

# Mini-reviews

There are four reviews in this section; two of these relate to prostate cancer, one to paediatric urology, and one to bladder function. The prostate cancer mini-reviews concern two important areas that are talking points in urological oncology. Multidisciplinary team management, which is a very attractive idea to many, remains controversial in the eyes of some. This concept is discussed in detail, as is another controversial idea, the use of high-intensity focused ultrasound in the treatment of prostate cancer.

## A systematic review of the reliability of frequency-volume charts in urological research and its implications for the optimum chart duration

Tet L. Yap, David C. Cromwell\* and Mark Emberton

*Clinical Effectiveness Unit, Royal College of Surgeons of England, and \*Institute of Urology and Nephrology, University College Hospital, London, UK*

Accepted for publication 7 July 2006

### OBJECTIVE

To determine how the reliability of frequency-volume charts (FVCs) vary with their duration when used to assess patients with lower urinary tract symptoms (LUTS) and whether the duration influences patient compliance.

### METHODS

Peer-reviewed studies involving patients with LUTS were searched systematically, with the selected studies assessed for their internal and external validity and statistical quality. Details of the patients and type of FVC used were summarized, and reliability coefficients and levels of compliance were extracted for commonly assessed FVC variables.

### RESULTS

In all, 13 studies were considered to meet the review criteria; they assessed the reliability of FVCs lasting 1, 2, 3 and 7 days. The reliability coefficients for 3- and 7-day FVCs were generally  $>0.8$ ; those for shorter charts tended to be lower, but strong conclusions could not be drawn due to study limitations. There was no obvious relationship between the duration of the FVC and the level of compliance.

### CONCLUSIONS

Strong recommendations cannot be made about what duration of an FVC should be used to assess or monitor patients with LUTS. The current consensus on using FVCs of  $\geq 3$  days seems to be the most defensible policy, but more research of high quality is required, especially into the relationship of FVC duration with compliance.

### KEYWORDS

frequency-volume chart, duration, systematic review, reliability studies

### INTRODUCTION

Frequency-volume charts (FVCs) are noninvasive tools that allow the assessment of LUTS outside the setting of a clinical interview. They are recommended by the WHO as a diagnostic test in the routine assessment of men with LUTS suggestive of BPH, and are regarded by the most recent AUA BPH guidelines as being especially useful when assessing LUTS where nocturia is the predominant feature. FVCs can also be used to monitor therapeutic outcomes for interventions such as bladder training [1].

There are various formats that the FVC can take [1]. Early FVCs were simple forms on which patients recorded voiding frequency for a fixed number of days [1]. Most FVCs now require patients to record the time, volume and type of every drink they take, and the time and amount of urine at each voiding episode. From these data, several variables are derived, e.g. the 24-h voiding frequency, nocturnal frequency and mean voided volume. The ICS has agreed definitions for these variables to ensure that practice is consistent and research is comparable [2]. Episodes of urgency and incontinence (with the number of incontinence pads used) can also be recorded on the FVC, depending on its intended use. Some authors classify more detailed FVCs as 'urinary diaries'.

There has been much debate about the optimum duration of FVCs; the number of days recommended in previous reports was 1 to >7 [3–6]. The main issue at the centre of the debate is the need to balance patient compliance with the reliability of derived variables; as the duration of the FVC increases, the number of days completed by patients (compliance) can decrease. This is unsurprising, as it is labour-intensive to record every intake and voiding episode. However, reliability, an indication of the amount of error in any measurement, improves as the duration of a FVC lengthens [7,8]. We report a systematic review of studies that assessed the reliability of FVCs, with the aim of determining the optimum FVC duration for assessing the voiding behaviour of patients with LUTS.

## METHODS

For this study, we used 'FVC' to encompass the variety of terms encountered in medical reports to describe this approach to measuring voiding behaviour. Common terms include micturition chart, voiding chart, bladder chart, urinary chart, frequency chart, frequency-severity chart and FVC. These terms will also be encountered with 'chart' being replaced by 'diary'.

We developed a structured search strategy (Appendix 1) to identify relevant studies in the Medline, Embase and Cinahl databases, published between January 1966 and December 2005. The databases were searched by one reviewer (T.L.Y.). Using information in listed titles and abstracts, potentially relevant

studies were then selected and the full paper obtained. The review was limited to published peer-reviewed articles or book chapters that had evaluated FVCs in patients with urinary problems. The studies were also required to have used a repeated-measures, test-retest reliability design and presented reliability statistics for some or all of the FVC variables, as defined by the ICS (24-h frequency of urination, episodes of nocturia, episodes of urgency, episodes of urinary incontinence, number of daytime voids, nocturnal volume, total intake volume/24 h, total voided volume/24 h, mean voided volume, largest voided volume).

Studies evaluating other measures of urinary symptoms (e.g. symptom scores) or that had incorporated FVCs to assess the validity of other measures of LUTS severity, were excluded. All of the articles retrieved were further examined for additional publications within their reference list. Eligibility was evaluated by two reviewers (T.L.Y. and D.C.) and disagreements were resolved by discussion.

A validated checklist of criteria for assessing the methodological quality of test-retest reliability studies on patient-rated scales like the FVC was not available. We therefore devised a list of criteria from previous systematic reviews investigating reliability in other fields of clinical research [9,10] (Appendix 2). The criteria stem primarily from diagnostic research studies including a validated tool for assessing the diagnostic accuracy and study quality [11,12]. These criteria were grouped into measures of external validity (criteria 1–3), internal validity (4–6), and criteria assessing statistical methods and description of patients (7–9). Papers were independently assessed by two reviewers (T.L.Y. and D.C.) with discrepancies again resolved by discussion. Questions were scored 'Yes' or 'No', with any case of unclear or no reporting given the latter response.

Data from the included publications were collected on participants (number enrolled, inclusion and exclusion criteria, distribution of age and gender, clinical characteristics, method of recruitment, compliance), type of FVC assessed (duration, components, variables derived), method of data collection, including whether or not patients were trained on how to use the FVC, and duration between assessments with the FVC. Data were also extracted on the statistical methods used for

analysis, the results of reliability analysis, and the conclusions on recommended FVC duration.

Evaluating the test-retest reliability of an instrument involves assessing the agreement among two or more measurements from the same person [13], and is generally summarized using a statistic that ranges between 0 (indicating no agreement) and 1 (perfect agreement). For continuous quantities, like mean voided volume, reliability is commonly measured using an intraclass correlation coefficient (ICC). In its simplest form, the variation in the measurements is assumed to consist of the actual variation between patients and the measurement error, and the ICC is calculated by dividing the variance of the patients by the variance of the patients plus the variance of the measurement error. An alternative statistic is the concordance correlation coefficient [14].

The Pearson correlation coefficient is often used to measure the reliability of continuous measures, although this is inappropriate [15]. This statistic measures the strength of a linear relationship between variables and not the degree to which they agree. It is possible for two variables to be highly correlated but be in poor agreement. Similarly, rank correlation coefficients are also inappropriate because they measure the similarity in the ordering of the values from two variables rather than their agreement.

A different approach to measuring reliability was proposed by Bland and Altman [15]. They recommended plotting the difference between measurements against their mean. This would identify any systematic bias or outliers and whether the size of the measurement error was related to the measurement value. They then proposed using limits of agreement, calculated as the mean of the errors  $\pm$  twice its SD, to measure reliability. Although unorthodox, it has two advantages. First, if the measurement errors increase with the values, the limits can be derived from the log of the values and expressed as a percentage of the mean. Second, the limits of agreement are not influenced by the size of variability in the sampled patients, making it less dependent to the population being studied. By contrast, standard reliability statistics for an instrument will vary between patient groups with different distributions of measurements.

## RESULTS

Searching of the Medline, Embase and Cinahl databases yielded 527 citations. Of these, 18 possibly relevant studies were retrieved as full articles [1,3–6,16–28]. Nine studies fulfilled all inclusion criteria [3–6,16–20]. Reference tracing and hand searching yielded 11 more possibly relevant studies [29–39], four of which met inclusion criteria [29–32].

The 16 studies were excluded for the following reasons. Two were review articles and provided no new data [34,35], and two were only available in abstract form [36,39]. One did not use any discernible form of FVC [28]. Nine did not assess the test-retest reliability of their FVC data [1,21–27,37] and two examined only incontinent episodes with no assessment of other urinary variables [33,38].

In all, 13 studies were included in the present review [3–6,16–20,29–32]. Three studies included patients with LUTS [3,16,17], two included groups of asymptomatic women [19,20], two included women attending clinics (urological [5] and urodynamic [32]) and six in patients with urinary incontinence (including two in stress incontinent women [29] and four in predominantly urge incontinent patients [4,6,18,30]). The study characteristics are summarized in Table 1.

There was considerable variation in the methods used in the articles reviewed (Table 2). Only one study was judged to satisfy all external criteria, primarily because no other study was judged to have enrolled a representative sample. Studies had used patients enrolled in a clinical trial, or had excluded patients with only partial FVC data or had omitted or described poorly the method of enrolment.

Internal validity was also poor overall, with only two studies meeting all criteria. Surprisingly, only five studies gave an adequate description of compliance rates. Most studies provided incomplete information about reasons for withdrawal, but compliance rates could not be derived for two studies.

None of the studies met all the statistical criteria because no study gave a justification for the number of patients enrolled in the study. Five studies had not used an appropriate study design. In four cases, the

two sets of compared measurements were not independent, while in the other case, the study compared the distribution of measurements within groups rather than individual measurements. Among the eight studies with appropriate designs, reliability estimates were given for a range of FVC variables, most commonly 24-h frequency, nocturia, mean voided volume and number of incontinent episodes per day (Table 3). A few trials also assessed daytime frequency, total voided volume, largest voided volume and number of urgent episodes per day. The interpretation of the reliability statistics was hampered in two of these eight studies because the distribution of values among patients was not presented [16,17].

Three studies assessed the reliability of FVC variables in patients predominantly with LUTS. However, one used measurements that were not independent and was therefore not considered further. The two remaining studies were focused primarily on 3-day FVCs, and both reported reliability coefficients of  $>0.8$  for the assessed variables, suggesting only a small portion of the variation between values was due to measurement error. Unfortunately, the results are not directly comparable, as they evaluated different sets of variables and derived different statistics. Bryan and Chapple [16] concluded that 3-day FVCs showed acceptable reliability but the study had various weaknesses. More importantly, the exclusion criteria and recruitment procedure for patients were unclear, and comparison with other studies is difficult because the voiding behaviour of patients was only described graphically. Of the 61 patients enrolled, 51 completed both sets of FVCs.

Groutz *et al.* [17] also provided reliability estimates for 1-day FVCs, finding that, for 24-h frequency, the reliability decreased from 0.83 to 0.67, and for incontinent episodes, from 0.86 to 0.81. Groutz *et al.* concluded that a 24-h chart was a reliable instrument, and increasing the FVC duration to 3 days increased the reliability but was associated with decreased compliance. Compliance was 97% across all variables but was reported to decrease from 1 to 3 days for some. Exact values for 24-h frequency and incontinent episodes were not provided.

Five studies evaluated the reliability of FVCs in patients with predominantly incontinence, using FVCs for which the duration was

1–7 days. For FVCs completed for  $\geq 3$  days, the reliability coefficients were generally  $>0.8$ , and it was generally concluded that FVCs of such duration provide reliable results. The reliability estimates were similar for patients with chiefly stress incontinence [29,31] or urge incontinence [4,6,18,30]. However, the studies provided little cumulative evidence because they used different reliability statistics, assessed different sets of FVC variables, and enrolled different patient populations. The studies also provide little evidence on how compliance might decrease with duration of the FVC, as all reported levels of compliance of  $>90\%$ .

The two studies by Larsson *et al.* [29,30] used 1- and 2-day charts, and reported reliability using limits of agreement, expressed as a proportion of the average measurement value. These were the only studies to report that the difference between measurements increased with the average of the measurements. The reliability of variables derived from 24-h FVCs was described as poor, but a definitive recommendation about the duration of an FVC was not provided.

Two articles assessed patients from urological assessment clinics, with no specific pathological diagnosis [5,32]. However, the exact nature of this recruitment process was unclear in both papers. Schick *et al.* [5] used retrospective data but censored incomplete FVC data. This made it unclear what the original sample size and compliance rate was. Also, the lack of patient characteristics meant that the voiding behaviour of this cohort, and the stability of their condition, could not be accurately assessed. Finally, the statistics used to calculate test-retest reliability (ANOVA-derived [5] and rank [32] correlations) were inappropriate, as detailed in the present methods.

Of the two trials using asymptomatic patients, only one used an appropriate statistical measure to determine reliability [20]. This study evaluated the reliability of 1-day and 2-day FVCs. For 24-h frequency and mean voided volume, the Pearson correlation coefficients were 0.60 and 0.77, respectively, when derived from a 1-day FVC. The coefficients were higher for a 2-day FVC and the authors concluded that reliability can be improved with greater FVC duration, although they did not make a clear recommendation about the duration. However, the results do

TABLE 1 Description of reliability studies of FVCs according to patients' main symptom group

Ref	Description of subjects	FVC duration	Analysis		Compliance rate, %	FVC duration recommended
			D vs D	Statistic		
<b>Patients with predominantly LUTS</b>						
[16]	Prospective cohort of 63 with urgency; failed conservative therapy, offered neuromodulation. Mean age: IDO 47.3 (24–76); with NDO: 47.3 (24–76). 21 (33%) men (UK)	Two 3D (interval $\geq 1$ week)	3 vs 3	PC	81	3D
[3]	Prospective cohort of 180 men >50 year with LUTS due to BPH. Mean age 64.9 (50–89) (The Netherlands)	Three 1D (for 3D normal activity)	1 vs 3	RC	57	1D
[17]	Prospective cohort of 106, $\geq 18$ years referred for LUTS evaluation at 8 urological practices. Median age 64 (22–84), 92 (87%) men (USA)	Two 3D (interval 1 week, recorded on same days)	3 vs 3 2 vs 2 1 vs 1	CCC	97	1D
<b>Patients with predominantly incontinence</b>						
[18]	154 >20 years old from 5 urological clinics; history of urge or mixed UI with urge as primary component. Mean age (men) 53.8; (women) 54.6. 21 (14%) men (USA)	Two 7D (with interval $\geq 1$ week)	7 vs 7 4 vs 4 3 vs 3	ICC	94	7D
[30]	70 women with IDI from multicentre study of effects of terodiline. Median age 46 (18–76) (UK)	One 7D (data from first 4 days)	1 vs 1 2 vs 2	LA	89	?
[29]	83 women retrospectively recruited from a continence clinic with symptom and urodynamic diagnosis of genuine stress UI. Median age 47 (31–69)	Two 1D (filled as a 48-h chart)	1 vs 1	LA	98	?
[31]	140 women with urodynamic genuine stress UI from a multicentre trial assessing a urethral insert. Mean age 53.6 (27–78) (USA)	Two 7D (interval 4 weeks)	7 vs 7 3 vs 4	PC	99	3D
[4]	Retrospective, 98 women $\geq 18$ years in urological clinic with urge UI. Mean age 54.5 (18–81) (USA)	2D or three 1D	1 vs 2/3	RC	61	1D
[6]	50 women aged $\geq 55$ years with UI from a clinical trial of behavioural management for UI. Mean age 65.1 (55–86) (USA)	Two 7D (collected consecutively)	7 vs 7	PC	100	7D
<b>Other patient groups</b>						
[32]	150 women at a urodynamic clinic for assessing urinary symptoms. No reported age data (UK)	One 5D	1 vs 5	RC	n/a	?
[5]	Retrospective study of 84 women presenting to urology clinic 'regardless of pathology or clinical profile initially justifying urological consult'. Mean age 50.5 (18–77) (Canada)	One 7D	1–6 vs 7	Corr from ANOVA	n/a	4D
<b>Asymptomatic patients</b>						
[19]	137 asymptomatic women originally recruited for a study to determine the normal ranges of FVC variables. Median age 39 (18–91) (USA)	Two 1D (interval of 7–9 months)	1 vs 1	Median difference SEM, range	63	1D
[20]	70 asymptomatic women. Mean age 43 (19–81) (Sweden)	One 2D (subgroup of 16 repeated 2D after 1 month)	1 vs 1 2 vs 2	PC, CV	n/a	2D

D, day; I,NDO(I), idiopathic, neurogenic detrusor overactivity (instability); UI, urinary incontinence; PC, Pearson correlation; n/a, not available from the article; CV, coefficient of variation; RC, rank correlation; LA, limits of agreement; CCC, concordance correlation coefficient.

not provide strong evidence. Only 16 of the 151 patients who provided data for the 1-day FVC assessment provided data for the 2-day FVC assessment. It was unclear

on what criteria this subset was selected, and what the characteristics were of this smaller cohort. Compliance data were also lacking because the original sample

size was not stated, and values quoted in some tables (151) were inconsistent with those quoted in the text for completed charts (142).

TABLE 2 A description of the methodological designs of published reliability studies of FVCs (according to patient group). The numbered categories are from Appendix 2

Reference	External validity			Internal validity			Statistical methods		
	1	2	3	4	5	6	7	8	9
<b>Patients with predominantly LUTS</b>									
Bryan and Chapple [16]	N	N	Y	N	Y	N	N	Y	N
Gisolf <i>et al.</i> [3]	N	Y	Y	N	Y	N	Y	N	N
Groutz <i>et al.</i> [17]	Y	Y	Y	N	Y	Y	Y	Y	N
<b>Patients with predominantly incontinence</b>									
Brown <i>et al.</i> [18]	N	Y	Y	Y	Y	Y	Y	Y	N
Larsson <i>et al.</i> [30]	N	Y	Y	N	Y	Y	Y	Y	N
Larsson and Victor [29]	N	N	Y	N	Y	Y	Y	Y	N
Nygaard and Holcomb [31]	N	N	Y	N	Y	N	Y	Y	N
Wyman <i>et al.</i> [6]	N	Y	Y	Y	Y	Y	N	Y	N
Van Melick <i>et al.</i> [4]	N	Y	Y	N	Y	N	Y	N	N
<b>Other patient groups</b>									
Barnick [32]	N	N	N	N	N	N	N	N	N
Schick <i>et al.</i> [5]	N	Y	Y	N	N	N	N	N	N
<b>Asymptomatic patients</b>									
Fitzgerald and Brubaker [19]	N	Y	Y	N	Y	N	Y	N	N
Larsson and Victor [20]	N	N	Y	N	Y	N	Y	Y	N

Y = Yes, N = No.

## DISCUSSION

There is debate about what duration of FVC produces reliable measurements of urinary symptoms in patients with LUTS, and has an acceptable level of compliance. The present review suggests that reliable measurements are obtained from charts over 3–7 days (Table 3). Reliability appears to be lower for FVCs of shorter duration but it is unclear whether or not the reliability of the measurements is still adequate, because the studies are vague about the circumstances in which the FVC might be used. In addition, it is not clear how patient compliance is related to the duration of the FVC. One study of a 7-day FVC reported a compliance rate of 100%, while studies of shorter FVCs reported levels of compliance of <70%. Also, no study explicitly reported levels of compliance for each day on which data were collected.

In this review we also found considerable variation in the methodological quality of papers assessing the reliability of FVCs. Only two reports satisfied all criteria for internal validity, while only one fulfilled all external validity criteria. No study met all statistical criteria and just eight of the 13 used an appropriate method.

There is currently no widely used set of criteria for assessing reliability studies. Our criteria were developed from items commonly used to assess diagnostic studies [40]. While other reviews might adopt a different set of criteria, it is unlikely that our conclusions are sensitive to the adopted set. The criteria all relate to potential sources of bias or factors that can limit the extent to which the results can be applied in other situations. Nonetheless, we recognize that these criteria should be further developed and validated. Like all reviews, the present might include publication bias. We did not include publications that were not peer-reviewed, or abstracts, but it is unlikely that these sources would contain any major study. We also tried to limit this bias by searching the references in located articles, but it is possible that some studies were missed.

An important methodological aspect of the studies was the process of enrolling patients. The sample of patients analysed should reflect those who will be asked to complete the FVC, otherwise the results might be affected by selection bias. This might have arisen in five of the 13 studies because they used data from a clinical trial and the recruited group of patients might be more homogenous than those usually seen in clinical practice. In

addition, selection bias cannot be excluded for three studies that enrolled patients from clinics, because the recruitment process was not clearly described [3,18,32].

Selection bias might have also arisen in three studies that excluded patients with incomplete FVCs from the analysis [3,5,30]. To assess the size of any selection bias, the studies should ideally evaluate what effect including and excluding these incomplete FVCs had on the results. This is particularly important when reliability is measured using a statistic such as an ICC, because its value depends on the distribution of scores and the size of the measurement error, and this selection bias might affect both quantities.

Assessing the stability of voiding characteristics during the study is another important methodological criterion for test-retest reliability studies. There is a natural variation in voiding behaviour and, if the time between measurements is too long, changes in symptoms might cause the degree of measurement error to be overestimated. Most reviewed studies took measurements over 1–4 weeks, although the interval for one study was >6 months [21]. The effect of these different intervals is unclear and requires further investigation.

Changes in voiding behaviour could also arise from asking patients to complete the FVC; this would increase the measurement error, as would variation in how patients complete a FVC. Instructing patients not to change their routines and teaching them how to complete the FVC might reduce these types of measurement error. Such instruction might also improve patient compliance [18] and differences in teaching regimens among the studies could be one reason why there was no clear relationship between the duration of a FVC and the level of patient compliance. Only three studies stated that patients were taught how to complete a FVC [6,18,20]. Another difference in the methods adopted by the studies was the definition of the FVC variables. The most recent ICS definitions should have been used, to allow the results to be compared, but few studies cited these explicitly. Some studies also provided no information on how nocturnal and daytime voids were categorized. Brown *et al.* [18] standardized daytime and nocturnal frequency to an 8-h period but did not justify this unusual approach.

**TABLE 3** Reliability estimates (95% CI) based on 24-h frequency, nocturia, mean voided volume and number of incontinent episodes per day, with the mean (SD) values of the variables, from articles with acceptable statistical methods

Reference	Duration of FVC, days	24-h frequency	Nocturia	Mean voided volume	Incontinent episodes/day	Reliability statistic
<b>Patients with predominantly LUTS</b>						
Bryan and Chapple [16]	3	0.90 (0.83, 0.94)		0.86 (0.77, 0.92)		PC*
Groutz <i>et al.</i> [17]	3	0.83 (n/a)	0.61 (n/a)		0.86 (n/a)	CCC
	1	0.67 (n/a)			0.81 (n/a)	
mean (SD)	3	10.5 (3.9)	1.5 (1.3)		2.5 (3.0)	
<b>Patients with predominantly incontinence</b>						
Brown <i>et al.</i> [18]	7	0.82 (0.76, 0.82)	0.81 (0.75,0.86)		0.83 (0.78, 0.88)	ICC
	3	0.81 (0.74, 0.87)	0.70 (0.60,0.77)		0.81 (0.75, 0.87)	
mean (SD)	7	10.2 (2.7)	1.3 (1.0)		2.4 (2.2)	
Larsson <i>et al.</i> [30]	2	0.72–1.38†		0.72–1.39†		LA
	1	0.64–1.57†		0.68–1.47†		
median	2	9.5 (n/a)		170 (n/a)		
Larsson and Victor [29]	1	0.60–1.67†		0.56–1.79†		
median	1	6.5 (n/a)		220 (n/a)		
Nygaard and Holcomb [31]	7	0.91 (0.87,0.93)			0.83 (0.77, 0.88)	PC
	3	0.89 (0.85, 0.92)			0.91 (0.87, 0.93)	
mean (SD)	7	8.0 (2.8)			3.0 (2.5)	
Wyman <i>et al.</i> [6]	7		0.86 (0.76,0.92)		0.91 (0.85, 0.95)	PC
mean (SD)	7		1.2 (0.8)		2.9 (3.1)	
<b>Asymptomatic patients</b>						
Larsson and Victor [20]	1	0.60 (0.49,0.69)		0.77 (0.70,0.83)		PC
	2†	0.83 (0.56,0.93)		0.93 (0.82,0.98)		
mean (SD)	1	5.8 (1.4)		250 (7.9)		

\*also expressed the reliability estimate in terms of LA for 24-h frequency ( $\pm 2.5$  voids) and mean voided volume ( $\pm 25$  mL). †LA indicating in 95% of cases, the second measurement of the variable is within this range times the first measurement. ‡2-day test in a smaller cohort (16) compared to 1-day test (in 151); n/a, not available from the article; PC, Pearson correlation; LA, limits of agreement; CCC, concordance correlation coefficient.

It was somewhat surprising to find so much variation in the design of the test-retest analysis. Only four studies followed the standard design and collected measurements from two distinct FVCs. The other nine created two sets of measurements from one FVC. This is not problematic in itself, but five of these nine studies included the same data in both sets of measurements, which will artificially increase the observed level of agreement. Another weakness in the methods in all studies was that none justified their sample size, despite there being such formulae for reliability studies [41]. The sample size was 16–300 patients, and the wide CIs for some studies suggest these were under-powered. However, not all studies included CIs for the reliability coefficients. We would argue that such information is essential, to draw inferences about the likely increase in reliability from FVCs of different duration.

In conclusion, deficiencies in the studies that have assessed the reliability of FVCs means

that strong recommendations cannot be made about the optimum duration to use when assessing or monitoring patients with LUTS. Some variables of 3-day FVCs were shown to be reliable in patients with predominantly LUTS, and 3- and 7-day FVCs showed similarly high reliability in patients with predominantly incontinence. This suggests that the current consensus on using FVCs of  $\geq 3$  days might be the most defensible policy. However, while it is 'common sense' to think that compliance rates decline as the duration of a FVC increases, we found no evidence of a link between compliance rates and FVC duration.

As shown by this review, the studies evaluating the reliability of FVCs of different duration contain many weaknesses, some of which might cause reliability to be overestimated. Thus, we propose that future research:

- Ensures that representative patients with LUTS or incontinence are enrolled as participants.

- Gives proper consideration to the stability and description of key patient characteristics and voiding behaviour.
- Uses proper statistical methods for describing test-retest reliability, with adequate sample size calculations and an indication of precision (95% CI).
- Carefully monitors compliance characteristics, with reasons for noncompliance sought.

#### CONFLICT OF INTEREST

None declared. Source of funding: Drexler Foundation, Royal College of Surgeons of England.

#### REFERENCES

- 1 Abrams P, Klevmark B. Frequency-volume charts: an indispensable part of lower urinary tract assessment. *Scand J Urol Nephrol Suppl* 1996; **179**: 47–53

- 2 **Abrams P, Cardozo L, Fall M et al.** The standardisation of terminology of lower urinary tract function: report from the Standardisation Sub-committee of the International Continence Society. *Neurourol Urodyn* 2002; **21**: 167–78
- 3 **Gisolf KW, van Venrooij GE, Eckhardt MD, Boon TA.** Analysis and reliability of data from 24-hour frequency-volume charts in men with lower urinary tract symptoms due to benign prostatic hyperplasia. *Eur Urol* 2000; **38**: 45–52
- 4 **van Melick HH, Gisolf KW, Eckhardt MD, van Venrooij GE, Boon TA.** One 24-hour frequency-volume chart in a woman with objective urinary motor urge incontinence is sufficient. *Urology* 2001; **58**: 188–92
- 5 **Schick E, Jolivet-Tremblay M, Dupont C, Bertrand PE, Tessier J.** Frequency-volume chart: the minimum number of days required to obtain reliable results. *Neurourol Urodyn* 2003; **22**: 92–6
- 6 **Wyman JF, Choi SC, Harkins SW, Wilson MS, Fantl JA.** The urinary diary in evaluation of incontinent women: a test-retest analysis. *Obstet Gynecol* 1988; **71**: 812–7
- 7 **Lamping DL, Schroter S, Marquis P, Marrel A, Duprat-Lomon I, Sagnier PP.** The community-acquired pneumonia symptom questionnaire: a new, patient-based outcome measure to evaluate symptoms in patients with community-acquired pneumonia. *Chest* 2002; **122**: 920–9
- 8 **Lamping DL, Schroter S, Kurz X, Kahn SR, Abenheim L.** Evaluation of outcomes in chronic venous disorders of the leg: development of a scientifically rigorous, patient-reported measure of symptoms and quality of life. *J Vasc Surg* 2003; **37**: 410–9
- 9 **van Trijffel E, Anderegg Q, Bossuyt PM, Lucas C.** Inter-examiner reliability of passive assessment of intervertebral motion in the cervical and lumbar spine: a systematic review. *Man Ther* 2005; **10**: 256–69
- 10 **Audige L, Bhandari M, Kellam J.** How reliable are reliability studies of fracture classifications? A systematic review of their methodologies. *Acta Orthop Scand* 2004; **75**: 184–94
- 11 **Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J.** Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004; **140**: 189–202
- 12 **Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J.** The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003; **3**: 25
- 13 **Streiner DL, Norman GR.** Reliability. In Streiner DL, Norman GR eds, *Health Measurement Scales*, 3rd edn, Chapt. 8. Oxford: Oxford University Press, 2003: 127
- 14 **Lin LI.** A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; **45**: 255–68
- 15 **Bland JM, Altman DG.** Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **1**: 307–10
- 16 **Bryan NP, Chapple CR.** Frequency-volume charts in the assessment and evaluation of treatment: how should we use them? *Eur Urol* 2004; **46**: 636–40
- 17 **Groutz A, Blaivas JG, Chaikin DC et al.** Noninvasive outcome measures of urinary incontinence and lower urinary tract symptoms: a multicenter study of micturition diary and pad tests. *J Urol* 2000; **164**: 698–701
- 18 **Brown JS, McNaughton KS, Wyman JF et al.** Measurement characteristics of a voiding diary for use by men and women with overactive bladder. *Urology* 2003; **61**: 802–9
- 19 **Fitzgerald MP, Brubaker L.** Variability of 24-hour voiding diary variables among asymptomatic women. *J Urol* 2003; **169**: 207–9
- 20 **Larsson G, Victor A.** Micturition patterns in a healthy female population, studied with a frequency/volume chart. *Scand J Urol Nephrol Suppl* 1988; **114**: 53–7
- 21 **Homma Y, Ando T, Yoshida M et al.** Voiding and incontinence frequencies: variability of diary data and required diary length. *Neurourol Urodyn* 2002; **21**: 204–9
- 22 **Dmochowski RR, Sanders SW, Appell RA, Nitti VW, Davila GW.** Bladder-health diaries: an assessment of 3-day vs 7-day entries. *BJU Int* 2005; **96**: 1049–54
- 23 **Matthiesen TB, Rittig S, Mortensen JT, Djurhuus JC.** Nocturia and polyuria in men referred with lower urinary tract symptoms, assessed using a 7-day frequency-volume chart. *BJU Int* 1999; **83**: 1017
- 24 **Mazurick CA, Landis JR.** Evaluation of repeat daily voiding measures in the National Interstitial Cystitis Data Base Study. *J Urol* 2000; **163**: 1208–11
- 25 **Ku JH, Jeong IG, Lim DJ, Byun SS, Paick JS, Oh SJ.** Voiding diary for the evaluation of urinary incontinence and lower urinary tract symptoms: prospective assessment of patient compliance and burden. *Neurourol Urodyn* 2004; **23**: 331–5
- 26 **Sommer P, Bauer T, Nielsen KK et al.** Voiding patterns and prevalence of incontinence in women. A questionnaire survey. *Br J Urol* 1990; **66**: 12–5
- 27 **Jaffe JS, Ginsberg PC, Silverberg DM, Harkaway RC.** The need for voiding diaries in the evaluation of men with nocturia. *J Am Osteopath Assoc* 2002; **102**: 261–5
- 28 **Russell EB, Lee AJ, Garraway WM, Prescott RJ.** Use of a 7-day diary for urinary symptom recording. *Eur Urol* 1994; **26**: 227–32
- 29 **Larsson G, Victor A.** The frequency/volume chart in genuine stress incontinent women. *Neurourol Urodyn* 1992; **11**: 23–31
- 30 **Larsson G, Abrams P, Victor A.** The frequency/volume chart in detrusor instability. *Neurourol Urodyn* 1991; **10**: 533–543
- 31 **Nygaard I, Holcomb R.** Reproducibility of the seven-day voiding diary in women with stress urinary incontinence. *Int Urogynecol J Pelvic Floor Dysfunct* 2000; **11**: 15–7
- 32 **Barnick C.** Frequency/volume charts. In Cardozo, L ed., *Urogynecology: the King's Approach*. New York: Churchill Livingstone, 1997: 104
- 33 **Robb SS.** Urinary incontinence verification in elderly men. *Nurs Res* 1985; **34**: 278–82
- 34 **Lose G, Fantl JA, Victor A et al.** Outcome measures for research in adult women with symptoms of lower urinary tract dysfunction. *Neurourol Urodyn* 1998; **17**: 255–62
- 35 **Burton JR.** Managing urinary incontinence – a common geriatric problem. *Geriatrics* 1984; **39**: 46–51
- 36 **Addla S, Adeyoju A, Neilson D.** Frequency-volume charts – comparison of one, three and seven day charts. *International Continence Society (ICS) Proceedings* 2003: 165
- 37 **Elser DM, Fantl JA, McClish DK.** Comparison of 'subjective' and 'objective' measures of severity of urinary incontinence in women. Program for

- Women Research Group. *Neurorol Urodyn* 1995; **14**: 311–6
- 38 **Locher JL, Goode PS, Roth DL, Worrell RL, Burgio KL.** Reliability assessment of the bladder diary for urinary incontinence in older women. *J Gerontol A Biol Sci Med Sci* 2001; **56**: M32–5
- 39 **Lentz GL, Stanton SL.** Urinary diary: designing a shorter diary. *Int Urogynecol J* 1992; **3**: 69
- 40 **Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J.** Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Med Res Methodol* 2006; **6**: 9
- 41 **Walter SD, Eliasziw M, Donner A.** Sample size and optimal designs for reliability studies. *Stat Med* 1998; **17**: 101–10

**Correspondence:** Tet L. Yap, Clinical Effectiveness Unit, Royal College of Surgeons of England, 35–43 Lincoln's Inn Fields, London WC2A 3PE, UK.  
e-mail: tetyap@doctors.net.uk

**Abbreviations:** FVC, frequency-volume chart; ICC, intra-class correlation coefficient.

## APPENDIX 1

Search strategy used for studies on reliability of FVCs in MEDLINE using medical subject headings (mh), MESH terms (Mesh) and text words (tw). Similar strategy without abbreviations used for CINAHL search.

((((urin\*[tw] AND diary[tw]) OR (urin\*[tw] AND chart[tw]) OR (micturition[tw] AND diary[tw])

OR (micturition[tw] AND chart[tw]) OR (voiding[tw] AND diary[tw]) OR (voiding[tw] AND chart[tw]) OR (frequency[tw] AND volume[tw] AND chart[tw]) OR (frequency[tw] AND chart[tw]) OR (volume [tw]AND chart[tw]) AND medline[sb] AND Humans[Mesh])) OR ((frequency-volume charts))

AND

((reliability[tw]) OR (test retest[tw]) OR (duration[tw]) OR (observer variation[mh]) OR (reproducibility[tw]) OR (concordance[tw]) OR (repeatability [tw] OR agreement[tw] OR variation\*[tw] OR variabilit\*[tw]))))

## APPENDIX 2

The criteria list for assessing the methodological quality of studies on the reliability of FVCs

1. Were cases representative of the patients who would receive the FVC? The cases were considered representative when authors reported that selection was based on a consecutive series of patients, or patients randomly selected from a list. If selection was based on the quality of the returned FVCs (e.g. whether they were incomplete or wrongly completed), the sample was considered not representative. Trials that used patients enrolled in other trials are also deemed not representative.
2. Were selection criteria of the cases clearly described? The selection criteria were considered adequately described if authors

provided clear and valid inclusion and exclusion criteria. Criteria deemed valid included those describing patient symptoms relevant to the trial, age of patient, methods of diagnosis, any treatment received or any causes of symptoms.

3. Was the description or example provided of the FVC used? Clear descriptions of the components of the FVC that were used, including sleep/wake time recordings and types of chart variables analysed, were sought.

4. Were the cases taught how to complete the FVC?

5. Were the characteristics of the cases under study stable during research?

6. Were case compliance and withdrawals adequately described? Case compliance and withdrawals were judged to be adequately described if case withdrawals and values for incompletely or wrongly completed FVCs were clearly reported, and reasons for any noncompliance.

7. Were key characteristics of patients presented? The reliability of the FVC depends upon the population in which it is used. Descriptions should, as a minimum, include all of the following key characteristics: description of voiding behaviour; mean and distribution of symptoms; and age.

8. Was an appropriate method used for calculating reliability?

9. Was the size of the sample justified by the authors?