Abdominal Radiology (2019) 44:2535–2544
https://doi.org/10.1007/s00261-019-01970-z

PELVIS

Precision and accuracy of magnetic resonance imaging for lobar classification of benign prostatic hyperplasia

Neil F. Wasserman[1,2] · Eric Niendorf[1] · Benjamin Spilseth[1]

Published online: 30 March 2019

Abstract

Purpose To validate the application of a magnetic resonance imaging (MRI)-based lobar classification of benign prostatic hyperplasia (BPH) for use in research and clinical management.

Methods Two radiologists with 5 and 11 years post-fellowship experience were trained in the lobar classification of BPH using an internally developed atlas of prostate anatomy with example MRI images edited by a third senior radiologist designated as the "administrator" of the study. A study population of 140 patients referred to a tertiary academic medical center with known or suspected prostate cancer was selected by the administrator to test the interrater reliability (IRR; precision) of the classification as well as accuracy of the two readers compared to the administrator as the "gold" standard. The intrarater reliability of repeat readings of the administrator was also examined. Percentage of agreement, proportion of agreement, and Cohen's κ were applied. This was a retrospective IRB-approved study.

Results IRR (precision) between the two interpreting radiologists was 64% agreement, corresponding to unweighted κ of 0.52. Composite proportion of agreement across all BPH types (categories) for interpreting radiologists was 0.67. Observer accuracy was 62% agreement, unweighted κ 0.49, for observer 1 and 67%, unweighted κ 0.58, for observer 2. Intrarater reliability for the administrator was 87% agreement, unweighted κ 0.81 with composite proportion of agreement across all categories of 0.87.

Conclusions MRI lobar classification of BPH is a reproducible and reliable tool for research and clinical applications.

Keywords Prostate · BPH · Lobar classification

Introduction

Benign prostatic hyperplasia (BPH) as a cause of lower urinary tract symptoms (LUTS) affects more than 20% of American men aged 30-79 years or roughly 15 million men [1]. This prevalence appears to increase with the increasing age, as approximately 80% of men are affected by BPH/LUTS by 70 years of age [2].

The annual cost of treatment for BPH in the US alone is estimated to be $1 billion annually [3]. BPH/LUTS is the second-most common reason for being seen by an urologist behind urinary tract infection, and the most common for males accounting for more than twice the number of diagnoses than prostatic cancer [4].

The variety of therapeutic options available for management of this disorder is complicated by the fact that, pathophysiologically, BPH is not one disorder, and a single management choice will not be successful in all patients with BPH/LUTS. The First International Consultation of BPH recommended "future studies should seek determinants, or predictors of BPH parameters," calling for "larger studies with adequate power for subgroup analysis" [5]. The most common method to stratify BPH for both research and clinical management is by total prostatic weight as a determinant for the surgical decision to perform a simple prostatectomy or holmium laser resection if larger than 80 g, or transurethral resection or minimally invasive surgery if smaller [6]. More recently, classification by weight over 40–50 g versus under was used to decide whether to treat medically with 5-alpha reductase inhibitors (5-ARIs) for larger glands or alpha-1-receptor blockers for smaller

✉ Neil F. Wasserman
wasse001@umn.edu

[1] Department of Radiology, University of Minnesota Medical School, Mayo Mail Code 292, 420 Delaware Street S.E., Minneapolis, MN 55455, USA

[2] Minneapolis, USA

ones [6, 7]. This classification was based on the supposition that larger glands were primarily caused by growth of glandular areas resulting in mechanical obstruction of the urethra by bulk, while it was presumed that patients with LUTS who had smaller or no enlargement were affected by muscular (dynamic) mechanisms. Randall [8] was the first to propose gross lobar classification of the prostate followed later by an analogous classification based on transrectal ultrasound (TRUS) and magnetic resonance imaging (MRI) [9, 10]. The latter investigators emphasize a continuing need to find phenotypes or genotypes of BPH in order to tailor treatment with greater precision.

There has been recent interest in developing and utilizing a lobar classification for BPH in assessing results of therapeutic strategies [11]. An illustrated summary of this classification can be found in Fig. 1a–g. Some have already adopted this classification in clinical studies [12, 13], but its validity as an instrument for use to measure the effects of lower urinary tract obstruction has not been tested. Before widely applying the classification for research or clinical use, it is essential to assess its level of inter- and intraobserver reproducibility (precision) and accuracy (validity) [14].

This lobar classification system has the potential to add another useful tool in the assessment, treatment, and

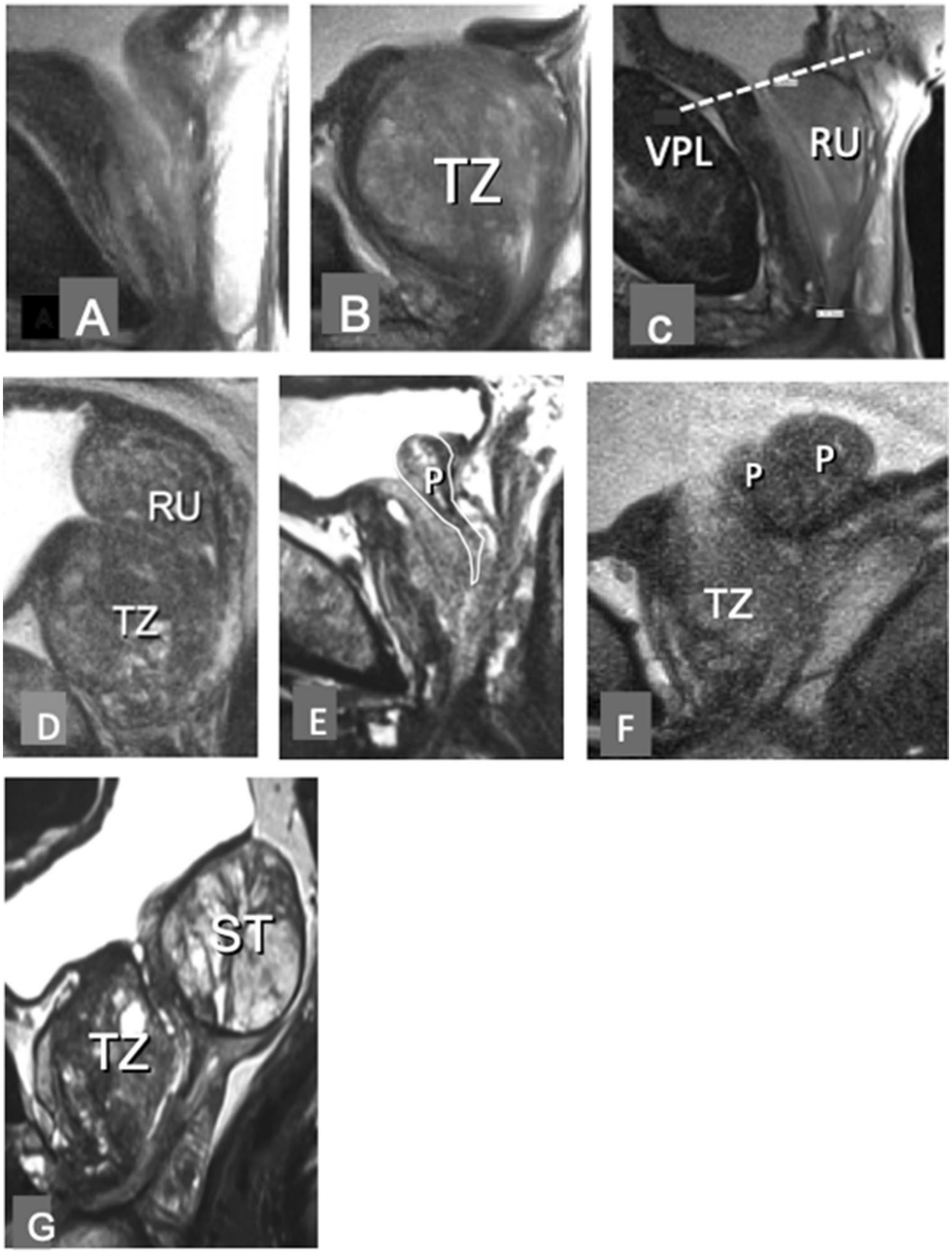


Fig. 1 Lobar classification of BPH. **a** Type 0 (normal); **b** Type 1 = TZ (bilateral transition zone) enlargement; **c** Type 2 = RU (retrourethral lobe) enlargement, dashed line is VPL; **d** Type 3 = TZ plus RU enlargement; **e** Type 5 = Type 4 plus Type 1 and/or Type 2, or Type 6; **f** Type 6 = subtrigonal (ST) or other ectopic; and **g** Type 7 = Type 6 plus any other type

monitoring of BPH. In order to proceed with confidence, researchers and clinicians must know that the classification is reproducible. Therefore, we hypothesize that the MRI lobar classification of BPH can be performed at a level of intra- and interobserver reliability comparable to other clinical assessment classifications, such as the Prostate Imaging Reporting and Data System (PIRADS) and Gleason scoring, so that it may to be used in the research and clinical setting.

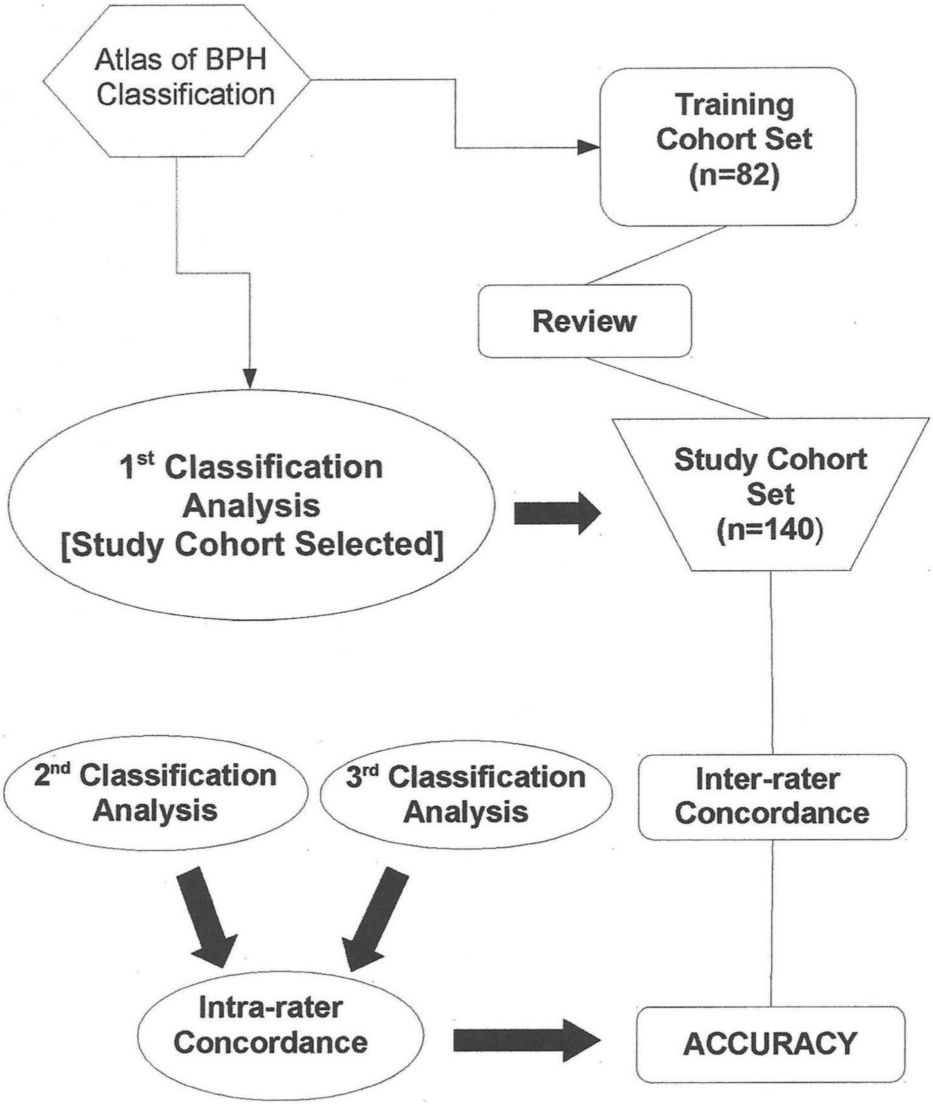
Materials and methods

Study design and population

This IRB approved retrospective study consisted of a population selected by a study administrator (NW) who is a board-certified radiologist with 40 years of experience studying prostate anatomy and disorders. 1500 Consecutive separate patients with MRIs obtained at a single institution between 10 January 2010 and 23 August 2016 were reviewed. The MRIs were originally obtained for the diagnosis or staging of known or suspected prostate cancer. 140 subjects, with mean age of 64 (31–82), were selected from the 1500 reviewed patients to constitute our test cohort satisfying statistical power analysis. Patients with a history of prior prostate surgery or radiation were excluded, as were examinations with very poor quality T2-weighted images (T2WIs) and those in which cancer precluded evaluation of image boundaries, as determined by the study administrator.

The overall experimental algorithm is illustrated in Fig. 2. The administrator classified the cohort on four occasions. The first one was used to select the initial study cohort. The population was chosen to be reasonably representative of a variety of lobar classification types as determined by the study administrator, including those that might not be expected in a consecutive dataset due to the low prevalence

Fig. 2 Study algorithm. All participants used atlas of BPH classification (upper left) as primary source. The left side (ellipses) comprises the administrator/expert duties. The right side (rounded rectangles) comprises the interpreters' duties. These activities merge at the bottom with "accuracy"


Atlas of BPH Classification
Training Cohort Set (n=82)
Review
1st Classification Analysis [Study Cohort Selected]
Study Cohort Set (n=140)
2nd Classification Analysis
3rd Classification Analysis
Inter-rater Concordance
Intra-rater Concordance
ACCURACY

of some classification subtypes. Following initial selection of the population, the second and third rounds of lobar classification were repeated by the administrator on the same patients after 6–12-month interval to allow for the analysis of *intraobserver* concordance. The administrator was blinded to the first- and second-round data while classifying the third round.

The administrator reevaluated his second and third analyses to reconcile inconsistencies of individual case classifications in order to comprise the final study dataset for testing interrater correlation (precision) and accuracy. The final dataset comprised 49 Type 1, 2 Type 2, 58 Type 3, 6 Type 5, 1 Type 6, and 6 Type 7 BPH as determined by final analysis of the administrator. We found no solitary pedunculated Type 4 cases.

After initial training, using a completely independent dataset that did include examples of solitary Type 4 BPH, two fellowship-trained abdominal radiologists with 5 and 11 years of experience interpreting pelvic MRI independently reviewed and classified the study population. The interpreting radiologists had no prior knowledge of the BPH types represented in the experimental dataset. They were not blinded to patient age, but were blinded to clinical history. All readers had access to all images in each MRI in addition to the T2-weighted series. Original prostate volume calculations were not visible on the study images.

Training of interpreting radiologists

The training material consisted of an "atlas" of MRI lobar classification of BPH based on prior published work and containing numerous example cases that were not previously used. The interpreting radiologists and administrator reviewed and discussed the atlas prior to interpreting the study cases, and the atlas was available for use during the study image classification. Prior to interpreting the final dataset, a training dataset consisting of 82 MRIs selected by the administrator from MRI examinations obtained between January and December 2014 was independently reviewed and classified by the administrator and the interpreting radiologists. After reviewing these results, the interpreting radiologists thoroughly discussed all instances of discordance and sources of error with the administrator resulting in a shared understanding of the classification process.

MRI techniques

Prostate MRI was performed at 3T with a pelvic phased-array coil, 27 with and 113 without endorectal coil (Magnetom Skyra or TrioTim; Siemens Healthcare, Erlangen, Germany). T2W FSE imaging was performed in the axial plane (T_R/T_E 3700/80 ms, N_{EX} 3, 3 mm slice thickness, no interslice gap, flip angle 160°, FOV 140 mm, matrix 320×256) and coronal plane (T_R/T_E 4030/100 ms, N_{EX} 2, 3 mm slice thickness, no interslice gap, flip angle 122°, FOV 180 mm, matrix 320×256). T2W 3D SPACE images were obtained (T_R/T_E 1400/101 ms, flip angle 135°, $256 \times 256 \times 205$ matrix, 180 mm FOV). In addition to 1 mm axial images, T2W 3D SPACE images were reconstructed at 3 mm in the axial, sagittal, and coronal planes. In addition, dynamic contrast-enhanced T1-weighted (3 mm slice thickness, T_R/T_E 4.9/1.8 ms, 224×156 matrix, 250 mm FOV, temporal resolution < 10 s), and diffusion-weighted images at b values of 50, 800, and 200 s/mm^3 were performed.

Variables evaluated by interpreting radiologists and administrator

All images were reviewed on high-resolution monitor PACS using Philips Intellispace v4.4 software. Each observer classified patient prostate lobar morphologies into normal and one of seven BPH categories (Fig. 1). Total prostate volumes ≤ 25 cm^3 were classified as Type 0, despite the fact that many of these patients clearly demonstrated some nodular hyperplasia of the transition zone (TZ). Prostate volumes were calculated using three-plane measurements agreed upon by the three radiologists and included in the training atlas. Type 1 BPH comprised bilateral TZ enlargement. Type 2 BPH showed only enlargement of the retrourethral lobe (RU). Type 3 included those with bilateral TZ plus RU hyperplasia wherein the maximal antero-posterior diameter of the RU was ≥ 10 mm , as measured from its posterior boundary to the posterior boundary of the prostatic urethra. Solitary pedunculated intraurethral or intravesical nodules were considered Type 4. The presence of any such nodules in conjunction with any other lobar enlargements constituted Type 5. Subtypes were also noted. (For example, a Type 5 might comprise Types 1 plus 4 or Types 1, 4, and 6 or Types 3 plus 4, etc.). Type 6 was defined as solitary subtrigonal or ectopic BPH. Type 6 distributions, even when large, rarely contribute to bladder outlet obstruction. Any other configuration combination was classified as Type 7 BPH. Type 7 might comprise Types 1–5 plus 6 (Fig. 1g).

Three-plane measurements were recorded and prostatic volume calculated using the prolate ellipsoid formula (maximal AP diameter × maximal transverse diameter × length × π/6) for distinguishing Type 0 from Type 1. Measurement of the RU lobe was made to determine its maximal antero-posterior diameter. This occurred either above or below the vesico-prostatic line (VPL; Fig. 1c). These lobar classification designations were used for calculation of intra- and interobserver correlation.

Statistics

Intraobserver correlation was determined by comparing the second- and third-round administrator analyses. In an attempt to establish a "ground truth" dataset for comparison to the two interpreters, a fourth lobar classification reconciling results of the second and third, was performed by the administrator acting as his own arbiter. Accuracy was determined by comparing the two interpreting radiologist results with the administrator's "ground truth." Interobserver variability between the two interpreting radiologists was evaluated to determine precision.

Categorical data were compared for case-to-case matching by percent and the unweighted Cohen κ statistic [7] at 0.95 confidence interval (CI). The latter accounts for peripherals that might include guesses favoring the more prevalent BPH types in the study population [14, 15]. CIs for proportions were calculated according to the Wilson efficient-score method, corrected for continuity (http://Vassarstats.net). Following calculation of classification concordance, the administrator analyzed the results for sources of mismatch.

Table 2 Interrater agreement

Proportions of agreement				0.95 CI of observed	
Types	Maximum possible	Chance expected	Observed	Lower limit	Upper limit
0	0.67	0.03	0.54	0.26	0.80
1	0.74	0.12	0.51	0.35	0.67
2	0.6	0.01	0.33	0.06	0.76
3	0.81	0.33	0.61	0.45	0.71
4	0	0	0	0	0.69
5	0.77	0.05	0.35	0.15	0.61
6	0	0	0	0	0.95
7	1	0.04	0.29	0.10	0.58
Composite	0.86	0.31	0.67	0.59	0.75

Proportions of agreement* of two interpreters across all lobar BPH types. Confidence intervals for proportions are calculated according to the Wilson efficient-score method, corrected for continuity

CI confidence limits

*Proportions of Agreement represent percentiles corrected for chance decisions. Birkimer JC, Brown JH. Back to basics: percentage agreement measures are adequate, but there are easier ways. J Applied Behavior Analysis 1979:12:535–543

Results

κ Scores in all tables are reported with their 0.95 CIs. Interrater reliability (IRR) between the two interpreting radiologists resulted in an unweighted κ of score of 0.52 (Table 1) corresponding to 64% agreement. Composite proportion of agreement across all BPH types (categories) for interpreting radiologists was 0.67 (Table 2).

Observer accuracy compared with the expert administrator was 62% agreement with unweighted κ 0.49 for observer 1 and 67% with unweighted κ 0.58, for observer 2 (Tables 3, 4). The results of mismatches along with characterization by administrator after discussion with interpreting radiologists are shown in Table 5 demonstrating sources of error.

Intrarater reliability for the administrator resulted in an unweighted κ 0.81 (Table 6) with 87% agreement. Composite proportion of agreement across all lobar BPH types was 0.87 (Table 7).

Discussion

Our results demonstrate almost perfect intrarater reliability for an expert radiologist. There was moderate interobserver agreement between interpreting mid-career radiologists, and moderate accuracy relative to a highly

Table 1 Cohen's unweighted κ* for interrater observers

Observed κ	Standard error	0.95 Confidence interval	
0.52		Lower limit	Upper limit
Method 1[†]	0.06	0.41	0.64
Method 2[††]	0.057	0.41	0.64
0.81	Maximum possible linear-weighted κ, given the observed marginal frequencies		
0.66	Observed as proportion of maximal possible		

Lower and upper limits of the 95% confidence interval for a proportion, according to two methods described by Robert Newcombe, both derived from a procedure outlined by E. B. Wilson in 1927

*Reference [15]

[†] Newcombe, Robert G. "Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods," *Statistics in Medicine*, **17**, 857–872 (1998)

[††] Wilson, E. B. "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association*, **22**, 209–212 (1927)

Table 3 Cohen's unweighted κ* for accuracy observer 1

Observed κ	Standard error	0.95 Confidence interval	
0.49		Lower limit	Upper limit
Method 1†	0.06	0.37	0.61
Method 2††	0.06	0.38	0.60
0.70	Maximum possible linear-weighted κ, given the observed marginal frequencies		
0.70	Observed as proportion of maximal possible		

Lower and upper limits of the 95% confidence interval for a proportion, according to two methods described by Robert Newcombe, both derived from a procedure outlined by E. B. Wilson in 1927

*Reference [15]

†Newcombe, Robert G. "Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods," *Statistics in Medicine*, **17**, 857–872 (1998)

††Wilson, E. B. "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association*, **22**, 209–212 (1927)

Table 4 Cohen's unweighted κ* for accuracy observer 2

Observed κ	Standard error	0.95 Confidence interval	
0.58		Lower limit	Upper limit
Method 1†	0.05	0.47	0.68
Method 2††	0.05	0.48	0.67
0.80	Maximum possible linear-weighted κ, given the observed marginal frequencies		
0.72	Observed as proportion of maximal possible		

Lower and upper limits of the 95% confidence interval for a proportion, according to two methods described by Robert Newcombe, both derived from a procedure outlined by E. B. Wilson in 1927

*Reference [15]

†Newcombe, Robert G. "Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods," *Statistics in Medicine*, **17**, 857–872 (1998)

††Wilson, E. B. "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association*, **22**, 209–212 (1927)

experienced "expert" radiologist that represented the "gold standard" in our study. Interrater agreement was the highest for normals, Types 1 and 3 BPH constituting the vast majority of lobar distributions in the general population. The interpretation of the meaning of κ scores has been debated in the literature. The generally accepted descriptions suggest that κ of < 0 represents less than chance agreement, 0.01–0.20 "slight" agreement, 0.21–0.40 "fair" agreement, 0.41–0.60 "moderate" agreement, 0.61–0.80 "substantial" agreement, and 0.81–0.99 "almost perfect" agreement [14, 16].

Our results compare favorably with other widely utilized clinical prostate classification systems. Studies of IRR of Gleason grading show overall κ scoring between pathologists from 0.28 [17] to 0.64 [18]. The best-achieved IRR was among expert uropathologists scoring overall κ between 0.47 and 0.64 [18]. Analyzed by *percent* agreement among pathologists, reported consensus scores range between 61% [19] and 78% [20]. Higher agreements are documented for pathologists who are genitourinary specialists compared with general pathologists and those with recent tutorials or exposure from professional meetings [20]. Studies have also shown that IRR can be improved with tutorials [21, 22].

There have been several attempts to estimate "accuracy" by comparing Gleason scoring by the experienced uropathologists against an expert consensus gold standard resulting in κ scores between 0.43 and 0.83; with 42% to 71% agreement [23–27]. Gleason, himself,[24], reported intraobserver agreement to be 50%.

IRR κ for PIRADS has been reported to be 0.43–0.73 [28–30] with percent agreement of 0.68 [30]. Determination of interrater accuracy using biopsy as the gold standard found κ to be 0.55 for T2WIs [31].

When, as is the case with prostate lobar classification, there is likely to be guessing by observers, the κ statistic is preferred over reliance on the percent of agreement [32].

Our results for IRR, intrareader reliability, and accuracy are in accordance with the above ranges for Gleason grading and PIRADS rating that have been considered acceptable for use in the research and in the clinical setting.

Analysis of our sources of error in IRR and accuracy of MRI BPH classification shows two categories of mistake:

Table 5 Sources of error in mismatches of types of lobar BPH

Measurement errors	Counts	
Borderline prostate volume (Type 0 vs. 1)	7	
Significant over measurement of prostate volume	4	
Mismeasurement of prostate volume due to difficulty locating prostatic urethra course	5	
Mismeasurement due to missed location posterior urethral wall	1	
Mismeasurement prostate volume below VPL	1	
Borderline RU lobe AP diameter	4	
Placement level		
Finding RL long axis (locating Veru)		1
Locating urethral course		(3)
Overmeasurement of RL	5	
		31
Failed observations		
Missed subtrigonal lobe (Type 6)	8	
Missed polypoid part of (Type 5)	3	
Overcalled polypoid (Type 4)		
Mistook cystitis cystica/glandularis for polypoid (Type 4)		1
Mistook cystic BPH nodule in RU for polypoid (Type 4)		2
Mistook asymmetrical Type 1 w/herniation for polypoid Type 4		2
Mistook herniating cancer nodule for polypoid (Type 4)		2
Mistook SV cyst for subtrigonal lobe (Type 6)		1
Mistook Type 3 for Type 4		1
Mistook Type 3 for Type 5 (3 + 4)		1
Mistook displaced upper TZ (Type 1) for RU (Type 3)		1
Mistook Type 2 for Type 3 because forgot to measure volume below the VPL		1
Missed identifying the RL		1
Mistook Type 1 for Type 5		1
Misinterpreted vas deferens cyst for subtrigonal (Type 6)		1
		26

VPL vesico-prostatic Line, *RL* retrourethral lobe, *TZ* transition zone

Table 6 Cohen's unweighted κ* for intrarater observations

Observed κ	Standard error	0.95 Confidence interval	
0.8134		Lower limit	Upper limit
Method 1[†]	0.045	0.7252	0.9016
Method 2[††]	0.0446	0.7259	0.9009
0.9378	Maximum possible linear-weighted κ, given the observed marginal frequencies		
0.87	Observed as proportion of maximal possible		

Lower and upper limits of the 95% confidence interval for a proportion, according to two methods described by Robert Newcombe, both derived from a procedure outlined by E. B. Wilson in 1927

*Reference [15]

[†] Newcombe, Robert G. "Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods," *Statistics in Medicine*, **17**, 857–872 (1998)

[††] Wilson, E. B. "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association*, **22**, 209–212 (1927)

error of measurement and failure to make observations (Table 4). Since the decision whether to classify Type 0 or 1 depends on a single volume threshold, minor differences in the measurement of boundaries can become magnified into greater differences in calculated volume. The same is true of the measurement threshold of 10 mm (or any finite number) in the decision to classify enlargement of the RU resulting in Type 2 or 3. The traditional selection of the > 25 cm^3 volume

Table 7 Intrarater agreement

Proportions of agreement				0.95 CI of observed	
Types	Maximum possible	Chance expected	Observed	Lower limit	Upper limit
0	0.86	0.06	0.86	0.56	0.98
1	0.96	0.24	0.80	0.66	0.90
2	0.336	0.006	0.33	0.02	0.98
3	0.96	0.24	0.80	0.66	0.90
4	0	0	0	0	0
5	0.86	0.03	0.63	0.26	0.90
6	1	0.004	1	0.06	1
7	0.8	0.02	0.50	0.14	0.86
Composite	0.96	0.31	0.87	0.79	0.93

Proportions of agreement* across all lobar BPH types. Confidence intervals for proportions are calculated according to the Wilson efficient-score method, corrected for continuity

CI confidence limits

*Proportions of Agreement represent percentiles corrected for chance decisions. Birkimer JC, Brown JH. Back to basics: percentage agreement measures are adequate, but there are easier ways. J Applied Behavior Analysis 1979:12:535–543

threshold for BPH and the choice of 10 mm AP diameter for the diagnosis of RU BPH are arbitrary. There is a need for better clinical correlative studies to justify those values. AP measurement of the RU is complicated by difficulties identifying the course of the prostatic urethra in some patients. This may be unavoidable, even with application of careful multiplanar localization techniques. It is especially true in asymmetrical TZ enlargement wherein the curved displacement of the prostatic urethra cannot be followed in one sagittal or parasagittal plane.

In the study design, rather than using consecutive patients, a decision was made to limit the number of the most common classifications and include a greater number of more rare types of BPH resulting in a better-distributed nonconsecutive series for the test cohort. The effect of including a higher ratio of less common BPH types, if anything, would be expected to lead to bias toward lower κ scores. The κ statistic or any statistic for categorical data can be heavily biased for positive correlation by large imbalances of numbers within each group (prevalence effect) [14]. In our case, we knew from previous experience that there would be far greater numbers of Type 0 (normal), Types 1 and 3 compared with other categories. This imbalance would be exaggerated if the datasets were consecutive. Modern κ statistics compensate to some extent for asymmetrical prevalence in the study population. We therefore believe that the interobserver agreement reported is a conservative measure and highly likely to be reproducible in a consecutive series.

Limitations of this study include certain other assumptions that were made. First, is it valid to use total prostate volume as a critical criterion for making the diagnosis of symptomatic BPH? Some patients with total volume of ≤ 25 cm^3 show nodular TZ hyperplasia disproportionate to overall prostate volume, while others have almost none. Meikle et al. [33] demonstrated a significant difference in symptom scores in patients with $< 25 \pm 0.085$ cm^3 total prostate volumes measured by TRUS. Those with AUA symptom scores < 10 showed a mean TPV volume of 24.6 cm^3 compared with 32.1 cm^3 for those with symptom scores ≥ 10.

Second, is the assumption that maximal AP diameter is the best method of deciding the significance of the RU lobe effect on LUTS, and thus defining RU enlargement? If so, what should the threshold be? We made the assumption that posterior impression on the prostatic urethra alone or in combination with anterolateral compression from the TZ might be important in exerting urethral stenotic effects favoring obstruction. This may or may not be true when the bulk of RU volume is below the VPL, but perhaps maximal RU lobe effects on voiding actually occur above the VPL level. Intravesical prostatic protrusion (IPP) has been demonstrated to correlate with lower urinary tract obstruction and responsiveness to treatment [34–37]. The mechanism may be due to narrowing of the bladder neck and vesicourethral junction by the herniated mass effect or, if sufficiently mobile, the tissue could act as a "flap" valve over the bladder outlet with detrusor contraction. Urologists need to know the extent of IPP and relationship to other lobar enlargements to guide treatment selection, since many minimally invasive surgical procedures do not ablate the RU (median lobe). Consideration of these factors may require future alterations in the MRI lobar classification of BPH to include subclassifications for IPP (herniation into the bladder).

Third, is the assumption that Type 4 pedunculated masses fully within the lumen of the lower urinary tract are major causes of obstruction, especially when the bulk of tissue is located above the VPL in the bladder neck causing blockage by a "ball-valve" mechanism? The absence of enough cases of Type 4 (pedunculated) BPH nodules in our modest-sized cohort limits our understanding of its obstructive potential. There were no *solitary* Type 4 cases in our study cohort, but analysis of Table 2 shows modest abilities of the interpreters to identify polypoid lesions in Type 5 BPH. The data show better IRR for the most common BPH Types 1, 2, and the absence of enlargement (Type 0). Because of their rarity, magnitudes of IRR for Types 2 and 4, in particular, will require much larger cohorts. Patients with RU lobe enlargement and a total volume of exceeding 25 cm^3 may fall into Type 2 or 3 categories. If the volume below the VPL is equal to or less that 25 cm^3, we arbitrarily categorize these as Type 2 and if greater than 25 cm^3, Type 3.

Selection bias by the administrator to include cases with technical and judgment challenges may have affected results, but as the administrator included more challenging cases than would be expected in a consecutive series, this would likely be in a negative correlative direction. Unnoticed confounding bias in case selection or measurement on the part of the senior radiologist (administrator) when creating the final test images cannot be excluded. The effects of the use of the ERC on the shape and measurements of the prostates were not evaluated but could have accounted for small errors of initial categorization of Type 0 versus 1 classification. ERC posterior compression deformity may have minimally reduced AP measurements of some RU lobes. Also, using the expert administrator judgment as the "gold standard" may lead to a source of bias. A consensus panel of experts would have been more acceptable, but there are not enough radiologists available at this time with experience using the classification to form such a panel. Therefore, no suitable alternative to the single administrator was readily available.

Finally, our experience suggests that judgments about BPH classification are complex and are associated with a significant learning curve. Our results in an academic setting may not be transferable to the community setting. However, it is probable that, as with Gleason grading, improvement in performance can be obtained with tutoring and continuing education exposure [21, 22]. The determination of intraobserver reliability, though strong, is severely limited by the use of a single interpreter weakening the result which should be taken only as preliminary and requiring further study.

Conclusion

This study appears to validate the operational capacity of a lobar classification of BPH to be applied in a research setting for further studies of BPH. IRR (precision) and accuracy perform at similar acceptable levels compared with other prostate diagnostic classification systems in clinical use. The actual performance of this lobar classification for clinical application will require further studies.

Funding This study was not supported by any grants or other financial assistance.

Compliance with ethical standards

Conflicts of interest Authors have no pertinent conflicts of interest to declare.

Ethical approval This study was approved by our Institutional Review Board for human research.

References

1. Maserejian N.N., Chen S., Chiu G.R., et al: Incidence of lower urinary tract symptoms in a population-based study of men and women. Urology (2013); 82:560–564
2. Saigal C, Joyce G, Geschwind S, et al. (2004) Methods. In Litwin MS, Saigal CS (eds): Urologic diseases in America. Washington, DC: US Government Publishing Office. pp 283–316
3. Taub DA, Wei JT (2006) The economics of benign prostatic hyperplasia and lower urinary tract symptoms in the United States. Curr Urol Rep 7:272–281
4. Vuichoud C, Loughlin KR. (2015) Benign prostatic hyperplasia: epidemiology, economics and evaluation. Can J Urol 22(Suppl 1):1–6
5. Barry MJ, Beckley S, Boyle P, et al. Importance of understanding the epidemiology and natural history of BPH. In: Cockett ATK, Aso Y, Chatelain C, et al (eds) (1991) Proceedings of the International consultation on benign prostatic hyperplasia (BPH). Paris: Scientific Communications International Ltd.; p. 37
6. Lee A, Lee HJ, Foo KT (2017) Can men with prostates sized 80 mL or larger be managed conservatively? Investig Clin Urol 58:359–364
7. Lepor H, Williford WO, Barry MJ, Brawer MK, Dixon CM, Gormley G, Haakenson C, Machi M, Narayan P, Padles RJ (1996) The efficiency of terazosin, finasteride or both in benign prostatic hyperplasia. N Engl J Med 335: 533–9.
8. Randall A (1931). Surgical pathology of prostatic obstructions. Baltimore (MD): Williams and Wilkins
9. Wasserman NF (2006) Benign prostatic hyperplasia: a review and ultrasound classification. Radiol Clin N Am 44:689–710ß
10. Wasserman NF, Spilseth B, Golzerian J, Metzger BJ (2015) Use of MRI for lobar classification of benign prostatic hyperplasia: potential phenotypic biomarkers for research on treatment. AJR 205:564–571
11. Golzarian J, Antunes AA, Bilhim T, Carnevale FC, Konety B, McVary KT, Parsons JK, Pisco JM, Siegel DN, Spies J, Wasserman N, Gowda N, Ahrar K (2014) Prostatic artery embolization to treat lower urinary tract symptoms related to benign prostatic hyperplasia and bleeding in patients with prostate cancer: proceedings from a multidisciplinary research consensus panel. Journal of Vascular and Interventional Radiology. 25(5):665–74 [Consensus Development Conference]
12. Guneyli S, Ward E, Peng Y, Nehal Yousuf A, Trilisky I, Westin C, Antic T, Oto A (2017) MRI evaluation of benign prostatic

hyperplasia: correlation with international prostate symptom score. J Magnetic Resonance Imaging 45:917–925

13. Grivas N, van der Roest R, Tillier C, Schouten D, van Muilekrom E, Schoots IHeijmink S (2017) Patterns of benign prostate hyperplasia based on magnetic resonance imaging are correlated with lower urinary tract symptoms and continence in men undergoing robot-assisted radical prostatectomy for prostate cancer. Urology
14. Viera AJ, Garrett JM (2005) Understanding interobserver agreement: the κ statistic. Fam Med. 37:360–363
15. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas. 20:37–46.
16. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
17. de la Talle A, Viellefond A, Berger N, Boucher N, De Fromont M, Fondimare A, Moliné V, Piron D, Sibony M, Staroz D, Triller M, Peltier E, Thiounn N, Rubin MA (2003) Evaluation of the interobserver reproducibility of Gleason grading of prostatic adenocarcinoma using tissue microarrays. Human Pathol 34:444–449
18. Allsbrook WC, Jr, Mangold KA, Johnson MH, et al (2001) Interobserver reproducibility of Gleason grading of prostatic carcinoma: Urologic pathologists. Hum Pathol 32:74–80
19. Allsbrook WC Jr, Mangold KA, Johnson MH, et al (2001) Interobserver reproducibility of Gleason grading of prostatic carcinoma: General pathologist. Human Pathol 32:81–88
20. Melia J, Moseley R, Ball RY, Griffiths DFR, Grigor K, Harnden P, Jarmulowicz M, McWilliam LJ, Montironi R, Waller M, Moss S, Parkinson MC (2006) A UK-based investigation of inter- and intra-observer reproducibility of Gleason grading of prostatic biopsies. Histopathology 48:644–654
21. Kronz JD, Silberman MA, Allsbrook WC Jr, et al (2000) Pathology residents' use of a web-based tutorial to improve Gleason grading of prostate carcinoma on needle biopsies. Hum Pathol 31:1044–1150
22. Egevad L (2001) Reproducibility of Gleason grading of prostate cancer can be improved by the use of reference images. Urology 57:291–295
23. Bain GO, Koch M, Hanson J (1982) Feasibility of grading carcinomas. Arch Pathol Lab Med 106:265–267
24. Gleason DF (1992) Histologic grading of prostate cancer: A perspective. Hum Pathol 23:273–279
25. Cintra ML, Billis A (1991) Histologic grading of prostatic adenocarcinoma: Intraobserver reproducibility of the Mostofi, Gleason and Bocking Systems. Int Urol Nephrol 23:449–454
26. Rousselet MC, Saint-Andre JP, Six P, et al (1986) Reproductibilite et valeur pronostique des grades histologiques de Gleason et de Gaeta dans les carcinomes de la prostate. Ann Urol 20:317–322
27. Harada M, Mostofi FK, Corle DK, et al (1977) Preliminary studies of histologic prognosis in cancer of the prostate. Cancer Treat Rep 61:223–225
28. Vaché T, Bratan F, Mège-Lechevallier F, Roche S, Rabilloud M, Rouvière, O. (2014) Characterization of prostate lesions as benign or malignant at multiparametric MR imaging: comparison of three scoring systems in patients treated with radical prostatectomy. Radiology 272:446–455
29. Muller, B. G., Shih, J. H., Sankineni, S. et al. (2015) Prostate Cancer: Interobserver Agreement and Accuracy with the Revised Prostate Imaging Reporting and Data System at Multiparametric MR Imaging. Radiology, 277: 741
30. Renard-Penna R, Mozer P, Cornud F, Barry-Delongchamp N, Bruguière E, Portalez D, Malavaud B (2015) Prostate imaging reporting and data system and Likert scoring system: multiparametric MR imaging validation study to screen patients for initial biopsy. Radiology 275:458–468
31. Schimmöller L, Quentin M, Arsov C, et al (2013) Inter-reader agreement of the ESUR score for prostate MRI using in-bore MRI-guided biopsies as the reference standard. Eur Radiol 23(11):3185–3190
32. McHugh ML. Interrater reliability: the κ statistic. (2012) Biochem Med (Zagreb); 22(3):276–282.
33. Meikle AW, Stephenson RA, Lewis CM, Middleton RG (1997) Effects of age and sex hormones on transition and peripheral zone volumes of prostate and benign prostatic hyperplasia in twins. J Clin Endocrinology and Metabolism. 82:571–575
34. Tan YH, Foo KT (2003) Intravesical prostatic protrusion predicts the outcome of a trial without catheter following acute urine retention. J Urol. 170:2339–2341
35. Shin SH, Kim JW, Kim JW, Oh MM, Moon DG (2013) Defining the degree of intravesical prostatic protrusion in association with bladder outlet obstruction. Korean J Urol 54:369–372
36. Lee SW, Cho JM, Kang JY, Yoo TK. (2010) Clinical and urodynamic significance of morphological differences in intravesical prostatic protrusion. Korean J Urol 51:694–699
37. Cumpanas A, Botoca M, Minciu R, Bucuras V (2013) Intravesical prostatic protrusion can be a predicting factor for the treatment outcome in patients with lower urinary tract symptoms due to benign prostatic obstruction treated with tamsulosin. Urology 81:859–863