon AutoReader as it prevents missing cells and accelerates analysis, potentially reducing costs and increasing efficacy for population screening.

The use of Ummon AutoReader in routine practice could reduce delays and increase throughput, thereby potentially alleviating patient anxiety while awaiting screening results. The high cell detection rate also suggests that diagnoses can be made with greater confidence. Furthermore, as slide exploration is a labor-intensive task, Ummon AutoReader improves the ergonomics of reading p16/Ki-67 smears, enhancing overall work comfort for pathologists.

Reference to patent here

References

Arbyn, M., Weiderpass, E., Bruni, L., Sanjosé, S. de, Saraiya, M., Ferlay, J., & Bray, F. (2020). Estimates of incidence and mortality of cervical cancer in 2018: A worldwide analysis. *The Lancet Global Health*, *8*(2), e191–e203.

https://doi.org/10.1016/S2214-109X(19)30482-6

Barré, S., Massetti, M., Leleu, H., & De Bels, F. (2017). Organised screening for cervical cancer in France: A cost-effectiveness assessment. *BMJ Open*, *7*(10), e014626. https://doi.org/10.1136/bmjopen-2016-014626

Bergeron, C., Ikenberg, H., Sideri, M., Denton, K., Bogers, J., Schmidt, D., Alameda, F., Keller, T., Rehm, S., Ridder, R., & PALMS Study Group. (2015). Prospective evaluation of p16/Ki-67 dual-stained cytology for managing women with abnormal Papanicolaou cytology: PALMS study results. *Cancer Cytopathology*, *123*(6), 373–381. https://doi.org/10.1002/cncy.21542

Bhatla, N., & Singhal, S. (2020). Primary HPV screening for cervical cancer. *Best Practice & Research Clinical Obstetrics & Gynaecology*, *65*, 98–108. https://doi.org/10.1016/j.bpobgyn.2020.02.008

Carozzi, F., Gillio-Tos, A., Confortini, M., Mistro, A. D., Sani, C., Marco, L. D., Girlando, S., Rosso, S., Naldoni, C., Palma, P. D., Zorzi, M., Giorgi-Rossi, P., Segnan, N., Cuzick, J., & Ronco, G. (2013). Risk of high-grade cervical intraepithelial neoplasia during follow-up in HPV-positive women according to baseline p16-INK4A results: A prospective analysis of a nested substudy of the NTCC randomised controlled trial. *The Lancet Oncology*, *14*(2), 168–176. https://doi.org/10.1016/S1470-2045(12)70529-6

Catarino, R., Petignat, P., Dongui, G., & Vassilakos, P. (2015). Cervical cancer screening in developing countries at a crossroad: Emerging technologies and policy choices. *World Journal of Clinical Oncology*, *6*(6), 281–290. https://doi.org/10.5306/wjco.v6.i6.281

Clarke, M. A., Cheung, L. C., Castle, P. E., Schiffman, M., Tokugawa, D., Poitras, N., Lorey, T., Kinney, W., & Wentzensen, N. (2019). Five-Year Risk of Cervical Precancer Following p16/Ki-67 Dual-Stain Triage of HPV-Positive Women. *JAMA Oncology*, *5*(2), 181–186. https://doi.org/10.1001/jamaoncol.2018.4270

Clarke, M. A., Wentzensen, N., Perkins, R. B., Garcia, F., Arrindell, D., Chelmow, D., Cheung, L. C., Darragh, T. M., Egemen, D., Guido, R., Huh, W., Locke, A., Lorey, T. S., Nayar, R., Risley, C., Saslow, D., Smith, R. A., Unger, E. R., Massad, L. S., & Enduring Consensus Cervical Cancer Screening and Management Guidelines Committee. (2024). Recommendations for Use of p16/Ki67 Dual Stain for Management of Individuals Testing Positive for Human Papillomavirus. *Journal of Lower Genital Tract Disease*, *28*(2), 124–130. https://doi.org/10.1097/LGT.0000000000000802

Cuschieri, K., Ronco, G., Lorincz, A., Smith, L., Ogilvie, G., Mirabello, L., Carozzi, F., Cubie, H., Wentzensen, N., Snijders, P., Arbyn, M., Monsonego, J., & Franceschi, S. (2018). Eurogin roadmap 2017: Triage strategies for the management of HPV-positive women in cervical screening programs. *International Journal of Cancer*, *143*(4), 735–745. https://doi.org/10.1002/ijc.31261

Dillner, J., Rebolj, M., Birembaut, P., Petry, K.-U., Szarewski, A., Munk, C., De Sanjose, S., Naucler, P., Lloveras, B., Kjaer, S., Cuzick, J., Van Ballegooijen, M., Clavel, C., & Iftner, T. (2008). Long term predictive values of cytology and human papillomavirus testing in cervical cancer screening: Joint European cohort study. *BMJ*, *337*(oct13 1), a1754–a1754. https://doi.org/10.1136/bmj.a1754

Gage, J. C., Schiffman, M., Katki, H. A., Castle, P. E., Fetterman, B., Wentzensen, N., Poitras, N. E., Lorey, T., Cheung, L. C., & Kinney, W. K. (2014). Reassurance Against Future Risk of Precancer and Cancer Conferred by a Negative Human Papillomavirus Test. *JNCI: Journal of the National Cancer Institute*, *106*(8), dju153. https://doi.org/10.1093/jnci/dju153

Jansen, E. E. L., Zielonke, N., Gini, A., Anttila, A., Segnan, N., Vokó, Z., Ivanuš, U., McKee, M., de Koning, H. J., de Kok, I. M. C. M., Veerus, P., Anttila, A., Heinävaara, S., Sarkeala, T., Csanádi, M., Pitter, J., Széles, G., Vokó, Z., Minozzi, S., … Priaulx, J. (2020). Effect of organised cervical cancer screening on cervical cancer mortality in Europe: A systematic review. *European Journal of Cancer*, *127*, 207–223. https://doi.org/10.1016/j.ejca.2019.12.013

Katki, H. A., Kinney, W. K., Fetterman, B., Lorey, T., Poitras, N. E., Cheung, L., Demuth, F., Schiffman, M., Wacholder, S., & Castle, P. E. (2011). Cervical cancer risk for women undergoing concurrent testing for human papillomavirus and cervical cytology: A population-based study in routine clinical practice. *The Lancet Oncology*, *12*(7), 663–672. https://doi.org/10.1016/S1470-2045(11)70145-0

McCambridge, J., Witton, J., & Elbourne, D. R. (2014). Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology*, *67*(3), 267–277. https://doi.org/10.1016/j.jclinepi.2013.08.015

Momenimovahed, Z., Mazidimoradi, A., Maroofi, P., Allahqoli, L., Salehiniya, H., & Alkatout, I. (2023). Global, regional and national burden, incidence, and mortality of cervical cancer. *Cancer Reports*, *6*(3), e1756. https://doi.org/10.1002/cnr2.1756

Ouh, Y.-T., Kim, H. Y., Yi, K. W., Lee, N.-W., Kim, H.-J., & Min, K.-J. (2024). Enhancing Cervical Cancer Screening: Review of p16/Ki-67 Dual Staining as a Promising Triage Strategy. *Diagnostics*, *14*(4), Article 4. https://doi.org/10.3390/diagnostics14040451

Perkins, R. B., Wentzensen, N., Guido, R. S., & Schiffman, M. (2023). Cervical Cancer Screening: A Review. *JAMA*, *330*(6), 547–558. https://doi.org/10.1001/jama.2023.13174

Petry, K. U., Barth, C., Wasem, J., & Neumann, A. (2017). A model to evaluate the costs and clinical effectiveness of human papilloma virus screening compared with annual papanicolaou cytology in Germany. *European Journal of Obstetrics, Gynecology, and Reproductive Biology*, *212*, 132–139. https://doi.org/10.1016/j.ejogrb.2017.03.029

Sawaya, G. F., Smith-McCune, K., & Kuppermann, M. (2019). Cervical Cancer Screening: More Choices in 2019. *JAMA*, *321*(20), 2018–2019. https://doi.org/10.1001/jama.2019.4595

Schmidt, D., Bergeron, C., Denton, K. J., Ridder, R., & European CINtec Cytology Study Group. (2011). p16/ki-67 dual-stain cytology in the triage of ASCUS and LSIL papanicolaou cytology: Results from the European equivocal or mildly abnormal Papanicolaou cytology study. *Cancer Cytopathology*, *119*(3), 158–166.

https://doi.org/10.1002/cncy.20140

Stoler, M. H., Schiffman, M., & for the Atypical Squamous Cells of Undetermined Significance–Low-grade Squamous Intraepithelial Lesion Triage Study (ALTS) Group. (2001). Interobserver Reproducibility of Cervical Cytologic and Histologic InterpretationsRealistic Estimates From the ASCUS-LSIL Triage Study. *JAMA*, *285*(11), 1500–1505. https://doi.org/10.1001/jama.285.11.1500

Vahadane, A., Peng, T., Albarqouni, S., Baust, M., Steiger, K., Schlitter, A. M., Sethi, A., Esposito, I., & Navab, N. (2015). Structure-preserved color normalization for histological images. *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, 1012–1015. https://doi.org/10.1109/ISBI.2015.7164042

Wang, H., & Chow, S.-C. (2007). Sample Size Calculation for Comparing Proportions. In *Wiley Encyclopedia of Clinical Trials*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780471462422.eoct005

Wentzensen, N., Clarke, M. A., Bremer, R., Poitras, N., Tokugawa, D., Goldhoff, P. E., Castle, P. E., Schiffman, M., Kingery, J. D., Grewal, K. K., Locke, A., Kinney, W., & Lorey, T. S. (2019). Clinical Evaluation of Human Papillomavirus Screening With p16/Ki-67 Dual Stain Triage in a Large Organized Cervical Cancer Screening Program. *JAMA Internal Medicine*, *179*(7), 881–888. https://doi.org/10.1001/jamainternmed.2019.0306

Wentzensen, N., Fetterman, B., Castle, P. E., Schiffman, M., Wood, S. N., Stiemerling, E., Tokugawa, D., Bodelon, C., Poitras, N., Lorey, T., & Kinney, W. (2015). P16/Ki-67 Dual Stain Cytology for Detection of Cervical Precancer in HPV-Positive Women. *JNCI: Journal of the National Cancer Institute*, *107*(12), djv257. https://doi.org/10.1093/jnci/djv257

Wentzensen, N., Lahrmann, B., Clarke, M. A., Kinney, W., Tokugawa, D., Poitras, N., Locke, A., Bartels, L., Krauthoff, A., Walker, J., Zuna, R., Grewal, K. K., Goldhoff, P. E., Kingery, J. D., Castle, P. E., Schiffman, M., Lorey, T. S., & Grabe, N. (2021). Accuracy and Efficiency of Deep-Learning–Based Automation of Dual Stain Cytology in

Cervical Cancer Screening. *JNCI: Journal of the National Cancer Institute*, *113*(1), 72–79. https://doi.org/10.1093/jnci/djaa066

Wentzensen, N., Schiffman, M., Palmer, T., & Arbyn, M. (2016). Triage of HPV positive women in cervical cancer screening. *Journal of Clinical Virology*, *76*, S49–S55. https://doi.org/10.1016/j.jcv.2015.11.015

Wentzensen, N., Schwartz, L., Zuna, R. E., Smith, K., Mathews, C., Gold, M. A., Allen, R. A., Zhang, R., Dunn, S. T., Walker, J. L., & Schiffman, M. (2012). Performance of p16/Ki-67 Immunostaining to Detect Cervical Cancer Precursors in a Colposcopy Referral Population. *Clinical Cancer Research*, *18*(15), 4154–4162. https://doi.org/10.1158/1078-0432.CCR-12-0270

Wright Jr, T. C., Stoler, M. H., Behrens, C. M., Sharma, A., Sharma, K., & Apple, R. (2014). Interlaboratory variation in the performance of liquid-based cytology: Insights from the ATHENA trial. *International Journal of Cancer*, *134*(8), 1835–1843. https://doi.org/10.1002/ijc.28514

Wright, T. C., Behrens, C. M., Ranger-Moore, J., Rehm, S., Sharma, A., Stoler, M. H., & Ridder, R. (2017). Triaging HPV-positive women with p16/Ki-67 dual-stained cytology: Results from a sub-study nested into the ATHENA trial. *Gynecologic Oncology*, *144*(1), 51–56. https://doi.org/10.1016/j.ygyno.2016.10.031

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). *Learning Transferable Architectures for Scalable Image Recognition* (arXiv:1707.07012). arXiv. https://doi.org/10.48550/arXiv.1707.07012