Interlaboratory Gleason grading variation affects treatment: a Dutch historic cohort study in 30 509 patients with prostate cancer

Rachel N Flach ,[1] Carmen van Dooijeweert,[2] Katja K H Aben,[3,4] Britt B M Suelmann ,[5] Peter-Paul M Willemse,[1] Paul J van Diest ,[2] Richard P Meijer[1]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/jcp-2021-208067).

[1]Department of Oncological Urology, UMC Utrecht, Utrecht, The Netherlands
[2]Department of Pathology, UMC Utrecht, Utrecht, The Netherlands
[3]Department of Research & Development, Netherlands Comprehensive Cancer Centre, Utrecht, The Netherlands
[4]Radboud Institute for Health Sciences, Radboud UMC, Nijmegen, Gelderland, The Netherlands
[5]Department of Medical Oncology, UMC Utrecht, Utrecht, The Netherlands

Correspondence to
Professor Paul J van Diest; p.j.vandiest@umcutrecht.nl

Received 23 November 2021
Accepted 11 June 2022

To cite: Flach RN, van Dooijeweert C, Aben KKH, *et al*. *J Clin Pathol* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/jclinpath-2021-208067

ABSTRACT

Aim Substantial variation in Gleason grading (GG) of prostate cancer (PCa) exists between Dutch pathology laboratories. This study investigates its impact on treatment strategies.

Methods Pathology reports of prostate needle biopsies and clinical data of patients with PCa diagnosed between 2017 and 2019 were retrieved from the Dutch nationwide network and registry of histopathology and cytopathology and The Netherlands Cancer Registry. We investigated the impact of grading variation on treatment strategy for patients whose grade was decisive in treatment choice. First, we evaluated the effect of grading practice (low, average or high grading) on active treatment (AT) versus active surveillance in patients with prostate-specific antigen (PSA) <10 ng/mL and cT1c/cT2a disease. Second, we assessed the association of grading practice with performance of pelvic lymph node dissection (PLND) in patients with PSA 10–20 ng/mL or cT2b disease. We used multivariable logistic regression to analyse the relation between laboratories' grading practices and AT or PLND.

Results We included 30 509 patients. GG was decisive in treatment strategy for 11 925 patients (39%). AT was performed significantly less often in patients diagnosed by laboratories that graded lower than average (OR=0.77, 95% CI 0.68 to 0.88). Conversely, patients received AT significantly more often when diagnosed in high-grading laboratories versus average-grading laboratories (OR=1.21, 95% CI 1.03 to1.43). PLND was performed significantly less often in patients diagnosed by low-grading versus average-grading laboratories (OR=0.66, 95% CI 0.48 to 0.90).

Conclusion Our study shows that the odds that a patient undergoes AT or PLND, depends on laboratories' grading practices in a substantial number of patients. This likely influences patient prognosis and outcome, necessitating standardisation of GG to prevent suboptimal patient outcome.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Prostate cancer grading (the Gleason grade) is subjected to considerable observer variation, both between individual pathologists and different pathology laboratories. It is unknown how many patients are affected, and how this affects their treatment strategy.

WHAT THIS STUDY ADDS

⇒ Gleason grade is decisive in treatment strategy for roughly 40% of all patients with prostate cancer in the Netherlands. Patients in higher grading laboratories are more likely to receive active treatment, than those in lower grading laboratories.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Research should focus on improving consistency in prostate cancer grading. Clinicians and pathologists should perform second readings more often for patients for whom grading is decisive in treatment strategy.

INTRODUCTION

Prostate cancer (PCa) is the most common cancer in European men and incidence numbers have tripled in 30 years, peaking to over 13 000 newly diagnosed cases of PCa in 2019 in The Netherlands.[1 2] PCa prognosis and treatment are based on histologic grading of PCa (Gleason grade (GG)).[3–5] Interobserver variation of GG has been reported multiple times, and genitourinary pathologists outperform general pathologists.[6–10] Several efforts have been made to improve this system. For example, the GG system is updated regularly, which in 2014 resulted in the introduction of grade groups that would better reflect clinical prognosis, according to the International Society of Urological Pathology (ISUP) conference.[3 11] Nevertheless, since the implementation of the ISUP grade groups, interobserver variation persists.[12] We recently even showed that this variation exists in daily clinical practice on a nationwide level between and within pathology laboratories.[13] Even though grade is crucial in treatment strategy and prognosis, it is unclear how this variation affects patients.

Localised PCa (ie, tumour confined to the prostate without any metastases) is subdivided into three risk groups, according to the European Association of Urology-European Society for Radiotherapy and Oncology-International Society of Geriatric Oncology (EAU-ESTRO-SIOG) risk stratification,[14] which determines treatment strategies. Low-risk patients are eligible for active surveillance (AS) (besides active treatment (AT)), whereas intermediate-risk and high-risk patients require

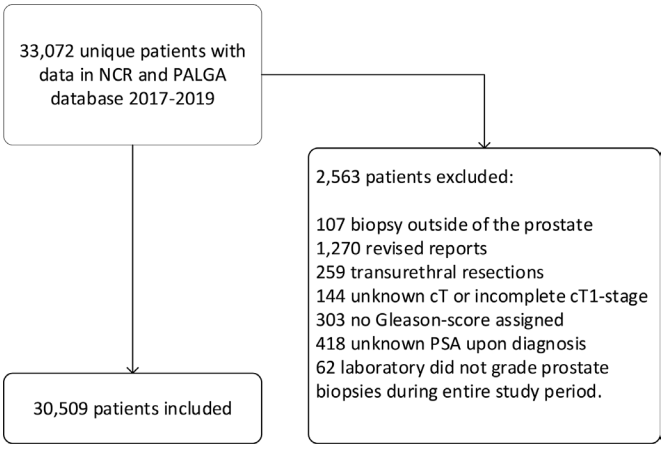

Figure 1 Flowchart with patient inclusion and exclusion criteria. NCR, Netherlands Cancer Registry; PSA, prostate-specific antigen.

immediate curative therapy, with or without pelvic lymph node dissection (PLND). Patients are assigned to a risk category based on a combination of GG, prostate-specific antigen (PSA) value(s), and tumour stage.[14] For a substantial proportion of these patients, GG is the determining factor for risk stratification and choice of treatment.[5]

In order to investigate the impact of interlaboratory variation in GG on treatment choice, we analysed national data from the nationwide network and registry of histopathology and cytopathology in the Netherlands (PALGA) and the Netherlands Cancer Registry (NCR) of over 30 000 patients with PCa.

METHODS

Study population

All pathology reports of prostate biopsies from patients with PCa between 1 January 2017 and 31 December 2019 were extracted from the PALGA database. Subsequently, these reports were matched to clinical data of patients diagnosed with PCa in 2017–2019, from the NCR, hosted by the Netherlands Comprehensive Cancer Organization (IKNL). As pathology reports also contained biopsies from patients primarily diagnosed before 2017, not all pathology reports could be linked to a patient. In the final stage of linking, 88% of all pathology reports could be linked to patients' records in the NCR. All data were pseudonymised by a trusted third party (ZorgTTP, Houten, The Netherlands) and did not contain identifiable patient data. All laboratories gave consent for storage and scientific use of their data in the PALGA database and were anonymised to the researchers. The scientific and privacy committees of PALGA and NCR approved this study. All data were handled in compliance with the General Data Protection Regulation Act.

Overall, we identified 33 072 unique patients with PCa after linking the PALGA and NCR databases. We excluded all patients with reports of biopsies taken outside the prostate (n=107). Furthermore, we excluded patients with second-opinion pathology reports of the same biopsy, as it was unclear whether treatment was based on the original or the revised report (n=1270). Since we were only interested in prostate biopsy pathology report because of the diagnostic and curative intent, we excluded all T1a/T1b tumours (n=259). Patients with missing PSA on diagnosis and cTX stage or unknown Gleason score were excluded, as it would not be possible to select patients with localised PCa for our analyses (n=865).

Finally, we only included laboratories in the analyses that graded biopsies of patients with PCa during all 3 study years, thereby excluding one laboratory with 62 patients. Overall, this resulted in a total population of 30 509 patients with PCa (figure 1).

Main objectives

Our study focused on two main objectives. The first objective was to identify the proportion of patients for whom grade would be the determining factor in localised PCa management. To identify these patients, we used both a strict and a lenient definition. Our strict definition follows current EAU guidelines. The EAU risk stratification identifies patients who are eligible for AS and patients who might benefit from AT (radical prostatectomy or radiotherapy with or without PLND).[5] According to these guidelines, GG is decisive for two groups of patients:

- Group 1 consists of patients with localised PCa and a PSA value <10 ng/mL and cT1c/cT2a stage. For these patients, GG differentiates low-risk from intermediate-risk PCa, as only patients with an ISUP grade 1 tumour are diagnosed with low-risk PCa.
- Group 2 consists of patients with localised PCa and with either a PSA value 10–20 ng/mL and cT stage <cT2c or patients with a cT2b stage and a PSA value <20 ng/mL. For these patients, GG is decisive in distinguishing intermediate-risk from high-risk PCa, as patients with an ISUP grade 4 or 5 tumour are diagnosed with high-risk PCa.

However, clinical T2-substaging (ie, differentiating between T2a, T2b and T2c) is not always considered to be very accurate.[15] Urologists might also consider cT2b/cT2c patients for AS, leading to an underestimation of the number of patients for whom grade is decisive. Therefore, we also applied a more lenient definition, in which we did not distinguish between different cT2 stages. Figure 2 shows a flow diagram of how the different groups were composed, using NCR data.

The second main objective was to evaluate the association between laboratory-specific grading practice and different PCa treatment strategies, while controlling for several patient-related factors. For this analysis, we excluded patients for whom data were incomplete for number of (positive) biopsies (n=2145), or for patients for whom only MRI-guided biopsies were taken (n=4589). Maximum tumour volume percentage contained missing variables for 8881 patients. This is caused by the fact that part of the laboratories use millimetres tumour length, which is not entered into the NCR database. Percentages of missing data per laboratory varied from 2% to 98%. As excluding all cases with missing volume percentages would lead to significant selection bias, excluding half of the laboratories, we decided to include these patients in the model, using the category 'missing'. This leads to a total of 25 920 patients at the basis of the second part of the analysis.

We analysed two different outcome measures for group 1 and group 2. For group 1, we used AT (yes vs no) as outcome measure (ie, radical prostatectomy, radiotherapy or other treatment). As data recorded in the NCR are based on data retrieved by data managers from medical files, in which the distinction between watchful waiting and active surveillance is not consistently used, we could not discriminate between these two entities. We, therefore, assumed that patients with a low-risk or intermediate-risk PCa with no AT were managed by AS. For group 2, we used PLND (yes vs no) as outcome measure. All analyses were performed again using the lenient definition. We presented the

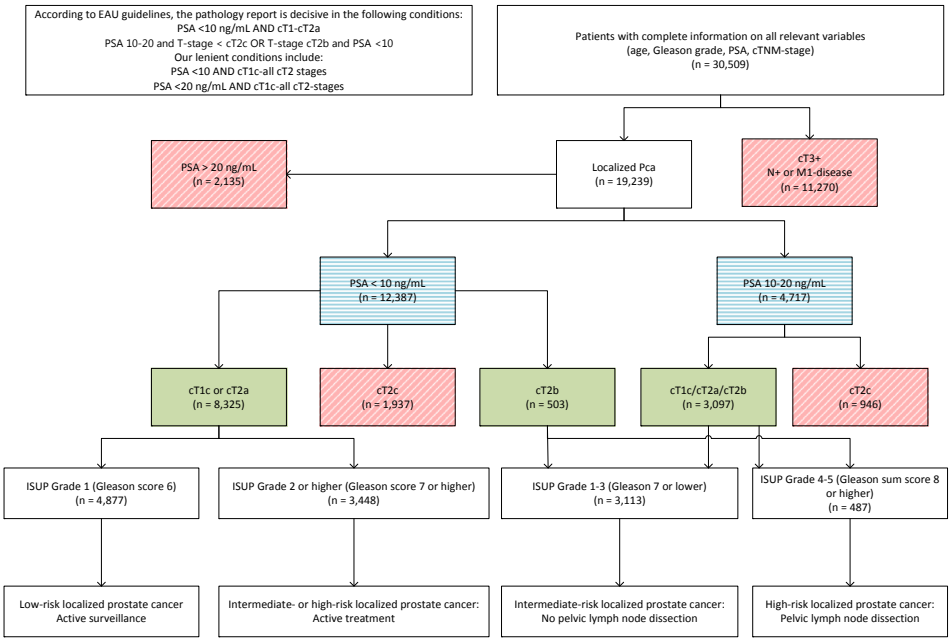

Figure 2 Flowchart with patients whose grade might impact on treatment strategy. The green full-coloured blocks indicate the groups of patients for whom grade is decisive in the EAU risk stratification. The blue horizontally striped blocks indicate the groups of patients for whom grade is decisive according to our lenient definition. The red diagonally striped blocks indicate the groups of patients for whom grade is not decisive. EAU, European Association of Urology; ISUP, International Society of Urological Pathology; PCa, prostate cancer; PSA, prostate-specific antigen.

results of the analyses using the lenient definition in the Supplementary Materials, as results were comparable.

We evaluated the impact of grading practice on AT, by categorising pathology laboratories as low-grading, average-grading or high-grading laboratories. To this end, we calculated the median percentage of ISUP grade 1 and interquartile ranges of ISUP grade 1 (Q1–Q3) of the ISUP grade 1 percentages per laboratory, as described in our previous paper.[13] Low-grading laboratories were defined as laboratories assigning a percentage of ISUP grade 1 higher than the third quartile (Q3). High-grading laboratories were categorised as laboratories assigning a percentage of ISUP grade 1 lower than the first quartile (Q1). Laboratories grading between Q1 and Q3 were categorised as average-grading laboratories. We followed a similar strategy for evaluating the impact on performance of PLND, but we used the percentages of ISUP grades 4 or 5 per laboratory, as grade 4 or higher defines high-risk PCa qualifying for PLND.[5]

Statistical analyses

We summarised patient, tumour and treatment characteristics using counts and proportions, means and SD as appropriate. We tested for differences between the different treatment groups in patient and tumour characteristics, using Mann-Whitney U test and χ^2 test as appropriate.

For evaluating the impact of interlaboratory variation on treatment strategy, we performed a multivariable logistic regression. We corrected for *a priori*-selected case-mix variables age, PSA, cT-stage, maximum tumour volume percentage, number of negative biopsies and number of positive biopsies. Adjusted ORs (aORs) and 95% CIs were calculated for AT versus no AT and PLND versus no PLND. The average-grading laboratories were used as a reference. *P* values <0.05 were considered statistically significant. All statistical analyses were performed using R V.4.0.3.[16]

RESULTS

Characteristics

Characteristics of the 30 509 included patients with PCa are shown in table 1. The overall mean age was 70 years and the largest group of patients had a cT1c tumour (31.9%). 19 239 patients were diagnosed with localised PCa, representing the large majority (63.1%). Lymph node metastases and distant metastases were present in 17.4%, respectively. Almost half of the patients had PSA values on diagnosis lower than 10 ng/mL (48.1%). AT was applied in roughly three quarters of all patients (76.1%), which mostly consisted of a radical prostatectomy (23.1%), hormonal and radiotherapy (20.3%) or hormonal therapy with or without additional chemotherapy (19.4%). PLND was performed in 4577 patients, representing 15% of all patients.

Risk group stratification

We identified 8325 (27.3%) and 12 387 (40.6%) patients for whom GG was determining in distinguishing low-risk from intermediate-risk PCa, according to our strict and lenient definitions, respectively (figure 2 and online supplemental table 1). For differentiating between intermediate-risk and high-risk PCa, we identified 3358 (11.0%) and 4694 (15.4%) patients for whom grade was the determining factor according to the strict and lenient definitions, respectively.

Active therapy versus no active therapy

For our multivariable logistic regression, we included 6818 patients in group 1 and 2907 patients in group 2. Roughly half of the patients in group 1 did not receive AT (50.9%) (table 2).

Patients diagnosed in low-grading laboratories received AT significantly less often compared with patients diagnosed in average-grading laboratories (tables 2 and 3). Patients in high-grading laboratories received AT significantly more often than

Table 1 Patient characteristics of a Dutch nationwide cohort of prostate cancer patients diagnosed with prostate biopsy between 2017 and 2019

Characteristics N, (%)	Total (N=30 509) (%)
Age (years), median (Q1–Q3)	70 (65–75)
Localised prostate cancer	19 239 (63,1)
T-stage	
cT1c	9737 (31.9)
cT2 unspecified	3179 (10.4)
cT2a	3228 (10.6)
cT2b	994 (3.3)
cT2c	4539 (14.9)
cT3a or higher	8835 (20.5)
Lymph node status	
cN0/Nx	25 194 (82.6)
cN1	5315 (17.4)
Metastases, N (%)	
cM0	25 189 (82.6)
cM1	5320 (17.4)
PSA (ng/mL)	
<10	14 668 (48.1)
10–20	6656 (21.8)
>20	9185 (30.1)
ISUP grade group (Gleason Score)	
1 (3+3)	9235 (30.3)
2 (3+4)	7017 (23.0)
3 (4+3)	4387 (14.4)
4 (8)	4335 (14.2)
5 (9–10)	5535 (18.1)
Primary treatment	
No active treatment	7298 (23.9)
Radical prostatectomy	6999 (23.0)
EBRT or brachytherapy	4003 (13.2)
EBRT and ADT*	6144 (20.2)
ADT alone or combined with chemotherapy	5894 (19.3)
Other	41 (0.1)
Pelvic lymph node dissection	
Yes	4577 (15.0)
No	25 932 (85.0)

*This group also contains patients with M1-disease.
ADT, androgen deprivation therapy; EBRT, external beam radiotherapy; ISUP, International Society of Urological Pathology; PSA, prostate-specific antigen.

those in average-grading laboratories. Over half of the patients were graded in either low-grading or high-grading laboratories (37.6% and 17.1%, respectively). Adjusted ORs for AT of low-grading and high-grading laboratories were 0.77 (95% CI 0.68 to 0.88) and 1.21 (95% CI 1.03 to 1.43), respectively, as compared with the reference of average-grading laboratories (table 3). The results for the lenient models were comparable to the strict model. These results are presented in the online supplemental table 1 and online supplemental table 2.

PLND versus no PLND

For PLND, the aORs showed a significant association for low-grading laboratories versus average-grading laboratories (tables 2 and 3). Patients from low-grading laboratories underwent significantly less often PLNDs than patients from average-grading laboratories (aORs 0.66 (95% CI 0.48 to 0.90) and 0.72 (95% CI 0.56 to 0.91) for the strict and lenient definitions, respectively). The frequency of PLNDs in high-grading laboratories was not statistically different from average-grading laboratories (aOR 0.92 (95% CI 0.68 to 1.25) (figure 3).

DISCUSSION

This nationwide study evaluated the impact of interlaboratory variation in grading of PCa on treatment in over 30 000 patients, among 40 Dutch laboratories.[13] It shows the importance of consistency in GG for clinical decision-making, as the chance that a patient undergoes AT or PLND, depends on laboratory grading practice in a substantial number of patients. As this likely influences patient prognosis and outcome, standardisation of GG is necessary to prevent suboptimal patient outcome.

We hypothesised that patients whose biopsies were assessed in higher grading laboratories were more likely to receive more aggressive treatment than those assessed in lower grading laboratories. We focused on patients for whom grade would be decisive in the EAU risk stratification.[5] We also provided a lenient model (including all cT2 stadia rather than just cT2a or cT2b) to describe patients whose treatment could alter by a different grade. Both models found a significant and clinically relevant association between receiving active therapy and higher grading laboratories for patients with low-risk or intermediate-risk PCa. For patients with intermediate-risk or high-risk PCa, however, only a significant association between PLND and low versus average grading practice existed.

GG would be decisive in a patient's risk stratification and subsequently their treatment strategy for 39.1% of all patients with PCa between 2017 and 2019 when applying a strict definition and for 56.1% of all patients when applying a lenient definition. The association of grading variation with treatment choice was most profound in patients whose grade would differentiate between low-risk or intermediate-risk PCa and subsequently their eligibility for either AS or AT. For patients with intermediate-risk or high-risk PCa, the impact of the grading practice was limited and not consistent. Only laboratories with a low proportion of ISUP grade 4 or grade 5 PCa had patients undergoing PLND significantly less often than average-grading laboratories, whereas we found no significant difference for high-grading versus average-grading laboratories nor between high-grading and low-grading laboratories.

Several possible explanations exist for this ambiguous result for PLND performance. First, it is possible that guideline adherence of PLND performance varies between hospitals. For example, a previous Dutch PCa cohort study (ProZIB; Dutch Acronym for Insight into Prostate Cancer Care) showed variation in guideline adherence between hospitals for appliance of radiotherapy and hormone therapy for both intermediate-risk and high-risk patients between hospitals, and to a lesser extent also for appliance of AS for low-risk patients with PCa.[17 18] It is possible that some surgeons, and, therefore, some centres, use lower or higher thresholds for performing PLND's. However, the vast majority of patients are discussed in multidisciplinary tumorboard meetings. Therefore, the influence of individual clinicians is probably limited. Also, the number of patients in the intermediate-risk or high-risk group is relatively low, potentially influencing the results.

Second, due to the retrospective nature of the data, we did not have information on of the specific type of imaging devices, which might have been a contributing factor in the decision to perform a PLND or not.

Table 2 Characteristics of group 1 and group 2—patients whose risk stratification depends on Gleason Grade

Group 1 strict definition*				
	Total, N=6818 (%)	**AT, N=3,345 (49.1)**	**no AT, N=3,473 (50.9)**	**P value¶**
Age, years (median (Q1–Q3))	68 (63–72)	68 (63–72)	68 (63–73)	0.01
PSA (ng/mL), median (Q1–Q3)	6.7 (5.3–8.1)	6,7 (5.4–8.2)	6.6 (5.2–8.0)	0.002
Clinical T-stage, N (%)				<0.001
cT1c	5283 (77.5)	2363 (70.6)	2920 (84.1)	
cT2a	1535 (22.5)	(982 (29.4)	553 (15.9)	
Number of negative biopsies, median (Q1–Q3)	7 (5–9)	6 (4–8)	8 (7–10)	<0.001
Number of positive biopsies, median (Q1–Q3)	2 (1–4)	4 (2–5)	2 (1–2)	<0.001
Maximum tumour volume percentage, N (%)				<0.001
<5% or less	1315 (19.3)	235 (7.0)	1080 (31.1)	
6%–25%	1697 (24.9)	782 (23.4)	915 (26.3)	
26%–50%	971 (14.2)	698 (20.9)	273 (7.9)	
51% or more	731 (10.7)	607 (18.1)	124 (3.6)	
Missing	2104 (30.9)	1023 (30.6)	1081 (31.1)	
Grading practice laboratory†, N (%)				<0.001
20 average-grading labs	3088 (45.3)	1541 (46.1)	1547 (44.5)	
10 low-grading labs	2562 (37.6)	1158 (34.6)	1404 (40.4)	
10 high-grading labs	1168 (17.1)	646 (19.3)	522 (15.0)	
Group 2 strict definition‡				
	Total, N=2907	**PLND, N=412**	**No PLND, N=2495**	
Age, years (median (Q1–Q3))	71 (66–75)	69 (64–72)	71 (66–76)	<0.001
PSA (ng/mL), median (Q1–Q3)	12.0 (10.9–14.7)	12.0 (10.0–15.0)	12.1 (11.0–14.7)	0.01
Clinical T-stage, N (%)				<0.001
cT1c	1810 (62.3)	165 (40.0)	1645 (65.9)	
cT2a	505 (17.4)	85 (20.6)	420 (16.8)	
cT2b	592 (20.4)	162 (39.3)	430 (17.2)	
Number of negative biopsies, median (Q1–Q3)	7 (5–9)	5 (4–7)	7 (5–9)	<0.001
Number of positive biopsies, median (Q1–Q3)	3 (2–5)	4 (3–6)	3 (1–4)	<0.001
Maximum tumour volume percentage, N (%)				<0.001
<5% or less	406 (14.0)	20 (4.8)	386 (15.5)	
6%–25%	609 (20.9)	51 (12.4)	558 (22.4)	
26%–50%	452 (15.5)	79 (19.2)	373 (14.9)	
51% or more	521 (17.9)	155 (37.6)	366 (14.7)	
Missing	919 (31.6)	107 (26.0)	812 (32.5)	
Grading practice laboratory§, N (%)				0.03
20 average-grading labs	1795 (61.7)	275 (66.7)	1520 (60.9)	
10 low-grading labs	597 (20.5)	66 (16.0)	531 (21.3)	
10 high-grading labs	515 (17.7)	71 (17.2)	444 (17.8)	

P=significant at<0.05.
*Group 1—PSA <10AND cT1c or cT2a
†Group 1 grading practice: average grading=laboratories' percentage of ISUP 1 falls within Q1–Q3 of national median, low-grading: laboratories' percentage of ISUP 1>Q3 of national median, high-grading=laboratories' percentage of ISUP 1<Q1 of national median.
‡Group 2 PSA 10–20 and T-stage <cT2 c OR T-stage cT2b and PSA <10.
§Group 2 grading practice: average grading: laboratories' percentage of ISUP 4/5 falls within Q1–Q3 of national median, low-grading: laboratories' percentage of ISUP 4/5<Q1 of national median, high grading: laboratories' percentage of ISUP 4/5>Q3 of national median.
¶Mann-Whitney U-test and χ^2 test as appropriate.
AT, active therapy; ISUP, International Society of Urological Pathology; PLND, pelvic lymph node dissection; PSA, prostate-specific antigen.

Moreover, many centres use the Memorial Sloan Kettering Cancer Center (MSKCC) or extent Briganti models for the decision to perform a PLND. We considered using the Briganti and MSKCC models as well for our analysis on PLND, as these models calculate the risk of lymph node involvement. Unfortunately, relevant variables (number of biopsies and volume percentage) were not consistently reported in our data set, hence, strictly applying these definitions was not possible.[19 20] It is possible that patient groups based on these models would have been slightly different from the patient groups formed by the EAU-risk stratification, which might also attribute to the ambiguous results for PLND versus no PLND. However, our results are robust in the AS versus AT group in both the strict model and the lenient model, and we expect any impact to be relatively small.

In total, 10% of tumours were labelled with an unspecified cT2 stage. Historically, clinical T stage by digital rectal examination (DRE) and transrectal ultrasound is known to be relatively inaccurate and inconsistent in up to 35.4%.[15 21] Unfortunately, it is largely unknown whether MRI or DRE was performed for

Table 3 Results of the multivariable logistic regression models for active treatment and pelvic lymph node dissection

Active treatment*	OR (CI)	P value
Grading practice laboratory†		
Average grading	ref	ref
Low grading	0.77 (0.68 to 0.88)	*<0.001*
High grading	1.21 (1.03 to 1.43)	*0.02*
Age	0.97 (0.96 to 0.98)	*<0.001*
PSA (ng/mL)	1.06 (1.02 to 1.09)	*<0.001*
Clinical T-stage		
cT1c	ref	ref
cT2a	2.33 (2.02 to 2.68)	*<0.001*
Number of negative biopsies	0.93 (0.90 to 0.95)	*<0.001*
Number of positive biopsies	1.79 (1.71 to 1.88)	*<0.001*
Maximum tumour volume percentage		
<5% or less	ref	
6%–25%	2.18 (1.81 to 2.63)	*<0.001*
26%–50%	3.98 (3.19 to 4.98)	*<0.001*
51% or more	5.25 (4.00 to 6.91)	*<0.001*
Missing	2.37 (1.97 to 2.86)	*<0.001*
Pelvic lymph node dissection‡	*OR (CI)*	*P-value*
Grading practice laboratory§		
Average grading	ref	ref
Low grading	0.66 (0.48 to 0.90)	*0.01*
High grading	0.92 (0.68 to 1.25)	0.6
Age	0.93 (0.92 to 0.95)	*<0.001*
PSA (ng/mL)	1.07 (1.03 to 1.11)	*<0.001*
Clinical T-stage		
cT1c	ref	ref
cT2a	2.12 (1.57 to 2.86)	*<0.001*
cT2b	3.48 (2.55 to 4.76)	*<0.001*
Number of negative biopsies	0.96 (0.92 to 1.01)	0.1
Number of positive biopsies	1.15 (1.08 to 1.23)	*0.001*
Maximum tumour volume percentage		
<5% or less	ref	
6%–25%	1.37 (0.80 to 2.41)	0.3
26%–50%	2.29 (1.35 to 4.03)	*0.003*
51% or more	3.51 (2.09 to 6.16)	*<0.001*
Missing	1.73 (1.05 to 2.98)	*0.04*

p=significant at<0.05
*Group 1—PSA <10AND cT1c or cT2a.
†Group 1 grading practice: average grading=laboratories' percentage of ISUP grade 1 falls within Q1–Q3 of national median, low grading: laboratories' percentage of ISUP grade 1>Q3 of national median, high grading=laboratories' percentage of ISUP grade 1<Q1 of national median.
‡Group 2 - PSA 10–20 and cT1c, cT2a or cT2b stadium.
§Group 2 grading practice: average grading: laboratories' percentage of ISUP grade 4/5 falls within Q1–Q3 of national median, low-grading: laboratories' percentage of ISUP grade 4/5<Q1 of national median, high grading: laboratories' percentage of ISUP grade 4/5>Q3 of national median.
ISUP, International Society of Urological Pathology; PSA, prostate-specific antigen; ref, reference category.

clinical staging. Uptake varies per centre, and using MRI prebiopsy can lead to less lower GG cases, as these would probably not be biopsied.[22] The strong positive association between a more advanced clinical T2 stage and AT is in line with the advice of the national and EAU guidelines or MSKCC model, making a distinction between cT2a and cT2b.[5 19] Interestingly, the difference in ORs between cT2c and cT2a was smaller than for cT2b and cT2a but could potentially be attributed to the relatively small group of cT2b tumours. Yet, the 10% of cT2-unknown understaged patients might signal a different practice, or be due to suboptimal reporting.

Our study shows comparable or higher uptake of AS, as compared with previous nationwide studies in AS uptake. Uptake of AS was slightly higher than in the 2015–2016 ProZIB-cohort, where up to 70% of the patients were treated with AS. Our results are comparable to the 74% uptake of AS in a Swedish cohort study.[18 23]

Our results implicate that during our study period, at least one out of five patients (3730 of 19 239) with localised PCa might have received a different treatment strategy if their sample had been graded in an average-grading laboratory, when focusing on AS vs AT. Within this group of patients, 2562 (13%) patients would have been more likely to have received AT, and 1168(6%) might have been eligible for AS. We used the average-grading laboratories, which may be considered arbitrary, as these laboratories are no gold standard, but it was considered the best reference. It is not possible to say which laboratory grades erroneously, and, therefore, which patients would have received undertreatment or overtreatment. However, the difference in grading practice that leads to a difference in treatment strategies is evident. In theory, it is possible that socioeconomic factors affect the grading practice in the Netherlands, but due to the high hospital density (in some regions up to 15 different hospitals in a 20 km radius), this is unlikely.[24] Unfortunately, we cannot indicate on individual patient level whose treatment was affected by grading variation, as we could not perform revisions of the needle biopsies, due to feasibility reasons. However, the nationwide design offered the possibility to analyse a more general trend. An important benefit of our approach is that these results apply to daily clinical practice.

Our results show that interlaboratory grading variation leads to variation in treatment. While undertreatment generally leads to poorer cancer survival, overtreatment induces treatment toxicity. The ProtecT study group measured both patient-reported outcome and patient survival and morbidity for AT modalities and AS. They showed that patient-reported outcome measures on bowel, sexual and urinary function were much better for patients with AS than with AT.[25 26] Conversely, AT seemed to be associated with lower PCa mortality than AS, but numbers of deaths were very low in both groups.[27] However, when the ProtecT study was performed, diagnostics of prostate cancer did not yet involve transrectal ultrasound (TRUS) or MRI-aimed biopsies, which aid in reducing sampling error and subsequently under diagnosis of the prostate biopsies. Also, the ProtecT-study randomised patients with PSA up to 20 ng/mL, which is not common practice according to the current guidelines. The current AS programmes might, therefore, have even better survival outcomes than those in the ProtecT-study. It is, therefore, important to correctly select patients for whom AS might be an option in order to reduce treatment-specific morbidity. Whether current grading variation practices also influence patient survival and morbidity may only be concluded after a longer period of follow-up, but it is not unlikely.

The current EAU guidelines state no recommendations regarding the use of re-evaluations of biopsies. In Dutch PCa practice, re-evaluations are rarely performed, as only 1262 reports (<4% of all reports, data not shown) were marked as a re-evaluation report. This has not improved after 2015–2016, as in the ProZIB cohort, 3% of all patients with PCa received a re-evaluation.[28] Especially for patients for whom grade could be the determining factor, a re-evaluation could have implications for both treatment strategy and patient outcome. Both Kuijpers

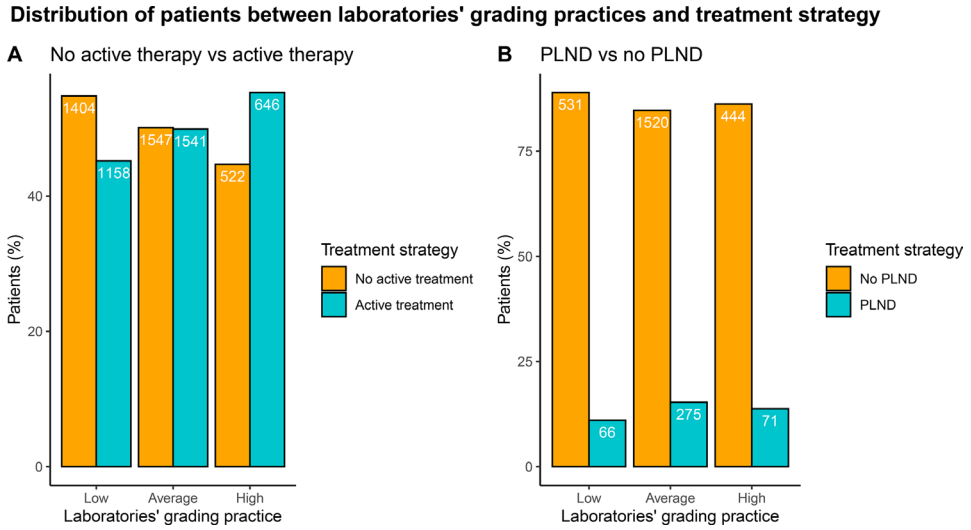

Figure 3 Distribution of patients between laboratories' grading practices and treatment strategy. (A) Bar chart for active surveillance uptake for patients with a cT1c/cT2a-tumour and a PSA <10 ng/mL for active treatment vs no active treatment. (B) Bar chart for PLND uptake for patients with a cT2b-tumour and a PSA <20 on diagnosis, or patients with a PSA between 10 and 20 ng/mL on diagnosis and cT-stage <cT2 c. Low: 10 laboratories grading >Q3 of % ISUP Grade 1 for active treatment or <Q1 ISUP Grade 4/5 for PLND. Average: 20 laboratories grading Q1-Q3 of % ISUP Grade 1 for active treatment or Q1-Q3 of ISUP Grade 4/5 for PLND. High: 10 laboratories grading <Q1 of % ISUP Grade 1 for active treatment or >Q3 of % ISUP Grade 4/5 for PLND. Absolute numbers of patients are displayed in white in the top of each bar. ISUP, International Society of Urological Pathology; PLND, pelvic lymph node dissection; PSA, prostate-specific antigen.

et al and Van Santvoort *et al* hypothesised that 4%–8% of patients in the Netherlands might receive a different treatment strategy after re-evaluation.[28 29]

Other interventions to reduce variation in GG should be investigated as well. Feedback reports and e-learning modules were successfully introduced for reducing grading variation for invasive breast cancer, ductal carcinoma in situ of the breast and colorectal adenomas.[30–32] Furthermore, artificial intelligence systems seem promising in reducing grading variation between pathologists, but currently lack sufficient validation or implementation in daily pathology practice.[33–35]

In conclusion, this study is, to our knowledge, the first to signal clinical implications of interlaboratory variation in PCa grading. The effect is most profound for patients whose GG will determine whether low-risk or intermediate-risk prostate cancer is diagnosed. This affects the choice between AT and AS. It is, therefore, likely that patient outcome is affected by inter-laboratory grading variation. Future work should focus on reducing this variation.

Take home messages

⇒ Interlaboratory variation in GG affects clinical decision-making for 39.1% of all patients with prostate cancer.
⇒ Patients in different institutions are more or less likely to undergo AT or PLND based on the laboratories' grading practice.
⇒ This likely affects patient outcome.

Handling editor Dhirendra Govender.

Contributors RNF: conceptualisation, methodology, formal analysis, investigation, data curation, writing original draft, visualisation. CvD: conceptualisation, writing—review and editing. KKHA: methodology, resources, writing—review and editing. BBMS: conceptualisation, writing—review and editing, funding acquisition. P-PMW: conceptualisation, writing—review and editing. PJvD: writing—review and editing, funding acquisition, guarantor. RPM: conceptualisation, writing—review and editing, funding acquisition, supervision.

Funding This research was funded by Quality Foundation of the Dutch Association of Medical Specialists (SKMS), Astellas Pharma BV and Pfizer BV. The funding sources had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing interests PJvD received research grant from Quality Foundation of the Dutch Associaton of Medical Specialists (SKMS). RPM received research grant from Astellas Pharma B.V. BBMS received research grant from Pfizer BV. All other authors declare no conflicts of interests.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. Data are available upon reasonable request at PALGA and the NCR.

ORCID iDs
Rachel N Flach http://orcid.org/0000-0002-4448-2895
Britt B M Suelmann http://orcid.org/0000-0002-7155-3337
Paul J van Diest http://orcid.org/0000-0003-0658-2745

REFERENCES

1 Ferlay J, Colombet M, Soerjomataram I, *et al*. Cancer incidence and mortality patterns in Europe: estimates for 40 countries and 25 major cancers in 2018. *Eur J Cancer* 2018;103:356–87.
2 Cijfers over kanker. Netherlands Cancercancer Registryregistry supplied by IKNL, 2020. Available: www.cijfersoverkanker.nl [Accessed 25 May 2020].
3 Pierorazio PM, Walsh PC, Partin AW, *et al*. Prognostic Gleason grade grouping: data based on the modified Gleason scoring system. *BJU Int* 2013;111:753–60.
4 Joniau S, Briganti A, Gontero P, *et al*. Stratification of high-risk prostate cancer into prognostic categories: a European multi-institutional study. *Eur Urol* 2015;67:157–64.
5 Mottet N, van den Bergh RCN, Briers E, *et al*. EAU-EANM-ESTRO-ESUR-SIOG guidelines on prostate Cancer-2020 update. Part 1: screening, diagnosis, and local treatment with curative intent. *Eur Urol* 2021;79:243–62.

6 Mikami Y, Manabe T, Epstein JI, *et al*. Accuracy of gleason grading by practicing pathologists and the impact of education on improving agreement. *Hum Pathol* 2003;34:658–65.
7 Allsbrook WC, Mangold KA, Johnson MH, *et al*. Interobserver reproducibility of gleason grading of prostatic carcinoma: general pathologist. *Hum Pathol* 2001;32:81–8.
8 Allsbrook WC, Mangold KA, Johnson MH, *et al*. Interobserver reproducibility of gleason grading of prostatic carcinoma: urologic pathologists. *Hum Pathol* 2001;32:74–80.
9 Oyama T, Allsbrook WC, Kurokawa K, *et al*. A comparison of interobserver reproducibility of gleason grading of prostatic carcinoma in Japan and the United States. *Arch Pathol Lab Med* 2005;129:1004–10.
10 McKenney JK, Simko J, Bonham M, *et al*. The potential impact of reproducibility of gleason grading in men with early stage prostate cancer managed by active surveillance: a multi-institutional study. *Journal of Urology* 2011;186:465–9.
11 Epstein JI, Egevad L, Amin MB. International Society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma definition of grading patterns and proposal for a new grading system. *Am J Surg Pathol* 2014;40:244–52.
12 Ozkan TA, Eruyar AT, Cebeci OO, *et al*. Interobserver variability in gleason histological grading of prostate cancer. *Scand J Urol* 2016;50:420–4.
13 Flach RN, Willemse P-PM, Suelmann BBM, *et al*. Significant inter- and intralaboratory variation in gleason grading of prostate cancer: a nationwide study of 35,258 patients in the Netherlands. *Cancers* 2021;13:5378.
14 Mottet N, Bellmunt J, Bolla M, *et al*. EAU-ESTRO-SIOG guidelines on prostate cancer. Part 1: screening, diagnosis, and local treatment with curative intent. *Eur Urol* 2017;71:618–29.
15 Reese AC, Sadetsky N, Carroll PR, *et al*. Inaccuracies in assignment of clinical stage for localized prostate cancer. *Cancer* 2011;117:283–9.
16 R Core Team. R: a language and environment for statistical computing, Published online 2020. Available: https://www.r-project.org/
17 Rijksen BLT, Pos FJ, Hulshof MCCM, *et al*. Variation in the prescription of androgen deprivation therapy in intermediate- and high-risk prostate cancer patients treated with radiotherapy in the Netherlands, and adherence to European association of urology guidelines: a population-based study. *Eur Urol Focus* 2021;7:332–9.
18 Jansen H, van Oort IM, van Andel G, *et al*. Immediate treatment vs. active-surveillance in very-low-risk prostate cancer: the role of patient-, tumour-, and hospital-related factors. *Prostate Cancer Prostatic Dis* 2019;22:337–43.
19 Prostate Cancer Nomograms. Dynamic prostate cancer nomogram: coefficients | Memorial Sloan Kettering cancer center. Available: https://www.mskcc.org/nomograms/prostate/pre_op/coefficients [Accessed 01 July 2021].
20 Briganti A, Larcher A, Abdollah F, *et al*. Updated nomogram predicting lymph node invasion in patients with prostate cancer undergoing extended pelvic lymph node dissection: the essential importance of percentage of positive cores. *Eur Urol* 2012;61:480–7.
21 Gosselaar C, Kranse R, Roobol MJ, *et al*. The interobserver variability of digital rectal examination in a large randomized trial for the screening of prostate cancer. *Prostate* 2008;68:985–93.
22 Siddiqui MM, Rais-Bahrami S, Turkbey B, *et al*. Comparison of MR/ultrasound fusion-guided biopsy with ultrasound-guided biopsy for the diagnosis of prostate cancer. *JAMA* 2015;313:390–7.
23 Loeb S, Folkvaljon Y, Curnyn C, *et al*. Uptake of active surveillance for very-low-risk prostate cancer in Sweden. *JAMA Oncol* 2017;3:1393–8.
24 Centraal bureau voor Statistiek. Regionale kerncijfers Nederland, 2022. Available: https://opendata.cbs.nl/statline/#/CBS/nl/dataset/70072ned/table?ts=1649768502110 [Accessed 12 Apr 2022].
25 Hamdy FC, Donovan JL, Lane JA, *et al*. 10-Year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer. *N Engl J Med* 2016;375:1415–24.
26 Donovan JL, Hamdy FC, Lane JA. Patient-Reported outcomes after monitoring, surgery,or radiotherapy for prostate cancer. *N Engl J Med*.
27 Neal DE, Metcalfe C, Donovan JL, *et al*. Ten-Year mortality, disease progression, and treatment-related side effects in men with localised prostate cancer from the protect randomised controlled trial according to treatment received. *Eur Urol* 2020;77:320–30.
28 van Santvoort BWH, van Leenders GJLH, Kiemeney LA, *et al*. Histopathological re-evaluations of biopsies in prostate cancer: a nationwide observational study. *Scand J Urol* 2020;54:463–9.
29 Kuijpers CCHJ, Burger G, Al-Janabi S, *et al*. Improved quality of patient care through routine second review of histopathology specimens prior to multidisciplinary meetings. *J Clin Pathol* 2016;69:866–71.
30 van Dooijeweert C, van Diest PJ, Baas IO, *et al*. Variation in breast cancer grading: the effect of creating awareness through laboratory-specific and pathologist-specific feedback reports in 16 734 patients with breast cancer. *J Clin Pathol* 2020;73:793–9.
31 van Dooijeweert C, Deckers IAG, de Ruiter EJ, *et al*. The effect of an e-learning module on grading variation of (pre)malignant breast lesions. *Mod Pathol* 2020;33:1961–7.
32 Madani A, Kuijpers CCHJ, Sluijter CE, *et al*. Decrease of variation in the grading of dysplasia in colorectal adenomas with a national e-learning module. *Histopathology* 2019;74:925–32.
33 Raciti P, Sue J, Ceballos R, *et al*. Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies. *Mod Pathol* 2020;33:2058–66.
34 Bulten W, Pinckaers H, van Boven H, *et al*. Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol* 2020;21:233–41.
35 Ström P, Kartasalo K, Olsson H, *et al*. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol* 2020;21:222–32.