Human Pathology (2013) **44**, 888–894

ELSEVIER

Human PATHOLOGY

www.elsevier.com/locate/humpath

Original contribution

Reliability of whole slide images as a diagnostic modality for renal allograft biopsies

Kuang-Yu Jen MD, PhD [a], Jean L. Olson MD [a], Sergey Brodsky MD [b], Xin J. Zhou MD [c,d], Tibor Nadasdy MD, PhD [b], Zoltan G. Laszik MD, PhD [a,*]

[a] *Department of Pathology, University of California, San Francisco, San Francisco, CA 94143, USA*
[b] *Department of Pathology, The Ohio State University Medical Center, Columbus, OH 43210, USA*
[c] *Renal Path Diagnostics, Pathologists BioMedical Laboratories, Lewisville, TX 75067, USA*
[d] *Department of Pathology, Baylor University Medical Center at Dallas, Dallas, TX 75246, USA*

Received 17 July 2012; revised 20 August 2012; accepted 22 August 2012

Keywords:
Digital whole slide images;
Telepathology;
Renal biopsy;
Kidney;
Allograft;
Transplant;
Diagnostic reliability

Summary The use of digital whole slide images (WSI) in the field of pathology has become feasible for routine diagnostic purposes and has become more prevalent in recent years. This type of technology offers many advantages but must show the same degree of diagnostic reliability as conventional glass slides. Several studies have examined this issue in various settings and indicate that WSI are a reliable method for diagnostic pathology. Since transplant pathology is a highly specialized field that requires not only accurate but rapid diagnostic evaluation of biopsy materials, this field may greatly benefit from the use of WSI. In this study, we assessed the reliability of using WSI compared to conventional glass slides in renal allograft biopsies. We examined morphologic features and diagnostic categories defined by the Banff 07 Classification of Renal Allograft Pathology as well as additional morphologic features not included in this classification scheme. We found that intraobserver scores, when comparing the use of glass slides versus WSI, showed substantial agreement for both morphologic features ($\kappa = 0.68$) and acute rejection diagnostic categories ($\kappa = 0.74$). Furthermore, interobserver reliability was comparable for morphologic features ($\kappa = 0.44$ [glass] vs 0.42 [WSI]) and acute rejection diagnostic categories ($\kappa =$ 0.49 [glass] vs 0.51 [WSI]). These data indicate that WSI are as reliable as glass slides for the evaluation of renal allograft biopsies.

1. Introduction

Traditionally, examination of glass slides under the microscope was the only available modality for pathologists to evaluate tissue for diagnostic purposes. More recently, advances in digital imaging technology, data storage, data compression, and network/internet data transfer have made digitized images of glass slides a realistic option for pathologists. Commonly referred to as "digital" or "virtual" slides, whole slide images (WSI) are created from scans of glass slides and are subsequently viewed on a computer monitor. Interactive software allows the pathologist to navigate the WSI and switch between magnification levels just as with the traditional microscope.

The recent progress in the availability of both hardware and software for the production of WSI has made digital pathology a feasible option in a variety of pathology practice settings including research, education, and routine diagnostics. The

* Corresponding author. Department of Pathology, University of California, San Francisco, San Francisco, CA 94143, USA.
E-mail address: Zoltan.Laszik@ucsfmedctr.org (Z. G. Laszik).

0046-8177/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.humpath.2012.08.015

availability of WSI offers many advantages including rapid evaluation by off-site pathologists, ease of consultation, widespread availability of teaching materials, and simple implementation of multicenter studies. The use of WSI at remote sites away from the primary pathology laboratory has been termed telepathology and may be most helpful in situations where immediate diagnostic feedback is crucial for patient management. Several groups have shown that this method is feasible for applications such as frozen section diagnoses and consultation [1-4].

An area of pathology that would greatly benefit from WSI is transplantation pathology [5]. Similar to frozen section diagnoses, transplant pathology commonly requires immediate diagnostic feedback to the transplant surgeon or clinical specialist in order to initiate timely treatment for possible transplant-threatening or life-threatening diseases. Although the use of WSI for the purpose of diagnostics has quickly emerged as a viable option in many pathology settings [6-10], the data demonstrating the feasibility and reliability of such technology in the field of renal allograft pathology are limited. In this study, we assessed the diagnostic reproducibility and accuracy of evaluating WSI on a computer monitor as compared to conventional glass slides viewed under the microscope for renal allograft biopsies. Our data indicate that these modalities show high concordance, and therefore, WSI represent a reliable method for rendering primary diagnoses in this setting.

2. Materials and methods

Six well-trained renal pathologists from three different institutions evaluated 25 kidney allograft biopsies from the University of California–San Francisco. The biopsies were chosen to represent a wide spectrum of pathologic changes typically seen in renal allograft biopsies. Each case included one hematoxylin/eosin and one periodic acid Schiff–stained 2-μm-thick paraffin section on glass slides, which were scanned using the Aperio ScanScope Scanner at high magnification (40×) to produce WSI (Aperio, Vista, CA). Each pathologist independently evaluated each case twice, on separate occasions at least two weeks apart: once using a microscope to view the glass slides and once using the WSI viewed on a computer monitor using the Aperio ImageScope software. The pathologists were not provided with any clinical information and had no knowledge of the previously reported diagnoses or the scoring of the other participating pathologists. The cases were also scrambled and coded to minimize bias. In order to best reflect realistic multi-institutional settings, no training sets or tutorials were provided for the pathologists prior to examination of the cases.

Twenty-two pre-specified morphologic features, of which 13 consisted of Banff scores based on the Banff 07 Classification of Renal Allograft Pathology, were categorically scored by each pathologist for both the conventional glass slides and the WSI [11,12]. The Banff scores included interstitial inflammation, interstitial inflammation in total parenchyma, tubulitis, interstitial fibrosis, tubular atrophy, peritubular capillaritis, intimal arteritis, vascular intimal thickening, hyaline arteriolar thickening, alternative hyaline arteriolar thickening, glomerulitis, mesangial matrix increase, and allograft glomerulopathy. The morphologic features that are not part of the Banff scoring system (non-Banff features) were scored based on presence or absence of that feature with no predefined thresholds as guidance and included segmental sclerosis, interstitial edema, interstitial hemorrhage, tubulitis within atrophic tubules, acute kidney injury, and isometric vacuolization. The pathologists also recorded the time spent evaluating each case.

Intraobserver reliability was assessed by using linearly weighted Cohen κ to compare scores for conventional glass slides versus WSI for each pathologist. Interobserver reliability between the 6 pathologists was evaluated using 2 methods: (1) average of every combination of pairwise linearly weighted Cohen κ and (2) linearly weighted Fleiss κ. The former (designated mean pair-wise Cohen κ) represents a summary statistic and is difficult to interpret in a statistical manner; however, this method was performed in order to compare our results with prior studies, which have commonly implemented this strategy. On the other hand, the latter method, the Fleiss κ, is a chance-corrected measure of agreement, which is an extension of Cohen κ for evaluation of reliability or agreement between more than two observers [13]. Thus, the Fleiss κ is a more comparable measure to Cohen κ and is essentially equivalent to Cohen κ when the number of observers equals 2 [14]. κ coefficients were interpreted as follows as recommended by Landis and Koch: <0.00, poor; 0.00-0.20, slight; 0.21-0.40, fair; 0.41-0.60, moderate; 0.61-0.80, substantial; >0.80, almost perfect [15]. $P < .05$ were considered significant.

3. Results

The biopsies examined in this study featured a wide spectrum of pathologic changes typically seen in renal allograft biopsies. These cases were chosen based on the original diagnostic pathology report. No reference pathologist was used in this study, and therefore, the 6 participating pathologists' scores were treated equally. The distribution of Banff scores, averaged across all 6 participating pathologists, is shown in Table 1.

3.1. Intraobserver reliability for evaluating Banff scores and non-Banff features of renal allograft biopsies using glass slides versus WSI

Since our goal was to assess whether using WSI is a reliable method for evaluating renal allograft biopsies, we first compared each pathologist's biopsy scores using WSI