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

\textsuperscript{a}Department of Urology, General Teaching Hospital and 1st Faculty of Medicine, Charles University in Praha, Praha, Czech Republic; \textsuperscript{b}Department of Urology, Royal Free London NHS Foundation Trust, London, United Kingdom; \textsuperscript{c}Department of Urology, Hospital Universitario Fundación de Alcorcón, Madrid, Spain; \textsuperscript{d}Hospital Motol and Second Faculty of Medicine, Charles University, Department of Urology, Prague, Czech Republic; \textsuperscript{e}Department of Urology and Paediatric Urology, Julius-Maximilians-University Würzburg, Würzburg, Germany; \textsuperscript{f}Department of Pathology, Groupe Hospitalier Pitie' – Salpêtrière, Assistance Publique Hopitaux de Paris, Faculty of Medicine Pierre et Marie Curie, Institut Universitaire de Cancérologie GRC5, University Paris 6, Paris, France; \textsuperscript{g}Department of Surgical Sciences, Urology, University of Turin, Turin, Italy; \textsuperscript{h}Academic Urology Unit, University of Aberdeen, Scotland, United Kingdom; \textsuperscript{i}Department of Urology, Royal Surrey County Hospital, Guildford, UK; \textsuperscript{j}Department of Urology, Fundació Puigvert, Universitat Autònoma de Barcelona, Barcelona, Spain; \textsuperscript{k}Department of Urology, Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands; \textsuperscript{l}Department of Urology, Groupe
Hospitalier Pitie´ – Salpê trière, Assistance Publique Hopitaux de Paris, Faculty of Medicine
Pierre et Marie Curie, Institut. Universitaire de Cancérologie GRC5, University Paris 6, Paris,
France; mDepartment of Urology, Vienna General Hospital, Medical University of Vienna,
Vienna, Austria; nEAU Guidelines Office Board, European Association of Urology, The
Netherlands; oDepartment of Medicine, Health Science Centre, McMaster University,
Hamilton, Ontario, Canada; pDepartment of Urology, Medizinische Universität Graz, Graz,
Austria
Keywords: Non-muscle invasive bladder cancer; Grade; 1973 WHO classification; 2004/2016 WHO classification; Prognosis; Recurrence; Progression; Repeatability; Reproducibility.

Abstract

Context: Tumour grade is an important prognostic indicator in non-muscle invasive bladder cancer (NMIBC). Histopathological classifications are limited by inter-observer variability (reproducibility) which may have prognostic implications. EAU NMIBC guidelines suggest

**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, overall and cancer-specific survival. For reproducibility, the primary outcome was inter-observer variability between pathologists. Secondary outcome was intra-observer variability (repeatability) by the same pathologist.

**Evidence synthesis:** Of 3,593 articles identified, 20 studies were included in the prognostic review; 3 were eligible for the reproducibility review. Increasing tumour grade in both classifications was associated with higher disease progression and recurrence rates. Progression rates in G1 patients were similar to those in low grade patients; progression rates in G3 patients were higher than in high grade patients. Survival data was limited. Reproducibility of the 2004/2016 system was marginally better than 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 sub-optimal. The 1973 system identifies more aggressive tumours. Intra- and inter-observer 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.

Tweet 140 characters: Current grade classifications are not optimal in #bladdercancer according to #eauguidelines systematic review of the literature
1. Introduction

Up to 70% of patients with non-muscle-invasive bladder cancer (NMIBC) have tumour recurrence and about 10–15% progress to muscle-invasive disease [1]. Accurate prediction of tumour recurrence and progression is important to determine appropriate therapy and follow-up. Tumour grade is an important predictor of tumour prognosis [2]. However, histopathological classifications are known to be limited by inter- and intra-observer variability which may have profound prognostic implications [3].

Current European Association of Urology (EAU) recommendations for grading of NMIBC indicate that both the 1973 World Health Organization (WHO) and the 2004/2016 WHO classifications should be used [4]. The 1973 classification distinguishes 3 different grades and evaluates microscopic features related to the degree of cellular atypia, necrosis and mitotic activity. Grade 1 (G1) carcinomas (well-differentiated) are defined as showing only mild degrees of cytological atypia and infrequent mitotic figures. Grade 3 (G3) (poorly-differentiated) carcinomas are defined as showing marked nuclear pleomorphism, loss of maturation from the base to the surface and mitotic activity. Grade 2 (G2) carcinomas (moderately-differentiated) are comprised of all tumours between these extremes [5]. The lack of clarity between the three grades may adversely affect prognostic prediction due to high intra- and inter-observer variability. Furthermore, there is a tendency to classify the majority of tumours in the middle group (grade 2) [6].

In an attempt to reduce variability and increase reproducibility, a new grading system based on more detailed histological criteria has been promoted since 1998 by the International Society of Urological Pathology (ISUP) and was subsequently adopted by WHO in 2004. The main aim was to standardize the classification and grading of urothelial
neoplasms, creating a uniform terminology for use by pathologists and urologists [7,8]. Under the 2004 system, some G1 lesions are classified as papillary urothelial neoplasms with low malignant potential (PUNLMP) and others are classified as low grade (LG); G2 lesions are classified as low- or high-grade urothelial carcinomas; G3 lesions as high-grade (HG) urothelial carcinomas (Figure 1). Recently an update of the 2004 WHO grading classification was published without substantial changes so 2004 WHO classification is now known as 2016 WHO classification [9].

By eliminating the heterogeneous moderately-differentiated (G2) category of the 1973 system, the 2004/2016 classification was expected to provide a more reproducible stratification of patients with differing prognoses and well-defined recommendations for treatment and follow-up. However, several studies have shown considerable inter-observer variability and its anticipated superior prognostic value is still a matter of debate [6,10].

This systematic review compares the prognostic performance and reproducibility of the 1973 WHO and 1998 ISUP/2004 WHO/2016 WHO grading systems for NMIBC.

2. Evidence acquisition

2.1. Search strategy

The protocols for both the prognostic and reproducibility reviews have been published (http://www.crd.york.ac.uk/PROSPERO; registration numbers CRD42015025045 and CRD42016029714); the search strategy is outlined in Supplement 1.

Databases including Medline, Embase, and the Cochrane Central Register of Controlled Trials were systematically searched from 1st January 1998 to 31st 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 onward were included. There were no language restrictions. A minimum follow-up of 3 months (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 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: adult patients (>18 years old) with primary or recurrent Ta/T1 urothelial carcinoma (UC) of the bladder who underwent a Transurethral Resection of Bladder Tumour (TURBT). All risk groups and adjuvant treatments were included. Exclusion criteria were: patients under 18 years; Muscle-Invasive Bladder Cancer (MIBC); clinical N+ or M+; grading based on radical cystectomy specimen; bladder biopsies only (as opposed to
The protocol allowed inclusion of studies with exclusion criteria if affected subjects 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, overall and cancer-specific survival. All outcomes were measured at least 3 months post-TURBT.

In the reproducibility review, the primary outcome was inter-observer variability (reproducibility) between pathologists. The secondary outcome was intra-observer 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, Statistical analysis [11]. The EAU NMIBC Guidelines Panel identified the three most important prognostic confounders as intravesical BCG (yes/no), stage (Ta/T1) and concomitant CIS (yes/no). 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 \( \kappa \)). Arbitrary guidelines characterize values of kappa greater than 0.75 as excellent concordance, 0.40 to 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 (Figure 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]. 3 studies involving 566 patients met the reproducibility inclusion criteria [3, 16, 33].

3.2. Characteristics of the 20 included studies
The baseline characteristics of included studies in prognostic review are detailed in Table 1. The three retrospective studies contained information on reproducibility or repeatability: Mangrud [16] - three pathologists independently reviewed both classifications, two pathologists repeated the classification for intra-observer variability, however only one pathologist assessed both grading systems. Van Rhijn [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 [33] reported reproducibility of both grading systems between four independent pathologists (Table 1).

3.3. Risk of bias 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 was 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 in T1 patients (Table 2 and 3).
3.4. Comparisons of prognostic outcome measures

For analysis of progression, recurrence, overall and cancer specific survival, most available information concerned the number of patients with an event during follow up and the percentage of patients with an event at a given point in time. There was little time-to-event data i.e. 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 of the studies provided endpoint information for each grading classification. In order to minimize the risk of bias when comparing 1973 to 2004/2016, the most reliable results were obtained when analysing only the studies that assessed both grading classifications. Thus, the two grading classifications are each assessed on the same set of patients so 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 6 studies, progression was defined as any increase in disease stage, including Ta to T1, while in 7 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

3.4.1.1.1.1. Comparisons only from studies that assessed both the 1973 and 2004/2016 classifications

Direct comparison of the two grading systems demonstrated progression by 1973 grade (G1 vs G2 vs G3) in 3% vs 9% vs 32%, whereas for 2004/2016 grade (PUNLMP vs LG vs HG), 1% vs 4% vs 25% progressed, respectively (Table 2).

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

3.4.1.1.1.2. Comparisons using all available data

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

3.4.1.1.2. Progression defined as any increase in disease stage
3.4.1.1.1.2.1. Comparisons only from studies that assessed both the 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% vs 8% vs 27% and (PUNLMP vs LG vs HG) 2% vs 4% vs 22%, respectively (Table 2).

In LG Ta patients, we found a higher progression rate in G2 patients as compared to G1 patients - 7% vs 1%.

3.4.1.1.2.2. Comparison using all available data

Progression rates were (G1 vs G2 vs G3) 3% vs 9% vs 28%, respectively and (PUNLMP vs LG vs HG) 2% vs 4% vs 19%, respectively.

3.4.1.2. Recurrence

Eight studies provided information on the number of patients with recurrence, but only 5 used both grading systems (Table 3).

3.4.1.2.1. Comparison of 5 studies that utilised both 1973 and 2004/2016 classifications

The pooled recurrence rates were (G1 vs G2 vs G3) 33% vs 42% vs 63% and (PUNLMP vs LG vs HG) 20% vs 38% vs 55%, respectively (Table 3).
The majority of patients in these 5 studies had Ta disease; a separate analysis in T1 patients was not possible [16, 20, 26, 30, 33]. A subgroup analysis of T1 high grade patients revealed a higher recurrence rate in the G3 patients compared with G2 (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% vs 44% vs 65%, respectively. For the 2004/2016 grade (PUNLMP vs LG vs HG), 28% vs 43% vs 58% recurred, respectively (Table 3).

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

3.4.1.3. Death Due to Bladder Cancer

Only 1 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 2 studies [18, 28] and only 1 study contributed to the analysis [31]. In this study, death rates for the best and
worst prognosis patients 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 inter-observer agreement and kappa values for the 1973 and 2004/2016 WHO
classifications are presented in Table 4.

The inter-observer agreement for the 1973 classification ranged from 38% to 89% (kappa
values from 0.003 to 0.68). Agreement in combined assessment of G1+2 vs G3 tumours in
two studies [3, 16] was higher than in separate assessment of G1 vs G2 vs G3 tumours (80-
89% vs 39-66%; kappa values 0.44-0.68 vs 0.15-0.68). The inter-observer agreement for the
2004/2016 classification ranged from 43% to 100% (kappa values 0.17 to 0.70). Only one
study assessed agreement between two pathologists in combined review of PUNLMP+LG vs
HG tumours [3]. It showed slightly better reproducibility than for a separate analysis of
PUNLMP vs LG vs 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 only 2004/2016 WHO
classification. Inter-observer agreement for the separate review of PUNLMP vs LG vs 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 vs
HG tumours ranged from 65% to 88% (kappa values from 0.30 to 0.73).

3.4.2.2. Repeatability
The intra-observer 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 intra-observer agreement for 1973 WHO grading classification ranged from 63% to 95% (kappa values 0.61 to 0.88). Repeatability for combined assessment of G1+G2 vs G3 tumours was slightly higher than for a separate analysis of G1 vs G2 vs G3 tumours (88-95% vs 63-81%, kappa values 0.64-0.88 vs 0.61-0.69). The intra-observer agreement for 2004/2016 WHO grading classification ranged from 71% to 93% (kappa values 0.56 to 0.83). In the only study that assessed the difference between combined and separate pathological review, the repeatability of group PUNLMP+LG vs HG was higher than in PUNLMP vs LG vs 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 both for separate review of PUNLMP vs LG vs HG (kappa values 0.55 and 0.81) and 85% and 97% for combined review of PUNLMP+LG vs HG (kappa values 0.70 and 0.91).

4. Discussion

4.1. 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 3 grades of the 2004/2016 classification compared to the 1973 classification. Progression rates in G1 patients were similar to LG patients, while those in G3 patients were higher than HG.
patients. We found a lower recurrence rate in PUNLMP versus G1 patients, but a higher recurrence rate in G3 compared to HG patients.

Reproducibility assessment was hindered by a paucity of available studies [3, 33]. In both studies the inter-observer reproducibility for G1 vs G2 vs G3 tumours was poor (kappa values 0.003 to 0.365), while the inter-observer reproducibility for PUNLMP vs LG vs HG was poor to fair (kappa values 0.17 to 0.516). Comparing the reproducibility of G1+G2 vs G3 and PUNLMP+LG vs 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 inter-observer reproducibility of the 2004/2016 classification may be slightly better than that of the 1973 classification, however the inter-observer 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 intra-observer repeatability for G1 vs G2 vs G3 for the two pathologists was good (kappa values 0.61-0.69), whereas the repeatability for PUNLMP vs LG vs HG was fair to good (kappa values 0.56-0.83). Moreover, repeatability for G1+G2 vs G3 and PUNLMP+LG vs HG was good to excellent (kappa values 0.88 and 0.80). One study [16] suggests that intra-observer repeatability of the 2004/2016 classification may be better than that of the 1973 classification, however another demonstrated no difference [3].

4.2. How do the review findings impact on 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 grade 2 category, seen as a “default diagnosis.” Pathologists tend to classify a majority of tumours into the middle group when using a 3-tier-grading system [36].

The 2004/2016 classification is based on better defined histological criteria. In theory, this should reduce inter- and intra-observer variability within a 2-tiered classification, with the addition of PUNLMP category. However, several studies have shown considerable inter-observer variability using the WHO 2004/2016 system [3, 16, 33].

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

4.2.1. 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 [36, 37]. 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% [38, 39].

4.2.2. HG category
The primary objective of the 2004/2016 system was to improve the stratification of patients according to the risk of progression [36]. 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 with otherwise similar risk factors for progression (prior recurrence rate, tumour multiplicity, size, stage, CIS). One of the advantages of the 1973 and WHO 1999 systems is the ability to identify the more aggressive tumours; dividing HG disease into G2 and G3 may avoid overtreatment. [16, 40].

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 (BGC, re-TUR etc.) and associated morbidity are unknown [40].

4.2.3. Papillary urothelial neoplasm of low malignant potential

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 [38]. The published incidence of PUNLMP ranges from 12–39%, with recurrence rates between 25 and 60% and stage progression rates between 2 and 8%, very similar to the low-grade carcinomas [30, 32, 42, 43].
Ten studies in this systematic review reported a total of 624 patients with PUNLMP and 1303 patients with G1 tumours [3, 17, 20, 26-28, 30-34]. Tumour recurrence occurred in 75 with PUNLMP and 111 patients G1 tumours (12% vs 9%).

Tumour progression of PUNLMP, defined as any stage increase, was reported in 8 studies [3, 17, 20, 26, 27, 31-33]. Progression was diagnosed in 6 of 354 PUNLMP patients and in 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 in 8 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, the molecular profile of PUNLMP and G1 categories is similar [34]. Consequently, patients diagnosed with PUNLMP should be followed-up in the same manner as patients with non-invasive G1 tumours.

### 4.2.4. T1 category

T1 tumours are rarely classified as low-grade [44]. As such, the 2004/2016 system does not allow differentiation of T1 tumours in sub-groups 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 low-grade (1.9%).

Recurrence and progression are more frequent in G3 than 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 [33].

4.3. 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 risks of bias. Heterogeneity in study designs, populations, treatment, definition of progression, incomplete reporting of outcome data and the 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 the 1973 and 2004/2016 classifications were assessed. This approach has minimized bias and is the major strength of the 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 intra-observer variability between WHO 1973 and 2004/2016 classifications.

5. Conclusions

Current three tiered WHO 1973 and 2004/2016 classifications systems for grade are not optimal. Intra- and inter-observer variability are slightly lower in 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, G3), the risks of recurrence and progression are higher than in the corresponding category of 2004/2016 WHO classification (PUNLMP, LG, 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 WHO and 2004/2016 WHO classifications.
Figure 1. Stratification of tumours according to grade in the WHO 1973 and 2004 classifications.

Classification WHO 2004

Classification WHO 1973

PUN-LMP = papillary urothelial neoplasia-low malignant potential, PUC-LG = papillary urothelial carcinoma-low grade, PUC-HG = papillary urothelial carcinoma-low grade
Figure 2. PRISMA diagram (applicable for both prognostic and reproducibility reviews)

* Three of those studies were also eligible for the reproducibility part
Figure 3 – (a) Risk of bias for included studies (n = 20). Green indicates low risk, red indicates high risk, and yellow indicates unclear risk.
Table 1 - Baseline study characteristics for the 20 comparative studies with 4505 patients.

<table>
<thead>
<tr>
<th>Author</th>
<th>Study Start</th>
<th>Follow Up (Month(s))</th>
<th>Uropathology</th>
<th>Patients Included</th>
<th>Patients Excluded</th>
<th>Age (mean)</th>
<th>Males/Females</th>
<th>T Category TV/T1</th>
<th>CI</th>
<th>Intravesical Therapy</th>
<th>of BC</th>
<th>G1</th>
<th>G2</th>
<th>G3</th>
<th>P UPL</th>
<th>MP</th>
<th>MS</th>
<th>MG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mangrut 2010a*, [16]</td>
<td>200 2-200 6</td>
<td>75.0</td>
<td>No</td>
<td>193</td>
<td>56</td>
<td>146/45</td>
<td>Ta and T1</td>
<td>2</td>
<td>2</td>
<td>193</td>
<td>44</td>
<td>98</td>
<td>51</td>
<td>0</td>
<td>11</td>
<td>9</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>Gontero 2014, [18]</td>
<td>199 2-200 6</td>
<td>71.6</td>
<td>Yes</td>
<td>131</td>
<td>60</td>
<td>66.3</td>
<td>112/19</td>
<td>Only Ta</td>
<td>5</td>
<td>65</td>
<td>65</td>
<td>0</td>
<td>10</td>
<td>5</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Nishiyama et al 2013, [19]</td>
<td>199 5-201 0</td>
<td>No</td>
<td>153</td>
<td>68.5</td>
<td>122/31</td>
<td>Ta and T1</td>
<td>49</td>
<td>24</td>
<td>2</td>
<td>89</td>
<td>62</td>
<td>0</td>
<td>37</td>
<td>11</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chen 2012, [20]</td>
<td>199 9-200 0</td>
<td>47.0</td>
<td>Yes</td>
<td>348</td>
<td>44</td>
<td>287/61</td>
<td>Ta and T1</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>12</td>
<td>5</td>
<td>17</td>
<td>6</td>
<td>47</td>
<td>40</td>
<td>22</td>
<td>3</td>
</tr>
<tr>
<td>Pellucchi 2011, [21]</td>
<td>200 4-200 8</td>
<td>25.0</td>
<td>Yes</td>
<td>270</td>
<td>162</td>
<td>220/50</td>
<td>Only Ta</td>
<td>270</td>
<td>87</td>
<td>18</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>27</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pellucchi 2012, [22]</td>
<td>200 4-200 8</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>26</td>
<td>8</td>
<td>0</td>
<td>12</td>
<td>4</td>
<td>14</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ohto 2011, [23]</td>
<td>198 9-200 0</td>
<td>49.0</td>
<td>Yes</td>
<td>310</td>
<td>79</td>
<td>71.7</td>
<td>239/71</td>
<td>Only T1</td>
<td>25</td>
<td>2</td>
<td>11</td>
<td>2</td>
<td>19</td>
<td>8</td>
<td>0</td>
<td>13</td>
<td>29</td>
<td>7</td>
</tr>
<tr>
<td>Van Rhijn 2010a, [25]</td>
<td>198 3-200 6</td>
<td>68.0</td>
<td>Yes</td>
<td>164</td>
<td>68.8</td>
<td>135/29</td>
<td>Ta and T1</td>
<td>5</td>
<td>5</td>
<td>26</td>
<td>16</td>
<td>4</td>
<td>0</td>
<td>74</td>
<td>90</td>
<td>0</td>
<td>37</td>
<td>12</td>
</tr>
<tr>
<td>Burger 2008b, [26]</td>
<td>198 5-200 2</td>
<td>48</td>
<td>Yes</td>
<td>109</td>
<td>60</td>
<td>97/12</td>
<td>Only Ta</td>
<td>6</td>
<td>58</td>
<td>46</td>
<td>5</td>
<td>6</td>
<td>77</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burger 2008b, [27]</td>
<td>-</td>
<td>Yes</td>
<td>211</td>
<td>0</td>
<td>173/50</td>
<td>T1 and T1</td>
<td>80</td>
<td>11</td>
<td>0</td>
<td>25</td>
<td>49</td>
<td>11</td>
<td>9</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schmid 2007, [28]</td>
<td>199 4-200 0</td>
<td>No</td>
<td>504</td>
<td>353</td>
<td>61.5</td>
<td>376/128</td>
<td>Only T1</td>
<td>29</td>
<td>5</td>
<td>15</td>
<td>4</td>
<td>55</td>
<td>179</td>
<td>22</td>
<td>4</td>
<td>73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kamel 2006, [29]</td>
<td>199 1-200 3</td>
<td>48</td>
<td>No</td>
<td>105</td>
<td>-</td>
<td>85/20</td>
<td>Only T1</td>
<td>1</td>
<td>5</td>
<td>25</td>
<td>0</td>
<td>61</td>
<td>44</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Yin 2004, [30]</td>
<td>199 5-200 0</td>
<td>No</td>
<td>84</td>
<td>0</td>
<td>69.4</td>
<td>-</td>
<td>Only Ta</td>
<td>32</td>
<td>46</td>
<td>5</td>
<td>3</td>
<td>12</td>
<td>53</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oesterhuis s 2002, [31]</td>
<td>199 9-200 0</td>
<td>No</td>
<td>320</td>
<td>79</td>
<td>66.6</td>
<td>295/64</td>
<td>Only T1</td>
<td>28</td>
<td>8</td>
<td>31</td>
<td>28</td>
<td>6</td>
<td>1</td>
<td>116</td>
<td>14</td>
<td>1</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>Samaratunga 2002, [32]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>134</td>
<td>-</td>
<td>65.7</td>
<td>95/39</td>
<td>Only T1</td>
<td>16</td>
<td>42</td>
<td>79</td>
<td>6</td>
<td>29</td>
<td>73</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Holmang 2001, [33]</td>
<td>198 7-198 9</td>
<td>No</td>
<td>363</td>
<td>317</td>
<td>-</td>
<td>Only Ta</td>
<td>3</td>
<td>0</td>
<td>25</td>
<td>5</td>
<td>95</td>
<td>13</td>
<td>95</td>
<td>16</td>
<td>0</td>
<td>10</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Van Rhijn 2010b*, [3]</td>
<td>198 3-200 1</td>
<td>Yes</td>
<td>175</td>
<td>-</td>
<td>64.9</td>
<td>120/44</td>
<td>Ta and T1</td>
<td>1</td>
<td>5</td>
<td>79</td>
<td>75</td>
<td>25</td>
<td>97</td>
<td>51</td>
<td>18</td>
<td>69</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>May M 2010*, [33]</td>
<td>199 7-200 4</td>
<td>Yes</td>
<td>200</td>
<td>-</td>
<td>68.6</td>
<td>149/51</td>
<td>Only Ta</td>
<td>0</td>
<td>82</td>
<td>10</td>
<td>9</td>
<td>1</td>
<td>14</td>
<td>9</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>van Rhijn 2014, [34]</td>
<td>198 6-200 6</td>
<td>Yes</td>
<td>325</td>
<td>0</td>
<td>66.4</td>
<td>254/71</td>
<td>Ta and T1</td>
<td>6</td>
<td>2</td>
<td>98</td>
<td>22</td>
<td>5</td>
<td>88</td>
<td>14</td>
<td>9</td>
<td>88</td>
<td>79</td>
<td>10</td>
</tr>
</tbody>
</table>
## Table 2: The distribution of the percent of patients with tumour progression

<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>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>study numb.</td>
<td>percent G1 patients</td>
<td>percent G2 patients</td>
<td>percent G3 patients</td>
<td>value</td>
<td>study numb.</td>
<td>percent G1 patients</td>
<td>percent G2 patients</td>
<td>percent G3 patients</td>
</tr>
<tr>
<td>Studies in which both 1973 and 1998/2004 can be compared</td>
<td>T2 or greater Increase in Stage</td>
<td>Chen 2012 [20], Burger 2008a [26], Samaratunag 2002 [32], Van Rhijn 2010b [5]</td>
<td>757</td>
<td>5.2</td>
<td>8.5</td>
<td>32.1</td>
<td>0.000</td>
<td>Chen 2012 [20], Burger 2008a [26], Samaratunag 2002 [32], Van Rhijn 2010b [5]</td>
<td>761</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>Any Increase in Stage</td>
<td>Mangrud 2014b [16], Chen 2012 [20], Burger 2008a [26], Burger 2008b [27], Samaratunag 2002 [32], Van Rhijn 2010b [5], May M 2010 [33]</td>
<td>1371</td>
<td>3.3</td>
<td>8.4</td>
<td>27.3</td>
<td>0.000</td>
<td>Mangrud 2014b [16], Chen 2012 [20], Burger 2008a [26], Burger 2008b [27], Samaratunag 2002 [32], Van Rhijn 2010b [5], May M 2010 [33]</td>
<td>1372</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>T2 or greater Increase in Stage</td>
<td>Chen 2012 [20], Pellucchi 2015 [22], Burger 2008a [26], Kamel 2006 [29], Samaratunag 2002 [32], Van Rhijn 2010b [5]</td>
<td>1128</td>
<td>3.2</td>
<td>9.8</td>
<td>29.5</td>
<td>0.000</td>
<td>Gontero 2014 [18], Chen 2012 [20], Pellucchi 2015 [22], Burger 2008a [26], Kamel 2006 [29], Samaratunag 2002 [32], Van Rhijn 2010b [5], May M 2010 [33]</td>
<td>1263</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>Any Increase in Stage</td>
<td>Mangrud 2014b [16], Chen 2012 [20], Pellucchi 2011 [21], Pellucchi 2015 [22], Burger 2008a [26], Burger 2008b [27], Kamel 2006 [29], Samaratunag 2002 [32], Van Rhijn 2010b [5], May M 2010 [33]</td>
<td>2012</td>
<td>2.9</td>
<td>8.9</td>
<td>27.6</td>
<td>0.000</td>
<td>Mangrud 2014b [16], Gontero 2014 [18], Chen 2012 [20], Pellucchi 2011 [21], Pellucchi 2015 [22], Burger 2008a [26], Burger 2008b [27], Kamel 2006 [29], Oosterhuis 2002 [31], Samaratunag 2002 [32], Holmang 2002 [17], Van Rhijn 2010b [5], May M 2010 [33]</td>
<td>2809</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>T2 patients only</td>
<td>Pellucchi 2011 [21], Burger 2008a [26], Samaratunag 2002 [32], May M 2010</td>
<td>706</td>
<td>3.7</td>
<td>7.4</td>
<td>35.0</td>
<td>0.000</td>
<td>Gontero 2014 [18], Pellucchi 2011 [21], Burger 2008a [26], Oosterhuis 2002 [31]</td>
<td>1506</td>
<td>1.6</td>
</tr>
<tr>
<td>T1 patients only</td>
<td>T2 or greater increase in Stage</td>
<td>Pelluchi 2015 [22], Kamel 2006 [29]</td>
<td>371</td>
<td>-</td>
<td>12.4</td>
<td>28.0</td>
<td>0.000</td>
<td>Pelluchi 2015 [22], Kamel 2006 [29]</td>
<td>371</td>
<td>-</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>---------------------------------</td>
<td>--------------------------------------</td>
<td>------</td>
<td>----</td>
<td>------</td>
<td>------</td>
<td>-------</td>
<td>--------------------------------------</td>
<td>------</td>
<td>----</td>
</tr>
<tr>
<td>G1 vs G3 in Ta LG tumours</td>
<td>Any increase in Stage</td>
<td>Pelluchi 2011 [21]</td>
<td>270</td>
<td>1.2</td>
<td>7.1</td>
<td>-</td>
<td>0.039</td>
<td>Pelluchi 2011 [21]</td>
<td>270</td>
<td>-</td>
</tr>
<tr>
<td>G2 vs G3 in T1 HG tumours</td>
<td>T2 or greater increase in Stage</td>
<td>Pelluchi 2015 [22], Kamel 2006 [29]</td>
<td>371</td>
<td>-</td>
<td>12.4</td>
<td>28.0</td>
<td>0.000</td>
<td>Pelluchi 2015 [22], Kamel 2006 [29]</td>
<td>371</td>
<td>-</td>
</tr>
</tbody>
</table>
### Table 3: The distribution of the percentage of patients with tumour recurrence

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Studies included</td>
<td>number of patients</td>
<td>percent G1 patients with recurrence</td>
<td></td>
<td>number of patients</td>
<td>percent G2 patients with recurrence</td>
<td></td>
<td>number of patients</td>
</tr>
<tr>
<td>Studies in which both 1973 and 1998/2004 can be compared</td>
<td>Mangrud 2014b [16], Chen 2012 [20], Burger 2008a [26], Yin 2004 [30], May M 2010 [33]</td>
<td>931</td>
<td>32.6</td>
<td>42.3</td>
<td>62.6</td>
<td>0.000</td>
<td>Mangrud 2014b [16], Chen 2012 [20], Burger 2008a [26], Yin 2004 [30], May M 2010 [33]</td>
<td>934</td>
</tr>
<tr>
<td>All studies with Recurrence Data</td>
<td>Mangrud 2014b [16], Chen 2012 [20], Pelluchi 2015 [22], Burger 2008a [26], Yin 2004 [30], May M 2010 [33]</td>
<td>1197</td>
<td>32.6</td>
<td>43.9</td>
<td>65.4</td>
<td>0.000</td>
<td>Mangrud 2014b [16], Chen 2012 [20], Pelluchi 2015 [22], Burger 2008a [26], Yin 2004 [30], Oosterhuis J.W.A. 2002 [31], Holmang 2001 [17], May M 2010 [33]</td>
<td>1865</td>
</tr>
<tr>
<td>Ta patients only</td>
<td>Burger 2008a [26], Yin 2004 [30], May M 2010 [33]</td>
<td>390</td>
<td>39.0</td>
<td>40.8</td>
<td>70.6</td>
<td>0.040</td>
<td>Burger 2008a [26], Yin 2004 [30], Oosterhuis J.W.A. 2002 [31], Holmang 2001 [17], May M 2010 [33]</td>
<td>988</td>
</tr>
<tr>
<td>T1 patients only</td>
<td>Pelluchi 2015 [22]</td>
<td>266</td>
<td>-</td>
<td>50.0</td>
<td>67.6</td>
<td>0.004</td>
<td>Pelluchi 2015 [22]</td>
<td>266</td>
</tr>
<tr>
<td>G2 vs G3 in T1 HG tumours</td>
<td>Pelluchi 2015 [22]</td>
<td>266</td>
<td>-</td>
<td>50.0</td>
<td>67.6</td>
<td>0.004</td>
<td>Pelluchi 2015 [22]</td>
<td>266</td>
</tr>
</tbody>
</table>
Table 4. Inter-observer reproducibility for the 1973 and 2004/2016 WHO classifications

<table>
<thead>
<tr>
<th>Study</th>
<th>Type of analysis</th>
<th>Agreement (95% CI)</th>
<th>Kappa (95% CI)</th>
<th>1973 WHO classification</th>
<th>Type of analysis</th>
<th>Agreement (95% CI)</th>
<th>Kappa (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mangrud 2014b [16]</td>
<td>G1 vs G2 vs G3</td>
<td>66% (59-73%)</td>
<td>0.68 (0.57-0.78)</td>
<td>LG</td>
<td>LG</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G1+G2 vs G3</td>
<td>69% (61-76%)</td>
<td>0.68 (0.56-0.80)</td>
<td>HG</td>
<td>G1</td>
<td>66%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G2</td>
<td>15%</td>
<td></td>
<td>G2 vs HG</td>
<td>LG vs HG</td>
<td>87% (81-91%)</td>
<td>0.70 (0.59-0.81)</td>
</tr>
<tr>
<td>Van Rhijn 2010b [3]</td>
<td>G1 vs G2 vs G3*</td>
<td>39-54%</td>
<td>0.15-0.32</td>
<td>PUNLMP vs LG vs HG*</td>
<td>43-66%</td>
<td>0.17-0.48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G1+G2 vs G3</td>
<td>80%</td>
<td>0.44-0.58</td>
<td>PUNLMP vs LG vs HG*</td>
<td>73-86%</td>
<td>0.46-0.72</td>
<td></td>
</tr>
<tr>
<td>May M 2010‡ [33]</td>
<td>G1 vs G2 vs G3‡</td>
<td>38-73%</td>
<td>0.003-0.365</td>
<td>PUNLMP vs LG vs HG‡</td>
<td>71-83%</td>
<td>0.296-0.516</td>
<td></td>
</tr>
</tbody>
</table>

* Pathologist A vs pathologist D (analysis of a total of four different combinations of two rounds of the grading assessment), † Pathologist A vs B vs C vs D (a total of six pairwise comparisons), ‡ only Ta tumours included

Table 5. Intra-observer repeatability for the 1973 and 2004/2016 WHO classifications

<table>
<thead>
<tr>
<th>Study</th>
<th>Pathologist (type of analysis)</th>
<th>Agreement (95% CI)</th>
<th>Kappa (95% CI)</th>
<th>Pathologist (type of analysis)</th>
<th>Agreement (95% CI)</th>
<th>Kappa (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mangrud 2014b [16]</td>
<td>A (G1 vs G2 vs G3)</td>
<td>68% (61-74%)</td>
<td>0.69 (0.59-0.79)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>A (G1+G2 vs G3)</td>
<td>88% (82-93%)</td>
<td>0.66 (0.54-0.79)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>B (G1 vs G2 vs G3)</td>
<td>63% (56-70%)</td>
<td>0.61 (0.48-0.74)</td>
<td>B (PUNLMP vs LG vs HG)</td>
<td>93% (88-96%)</td>
<td>0.83 (0.74-0.92)</td>
</tr>
<tr>
<td>Van Rhijn 2010b [3]</td>
<td>A (G1 vs G2 vs G3)</td>
<td>80%</td>
<td>0.67 (0.57-0.76)</td>
<td>A (PUNLMP vs LG vs HG)</td>
<td>71%</td>
<td>0.56 (0.46-0.66)</td>
</tr>
<tr>
<td></td>
<td>D (G1 vs G2 vs G3)</td>
<td>81%</td>
<td>0.69 (0.59-0.78)</td>
<td>D (PUNLMP vs LG vs HG)</td>
<td>82%</td>
<td>0.69 (0.60-0.78)</td>
</tr>
<tr>
<td></td>
<td>A (G1+G2 vs G3)</td>
<td>91%</td>
<td>0.64 (0.48-0.83)</td>
<td>A (PUNLMP vs LG vs HG)</td>
<td>86%</td>
<td>0.68 (0.57-0.80)</td>
</tr>
<tr>
<td></td>
<td>D (G1+G2 vs G3)</td>
<td>95%</td>
<td>0.88 (0.80-0.96)</td>
<td>D (PUNLMP vs LG vs HG)</td>
<td>90%</td>
<td>0.80 (0.72-0.89)</td>
</tr>
</tbody>
</table>
References:


