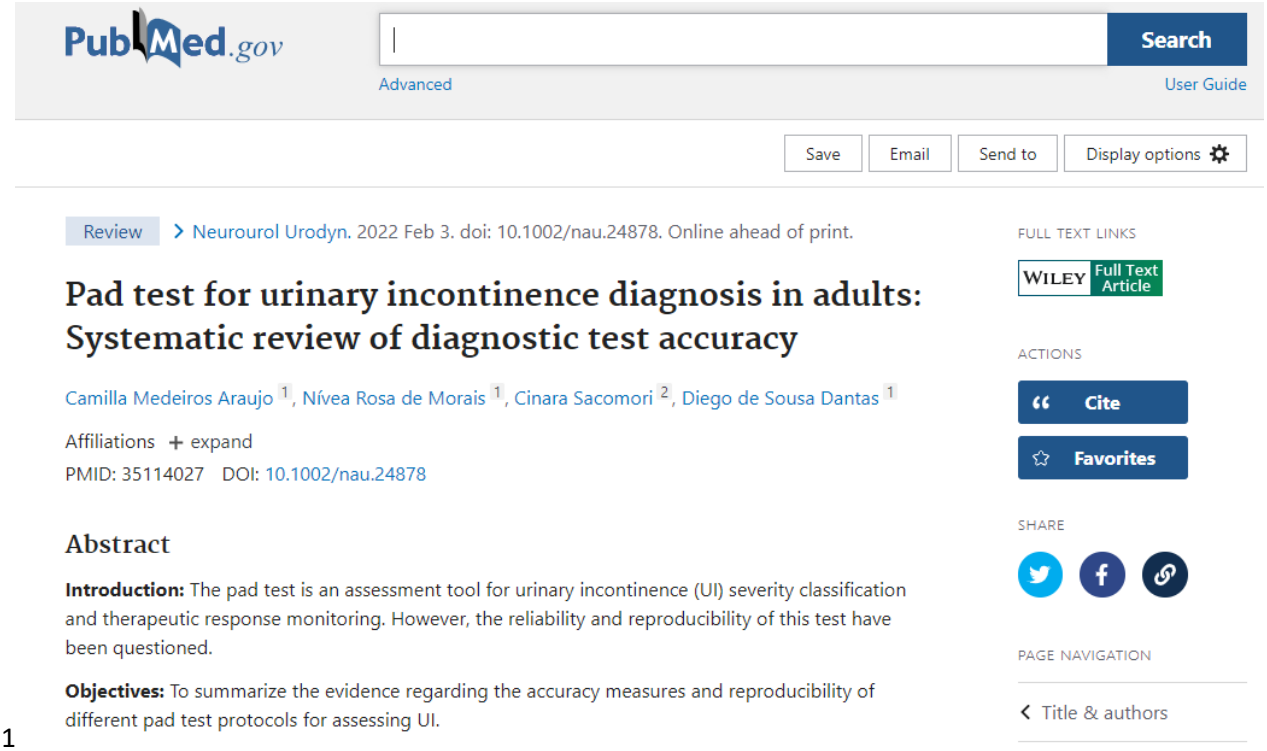

Accepted version, published at Neurourol Urodyn. 2022;1–13.

Pad test for urinary incontinence diagnosis in adults: Systematic review of diagnostic test accuracy

Camilla Medeiros Araujo1 | Nívea Rosa de Morais1 | Cinara Sacomori2 | Diego de Sousa Dantas1

1Physiotherapy Department, Universidade Federal de Pernambuco (UFPE), Recife, Brazil

2Facultad de Medicina Clinica Alemana, Universidad del Desarrollo, Santiago, Chile

Abstract

Introduction: The pad test is an assessment tool for urinary incontinence (UI) severity classification and therapeutic response monitoring. However, the reliability and reproducibility of this test have been questioned. **Objectives:** To summarize the evidence

regarding the accuracy measures and reproducibility of different pad test protocols for assessing UI more objectively. **Methods:** A systematic review of the diagnostic accuracy of this tool was performed (CRD42020219392). **Eligibility criteria**: Studies reporting data on the accuracy measures and reproducibility of the pad test when used for detecting UI in adult men and women. **Data Sources:** MEDLINE, Science Direct, Cochrane Database, Web of Science, LILACS, and Pedro databases. **Data extraction and synthesis:** Two reviewers independently screened the articles, extracted the data and evaluated the risk of bias using the QUADAS-2 tool. **Results**: From 1048 studies, 18 studies were included. Eight of these reported accuracy data, and 12 reported reproducibility properties. A total of 1070 individuals were analysed, whose mean age ranged from 20 to 90 years. The accuracy of the long-duration protocols was generally moderate to high (sensitivity, 60%–93%; specificity, 60%–84%). The 1-hour protocols obtained higher accuracy values than the long-duration protocols. The overall reproducibility was moderate to high ($\kappa \geq 0.66$). **Limitations:** The risk of bias among the studies was high and, due to different cut-off points adopted by studies, the bivariate model was not satisfied to perform a meta-analysis. **Discussion**: The 1-hour pad test was more accurate but less reproducible when compared to the long-duration tests. Pad test results should be used with caution in clinical practice.

Keywords: Urinary incontinence; Severity of illness index; Data accuracy; Reproducibility of results; Measures of association, exposure, risk or outcome.

Introduction

Urinary incontinence (UI) is a clinical condition that may affect men and women from children to older adults[1]. Diagnosis of UI severity is made clinically and with the urodynamic test as the gold standard in evaluating lower urinary tract function [2]. However, the limiting factors of the urodynamic test include its high cost and invasiveness [3]. Practising clinicians may also use several validated questionnaires, voiding diaries, and the pad test[4] to identify UI and associated symptoms.

The pad test is useful in clinical and scientific practice as an additional tool in the diagnosis and classification of UI severity, and as such indirectly influences the choice of treatment and therapeutic response monitoring[4],[5]. It is recommended by the International Continence Society (ICS) with grade C level 3 evidence[4],[5] standing out for being non-invasive, relatively easy, inexpensive, and for presenting a good correlation with UI assessment questionnaires [6].

Proposed by James et al. in 1971[7], the pad test was intended to assess UI by weighing pads. Ten years later, it was associated with stimulation exercises by Sutherst et al[8]. The more widespread protocols are divided into four types: the 20-min [9], standardized 1-hour [10], 24-hour [11], and 48-hour protocols [12]. However, the reliability and reproducibility of this test have been questioned, as the short duration tests are restricted to experimental situations on urethral insufficiency, offering a limited impression of daily urinary loss situations [13]. The longer duration tests (24-, 48-, and 72-hour) seem to be more reproducible, but with lower patient compliance [14]. In the long duration tests, patients are exposed to daily life situations that might increase intra-abdominal pressure

or provoke urgency. Meanwhile, in the short-duration tests patients are exposed to a standardized set of activities [14].

In clinical practice, the decision-making process to choose any of these tools should be based on parameters of test accuracy related to a gold standard. These parameters are expressed by sensitivity, specificity, positive and negative predictive values, and positive and negative likelihood ratios. Reproducibility is another relevant parameter to estimate the concordance of measures of test and recommend its use[15].

The measurement properties of the pad test have been investigated, producing conflicting results. Studies show a low correlation between the urodynamic examination results and the 24-hour pad test results [16]. The results of comparing the 1-hour and 24-hour protocols showed a good correlation (0.86) but poor agreement (Cohen's kappa coefficient $\kappa = 0.25$) in UI severity classification [13].

In this context, this review aims to summarize the evidence of the accuracy measures and reproducibility of different pad test protocols in detecting more objectively UI in men and women, as well as to investigate the quality of this evidence.

Methods

This systematic review was registered in the Prospective Register of Systematic Reviews (PROSPERO) under number CRD42020219392. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies[17], and the Cochrane Handbook for Systematic Review of Diagnostic Test Accuracy[18] were followed.

Eligibility Criteria

Any cross-sectional or prospective studies reporting data on the accuracy measures and reproducibility properties of any pad test protocol for objectively detecting UI or measuring loss of urine in adult women and men (≥ 18 years) were eligible for this systematic review. The studies included cover a period from 1971, the publication year of the first version of the pad test [7], to the search date (December 08, 2021). There was no language limitation.

Narrative and systematic reviews, letters to the editor, editorials, case studies, and case series, as well as studies that did not report data on accuracy measures, or that presented insufficient data to calculate these values, were excluded.

Diagnostic test

We included articles that reported the clinimetric properties of pad tests of 20 minutes, 1h, 24h, and 48 hours, as well as studies that dealt with pad test reproducibility measures. Urodynamic[2] and/or videourodynamic[19] studies were considered the gold standard in diagnosing UI.

Search for articles and data collection

One of the researchers searched for articles using MEDLINE via the EBSCO, Science Direct, Cochrane Database, Web of Science, PubMed, and Pedro databases. The following descriptors were used: pad test, accuracy, sensitivity, specificity, reproducibility of results [MeSH], reliability, and urinary incontinence, with the Boolean operators AND/OR. The search strategy (supplementary material) was differentiated for each database and involved combining the descriptors mentioned above. In addition, relevant articles identified from the references were included.

Two reviewers independently screened and analysed the eligibility of the articles. After excluding duplicate entries, the studies were screened by reading the titles and abstracts, and were assessed for eligibility criteria by reading the full text. A third reviewer resolved disagreements for the final consensus.

The researcher extracted the data based on an electronic questionnaire developed for the study containing the following: authors, year of publication, study design, sample size, participants' sex, mean age, UI type, type of pad test performed, true and false negatives and positives, sensitivity and specificity values with confidence intervals, statistical test, and interval between the performance of the tests in days. The data were then registered in tables.

Outcome measures

The primary outcomes were the accuracy values (sensitivity and specificity), and the secondary outcomes were the reproducibility measures (test-retest measures, intraclass correlation coefficient, agreement, and standard deviation).

Assessment of the risk of bias and applicability

The risk of bias (RoB) of studies investigating diagnostic accuracy was verified using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool[20], which comprises four domains: participant selection, index test, standardised reference test, and participant flow. Each domain was evaluated as to the ROB and also the clinical applicability concerning the participants and tests evaluated.

The Guideline for Reporting Reliability and Agreement Studies (GRRAS) was used to evaluate the RoB of the reproducibility measures. The GRRAS addresses 15 items related to the title and abstract, introduction, methodology, presentation of results, discussion, and auxiliary material [21]. A score was developed using the number of accomplished items divided by the total number of applicable items by each study. Higher scores demonstrate that the studies report more reliable results.

Data analysis

The outcomes were analysed using Review Manager (RevMan) (version 5.4.1; Cochrane Collaboration, 2020). Non-explicit sensitivity and specificity values were automatically calculated by RevMan from the number of true-positive, true-negative, false-positive, and false-negative cases. Thus, we recreated the 2×2 contingency tables and calculated sensitivity, specificity, and positive and negative predictive values, as well as the 95% confidence intervals (CIs).

The inter-rater agreement in the article selection phase was analysed using the calculated κ from the observed and random agreements, which was interpreted as follows: 0 to 0.20 (no agreement), 0.21 to 0.39 (minimal), 0.49 to 0.59 (weak), 0.60 to 0.79 (moderate), 0.80 to 0.90 (strong)and above 0.90 (near perfect) [22].

Results

Selection of studies

The process of study selection has been represented in a flow diagram (Figure 1). After the screening process, 18 studies were included in the qualitative synthesis, among which eight addressed accuracy measures, and 12 presented reproducibility measures.

The screening process showed moderate agreement between the evaluators (κ = 0.702).

Characteristics of the included studies

A total of 1070 individuals were included in the studies evaluating the accuracy of the pad test (Table 1), with 617 participants included in the studies that evaluated the reproducibility of the test (Table 2).

Risk of bias

The RoB and applicability are presented graphically in percentages and detailed by study (Figure 2).

The selection of participants was the primary source of bias in the included studies, as most studies, except that of Matharu et al. [23], only included volunteers with incontinence complaints.

The studies also showed biases in interpreting the test results. In three studies [24–26] it was not clearly described if the evaluator was blinded, i.e. whether the pad test was analysed without knowing the urodynamic test result. In another three studies [27–29], the participants performed the pad test after knowing the urodynamic study result.

Regarding flow and timing Bias, the interval between tests was poorly described in the studies by Karantanis et al. [24] and Price and Noblett [26]. Data were not completely

analysed in the study by Versi et al. [27] with only healthy and stress UI-diagnosed volunteers included in the urodynamic study.

The studies generally showed good applicability of concepts concerning participant selection, index testing, and reference testing. We highlight that only Timmermans et al. [25] included men in their sample. Only the study by Matharu et al.[23] showed low applicability because the researchers modified the 1-hour pad test protocol and excluded the activity of climbing a flight of stairs, considering the participants' physical limitations.

None of the 12 included studies [9,11,24,30–38] on pad test reproducibility met the criteria of all 15 items recommended by GRRAS (Table 3). The main inadequacies regarding methodology were related to sample selection and calculation, the number of raters and repeated measures, and whether the evaluations and interpretations occurred independently. Regarding the description of the results, most did not present the characteristics of the sample of raters and participants or included measures of statistical uncertainty.

Summary of results

The synthesis of the results is described by the type of pad test (Figure 3). A meta-analysis of the studies was not possible because of differences in the cut-off point between the long-duration studies, and the small number of short-duration studies which also presented some methodological limitations. Some studies did not present the sensitivity/specificity values [23–25,39]; therefore, they were calculated and presented in the Forest Plot (Figure 3).

48-hour pad test

Only the study by Versi et al.[27] analysed the accuracy properties of the 48-hour pad test among 105 participants. The cut-off point value for UI diagnosis at 15 g had a sensitivity of 92% (95% CI, 82%–97%) and specificity of 84% (95% CI, 68%–94%), compared to that of the urodynamic test.

Three studies[27,33,37] presented intra-examiner reproducibility analysis with coefficients ranging from 0.87 to 0.94.

24-hour pad test

Three studies[23,24,26] analysed the properties of the 24-hour pad test among a total of 504 participants. However, each study analysed a different cut-off point. When the 1-g cut-off point[24] was adopted, a sensitivity of 100% was estimated, but specificity could not be verified. For the 4-g cut-off point[23], sensitivity of 63% (95% CI, 57%–69%) and specificity of 64% (95% CI: 49%–77%) were estimated. At the 8-g cut-off point[26] sensitivity and specificity were both 60% (95% CIs, 45%–74% and 15%–95%, respectively).

The reproducibility was analysed by seven studies[11,27,31,33,34,37], calculating correlation coefficients between 0.66 and 0.9. Variabilities of 69.5–133.4 g was identified between consecutive pad tests, depending on the level of physical activity of the participant (95% CI, 80.4–186.5 g) [31].

1-hour pad test

Four studies[23,25,28,39], with a total of 719 participants, analysed the properties of the 1-hour pad test. The cut-off point value utilized was 1 g, and the sensitivity and specificity values ranged 34%–95% and 60%–100%, respectively**.**

The reproducibility of the 1-hour pad test was evaluated by two studies[32,38] with a correlation coefficient of 0.96 between the two measurements[32]. In contrast, a mean difference of 9.7 ± 29.7 g was identified by the Bland Altman analysis, with the second measurement varying from 66 g below to 46 g above the first test result [38].

20-min pad test

Only two studies[28,29] with a total of 183 participants analysed the properties of the 20-min pad test. The cut-off point value used was 1 g, and identified sensitivities between 46% (95% CI, 36%–56%) and 71% (95% CI, 60%–81%). Specificity could not be estimated or reported in the studies because of participant selection bias.

Three studies[9,30,36] analysed the reproducibility and presented correlation values between pad test measurements ranging from 0.75 to 0.94. The method error was 30% in the study by Hahn and Fall[9].

Discussion

This review summarized the data of the long- and short-duration pad tests accuracy and reproducibility in assessing UI and measure loss of urine in adults. Eighteen studies were identified, most of them involving women. The lack of studies and reference values for the male population draws attention because the pad test is recommended in assessing UI in men, having been adopted as the gold standard in the validation of instruments related to UI severity [13]; therefore there is still a need to study clinimetric properties of this test in men.

The cut-off values of the studies were similar only for the 20-min and 1-hour pad tests, allowing comparison between these tests. The included studies used a 1g cut-off in increase in pad weight to make a diagnosis according to the ICS standardization[14]. However, the lower limit should be interpreted with prudence because weight gains of up to 1.4 g or 4.4 g for 1 hour or 24-hr pad test, respectively, could be a result of sweating or vaginal discharge. If the findings are inconclusive, clinicians may use oral phenazopyridine to color the urine orange and distinguish the urine from other fluids[14].

The identified accuracy measures were quite variable and associated with a strong tendency to be overestimated due to research bias, mainly due to participant selection bias[20]. The wide confidence intervals concerning the specificity measure point to the imprecision of the test, probably due to the selection bias [40]. Several studies have used the pad test as the diagnostic gold standard[4,41]; however, they present divergent accuracy measures.

The accuracy of the 1-hour pad test was the most investigated; among the studies, the one by Costantini et al.[39] stood out as having the lowest RoB, presenting a sensitivity and specificity of 84% and 70%, respectively. The 20-minute pad test presented a sensitivity of only 46% in studies with a lower RoB, and its specificity could not be estimated due to selection bias.

Among the long-term protocols, the 24-hour pad test was the most investigated and presented both sensitivity and specificity values of approximately 60% in studies with a lower RoB. For the 1-g cut-off point[24], a sensitivity of 100% was estimated, but the studies involved only participants already diagnosed with UI, tending to overestimate the

test's specificity[20]. Only one study evaluated the 48-hour pad test and had the most robust results among the four tests (Figure 3) numerically; however, we highlighted the high RoB present in the study for the four domains evaluated.

Regarding reproducibility, the long duration tests seem to be more reproducible than the short duration tests, presenting more reliable measurements, ranging from 0.85 to 0.94, indicating strong levels of agreement[22]. Factors such as absorbency, evaporation, level of physical activity, and fluid intake may modify their results every 4–6 hours, thereby increasing costs [24,31].

The reproducibility of the short-duration pad tests was the least investigated; the 1-hour protocol showed stronger reliability (0.96) compared to the 20-min protocol (0.75). The 20-min protocol presented lower standard deviations, indicating better agreement, than the 1h protocol, but with variation coefficients of 157%–326%[36] which can compromise the reliability of measures. The Bland-Altman analysis is not appropriate for comparing repeated measurements, except by adding a random-model effect to the analysis, which was not carried out in the selected studies [42].

We suggest that, both for research and clinical practice, the results of the 1-hour pad test should be interpreted with caution, and repeated measurements should be performed, as recommended by the ICS [43]**,** especially in cases where the results do not correlate with the UI symptoms as reported by the patient[44]. The wide variation in agreement between repeated measurements affects the reproducibility of the test, as it may cause inaccuracies in classifying incontinence severity, consequently affecting the prognosis and evaluation of therapeutic response[45]. The association of the pad test with

other evaluation measures (urodynamics, questionnaires) is recommended for a better comprehension of UI and its underlying factors, as well as for the follow-up of its clinical evolution.

The main methodological limitations were the absence of sample calculation and poor description of the type of sampling, the number of evaluators, and repeated measures, as well as unclear evaluations and interpretations of the results, leading to a risk of overestimating the measurement and compromising the verification of the measurement error [21]. Regarding the included studies, most did not present the characteristics of the evaluators and subjects, contributing to the unfeasibility of analysing external validity. Most of them did not include measures of statistical uncertainty, providing limited measures and restricting the interpretation of results[46].

To the best of our knowledge, this is the first review that has synthesized the accuracy, reliability, and reproducibility of the pad test and assessed the quality of evidence, as a previous review only sought to verify the concepts and applications of the pad test[47]. We highlight the strong agreement in the selection processes of the studies and the robust methods used to synthesize and more accurately evaluate the available evidence and analyse the data. Urinary incontinence might be transitory or persistent with dynamic changes in the severity scores because it is a symptom of an underlying condition. The pad test could be an accessible tool to detect these changes.

The study limitations were as follows: the process of data extraction was made by a single researcher; no primary studies were developed with elderly populations or comparing different ages; the lack of a “gold standard” comparator is also a limitation as

not all the studies used the urodynamic test. In addition, UI is a dynamic condition, and the urodynamic test also evaluates an inherently variable biological system. The poor methodological description of the primary studies made it impossible to judge some biases on data extraction; and it was not possible to present meta-analysis for summarising sensitivity/specificity values because the studies presented significant methodological bias, different cut-off points between the long-duration protocols, and were insufficient to satisfy the bivariate model for meta-analysis[48].

This review demonstrated that there are significant biases in the methodological conduct of pad test accuracy and reproducibility studies, despite it being a widely disseminated and used clinical tool. Therefore, future research on the accuracy of the pad test should follow the recommendations of the Standards for Reporting of Diagnostic Accuracy Studies guidelines [49]. Reproducibility studies should follow the GRRAS checklist[21] for better designs and measurement safety.

Conclusion

The evidence regarding the use of the pad test as a more objective evaluation tool for urinary incontinence is divergent, and the studies present a high risk of bias. Thus, they should be interpreted with caution. Among the analysed protocols, the 1-hour pad test showed better accuracy but poorer reproducibility compared to the long-duration tests. The 20-min pad test was found to be the least recommendable because it has less support in the literature and low accuracy.

New studies with greater methodological rigor, incorporating man, and in compliance with international guidelines need to be developed to achieve more reliable results.

References

1. Gibson W, Wagg A. Incontinence in the elderly, “normal” ageing, or unaddressed pathology? *Nat Rev Urol*. 2017;14(7):440-448. doi:10.1038/nrurol.2017.53

2. Drake MJ, Doumouchtsis SK, Hashim H, Gammie A. Fundamentals of urodynamic practice, based on International Continence Society good urodynamic practices recommendations. *Neurourol Urodyn*. 2018;37(S6):S50-S60. doi:10.1002/nau.23773

3. Monteiro MV de C, Fonseca AMRM da, Silva Filho AL. Valor do estudo urodinâmico no tratamento da incontinência urinária. *Femina*. 2012.

4. Kopp DM, Bengtson AM, Tang JH, Chipungu E, Moyo M, Wilkinson J. Use of a postoperative pad test to identify continence status in women after obstetric vesicovaginal fistula repair: a prospective cohort study. *BJOG*. 2017;124(6):966-972. doi:10.1111/1471-0528.14505

5. Yi YA, Keith CG, Graziano CE, et al. Strong correlation between standing cough test and 24-hour pad weights in the evaluation of male stress urinary incontinence. *Neurourol Urodyn*. 2020;39(1):319-323. doi:10.1002/nau.24200

6. Abdel-fattah M, Barrington JW, Youssef M. The standard 1-hour pad test: does it have any value in clinical practice? *Eur Urol*. 2004;46(3):377-380. http://search.ebscohost.com/login.aspx?direct=true&db=mdc&AN=15306111&lang=pt-br&site=ehost-live&authtype=ip,cookie,uid.

7. James, ED; Flack, FC; Caldwell, KP; Martin M. Continuous measurement of urine loss and frequency in incontinent patients. Preliminary report. *Br J Urol*. 1971;43(2):233-237.

8. Sutherst JR, Brown MC, Richmond D. Analysis of the pattern of urine loss in women with incontinence as measured by weighing perineal pads. *Br J Urol*. 1986;58(3):273-278. http://search.ebscohost.com/login.aspx?direct=true&db=mdc&AN=3719248&lang=pt-br&site=ehost-live&authtype=ip,cookie,uid.

9. Hahn I, Fall M. Objective quantification of stress urinary incontinence: A short, reproducible, provocative pad-test. *Neurourol Urodyn*. 1991;10(5):475-481. doi:10.1002/nau.1930100503

10. Jørgensen L, Lose G, Andersen JT. One-Hour pad-weighing test for objective assessment of female urinary incontinence. *Obstet Gynecol*. 1987;69(1):39-42.

doi:10.1016/0020-7292(87)90383-3

11. Lose G, Jørgensen L, Thunedborg P. 24-hour home pad weighing test versus 1-hour ward test in the assessment of mild stress incontinence. *Acta Obstet Gynecol Scand*. 1989;68(3):211-215. doi:10.3109/00016348909020991

12. Jakobseny H, Vedel P, Andersen JT. Objective assessment of urinary incontinence: An evaluation of three different pad-weighing tests. *Neurourol Urodyn*. 1987;6(4):325-330. doi:10.1002/nau.1930060407

13. Soto González M, Da Cuña Carrera I, Lantarón Caeiro EM, Gutiérrez Nieto M, López García S, Ojea Calvo A. Correlation between the 1-hour and 24-hour pad test in the assessment of male patients with post-prostatectomy urinary incontinence. *Progrès en Urol*. 2018;28(11):536-541. doi:https://doi.org/10.1016/j.purol.2018.06.011

14. Krhut J, Zachoval R, Smith PP, et al. Pad weight testing in the evaluation of urinary incontinence. *Neurourol Urodyn*. 2014;33(5):507-510. doi:10.1002/nau.22436

15. Buehler AM, Ascef B de O, Oliveira Júnior HA de, Ferri CP, Fernandes JG. Rational use of diagnostic tests for clinical decision making. *Rev Assoc Med Bras*. 2019;65(3):452-459. doi:10.1590/1806-9282.65.3.452

16. Henderson JW, Kane SM, Mangel JM, et al. A Randomized Comparative Study Evaluating Various Cough Stress Tests and 24-Hour Pad Test with Urodynamics in the Diagnosis of Stress Urinary Incontinence. *J Urol*. 2018;199(6):1557-1564. doi:10.1016/j.juro.2017.11.073

17. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies. *JAMA*. 2018;319(4):388. doi:10.1001/jama.2017.19163

18. Deeks JJ, Wisniewski S DC. Guide to the contents of a Cochrane Diagnostic Test Accuracy Protocol. In: *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0. The*. ; 2013:3-15.

19. Jiang YH, Chen SF, Kuo HC. Role of videourodynamic study in precision diagnosis and treatment for lower urinary tract dysfunction. *Tzu Chi Med J*. 2020;32(2):121-130. doi:10.4103/tcmj.tcmj_178_19

20. Reitsma JB, Leeflang MMG, Sterne JAC, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med*. 2011;155(4):529-536.

21. Kottner J, Audigé L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J Clin Epidemiol*. 2011;64(1):96-106. doi:10.1016/j.jclinepi.2010.03.002

22. McHugh ML. Lessons in biostatistics interrater reliability : the kappa statistic. *Biochem Medica*. 2012;22(3):276-282.

23. Matharu GS, Assassa RP, Williams KS, et al. Objective Assessment of Urinary Incontinence in Women: Comparison of the One-Hour and 24-Hour Pad Tests. *Eur Urol*. 2004;45(2):208-212. doi:https://doi.org/10.1016/j.eururo.2003.09.006

24. Karantanis E, Allen W, Stevermuer TL, Simons AM, O'Sullivan R, Moore KH. The repeatability of the 24-hour pad test. *Int Urogynecol J Pelvic Floor Dysfunct*. 2005;16(1):63-68. http://search.ebscohost.com/login.aspx?direct=true&db=mdc&AN=15647965&lang=pt-br&site=ehost-live&authtype=ip,cookie,uid.

25. Timmermans L, Mélot C, Wespes E, Falez F. Analyse de la valeur du questionnaire ICIQ-sf et du test d'incontinence sur une heure pour le diagnostic d'incontinence urinaire objectivé par un bilan urodynamique. *Pelvi-perineologie*. 2011;6(3-4):162-165. doi:10.1007/s11608-011-0373-y

26. Price DM, Noblett K. Comparison of the cough stress test and 24-h pad test in the assessment of stress urinary incontinence. *Int Urogynecol J*. 2012;23(4):429-433. doi:10.1007/s00192-011-1602-1

27. Versi E, Orrego G, Hardy E, Seddon G, Smith P, Anand D. Evaluation of the home pad test in the investigation of female urinary incontinence. *BJOG An Int J Obstet Gynaecol*. 1996;103(2):162-167. doi:10.1111/j.1471-0528.1996.tb09669.x

28. Wu WY, Sheu BC, Lin HH. Comparison of 20-minute pad test versus 1-hour pad test in women with stress urinary incontinence. *Urology*. 2006;68(4):764-768. doi:10.1016/j.urology.2006.04.018

29. Wu W-Y, Sheu B-C, Lin H-H. Twenty-minute pad test: Comparison of infusion of 250ml of water with strong-desire amount in the bladder in women with stress urinary incontinence. *Eur J Obstet Gynecol Reprod Biol*. 2008;136(1):121-125. doi:https://doi.org/10.1016/j.ejogrb.2007.02.030

30. MacHold S, Olbert PJ, Hegele A, Kleinhans G, Hofmann R, Schrader AJ. Comparison of a 20-min pad test with the 1-hour pad test of the international continence society to evaluate post-prostatectomy incontinence. *Urol Int*. 2009;83(1):27-32. doi:10.1159/000224864

31. Malik RD, Cohn JA, Fedunok PA, Chung DE, Bales GT. Assessing variability of the 24-hour pad weight test in men with post-prostatectomy incontinence. *Int Braz J Urol*. 2016;42(2):327-333. doi:10.1590/S1677-5538.IBJU.2014.0506

32. Klarskov P, Hald T. Reproducibility and reliability of urinary incontinence assessment with a 60 min test. *Scand J Urol Nephrol*. 1984;18(4):293-298. doi:10.3109/00365598409180199

33. Victor A, Larsson G, Åsbrink AS. A simple patient-administered test for objective quantitation of the symptom of urinary incontinence. *Scand J Urol Nephrol*. 1987;21(4):277-279. doi:10.3109/00365598709180782

34. Mouritsen L, Berild G, Hertz J. Comparison of different methods for quantification

of urinary leakage in incontinent women. *Neurourol Urodyn*. 1989;8(6):579-586. doi:10.1002/nau.1930080604

35. Versi E, Cardozo L, Anand D, Cooper D. Symptoms analysis for the diagnosis of genuine stress incontinence. *Br J Obstet Gynaecol*. 1991;98(8):815-819. doi:10.1111/j.1471-0528.1991.tb13488.x

36. Siltberg H, Larsson G, Victor A. Reproducibility of a new method to determine cough-induced leak-point pressure in women with stress urinary incontinence. *Int Urogynecol J Pelvic Floor Dysfunct*. 1996;7(1):13-19. http://search.ebscohost.com/login.aspx?direct=true&db=mdc&AN=8798081&lang=pt-br&site=ehost-live&authtype=ip,cookie,uid.

37. 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(3 Part 1):698-701. doi:10.1016/S0022-5347(05)67284-9

38. Simons AM, Yoong WC, Buckland S, Moore KH. Inadequate repeatability of the one-hour pad test: the need for a new incontinence outcome measure. *BJOG*. 2001;108(3):315-319. http://search.ebscohost.com/login.aspx?direct=true&db=mdc&AN=11281474&lang=pt-br&site=ehost-live&authtype=ip,cookie,uid.

39. Costantini E, Lazzeri M, Bini V, Giannantoni A, Mearini L, Porena M. Sensitivity and specificity of one-hour pad test as a predictive value for female urinary incontinence. *Urol Int*. 2008;81(2):153-159. doi:10.1159/000144053

40. Bossuyt P, Davenport C, Deeks J, Hyde C, Leeflang M, Scholten R. Chapter 11 Interpreting results and drawing conclusions. In: *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*. ; 2013:1-31.

41. Tsui JF, Shah MB, Weinberger JM, et al. Pad count is a poor measure of the severity of urinary incontinence. *J Urol*. 2013;190(5):1787-1790. doi:10.1016/j.juro.2013.05.055

42. Doğan NÖ. Bland-Altman analysis: A paradigm to understand correlation and agreement. *Turkish J Emerg Med*. 2018;18(4):139-141. doi:10.1016/j.tjem.2018.09.001

43. Abrams P, Cardozo L, Wagg A, Wein A. *Epidemiology of Urinary Incontinence (UI) and Other Lower Urinary Tract Symptoms (LUTS), Pelvic Organ Prolapse (POP) and Anal Incontinence (AI)*.; 2016. https://www.ics.org/publications/ici_6/Incontinence_6th_Edition_2017_eBook_v2.pdf.

44. Soroka D, Drutz HP, Glazener CMA, Hay-Smith EJC, Ross S. Perineal pad test in evaluating outcome of treatments for female incontinence: a systematic review. *Int Urogynecol J Pelvic Floor Dysfunct*. 2002;13(3):165-175. http://search.ebscohost.com/login.aspx?direct=true&db=mdc&AN=12140710&lan

g=pt-br&site=ehost-live&authtype=ip,cookie,uid.

45. Hess AS, Hess JR. Understanding standard deviations and standard errors. *Transfusion*. 2016;56(6):1259-1261. doi:10.1111/trf.13625

46. O'Brien SF, Yi QL. How do I interpret a confidence interval? *Transfusion*. 2016;56(7):1680-1683. doi:10.1111/trf.13635

47. Maranhão TM de O, Bernardo-Filho M, Galvão LLLF, Micussi MTABC. Pad-test: uma revisão sobre conceitos e utilização. *Femina*. 2008:243-246.

48. Macaskill P, Gatsonis C, Deeks J, Harbord R, Takwoingi Y. Chapter 10 Analysing and Presenting Results. In: *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*. ; 2013.

49. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ*. 2016;6:1-17. doi:10.1136/bmjopen-2016- 012799

482Table 1. Extraction of data from studies on PadTest accuracy for 20-min, 1-hour, 24-hour and 48-hour pad tests

Author	Year	Sex (age; range or MD ± SD)	Type of UI	Design	n	Pad test	VP	VN	FP	FN	S (95% CI)	E (95% CI)
Versi et al.	1996	F (absent)	SUI (105), UUI/MUI (44)	Prospective	105**	48 h	57	31	12	5	0.93 (0.84–0.98)	0.84 (0.68–0.94)
Matharu et al.	2004	F (>40) *	SUI (132), UUI (83), MUI (79)	Cross-sectional	341	1 h	205	28	19	89	0.70 (0.64–0.75)	0.60 (0.44–0.74)
						24 h	186	30	17	108	0.63 (0.57–0.69)	0.64 (0.49–0.77)
Karantanis et al.	2005	F (57; 46–68)	SUI (66), MUI (30), UUI (6), others (6)	Cross-sectional	108	24 h	108	0	0	0	1.00 (0.97–1.00)	Not estimable
Price and Noblett	2012	F (56 ± 10)	SUI (53), UUI (2)	Prospective	55	24 h	30	3	2	20	0.60 (0.45–0.74)	0.60 (0.15–0.95)
Wu, Sheu and Lin	2008	F (53.6 ± 2,6)	SUI (64), UUI (1), MUI (18)	Cross-sectional	83	20 min	59	0	24	0	0.71 (0.60–0.81)	Not estimable
Wu, Sheu and Lin	2006	F (53.3 ± 12)	SUI (72), UUI (11), MUI (17)	Cross-sectional	100	20 min	46	0	54	0	0.46 (0.36–0.46)	Not estimable
						1 h	34	0	0	66	0.34 (0.25–0.44)	Not estimable
Costantini et al.	2008	F (69, 20–90)	SUI (43), MUI (45), UUI (16)	Cross-sectional	158	1 h	87	38	16	17	0.84 (0.75–0.90)	0.70 (0.56–0.82)
Timmermans et al.	2011	F/M (58 ± 15)	UUI (42), SUI (39), MUI (25)	Cross-sectional	120	1 h	101	14	0	5	0.95 (0.89–0.98)	1.00 (0.77–1.00)

483MD, mean difference; SD, standard deviation; UI, urinary incontinence; VP: true positive; VN, true negative; FP, false positive; FN, false
484negative; S, sensitivity; E, specificity, SUI, stress urinary incontinence; UUI, urge urinary incontinence; MUI, mixed urinary incontinence.
485*Patients over 40 years. **Only cases of SUI were analysed

486

487

488Table 2. Extraction of data from the studies on the reproducibility of the 20-min, 1-hour, 24-hour, and 48-hour pad test.

Author	Year	Sex (Age, range; or MD ± SD)	Type of UI	n	pad test	Reproducibility (95% CI)	Interval (days)	Statistical measure
Klarskov and Hald	1984	F/M (52, 17–75)	SUI, UUI, MUI	19	1 h	0.96	1–36	Correlation coefficient
Victor et al.	1987	F (50, 21–73)	Not Informed	15	24 h 48 h	0.66 0.90	6–28	ICC
Mouritsen et al.	1989	F (55, 40–61)	SUI (39), MUI (33)	54	24 h	0.87	**	Pearson Correlation
Lose et al.	1989	F (57, 20–79)	SUI (23), MUI (8)	31	24 h	0.82	0	Correlation
Hahn and Fall	1991	F (52, 30–79)	SUI	50	20 min	0.94	7	Pearson Correlation
Versi et al.	1996	F (absent)	SUI (62), UUI/MUI (44)	112	24 h 48 h	0.90 (0.87–0.94) 0.94 (0.93–0.95)	7	Pearson Correlation
Siltberg et al.	1996	F (49, 37–64)	SUI	21	20 min	0.75 -0.3 (±19.9)*	21	Correlation/ Bland-Altman method
Groutz et al.	2000	F/M (64, 22–84)	UUI (34%), SUI (22%), MUI (26%), OA (16%)	92 F/ 17 M	24 h 48 h	0.72 0.87	7	CCC
Simons et al.	2001	F (56, 44–66)	Not Informed	56	1 h	9.7 (± 29.7) *	3–10	Bland-Altman method
Karantanis et al.	2005	F (57, 48–68)	SUI(66), MUI (30), UUI (6), others (6)	104	24 h	0.88	0	Pearson Correlation
Machold et al.	2009	M (66, 53–80)	PPUI	20	20 min	0.76	7–14	Spearman correlation coefficient
Malik et al.	2016	M (64.5 ± 8.1)	PPUI	25	24 h	0.85 (0.73–0.93)	2	ICC

489MD, mean difference; SD, standard deviation; UI, urinary incontinence; SUI, stress urinary incontinence; UUI, urge urinary incontinence; MUI,
490mixed urinary incontinence; ICC, intraclass correlation coefficient; CCC, Lin's concordance correlation coefficient. OA, overactive bladder; PPUI,
491Post-prostatectomy urinary incontinence . *Mean difference (± SD), **not reported

492Table 3. Analysis of the methodological quality of the PadTest reproducibility studies included in the review, based on the Guideline for
493Reporting Reliability and Agreement Studies (GRRAS).

	Assessed Item (GRAAS)	Klarskov e Hald (1984)	Victor et al. (1987)	Mouritsen et al. (1989)	Lose et al. (1989)	Hahn e Fall (1991)	Versi et al. (1996)	Siltberg et al. (1996)	Groutz et al. (2000)	Simons et al. (2001)	Karantanis et al. (2005)	Mach old et al. (2009)	Malik et al. (2016)
Title and abstract	Identify in title or abstract that interrater/intrarater reliability or agreement was investigated	Y	Y	N	N	Y	Y	N	Y	Y	Y	Y	N
Introduction	Name and describe the diagnostic or measurement device of interest explicitly.	Y	N	Y	Y	N	Y	N	N	Y	Y	Y	Y
	Specify the subject population of interest	Y	Y	N	N	N	Y	Y	N	N	Y	Y	Y
	Specify the rater population of interest	N	N	N	N	N	N	N	Y	Y	Y	Y	N
	Describe what is already known about reliability and agreement and provide a rationale for the study	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	N	N
Methods	Explain how the sample size was chosen. State the determined number of raters, subjects/objects, and replicate observations	N	Y	N	N	N	U	U	U	Y	U	N	Y
	Describe the sampling method	N	N	N	N	N	N	N	Y	Y	Y	N	Y
	Describe the measurement/rating process	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	State whether measurements/ratings were conducted independently	N	N	U	U	U	U	U	U	U	U	N	U
	Describe the statistical analysis	N	N	Y	N	Y	U	Y	Y	N	Y	Y	Y
Results	State the actual number of raters and subjects/objects which were included and the number of replicate observations which were conducted.	Y	Y	N	N	N	Y	Y	U	N	Y	U	U
	Describe the sample characteristics of raters and subjects (e.g. training, experience)	N	N	N	N	N	N	N	N	N	N	N	N
	Report estimates of reliability and agreement including measures of statistical uncertainty.	N	N	N	N	N	Y	Y	N	Y	N	N	Y

Discussion	Discuss the practical relevance of results	N	N	N	N	N	Y	N	Y	Y	Y	Y	Y
Auxiliary	Provide detailed results if possible	N	N	N	N	N	N	N	N	N	N	N	N
Score (%)		**6/15 (40)**	**5/15 (33)**	**4/15 (27)**	**3/15 (33)**	**4/15 (27)**	**8/15 (53)**	**8/15 (53)**	**7/15 (47)**	**9/15 (60)**	**9/15 (60)**	**7/15 (47)**	**8/15 (53)**

494N:No, Y:Yes, U: Unclear

Figure 1.Flow diagram pad test accuracy

Figure 2.Risk of bias and applicability concerns summary and graph: Review authors judgments about each domain presented as percentages across included studies

Figure 3. Coupled forest plot from studies of PadTest accuracy compared to urodynamic test.