Original Article

Selection of Representative Histologic Slides in Interobserver Reproducibility Studies: Insights from Expert Review for Ovarian Carcinoma Subtype Classification

Marios A. Gavrielides[1], Brigitte M. Ronnett[2], Russell Vang[2], Fahime Sheikhzadeh[3], Jeffrey D Seidman[4]

[1]Division of Imaging, Diagnostics, and Software Reliability, Office of Science and Engineering Laboratories , Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, USA, (Currently at AstraZeneca, Precision Medicine and Biosamples, Gaithersburg, Maryland, USA), [2]Department of Pathology and Gynecology and Obstetrics, The Johns Hopkins Hospital, Baltimore, Maryland, USA, [3]Electrical and Computer Engineering Department, University of British Columbia, Vancouver, Canada, (Currently at Roche Diagnostics, San Francisco, California, USA), [4]Division of Molecular Genetics and Pathology, Office of In Vitro Diagnostics and Radiological Health, Office of Product Evaluation and Quality, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, USA

Submitted: 25-Jun-2020 Revised: 02-Sep-2020 Accepted: 28-Oct-2020 Published: 22-Mar-2021

Abstract

Background: Observer studies in pathology often utilize a limited number of representative slides per case, selected and reported in a nonstandardized manner. Reference diagnoses are commonly assumed to be generalizable to all slides of a case. We examined these issues in the context of pathologist concordance for histologic subtype classification of ovarian carcinomas (OCs). **Materials and Methods:** A cohort of 114 OCs consisting of 72 cases with a single representative slide (Group 1) and 42 cases with multiple representative slides (148 slides, 2–6 sections per case, Group 2) was independently reviewed by three experts in gynecologic pathology (case-based review). In a follow-up study, each individual slide was independently reviewed in a randomized order by the same pathologists (section-based review). **Results:** Average interobserver concordance varied from 100% for Group 1 to 64.3% for Group 2 (86.8% across all cases). Across Group 2, 19 cases (45.2%) had at least one slide classified as a different subtype than the subtype assigned from case-based review, demonstrating the impact of intratumoral heterogeneity. Section-based concordance across individual sections from Group 2 was comparable to case-based concordance for those cases indicating diagnostic challenges at the individual section level. Findings demonstrate the increased diagnostic complexity of heterogeneous tumors that require multiple section sampling and its impact on pathologist performance. **Conclusions:** The proportion of cases with multiple representative slides in cohorts used in validation studies, such as those conducted to evaluate artificial intelligence/machine learning tools, can influence diagnostic performance, and if not accounted for, can cause disparities between research and real-world observations and between research studies. Case selection in validation studies should account for tumor heterogeneity to create balanced datasets in terms of diagnostic complexity.

Keywords: Observer study, ovarian cancer, pathologist concordance, representative slides, subtype classification

Introduction

Observer studies in pathology are often conducted to examine interobserver concordance for a variety of diagnostic activities. These include traditional diagnostic tasks, such as rendering diagnoses in general, tumor subtyping, and assessing biomarkers,[1-3] as well as newer technologies and applications. The latter include evaluating whole slide imaging (WSI) for primary diagnosis,[4-10] assessing tools for clinical decision support[11] or computer-aided diagnosis,[12-14] and providing reference diagnoses for annotated datasets needed for the development and validation of image analysis technologies, including artificial intelligence and machine learning (AI/ML) algorithms.[15,16] The study design and related parameters of observer studies, such as the number and experience of observers, number and diversity of cases, reading protocols

Access this article online

Quick Response Code:

Website:
www.jpathinformatics.org

DOI:
10.4103/jpi.jpi_56_20

Address for correspondence: Dr. Marios A. Gavrielides,
One MedImmune Way, Gaithersburg, Maryland 20878, USA.
E-mail: marios.gavrielides@astrazeneca.com

For reprints contact: WKHLRPMedknow_reprints@wolterskluwer.com

How to cite this article: Gavrielides MA, Ronnett BM, Vang R, Sheikhzadeh F, Seidman JD. Selection of representative histologic slides in interobserver reproducibility studies: Insights from expert review for ovarian carcinoma subtype classification. J Pathol Inform 2021;12:15.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2021/12/1/15/311692

and procedures, and instructions to observers, among others, need to be carefully selected so that the findings from such preclinical studies can be comparable to observations in clinical practice.

One such study design parameter is the selection of histologic sections for pathology review from each case within an observer study. Even though all tissue materials per case are reviewed in routine clinical practice,[17] observer studies often utilize a limited number of representative sections per case due to dependence on the time and availability of multiple pathologists and the need for an adequate number of independent cases. The approach for selecting representative sections for pathology review is not standardized. Among the observer studies focusing on the validation of WSI for primary diagnosis, for instance, some studies reported utilizing all available sections per case,[5,6,10,18-20] others utilized a single representative section per case,[21-23] while for others, the number of representative sections varied from case to case within the study, ranging from a single-section to multiple sections per case.[4,7,24] For the latter category, the distribution of the number of sections across cases has not been reported. Some authors have reported the mean[7] or range[4,7] of the number of representative sections per case in their dataset.

Differences in the section selection approach could contribute to decreased reproducibility between research studies as well as between research studies and clinical practice. One issue relates to the spectrum of diagnostic difficulty of cases reviewed within a study. Cases that can be represented with single sections likely present typical histology across all sections and could pose lesser diagnostic challenges compared to heterogeneous tumors. That would lead to discrepancies between findings of studies having different proportions of such cases. The selection of representative slides has been identified as a major drawback in validation studies of WSI,[25] but its effect on pathologist performance has not been investigated, to the best of our knowledge. Another issue related to the selection of representative slides is the generalizability of case diagnosis to the individual sections of that case. This can be a factor when establishing reference diagnoses in validation studies of AI/ML software for clinical decision support. Algorithms developed and tested on representative sections from such tumors with inaccurate labeling could be biased toward making certain diagnoses over others.

In this study, we examine the tissue selection issues discussed above in the context of pathologist concordance for histological subtype classification of ovarian carcinomas (OCs). Epithelial ovarian cancer is not one disease; it comprises at least five distinct histological subtypes,[26] namely high-grade serous carcinoma (HGSC), low-grade serous carcinoma (LGSC), clear cell carcinoma (CCC), endometrioid carcinoma (EC), and mucinous carcinoma (MUC), and a few less common variant subtypes. OCs often show intratumoral heterogeneity as evident by admixtures of different morphological patterns side by side within an individual neoplasm,[27] making sample selection for histology review, particularly important and vulnerable to bias.[28] The focus of the study was two-fold; the first objective was to quantify interobserver concordance for a panel of experienced gynecologic pathologists across cases consisting of single and multiple representative sections to examine whether and how they provided different diagnostic challenges. The second objective was to examine how often individual sections from OC cases shared the same subtype classification as the reference subtype diagnoses for those cases.

Materials and Methods

Ovarian carcinoma cohort

A cohort of 114 OCs which was described in Seidman *et al.*[29] was used in this study. It was created by accruing 60 consecutive cases from the gynecology and gynecologic oncology service of a large community and tertiary care hospital (MedStar Washington Hospital Center, Washington, DC, USA) and then enriching for the less common subtypes (non-HGSCs), also as they appeared consecutively. The cases included 72 single-section and 42 multi-section cases (148 sections, 2–6 sections per case), for a total of 220 sections derived from formalin-fixed, paraffin-embedded tumor tissue and stained with hematoxylin and eosin (H and E). The sections were selected by the senior author (JDS) to include the invasive component of tumors. The number of selected slides per case was determined as needed to represent the histology of that case.

Study design

As previously reported,[29] three experienced gynecologic pathologists using a light microscope reviewed independently all available sections (slides) per case (case-based review) of the OC cohort and reported the histologic subtype for each case based on the 2014 World Health Organization guidelines for the classification of gynecological tumors.[30] Reference subtype diagnosis for each case was defined by the majority consensus diagnosis (at least 2 out of 3 pathologists agreeing on the subtype).

About a year following the case-based review, the same three pathologists conducted independent randomized order review of all individual sections in the cohort and were asked to diagnose the subtype of each particular section (individual slide) viewed (section-based review). Multiple review sessions were conducted to complete the study since a maximum of 40 sections per a 2-h session was allowed to mitigate fatigue. Randomizing the order of sections for review, as well as ensuring a minimum of 2 weeks between sessions, reduced the probability of sections from the same case being reviewed consecutively. Two weeks was deemed adequate since different sections were reviewed at each session. Each of the reviewers was particularly cognizant of the recently emphasized problematic areas in OC subtype classification as described in Seidman *et al.*[29]

At the completion of the section-based study, a follow-up consensus meeting was held to review discrepant cases (those for which there was no unanimous agreement from the independent reviews), regarding their histological subtype classification. The consensus meeting consisted of a review of discrepant cases at a multi-head microscope and a discussion on the results of the case-based and section-based studies aiming to describe specific reasons for discrepancies.

Statistical analysis

Primary outcome measures included interexpert concordance across all cases, cases with single slides, and cases with multiple slides, as well as across all individual sections. Interexpert concordance was measured by the mean and standard deviation of pairwise percent agreement between pairs of observers classifying the subtype of each case. The measure used to assess the generalizability of reference diagnosis was the proportion of cases having at least one section with a different subtype classification by the expert panel compared to the reference diagnosis by the same panel.

Results

Distribution of subtypes based on case-based and section-based review

The subtype diagnoses from case-based review of the cohort are tabulated in Table 1. For 93 of the 114 cases (81.6%), tumor subtype diagnosis was unanimous. For 18 of the 114 cases (15.8%), subtype diagnosis was not unanimous, with one expert providing a different classification while the other two experts were in agreement (majority consensus). For 3 cases out of 114 (2.6%), subtype could not be determined (referred to as undetermined) since all three experts had reported different subtype classifications. Only 21 of 42 (50%) cases with multiple sections had unanimous agreement, compared to 72 of 72 (100%) cases for which there was a single representative section.

Table 2 shows the distribution of subtype diagnoses across all individual sections in the cohort resulting from independent

Table 1: Per case subtype classification by independent review of panel of gynecologic pathologists

Ovarian carcinoma subtype classification	Number of cases by unanimous agreement (single-section/ multiple section)	Number of cases by majority agreement (single-section/multiple section)	Percentage of cases with unanimous agreement (%)	Percentage of cases with majority agreement (%)	Total number of cases in subtype category
High-grade serous	41 (39/2)	4 (0/4)	91.1	8.9	45
Low-grade serous	8 (6/2)	2 (0/2)	80	20.0	10
Mucinous	11 (6/5)	3 (0/3)	78.6	21.4	14
Endometrioid	9 (7/2)	1 (0/1)	90	10	10
Clear cell	11 (8/3)	0	100	0	11
Carcinosarcoma	8 (3/5)	5 (0/3)	61.5	38.5	13
Seromucinous	0	2 (0/2)	0	100	2
Brenner	5 (3/2)	0	100	0	5
Mixed	0	1 (0/3)	0	100	1
Undetermined*					3
Total	93 (72/21)	18 (0/18)	81.6	15.8	114

Majority agreement refers to subtype classification based on agreement by two out of three pathologists. *Undetermined refers to 3 cases (2.6%) where the three pathologists provided three different classifications (no agreement)

Table 2: Per section subtype classification by independent review of panel of gynecologic pathologists

Ovarian carcinoma subtype classification	Number of sections by unanimous agreement (% of total number of subtype sections)	Number of sections by majority agreement (% of total number of subtype sections)	Number of sections with no agreement	Total number of sections
High-grade serous	59 (75.6)	19 (24.4)		78
Low-grade serous	16 (80.0)	4 (20.0)		20
Mucinous	29 (67.4)	14 (32.6)		43
Endometrioid	9 (56.3)	7 (43.7)		16
Clear cell	16 (76.2)	5 (23.8)		21
Carcinosarcoma	11 (78.6)	3 (21.4)		14
Seromucinous	4 (50.0)	4 (50.0)		8
Brenner	7 (100)	0		7
Undetermined*			12	12
Total	151 (68.9)	56 (25.6)	12 (5.5)	219

Majority agreement refers to subtype classification based on agreement by two out of three pathologists. *Undetermined here refers to sections where the three pathologists provided three different classifications (no agreement) review. Sections from mucinous, endometrioid, and seromucinous carcinoma cases had the highest rates of nonunanimous agreement (32.6%, 43.7%, and 50% respectively).

Analysis of interobserver concordance for case-based subtype diagnosis

Table 3 shows observer concordance results averaged across pairs of experts for case-based OC subtype classification. Results show an overall mean agreement of 86.8% across all cases in the cohort. However, looking at subgroups of cases, the mean agreement was 100% for cases with a single representative section and 64.3% for cases with multiple slides. This finding indicates that the cases with single representative sections in this study were less challenging, suggesting typical histology for each subtype, compared to cases that needed multiple representative slides due to intratumoral heterogeneity and presented a more challenging diagnostic task. The finding also suggests that the distribution of single slide and multi-slide cases within a cohort could affect the overall assessment of interobserver concordance.

Analysis of interobserver concordance for section-based subtype diagnosis

Results of pathologist concordance analysis for section-based OC subtype classification are shown in Table 4. The average concordance rate across all individual sections in the cohort was 77.3% and dropped to 66.2% across sections from multi-section cases. Note that the mean agreement on the single representative sections from Table 3 was 100%. It was also observed that section-based interobserver concordance across individual sections from cases with multiple sections was comparable to case-based concordance for those cases (66.2% and 64.3%, respectively). This finding suggests that differences in pathologist concordance rates between cases with single or multiple slides were not due to variability from combining information from multiple slides into an overall diagnosis, but more likely due to cases represented by a single section being typical and less challenging than cases with multiple sections, even at the individual section level. Figure 1 illustrates a case for which a section had

Table 3: Pathologist case-based concordance across the whole cohort, across cases with multiple representative sections, and across cases with single representative sections

Majority classification	Analysis set					
	All cases in cohort		Cases with multiple representative sections		Cases with single representative sections	
	Number of cases	Average concordance (95% CI)	Number of cases	Average concordance (95% CI)	Number of cases	Average concordance (95% CI)
All subtypes	114	86.8 (80.4-92.1)	42	64.3 (50.0-77.0)	72	100 (100-100)
High-grade serous	45	94.1 (84.4-100.0)	6	55.6 (16.7-100.0)	39	100 (100-100)
Low-grade serous	10	86.7 (60.0-100.0)	4	66.7 (25.0-100.0)	6	100 (100-100)
Mucinous	14	85.7 (64.3-100.0)	8	75 (45.8-100.0)	6	100 (100-100)
Endometrioid	10	93.3 (73.3-100.0)	3	77.8 (33.3-100.0)	7	100 (100-100)
Clear cell	11	100.0 (100.0-100.0)	3	100 (100.0-100.0)	8	100 (100-100)
Carcinosarcoma	13	74.4 (46.2-100.0)	10	66.7 (30.0-100.0)	3	100 (100-100)
Other (SM, BR, mixed)	8	75.0 (45.8-100.0)	5	60.0 (20.0-93.3)	3	100 (100-100)
Undetermined	3	0	3	0	0	N/A

SM: Seromucinous, BR: Brenner, CI: Confidence interval, N/A: Not available

Table 4: Pathologist section-based concordance across all sections in the cohort and across sections from cases with multiple representative sections

	Analysis set			
	Sections from all cases in cohort		Sections from cases with multiple representative sections	
	Number of sections	Average concordance (95% CI)	Number of sections	Average concordance (95% CI)
All subtypes	220	77.3 (71.4-83.0)	148	66.2 (58.1-74.1)
High-grade serous	78	83.8 (75.2-91.0)	39	67.5 (53.0-80.3)
Low-grade serous	20	86.7 (70.0-98.3)	14	81.0 (57.1-97.6)
Mucinous	44	77.3 (61.4-90.9)	38	73.7 (56.1-87.7)
Endometrioid	16	70.8 (43.8-93.8)	9	48.1 (11.1-88.9)
Clear cell	21	84.1 (61.9-100.0)	13	74.4 (38.5-100.0)
Carcinosarcoma	14	85.7 (66.7-100.0)	11	81.8 (57.6-100.0)
Seromucinous	8	66.7 (37.5-91.2)	8	66.7 (37.5-91.7)
Brenner	7	100 (100.0-100)	4	100.0 (100-100)
Undetermined	12	0	12	0

CI: Confidence interval

different classifications by each of the three observers, likely attributable due to overlapping histologic features that can be encountered in different tumor subtypes.[27] In Figure 1, columnar cells, cytoplasmic clearing, and notable nuclear atypia with some hobnail morphology can be seen in serous, clear cell, and endometrioid differentiation.

Comparison between section and case diagnoses

Analysis of the association between the section-based and reference (case-based) diagnoses [Table 5] showed that out of the 42 cases with multiple sections, 19 cases (45.2%) had at least one section that the majority of the pathologists classified as a different subtype than the subtype assigned from case-based review. This finding includes three undetermined cases (for which each pathologist reported a different subtype from case-based review) that had individual sections with a majority classification.

Across subtypes, carcinosarcomas (CSs) had the largest percentages of cases with differing section classifications, probably due to diagnosis often based on small foci of a sarcomatous component that where present in only a single section within a case as can be seen in the example shown in Figure 2. Another example of intratumoral heterogeneity affecting the generalizability of a case-based diagnosis to all sections of the case is shown in Figure 3. The case-based consensus diagnosis was low-grade serous; however, the majority consensus diagnosis was LGSC for one section but HGSC for another section – a distinction that has therapeutic and prognostic consequences.

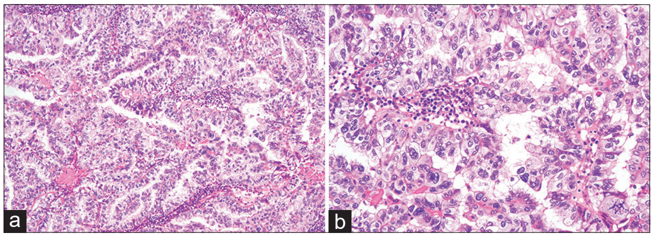

Figure 1: Glandular and papillary areas with more columnar cells having cytoplasmic clearing and notable nuclear atypia, from a section for which there was no consensus, as it was assessed as mixed clear cell and endometrioid by one observer, as clear cell by the second, and as high-grade serous by the third. The case from which this section was extracted received a majority consensus (2/3) classification of high-grade serous. (a and b) Images acquired with 10x and 20x objectives, respectively

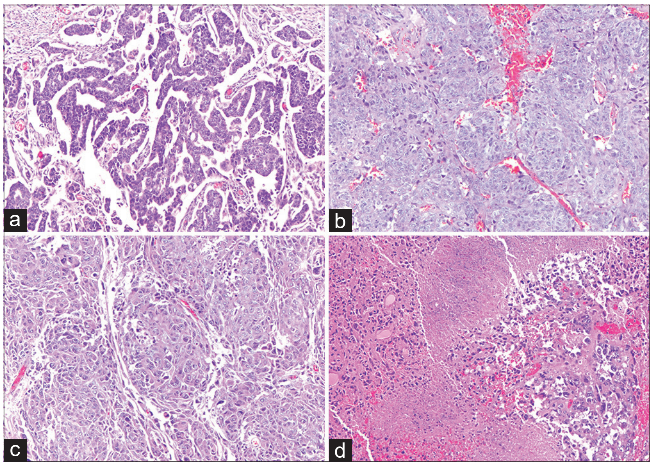

Figure 2: Example of case based consensus diagnosis (carcinosarcoma) differing from diagnoses on individual sections. (a) One slide was unanimously assessed as high-grade serous (3/3) in the section-based reads. (b and c) Other slides had a section-based consensus (2/3) interpretation of high-grade serous, with a minority assessment in both of these being endometrioid, likely due to some suggestion of squamous features. (d) Another slide had a section-based consensus (2/3) of carcinosarcoma, but one observer assessed this as unclassified. Despite the heterogeneity, the case-based (combined assessment of features in a-d) consensus (2/3) was carcinosarcoma, with the minority assessment being high-grade serous

Table 5: Intratumoral heterogeneity as indicated by number of cases containing at least one section for which per section and per case majority classification were different

Subtype classification based on case-based majority consensus	Number of cases	Number of cases including at least 1 differing section-based subtype classification	Differing section-based subtype classification
High-grade serous	6	2	CCC (1), UND (1)
Low-grade serous	4	1	HGSC (1)
Mucinous	8	1	UND (1)
Endometrioid	3	2	MUC (1), HGSC (1)
Clear cell	3	1	UND (1)
Carcinosarcoma	10	8	HGSC (6), CCC (2)
Seromucinous	2	0	
Brenner	2	0	
Mixed CCC/EC	1	1	CCC (1)
Undetermined	3	3*	MUC (1), HGSC (1), LGSC (1)
Total	42	19	

Only cases with multiple sections are included in this analysis. *For each of the 3 undetermined cases, there was at least 1 section for which there was majority consensus subtype classification by the panel of experts. Mixed CCC/EC: CCC with minor EC, UND: Undetermined (no consensus classification), CCC: Clear cell carcinoma, EC: Endometrioid carcinoma, HGSC: High-grade serous carcinoma, LGSC: Low-grade serous carcinoma, MUC: Mucinous carcinoma

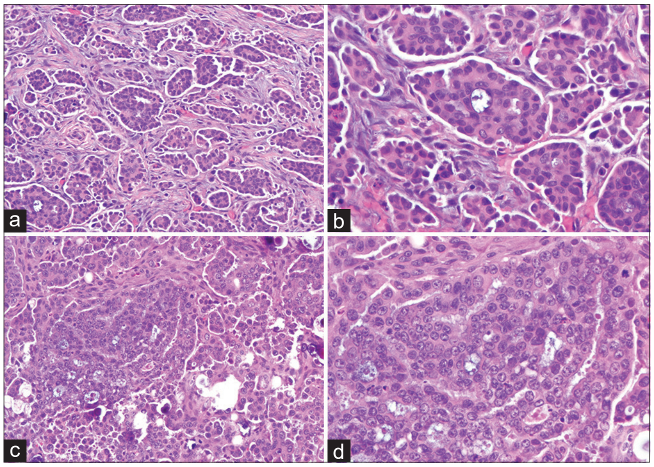

Figure 3: A case for which there was unanimous consensus that the type of differentiation was serous but some lack of agreement regarding low grade versus high grade. (a and b) Papillary serous epithelial fragments in one slide were unanimously interpreted as low-grade in the section-based reads. (c and d) In another slide, focal areas with somewhat more solid growth and moderate nuclear atypia with more evident mitotic activity accounted for a majority consensus interpretation of high grade in the section-based reads. The case-based read (combined assessment of features in a-d) received a majority consensus interpretation of low grade

Diagnostic challenges for ovarian subtype classification identified at the consensus meeting

Among the 21 cases with nonunanimous classification, the most common discrepancies were between MUC versus EC (5 cases), HGSC versus CS (5 cases), HGSC versus LGSC (3 cases), HGSC versus EC (2 cases), and HGSC versus CCC (2 cases). The classifications by each observer, along with brief summary of discrepancy reasons identified during the consensus meeting for each case, are shown in Table 6. The simultaneous review on a multi-head microscope and discussion of the 21 cases with nonunanimous subtype classifications highlighted a variety of diagnostic issues for this task.

One issue was placing different importance on the presence of certain histologic features that were indicative of different subtypes. The identification of morphology, suggestive of endometriosis as a supportive pattern for EC, differed substantially between two reviewers and likely affected diagnosis of EC versus MUC in two cases [Figure 4] and was also a likely factor in the disagreement between classification as EC versus HGSC in another case [Figure 5]. Related to this was the consideration of the rarely reported but well-known evolution of EC to HGSC when discussing the possible presence of morphology, suggestive of endometriosis in HGSC. Loss of intracytoplasmic mucin was observed in two cases, which might have also contributed to a reviewer favoring a diagnosis of EC over MUC. The presence of some overlapping features of clear cell, serous, and endometrioid features also led to some disagreement, regarding classification as CCC versus HGSC [Figure 1].

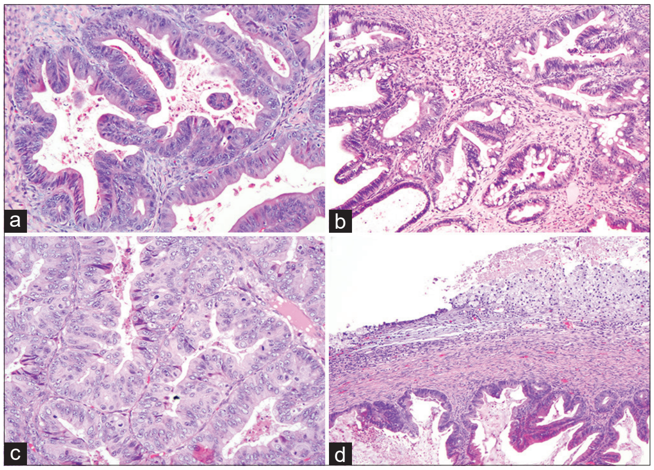

Figure 4: Two cases for which there was lack of unanimous consensus concerning endometrioid versus mucinous differentiation. (a and b) Different slides from one case demonstrate glands for which there was debate regarding assessment as endometrioid versus mucinous (somewhat mucin-depleted), (a) as well as glands with definitive mucinous features (b), leading to some disagreement regarding classification as endometrioid with mucinous features versus pure mucinous differentiation. Case-based and section-based consensus diagnoses were mucinous (2/3 for both). (c) In another example, there was also debate regarding endometrioid versus mucinous features, attributable to the manner in which enlarged atypical nuclei can lead to a mucin-depleted appearance in mucinous glands and impart an endometrioid appearance. Case-based consensus diagnosis was mucinous (2/3), and section-based assessment was unanimously mucinous (3/3). (d) For the first case, a cystic component with a partially denuded epithelial lining and a histiocytic reaction might be interpreted as an endometriotic cyst and used as evidence for interpretation as endometrioid with mucinous differentiation; but given the non-specific appearance of the epithelium and lack of definitive endometrial-type stroma, this might be a nonspecific finding related to gland rupture

A second issue was the use of different thresholds for identifying histologic features or detecting minor foci, with a particular feature that usually suggests a particular cell/tumor type. This was evident for the identification of a sarcomatous component in three cases with differing diagnosis of CS versus HGSC [Figure 6]. Variable detection of small foci of a squamous component also may have contributed to differing diagnoses of EC versus MUC in five cases. This issue may have also had a role in the lack of consensus classification for two undetermined cases. In these cases, the three pathologists viewed the same foci simultaneously but disagreed whether certain features were presented conclusively.

A third issue was the presence of focal transitional features, assessed by some observers as small foci of high-grade atypia in LGSC, which likely contributed to some disagreement in two cases with differing diagnoses of HGSC versus LGSC [Figure 3].

Finally, the effect of technical factors may have played a role in two cases, including one case where the staining quality likely affected the ability to identify intracytoplasmic mucin and a case with frozen section artifact, which may have contributed to the missed identification of a small sarcomatous component.

Table 6: Subtype classifications for 21 discrepant cases based on independent reviews and consensus meeting

Observer #1	Observer #2	Observer #3	Majority classification from independent reviews	Notes on case
MUC	EC	MUC	MUC	Squamous in minor component, proportion of EC/MUC features different in different sections
CCC	HGSC	HGSC	HGSC	High-grade atypia and mitotic count but lacking typical HGSC architectural patterns
EC	MUC	MUC	MUC	Disagreement in importance of squamous or endometriosis presence
MUC	EC	EC	EC	Disagreement in importance of squamous or endometriosis presence
CS	CS	HGSC	CS	Small focus of sarcoma, too small for CS diagnosis
CS	CS	HGSC	CS	Disagreement in threshold for CS
CS	CS	HGSC	CS	Sarcoma foci not identified in consensus review
LGSC	HGSC	LGSC	LGSC	One expert had based diagnosis on one focus of high-grade atypia
SM	SM	EC	SM	Lacking columnar endocervical component
CS	CS	HGSC	CS	Frozen section artifact present, one expert had missed unequivocal sarcomatous focus
EC/MUC	MUC	EC/SM	UND	Disagreement about the presence of squamous differentiation
EC	SM	EC/HGSC	UND	Several patterns supporting several subtypes; disagreement about the presence of squamous component
CCC	HGSC	HGSC	HGSC	Disagreement about clear cell component; high-grade component may be unclassified
EC	HGSC	HGSC	HGSC	Transitional cell-like features
CCC	CCC/EC	CCC/EC	CCC/EC	Agreement about sufficient minor component for mixed classification
CS	CS	HGSC	CS	Consensus review did not confirm presence of sarcoma
EC/HGSC	HGSC	LGSC	UND	Transitional cell-like features
EC/HGSC	HGSC	HGSC	HGSC	Disagreement in importance of presence of endometriosis
SM	SM	LGSC	SM	Intracellular mucin difficult to identify due to staining
LGSC	HGSC	LGSC	LGSC	Small component of HGSC
EC	MUC	MUC	MUC	Disagreement on importance of the presence of endometriosis

/: Mixed subtypes, UND: Undetermined (no subtype could be assigned to case), HGSC: High-grade serous carcinoma, LGSC: Low-grade serous carcinoma, MUC: Mucinous carcinoma, CCC: Clear cell carcinoma, EC: Endometrioid carcinoma, CS: Carcinosarcoma, SM: Seromucinous carcinoma, BR: Brenner

Discussion

Significant focus has been placed recently on the reproducibility of research and clinical findings, particularly in terms of identified biases, regarding the performance of AI/ML algorithms for diagnostic tasks.[31] It is important to identify and control research study parameters that contribute to disparities between expected results from preclinical studies and real-world observations in clinical practice. In this study, we examined an under-examined study design parameter, the selection of representative tumor sections for diagnostic pathology review, in the context of OC histological subtyping.

Differences in the section selection approach could contribute to decreased reproducibility between research studies, as well as between research studies and clinical practice. One reason relates to the spectrum of diagnostic difficulty of cases reviewed within a study. One of the recommendations made by the College of American Pathologists Pathology and Laboratory Quality Center for the validation of WSI was the inclusion of "easy and difficult cases."[32] The number of representative slides needed to represent the histology of a particular case can be related to the diagnostic difficulty for that case. Cases that can be adequately represented by a single section likely present typical histology across all sections. For heterogeneous tumors, a single representative section would likely be chosen as one that best (or more typically) matches the reference diagnosis. A dataset consisting primarily of single-section cases might not reflect the histologic diversity and tumor heterogeneity of cancer cases observed in clinical practice and not provide an adequate sample of challenging, nontypical cases. Observer diagnostic performance or the performance of algorithms for such datasets could be overestimated and not provide a realistic measure of expected performance in clinical practice. Similarly, observer or algorithm performance might vary across datasets with different distributions of the number of representative sections per case. The need for including challenging cases was emphasized in a recent study reporting on the training of pathologists to adopt digital pathology.[33] However, rarely do studies report on the distribution of easy and difficult or typical and challenging cases. In a recently published study,[11] it was reported that agreement between pathology trainees and the reference diagnosis on OC subtype classification determined by a panel of three experts was on an average of 72% for cases where the reference diagnosis was unanimous but dropped to 42% for cases with nonunanimous reference diagnosis. It can be deduced from that study that the

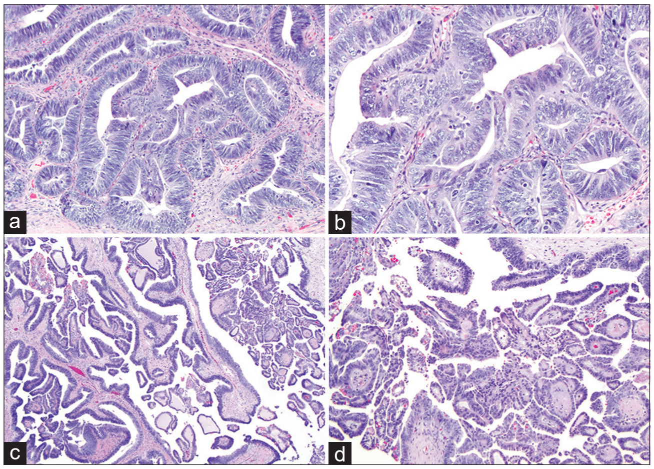

Figure 5: A case for which there was debate concerning whether the tumor was purely high-grade serous or a mixed endometrioid and high-grade serous carcinoma. (a and b) Glands have an endometrioid appearance (a), but there are notable atypia and mitotic activity (b), raising the question of endometrioid versus high-grade serous differentiation. (c and d) Villoglandular and papillary architecture can suggest both endometrioid and serous differentiation. Case-based and section-based consensus diagnoses were high-grade serous (2/3)

cases with nonunanimous agreement were more diagnostically challenging.

A similar observation on the importance of challenging case inclusion was made in this study; average interobserver case-based agreement was 100% for cases that were represented by a single section and was reduced to 64% for cases with multiple sections. The drop in interobserver agreement reflects the histologic difficulty and heterogeneity of the cases represented with multiple sections compared to cases that were likely typical since they could be represented by a single section. The diagnostic difficulty of the cases with multiple sections in our study was also evident by the finding that interobserver concordance when reviewing individual sections from these cases in a randomized order was comparable to concordance for full case review (66.2% vs. 64.3% respectively). Based on the observations from a consensus meeting to discuss discrepant cases, intratumoral heterogeneity was often present at the single section level, in the form of overlapping histologic findings that supported different subtype diagnoses. Pathologists (even the experienced gynecologic pathologists recruited for this study) often disagreed on the interpretation of such features. Experts used different thresholds for identifying certain features or for determining whether the presence of such a feature is important for subtype classification. Such diagnostic challenges for OC subtyping were observed in 18% of cases; not including these challenging cases could substantially overestimate pathologist concordance for this task. Similarly, the performance of an image-based algorithm that was evaluated on a dataset of primarily typical cases could drop significantly when deployed in clinical practice where challenging cases are commonly seen.

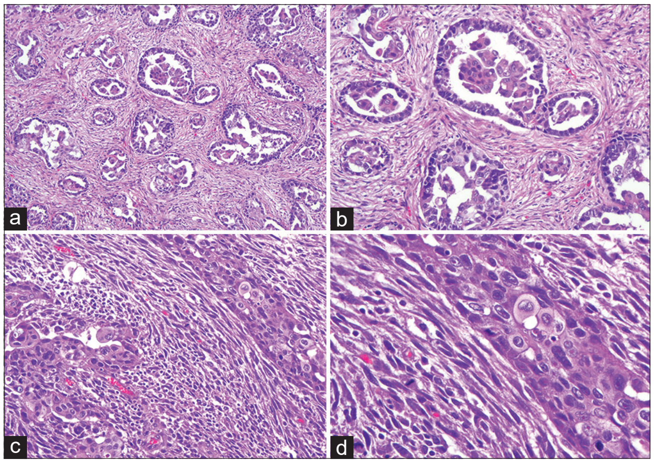

Figure 6: A case for which there was debate primarily concerning carcinosarcoma versus high-grade serous but also some consideration of endometrioid differentiation. (a and b) In one slide containing glandular epithelial elements displaying notable nuclear atypia and surrounding desmoplastic stroma, there was no section-based consensus, as this was assessed as endometrioid by one observer, as carcinosarcoma by the second, and as high-grade serous by the third. (c and d) In another slide, the section-based consensus (2/3) was high-grade serous, with one observer diagnosing carcinosarcoma. These discrepancies can be attributed to the observation that both sections were dominated by glandular elements within desmoplastic stroma as seen in a and b, but small regions with cytologically malignant spindled stroma were present. The case-based consensus (2/3) interpretation was carcinosarcoma, with the minority diagnosis of high-grade serous carcinoma attributable to failure to identify or place less importance on the presence of the small foci with a malignant mesenchymal component

The findings above suggest that studies reporting on observer studies examining primary diagnosis should be accompanied by a measure of the distribution of typical and challenging cases since this parameter could potentially affect the reproducibility of their findings. When examining tissue from resected tumors, reporting the approach for selecting representative sections and the distribution of number of sections per case across a cohort could provide some insight on case difficulty, as was observed in this study. Defining tumor cases as typical or challenging, which would be even more informative, is not trivial; however, some approaches could be to include a percentage of cases for which experienced pathologists were not unanimous in agreement or to include a percentage of cases presenting with overlapping histologic findings which present a number of diagnostic challenges as we have shown.

Another issue related to the selection of representative slides is the generalizability of case diagnosis to the individual sections of that case. Due to the need for a large number of annotated samples for developing and testing of AI/ML algorithms for computer-aided diagnosis and clinical decision support, assumptions are often made that reference diagnoses from patient medical records can be generalized to all sections and corresponding WSIs from that case. In that scenario, whole slide images from all sections of a case are assigned the same label, which is then used for the training (iterative parameter optimization) of a particular AI/ML method. Our study finding that 41% of cases with multiple sections included at least one section for which the majority consensus diagnosis was different from the case‑based diagnosis suggest that assumptions of generalizability may not be warranted, especially in the context of complex, heterogeneous tumors where different tumor regions could support different diagnoses. Algorithms developed and tested on representative sections from such tumors with inaccurate labeling could be biased toward making certain diagnoses over others. Better understanding of the impact of tissue heterogeneity on reference diagnosis could inform developers to account for such effects. Incomplete labeling (reference diagnosis) of such images, which for instance could ignore the presence of a minor component supporting a different subtype, or could be based on an inexperienced reviewer's diagnosis, or even assumptions made by nonpathologist investigators making such selections, could potentially introduce biases hindering the ability of such a method to generalize results of a validation study to an unknown population. The findings of our study support the need to account for tumor heterogeneity for such applications. Approaches such as multiple instance learning have been proposed to address such limitations in histology sample labeling;[34] however, more research is needed to demonstrate their effectiveness.

The finding in our study on the extent of intratumoral heterogeneity in OCs is in line with the current understanding on the heterogeneous nature of OC.[35] As noted by Berman, "tumor heterogeneity poses an enormous challenge to the successful treatment of certain classes of tumors … tumor heterogeneity is equivalent to the existence of multiple tumors within a single tumor mass."[36] Tumor heterogeneity creates logistical and practical challenges in the treatment of cancer patients. Cancer patients frequently travel and visit multiple healthcare providers. Typically, sections of their tumors are examined in multiple centers by multiple pathologists. In addition, clinical trial enrollment is often based on central pathology review. Frequently, only one or a small number of sections are used for these secondary evaluations, and such sections are selected in nonuniform and variable ways depending on which pathologist is available at the time the request is received by the pathology department or laboratory. Our findings highlight the potential clinical problems in using one or a small number of sections for diagnosis of ovarian cancers. A recently published article[37] provided a general perspective to tissue sampling issues in pathology.

A limitation of our study was that it included only experienced gynecologic pathologists in tertiary care centers with large gynecologic oncology divisions. Findings would likely be different with an observer population of community hospital pathologists due to the relatively low incidence of ovarian cancer and limited experience with the different subtypes, especially the less common ones. Another limitation is that our study only used histologic criteria; immunohistochemistry (IHC) was not used to refine this classification. It is established that IHC improves the interobserver reproducibility of OC classification.[2] However, the inclusion of IHC was beyond the scope of this work which focused on examining issues related to the selection of representative hematoxylin and eosin (H&E)‑stained sections in observer studies and the impact of histologic intratumoral heterogeneity. Histologic review of H&E‑stained tissue remains the standard method for primary diagnosis, especially in low‑resource settings where there is limited access to ancillary testing in pathology.[38] Many of the new technologies mentioned in this manuscript have been applied toward image analysis and computer‑aided diagnosis of histologic data. The findings of our study could contribute to improving study designs for the validation of such technologies so that they yield more reproducible and clinically meaningful results that can translate into real‑world practice.

Conclusions

Findings from this study demonstrated substantial differences in interobserver concordance for the histologic subtype diagnosis of OC between cases represented by single slides and cases represented by multiple slides. The approach for selecting sections from tumor cases for histology review in observer studies should be carefully controlled to create balanced datasets in terms of diagnostic difficulty and reduce potential disparities between research and clinical observations. The study also found that the case‑based diagnosis did not generalize to individual sections for 41% of cases with multiple sections. This finding supports the need to account for tumor heterogeneity when determining reference standard in datasets for developing and evaluating diagnostic algorithms.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Elmore JG, Longton GM, Carney PA, Geller BM, Onega T, Tosteson AN, *et al*. Diagnostic concordance among pathologists interpreting breast biopsy specimens. JAMA 2015;313:1122‑32.
2. Köbel M, Bak J, Bertelsen BI, Carpen O, Grove A, Hansen ES, *et al*. Ovarian carcinoma histotype determination is highly reproducible, and is improved through the use of immunohistochemistry. Histopathology 2014;64:1004‑13.
3. Gavrielides MA, Conway C, O'Flaherty N, Gallas BD, Hewitt SM. Observer performance in the use of digital and optical microscopy for the interpretation of tissue‑based biomarkers. Anal Cell Pathol (Amst) 2014;2014:157308.
4. Mukhopadhyay S, Feldman MD, Abels E, Ashfaq R, Beltaifa S, Cacciabeve NG, *et al*. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: A multicenter blinded randomized noninferiority study of 1992 cases (Pivotal Study). Am J Surg Pathol 2018;42:39‑52.
5. Campbell WS, Hinrichs SH, Lele SM, Baker JJ, Lazenby AJ, Talmon GA, *et al*. Whole slide imaging diagnostic concordance with light microscopy for breast needle biopsies. Hum Pathol 2014;45:1713‑21.
6. Campbell WS, Lele SM, West WW, Lazenby AJ, Smith LM, Hinrichs SH. Concordance between whole‑slide imaging and light microscopy for routine surgical pathology. Hum Pathol 2012;43:1739‑44.

7. Villa I, Mathieu MC, Bosq J, Auperin A, Pomerol JF, Lacroix-Triki M, *et al*. Daily biopsy diagnosis in surgical pathology: Concordance between light microscopy and whole-slide imaging in real-life conditions. Am J Clin Pathol 2018;149:344-51.
8. Williams BJ, DaCosta P, Goacher E, Treanor D. A systematic analysis of discordant diagnoses in digital pathology compared with light microscopy. Arch Pathol Lab Med 2017;141:1712-8.
9. Goacher E, Randell R, Williams B, Treanor D. The diagnostic concordance of whole slide imaging and light microscopy: A systematic review. Arch Pathol Lab Med 2017;141:151-61.
10. Snead DR, Tsang YW, Meskiri A, Kimani PK, Crossman R, Rajpoot NM, *et al*. Validation of digital pathology imaging for primary histopathological diagnosis. Histopathology 2016;68:1063-72.
11. Gavrielides MA, Miller M, Hagemann IS, Abdelal H, Alipour Z, Chen JF, *et al*. Clinical decision support for ovarian carcinoma subtype classification: A pilot observer study with pathology trainees. Arch Pathol Lab Med 2020;144 :869-77.
12. Wei JW, Tafe LJ, Linnik YA, Vaickus LJ, Tomita N, Hassanpour S. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. Sci Rep 2019;9:3358.
13. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. J Pathol Inform 2016;7:29.
14. Gavrielides MA, Gallas BD, Lenz P, Badano A, Hewitt SM. Observer variability in the interpretation of HER2/neu immunohistochemical expression with unaided and computer-aided digital microscopy. Arch Pathol Lab Med 2011;135:233-42.
15. Akbar S, Peikari M, Salama S, Panah AY, Nofech-Mozes S, Martel AL. Automated and manual quantification of tumour cellularity in digital slides for tumour burden assessment. Scientific reports 2019;9:1-9.
16. Gecer B, Aksoy S, Mercan E, Shapiro LG, Weaver DL, Elmore JG. Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks. Pattern Recognit 2018;84:345-56.
17. Safrin RE, Bark CJ. Surgical pathology sign-out. Routine review of every case by a second pathologist. Am J Surg Pathol 1993;17:1190-2.
18. Bauer TW, Schoenfield L, Slaw RJ, Yerian L, Sun Z, Henricks WH. Validation of whole slide imaging for primary diagnosis in surgical pathology. Arch Pathol Lab Med 2013;137:518-24.
19. Gage JC, Joste N, Ronnett BM, Stoler M, Hunt WC, Schiffman M, *et al*. A comparison of cervical histopathology variability using whole slide digitized images versus glass slides: Experience with a statewide registry. Hum Pathol 2013;44:2542-8.
20. Loughrey MB, Kelly PJ, Houghton OP, Coleman HG, Houghton JP, Carson A, *et al*. Digital slide viewing for primary reporting in gastrointestinal pathology: A validation study. Virchows Arch 2015;467:137-44.
21. van der Post RS, van der Laak JA, Sturm B, Clarijs R, Schaafsma HE, van Krieken JH, *et al*. The evaluation of colon biopsies using virtual microscopy is reliable. Histopathology 2013;63:114-21.
22. House JC, Henderson-Jackson EB, Johnson JO, Lloyd MC, Dhillon J, Ahmad N, *et al*. Diagnostic digital cytopathology: Are we ready yet? J Pathol Inform 2013;4:28.
23. Vyas NS, Markow M, Prieto-Granada C, Gaudi S, Turner L, Rodriguez-Waitkus P, *et al*. Comparing whole slide digital images versus traditional glass slides in the detection of common microscopic features seen in dermatitis. J Pathol Inform 2016;7:30.
24. Ordi J, Castillo P, Saco A, Del Pino M, Ordi O, Rodríguez-Carunchio L, *et al*. Validation of whole slide imaging in the primary diagnosis of gynaecological pathology in a University Hospital. J Clin Pathol 2015;68:33-9.
25. Hanna MG, Reuter VE, Hameed MR, Tan LK, Chiang S, Sigel C, *et al*. Whole slide imaging equivalency and efficiency study: Experience at a large academic center. Mod Pathol 2019;32:916-28.
26. Köbel M, Kalloger SE, Boyd N, McKinney S, Mehl E, Palmer C, *et al*. Ovarian carcinoma subtypes are different diseases: Implications for biomarker studies. PLoS Med 2008;5:e232.
27. McCluggage WG. Morphological subtypes of ovarian carcinoma: A review with emphasis on new developments and pathogenesis. Pathology 2011;43:420-32.
28. Hoang LN, Zachara S, Soma A, Köbel M, Lee CH, McAlpine JN, *et al*. Diagnosis of ovarian carcinoma histotype based on limited sampling: A prospective study comparing cytology, frozen section, and core biopsies to full pathologic examination. Int J Gynecol Pathol 2015;34:517-27.
29. Seidman JD, Vang R, Ronnett BM, Yemelyanova A, Cosin JA. Distribution and case-fatality ratios by cell-type for ovarian carcinomas: A 22-year series of 562 patients with uniform current histological classification. Gynecol Oncol 2015;136:336-40.
30. Kurman RJ, Carcangiu ML, Young RH, Herrington CS, WHO Classification of Female Reproductive Organs. 4th ed.. Lyon, France: International Agency for Research on Cancer; 2014.
31. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, *et al*. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020;368:m689.
32. Pantanowitz L, Sinard JH, Henricks WH, Fatheree LA, Carter AB, Contis L, *et al*. Validating whole slide imaging for diagnostic purposes in pathology: Guideline from the college of American pathologists pathology and laboratory quality center. Arch Pathol Lab Med 2013;137:1710-22.
33. Williams BJ, Hanby A, Millican-Slater R, Nijhawan A, Verghese E, Treanor D. Digital pathology for the primary diagnosis of breast histopathological specimens: An innovative validation and concordance study on digital pathology validation and training. Histopathology 2018;72:662-71.
34. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, *et al*. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25:1301-9.
35. Blagden SP. Harnessing pandemonium: The clinical implications of tumor heterogeneity in ovarian cancer. Front Oncol 2015;5:149.
36. Berman JJ. Neoplasms: Principles of Development and Diversity. Burlington, Massachusetts: Jones & Bartlett Learning; 2009.
37. McCall SJ, Dry SM. Precision pathology as part of precision medicine: Are we optimizing patients' interests in prioritizing use of limited tissue samples? JCO Precision Oncol 2019;3:1-6.
38. Adeyi OA. Pathology services in developing countries-the West African experience. Arch Pathol Lab Med 2011;135:183-6.