Upper Tract Urothelial Carcinoma Grade Prediction Based on the Ureteroscopic Appearance: Caution Should be Taken

Jan Erik Freund, Jaap D. Legemate, Joyce Baard, Kasra Saeb-Parsy, Oliver Wiseman, Steeve Doizi, Esteban Emiliani, Alberto Breda, Bart J. Boedt, Ernst P. van Haarst, Mariska M.G. Leeflang, and Guido M. Kamphuis

OBJECTIVES
To investigate the diagnostic accuracy, inter-rater and intrarater agreement of grade predictions based on the visual appearance of papillary upper tract urothelial carcinoma (UTUC) during digital ureteroscopy.

METHODS
Nine urologists predicted the histopathologic grade of 64 papillary UTUC (low-grade vs high-grade) by assessing the visual appearance of the tumors in videos from digital ureteroscopy. The diagnostic accuracy was estimated by comparing the grade predictions with the histopathology from colocalized biopsies. Inter-rater agreement was assessed by pairwise inter-rater percentage agreement and Fleiss Kappa analysis. The videos were rated in a random order again 30 days after the first assessment to evaluate the intrarater percentage agreement.

RESULTS
Low-grade tumors were predicted correctly in 37%-85% of the cases with a median concordance of 59% for questionnaire 1 and 66% for questionnaire 2. High-grade tumors were predicted correctly in 26%-91% of the cases with a median concordance of 52% and 61% for each questionnaire. The median pairwise inter-rater percentage agreement was 66% for both questionnaires with a Fleiss Kappa of 0.29 and 0.38, respectively. The median intrarater percentage agreement was 81%.

CONCLUSION
The histopathologic grade of UTUC is essential to the risk-stratification for treatment selection. Predictions of the histopathologic grade based on the visual appearance of papillary UTUC with digital ureteroscopy are often incorrect in comparison with biopsy results and yield low inter-rater agreement. Urologists must be aware of these limitations in the assessment of the tumor grade during digital ureteroscopy to warrant good clinical practice.

INTRODUCTION
Histopathologic assessment of urothelial carcinomas enables the subdivision into low-grade and high-grade tumors. The histopathologic grade is a decisive factor for the risk-stratification of upper tract urothelial carcinoma (UTUC), stratifying the disease into low-risk and high-risk groups. Adequate risk-stratification is necessary for treatment selection as endoscopic treatment is generally reserved for low-risk tumors and surgical resection is indicated for high-risk tumors.

At present, treatment selection is generally based on radiologic imaging, cytology, and diagnostic ureteroscopy (URS) together with ureteroscopic biopsies of the suspicious tumor.

Nevertheless, during diagnostic ureteroscopy when the visual appearance of upper tract tumors is perceived as low-grade UTUC, endourologists might be tempted to perform immediate endoscopic treatment after taking tissue biopsies. As tumor biopsies may cause bleeding and can reduce the visibility during ureteroscopy, endourologists might also limit the number of biopsies in case of multiple tumors when similar visual appearances are perceived and grade homogeneity is assumed. Furthermore, in situations without histopathologic certainty, when biopsies are inconclusive (10%-15% of ureteroscopic biopsies), endourologists might be tempted to base clinical decision-making on the perceived tumor grade based on the ureteroscopic appearance of UTUC. Clinical decision-making based on
ureteroscopic grade predictions is not advocated in any clinical guidelines. To our knowledge, studies on the diagnostic limitations of grade predictions based on the visual appearance of UTUC during digital ureteroscopy are lacking. Therefore, the objective of the present study is to investigate the diagnostic accuracy as well as the intra- and inter-rater agreement of grade predictions based on the visual appearance of papillary UTUC with digital ureteroscopy. With the present study, we do not aim to change the current diagnostic paradigm nor advocate clinical decision-making based on ureteroscopic assessment. Instead, we aim to investigate the means of visual UTUC grade assessment during digital ureteroscopy. The results of this study will help us understand the limits of visual assessment during diagnostic ureteroscopy in UTUC.

MATERIALS AND METHODS

Study execution
Nine urologists from the Netherlands, Spain, France, and the United Kingdom predicted the histopathologic grade based on the visual appearance of biopsy-proven UTUC in 64 videos from digital ureteroscopy. The videos were embedded in a questionnaire on an encrypted web-based platform (Data Management System, T&S Innovations, Utrecht, the Netherlands). Thirty days after completing the questionnaire, the urologists were invited for a second questionnaire with the same videos in a random order.

The questionnaires’ interface is illustrated in Figure 1. The tumor location of the depicted UTUC was given to aid the orientation of the raters. All raters were informed that the depicted tumors were histopathologic confirmed papillary urothelial carcinomas. The raters were blinded to tumor grade, the proportion of low-grade and high-grade cases and any other clinical information. Each video sequence could be viewed unlimitedly by each rater. After visual assessment, the raters predicted the tumor grade in a dichotomous fashion (low-grade or high-grade). The raters were asked to score the video quality concerning grade assessability on a 3-point Likert scale (low = 1, moderate = 2, high = 3) for each video before advancing to the next video. Additionally, the urologists reported their experience with cystoscopy and ureteroscopy of urothelial carcinomas.

The local institutional review board granted a waiver for this study as no additional activities in human subjects were involved in this study. Opt-out informed consent was obtained from all patients.

Power calculation
The sample size calculation for the agreement analysis was based on an estimated intra- and inter-rater percentage agreement of 80% for grade predictions. The results of El-Hakim et al, who investigated the diagnostic accuracy of grade predictions based on preoperative clinical information and the visual appearance of UTUC during fiberoptic ureteroscopies, were used to estimate the agreement. The power calculation resulted in an estimated sample size of 62 videos with 80% power and 10% alpha.

Video acquisition
The 64 videos were collected retrospectively from a video-data-base of the Amsterdam University Medical Centers. This video-database was established for internal reviewing. The selected videos were acquired from 62 consecutive ureteroscopic procedures in 44 patients where a colocalized ureteroscopic biopsy of the depicted tumor confirmed urothelial carcinoma and the histopathologic grade.

All ureteroscopic procedures were performed with digital flexible ureteroscopes (52 procedures with Karl Storz Flex XC, 12 procedures with Olympus V2) in white light mode. Recording of the videos was performed with the Medicapture medical USB300 system and the Karl Storz Endoscope TC200 system. For all 64 cases, a high definition video sequence with a duration of 6-10 seconds was created with Apple iMovie v.9.0.9. The video sequences were chosen as such that the biopsied tumors were most optimally visualized from a distance and in detail.

Histopathologic reference standard
The histopathologic grade (WHO 2004) of colocalized biopsies from the depicted tumors was used as the reference standard. The ureteroscopic biopsies were acquired with 1.1 mm 3F Flexible Ureteroscopic Piranha Biopsy Forceps or a ZeroTip 0.63 mm Nitinol Stone Retrieval Basket (both Boston Scientific, Marlborough, MA). Histopathologic assessment was performed by an uropathologist according to the standard clinical protocol. The cohort consisted of 41 (64%) low-grade and 23 (36%) high-grade tumors. The included tumors were located in 18 of 64 (28%) and in 46 of 64 (72%) of the cases in the ureter or in the pyelocalyceal system, respectively.
**RESULTS**

**Experience of the raters**

The experience of the 9 urologists with ureteroscopic UTUC assessment ranged from 3-14 years (mean 8.3 years) with a mean of 38 ureteroscopies for UTUC per year (SD 25.3). For cystoscopic assessment of urothelial carcinoma of the bladder (UCB), their experience ranged from 6-21 years (mean 11.9 years) with a mean of 142 cystoscopies per year (SD 96.7). (For the total number of endoscopic assessments of UC by each rater see Table S1 in supplements.) There was no correlation between the total experience with urothelial carcinoma assessment and overall accuracy of grade predictions (Spearman’s rank correlation coefficient = 0.23, p= 0.56).

**Diagnostic accuracy**

The overall accuracy of grade predictions and the percentages of correctly predicted low-grade and high-grade tumors are illustrated in Figures 2A-C. The median overall accuracy was 59% (IQR 12%) for questionnaire 1 and 64% (IQR 9%) for questionnaire 2. The median percentage of correctly predicted low-grade tumors was 59% (IQR 26%) for questionnaire 1 and 66% (IQR 26%) for questionnaire 2. For high-grade tumors, the median percentage of correct predictions was 52% (IQR 26%) for questionnaire 1 and 61% (IQR 26%) for questionnaire 2.

**Inter-rater agreement**

The pairwise inter-rater percentage agreement for questionnaire 1 ranged from 48% to 83% with a median of 66% (IQR 18%) (see Table S2 in supplements). This median value is below the threshold of acceptable percentage agreement. The Fleiss Kappa was 0.29 (95%CI: 0.28-0.30) for questionnaire 1 and 0.38 (95%CI: 0.37-0.39) for questionnaire 2. Hence, the overall inter-rater agreement was poor for both questionnaires.

**Video quality ratings**

Based on the 3-point Likert scale ratings, the overall mean video quality was 2.2 (SD 0.70) for questionnaire 1, and 2.3 (SD 0.68) for questionnaire 2. The overall mean video quality for cases with a correct grade prediction was 2.3 (SD 0.78). For cases with an incorrect grade prediction, the overall mean video quality was 2.2 (SD 0.76).

**DISCUSSION**

The histopathologic tumor grade is a decisive factor for the risk-stratification of UTUC. As cytology and radiologic imaging may lack diagnostic certitude, ureteroscopy has become an essential tool to increase the diagnostic capabilities for UTUC. Nevertheless, histopathologic grading of ureteroscopic biopsies is not always conclusive. Moreover, biopsies may not always be taken. In light of the results of this study, grade predictions based solely on the visual appearance of UTUC with digital ureteroscopy are too limited to provide diagnostic certainty of the histopathologic grade. Despite acceptable inter-rater agreement, grade predictions were incorrect in one-third to half of all cases with poor inter-rater agreement. Therefore, additional clinical information, such as prior UTUC grade, urine cytology, and radiologic findings should be consulted for clinical decision-making if a re-URS is not wanted. In case of ureteroscopic treatment without histopathologic certainty of the tumor grade, we would advocate a follow-up URS at a short interval (approximately 6 weeks).

To our knowledge, this is the first study to investigate the diagnostic accuracy of grade predictions solely based on the visual appearance of UTUC in ureteroscopic images. A comparison of the diagnostic accuracy of grade predictions can be performed with the study by El-Hakim et al, which, however, had a different study design. In their study, operating room reports were checked retrospectively to investigate the diagnostic accuracy of grade predictions based on preoperative clinical findings together with the visual appearance of UTUC during fiberoptic ureterorenoscopies.
Figure 2. (A) Overall accuracy of grade predictions. (B) Percentage of correctly predicted low-grade tumors. (C) Percentage of correctly predicted high-grade tumors. (Color version available online.)
The extent and variation in additional clinical information available for grade predictions and potential inter-rater variability were not reported.6 Forty cases were included in their final analysis, resulting in an overall accuracy of 70%. The median overall accuracy of our study was slightly lower for both questionnaires (59% and 64%).6 El-Hakim et al. reported a percentage of correctly predicted high-grade tumors of 47%, while we found a median percentage of correct predictions of 52% and 61% for the individual questionnaires. In contrast, the percentage of correctly predicted low-grade tumors by El-Hakim was 87%, while we identified a median percentage of 59% and 66% for the individual questionnaires. It is remarkable that the percentage of correctly predicted high-grade tumors by El-Hakim is lower despite the availability of additional clinical information. On the other hand, the value of additional clinical information may seem of additional value when looking at the difference in correctly predicted low-grade tumors. However, with the limited information at hand, one cannot draw conclusions about the origin of these differences.

Variations in the diagnostic accuracy of tumor grade predictions during cystoscopy of UCB have also been reported. Based on the visual appearance of UCB alone, Liem et al. correctly predicted the grade in 54% of low-grade UCB and in 67% of high-grade UCB.9 In contrast, Cina et al. correctly identified 91% of all low-grade UCB and Herr et al. correctly identified 93% of all grade 1 UCB.10,11 Yet in the latter studies, additional clinical information was available. Furthermore, the absence of intermediate grade 2 tumors in the study by Herr et al. may have led to more distinctive visual differences between low-grade and high-grade tumors.11

In our study, the inter-rater agreement was poor and hence highlighting the limitations of grade predictions based on the ureteroscopic appearance of the tumor.7,8 In contrast, Herr et al. observed an excellent agreement between raters for grade predictions of UCB.11 This difference in inter-rater agreement may also be explained by the availability of additional clinical information in their study. One may also argue that the difference in inter-rater agreement could arise from differences in personal experience of the raters. In our study, however, the urologists’ experience with assessment of urothelial carcinoma was not correlated with the overall accuracy of UTUC grade predictions.

The quality of the videos was assessed to evaluate if the video quality was a confounder for grade predictions. Overall, the video quality concerning the assessability of the tumor grade was rated as moderate with a tendency towards high quality. The video quality of the correctly predicted tumors was very similar to the image quality of the falsely predicted tumors. Therefore, it seems that the video quality did not limit nor influence grade prediction.

The outcome of this study shows that the visual appearance of UTUC during digital ureteroscopy alone is insufficient for accurate grade prediction. As a result, it seems disputable whether the histopathologic grade is distinctly characterized by the phenotypes of low-grade and high-grade UTUC. Nonetheless, future studies should investigate if widely available image enhancement techniques, such as Narrow Band Imaging or Image1S, may aid the assessment of the visual appearance. Moreover, convolutional neural networks and other image recognition software may be able to identify image patterns of UTUC that allow for accurate grade prediction with digital ureteroscopy.

Limitations
The use of ureteroscopic biopsies as the reference standard is a point of discussion as biopsies may underestimate the histopathologic grade in comparison to resection specimens.4,11 Therefore, one may argue that histopathology from surgical specimen is a superior reference standard. Yet, the use of resection specimens for comparison may introduce selection bias due to treatment selection based on the histopathologic grade. Moreover, with regard to tumor grade heterogeneity between spatially separated unilateral tumors and tumor progression, colocalized biopsies may be more suitable as a reference standard for the assessment of focal regions of interest than histopathology from representative sampling of resection specimens.13,14 In any case, the limitations of the reference standards may confine the accuracy of comparison.

To achieve the aim of this study, the raters were blinded to additional clinical information. Therefore, the findings of this study can be considered as the baseline diagnostic accuracy for grade predictions based on the visual appearance of UTUC. The additive diagnostic value of additional clinical information for grade predictions remains to be determined. Moreover, the assessment of the tumor architecture was not considered in this study, which could influence the grade prediction of small, almost sessile, papillary tumors.

Despite careful selection of the video sequences for optimal visualization of the tumors in close-up and from a distance with the unlimited play-back time for each rater, the restriction in video length may have limited the visualization of the tumors. More extensive videos from more angles and visualization of the complete upper tract might aid grade prediction.

CONCLUSION
The histologic grade of UTUC is a decisive factor for the risk-stratification of the disease. The visual appearance of papillary UTUC with digital ureteroscopes, however, does not allow for accurate predictions of the histopathologic grade. The diagnostic accuracy of grade predictions is limited and of low inter-rater agreement. Urologists must be aware of these limitations in the ureteroscopic assessment of UTUC to warrant good clinical practice.

CONFLICT OF INTEREST
J.E. Freund declares no conflicts of interests. J.D. Legemate declares no conflicts of interests. J. Baard declares no conflicts of interests. K. Saeb-Parsy declares no conflicts of interests. O. Wiseman is a consultant for Boston Scientific, Coloplast, EMS, and received a research grant from Coloplast. S. Doizi is a consultant for Coloplast. E. Emiliani declares no conflicts of

Wiseman is a consultant for BostonScientific, Coloplast, EMS, and received a research grant from Coloplast. S. Doizi is a consultant for Coloplast. E. Emiliani declares no conflicts of

Wiseman is a consultant for BostonScientific, Coloplast, EMS, and received a research grant from Coloplast. S. Doizi is a consultant for Coloplast. E. Emiliani declares no conflicts of
interests. A. Breda is a consultant for Cook, Karl Storz and Intuitive. B.J. Boodt declares no conflicts of interests. E.P. van Haast declares no conflicts of interests. M.M.G. Leeflang declares no conflicts of interests. G.M. Kamphuis is a consultant for Boston Scientific and Olympus.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found in the online version at https://doi.org/10.1016/j.urology.2019.05.046.

References


EDITORIAL COMMENT

Until recently, risk stratification for upper tract urothelial carcinoma (UTUC) was sparsely performed, often based on only findings of a filling defect or ureteroscopic visual appearance. In this study, the authors sought to define the accuracy of UTUC grade prediction based solely on the tumor visual appearance using digital ureteroscopy. While visual appearance of UTUC is not a defined factor for clinical decision-making, there are scenarios where presumptions of tumor risk could be made based on the appearance with little data to support such assessment. Nine expert urologists at multiple European centers graded 64 biopsy-proven papillary tumors depicted on videos of the ureteroscopy. Sessile tumors were excluded. Visual low- versus high-grade assessment was compared to the biopsy grade. Inter-reader and intrareader variability was reported, with the latter assessed by repeating the exercise 30 days after the initial viewing. The median percentage of correctly predicting low and high-grade tumors was 59% and 52%, respectively. These numbers increased slightly at the second viewing (66% for low and 61% for high-grade tumors). These findings should not be a surprise. Urologists are not good at accurately predicting grade and stage of bladder tumors based on cystoscopic impression.1,2 1,2 The current study does not report the pathologic grade from the subset of patients undergoing nephroureterectomy. The gold standard in this study is somewhat flawed and perhaps, a different correlation exists between ureteroscopic impression and pathologic grade.

Nevertheless, this study shows that grade determination by visual inspection alone ought not to be used in clinical decision-making. For that matter, any unimodal method is likely to have major pitfalls. Contemporary practice should now be more sophisticated, requiring evaluation of imaging, ureteroscopic architecture, biopsy, and sometimes selective cytology results. A new clinical nomogram even requires hemoglobin, an independent predictor, for accurately stratifying patients. Modern practice thus requires not a highly fallible staging, visual inspection, or reliance on any one single finding, but multimodal risk stratification using either the European consensus guidelines 1 or a recent validated clinical nomogram.2

Lambros Stamatakis, Suren F. Matin, Department of Urology, Georgetown University School of Medicine, Washington DC; Department of Urology, MD Anderson Cancer Center, Houston, TX

References