PATHOLOGICA 2023;115:221-226;
DOI: 10.32074/1591-951X-896

Original Article

Diagnostic concordance between traditional and digital workflows. A study on 1427 prostate biopsies

Evelin Torresani[1], Maria Adalgisa Gentilini[2], Stefano Grassi[3], Luca Cima[1], Irene Pedrolli[1], Tommaso Cai[4], Marco Puglisi[4], Valentino Vattovani[4], Bianca Guadin[5], Matteo Brunelli[6], Claudio Doglioni[3], Mattia Barbareschi[1,7]

[1] Unit of Surgical Pathology, Santa Chiara Hospital, APSS, Trento, Italy; [2] Epidemiology and Clinical Evaluation Unit, APSS, Trento, Italy; [3] Department Pathology, Vita e salute University, San Raffaele Hospital, Milano, Italy; [4] Unit of Urology, Santa Chiara Hospital, APSS, Trento, Italy; [5] Experimental and Applied Biology, University of Pavia, Italy; [6] Pathology Unit, Department of Diagnostics and Public Health, University and Hospital Trust of Verona, Verona, Italy ; [7] Centre for Medical Sciences - CISMed, University of Trento, Italy

Summary

Objective. To evaluate intra-observer diagnostic reproducibility using traditional slides (TS) versus whole slide images (WSI).

Methods. TS and WSI of 1427 prostatic biopsies (107 consecutive patients) were evaluated by a single pathologist. Agreement between readings was evaluated with Gwet's Agreement coefficient (AC) and Landis and Koch benchmark scale.

Results. The positive/negative agreement between the readings was almost perfect (AC_1= 0.962; 95% CI[0.949,0.974]), with method independent distribution of discrepancies. Among positive biopsies, 212 had identical Gleason score (GS) on TS and WSI and discordant GS in 69 cases (AC_2 = 0.932; 95% CI[0.907, 0.956]). Concordant negative and positive patient classification was observed in 39 and 64 cases, respectively; two cases were assigned to the positive group on TS and 2 on WSI configuring an almost perfect agreement (AC_1=0.929; 95% C1[0.860, 0.998]). ISUP Grade group (ISUP GG) agreement was evaluated in the 60 concordantly positive cases: in 45 cases it was identical on TS and WSI; in 10 biopsies the discrepancy implied a modification of the assigned ISUP GG of ≤ 1 class and in 5 the discrepancy implied a modification of 2 classes. Gwet's agreement coefficient was (95% CI [0.834, 0.962]), i. e.: almost perfect agreement.

Conclusions. Our data show almost perfect agreement between digital and traditional diagnostic activity in a routine setting, confirming that digital pathology can be safely introduced into routine workflows.

Key words: digital pathology, prostate, prostatic neoplasms, histopathology, human

Received and accepted: June 4, 2023

Correspondence
Mattia Barbareschi
S. Chiara Hospital, Trento, Largo Medaglie Oro 9, 38122 Trento
E-mail: mattia.barbareschi@apss.tn.it

How to cite this article: Torresani E, Gentilini MA, Grassi S, et al. Diagnostic concordance between traditional and digital workflow. A study on 1427 prostatic biopsies.. Pathologica 2023;115:221-226. https://doi.org/10.32074/1591-951X-896

Introduction

Digital pathology (DP), based on the substitution of traditional glass slides (TS) with their digital counterpart, whole slide images (WSI), is prone to modify most of our routine workflow. Borne almost 20 years ago, DP was initially considered an exotic tool, but it rapidly suggested a new world of possibilities [1]. In the last few years DP has acquired the technological maturity to allow its extensive introduction in routine practice, even if few laboratories have adopted a fully digital diagnostic workflow [2]. DP has many advantages over traditional pathology (TP). DP can increase safety of patients and quality of diagnostic activity, including: direct link between images and personal data sheets, possibility to easily achieve distant consults by well recognized experts, compare images of the same

slides stained with different immunostains, compare previous slides of the same patient, upload WSI in the patient's electronic fascicle. Moreover the most promising advantage will be the introduction of artificial intelligence (AI) algorithms which will enhance our diagnostic ability. AI will probably become an extremely useful tool to support pathologists in reducing false positives and especially false negative diagnosis and will help to reduce (or probably eliminate) inter/intra- observer disagreements when qualitative parameters are to be taken into account.

Despite the enthusiasm for DP, there are still limited data comparing the diagnostic ability of pathologists working with TS and WSI. For pathologists, reading WSI implies the modification of a long-standing practice and a different cognitive approach with the adaptation to a new working environment.

As a preliminary step towards the implementation of DP and AI in our routine practice, we decided to evaluate intra-observer reproducibility of routine diagnostic activity based on WSI and TS of a series of prostatic biopsies. The aim of the present study is to verify whether the use of WSI can modify the diagnostic performance of pathologists working in a routine setting, using prostate biopsies as a testing model. Prostate biopsies were selected as a testing model because the diagnostic reproducibility may be focused on different levels of reporting, ranging from the distinction of positive versus negative cases to more subjective evaluations such as perineural invasion and the Gleason score, which are known to be subject to severe interobserver variation [3].

Materials and methods

We selected a consecutive series of 107 patients who underwent prostate biopsy between January 2021 and March 2021 in the Hospitals of the *Azienda Provinciale per i Servizi Sanitari, of Trento* (Italy). Each patient underwent at least 8 core biopsies which were individually processed and paraffin embedded corresponding to a total of 1431 slides, each containing 2 consecutive sections of the same core. The sections had been stained with conventional hematoxylin and eosin (H&E) stains using an automated stainer (Leica ST5010 Autostainer XL, https://www.leicabiosystems.com). All initial H&E sections were digitized using a high throughput device (P1000, 3DHitstech, Epredia, Italy). Each original TS (not additional levels nor immunostains) and corresponding WSI were evaluated by a single pathologist with a 4-week wash-out period between the two readings. The WSI reading was done on a usual office 32-inch monitor using the "Slide viewer 2.6" platform (3DHistech). Diagnosis was made according to the 2022 WHO classification [4,5]. For each reading done on TS and on WSI, we collected the following parameters: normal tissue/atypical small acinar proliferation/cancer, length of the cancer foci, percentage of cancer relative to the lengths of the tissue cores, major and minor Gleason Scores (GS) and their relative percentage, and perineural invasion. In cases with 3+4 or 4+3 GS and neoplastic foci larger than 1 mm in maximum diameter, we recorded the percentage of pattern 4 in 10% incremental groups. In biopsies with pattern 4 we recorded the presence of cribriform pattern. This data collection gave origin to 2 different datasets, containing up to 9989 single items. The ISUP grade group (ISUP GG) was recorded for each patient with at least 1 neoplastic biopsy. For statistical evaluation of diagnostic agreement, atypical small acinar proliferation (ASAP) was grouped with positive cases.

Statistical analysis

The characteristics of the diagnostic datasets are presented as mean and standard deviation for continuous variables, median and range for ordinal variables, absolute frequencies and percentages for nominal variables. The agreement between the pairs of the diagnostic sets was calculated with the Gwet's Agreement coefficient (AC) for the nominal variables and the Gwet's weighted AC for the ordinal variables. ACs were calculated with 95% confidence interval. The degree of agreement of the ACs was compared with the Landis and Koch benchmark scale. McNemar and chi-squared test was used to compare categorical variables. Statistical significance was defined with alpha equal to 0.05. All analyses were performed with SAS 9.4. Agreement graphs were elaborated with SAS 9.4 and other graphs with MS Excel.

Results

TS readings

Out of 1431 TS, 294 (21%) were diagnosed as prostate adenocarcinoma, 32 as ASAP (2%) and 1105 (77%) as negative. In 2 of the 294 positive slides GS couldn't be assigned. Out of 159 biopsies with 3+4 and 4+3 GS the percentage of pattern 4 was recorded in 139; in 20 biopsies the percentage could not be evaluated due to artifacts or limited amount of neoplas-

tic tissue. A cribriform pattern was found in 92 (43%) out of the 214 biopsies containing pattern 4. Median tumor length was 4.6 mm +/- 4.2 SD. Percentage of cancer relative to biopsy length was 38.3% +/- 33.1 SD. Perineural invasion was found in 49 (17%) of 294 positive biopsies.
The 294 positive biopsies corresponded to 66 patients (62% of the series of 107 patients).

WSI readings

Out of 1431 biopsies, 4 could not be properly scanned, leaving 1427 WSI available; 290 (20%) were diagnosed as prostate adenocarcinoma, 41 as ASAP (3%) and 1096 (77%) as negative. Out of 155 biopsies with 3+4 and 4+3 GS the percentage of pattern 4 was recorded in 134; in 21 biopsies the percentage could not be evaluated due to a limited amount of neoplastic tissue. A cribriform pattern was found in 99 (46%) out of the 213 biopsies containing pattern 4. Median tumor length was 4.6 mm +/- 4.3 SD. Percentage of cancer relative to biopsy length was 38.4% +/- 32.7 SD. Perineural invasion was found in 40 (14%) of 290 positive biopsies.
The 290 positive biopsies were from 66 patients (62% of the series of 107 patients).

Agreement of TS vs WSI readings at the biopsy level

The agreement between the two readings was evaluated on 1427 of 1431 slides as 4 WSI was lacking (Tab. I, Fig. 1). 1082 cases were diagnosed as negative and 310 as positive in the two readings. 14 cases were positive only on TS readings and 21 were positive only on WSI readings. Gwet's agreement coefficient was 0.962 (0.949-0.974 95% CI), which corresponds to almost perfect agreement according to the Landis and Koch benchmark scale. Discordant cases were evaluated with the McNemar test, which was statistically significant ($\chi^2(1) = 1.4$, $p = 0.237$), pointing to method independent distribution. The critical revision of these discrepant cases showed that in real practice these biopsies would have been submitted to additional ancillary techniques or further sections, which was excluded by the design of the present study.
GS was evaluable in 281 paired biopsies: in 212 cases the GS was identical on TS and WSI. Discordant GS evaluations were recorded in 69 cases: in 53 biopsies the discrepancy implied a modification of the assigned ISUP GG of ≤ 1 and in 17 the discrepancy implied a modification of the assigned ISUP GG of > 1. Gwet's weighted agreement coefficient was 0.932 (0.907-0.956 95% CI), which corresponds to almost perfect agreement according to the Landis and Koch benchmark scale (Tab. II, Fig. 2).
Perineural invasion detection was investigated in 282 cases: 226 were negative and 32 were positive both on TS and WSI, while 24 biopsies showed discrepant results, with 17 cases positive on TS and 7 only in WSI. Gwet's agreement coefficient was 0.885 (0.837-0.932 95% CI), which corresponds to almost perfect agreement according to the Landis and Koch benchmark scale. Discordant cases were evaluated with the McNemar test, which was statistically significant ($\chi^2(1) = 4.17$, $p = 0.041$), pointing to a method-related distribution. (Tab. III, Fig. 3).
Analysis of the percentage of adenocarcinoma tissue in comparison with biopsy length identified with the two methods showed no difference between the two methods (Wilkoxon rango test -12.5, $p = 0.975$).

Agreement of TS vs WSI readings at the patient level

Diagnostic agreement was evaluated considering the 107 patients included in the study. Concordant negative and positive patient classification was observed

Table I. Positive vs negative agreement on 1427 slides between TS and WSI readings (4 missing cases).

	WSI Positive	WSI Negative	Total
TS Positive	310	14	324
TS Negative	21	1082	1103
Total	331	1096	1427

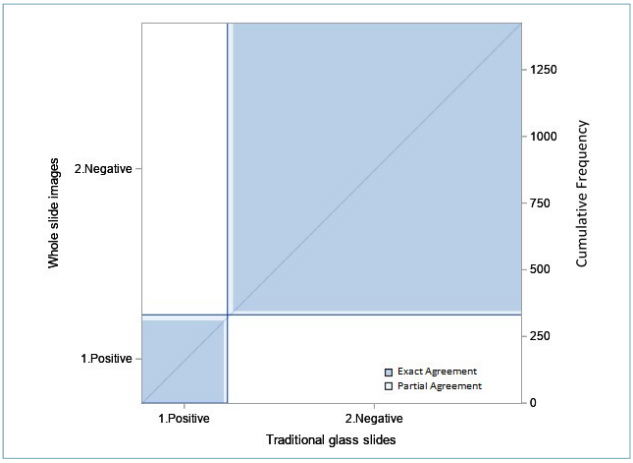

Figure 1. Diagnostic agreement for all 1427 slides. In ordinate are WSI paired with TS on abscissa, subdivided as positive or negative; dark blue squares represent perfectly concordant diagnoses.

Table II. Gleason score agreement on 281 positive slides between TS and WSI readings.

		WSI GS	WSI GS	WSI GS	WSI GS	WSI GS	WSI GS	WSI GS		
TS GS		**1.3+3**	**2.3+4**	**3.4+3**	**4.4+4**	**5.3+5**	**7.4+5**	**8.5+4**	**Total**	**Percent**
TS GS	**1.3+3**	58	13	0	0	0	0	0	71	[25,3%]
TS GS	**2.3+4**	12	78	3	0	3	0	0	96	[34,2%]
TS GS	**3.4+3**	0	10	39	1	1	9	0	60	[21,4%]
TS GS	**4.4+4**	1	0	5	23	0	3	0	32	[11,4%]
TS GS	**5.3+5**	0	2	0	0	0	0	0	2	[0,7%]
TS GS	**7.4+5**	0	0	2	0	0	13	0	15	[5,3%]
TS GS	**8.5+4**	0	0	0	0	0	4	1	5	[1,8%]
	Total	71	103	49	24	4	29	1	281	[100%]
	Percent	[25,3%]	[36,7%]	[17,4%]	[8,5%]	[1,4%]	[10,3%]	[0,4%]	[100%]	

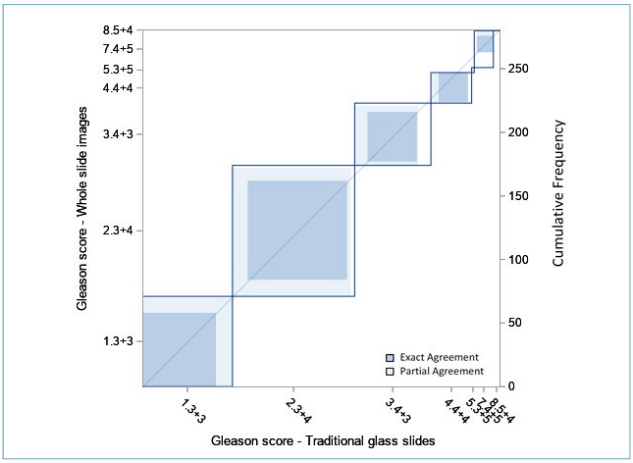

Figure 2. Gleason score diagnostic agreement for 281 slides. In ordinate are WSI data paired with TS on abscissa, subdivided on the basis of increasing Gleason score; dark blue squares represent perfectly concordant diagnoses.

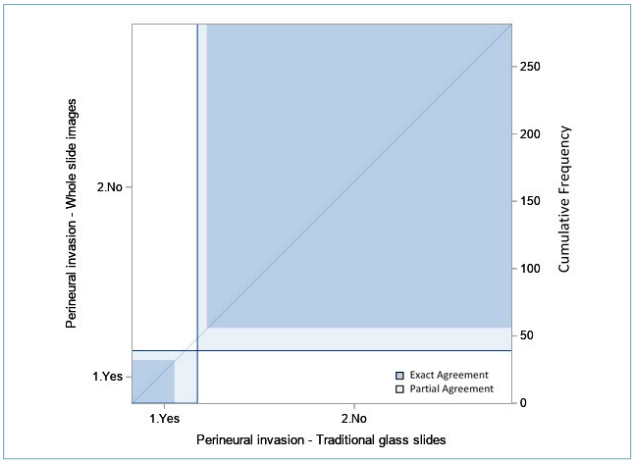

Figure 3. Perineural invasion diagnostic agreement for 282 slides. In ordinate are WSI data paired with TS on abscissa, subdivided as positive or negative; dark blue squares represent perfectly concordant diagnoses.

Table III. Perineural invasion agreement on 282 positive slides between TS and WSI reading.

	WSI Positive	WSI Negative	Total
TS Positive	32	17	49
TS Negative	7	226	233
Total	39	243	282

in 39 and 64 cases, respectively (Tab. IV). Two cases were assigned to the positive group only using TS and 2 using only WSI. Gwet's agreement coefficient was 0.929 (0.860-0.998 95% CI), which corresponds to almost perfect agreement according to the Landis and Koch benchmark scale. Discordant cases were evaluated with the McNemar test, which was not statistically significant ($\chi^2(1) = 0$, $p = 0.1$), pointing to method independent distribution (Fig. 4).

ISUP GG agreement was evaluated in 60 cases: in 45 cases the GS was identical on TS and WSI (Tab. V); in 10 cases the discrepancy implied a modification of the assigned ISUP GG of ≤ 1 class and in 5 the discrepancy implied a modification of the assigned ISUP GG of 2 classes. Gwet's weighted agreement coefficient was 0.898 (0.834-0.962 95% CI), which corresponds to almost perfect agreement according to the Landis and Koch benchmark scale (Fig. 5).

Table IV. Positive versus negative patient agreement between TS and WSI reading.

	WSI Positive	WSI Negative	Total
TS Positive	64	2	66
TS Negative	2	39	41
Total	66	41	107

Gwet's agreement coefficient: 0.929 (0.860-0.998 95% CI)

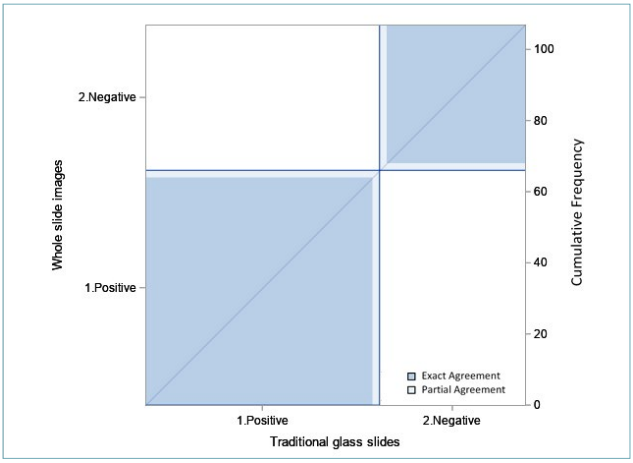

Figure 4. Diagnostic agreement for all 107 patients. In ordinate are WSI paired with TS on abscissa, subdivided as positive or negative; dark blue squares represent perfectly concordant diagnoses.

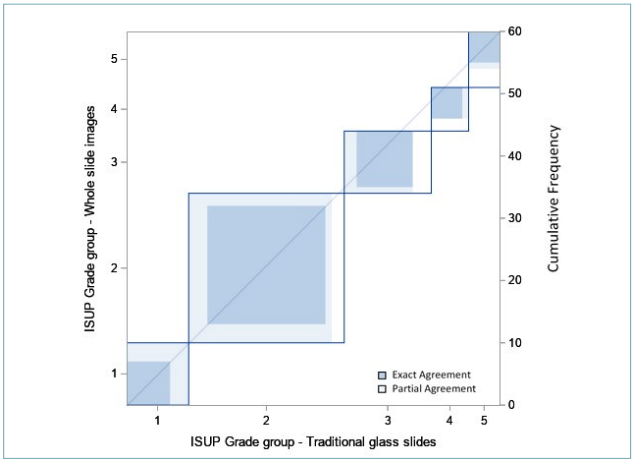

Figure 5. ISUP Grade group diagnostic agreement for all 107 patients. In ordinate are WSI paired with TS on abscissa, subdivided on the basis of increasing ISUP Grade group; dark blue squares represent perfectly concordant diagnoses.

Discussion

Our study strongly supports the reliability of digital diagnostic workflow. The almost perfect agreement between digital and traditional diagnostic activity in a routine setting confirms previous data on more limited series of prostatic biopsies [6] that DP can be safely introduced in the routine workflow in our surgical pathology laboratories.

The study, performed in keeping with international guidelines for WSI workflow validation [7], confirms that WSI evaluation in routine workflow provides the same level of morphological information as TS. We showed extremely high agreement in distinguishing negative from positive biopsies and in Gleason score evaluation, which is more subtle and known to be prone to inter- and even of intra-observer variability [8-10]. Our agreement for Gleason score was higher than previously reported, probably reflecting an increase in quality of the images [11].

With the exception of perineural invasion detection, the few discordant cases in our study were not related to the analytical method (TS vs WSI) but represent the normal intrinsic difficulties in histopathological evaluation. The critical revision of these discrepant cases showed that in real practice these cases would have been analyzed with additional ancillary techniques or further sections, whose use was excluded by the design of the present study. Gleason score discrepancies were very limited and reasonably related to the known intrinsic difficulty in the identification and quantification of Gleason patterns. Only detection of perineural invasion, although statistically almost perfectly correlated in the two methods, showed some discrepancies which were method-related, with slightly more cases detected using TS, suggesting that this could be a critical aspect which needs particular attention when reading WSI. At any rate, it is well known that diagnostic discrepancies in perineural invasion evaluation are quite frequently described, with reported

Table V. ISUP GG agreement between TS and WSI reading.

		WSI ISUP GG	WSI ISUP GG	WSI ISUP GG	WSI ISUP GG	WSI ISUP GG		
		1	**2**	**3**	**4**	**5**	**Total**	**Percent**
TS ISUP GG	**1**	7	3	0	0	0	10	[16.7%]
TS ISUP GG	**2**	3	19	1	2	0	25	[41.7%]
TS ISUP GG	**3**	0	2	9	0	3	14	[23.3%]
TS ISUP GG	**4**	0	0	0	5	1	6	[10%]
TS ISUP GG	**5**	0	0	0	0	5	5	[8.3%]
	Total	10	24	10	7	9	60	[100%]
	Percent	[16.7%]	[40%]	[16.7%]	[11.7%]	[15%]	[100%]	

kappa values for interobserver variability in the 0.67-0.75 range [12].

The pathologist who did the diagnostic evaluation (ET) in the present study had not been previously involved in a complete DP workflow and the present activity was her first real life experience in reading a large series of WSI. This underscores the fact that the transition from a traditional to a digital workflow does not need any specific lengthy training and may be rapidly implemented in our services. Indeed, several experiences around the world have shown that once pathologists accept the DP workflow, the transition to the new environment is quite easy and rapid [13], and that DP in-house validation can rely on limited sets of cases [14].

Conclusions

In conclusion, our study confirms that a completely digital pathology workflow can be considered completely equivalent to the traditional one, further supporting the ease of such transition. In our opinion, digital pathology will be the future of our discipline and will provide many advantages in terms of workflow and diagnostic quality, especially with the introduction of artificial intelligence (AI) supported diagnostic algorithms [15], which has the promise to improve turnaround time, help to cope with an increasing workload in face of decreasing number of pathologists, and increase diagnostic reliability and reproducibility, especially in those fields where interobsever diagnostic variability is high.

Conflicts of interest

The Authors declare no conflict of interest.

Funding

The research has been founded by the Azienda Provinciale per i Servizi Sanitari, "Fondo di ricerca".

Authors' contributions

ET: principal investigator of the study: reading of the slides and use of the artificial intelligence algorithm, data interpretation and manuscript preparation; MB: project of the study, supervision of the study, data interpretation, quality control and manuscript preparation; MAG: statistical analysis data interpretation, and manuscript preparation. All other authors contributed to data interpretation and manuscript preparation.

Ethical consideration

Ethical approval has been obtained by the APSS ethical committee.

References

1 Barbareschi M, Demichelis F, Forti S, et al. Digital pathology: science fiction? Int J Surg Pathol 2000;8:261-263. https://doi.org/10.1177/106689690000800401

2 Fraggetta F, Pantanowitz L. Going fully digital: utopia or reality? Pathologica 2018;110:1-2.

3 Flach RN, Willemse PM, Suelmann BBM, et al. Significant inter- and intralaboratory variation in gleason grading of prostate cancer: a nationwide study of 35,258 patients in the Netherlands. Cancers (Basel) 2021;13:5378. https://doi.org/10.3390/cancers13215378

4 Surintrspanont J, Zhou M. Prostate Pathology: What is New in the 2022 WHO Classification of Urinary and Male Genital Tumors? Pathologica. 2022;115:41-56. https://doi.org/10.32074/1591-951X-822

5 Board WCoTE. Urinary and Male Genital Tumours. 5th Edition. Lyon (France): International Agency for Research on Cancer 2022.

6 Rodriguez-Urrego PA, Cronin AM, Al-Ahmadie HA, et al. Interobserver and intraobserver reproducibility in digital and routine microscopic assessment of prostate needle biopsies. Hum Pathol 2011;42:68-74. https://doi.org/10.1016/j.humpath.2010.07.001.

7 Evans AJ, Brown RW, Bui MM, et al. Validating whole slide imaging systems for diagnostic purposes in pathology. Arch Pathol Lab Med 2022;146:440-450. https://doi.org/10.5858/arpa.2020-0723-CP

8 Oyama T, Allsbrook WC Jr, Kurokawa K, et al. A comparison of interobserver reproducibility of Gleason grading of prostatic carcinoma in Japan and the United States. Arch Pathol Lab Med 2005;129:1004-10. https://doi.org/10.5858/2005-129-1004-ACOIRO

9 Ozkan TA, Eruyar AT, Cebeci OO, et al. Interobserver variability in Gleason histological grading of prostate cancer. Scand J Urol 2016;50:420-424. https://doi.org/10.1080/21681805.2016.1206619

10 van Santvoort BWH, van Leenders GJLH, Kiemeney LA, et al. Histopathological re-evaluations of biopsies in prostate cancer: a nationwide observational study. Scand J Urol 2020;54:463-469. https://doi.org/10.1080/21681805.2020.1806354

11 Rodriguez-Urrego PA, Cronin AM, Al-Ahmadie HA, et al. Interobserver and intraobserver reproducibility in digital and routine microscopic assessment of prostate needle biopsies. Hum Pathol 2011;42:68-74. https://doi.org/10.1016/j.humpath.2010.07.001

12 Egevad L, Delahunt B, Samaratunga H, et al. Interobserver reproducibility of perineural invasion of prostatic adenocarcinoma in needle biopsies. Virchows Arch 2021;478:1109-1116. https://doi.org/10.1007/s00428-021-03039-z

13 Rojansky R, Jhun I, Dussaq AM, et al. Rapid deployment of whole slide imaging for primary diagnosis in surgical pathology at Stanford medicine: responding to challenges of the COVID-19 pandemic. Arch Pathol Lab Med 2023;147:359-367. https://doi.org/10.5858/arpa.2021-0438-OA

14 Evans AJ, Brown RW, Bui MM, et al. Validating whole slide imaging systems for diagnostic purposes in pathology. Arch Pathol Lab Med 2022;146:440-450. https://doi.org/10.5858/arpa.2020-0723-CP

15 Flach RN, Fransen NL, Sonnen AFP, et al. Implementation of Artificial Intelligence in diagnostic practice as a next step after going digital: the UMC Utrecht perspective. Diagnostics (Basel) 2022;12:1042. https://doi.org/10.3390/diagnostics12051042