ORIGINAL ARTICLE

Poor sensitivity and consistency of microscopy in the diagnosis of low grade non-gonococcal urethritis

R Smith, A J Copas, M Prince, B George, A S Walker, S T Sadiq

Sex Transm Infect 2003;**79**:487–490

See end of article for authors' affiliations

Correspondence to:
Dr Tariq Sadiq,
Department of GUM, St George's Hospital Medical School, Blackshaw Road, London SW17 0RE, UK;
s.sadiq@sghms.ac.uk

Accepted for publication 18 June 2003

Objectives: To determine the reliability of the diagnosis of non-gonococcal urethritis (NGU), and the variation between and within microscopists, from urethral smears at a large London genitourinary medicine clinic.

Methods: A senior microscopist (SM) preselected 60 Gram stained urethral smear slides, 20 negative (<5 polymorphs/hpf), 20 low grade NGU (5–20 p/hpf), and 20 high grade NGU (>20 p/hpf). Ten experienced microscopists, blinded to these initial grades, examined all slides giving each a polymorph score. After relabelling and randomly changing their order, the slides were re-examined by the same microscopists. Finally, the SM determined whether the study had resulted in loss of cells from any of the slides. The SM's initial grading and the consensus among microscopists provide two gold standards for analysis.

Results: Nine low grade and five high grade slides were removed from analysis because of loss of cells. By SM standard, considering microscopists' readings as simply non-NGU (<5 p/hpf) or NGU (⩾5 p/hpf), 97% from negative slides were correct (variation 93–100 across microscopists), 68% from low grade slides (45–95), and 94% from high grade slides (83–100). Consistency between repeat readings by the same microscopist was 96% for negatives, 75% for low grade and 89% for high grade slides. Results were similar by consensus standard.

Conclusions: There was considerable variation between and within microscopists in the diagnosis of NGU. Sensitivity was strongly related to grade of urethritis, with an appreciable proportion of low grade urethritis falsely diagnosed as negative. With increasing attendances for sexual health screening, a false positive rate of only 3% may lead to many false diagnoses.

Non-gonococcal urethritis (NGU) is one of the most commonly diagnosed conditions in males attending genitourinary medicine (GUM) clinics in the United Kingdom. Although *Chlamydia trachomatis* may account for 30–50% of cases, *Mycoplasma genitalium*, *Ureaplasma urealyticum*, and *Trichomonas vaginalis* can all cause urethritis, and multiple other pathogens have also been implicated.[1] The diagnosis in symptomatic males or in the presence of discharge is generally straightforward but is more difficult in men without symptoms or signs. In these cases, clinicians rely on the presence of excess polymorphs in the anterior urethra to establish the diagnosis.

UK guidelines for the management of NGU state that diagnosis must be confirmed either by demonstrating more than four polymorphonuclear leucocytes per high power field (p/hpf) or more than 9 p/hpf on a Gram stained smear from the anterior urethra or from the sediment of a sample of first pass urine, respectively.[2] The high power field represents a magnification of ×1000 using a light microscope (combining a ×100 objective lens with a ×10 eyepiece). The use of 5 p/hpf as a cut off between normal and NGU has historically come from research studies at a time when sensitive and specific molecular techniques for detecting pathogens had not been developed and there is some difficulty when attempting to extend the findings of these studies to the asymptomatic patient.[3–9]

Additionally, diagnosis of NGU by microscopy will vary depending on a number of factors, including the time since last passing urine, the method by which the urethral specimen is taken and applied to the slide, and the subjectivity of the microscopist examining it. Previous research has found a considerable degree of observer variation in the microscopic interpretation of urethral smears changing the diagnosis made by an initial reader in up to 40% of patients.[10]

The aim of this study was to assess the accuracy of diagnosis of NGU by microscopic analysis of urethral smears among a large number of trained observers in a London GUM clinic when compared to a gold standard. A further aim was to determine the degree of variation both between and within microscopists in giving polymorph counts from urethral smears and how this variation might influence the reliability of NGU diagnosis.

METHOD

The senior microscopist (SM) responsible for microscopy training at the Mortimer Market Centre selected 60 Gram stained urethral smear slides, which, in his opinion, represented three grades of polymorph values: 20 with less than 5 p/hpf (negative), 20 with values between 5 and 20 p/hpf (low grade urethritis), and 20 with more than 20 p/hpf (high grade urethritis). The slides were allocated a random order and randomly relabelled 1–60 by one of the investigators. From among approximately 40 experienced staff routinely undertaking microscopy at the clinic, 10 were randomly selected to take part in the study and none declined. To get a polymorph score microscopists had been trained to examine the area of cellular material on slides initially by low power, and then to select three high power fields within this area that represented the highest concentration of polymorphs. The average score from these fields was then taken as the score.

Slides were examined by these microscopists during allocated individual sessions but under normal clinic conditions. Polymorph scores were recorded as <5 p/hpf, 5–20 p/hpf, or >20 p/hpf for each slide. Scores were collected

Table 1 Cross tabulation of senior microscopist (SM) grades and consensus grades

	Consensus grading			
SM grading	**Negative**	**Low grade**	**High grade**	**Total**
Negative	20	0	0	20
Low grade	3	7	1	11
High grade	0	5	10	15
Total	23	12	11	46

at the end of each session to eliminate conferring. After every microscopist had read every slide once, the slides were randomly relabelled and their order changed. The same microscopists were then asked to re-examine and score every slide again.

The original identities and categories of the slides were then restored and the SM determined whether the study protocol had resulted in loss of polymorphs from any of the slides. Throughout the study, the investigators analysing the data were blinded to the identity of the individual microscopists taking part.

The initial grading of the slides provides a gold standard with which to measure the performance of the 10 microscopists. A second gold standard was the consensus, where the grade of each slide was determined by the majority agreement of the readings (two per slide) from the 10 microscopists. For example, if across the 20 readings for a slide 12/20 scored a slide as <5 p/hpf, 6/20 as 5–20 p/hpf, and 2/20 as >20 p/hpf the consensus grade for that slide was <5 p/hpf or negative.

The data were initially analysed using STATA 6. However, the data are cross classified by microscopist and slide. For this reason we fitted a random effects logistic regression model to the data for each of the two binary outcomes considered. The first, defined for each reading, is correct diagnosis relative to gold standard, performed separately for each standard. The second is consistency of reading, defined for each slide/microscopist pair, determined by whether the two diagnoses provided for each slide by each microscopist were in agreement or not. For each model, random effects were included for each slide and for each microscopist and fixed effects for low and high slide grade (by gold standard) relative to negative. The models were fitted using MLwiN (version 1.1).[11] The resulting odds ratios are conditional, relating to the odds of the outcome for a different slide grade for the same microscopist. The variance between the random effects of the microscopists was estimated for each model. As an informal test, when this variance was greater than twice its standard error we declared statistically significant inter-microscopist variation—that is, variation larger than could have occurred by chance alone.

RESULTS

Nine low grade and five high grade slides were removed from analysis because of loss of cells. Of the remaining 46 slides, each was read twice by 10 microscopists, with the exception of one slide that was read once by all microscopists but was misplaced before the second reading could be performed by six of the microscopists. Hence the analysis is based on 914 readings in total.

For each slide, the grade of urethritis given by the SM was compared to the consensus grade of the 10 microscopists over both readings (see methods) of the slide (table 1). The two gradings were in agreement for 80% (37/46) of slides. Of the nine discrepant slides, eight were given a higher reading by the SM than the consensus. Of the 11 low grade slides by SM, three were negative by consensus, and 5/15 high grade slides by SM were low grade by consensus. Only the former three slides would have been diagnosed differently by SM and consensus.

We then examined the proportion of times the different grades of slides were correctly scored by each microscopist and overall as either non-NGU (<5 p/hpf) or NGU (⩾5 p/hpf) relative to the two gold standards over both phases of the study (table 2).

Overall, relative to the SM and consensus standards, 97% and 94% of readings for negative slides were correctly identified as non-NGU, respectively; 68% and 84% of readings for low grade slides were correctly identified as NGU, respectively; 94% and 96% of readings for high grade slides were correctly identified as NGU, respectively. This difference between slide grades was statistically significant ($p<0.001$, SM standard; $p = 0.02$, consensus standard) (see table 3). The odds ratios (95% confidence interval; CI) of correct diagnosis for low grade relative to negative were 0.030 (0.007 to 0.150) by SM standard and 0.24 (0.08 to 0.74) by

Table 2 Proportion of readings for each slide category correctly diagnosed as non-NGU or NGU

	% (n) of readings for negative slides correctly diagnosed as non-NGU		% (n) of readings for low grade slides correctly diagnosed as NGU		% (n) of readings for high grade slides correctly diagnosed as NGU	
Microscopist	**SM standard (n = 39/40*)**	**Consensus standard (n = 45/46*)**	**SM standard (n = 22)**	**Consensus standard (n = 24)**	**SM standard (n = 30)**	**Consensus standard (n = 22)**
1	100 (39/39)	100 (45/45)	64 (14)	71 (17)	83 (25)	100 (22)
2	97 (38/39)	96 (43/45)	55 (12)	75 (18)	93 (28)	95 (21)
3	97 (38/39)	93 (42/45)	77 (17)	92 (22)	97 (29)	100 (22)
4	100 (39/39)	98 (44/45)	50 (11)	79 (19)	100 (30)	95 (21)
5	95 (38/40)	83 (38/46)	95 (21)	100 (24)	100 (30)	100 (22)
6	93 (37/40)	91 (42/46)	77 (17)	96 (23)	97 (29)	100 (22)
7	95 (37/39)	91 (41/45)	82 (18)	96 (23)	97 (29)	100 (22)
8	93 (37/40)	93 (43/46)	45 (10)	67 (16)	83 (25)	86 (19)
9	100 (39/39)	98 (44/45)	59 (13)	75 (18)	90 (27)	95 (21)
10	100 (40/40)	93 (43/46)	77 (17)	92 (22)	97 (29)	95 (21)
Range	93–100%	83–100%	45–95%	67–100%	83–100%	86–100%
Overall:	97.0% (382/394)	93.6% (425/454)	68.2% (150/220)	84.2% (202/240)	93.7% (281/300)	96.4% (212/220)

*Denominators vary because one slide was read once by all microscopists but misplaced before the second reading could be performed by six of the microscopists.

Table 3 Results of a logistic regression to examine the effect of slide grade on the odds relative to negative grade of (a) correct diagnosis and (b) within microscopist slide consistency

	Correct diagnosis odds ratio (95% CI)		Consistency of readings odds ratio (95% CI)	
Type of slide	SM grade	Consensus	SM	Consensus
Negative	1	1	1	1
Low grade	0.032 (0.007 to 0.150)	0.24 (0.08 to 0.74)	0.11 (0.05 to 0.27)	0.20 (0.10 to 0.41)
High grade	0.42 (0.08 to 2.04)	1.74 (0.39 to 7.81)	0.33 (0.13 to 0.81)	0.91 (0.37 to 2.27)
*p Value	<0.001	0.02	<0.001	<0.001

Odds ratios are shown for both senior microscopist (SM) and consensus standards of slide.
*p Value for the overall difference between slide grades.

consensus standard. Thus, there was poor sensitivity of microscopy for low grade urethritis even relative to the consensus standard. This finding also held for each microscopist when the SM standard was taken—in that the proportion correctly identified was lowest for low grade slides. When the consensus standard was taken, for 7/10 microscopists the proportion of readings that were correct was lowest for low grade slides, and for the remaining 3/10 the readings were least correct for negative slides. The proportion of slides correctly diagnosed varied substantially across microscopists for low grade slides, as indicated by the range: 45–95% by SM standard and 67–100% by consensus. The overall variability between microscopists on the log odds scale was, however, not statistically significant by either standard.

We then examined the consistency within microscopist and slide pairs—that is, how well microscopists agreed with themselves between the two readings of the same slides. If the microscopist had agreed with his/her previous reading, irrespective of the grade of slide, then the pair of readings was deemed consistent. The proportion of consistent pairs within each category by SM and consensus standards was then calculated (table 4) and found to be lowest for low grade urethritis at 75% by both gold standards. This difference between slide grades was statistically significant (table 3), p<0.001 by both standards, and the odds ratios (95% CIs) for consistent readings for low grade relative to negative were 0.11 (0.05 to 0.27) by SM standard and 0.20 (0.10 to 0.41) by consensus. The proportion of slides for which consistency occurred varied substantially across microscopists for low grade slides: being 45–100% and 50–100% by SM and consensus standards, respectively. However, the overall variability between microscopists was not statistically significant by either standard.

DISCUSSION

We were broadly satisfied with our sensitivity/specificity values and consistency in reading negative and high grade slides. However, even the specificity found of 97.0/93.6% taking the gold standard as SM/consensus reading, respectively, implies that 3.0/6.4% of patients without NGU will be falsely diagnosed with this potential sexually transmitted infection (STI), treated accordingly and asked to notify partners. With greater numbers of asymptomatic patients attending GUM clinics for screening in the United Kingdom, this research suggests that significant numbers of men may be falsely diagnosed as NGU. It would thus seem appropriate for the social and psychological consequences of being given a diagnosis of NGU to be more widely examined. These results would also not support the wide use of microscopy for the diagnosis of NGU in low prevalence settings such as primary care.

Table 4 Proportion of slide microscopist pairs where there was consistency between the two readings by the same microscopist, by senior microscopist (SM), and consensus standard of slide

		Proportion of slide microscopist pairs where the two readings were consistent
Type of slide	Slide gold standard	% (n), range across microscopists
Negative	SM	96 (186/194), 85–100
	Consensus	93 (209/224), 86–100
Low grade	SM	75 (82/110), 45–100
	Consensus	75 (90/120), 50–100
High grade	SM	89 (133/150), 67–100
	Consensus	93 (102/110), 73–100

Of substantial concern is the very low sensitivity and consistency achieved for low grade NGU—that is, counts between 5–20 p/hpf. These counts are always going to be a problem as many of the values lie near the cut-off value of 5 p/hpf, but may constitute an increasing proportion of NGU diagnoses seen in GUM clinics. In particular, if we take the SM grading as the gold standard then our sample sensitivity of 68.2% implies that for every 100 males attending with low grade NGU, 32 would not be given an NGU diagnosis. Even if 50% of men with low grade NGU are positive for *C trachomatis*, which is likely to be an overestimate, then 16 of the 32 men would have been recalled for treatment of chlamydia infection. This still leaves 16 men untreated out of every 100 men with low grade NGU. Our study was not large enough to examine how the sensitivity varies within the range 5–20 p/hpf, and this remains an area for further research.

Consistency between the two readings of a slide by the same microscopist was lowest for low grade urethritis smears by both standards, suggesting that variation within microscopist for a given slide is a major factor in the false negative diagnosis of low grade urethritis slides. For such slides, apparently substantial variation between microscopists was also observed: the range of sensitivity across the 10 microscopists being 45–95%/67–100%, and the range of consistency between repeated readings being 45–100%/50–100%. The overall variation between microscopists in both outcomes was, however, not statistically significant and further research with larger numbers of microscopists might better establish the magnitude of inter-microscopist variability.

Our study was conducted in one clinic and hence the generalisability of our findings cannot be assessed. As the microscopists were all experienced staff, it seems unlikely that our findings would not be broadly reproduced in other settings. It is important to note, however, that polymorph scores at our clinic were obtained by averaging over three high power fields with the greatest concentration of polymorphs as opposed to five, as suggested by the national guidelines. It is possible that this difference might lead to differences in variability of NGU diagnosis. Some slides (mainly low grade) were removed from the analysis because

Key messages

- The sensitivity of microscopy for diagnosing low grade non-gonococcal urethritis (NGU) is poor.
- There is considerable intraobserver variation in diagnosing NGU on slides from patients with low grade NGU.
- In low prevalence settings microscopy may lead to a large number of false diagnoses of NGU.

of loss of cells and this may have influenced our results. However, such slides, considered low grade initially by the SM but negative on his final reading, might be more likely to be misread as negative. Consequently, if removal has introduced a bias, it would be to inflate the sensitivity for low grade slides. Taking the consensus reading as gold standard may be considered to lead to inappropriately high sensitivity and specificity estimates. Clearly, however, our finding of relatively low sensitivity for low grade slides extends even to this gold standard. In addition, we have found poor consistency between repeat readings for low grade slides under both gold standards.

With such variability of readings of urethral smears, the case for the routine use of validated diagnostic tests for detecting causes of NGU other than *C trachomatis* is enhanced. Moreover, this study emphasises the need to take into account the clinical picture, including sexual history and symptoms, as well as polymorph count on urethral slides when making a diagnosis of NGU. With over 500 000 diagnoses of NGU a year in the United Kingdom,[12] this research raises some important public health concerns and we would strongly encourage further research to be undertaken. This should address the sensitivity and specificity of NGU diagnosis (particularly for low grade urethritis) across different settings, identify how the performance may be improved in practice, and examine the social and psychological impact of the diagnosis.

ACKNOWLEDGEMENTS

The authors acknowledge with gratitude the considerable efforts of the microscopists in this study.

CONTRIBUTORS

AJC, RS, and STS wrote the initial draft; STS conceived the project and together with RS, MP, and AJC designed it; MP selected the slides and together with BG organised microscopists; RS supervised collection of the data and randomised slides. Statistical analysis was provided by AJC and ASW.

Authors' affiliations

R Smith, B George, S T Sadiq, Mortimer Market Centre, Camden Primary Care Trust, London, UK
A J Copas, Department of Sexually Transmitted Diseases, RFUCMS UCL, London, UK
M Prince, University College Hospitals NHS Trust, London, UK
A S Walker, MRC Clinical Trials Unit, London, UK

Funding: Internal funding.

Conflicts of interest: None.

REFERENCES

1 **Taylor-Robinson D**. The history of nongonococcal urethritis. Thomas Parran Award Lecture. [erratum appears in *Sex Transm Dis* 1996;**23**:170.] *Sex Transm Dis* 1996;**23**:86–91.
2 **Clinical Effectiveness Group**. National guideline for the management of non-gonococcal urethritis. London: Clinical Effectiveness Group (Association of Genitourinary Medicine and the Medical Society for the Study of Venereal Diseases), 2002 (www.mssvd.org.uk/PDF/CEG2001/NGU).
3 **Holmes KK**, Handsfield HH, Wang SP, *et al*. Etiology of nongonococcal urethritis. *N Engl J Med* 1975;**292**:1199–205.
4 **Swartz SL**, Kraus SJ, Herrmann KL, *et al*. Diagnosis and etiology of nongonococcal urethritis. *J Infect Dis* 1978;**138**:445–54.
5 **Desai K**, Robson HG. Comparison of the Gram-stained urethral smear and first-voided urine sediment in the diagnosis of nongonococcal urethritis. *Sex Transm Dis* 1982;**9**:21–5.
6 **Bowie WR**. Comparison of Gram stain and first-voided urine sediment in the diagnosis of urethritis. *Sex Transm Dis* 1978;**5**:39–42.
7 **Arya OP**, Mallinson H, Andrews BE, *et al*. Diagnosis of urethritis: role of polymorphonuclear leukocyte counts in gram-stained urethral smears. *Sex Transm Dis* 1984;**11**:10–7.
8 **Landis SJ**, Stewart IO, Chernesky MA, *et al*. Value of the gram-stained urethral smear in the management of men with urethritis. *Sex Transm Dis* 1988;**15**:78–84.
9 **Terry PM**, Holland S, Olden D, *et al*. Diagnosing non-gonococcal urethritis: the gram-stained urethral smear in perspective. *Int J STD AIDS* 1991;**2**:272–5.
10 **Willcox JR**, Adler MW, Belsey EM. Observer variation in the interpretation of Gram-stained urethral smears: implications for the diagnosis of non-specific urethritis. *Br J Vener Dis* 1981;**57**:134–6.
11 **Goldstein H**, Rasbash J, Plewis I, *et al*. *A User's Guide to MlwiN. Version 1.0*. London: Multilevel Models Project: Institute of Education, 1998.
12 **Lamagni TL**, Hughes G, Rogers PA, *et al*. New cases seen at genitourinary medicine clinics: England 1998. *Commun Dis Rep* 1999;(CDR Suppl 9):S1–12.