Guidelines


Viktor Soukup a,*, Otakar Čapoun a,*, Daniel Cohen b, Virginia Hernández c, Marek Babjuk d, Max Burger e, Eva Compérat f, Paolo Gontero g, Thomas Lam h,i, Steven MacLennan h, A. Hugh Mostafid j, Joan Palou k, Bas W.G. van Rhijn l, Morgan Rouprêt m, Shahrokh F. Shariat n, Richard Sylvester o, Yuhong Yuan p, Richard Zigeuner q

a Department of Urology, General Teaching Hospital and 1st Faculty of Medicine, Charles University in Praha, Praha, Czech Republic; b Department of Urology, Royal Free London NHS Foundation Trust, London, UK; c Department of Urology, Hospital Universitario Fundación de Alcorcón, Madrid, Spain; d Hospital Motol and Second Faculty of Medicine, Charles University, Department of Urology, Prague, Czech Republic; e Department of Urology, Caritas-St. Josef Medical Center, University of Regensburg, Germany; f Department of Pathology, Hôpitaux Universitaires de l’Est-Parisien HUEP, Assistance Publique Faculty of Medicine Pierre et Marie Curie, Institut Universitaire de Cancérologie GRCs, University Paris 6, Paris, France; g Department of Surgical Sciences, Urology, University of Turin, Turin, Italy; h Academic Urology Unit, University of Aberdeen, Scotland, UK; i Department of Urology, Aberdeen Royal, 12 Infirmary, Aberdeen, Scotland; j Department of Urology, Royal Surrey County Hospital, Guildford, UK; k Department of Urology, Fundación Puigvert, Universitat Autònoma de Barcelona, Barcelona, Spain; l Department of Urology, Netherlands Cancer Institute—Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands; m Department of Urology, Groupe Hospitalier Pitié–Salpêtrière, Assistance Publique Hopitaux de Paris, Faculty of Medicine Pierre et Marie Curie, Institut Universitaire de Cancérologie GRCs, University Paris 6, Paris, France; n Department of Urology, Vienna General Hospital, Medical University of Vienna, Vienna, Austria; o EAU Guidelines Office Board, European Association of Urology, The Netherlands; p Department of Medicine, Health Science Centre, McMaster University, Hamilton, Ontario, Canada; q Department of Urology, Medizinische Universität Graz, Graz, Austria

Article info

Article history:
Accepted April 12, 2017

Associate Editor:
James Catto

Keywords:
Non–muscle-invasive bladder cancer
Grade
1973 World Health Organization classification

Abstract

Context: Tumour grade is an important prognostic indicator in non–muscle-invasive bladder cancer (NMIBC). Histopathological classifications are limited by interobserver variability (reproducibility), which may have prognostic implications. European Association of Urology NMIBC guidelines suggest concurrent use of both 1973 and 2004/2016 World Health Organization (WHO) classifications.

Objective: To compare the prognostic performance and reproducibility of the 1973 and 2004/2016 WHO grading systems for NMIBC.

Evidence acquisition: A systematic literature search was undertaken incorporating Medline, Embase, and the Cochrane Library. Studies were critically appraised for risk of bias (QUIPS). For prognosis, the primary outcome was progression to muscle-invasive or metastatic disease. Secondary outcomes were disease recurrence, and overall and cancer-specific survival. For reproducibility, the primary outcome was interobserver variability between pathologists. Secondary outcome was intraobserver variability (repeatability) by the same pathologist.

* Corresponding authors. Department of Urology, General Teaching Hospital and 1st Faculty of Medicine, Charles University in Praha, Praha, Czech Republic. Tel. +420 224967847; Fax: +420 224967102. E-mail addresses: viktor.soukup@seznam.cz (V. Soukup), otakar.capoun@seznam.cz (O. Čapoun).
Evidence synthesis: Of 3593 articles identified, 20 were included in the prognostic review; three were eligible for the reproducibility review. Increasing tumour grade in both classifications was associated with higher disease progression and recurrence rates. Progression rates in grade 1 patients were similar to those in low-grade patients; progression rates in grade 3 patients were higher than those in high-grade patients. Survival data were limited. Reproducibility of the 2004/2016 system was marginally better than that of the 1973 system. Two studies on repeatability showed conflicting results. Most studies had a moderate to high risk of bias.

Conclusions: Current grading classifications in NMIBC are suboptimal. The 1973 system identifies more aggressive tumours. Intra- and interobserver variability was slightly less in the 2004/2016 classification. We could not confirm that the 2004/2016 classification outperforms the 1973 classification in prediction of recurrence and progression.

Patient summary: This article summarises the utility of two different grading systems for non–muscle-invasive bladder cancer. Both systems predict progression and recurrence, although pathologists vary in their reporting; suggestions for further improvements are made.

© 2017 European Association of Urology. Published by Elsevier B.V. All rights reserved.
2. Evidence acquisition

2.1. Search strategy

Protocols for both the prognostic and reproducibility reviews have been published (http://www.crd.york.ac.uk/PROSPERO; registration numbers CRD420150250045 and CRD42016029714); the search strategy is outlined in the Supplementary material.

Databases including Medline, Embase, and the Cochrane Central Register of Controlled Trials were systematically searched from 1 January 1998 to 31 December 2015. All abstracts and full-text articles were independently screened by at least two reviewers. Disagreement was resolved by discussion with an independent arbiter. The search was complemented by additional sources including the reference lists of included studies and a panel of experts (EAU NMIBC Panel).

2.2. Types of study designs

Prospective and retrospective studies comparing the two grading systems were included. Only studies published from 1998 onwards were included. There were no language restrictions. A minimum follow-up of 3 mo (recurrence and/or progression) was required for inclusion in the prognostic review. Reproducibility assessment by two or more pathologists required use of identical specimens and grading systems. For the assessment of the repeatability of a grading system by the same pathologist, each pathologist or group of pathologists had to assess identical specimens using the same grading system at more than one time point.

2.3. Types of participants

Study inclusion criteria were as follows: adult patients (>18 yr old) with primary or recurrent Ta/T1 urothelial carcinoma of the bladder who underwent a transurethral resection of bladder tumour (TURBT). All risk groups and adjuvant treatments were included. Exclusion criteria were as follows: patients under 18 yr; muscle-invasive bladder cancer (MIBC); clinical N+ or M+; grading based on radical cystectomy specimen; and bladder biopsies only (as opposed to TURBT). The protocol allowed inclusion of studies with exclusion criteria if affected patients constituted <10% of the study population.

2.4. Type of outcome measures

In the prognostic review, the primary outcome was progression to muscle-invasive or metastatic stage. Secondary outcomes were bladder recurrence, and overall and cancer-specific survival. All outcomes were measured at least 3 mo post-TURBT.

In the reproducibility review, the primary outcome was interobserver variability (reproducibility) between pathologists. The secondary outcome was intraobserver variability (repeatability) by the same pathologist and reliability (variability due to heterogeneity of patient populations).

2.5. Assessment of risk of bias

As recommended by the Cochrane Prognosis Methods Group, the risk of bias (RoB) in the included studies was assessed using the QUIPS tool across six domains: study participation, attrition, prognostic factor measurement, outcome measurement, confounders, and statistical analysis [11]. The EAU NMIBC Guidelines Panel identified intravesical BCG (yes/no), stage (Ta/T1), and concomitant carcinoma in situ (CIS) (yes/no) as three most important prognostic confounders. The Cochrane Collaboration recommends not to combine domains or give overall summary scores [12]. We used Revman 5.3 software to generate graphs showing RoB for each domain, within and across studies.

2.6. Data extraction and analysis

In the prognostic review, outcome events along with all unadjusted (univariate) and adjusted (multivariable) measures of association, such as odds ratios and hazard ratios, were extracted, including those in subgroups of interest.

In the reproducibility review, all outcomes of reproducibility, repeatability, and reliability, both overall and in subgroups of interest, were extracted. Assessment of concordance was evaluated using Cohen’s kappa statistic (coefficient $\kappa$). Arbitary guidelines characterise values of kappa $>0.75$ as excellent concordance, $0.40–0.75$ as fair to good, and below $0.40$ as poor [13].

3. Evidence synthesis

3.1. Quantity of evidence identified

The study selection process is outlined in the Preferred Reporting Items for Systematic Reviews and Meta-analysis (PRISMA) flow diagram (Fig. 2). A total of 3593 abstracts were reviewed for both prognostic performance and reproducibility, of which 34 full texts were retrieved for further screening. Ultimately, 22 eligible studies were identified; however, two studies [14,15] were excluded as subsequent publications provided updated data [16,17]. Finally, 20 studies recruiting a total of 4505 patients met the inclusion criteria for prognostic performance [3,16–34]. Three of these studies involving 566 patients met the reproducibility inclusion criteria [3,16,33].

3.2. Characteristics of the 20 included studies

The baseline characteristics of studies included in the prognostic review are detailed in Table 1. The three retrospective studies contained information on reproducibility or repeatability: Mangrud et al [16]—three pathologists independently reviewed both classifications and two pathologists repeated the classification for intraobserver variability; however, only one pathologist assessed both grading systems. Van Rhijn et al [3]—two pathologists (A + D) reviewed both classifications on four separate occasions (both systems twice), allowing a direct comparison of the two
grading systems. In addition, four pathologists (A + B + C + D) reviewed the slides for the 2004/2016 WHO classification on two separate occasions. May et al [33] reported reproducibility of both grading systems between four independent pathologists (Table 1).

3.3. RoB and confounding assessment of the included studies

Figure 3 presents the RoB summary for the 20 included trials [3,16–34]. We found the highest RoB in study attrition (incomplete outcome data), study confounders (validity, reliability, and similarity of measurement), and study participation (representativeness of the study sample) [10]. The risk of reporting bias (selective reporting) was high in less than one-third of studies. The risks of bias in prognostic factor (tumour grade) measurement and outcome measurement (adequacy of outcome measurement) were low.

For the three most important prognostic confounders, tumour stage was well described, but presence of CIS and use of adjuvant treatment were incompletely reported (Table 1). Therefore, it was difficult to factor these last two confounders into the analyses. Some subgroup analyses were performed in Ta and T1 patients (Tables 2 and 3).

3.4. Comparisons of prognostic outcome measures

For analysis of progression, recurrence, and overall and cancer specific survival, most available information concerned the number of patients with an event during...
Table 1 – Baseline study characteristics for the 20 comparative studies with 4505 patients

<table>
<thead>
<tr>
<th>First author (year)</th>
<th>Study start-end</th>
<th>Follow-up (median; mo)</th>
<th>Uropathologist</th>
<th>Patients included</th>
<th>Patients excluded</th>
<th>Age (mean)</th>
<th>Males/ females</th>
<th>T category</th>
<th>CIS</th>
<th>Intravesical chemotherapy</th>
<th>BCG</th>
<th>G1</th>
<th>G2</th>
<th>G3</th>
<th>PUNLMP</th>
<th>LG</th>
<th>HG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mangrud (2014) [16]</td>
<td>2002–2006</td>
<td>75.0</td>
<td>No</td>
<td>193</td>
<td>56</td>
<td>148/45</td>
<td>Ta and T1</td>
<td>22</td>
<td>193</td>
<td>44</td>
<td>98</td>
<td>51</td>
<td>0</td>
<td>119</td>
<td>74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gontero (2014) [18]</td>
<td>1992–2006</td>
<td>71.6</td>
<td>Yes</td>
<td>131</td>
<td>60</td>
<td>66.3</td>
<td>112/19</td>
<td>5</td>
<td>Only Ta</td>
<td>65</td>
<td>65</td>
<td>0</td>
<td>105</td>
<td>26</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pellucchi (2011) [21]</td>
<td>2004–2008</td>
<td>25.0</td>
<td>Yes</td>
<td>270</td>
<td>162</td>
<td>71</td>
<td>239/71</td>
<td>Only T1</td>
<td>71</td>
<td>266</td>
<td>0</td>
<td>124</td>
<td>142</td>
<td>0</td>
<td>266</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pellucchi (2015) [22]</td>
<td>2004–2011</td>
<td>19.0</td>
<td>Yes</td>
<td>266</td>
<td>412</td>
<td>67.6</td>
<td>237/29</td>
<td>Only T1</td>
<td>71</td>
<td>266</td>
<td>0</td>
<td>124</td>
<td>142</td>
<td>0</td>
<td>266</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Otto (2011) [23]</td>
<td>1989–2006</td>
<td>49.0</td>
<td>Yes</td>
<td>310</td>
<td>39</td>
<td>71.7</td>
<td>239/71</td>
<td>Only T1</td>
<td>252</td>
<td>0</td>
<td>112</td>
<td>198</td>
<td>0</td>
<td>13</td>
<td>297</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Van Rhijn (2010) [25]</td>
<td>1983–2006</td>
<td>68.0</td>
<td>Yes</td>
<td>164</td>
<td>-</td>
<td>68.6</td>
<td>135/29</td>
<td>Ta and T1</td>
<td>26</td>
<td>164</td>
<td>0</td>
<td>74</td>
<td>90</td>
<td>0</td>
<td>127</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burger (2008) [27]</td>
<td>–</td>
<td>221</td>
<td>Yes</td>
<td>171</td>
<td>50</td>
<td>Ta and T1</td>
<td>15</td>
<td>86</td>
<td>110</td>
<td>49</td>
<td>119</td>
<td>50</td>
<td>0</td>
<td>77</td>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oosterhuis (2002) [31]</td>
<td>1979–2000</td>
<td>63</td>
<td>No</td>
<td>320</td>
<td>39</td>
<td>66.6</td>
<td>295/64</td>
<td>Only Ta</td>
<td>28</td>
<td>8</td>
<td>31</td>
<td>286</td>
<td>1</td>
<td>116</td>
<td>141</td>
<td></td>
<td></td>
</tr>
<tr>
<td>May (2010) [33]</td>
<td>1997–2004</td>
<td>Yes</td>
<td>200</td>
<td>68.6</td>
<td>149/51</td>
<td>Only Ta</td>
<td>0</td>
<td>82</td>
<td>109</td>
<td>9</td>
<td>1</td>
<td>149</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>van Rhijn (2015) [34]</td>
<td>1986–2006</td>
<td>Yes</td>
<td>325</td>
<td>66.4</td>
<td>254/71</td>
<td>Ta and T1</td>
<td>62</td>
<td>98</td>
<td>225</td>
<td>88</td>
<td>149</td>
<td>88</td>
<td>79</td>
<td>101</td>
<td>145</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CIS = carcinoma in situ; G1 = grade 1; G2 = grade 2; G3 = grade 3; HG = high grade; LG = low grade; PUNLMP = papillary urothelial neoplasm with low malignant potential.

* Studies included in the reproducibility part.
follow-up and the percentage of patients with an event at a given point in time. There were little time-to-event data, that is, time to recurrence, hazard ratios, p values, and multivariable adjustments. The main analysis is thus based on a comparison of the overall percentage of patients with an event during follow-up. The data from each study was combined to obtain an overall estimate and compared using a Pearson chi-square test. This was not possible for the percentage of patients with an event at a given point in time.

While it was possible to independently compare the outcomes for the categories within each of the two grading classifications, 1973 (G1 vs G2 vs G3) and 2004/2016 (PUNLMP vs LG vs HG), not all the studies provided end point information for each grading classification. In order to minimise the RoB when comparing 1973 with 2004/2016, the most reliable results were obtained when analysing only the studies that assessed both grading classifications. Thus, each of the two grading classifications is assessed on the same set of patients so that there are no differences between the two classifications concerning patient follow-up, characteristics, or treatment. Sensitivity analyses were carried out using all available information for each grading classification.

3.4.1. Prognostic outcomes
3.4.1.1. Progression. Overall, 13 studies provided data on progression. In six studies, progression was defined as any increase in disease stage, including Ta–T1, while in seven studies, it was defined as an increase to stage T2 or greater. In two studies [18,32] where data for both definitions were available, information on an increase to T2 or greater was used.

3.4.1.1.1. Progression defined as muscle-invasive or metastatic disease. Comparisons only from studies that assessed both the 1973 and 2004/2016 classifications. A direct comparison of the two grading systems demonstrated progression by 1973 grade (G1 vs G2 vs G3) in 3% versus 9% versus 32%, whereas for 2004/2016 grade (PUNLMP vs LG vs HG), 1% versus 4% versus 25% progressed, respectively (Table 2).

A separate subgroup analysis of HG T1 disease showed a higher progression rate in G3 versus G2: 28% versus 12%.

3.4.1.1.2. Comparisons using all available data. The overall percentage of patients with progression varied between grades within each classification; for the 1973 grade (G1 vs G2 vs G3), 3% versus 10% versus 29% progressed, respectively; for the 2004/2016 grade (PUNLMP vs LG vs HG), 1% versus 4% versus 19% progressed, respectively (Table 2).

3.4.1.2. Progression defined as any increase in disease stage
3.4.1.2.1. Comparisons only from studies that assessed both 1973 and 2004/2016 classifications. When defining progression as any stage increase, including Ta to T1, progression was observed in (G1 vs G2 vs G3) 3% versus 8% versus 27% and (PUNLMP vs LG vs HG) 2% versus 4% versus 22%, respectively (Table 2).

In LG Ta patients, we found a higher progression rate in G2 patients as compared with G1 patients: 7% versus 1%.

3.4.1.2.2. Comparison using all available data. Progression rates were (G1 vs G2 vs G3) 3% versus 9% versus 28%, and (PUNLMP vs LG vs HG) 2% versus 4% versus 19%, respectively.

3.4.1.2. Recurrence. Eight studies provided information on the number of patients with recurrence, but only five used both grading systems (Table 3).
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>T1 patients only</td>
<td>T2 or greater increase in stage</td>
<td>Pellucchi (2015) [22], Kamel (2006) [29]</td>
<td>371</td>
<td>–</td>
<td>12.4</td>
<td>28.0</td>
<td>0.000</td>
<td>Pellucchi (2015) [22], Kamel (2006) [29]</td>
<td>371</td>
<td>–</td>
<td>–</td>
<td>20.2</td>
</tr>
<tr>
<td>G1 vs G2 in Ta LG tumours</td>
<td>Any increase in stage</td>
<td>Pellucchi (2011) [21]</td>
<td>270</td>
<td>1.2</td>
<td>7.1</td>
<td>–</td>
<td>0.039</td>
<td>Pellucchi (2011) [21]</td>
<td>270</td>
<td>–</td>
<td>5.2</td>
<td>–</td>
</tr>
</tbody>
</table>

G1 = grade 1; G2 = grade 2; G3 = grade 3; HG = high grade; LG = low grade; PUNLMP = papillary urothelial neoplasm with low malignant potential.
Table 3 - Distribution of the percentage of patients with tumour recurrence

<table>
<thead>
<tr>
<th>Type of analysis</th>
<th>1973 grade—studies included</th>
<th>1973 grade—number of patients</th>
<th>1973 grade—percent of G1 patients with recurrence</th>
<th>1973 grade—percent of G2 patients with recurrence</th>
<th>1973 grade—percent of G3 patients with recurrence</th>
<th>Pearson x² test p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ta patients only</td>
<td>Mangrud (2014) [16], Chen (2012) [20], Burger (2008) [26], Yin (2004) [30], May (2010) [33]</td>
<td>931</td>
<td>32.6</td>
<td>42.3</td>
<td>62.6</td>
<td>0.000</td>
</tr>
<tr>
<td>Ta patients only</td>
<td>Mangrud (2014) [16], Chen (2012) [20], Burg (2008) [26], Yin (2004) [30], May (2010) [33]</td>
<td>934</td>
<td>20.3</td>
<td>38.0</td>
<td>54.7</td>
<td>0.000</td>
</tr>
<tr>
<td>All studies with recurrence data</td>
<td>Mangrud (2014) [16], Chen (2012) [20], Pellucchi (2015) [22], Burger (2008) [26], Yin (2004) [30], May (2010) [33], Oosterhuis (2002) [31], Holmäng (2001) [17], May (2010) [33]</td>
<td>1197</td>
<td>32.6</td>
<td>43.9</td>
<td>65.4</td>
<td>0.000</td>
</tr>
<tr>
<td>Ta1 patients only</td>
<td>Burger (2008) [26], Yin (2004) [30], May (2010) [33]</td>
<td>390</td>
<td>39.0</td>
<td>40.8</td>
<td>70.6</td>
<td>0.040</td>
</tr>
<tr>
<td>Ta2 patients only</td>
<td>Pellucchi (2015) [22]</td>
<td>988</td>
<td>28.3</td>
<td>52.0</td>
<td>60.5</td>
<td>0.000</td>
</tr>
</tbody>
</table>

G1 = grade 1; G2 = grade 2; G3 = grade 3; HG = high grade; LG = low grade; PUNLMP = papillary urothelial neoplasm with low malignant potential.

3.4.1.2.1. Comparison of five studies that utilised both 1973 and 2004/2016 classifications. The pooled recurrence rates were (G1 vs G2 vs G3) 33% versus 42% versus 63% and (PUNLMP vs LG vs HG) 20% versus 38% versus 55%, respectively (Table 3).

The majority of patients in these five studies had Ta disease; a separate analysis in T1 patients was not possible [16,20,26,30,33]. A subgroup analysis of T1 HG patients revealed a higher recurrence rate in G3 patients compared with G2 patients (68% vs 50%) [22].

3.4.1.2.2. Comparisons using all available data. The percentage of patients with recurrence using the 1973 grade (G1 vs G2 vs G3) was 33% versus 44% versus 65%, respectively. For the 2004/2016 grade (PUNLMP vs LG vs HG), recurrence occurred in 28% versus 43% versus 58%, respectively (Table 3).

Separate analysis of Ta patients revealed higher recurrence rates in G3 disease (G1 vs G2 vs G3): 39% versus 41% versus 71%, respectively; In Ta patients, PUNLMP patients have lower recurrence rates than LG or HG patients: 28% versus 52% versus 60%, respectively. No comparisons were possible in T1 patients (Table 3).

3.4.1.3. Death due to bladder cancer. Only one study provided limited information regarding death due to bladder cancer, so no conclusions could be drawn [29].

3.4.1.4. Death due to any cause. Information on all-cause mortality was available on a limited basis in two studies [18,28], and only one study contributed to the analysis [31]. In this study, death rates for patients with the best and worst prognosis seem to be similar in the two grading classifications, but no conclusions can be drawn.

3.4.2. Reproducibility and repeatability outcomes

3.4.2.1. Reproducibility. The interobserver agreement and kappa values for the 1973 and 2004/2016 WHO classifications are presented in Table 4.

The interobserver agreement for the 1973 classification ranged from 38% to 89% (kappa values from 0.003 to 0.68). Agreement in combined assessment of G1 + G2 versus G3 tumours in two studies [3,16] was higher than in separate assessment of G1 versus G2 versus G3 tumours (80–89% vs 39–66%; kappa values 0.44–0.68 vs 0.15–0.68). The interobserver agreement for the 2004/2016 classification ranged from 43% to 100% (kappa values 0.17–0.70). Only one study assessed the agreement between two pathologists in a combined review of PUNLMP + LG versus HG tumours [3]. It showed slightly better reproducibility than for a separate analysis of PUNLMP versus LG versus HG tumours (73–86% vs 43–66%, kappa values 0.46–0.72 vs 0.17–0.48). In this study, two additional pathologists assessed slides according to 2004/2016 WHO classification only. The interobserver agreement for the separate review of PUNLMP versus LG versus HG tumours between these two pathologists and with the latter two pathologists ranged from 38% to 74% (kappa values from 0.13 to 0.58) and for combined review of PUNLMP + LG versus HG tumours ranged from 65% to 88% (kappa values from 0.30 to 0.73).

3.4.2.2. Repeatability. The intraobserver repeatability and kappa values for the 1973 and 2004/2016 WHO classifications are presented in Table 5. Only two studies assessed the repeatability of both grading systems [3,15]. The intraobserver agreement for the 1973 WHO grading classification ranged from 63% to 95% (kappa values 0.61–0.88). Repeatability for combined assessment of G1 + G2 versus G3 tumours was slightly higher than that for a separate analysis of G1 versus G2 versus G3 tumours (88–95% vs 63–81%, kappa values 0.64–0.88 vs 0.61–0.69). The intraobserver agreement for the 2004/2016 WHO grading classification ranged from 71% to 93% (kappa values 0.56–0.83). In the only study that assessed the difference between combined and separate pathologist reviews, repeatability of group PUNLMP + LG versus HG was higher than that of PUNLMP versus LG versus HG (86–90% vs 71–82%, kappa values 0.68–0.80 vs 0.56–0.69) [3]. In this study, two additional pathologists assessed slides twice using the 2004/2016 WHO classification with 72% and 88% agreement for separate review of PUNLMP versus LG versus HG (kappa values 0.55 and 0.81, respectively), and 85% and 97% agreement for combined assessment of PUNLMP + LG versus HG.

Table 4 - Interobserver reproducibility for the 1973 and 2004/2016 WHO classifications

<table>
<thead>
<tr>
<th>Study</th>
<th>1973 WHO classification</th>
<th>2004/2016 WHO classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Type of analysis</td>
<td>Agreement (95% CI)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G1 vs G2 vs G3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>66% (59–73%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G1 + G2 vs G3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>89% (83–93%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>89%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>56%</td>
</tr>
<tr>
<td>van Rhijn (2010)</td>
<td>G1 vs G2 vs G3</td>
<td>39–54%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PUNLMP vs LG vs HG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G1 + G2 vs G3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PUNLMP + LG vs HG</td>
</tr>
<tr>
<td>May (2010)</td>
<td>G1 vs G2 vs G3</td>
<td>38–73%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PUNLMP vs LG vs HG</td>
</tr>
</tbody>
</table>

CI = confidence interval; G1 = grade 1; G2 = grade 2; G3 = grade 3; HG = high grade; LG = low grade; PUNLMP = papillary urothelial neoplasm with low malignant potential; WHO = World Health Organization.

* Pathologist A versus pathologist D (analysis of a total of four different combinations of two rounds of the grading assessment).

** Only Ta tumours included.

---

agreement for combined review of PUNLMP + LG versus HG (kappa values 0.70 and 0.91, respectively).

3.5. Discussion

3.5.1. Principal findings

This study demonstrates that both classifications identify patients at risk of tumour progression and recurrence; the risk rises significantly with increasing grade.

Additionally, we found that the 2004/2016 classification identifies patients with generally better prognosis. Our analysis demonstrates lower progression rates in all three grades of the 2004/2016 classification compared with the 1973 classification. Progression rates in G1 patients were similar to those in LG patients, while the rates in G3 patients were higher than those in HG patients. We found a lower recurrence rate in PUNLMP versus G1 patients, but a higher recurrence rate in G3 compared with HG patients.

Reproducibility assessment was hindered by a paucity of available studies [3,33]. In both studies, the interobserver reproducibility for G1 versus G2 versus G3 tumours was poor (kappa values 0.003–0.365), while the interobserver reproducibility for PUNLMP versus LG versus HG was poor to fair (kappa values 0.17–0.516). Comparing the reproducibility of G1 + G2 versus G3 and PUNLMP + LG versus HG tumours, kappa values were slightly higher for the 2004/2016 classification (0.44–0.58 vs 0.46–0.72). These findings suggest that the interobserver reproducibility of the 2004/2016 classification may be slightly better than that of the 1973 classification; however, the interobserver kappa values for both systems are disappointingly low.

The repeatability of both 1973 and 2004/2016 classifications was assessed in two studies [3,16]. In general, the intraobserver repeatability for G1 versus G2 versus G3 for the two pathologists was good (kappa values 0.61–0.69), whereas the repeatability for PUNLMP versus LG versus HG was fair to good (kappa values 0.56–0.83). Moreover, repeatability for G1 + G2 versus G3 and PUNLMP + LG versus HG was good to excellent (kappa values 0.88 and 0.80, respectively). One study [16] suggests that the intraobserver repeatability of the 2004/2016 classification may be better than that of the 1973 classification; however, another demonstrated no difference [3].

3.5.2. How do the review findings impact clinical practice and further research?

To address this, a discussion of the background, rationale, and critique of both grading systems is essential. Tumour grade is routinely used to determine prognosis, treatment, and follow-up of patients with NMIBC. Ideally, a grading system has to be practical, reproducible, and prognostically valid. EAU guidelines currently advocate the simultaneous use of both 1973 and 2004/2016 WHO classifications for grade because the 2004/2016 classification has not been sufficiently validated against the 1973 system [4].

Although the 1973 classification is well understood by clinicians, it has been criticised for a poorly defined G2 category, seen as a “default diagnosis”. Pathologists tend to classify a majority of tumours into the middle group when using a three-tier grading system [35].

The 2004/2016 classification is based on better-defined histological criteria. In theory, this should reduce inter- and intraobserver variability within a two-tiered classification, with the addition of PUNLMP category. However, several studies have shown considerable interobserver variability using the WHO 2004/2016 system [3,16,33].

There are several groups that are problematic for both grading systems:

3.5.2.1. G2 category. A high percentage of NMIBC is classified as G2 disease; previous studies have suggested that this is due to a lack of a clear definition of this category [8,36]. The proportion of G2 tumours in the 20 studies analysed in this systematic review was 50%; G1 tumours comprised 29% and G3 tumours 21%. This confirms the tendency to classify most patients as G2 in the 1973 classification and corresponds to the incidence of G2 tumours reported in the literature, which varies from 13% to 69% [37,38].

3.5.2.2. HG category. The primary objective of the 2004/2016 system was to improve the stratification of patients according to the risk of progression [8]. However, the inclusion of some G2 patients significantly enlarges the high-risk group. The percent of patients with HG tumours was two-fold higher (1887 cases, 42%) than those with G3 tumours (929 cases, 21%; Table 1). Treating HG tumours the same as G3 disease could lead to overtreatment of patients. Please cite this article in press as: Soukup V, et al. Prognostic Performance and Reproducibility of the 1973 and 2004/2016 World Health Organization Grading Classification Systems in Non-muscle-invasive Bladder Cancer: A European Association of Urology Non-muscle Invasive Bladder Cancer Guidelines Panel. Eur Urol (2017), http://dx.doi.org/10.1016/j.eururo.2017.04.015
with otherwise similar risk factors for progression (prior recurrence rate, tumour multiplicity, size, stage, and CIS). One of the advantages of the 1973 and 1999 WHO systems is the ability to identify more aggressive tumours: dividing HG disease into G2 and G3 may avoid overtreatment [16,39].

Implementation of the 2004/2016 system has been demonstrated to cause grade migration, with significantly more Ta cases graded as HG tumours; the resulting costs of overtreatment (BCG, re-transurethral resection, etc.) and associated morbidity are unknown [40].

3.5.2.3. Papillary urothelial neoplasm of low malignant potential. PUNLMP is defined as a papillary urothelial tumour that resembles exophytic urothelial papilloma but shows increased cellular proliferation exceeding the thickness of normal urothelium [8]. The introduction of this new category in the 2004/2016 WHO classification aimed to avoid labelling these patients with the term “cancer” to decrease psychosocial and economic burdens [37]. The published incidence of PUNLMP ranges from 12% to 39%, with recurrence rates between 25% and 60% and stage progression rates between 2% and 8%, very similar to the LG carcinomas [30,32,41,42].

Ten studies in this systematic review reported a total of 624 patients with PUNLMP and 1303 with G1 tumours [3,17,20,26–28,30–34]. Tumour recurrence occurred in 75 patients with PUNLMP and 111 G1 tumours (12% vs 9%). Tumour progression of PUNLMP, defined as any stage increase, was reported in eight studies [3,17,20,26,27,31–33]. Progression was diagnosed in six of 354 PUNLMP patients and 16 of 704 G1 patients (1.7% vs 2.3%). Progression to muscle-invasive disease from PUNLMP is very rare; it was found in one of 93 PUNLMP patients (1.1%) and eight of 250 G1 patients (3.2%).

Our study supports existing data demonstrating that progression of PUNLMP to muscle-invasive tumour is rare. The risk of recurrence and stage increase is comparable in PUNLMP and G1 patients. Moreover, molecular profiles of PUNLMP and G1 categories are similar [34]. Consequently, patients diagnosed with PUNLMP should be followed up in the same manner as patients with noninvasive G1 tumours.

3.5.2.4. T1 category. T1 tumours are rarely classified as LG [43]. As such, the 2004/2016 system does not allow differentiation of T1 tumours in subgroups with distinct prognoses [23].

Distribution of 2004/2016 WHO grade in the subgroup of T1 patients was reported in three studies included in our systematic review [22,23,29]. Of 681 T1 tumours, only 13 were classified as LG (1.9%). Recurrence and progression are more frequent in G3 than in HG tumours. Dividing HG T1 disease into G2 and G3, a higher recurrence rate (50% vs 68%) was found in one study [22] and a higher progression rate (12% vs 28%) was reported in two studies [22,29]. On the basis of these findings, the 1973 system may provide more accurate prognostic information in pT1 tumours. One solution may be the creation of new classification for grade, including elements from both 1973 and 2004/2016 systems, as suggested by van Rhijn et al [34].

3.5.3. Limitations and strengths of the review

Although this systematic review gives the best evidence we have so far, the quality of the evidence obtained was low, based on the absence of well-designed prospective studies with low RoBs. Heterogeneity in study designs, populations, treatment, definition of progression, incomplete reporting of outcome data, and lack of individual patient data limited the analyses that could be done and made meta-analysis inappropriate.

The main analysis in this systematic review is based on the studies for which both 1973 and 2004/2016 classifications were assessed. This approach has minimised bias and is the major strength of this review. Regarding the reproducibility part of the review, one study [16] appeared to present the overall global agreement and global kappa statistics, and not the agreement between pairs of pathologists as was done in the other two studies. Moreover, only two studies with a total of three pathologists assessed the intraobserver variability between the 1973 and 2004/2016 WHO classifications.

4. Conclusions

The current three-tiered 1973 and 2004/2016 WHO classification systems for grade are not optimal. Intra- and interobserver variability are slightly lower in the 2004/2016 WHO classification but still too high. We could not confirm that the 2004/2016 WHO classification outperforms the 1973 classification in predicting the risk of recurrence and progression. Each classification identifies different risk groups of NMIBC patients. In each category of the 1973 WHO classification (G1, G2, and G3), the risks of recurrence and progression are higher than in the corresponding category of the 2004/2016 WHO classification (PUNLMP, LG, and HG). A significant weakness of the 2004/2016 classification is that it gives almost no prognostic information in T1 patients, nearly all of whom are classified as HG. Prospective international multicentre studies and individual patient data analyses are needed to better assess the real prognostic value of the 1973 and 2004/2016 WHO classifications.

Author contributions: Viktor Soukup and Otakar Čapoun had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Soukup, Čapoun, Hernández, Babjuk, Burger, Compérat, Lam, MacLennan, van Rhijn, Rouprêt, Shariat, Sylvester, Zigeuner.

Acquisition of data: Soukup, Čapoun, Cohen, Hernández, Yuan.

Analysis and interpretation of data: Sylvester, Soukup, Čapoun, Cohen, Hernández.

Drafting of the manuscript: Soukup, Čapoun, Cohen, Hernández, Sylvester.

Critical revision of the manuscript for important intellectual content: Babjuk, Compérat, Zigeuner, Sylvester, Burger, Mostafid, van Rhijn, Gontero, Palou, Shariat.

Statistical analysis: None.

Obtaining funding: None.

Administrative, technical, or material support: MacLennan, Lam, Yuan.

Supervision: None.

Other: None.

Financial disclosures: Viktor Soukup and Otakar Čapoun certifies that all conflicts of interest, including specific financial interests and relationships and affiliations relevant to the subject matter or materials discussed in the manuscript (eg, employment/affiliation, grants or funding, consultancies, honoraria, stock ownership or options, expert testimony, royalties, or patents filed, received, or pending), are the following: O. Čapoun—honoraria or consultation fees from Janssen; company speaker honorarium from Janssen, Ipsen, Astellas, Bayer, GSK, Janssen Pharmaceuticals; trial participation in Janssen, Astra Zeneca, Aragon Pharmaceuticals: fellowship and travel grants from Janssen, Ipsen, Astellas, IPSEN, GSK. T. Lam—company consultant for Pfizer, GSK, Astellas, Ipsen; company speaker honorarium from Pfizer, GSK, Astellas, Ipsen. M. Babjuk—company consultant for Astellas, Ipsen; company speaker honorarium from Ferring, GSK; trial participation in Sotio. R. Zigeuner—company speaker honorarium from Pfizer, Bayer Healthcare, Roche, Novartis, Glaxo Smithkline, Amgen, Ipsen; fellowship and travel grants from Bayer Healthcare, Pfizer, Amgen, Novartis, Glaxo Smithkline, Astellas, Takeda; grants/research supports from Bayer Healthcare; company consultant for Pfizer. M. Burger—company consultant for Astellas, BMS, Ipsen Pharma, Janssen, Pfizer, Springer, Thieme, Cepheid, Merck Sharp Dome; company speaker honorarium from Astellas, BMS, Ipsen Pharma, Janssen, Pfizer, Springer, Thieme, Cepheid, Merck Sharp Dome; company speaker honorarium from Prostrakan; spouse/partner: Combat Medical; trial participation in MSD. B. van Rhijn—company consultant for Astellas 2015. P. Gontero—company consultant for Andromedical; trial participation in Astellas. J. Palou—company consultant for Olympus, ALLERGAN, IBSA, Olympus; company speaker honorarium from SANOFI-PASTEUR, General Electric; trial participation in Combat Medical, Presuryg. S. Shariat—company consultant for Astellas, Olympus, Wolff, Ipsen, Cepheid, Janssen; company speaker honorarium from Lilly, Astellas, Ipsen, Olympus, Wolff, Janssen; participation in Alere Inc. trials on NMP22; participation in a company-sponsored supporter’s bureau: BMS, MSD, Roche, Ipsen, Olympus; other: BMS, Janssen, Wolff, Olympus, Astellas, MSD; honoraria or consultation fees from Astellas, Olympus, Wolff, Ipsen, Janssen, Roche; trial participation in Roche, MSD, BMS; grants/research supports from Astellas, Sanofi; patents: 2001 Shariat S. and Slawin K.: Methods to determine prognosis after therapy for prostate cancer. U.S. patent application serial number: Docket#60/266,976. Filed May 31, 2001; 2001 Shariat S., Lerner S. and Slawin K.: Methods to determine prognosis after therapy for bladder cancer. U.S. patent application serial number: Docket#675,003US1. Filed June 1, 2001; 2002 Shariat S., Slawin K., Kattan M., and Scardino P.: Pre- and posttreatment nomograms for predicting recurrence in patients with clinically localized prostate cancer that includes the blood markers interleukin-6 soluble receptor and transforming growth; 2003 Slawin K., Kattan M., Shariat S., Stephenson A., and Scardino P.: Nomogram for predicting outcome of salvage radiotherapy for suspected local recurrence of prostate cancer after radical prostatectomy. US patent application serial number: Docket#5. Fi: 2005 Shariat S. M. Rouprét—company consultant for Lilly, GSK, Astellas, Takeda, Sanofi Pasteur, Ipsen; trial participation in GSK; company speaker honorarium from Roche, Zambron; trial participation in Pfizer, Roche. V. Soukup, E. Comperat, R. Sylvester, Steven MacLennan, Cathy Yuhong Yuan have nothing to declare.

Funding/Support and role of the sponsor: None.

References


