Health Technology Assessment 2006; Vol. 10: No. 6

Systematic review and evaluation of methods of assessing urinary incontinence

JL Martin, KS Williams, KR Abrams, DA Turner, AJ Sutton, C Chapple, RP Assassa, C Shaw and F Cheater

February 2006

Health Technology Assessment
NHS R&D HTA Programme

How to obtain copies of this and other HTA Programme reports.
An electronic version of this publication, in Adobe Acrobat format, is available for downloading free of charge for personal use from the HTA website (http://www.hta.ac.uk). A fully searchable CD-ROM is also available (see below).

Printed copies of HTA monographs cost £20 each (post and packing free in the UK) to both public **and** private sector purchasers from our Despatch Agents.

Non-UK purchasers will have to pay a small fee for post and packing. For European countries the cost is £2 per monograph and for the rest of the world £3 per monograph.

You can order HTA monographs from our Despatch Agents:

– fax (with **credit card** or **official purchase order**)
– post (with **credit card** or **official purchase order** or **cheque**)
– phone during office hours (**credit card** only).

Additionally the HTA website allows you **either** to pay securely by credit card **or** to print out your order and then post or fax it.

Contact details are as follows:

HTA Despatch
c/o Direct Mail Works Ltd
4 Oakwood Business Centre
Downley, HAVANT PO9 2NP, UK

Email: orders@hta.ac.uk
Tel: 02392 492 000
Fax: 02392 478 555
Fax from outside the UK: +44 2392 478 555

NHS libraries can subscribe free of charge. Public libraries can subscribe at a very reduced cost of £100 for each volume (normally comprising 30–40 titles). The commercial subscription rate is £300 per volume. Please see our website for details. Subscriptions can only be purchased for the current or forthcoming volume.

Payment methods

Paying by cheque
If you pay by cheque, the cheque must be in **pounds sterling**, made payable to *Direct Mail Works Ltd* and drawn on a bank with a UK address.

Paying by credit card
The following cards are accepted by phone, fax, post or via the website ordering pages: Delta, Eurocard, Mastercard, Solo, Switch and Visa. We advise against sending credit card details in a plain email.

Paying by official purchase order
You can post or fax these, but they must be from public bodies (i.e. NHS or universities) within the UK. We cannot at present accept purchase orders from commercial companies or from outside the UK.

How do I get a copy of *HTA on CD*?

Please use the form on the HTA website (www.hta.ac.uk/htacd.htm). Or contact Direct Mail Works (see contact details above) by email, post, fax or phone. *HTA on CD* is currently free of charge worldwide.

The website also provides information about the HTA Programme and lists the membership of the various committees.

Systematic review and evaluation of methods of assessing urinary incontinence

JL Martin,[1*] KS Williams,[1] KR Abrams,[1] DA Turner,[1] AJ Sutton,[1] C Chapple,[2] RP Assassa,[3] C Shaw[4] and F Cheater[5]

[1] Department of Health Sciences, University of Leicester, UK
[2] Urology Research, Royal Hallamshire Hospital, Sheffield, UK
[3] Pinderfields and Pontefract General Infirmary, UK
[4] Department of General Practice, University of Wales, College of Medicine, UK
[5] School of Healthcare Studies, University of Leeds, UK

* Corresponding author

Declared competing interests of authors: none

Published February 2006

This report should be referenced as follows:

Martin JL, Williams KS, Abrams KR, Turner DA, Sutton AJ, Chapple C, *et al*. Systematic review and evaluation of methods of assessing urinary incontinence. *Health Technol Assess* 2006;**10**(6).

Health Technology Assessment is indexed and abstracted in *Index Medicus*/MEDLINE, *Excerpta Medica*/EMBASE and *Science Citation Index Expanded (SciSearch®)* and *Current Contents®*/Clinical Medicine.

NHS R&D HTA Programme

The research findings from the NHS R&D Health Technology Assessment (HTA) Programme directly influence key decision-making bodies such as the National Institute for Health and Clinical Excellence (NICE) and the National Screening Committee (NSC) who rely on HTA outputs to help raise standards of care. HTA findings also help to improve the quality of the service in the NHS indirectly in that they form a key component of the 'National Knowledge Service' that is being developed to improve the evidence of clinical practice throughout the NHS.

The HTA Programme was set up in 1993. Its role is to ensure that high-quality research information on the costs, effectiveness and broader impact of health technologies is produced in the most efficient way for those who use, manage and provide care in the NHS. 'Health technologies' are broadly defined to include all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care, rather than settings of care.

The HTA Programme commissions research only on topics where it has identified key gaps in the evidence needed by the NHS. Suggestions for topics are actively sought from people working in the NHS, the public, service-users groups and professional bodies such as Royal Colleges and NHS Trusts.

Research suggestions are carefully considered by panels of independent experts (including service users) whose advice results in a ranked list of recommended research priorities. The HTA Programme then commissions the research team best suited to undertake the work, in the manner most appropriate to find the relevant answers. Some projects may take only months, others need several years to answer the research questions adequately. They may involve synthesising existing evidence or conducting a trial to produce new evidence where none currently exists.

Additionally, through its Technology Assessment Report (TAR) call-off contract, the HTA Programme is able to commission bespoke reports, principally for NICE, but also for other policy customers, such as a National Clinical Director. TARs bring together evidence on key aspects of the use of specific technologies and usually have to be completed within a short time period.

Criteria for inclusion in the HTA monograph series

Reports are published in the HTA monograph series if (1) they have resulted from work commissioned for the HTA Programme, and (2) they are of a sufficiently high scientific quality as assessed by the referees and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

The research reported in this monograph was commissioned by the HTA Programme as project number 99/29/02. The contractual start date was in April 2002. The draft report began editorial review in December 2003 and was accepted for publication in April 2005. As the funder, by devising a commissioning brief, the HTA Programme specified the research question and study design. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' report and would like to thank the referees for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

The views expressed in this publication are those of the authors and not necessarily those of the HTA Programme or the Department of Health.

ISSN 1366-5278

Published by Gray Publishing, Tunbridge Wells, Kent, on behalf of NCCHTA.
Printed on acid-free paper in the UK by St Edmundsbury Press Ltd, Bury St Edmunds, Suffolk.

Abstract

Systematic review and evaluation of methods of assessing urinary incontinence

JL Martin,[1*] KS Williams,[1] KR Abrams,[1] DA Turner,[1] AJ Sutton,[1] C Chapple,[2] RP Assassa,[3] C Shaw[4] and F Cheater[5]

[1] Department of Health Sciences, University of Leicester, UK
[2] Urology Research, Royal Hallamshire Hospital, Sheffield, UK
[3] Pinderfields and Pontefract General Infirmary, UK
[4] Department of General Practice, University of Wales, College of Medicine, UK
[5] School of Healthcare Studies, University of Leeds, UK
* Corresponding author

Objectives: To identify and synthesise studies of diagnostic processes of urinary incontinence and to construct an economic model to examine the cost-effectiveness of simple, commonly used primary care tests.
Data sources: The electronic databases MEDLINE (1966–2002), CINAHL (1982–2002) and EMBASE (1980–2002).
Review methods: Studies were selected and assessed using the Quality Assessment of Diagnostic Studies (QUADAS) tool. Studies that reported the results of applying the same diagnostic procedure using the same threshold value (cut-off) were pooled using a random effects meta-analysis model to produce pooled estimates of sensitivity, specificity and diagnostic odds ratio together with 95% confidence intervals.
Results: In total, 6009 papers were identified from the literature search, of which 129 were deemed relevant for inclusion in the review, and these papers compared two or more diagnostic techniques. The gold-standard diagnostic test for urinary incontinence with which each reference test was compared was multichannel urodynamics. In general, reporting in the primary studies was poor; there was a lack of literature in the key clinical areas and minimal literature dealing with diagnosis in men. Only a limited number of studies could be combined or synthesised, providing the following results when compared with multichannel urodynamics. A clinical history for diagnosing urodynamic stress incontinence (USI) in women was found to have a sensitivity of 0.92 and specificity of 0.56 and for detrusor overactivity (DO) a sensitivity of 0.61 and specificity of 0.87. For validated scales, question 3 of the Urogenital Distress Inventory was found to have a sensitivity of 0.88 and specificity of 0.60. Seven studies compared a pad test with multichannel urodynamics; however, four different pad tests were studied and therefore it was difficult to draw any conclusions about diagnostic accuracy. Of the four studies comparing urinary diary with multichannel urodynamics, only one presented data in a format that allowed sensitivity and specificity to be calculated. Their reported values of 0.88 and 0.83 suggest that a urinary diary may be effective in the diagnosis of DO in women. Examination of the incremental cost-effectiveness of three primary care tests used in addition to history found that the diary had the lowest cost-effectiveness ratio of between £35 and £77 per extra unit of effectiveness (or case diagnosed). Imaging by ultrasound to determine leakage was found to be effective in the diagnosis of USI in women, with a sensitivity of 0.94 and specificity of 0.83.
Conclusions: This is the first systematic review of methods for diagnosing urinary incontinence. As reporting of the primary studies was poor, clinical interpretation was often difficult because few studies could be synthesised and conclusions made. The report found that a large proportion of women with USI can be correctly diagnosed in primary care from clinical history alone. On the basis of diagnosis the diary appears to be the most cost-effective of the three primary care tests (diary, pad test and validated scales) used in addition to clinical history. Ultrasound imaging may offer a valuable alternative to urodynamic investigation. The clinical stress test is effective in the diagnosis of USI. Adaptation of such a test so that it

could be performed in primary care with a naturally filled bladder may prove clinically useful. If a patient is to undergo an invasive urodynamic procedure, multichannel urodynamics is likely to give the most accurate result in a secondary care setting. There is a dearth of literature on the diagnosis of urinary incontinence in men, with no studies meeting the study criteria for data extraction in the diagnosis of bladder outlet obstruction. There is a need for large-scale, high-quality primary studies evaluating the use of a number of diagnostic methods in a primary care setting to be undertaken so that the results of this systematic review can be verified or not. Such studies should include not only an assessment of clinical effectiveness, in this case diagnostic accuracy, but also an assessment of costs and quality of life/satisfaction to inform future health policy decisions. Studies carried out should be reported to a better standard. The recommendations of the Standards for Reporting Diagnostic Accuracy (STARD) initiative should be followed to ensure the accuracy and completeness of reporting design and results.

Contents

List of abbreviations

AUA	American Urological Association
AUC	area under the curve
BIDI	Bladder Instability Discriminant Index
BMI	body mass index
BND	bladder neck descent
BOO	bladder outlet obstruction
BPH	benign prostatic hyperplasia
CI	confidence interval
CRD	Centre for Reviews and Dissemination
df	degrees of freedom
DIS	Detrusor Instability Score
DO	detrusor overactivity
DOR	diagnostic odds ratio
DUEC	distal urethral conductance
F	female
ICS	International Continence Society
IIQ	Incontinence Impact Questionnaire
ISQ	Incontinence Screening Questionnaire
LUTS	lower urinary tract symptoms
M	male
MCU	multichannel urodynamics
MSSU	midstream specimen of urine
MUI	mixed urinary incontinence
PVRV	postvoid residual volume (of urine)
QALY	quality-adjusted life-year
QUADAS	Quality Assessment of Diagnostic Studies
RAP	Resident Assessment Protocol
ROC	receiver operating characteristic
SCU	single-channel urodynamics
SE	standard error
sEMG	surface electromyography
sROC	summary receiver operating characteristic
STARD	Standards for Reporting of Diagnostic Accuracy
SUI	stress urinary incontinence
UDI	Urogenital Distress Inventory
UI	urinary incontinence
UPP	urethral pressure profile
USI	urodynamic stress incontinence
UUI	urge urinary incontinence

All abbreviations that have been used in this report are listed here unless the abbreviation is well known (e.g. NHS), or it has been used only once, or it is a non-standard abbreviation used only in figures/tables/appendices in which case the abbreviation is defined in the figure legend or at the end of the table.

Executive summary

Background

Although urinary incontinence is not life threatening, it can have enormous costs to individuals and the health service in terms of expenditure and impact on quality of life. Epidemiological studies have demonstrated that urinary incontinence is a very common symptom, with a reported prevalence of any urinary incontinence (in those aged 40 and over) of 34% for women and 14% for men.

Pathways to diagnostic assessment are inconsistent, with some individuals being assessed and treated in primary care settings by GPs and nurses, and others being referred directly to a variety of specialists in secondary care (e.g. physiotherapists, gynaecologists and urologists) without any assessment or treatment. Assessment can be undertaken at a number of levels using different combinations of tests.

It is particularly important when implementing certain treatment interventions (e.g. medication that may have side-effects) that a diagnosis is made to determine the most effective treatment intervention, and it is imperative before surgical intervention. If a diagnosis is not made, then inappropriate and unnecessary interventions may be implemented. Two types of diagnosis can be made: symptomatic diagnosis and condition-specific diagnosis. In general, symptomatic diagnoses are made in primary care using clinical history-taking, urinary diaries, pad tests and validated symptom scales. Condition-specific diagnoses are made in secondary care using urodynamic techniques. The use of diagnostic assessment methods is influenced by the clinical setting and the expertise of the individual undertaking the assessment. The evidence available on the accuracy and acceptability of these diagnostic processes is inconsistent and variable.

Objectives

This systematic review aimed to:

- identify, appraise and summarise the published evidence relating to different methods of diagnostic assessment of male and female urinary incontinence: specifically urodynamic stress incontinence (USI) and detrusor overactivity (DO)
- quantitatively synthesise the extracted evidence using meta-analysis methods (where possible) or pooling of individual sensitivity and specificity data
- construct an economic model to examine the cost-effectiveness of simple, commonly used primary care tests
- identify gaps in the literature
- prioritise future clinical and research questions.

Methods

Data sources

The online bibliographic databases MEDLINE (1966–2002), CINAHL (1982–2002) and EMBASE (1980–2002) were used to obtain the literature. The search strategy was based on the Cochrane and NHS Centre for Reviews and Dissemination strategies for identifying studies of diagnostic performance.

Study selection

Study selection comprised a three-stage process using defined inclusion and exclusion criteria. All records were assessed for relevance by the first investigator on the basis of the abstract, or if the abstract was not available then title only. Papers were considered relevant to the systematic review if they considered the evaluation, appropriateness and/or cost of diagnostic assessment in the following categories:

- clinical history-taking
- simple investigations including validated scales, diaries and pad tests
- advanced (invasive) investigations (e.g. urodynamics).

To be included, a paper had to provide a quantitative comparison between two or more different methods of diagnosing urinary incontinence.

Data extraction

A panel consisting of at least three members of the review team, including at least one statistician,

discussed all papers identified as of potential relevance. The panel determined whether study data were presented in a suitable format to calculate sensitivity and specificity.

Quality assessment

All relevant papers were assessed for quality using Quality Assessment of Diagnostic Studies (QUADAS), a tool designed specifically for studies on diagnostic accuracy. An initial pilot study on four papers resulted in a number of clarifications being added to the instructions of the QUADAS tool to ensure consistency between assessors. Seven of the authors performed the full quality assessment process, with 10% of the papers being assessed by two authors to test for inter-reader agreement.

Data synthesis

Studies that reported the results of applying the same diagnostic procedure using the same threshold value (cut-off) were pooled using a random effects meta-analysis model to produce pooled estimates of sensitivity, specificity and diagnostic odds ratio together with 95% confidence intervals.

Results

In total, 6009 papers were identified from the literature search, of which 129 were deemed relevant for inclusion in the review, and these papers compared two or more diagnostic techniques. The gold-standard diagnostic test for urinary incontinence with which each reference test was compared was multichannel urodynamics.

In general, reporting in the primary studies was poor; there was a lack of literature in the key clinical areas and minimal literature dealing with diagnosis in men. Only a limited number of studies could be combined or synthesised, providing the following results when compared with multichannel urodynamics. A clinical history for diagnosing USI in women was found to have a sensitivity of 0.92 and specificity of 0.56 and for DO a sensitivity of 0.61 and specificity of 0.87. For validated scales, question 3 of the Urogenital Distress Inventory was found to have a sensitivity of 0.88 and specificity of 0.60. Seven studies compared a pad test with multichannel urodynamics; however, four different pad tests were studied and therefore it was difficult to draw any conclusions about diagnostic accuracy. Of the four studies comparing urinary diary with multichannel urodynamics, only one presented data in a format that allowed sensitivity and specificity to be calculated. Their reported values of 0.88 and 0.83 suggest that a urinary diary may be effective in the diagnosis of DO in women. Examination of the incremental cost-effectiveness of three primary care tests used in addition to history found that the diary had the lowest cost-effectiveness ratio of between £35 and £77 per extra unit of effectiveness (or case diagnosed). Imaging by ultrasound to determine leakage was found to be effective in the diagnosis of USI in women, with a sensitivity of 0.94 and specificity of 0.83.

Conclusions

This is the first systematic review of methods for diagnosing urinary incontinence. As reporting of the primary studies was poor, clinical interpretation was often difficult because few studies could be synthesised and conclusions made. The following information could be deduced from the available data.

- A large proportion of women with USI can be correctly diagnosed in primary care from clinical history alone.
- On the basis of diagnosis the diary appears to be the most cost-effective of the three primary care tests (diary, pad test and validated scales) used in addition to clinical history.
- Ultrasound imaging may offer a valuable alternative to urodynamic investigation.
- The clinical stress test is effective in the diagnosis of USI. Adaptation of such a test so that it could be performed in primary care with a naturally filled bladder may prove clinically useful.
- If a patient is to undergo an invasive urodynamic procedure, multichannel urodynamics is likely to give the most accurate result in a secondary care setting.
- There is a dearth of literature on the diagnosis of urinary incontinence in men, with no studies meeting the study criteria for data extraction in the diagnosis of bladder outlet obstruction.

Implications for healthcare

- There is currently a lack of high-quality research in clinically relevant areas to inform clinical practice.
- Most diagnostic methods can be undertaken in primary or secondary care.
- Simple investigations (e.g. pad test and diary) may offer useful information on severity which, when combined with history, may provide

sufficient information to commence primary care interventions (which are low cost and low risk).

Recommendations for research

Given the demographics of the UK population and the reported high prevalence of any urinary incontinence in the community-dwelling population, there will be an increasing burden placed on primary (and secondary) care services in terms of the diagnostic assessment and appropriate treatment of incontinence. Therefore, identifying which are the most clinically accurate and cost-effective diagnostic methods is of crucial importance.

There is a need for large-scale, high-quality primary studies evaluating the use of a number of diagnostic methods in a primary care setting to be undertaken so that the results of this systematic review can be verified or not. Such studies should include not only an assessment of clinical effectiveness, in this case diagnostic accuracy, but also an assessment of costs and quality of life/satisfaction to inform future health policy decisions.

Studies carried out should be reported to a better standard. The recommendations of the Standards for Reporting of Diagnostic Accuracy (STARD) initiative should be followed to ensure the accuracy and completeness of reporting design and results.

Chapter 1

Introduction and background

Background

Urinary incontinence has been defined by the International Continence Society (ICS) as "the complaint of any involuntary leakage of urine".[1] They suggest that such leakage should be further described by specifying type (distinguishing between stress, urge and mixed urinary incontinence), frequency, severity, precipitating factors, social impact, effect on hygiene and quality of life, measures used to contain leakage and whether the individual seeks or desires help for incontinence. Although urinary incontinence is not life threatening, it can have enormous costs to individuals and the health service in terms of expenditure and impact on quality of life. Epidemiological studies have demonstrated that urinary incontinence is a very common symptom; McGrother and colleagues report a prevalence of any urinary incontinence (in those aged 40 years and over) of 34% for women and 14% for men. The proportion of people finding that these symptoms impact on their lives is estimated to be around 29% for women and 14% for men.[2]

Aetiology

Three types of incontinence can be identified, depending on the symptoms of the presenting patient. These terms are commonly used in scientific studies and the definitions are taken from the current ICS Standardisation Report[1] to describe symptomatic diagnoses.

- Stress urinary incontinence (SUI) is the complaint of involuntary leakage on effort or exertion, or on sneezing or coughing.
- Urge urinary incontinence (UUI) is the involuntary leakage of urine accompanied or immediately preceded by urgency.
- Mixed urinary incontinence (MUI) is the complaint of involuntary leakage associated with urgency and also with exertion, effort, sneezing or coughing.

When the symptoms of incontinence are confirmed by urodynamic investigation then two types of incontinence can be diagnosed:

- Urodynamic stress incontinence (USI) is the involuntary leakage of urine during increased abdominal pressure in the absence of a detrusor contraction. This replaces the commonly used term genuine stress incontinence.[1]
- Detrusor overactivity (DO) is involuntary detrusor contractions during the filling phase, which may be spontaneous or provoked. This term replaces detrusor instability.

Cost and social problems

Urinary incontinence has an enormous cost to individuals and health services in terms of expenditure and impact on quality of life. A study investigating the cost of urinary storage disorders to the UK estimated that the total cost of treating urinary storage disorders in community-dwelling adults over the age of 40 was £536 million in 1999/2000 prices. In addition, there is an estimated cost of £207 million that is borne by the individual for managing their symptoms (£29 million and £178 million for men and women, respectively).[3]

In addition to the economic costs, urinary incontinence has a serious impact on the quality of life of sufferers. Effects have been shown to include depression,[4] anxiety[5] and poor life satisfaction.[6] All types of leakage have a detrimental effect on daily activities and overactive bladder symptoms in particular have been shown to be distressing for young women.[7]

Assessment and diagnosis

Diagnosis of urinary incontinence usually begins with an assessment of the symptoms in a clinical history. There are several different symptoms of urinary incontinence, depending on the circumstances under which people leak urine. Diagnosis may involve methods of assessing the severity and pattern of leakage, using methods such as pad tests and urinary diaries. Pad tests largely measure the severity of leakage, while the diary assesses the severity of frequency and leakage. Increased frequency and incontinence

recorded in a diary may be indicative of UUI and a positive pad test may indicate SUI.

Assessment procedures tend to be sequential, beginning with the recording of symptoms in a patient history, which may be indicative of a particular underlying condition. Linked in with these sequential assessment procedures are often clinical treatment interventions; these may be implemented and then further assessment processes undertaken depending on the success of the intervention.

Methods of diagnostic assessment can be broadly divided and sequentially ordered into five groups:

- clinical history-taking, including nature, duration and reported severity of symptoms, functional and mental status, relevant medical, surgical and gynaecological history, impact of symptoms on quality of life and exacerbating factors including diet, fluid and medications
- validated scales, which measure the severity of symptoms and impact of symptoms on quality of life
- physical examination, including abdominal, perineal, rectal, neurological and measurement of body mass index (BMI)
- simple investigations, including urinalysis, midstream specimen of urine (MSSU), measurement of postvoid residual volume (PVRV), provocation stress test, frequency–volume charts and pad tests
- advanced investigations, including urodynamics.

Pathways to diagnostic assessment are inconsistent, with some individuals being assessed and treated in primary-care settings by GPs and nurses, and others being referred directly to a variety of specialists in secondary care (e.g. physiotherapists, gynaecologists, urologists, geriatricians or specialist nurses based in secondary care) without any assessment or treatment. Although algorithms for the assessment and treatment of urinary incontinence have been recommended, the most appropriate healthcare worker to conduct such assessments has not been identified, nor has their ideal location.[8] For example, a symptomatic diagnosis conducted by a nurse in a health centre will have a different service cost to a condition-specific diagnosis conducted by a specialist in hospital using urodynamic equipment.

Assessment can be undertaken at a number of levels using different combinations of screening tests. *Figure 1* illustrates assessment processes in clinical practice and how they are interrelated with initiation of treatment. There are also likely overlaps of investigative methods being used at different points in the care pathway.

It is particularly important when implementing certain treatment interventions (e.g. medication that may have side-effects) that a diagnosis is made to determine the most effective treatment intervention, and of course it is imperative before surgical intervention. If a diagnosis is not made, then inappropriate and unnecessary interventions may be implemented. As has already been mentioned, there are two levels of diagnosis: symptomatic diagnosis and condition-specific diagnosis. In general, symptomatic diagnoses take place in primary care and condition-specific in secondary care, where urodynamic investigations are available. In primary care the diagnosis of urinary incontinence is dependent on history-taking, physical examination and simple investigations including frequency–volume charts, pad tests, urinalysis and estimation of PVRV. The choice of diagnostic assessment method is influenced by the clinical setting (primary/secondary care) and by the expertise of the professional conducting the diagnostic test. To date, research has focused on the clinical effectiveness of condition-specific diagnosis. Little attention has been paid to the effectiveness of symptomatic diagnosis, despite this being the basis of all treatment in primary care.

The term urodynamics relates to the study of pressure–flow relationships in the urinary tract and provides a functional assessment of the lower urinary tract to provide objective explanations for urinary symptoms or dysfunction.[9] Urodynamic tests include such minimally invasive tests as frequency–volume charts, but more commonly refer to cystometry, urethral pressure measurement, pressure–flow studies, videourodynamics and ambulatory monitoring.[9] The aim of clinical urodynamics is to reproduce symptoms while making precise measurements to identify the underlying cause for the symptoms and to quantify the pathophysiological processes.[10] Urodynamic tests are invasive, usually involving catheterisation of the bladder and the measurement of pressure in the urethra, bladder and abdomen. A significant number of people who undergo urodynamics find it embarrassing, painful or distressing.[11]

Full descriptions of urodynamic techniques can be found in a number of recent publications.[9,12]

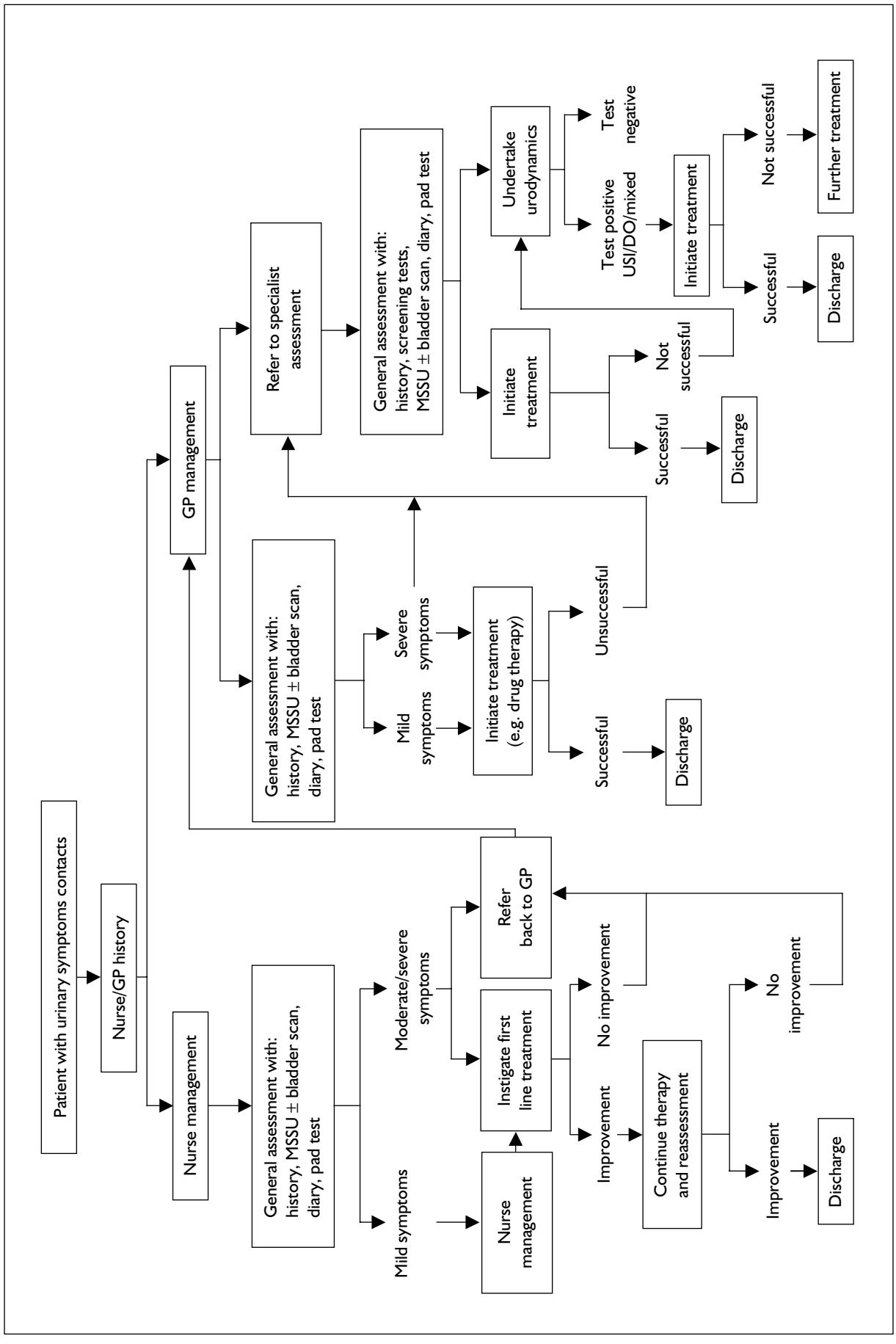


FIGURE 1 *Flowchart of assessment processes in clinical practice*

Although there are concerns about accuracy and reproducibility, urodynamics is still regarded as the gold-standard method for diagnosing urinary incontinence and is usually a necessary procedure before surgery is performed.[13]

Aims and objectives

This systematic review aims to:

- identify, appraise and summarise the published evidence relating to different methods of diagnostic assessment of male and female urinary incontinence: specifically USI and DO
- quantitatively synthesise the extracted evidence using meta-analysis methods (where possible) or pooling of individual sensitivity and specificity data
- develop an illustrative flowchart of diagnostic processes for urinary incontinence in current clinical practice, and construct an economic model to examine the cost-effectiveness of simple, commonly used primary-care tests
- identify gaps in the literature; and prioritise future clinical and research questions.

Chapter 2
Methods

General methodology

The systematic review followed the guidelines contained in NHS Centre for Reviews and Dissemination (CRD) Report 4[14] and aimed to appraise and summarise the published evidence relating to the different methods of diagnostic assessment in male and female urinary incontinence within the subgroups of diagnostic tests described in Chapter 1:

- clinical history-taking
- validated scales
- physical examination
- simple investigations
- advanced investigations.

The review examined the evidence of these subgroups of tests in relation to:

- clinical use, including sensitivity, specificity and positive predictive values of different diagnostic assessment methods when compared with the gold standard of multichannel urodynamics
- economic modelling.

The overall philosophy of the systematic review was to maintain breadth, synthesising the evidence where appropriate using quantitative techniques and providing economic modelling of costs of diagnostic methods.

Search strategy

The online bibliographic databases MEDLINE (January 1966 to December 2002), CINAHL (January 1982 to December 2002) and EMBASE (January 1980 to December 2002) were used to obtain the literature. The search strategy was based on the Cochrane and NHS CRD strategies for identifying studies of diagnostic performance, and the information officers at these centres were consulted during this process. A number of keywords was identified based on possible diagnostic tests and possible permutations of their names (*Table 1*). A paper was included if a word from {Diagnostic filter} OR {Diagnostic test} AND {Incontinence term} was found anywhere in the title or abstract or used as a MeSH heading. The search results were limited to humans, reports in the English language and adults (>19 years) only. The full search strategies can be seen in Appendix 1.

TABLE 1 *Keywords used in literature search of MEDLINE, EMBASE and CINAHL*

{Urinary incontinence}	**{Diagnostic test}**
Urinary incontinence	Urodynamics
Urge incontinence	Provocation stress test
Stress incontinence	Frequency volume chart
Leakage AND urin* (keyword)	Urinalysis
	Post-void residual volume
	Mid-stream specimen
	MSSU
	Pad tests OR pad testing OR pad test
	Urinalysis
	Midstream sample of urine

{Diagnostic filter}		
Sensitivity	Summary receiver operating characteristic/curve	Predictive value
Specificity	Diagnostic errors	Predictive standards
Predictive value of tests	Likelihood ratio	Predictive models
Reference values	Likelihood function	Criteria test
Reference standard	False positives	Validated standard
'Gold standard'	False negatives	Work-up bias
		Observer bias/variation

First exclusion process

All records were entered into a bibliographic referencing software program (Procite). Duplicate papers were identified and deleted. The remaining papers were assessed for relevance by the first investigator on the basis of the abstract, or if the abstract was not available then the title only. A sample (10%) was also assessed for potential relevance by the second investigator; agreement between the two readers was 99%.

Inclusion

Papers were considered relevant to the systematic review if they considered the evaluation, appropriateness and/or cost of diagnostic assessment in the five categories identified:

- clinical history-taking
- validated scales
- physical examination
- simple investigations
- advanced investigations.

To be included, a paper had to provide a quantitative comparison between two different methods of diagnosing urinary incontinence.

Exclusion

Any papers that fell into the following categories were excluded from the review:

- diagnosis of children
- reports in a non-English language
- case reports
- letters
- reviews (non-primary research)
- papers investigating interventional procedures where diagnostic tests were used as outcome measures.

All of the abstracts were read by the first investigator and classified as relevant, not relevant or unclear. A second investigator who was blinded to the initial classifications then read 20% of the relevant records, 10% of the not relevant records and 100% of the unclear records. Any discrepancies were discussed. Agreement between the two investigators was 98%. Full copies of those papers identified as either relevant or unclear were obtained.

Second exclusion process

Once obtained, full copies of papers identified as of potential relevance were read by the first investigator and classified as relevant, not relevant or unclear on the basis of the same inclusion and exclusion criteria. The same second investigator, again blinded, read 20% of the relevant, 20% of the not relevant and all of the unclear papers, and any discrepancies were discussed. Agreement between the two investigators was 96%.

Categorisation of studies

Owing to the large number of tests used for the diagnosis of urinary incontinence and, hence, the number of possible comparisons, a matrix was constructed to organise the literature (see *Table 2* in Chapter 3). Each relevant paper was assigned to a box in the matrix according to the two diagnostic tests compared (or boxes if more than two tests were compared).

Quality assessment

The recent growth in systematic reviews of diagnostic tests has resulted in the need for methods to assess the quality of diagnostic studies. In response to this, a project was funded by the HTA programme to develop a quality tool specifically for these types of studies, the Quality Assessment of Diagnostic Studies (QUADAS) tool,[15] which was used for the quality assessment component of the review. The tool consists of 14 questions regarding the quality of the study and quality of reporting (Appendix 2).

Pilot study

As the QUADAS tool was a recently developed instrument, a pilot quality assessment exercise was undertaken to ascertain whether it required amending or extending for the specific remit of the review. Four papers[16–19] identified as potentially relevant for inclusion in the review were assessed for quality by five of the project investigators using the original QUADAS tool. The investigators were asked to report any questions that they felt required clarification or expanding, or that were not relevant.

Several clarifications were added to the instructions based on the recommendations from the pilot study (Appendix 3). These included directives that no assumptions should be made, for example when judging the period between the two tests. This is rarely explicitly stated and it is tempting (and probably correct) to assume that the period between tests is short. Following advice from a clinical member of the project team, further information was provided for assessing the quality of papers that investigated urodynamic procedures, including the minimum amount of detail required

for replication of urodynamics to be possible. Information was also added to clarify the quality assessment of other questions on validity of the sample and appropriate reference standards.

Full quality assessment process

Seven members of the investigation team took part in this process, each assessing approximately 30 papers. Ten per cent of the papers were assessed by two different investigators to check the inter-rater reliability of the tool; the remaining 90% were assessed only by one investigator. This procedure also served as a final filter for relevance and investigators were asked to highlight any studies that they felt were not relevant to the review. These studies were discussed by two investigators and if not relevant were excluded from the review.

Data extraction

All papers identified as of potential relevance were discussed by a panel consisting of at least three members of the review team, including at least one statistician. The panel determined whether study data were presented in a suitable way to allow a cross-tabulation of the results or sensitivity and specificity to be calculated. The authors of studies that did not present sufficient data for inclusion in any meta-analysis were contacted by letter and asked to provide further details (Appendix 4). In order to aid this procedure and maximise the response, forms were sent with template data tables to aid the authors in providing data in either a cross-tabulation form or individual patient data (Appendix 5). A website was also set up to give authors further information about the project and examples of the data required (http://www.prw.le.ac.uk/research/hta/) (Appendix 6).

While members of the project team were assessing the quality of papers they also recorded other details. This included the size, gender and age of the sample, the care setting where the study was performed and the country (Appendix 7).

Data synthesis

Studies that reported the results of applying the same diagnostic procedure using the same threshold value (cut-off) were pooled using a random effects meta-analysis model (which reduced to a fixed effect model when the between-study variability was estimated to be 0) to produce pooled estimates of the sensitivity, specificity and diagnostic odds ratio (DOR), together with associated 95% confidence intervals (CIs). Tests for heterogeneity were carried out for each outcome and are reported. On the basis of the pooled sensitivity and specificity the positive likelihood ratio was calculated, together with associated 95% CI. A positive likelihood ratio can be used to assess the impact on diagnosis of a positive test result for an individual, although values greater than 10 are usually considered necessary for a test to provide convincing diagnostic evidence.[18] Pooling sensitivity and specificity separately assumes that the diagnostic threshold is the same in each study. Pooling DORs relaxes this assumption by assuming that the studies relate to the same symmetrical receiver operating characteristic (ROC) curve. The DOR has been put forward as a useful single indicator of test performance, which indicates the strength of the association between test results in disease (in much the same way as the odds ratio is used in epidemiology to express the association between exposure and disease). For a thorough explanation of the use of odds ratios in diagnostic applications, including their application to meta-analysis, see Glas and colleagues.[20]

The empirical study sensitivities and specificities and corresponding pooled estimates are plotted in ROC space to aid the simultaneous interpretation. The ROC curve corresponding to the pooled DORs is also presented together with the area under the curve (AUC) for the ROC curve and associated 95% CI.[21] The symmetric ROC curve determined by the pooled DOR is given by

$$\text{Sensitivity} = 1 \Bigg/ \left[1 + \frac{1}{\text{DOR} \times \left(\frac{1 - \text{Specificity}}{\text{Specificity}} \right)} \right]$$

The intention was, where between study heterogeneity existed, to explore it using meta-regression investigating potential associations between study characteristics (such as population under study, country of study and quality of study) on the DOR scale, but this proved to be infeasible owing to the low numbers of studies identified for each separate outcome of interest.

All analyses were performed using Stata version 7.0 (Stata Corporation, College Station, TX, USA, 2002) and MetaDiSc (www.hrc.es/investigation/metadisc.html – a new freely available program for carrying out meta-analysis of diagnostic test performance studies).

Whether studies reported sufficient data for meta-analysis or not, an attempt was made to undertake a narrative synthesis of all relevant papers identified.

Chapter 3
Results

Studies identified

A flowchart of the studies is shown in *Figure 2*. In total, 6099 papers were identified from MEDLINE (2913), CINAHL (411) and EMBASE (2775). Of these, 1479 duplicate papers were identified and deleted: 11 from MEDLINE, 111 from CINAHL and 1357 from EMBASE. There was a large amount of overlap between the studies identified by MEDLINE and EMBASE, with 49% of the studies identified by EMBASE also being identified by MEDLINE.

The deletion of duplicate papers left 4620 individual papers. After the first exclusion process 490 records were identified as of potential relevance and full copies of the papers were obtained. After the second exclusion process 197 different, original papers appeared to meet the inclusion criteria. These potentially reported the quantitative comparison of two or more diagnostic tests used for the detection of urinary incontinence. After more detailed reading of each paper during the data extraction and quality assessment processes a further 76 papers were found not to meet the inclusion criteria of the review and were excluded.[22–96]

Results of contacting authors

Twenty-four studies were identified as being of potential interest but with insufficient data presented in the written paper to enable any summary measures of diagnostic accuracy to be calculated.[19,32,43,52,80,97–116] The lead authors of these studies were contacted by letter and asked to provide further information. Four authors responded with all of the requested data.[97,98,104,107] The data from the other 19 studies were included in the review in the form presented in the paper.

Categorisation of papers

The completed matrix showing the distribution of the literature can be seen in *Table 2*. The majority of the published studies deal with the most commonly used diagnostic tests: urodynamics, pad test, urinary diary, clinical history and ultrasound imaging, although a small number of studies investigated less common tests.

A separate matrix was constructed to organise the literature that compared the different urodynamic tests (*Table 3*).

Quality assessment

Pilot study

Agreement between investigators for the various questions ranged from 0.65 to 1.00 (*Table 4*). A common problem encountered was in the lack of clarity in reporting. This led to investigators making, probably correct, assumptions about factors such as blinding of experimenters and periods between diagnostic tests.

Full quality assessment

The results of the full quality assessment procedure are displayed in *Table 5*. There was a lot of variation in terms of the quality of the studies and also the quality of the reporting. The items that resulted in the most favourable ratings were questions 3, 5, 6 and 7, which were all concerned with the quality of study design: specifically, whether an appropriate reference standard test was used (question 3: 84% of papers rated as 'yes'), whether all patients underwent identical diagnostic procedures (questions 5 and 6: 91% and 86% of papers rated as 'yes') and whether the two diagnostic tests were independent of each other (question 7: 77% of papers rated as 'yes').

Several items were poorly described in the papers: 39% of the papers did not clearly describe the selection criteria used in the study, therefore it is not possible to judge how appropriate the sample was. Questions 9a and 9b dealt with the issue of blinding; for the majority of the studies it was unclear whether the reference (79%) or index tests (83%) were interpreted without knowledge of the other test. Sixty-one per cent of the papers did not report whether there were any uninterpretable or intermediate results (question 11) and 67% of the papers did not report whether there were any withdrawals from the study (question 12). Question 4 dealt with the period between the two

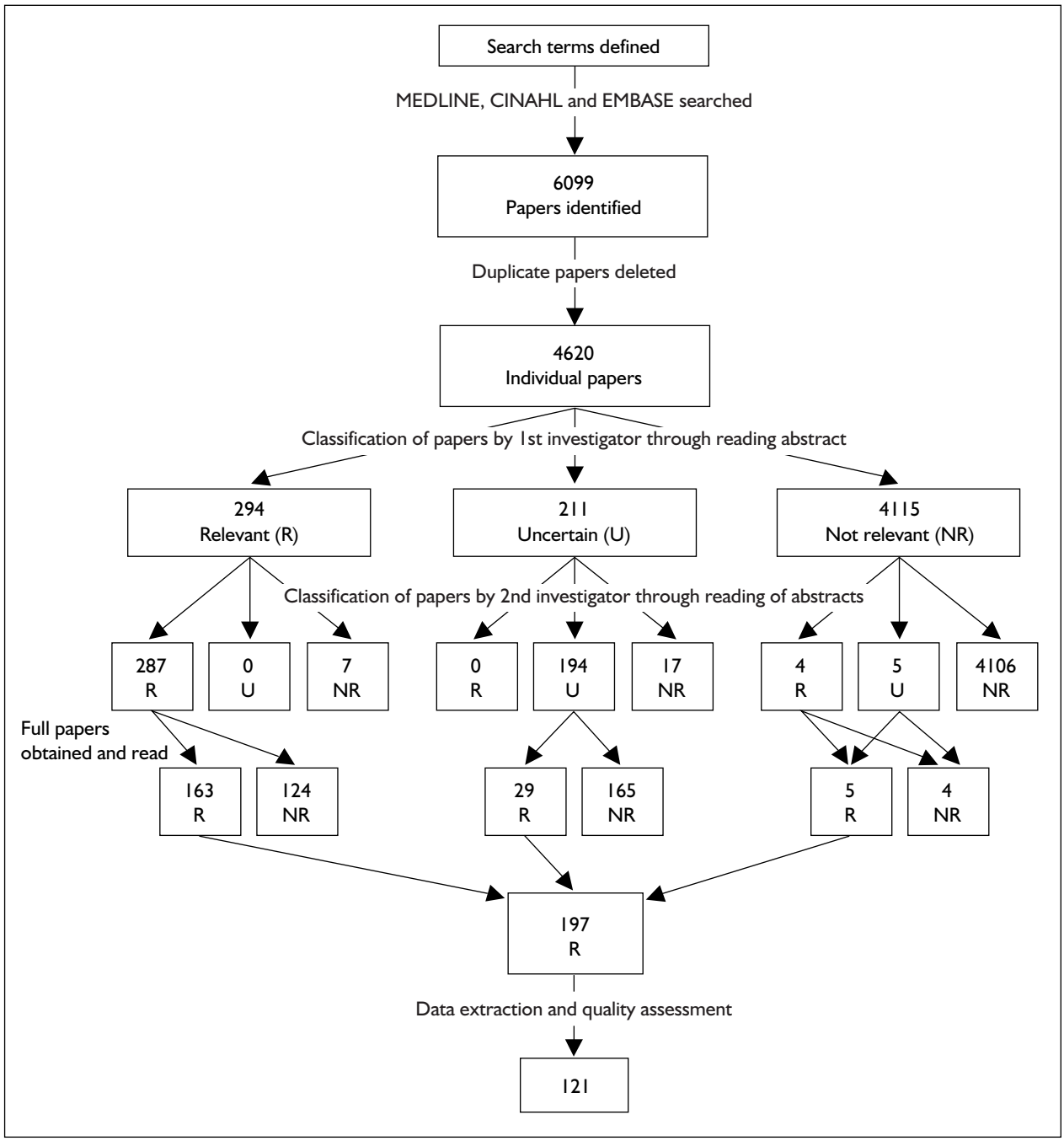


FIGURE 2 *Flowchart of literature*

diagnostic tests; 64% of the papers did not report this and although it is likely that in a lot of cases tests were performed either on the same day or within a few days this could not be assumed.

The responses to the other questions (1, 8a, 8b and 10) showed that for these items quality of both study design and reporting were good: 64% of the studies included a representative spectrum of patients, 64% and 59%, respectively, described the index and reference test in sufficient detail for replication and 79% provided the same clinical data as would be available in practice.

To check for inter-rater agreement 16 of the 121 papers were quality assessed by two separate investigators. The results of this did not allow a kappa test to be performed and therefore the proportion of agreement between the assessors was calculated for each question (*Table 6*). The proportion of agreement between raters ranged from 0.50 (identical ratings were given half of the

TABLE 2 *Matrix showing the distribution of literature that met the inclusion criteria*

	Urodynamics	History	Scales	Pads	Diary	Battery	sEMG
History	42		1	6	3		1
Scales	8		1				
Pads	7		4				
Diary	4				2		
Paper towel test		1					
Physical examination		1				1	2
Q-tip test	4						
Algorithm	3						
Battery	2	1					
Conductance	1						
Ultrasound	9						
Urodynamics	37						

TABLE 3 *Matrix showing comparison of urodynamic tests*

	Multichannel urodynamics	Clinical stress test	Single-channel cystometry
Imaging	5		
Stress tests	6		
Single-channel urodynamics	8		
Ambulatory	6		
UPP	5	1	
Flow measurement	1		
Cystometry by foetal monitor	2		
Ice-water test	1		
Fluid-bridge test			1
Stop test			1

TABLE 4 *Quality assessment pilot study: agreement between investigators*

Item		Proportion of agreement
1.	Was the spectrum of patients representative of the patients who will receive the test in practice?	0.65
2.	Were selection criteria clearly described?	0.90
3.	Is the reference standard likely to correctly classify the target condition?	1.00
4.	Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?	0.70
5.	Did the whole sample or a random selection of the sample receive verification using a reference standard of diagnosis?	1.00
6.	Did patients receive the same reference standard regardless of the index test result?	1.00
7.	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	0.85
8a.	Was the execution of the index test described in sufficient detail to permit replication of the test?	0.69
8b.	Was the execution of the reference standard described in sufficient detail to permit its replication?	0.69
9a.	Were the index test results interpreted without knowledge of the results of the reference standard?	0.65
9b.	Were the reference standard results interpreted without knowledge of the results of the index test?	0.65
10.	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	0.74
11.	Were uninterpretable/intermediate test results reported?	0.50
12.	Were withdrawals from the study explained?	0.75

TABLE 5 *Summary of quality assessment*

Item		Yes	No	Unclear
1.	Was the spectrum of patients representative of the patients who will receive the test in practice?	78	12	31
2.	Were selection criteria clearly described?	65	47	9
3.	Is the reference standard likely to correctly classify the target condition?	102	1	18
4.	Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?	43	0	78
5.	Did the whole sample or a random selection of the sample receive verification using a reference standard of diagnosis?	110	6	5
6.	Did patients receive the same reference standard regardless of the index test result?	104	2	15
7.	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	93	2	26
8a.	Was the execution of the index test described in sufficient detail to permit replication of the test?	78	25	18
8b.	Was the execution of the reference standard described in sufficient detail to permit its replication?	71	33	17
9a.	Were the index test results interpreted without knowledge of the results of the reference standard?	23	3	95
9b.	Were the reference standard results interpreted without knowledge of the results of the index test?	10	11	100
10.	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	95	26	0
11.	Were uninterpretable/intermediate test results reported?	22	25	74
12.	Were withdrawals from the study explained?	24	15	82

TABLE 6 *Full quality assessment: agreement between investigators*

Item		Proportion of agreement
1.	Was the spectrum of patients representative of the patients who will receive the test in practice?	0.62
2.	Were selection criteria clearly described?	0.75
3.	Is the reference standard likely to correctly classify the target condition?	0.75
4.	Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?	1.00
5.	Did the whole sample or a random selection of the sample receive verification using a reference standard of diagnosis?	0.87
6.	Did patients receive the same reference standard regardless of the index test result?	0.75
7.	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	0.87
8a.	Was the execution of the index test described in sufficient detail to permit replication of the test?	0.75
8b.	Was the execution of the reference standard described in sufficient detail to permit its replication?	0.62
9a.	Were the index test results interpreted without knowledge of the results of the reference standard?	0.87
9b.	Were the reference standard results interpreted without knowledge of the results of the index test?	0.75
10.	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	0.75
11.	Were uninterpretable/intermediate test results reported?	0.50
12.	Were withdrawals from the study explained?	0.75

time) to 1.00 (perfect agreement). Questions 4, 5, 7 and 9a resulted in a high level of agreement of 0.87 or above. Questions 1, 8b and 11 resulted in low levels of agreement of 0.62 or below. Disagreements were resolved by a third person reading the paper.

Studies identified: key characteristics

Where it is possible to undertake a meta-analysis or pool the results from a group of papers this will be reported in the text and tables. For those studies that could not be combined individual study results are reported. The shaded text in the tables illustrates the studies that presented data in a form that did not allow summary measures of diagnostic accuracy to be calculated. *Table 7* presents a summary of data and results of diagnostic accuracy for index tests compared with multichannel urodynamics.

Clinical history compared with urodynamics

USI in women

Twenty-one studies compared the diagnosis of USI in women by clinical history-taking and urodynamics (*Table 8*). Nineteen were performed in secondary care, one in primary care[125] and one did not specify where the study was performed.[119] All of these studies used the presence or absence of stress incontinence symptoms as their index test compared with the reference test of multichannel urodynamics, except for one study that used single-channel urodynamics as the reference standard.[165]

Fifteen studies provided a full cross-tabulation of results and the data from these studies were combined to provide a pooled sensitivity of 0.92 (95% CI 0.91 to 0.93) and specificity of 0.56 (95% CI 0.53 to 0.60) for the diagnosis of USI in women using a clinical history (*Figure 3*). Although all of these studies used symptoms of stress incontinence, it is probable that different amounts and types of questions were used and different cut-offs applied. Care should be taken, therefore, when interpreting the results. The positive likelihood ratio associated with the pooled sensitivity and specificity is 2.09 (95% CI 1.83 to 2.35) and the AUC for the ROC curve corresponding to the pooled DOR is 0.83 (95% CI 0.71 to 0.95) (*Figure 3*).

In addition to the pooled studies, three studies report sensitivity and specificity only.[161–163] These studies report sensitivities of 0.66, 0.96 and 0.56 and specificities of 0.63, 0.23 and 0.70, respectively. One study reported significantly higher stress symptoms in the USI-confirmed group than in the non-USI group[16] and one study reported that multichannel urodynamics confirmed USI in 89% of patients with stress incontinence symptoms.[164]

In addition, one study compared stress incontinence symptoms with single-channel urodynamics; a sensitivity of 0.92 and specificity of 0.39 is reported.[165]

DO in women

Fourteen studies compared the diagnosis of DO by clinical history and urodynamics (*Table 9*). Thirteen studies were performed in secondary care and one in primary care.[125]

Eight studies provided a full cross-tabulation of results and the data from these studies were combined to provide a pooled sensitivity of 0.61 (95% CI 0.57 to 0.65) and specificity of 0.87 (95% CI 0.85 to 0.89) for the diagnosis of DO in women by clinical history (*Figure 4*). Although all of these studies used symptoms of urge incontinence it is probable that different amounts and types of questions were used and different cut-offs applied. Again, care should be taken therefore when interpreting the results. The positive likelihood ratio associated with the pooled sensitivity and specificity is 4.69 (95% CI 4.05 to 5.33) and the AUC for the ROC curve corresponding to the pooled DOR is 0.83 (95% CI 0.69 to 0.97) (*Figure 4*).

In addition, two studies compared diagnosis by history with multichannel urodynamics in elderly women (*Figure 5*), resulting in a pooled sensitivity of 0.27 (95% CI 0.16 to 0.42) and specificity of 0.94 (95% CI 0.91 to 0.97).

Four papers presented only sensitivity and specificity from their studies.[162,163,166,167] The reported sensitivities were 0.70, 0.56, 0.40 and 0.53 and specificities 0.35, 0.70, 0.74 and 0.94, respectively.

Diagnosis of DO and USI in men

Three studies compared diagnosis made by clinical history and urodynamics in men (*Table 10*). In a post-prostatectomy population one study reports clinical history to be 1.00 sensitive and 0.50 specific for diagnosing USI and 0.50 sensitive and 0.77 specific for diagnosing DO.[168] One study reports a sensitivity of 0.73 and specificity of 0.60 for the diagnosis of DO by clinical history[169] and

TABLE 7 *Summary of data and results of diagnostic accuracy for index tests compared with multichannel urodynamics*

Reference	TP	FP	FN	TN	Sensitivity	95% CI	Specificity	95% CI	DOR	95% CI
Clinical history for USI in women										
Cundiff[117]	416	60	17	42	0.96	0.94 to 0.98	0.41	0.32 to 0.51	17.13	9.2 to 32.0
De Muylder[118]	228	58	14	108	0.94	0.91 to 0.97	0.65	0.57 to 0.72	30.33	16.20 to 56.76
Diokno[119]	65	14	30	52	0.68	0.58 to 0.78	0.79	0.67 to 0.88	8.05	3.87 to 16.73
Diokno[120]	145	40	9	6	0.94	0.89 to 0.97	0.13	0.05 to 0.26	2.42	0.81 to 7.19
FitzGerald[121]	187	51	22	33	0.90	0.85 to 0.93	0.39	0.29 to 0.51	5.5	2.95 to 10.25
Ishiko[122]	152	4	14	28	0.92	0.86 to 0.95	0.88	0.71 to 0.97	76	23.31 to 247.8
Korda[123]	451	39	52	24	0.90	0.87 to 0.92	0.38	0.26 to 0.51	5.34	2.98 to 9.57
Kujansuu[124]	46	20	11	43	0.81	0.68 to 0.90	0.68	0.55 to 0.79	8.99	3.86 to 20.93
Lagro-Janssen[125]	76	9	3	15	0.96	0.89 to 0.99	0.63	0.41 to 0.81	42.22	10.21 to 174.5
Niecestro[126]	13	17	3	32	0.81	0.54 to 0.96	0.65	0.50 to 0.78	8.16	2.04 to 32.63
Oulsander[127]	82	31	5	17	0.94	0.87 to 0.98	0.35	0.22 to 0.51	8.99	3.06 to 26.47
Ramsay[128]	72	28	19	81	0.79	0.69 to 0.87	0.74	0.65 to 0.82	10.96	5.65 to 21.28
Sand[129]	114	43	0	66	1.00	0.97 to 1.00	0.61	0.51 to 0.70	350.1	21.20 to 5780.2
Sandvik[130]	179	26	4	27	0.98	0.95 to 0.99	0.51	0.37 to 0.65	46.5	15.05 to 143.5
Sunshine[131]	73	14	0	15	1.00	0.95 to 1.00	0.52	0.33 to 0.71	157.1	8.89 to 2776.8
Pooled (RE)					**0.92**	**0.91 to 0.93**	**0.56**	**0.53 to 0.60**	**14.34**	**8.68 to 23.68**
LR+						**2.09 (95% CI 1.83 to 2.35)**				
Clinical history for DO in women										
Ishiko[122]	25	6	4	154	0.86	0.68 to 0.96	0.96	0.92 to 0.99	160.42	42.26 to 608.9
Cundiff[117]	42	17	60	416	0.41	0.32 to 0.51	0.96	0.94 to 0.98	17.129	9.17 to 32.0
Sandvik[130]	23	8	18	187	0.56	0.40 to 0.72	0.96	0.92 to 0.98	29.868	11.68 to 76.36
De Muylder[118]	147	91	89	81	0.62	0.56 to 0.69	0.47	0.40 to 0.55	1.47	0.99 to 2.19
Lagro-Janssen[125]	11	4	7	81	0.61	0.36 to 0.83	0.95	0.88 to 0.99	31.821	8.00 to 126.6
Sand[129]	10	3	20	185	0.33	0.17 to 0.53	0.98	0.95 to 1.00	30.833	7.83 to 121.4
FitzGerald[121]	10	21	27	235	0.27	0.14 to 0.44	0.92	0.88 to 0.95	4.145	1.77 to 9.72
Cantor[132]	107	53	11	43	0.91	0.84 to 0.95	0.45	0.35 to 0.55	7.892	3.77 to 16.53
Pooled (RE)					**0.61**	**0.57 to 0.65**	**0.87**	**0.85 to 0.89**	**14.72**	**4.87 to 44.5**
LR+						**4.69 (95% CI 4.05 to 5.33)**				
Clinical history for DO in elderly women[a]										
Diokno[120]	2	6	12	180	0.14	0.02 to 0.43	0.97	0.93 to 0.99	5.00	0.91 to 27.47
Ouslander[127]	12	10	25	88	0.32	0.18 to 0.50	0.90	0.82 to 0.95	4.22	1.63 to 10.92

continued

TABLE 7 *Summary of data and results of diagnostic accuracy for index tests compared with multichannel urodynamics (cont'd)*

Reference	TP	FP	FN	TN	Sensitivity	95% CI	Specificity	95% CI	DOR	95% CI
Validated scale:										
UDI-6 scale for USI in women										
Lemack[97]	39	30	7	52	0.85	0.71 to 0.94	0.63	0.52 to 0.74	9.66	3.84 to 24.27
FitzGerald[121]	135	22	18	27	0.88	0.82 to 0.93	0.55	0.40 to 0.69	9.21	4.36 to 19.44
Pooled (RE)					**0.87**	**0.82 to 0.92**	**0.60**	**0.51 to 0.69**	**9.38**	**5.25 to 16.77**
LR+						**2.18 (95% CI 1.49 to 2.86)**				
DIS scale for USI in women										
Klovning[133]	92	20	61	68	0.60	0.52 to 0.68	0.77	0.67 to 0.85	5.13	2.83 to 9.29
Pad test										
ICS 1-hour pad test for any leakage in women										
Jorgensen[134]	16	18	1	14	0.94	0.73 to 0.99	0.44	0.28 to 0.61	12.44	1.47 to 105.5
48 hour pad test for USI in women										
Versi[135]	57	12	5	31	0.92	0.82 to 0.97	0.72	0.57 to 0.83	29.45	9.50 to 91.3
Urinary diary for DO in women										
Contreras Ortiz[136]	23	33	3	158	0.88	0.71 to 0.96	0.83	0.77 to 0.87	36.71	10.41 to 129.4
Q-tip test for USI in women[a]										
Bergman[137]	38	27	13	37	0.75	0.60 to 0.86	0.58	0.45 to 0.70	4.01	1.80 to 8.93
Montz[138]	35	16	31	18	0.53	0.40 to 0.65	0.53	0.35 to 0.70	1.27	0.55 to 2.91
Pooled (RE)									**2.27**	**0.74 to 6.99**
Ultrasound: observed leakage for USI in women										
Dietz[139]	10	3	6	18	0.63	0.35 to 0.85	0.86	0.64 to 0.97	10.00	2.05 to 48.89
Dietz[140]	66	9	13	29	0.84	0.74 to 0.91	0.76	0.60 to 0.89	16.36	6.29 to 42.5
Dietz[141]	33	2	2	15	0.94	0.81 to 0.99	0.88	0.64 to 0.99	123.8	15.89 to 964.0
Quinn[142]	87	6	3	28	0.97	0.91 to 0.99	0.82	0.66 to 0.93	135.3	31.8 to 576
Pooled (RE)					**0.89**	**0.84 to 0.93**	**0.82**	**0.73 to 0.89**	**36.784**	**10.19 to 132.8**
LR+						**4.94 (95% CI 3.88 to 6.01)**				
Ultrasound: bladder neck descent for USI in women										
Chen[98]	27	15	10	50	0.73	0.56 to 0.86	0.77	0.65 to 0.87	9.00	3.56 to 22.74
Bergman[143]	38	2	6	45	0.86	0.73 to 0.95	0.96	0.86 to 1.00	142	27.16 to 747
Bergman[144]	30	3	2	26	0.94	0.79 to 0.99	0.90	0.73 to 0.98	130	20.14 to 839
Pooled (RE)					**0.84**	**0.76 to 0.90**	**0.86**	**0.79 to 0.91**	**49.24**	**6.27 to 386**
LR+						**6.00 (95% CI 4.72 to 7.28)**				

continued

TABLE 7 *Summary of data and results of diagnostic accuracy for index tests compared with multichannel urodynamics (cont'd)*

Reference	TP	FP	FN	TN	Sensitivity	95% CI	Specificity	95% CI	DOR	95% CI
X-ray: observed leakage for USI in women										
Pelsang[145]	37	29	24	69	0.61	0.47 to 0.73	0.70	0.60 to 0.79	3.67	1.87 to 7.19
Scotti[146]	53	18	35	68	0.60	0.49 to 0.71	0.79	0.69 to 0.87	5.72	2.92 to 11.21
Pooled (RE)					**0.60**	**0.52 to 0.68**	**0.74**	**0.68 to 0.81**		**2.85 to 7.37**
LR+						**2.31 (95% CI 1.62 to 3.00)**				
X-ray: bladder neck descent for USI in women										
Grischke[147]	20	20	14	30	0.59	0.41 to 0.75	0.60	0.45 to 0.74	2.14	0.88 to 5.20
Bergman[148]	32	15	0	12	1.00	0.89 to 1.00	0.44	0.26 to 0.65	52.4	2.91 to 943
Pooled (RE)					**0.79**	**0.67 to 0.88**	**0.55**	**0.43 to 0.66**	**8.11**	**0.30 to 222**
LR+						**1.76 (95% CI 0.90 to 2.61)**				
Full bladder clinical stress test for USI in women										
Hsu[149]	29	1	2	9	0.94	0.79 to 0.99	0.90	0.56 to 1.00	130.5	10.56 to 1612
Kadar[150]	14	5	4	14	0.78	0.52 to 0.94	0.74	0.49 to 0.91	9.8	2.17 to 44.32
Scotti[17]	68	10	13	54	0.84	0.74 to 0.91	0.84	0.73 to 0.92	28.25	11.50 to 69.37
Pooled (RE)					**0.85**	**0.78 to 0.91**	**0.83**	**0.74 to 0.90**	**25.42**	**8.66 to 74.6**
LR+						**5.00 (95% CI 3.79 to 6.21)**				
Single-channel urodynamics										
Standing SCU for DO in women										
Sand[151]	43	15	8	34	0.84	0.71 to 0.93	0.69	0.55 to 0.82	12.18	4.62 to 32.10
Sand[152]	15	8	46	134	0.25	0.15 to 0.37	0.94	0.89 to 0.98	5.46	2.17 to 13.72
Sutherst[153]	35	7	0	58	1.00	0.90 to 1.00	0.89	0.79 to 0.96	553	30.69 to 9993
Pooled (RE)					**0.63**	**0.55 to 0.71**	**0.88**	**0.84 to 0.92**	**19.03**	**3.34 to 108.5**
LR+						**12.00 (95% CI 10.58 to 13.42)**				
Supine SCU for DO in elderly women										
Fonda[154]	17	7	4	15	0.81	0.58 to 0.95	0.68	0.45 to 0.86	9.11	2.22 to 37.34
Ouslander[155]	61	10	23	43	0.73	0.62 to 0.82	0.81	0.68 to 0.91	11.40	4.93 to 26.38
Pooled (RE)					**0.74**	**0.65 to 0.82**	**0.77**	**0.66 to 0.86**	**10.75**	**5.23 to 22.12**
LR+						**3.57 (95% CI 2.41 to 4.73)**				
Supine SCU for DO in elderly men										
Fonda[154]	20	0	1	6	0.95	0.76 to 1.00	1.00	0.54 to 1.00	177.7	6.42 to 4914
Ouslander[155]	21	1	4	5	0.84	0.64 to 0.96	0.833	0.36 to 1.00	26.25	2.39 to 288
Pooled (RE)					**0.89**	**0.76 to 0.96**	**0.92**	**0.62 to 1.00**	**50.58**	**7.24 to 353**
LR+						**18.20 (95% CI 12.62 to 23.78)**				

continued

TABLE 7 *Summary of data and results of diagnostic accuracy for index tests compared with multichannel urodynamics (cont'd)*

Reference	TP	FP	FN	TN	Sensitivity	95% CI	Specificity	95% CI	DOR	95% CI
Supine SCU for USI in women										
Resnick[156]	19	10	3	64	0.86	0.67 to 0.95	0.86	0.77 to 0.92	40.53	10.11 to 162.4
Ambulatory urodynamics for USI in women										
Davis[157]	7	38	2	3	0.78	0.45 to 0.94	0.07	0.03 to 0.19	0.28	0.039 to 1.97
Sitting UPP for USI in women										
Swift[158]	32	1	33	42	0.49	0.37 to 0.62	0.98	0.88 to 1.00	40.73	5.29 to 313
Richardson[159]	30	43	5	59	0.86	0.70 to 0.95	0.58	0.48 to 0.68	8.23	2.95 to 22.95
Pooled (RE)					**0.62**	**0.52 to 0.72**	**0.70**	**0.61 to 0.77**	**14.46**	**3.06 to 68.2**
LR+						**9.38 (95% CI 6.91 to 11.84)**				
Supine UPP for USI in women										
Versi[160]	54	21	16	81	0.77	0.66 to 0.85	0.79	0.71 to 0.86	13.02	6.24 to 27.17

[a] Not pooled owing to excessive clinical heterogeneity.

DIS, Detrusor Instability Score; FN, false negative; FP, false positive; RE, random effects; TN, true negative; TP, true positive; UDI, Urogenital Distress Inventory; UPP, urethral pressure profile.

TABLE 8 *Clinical history compared with urodynamics for USI in women*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Statistical tests	Main findings
Cundiff[117]	535	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	History (stress symptoms)	Full contingency table	Sensitivity = 0.96 Specificity = 0.41
Niecestro[126]	66	F	Secondary	Referred for UD	USI	Multichannel urodynamics	History (stress symptoms)	Full contingency table	Sensitivity = 0.81 Specificity = 0.65
Diokno[119]	456	F	Not specified	Symptoms of UI + controls	USI	Multichannel urodynamics	History (stress symptoms)	Full contingency table	Sensitivity = 0.68 Specificity = 0.79
Ishiko[122]	198	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	History (stress symptoms)	Full contingency table	Sensitivity = 0.92 Specificity = 0.88
Sandvik[130]	250	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	History (stress symptoms)	Full contingency table	Sensitivity = 0.98 Specificity = 0.51
De Muylder[118]	408	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	History (stress symptoms)	Full contingency table	Sensitivity = 0.94 Specificity = 0.65
Lagro-Janssen[125]	103	F	Primary	Symptoms of UI	USI	Multichannel urodynamics	History (stress symptoms)	Full contingency table	Sensitivity = 0.96 Specificity = 0.63
Sand[129]	188	F	Secondary	LUTS	USI	Multichannel urodynamics	History (stress symptoms)	Full contingency table	Sensitivity = 1.00 Specificity = 0.61
Diokno[120]	200	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	History (stress symptoms)	Full contingency table	Sensitivity = 0.94 Specificity = 0.13
Ouslander[127]	135	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	History (stress symptoms)	Full contingency table	Sensitivity = 0.94 Specificity = 0.35
Kujansuu[124]	121	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	History (stress symptoms)	Full contingency table	Sensitivity = 0.80 Specificity = 0.68
FitzGerald[121]	293	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	History (stress symptoms)	Full contingency table	Sensitivity = 0.89 Specificity = 0.39
Sunshine[131]	109	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	History (stress symptoms)	Full contingency table	Sensitivity = 1.00 Specificity = 0.52
Korda[123]	566	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	History (stress symptoms)	Full contingency table	Sensitivity = 0.90 Specificity = 0.38
Ramsay[128]	200	F	Secondary	Positive urodynamics	USI/DO	Multichannel urodynamics	History (stress symptoms)	Full contingency table	Sensitivity = 0.79 Specificity = 0.74

continued

TABLE 8 *Clinical history compared with urodynamics for USI in women (cont'd)*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Statistical tests	Main findings
Weidner[161]	950	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	History (stress symptoms)	Sensitivity and Specificity	Sensitivity = 0.66 Specificity = 0.63
Clarke[162]	100	F	Secondary	LUTS	USI	Multichannel urodynamics	History (stress symptoms)	Sensitivity and specificity	Sensitivity = 0.96 Specificity = 0.23
Bergman[163]	154	F	Secondary	Symptoms of UI + controls	USI/DO	Multichannel urodynamics	History (stress symptoms)	Sensitivity and specificity	Sensitivity = 0.56 Specificity = 0.70
Amundsen[16]	115	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	History (stress symptoms)	Difference between USI and non-USI groups	Significantly higher SI symptoms in USI group
Ng[164]	28	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	History (stress symptoms)	Agreement between two methods	USI confirmed in 89% of those with stress symptoms
Le Coutour[165]	154	F	Secondary	Symptoms of UI	USI	Singlechannel urodynamics	History (stress symptoms)	Full contingency table	Sensitivity = 0.92 Specificity = 0.39

F, female; LUTS, lower urinary tract symptoms; UI, urinary incontinence.
The shaded area in *Tables 8–34* indicates studies that presented data in a form that did not allow summary measures of diagnostic accuracy to be calculated.

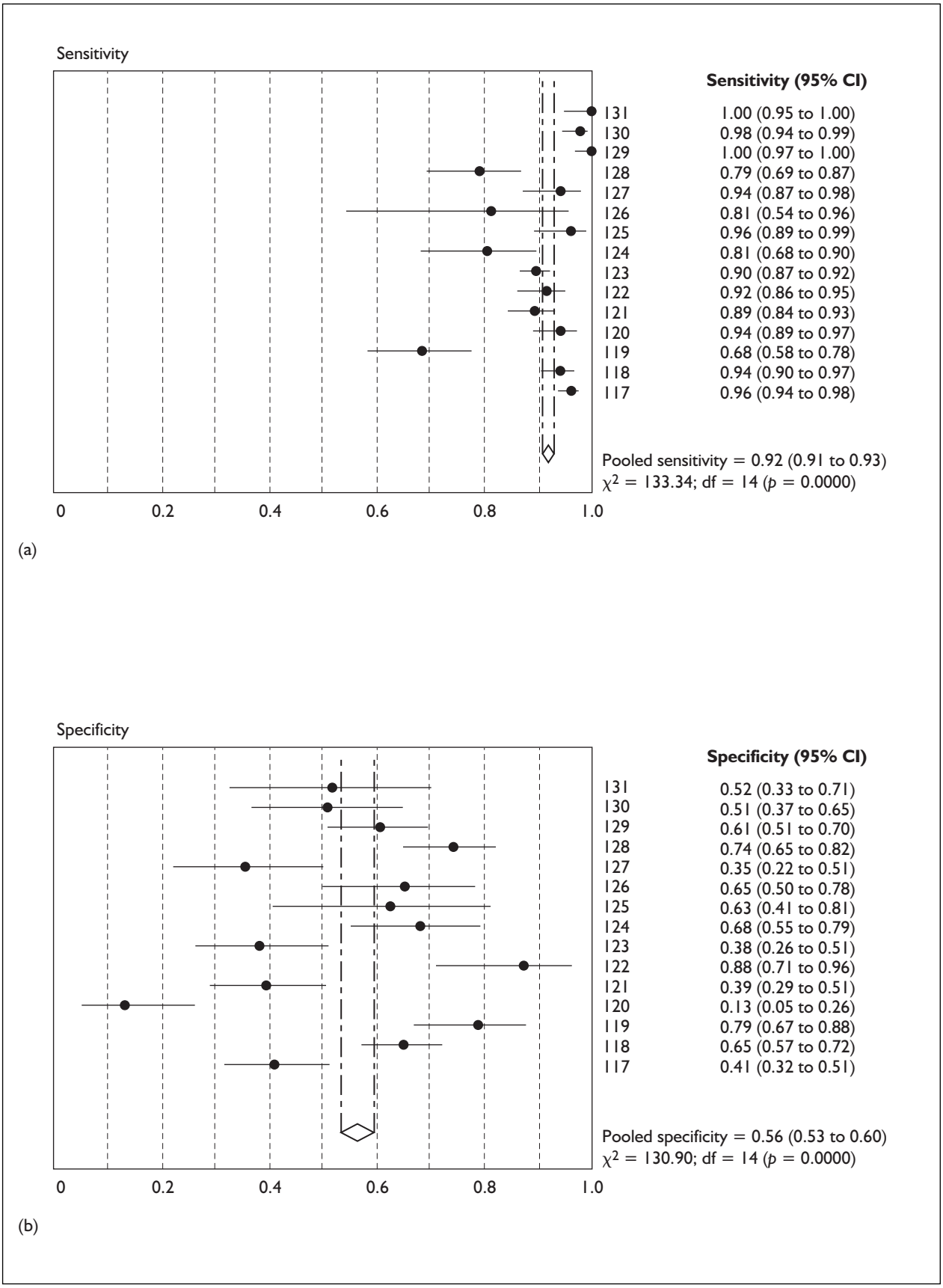
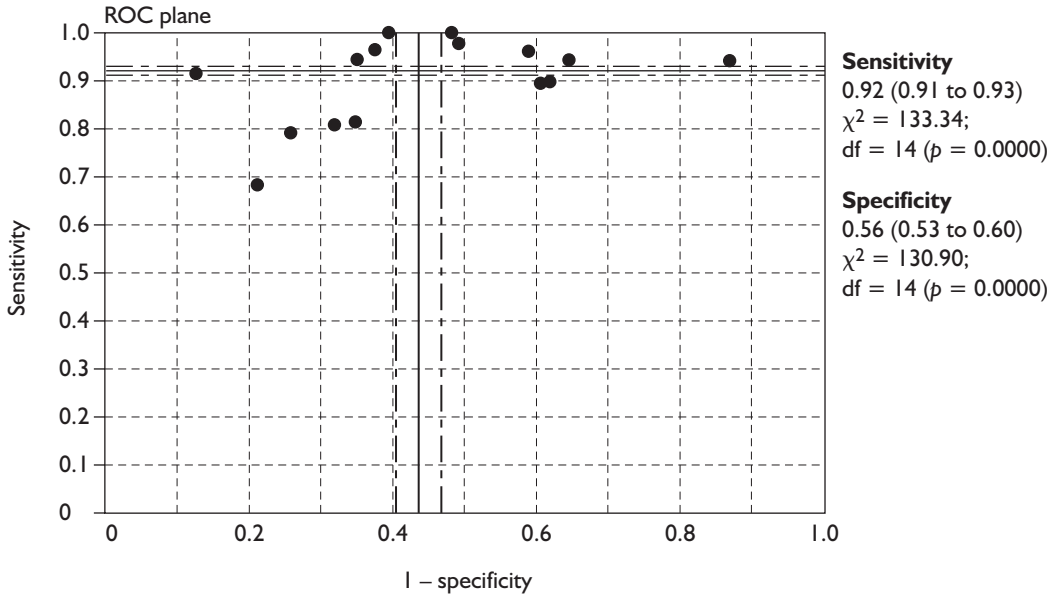
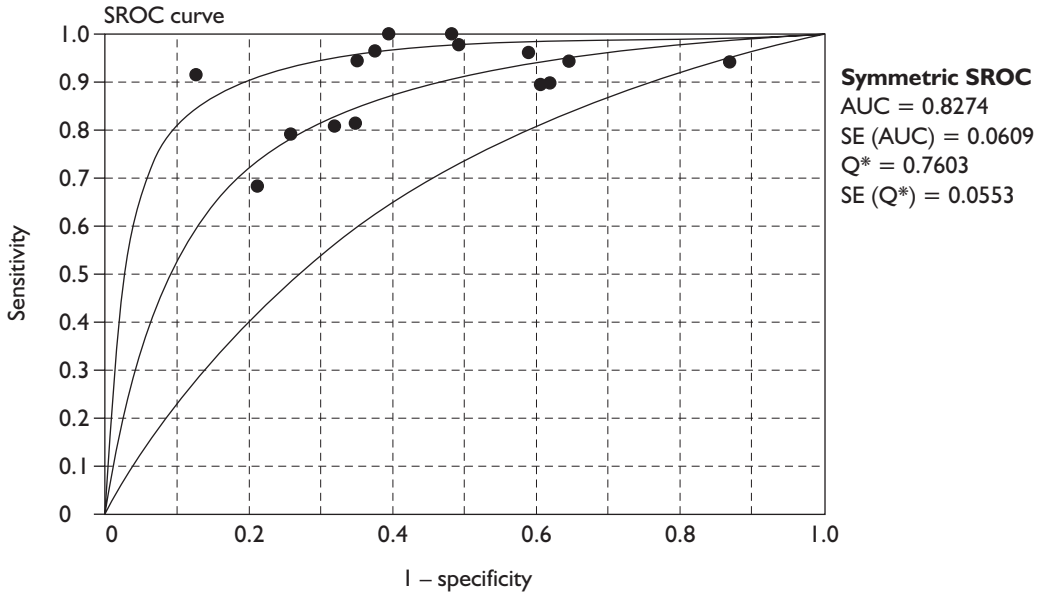


FIGURE 3 *Pooled random effect results: clinical history versus multichannel urodynamics (MCU) for diagnosis of USI in women. (a) Independently pooled sensitivity; (b) independently pooled specificity; (c) sensitivity and specificity for each study and pooled estimates plotted in ROC space; (d) pooled DOR (random effect) plotted in ROC space. SROC, summary receiver operating characteristic.*



(c)



(d)

Pooled (random effect) DOR 14.339 (95% CI 8.682 to 23.681)
Heterogeneity $\chi^2 = 62.07$ (df = 14) $p = 0.000$

FIGURE 3 (cont'd) *Pooled random effect results: clinical history versus multichannel urodynamics (MCU) for diagnosis of USI in women. (a) Independently pooled sensitivity; (b) independently pooled specificity; (c) sensitivity and specificity for each study and pooled estimates plotted in ROC space; (d) pooled DOR (random effect) plotted in ROC space. SROC, summary receiver operating characteristic.*

TABLE 9 *Clinical history compared with urodynamics for DO in women*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Statistical tests	Main findings
Ishiko[122]	198	F	Secondary	Symptoms of UI	DO	Multichannel urodynamics	History (urge symptoms)	Full contingency table	Sensitivity = 0.86 Specificity = 0.96
Cundiff[117]	535	F	Secondary	Symptoms of UI	DO	Multichannel urodynamics	History (urge symptoms)	Full contingency table	Sensitivity = 0.41 Specificity = 0.96
Sandvik[130]	250	F	Secondary	Symptoms of UI	DO	Multichannel urodynamics	History (urge symptoms)	Full contingency table	Sensitivity = 0.56 Specificity = 0.96
De Muylder[118]	408	F	Secondary	Symptoms of UI	DO	Multichannel urodynamics	History (urge symptoms)	Full contingency table	Sensitivity = 0.62 Specificity = 0.47
Lagro-Janssen[125]	103	F	Primary	Symptoms of UI	DO	Multichannel urodynamics	History (urge symptoms)	Full contingency table	Sensitivity = 0.61 Specificity = 0.95
Sand[129]	188	F	Secondary	LUTS	DO	Multichannel urodynamics	History (urge symptoms)	Full contingency table	Sensitivity = 0.33 Specificity = 0.98
FitzGerald[121]	293	F	Secondary	Symptoms of UI	DO	Multichannel urodynamics	History (urge symptoms)	Full contingency table	Sensitivity = 0.27 Specificity = 0.92
Cantor[132]	214	F	Secondary	Symptoms of UI	DO	Multichannel urodynamics	History (urge symptoms)	Full contingency table	Sensitivity = 0.91 Specificity = 0.45
Diokno[120]	200	F	Secondary	Symptoms of UI (elderly)	DO	Multichannel urodynamics	History (urge symptoms)	Full contingency table	Sensitivity = 0.14 Specificity = 0.97
Ouslander[127]	135	F	Secondary	Symptoms of UI (elderly)	DO	Multichannel urodynamics	History (urge symptoms)	Full contingency table	Sensitivity = 0.32 Specificity = 0.90
Clarke[162]	100	F	Secondary	LUTS	DO	Multichannel urodynamics	History	Sensitivity and specificity	Sensitivity = 0.70 Specificity = 0.35
Bergman[163]	154	F	Secondary	Symptoms of UI + controls	USI/DO	Multichannel urodynamics	History (range of symptoms)	Sensitivities and specificities	Mean Sensitivities = 56 ± 17 Specificities = 70 ± 23
Petros[166]	113	F	Secondary	Symptoms of UI	DO	History	Multichannel urodynamics	Sensitivity and Specificity	Sensitivity = 0.40 Specificity = 0.74
Van Doorn[167]	228	F	Secondary	Symptoms of UI	DO	Ambulatory urodynamics	History (urge symptoms)	Full contingency table	Sensitivity = 0.53 Specificity = 0.94

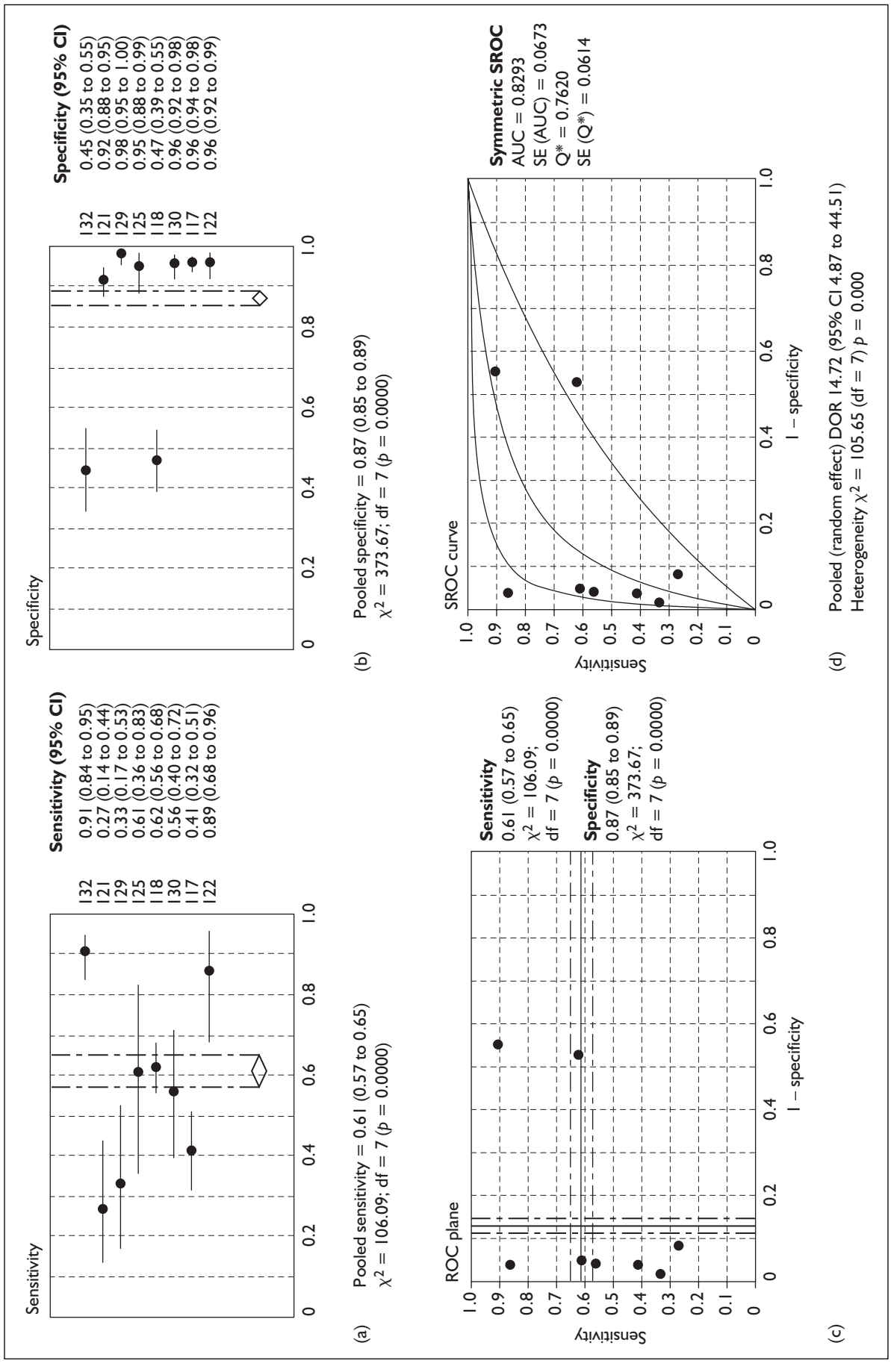


FIGURE 4 *Pooled random effect results: clinical history versus MCU for diagnosis of DO in women. (a) Independently pooled sensitivity; (b) independently pooled specificity; (c) sensitivity and specificity for each study and pooled estimates plotted in ROC space; (d) pooled DOR (random effect) plotted in ROC space.*

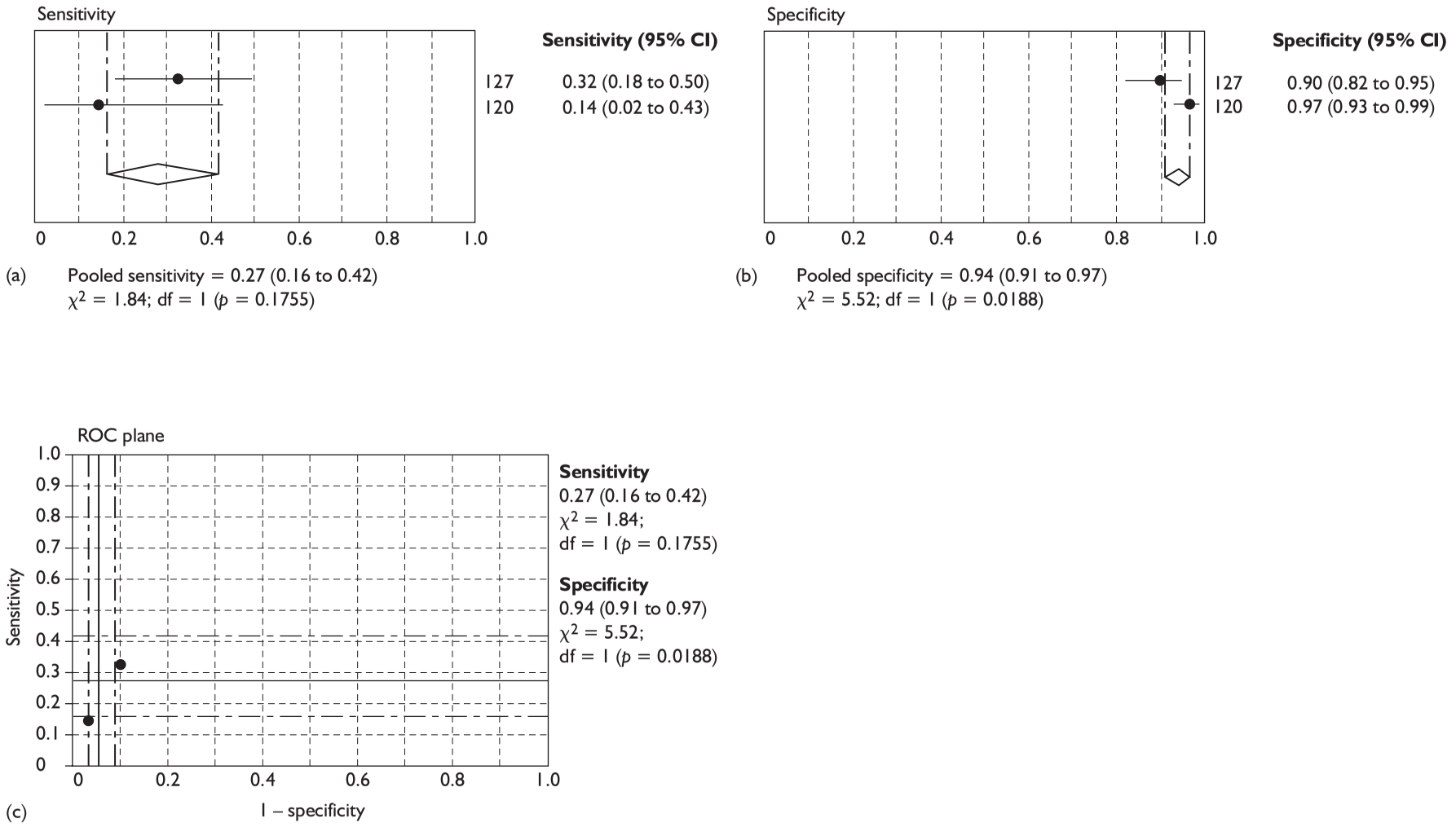


FIGURE 5 *Pooled random effect results: clinical history versus MCU for diagnosis of DO in elderly women. (a) Independently pooled sensitivity; (b) independently pooled specificity; (c) sensitivity and specificity for each study and pooled estimates plotted in ROC space.*

one study reports a higher incidence of urge symptoms in a urodynamically confirmed DO group compared with a urodynamically normal group.[170]

Diagnosis of USI and DO in a mixed population

Three studies compared diagnosis by clinical history and multichannel urodynamics in a mixed population (*Table 11*). One study reports a sensitivity of 1.00 and specificity of 0.95 for the diagnosis of USI by history taking of stress incontinence symptoms.[171] One study reports an agreement of 93% (USI) and 63% (DO) between the two methods,[172] and one reports an agreement of 60% for the diagnosis of USI.[173]

Validated scale compared with clinical history

One study compared the association of the Incontinence Impact Questionnaire (IIQ-6) and the Urogenital Distress Inventory (UDI-7) with various incontinence symptoms (*Table 12*). Correlation coefficients of between 0.24 and 0.69 were found.

Validated scale compared with validated scale

One study compared the association between the long and short forms of the IIQ and the UDI (*Table 13*). These scales measure the life impact and symptom distress of urinary incontinence in women. Correlations of $r = 0.93$ (UDI) and 0.97 (IIQ) were found between the two forms of the questionnaires, indicating that the shortened versions are equally as valid for the measurement of these quality of life symptoms.

Validated scale compared with pad test

Four papers reported a comparison of a validated scale with a pad-test (*Table 14*). All four studied only female patients.

Three papers did not present data in a way that allowed sensitivity and specificity to be calculated.[107,113,114] Attempts to contact the authors resulted in one response with the full data requested.[107]

One study investigated the association between the UDI and IIQ long form with a 1-hour pad test.[113] These scales were developed to assess the impact of urinary incontinence on activity and emotions and the degree to which symptoms of incontinence are distressing. Data were not presented in a way that allowed sensitivity and specificity to be calculated. However, the authors present an ROC analysis that shows that there was a 54% and 51% probability of correctly classifying incontinence as measured by the pad test for the IIQ and UDI, respectively.

One paper aimed to validate further the Sandvik severity index, this time with the association with a 48-hour pad test.[114] Insufficient data were presented to allow sensitivity or specificity to be calculated. The correlation between the severity index and leakage of the pad test was $r = 0.36$ ($p < 0.001$).

One study investigated a new screening questionnaire designed for women in primary care, the Incontinence Screening Questionnaire (ISQ) and compared it against the 48-hour pad test.[176] This resulted in a sensitivity of 0.65 and a specificity of 0.80 (cut-off for pad test = 7 g, positive ISQ = responded positively to at least one of the eight items).

One study evaluated the Sandvik scale, a three or four-level severity scale, against the 24-hour pad test.[107] When contacted, the author sent individual patient data for 315 cases allowing numerous cut-off points to be used. Based on positive cut-off point of above 1 for the severity scale and 7 g for the pad test, the scale was found to be 0.74 sensitive and 0.76 specific.

Validated scales compared with urodynamics

Eight studies compared the use of validated scales with standard multichannel urodynamics for the diagnosis of urinary incontinence (*Table 15*). Six studies investigated female patients[97,106,121,133,177,178] and two studied male patients.[101,179] Six separate scales were studied by the eight studies in this group.

Three papers studied the UDI.[97,121,178] Two papers used the response on question 3 of the short form of the scale (Are you bothered by urinary leakage caused by physical exercise?) to predict urodynamic diagnosis of USI.[97,121] These studies report sensitivities of 0.85 and 0.88 and specificities of 0.63 and 0.55. Owing to the homogeneity of these papers it was possible to combine the data to produce a pooled sensitivity of 0.87 (95% CI 0.82 to 0.92) and specificity of 0.60 (95% CI 0.51 to 0.69) for the diagnosis of USI from question 3 of the UDI-6 (*Figure 6*). One other paper reported a correlation of $r = 0.54$ between diagnosis using multichannel urodynamics and score on the UDI.[179]

One paper investigated the use of a Detrusor Instability Score (DIS),[133] a ten-question scale

TABLE 10 *Clinical history compared with urodynamics in men*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Statistical tests	Main findings
Ficazzola[168]	60	M	Secondary	Post-prostatectomy incontinence	USI/DO	Multichannel urodynamics	History	Full contingency table	USI: Sensitivity = 1.00 Specificity = 0.50 DO: Sensitivity = 0.50 Specificity = 0.77
Ding[169]	126	M	Secondary	LUTS	DO	Multichannel urodynamics	History	Sensitivity and specificity	Sensitivity = 0.73 Specificity = 0.60
Hyman[170]	160	M	Not specified	LUTS	DO	Multichannel urodynamics	History	Differences between diagnostic groups	Higher incidence of urge symptoms associated with DO group

M, male.

TABLE 11 *Clinical history compared with urodynamics in a mixed population*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Statistical tests	Main findings
Porru[171]	46	Mixed	Secondary	Symptoms of UI	USI	Multichannel urodynamics	History	Full contingency table (symptoms of USI)	Sensitivity = 1.00 Specificity = 0.95
Gray[172]	148	Mixed	Secondary	Symptoms of UI	USI/DO	Multichannel urodynamics	History	Agreement between methods	USI: 93% DO: 63%
De Bolla[173]	82	Mixed	Secondary	Symptoms of UI	USI/DO	Multichannel urodynamics	History	Agreement between methods	Agreement = 60%

TABLE 12 *Validated scale compared with clinical history*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Robinson[174]	384	F	Primary	Symptoms of UI (elderly)	Any leakage	History (various measures of incontinence)	IIQ-7 UDI-6	Correlation	Range r = 0.24–69 to 0.69

TABLE 13 *Validated scale compared with validated scale*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Statistical tests	Main findings
Uebersax[175]	162	F	Clinical trial for UI	Diagnosis of UI	USI/DO	UDI (long form) IIQ (long form)	UDI (short form) IIQ (short form)	Correlation	UDI: 0.93 IIQ: 0.97

TABLE 14 *Validated scale compared with pad test*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Gunthorpe[176]	89	F	Primary	Attending GP's surgery	All	48-hour pad test (cut-off = 7 g)	ISQ	Full contingency table	Sensitivity = 0.65 Specificity = 0.80
Sandvik[107]	315	F	Secondary	Symptoms of UI	All	24-hour pad test	Sandvik severity index	Individual patient data	Sensitivity = 0.74 Specificity = 0.76 (optimum)
Harvey[113]	150	F	Clinical trial	USI or DO positive	USI/DO	1-hour pad test (cut-off = 2 g)	IIQ long form	Correlation	r = 0.18
Hanley[114]	237	F	Primary + secondary	Symptoms of UI	All	48-hour pad test	Sandvik severity index	Differences between groups	Significant difference between severity groups

TABLE 15 *Validated scale compared with urodynamics*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Lemack[97]	128	F	Secondary	LUTS	USI	Multichannel urodynamics	UDI-6 (question 3)	Full contingency table	Sensitivity = 0.85 Specificity = 0.63
FitzGerald[121]	293	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	IIQ-7 UDI-6 (question 3)	Full contingency table	Sensitivity = 0.88 Specificity = 0.55
Klovning[133]	250	F	Secondary	Referred for 'urogenital dysfunction'	USI	Multichannel urodynamics	DIS	Full contingency table ROC curve	Sensitivity = 0.60 Specificity = 0.77 (optimum)
Haeusler[177]	1938	F	Secondary	Referred for urodynamics	USI/DO	Multichannel urodynamics	Gaudenz incontinence questionnaire (women)	Sensitivity and specificity	USI: Sensitivity = 0.56 Specificity = 0.45 DO: Sensitivity = 0.62 Specificity = 0.56
Shumaker[178]	162	F	Clinical trial	Symptoms of UI	USI/DO	Multichannel urodynamics	UDI IIQ	Correlation	$r = 0.54$
Nitti[179]	50	M	Secondary	Symptoms of voiding dysfunction	DO	Multichannel urodynamics	AUA symptom index (men)	Difference in score between diagnostic groups	No significant difference between DO group and non-DO group
Nitti[101]	83	M	Secondary	Symptoms of BPH	DO	Multichannel urodynamics	AUA men	Difference in score between diagnostic groups	Higher irritative symptoms scores associated with DO
Nager[106]	52	F	Secondary	USI	USI	Multichannel urodynamics	Quality of life questionnaire	Correlation	No statistically significant correlation

BPH, benign prostatic hyperplasia.

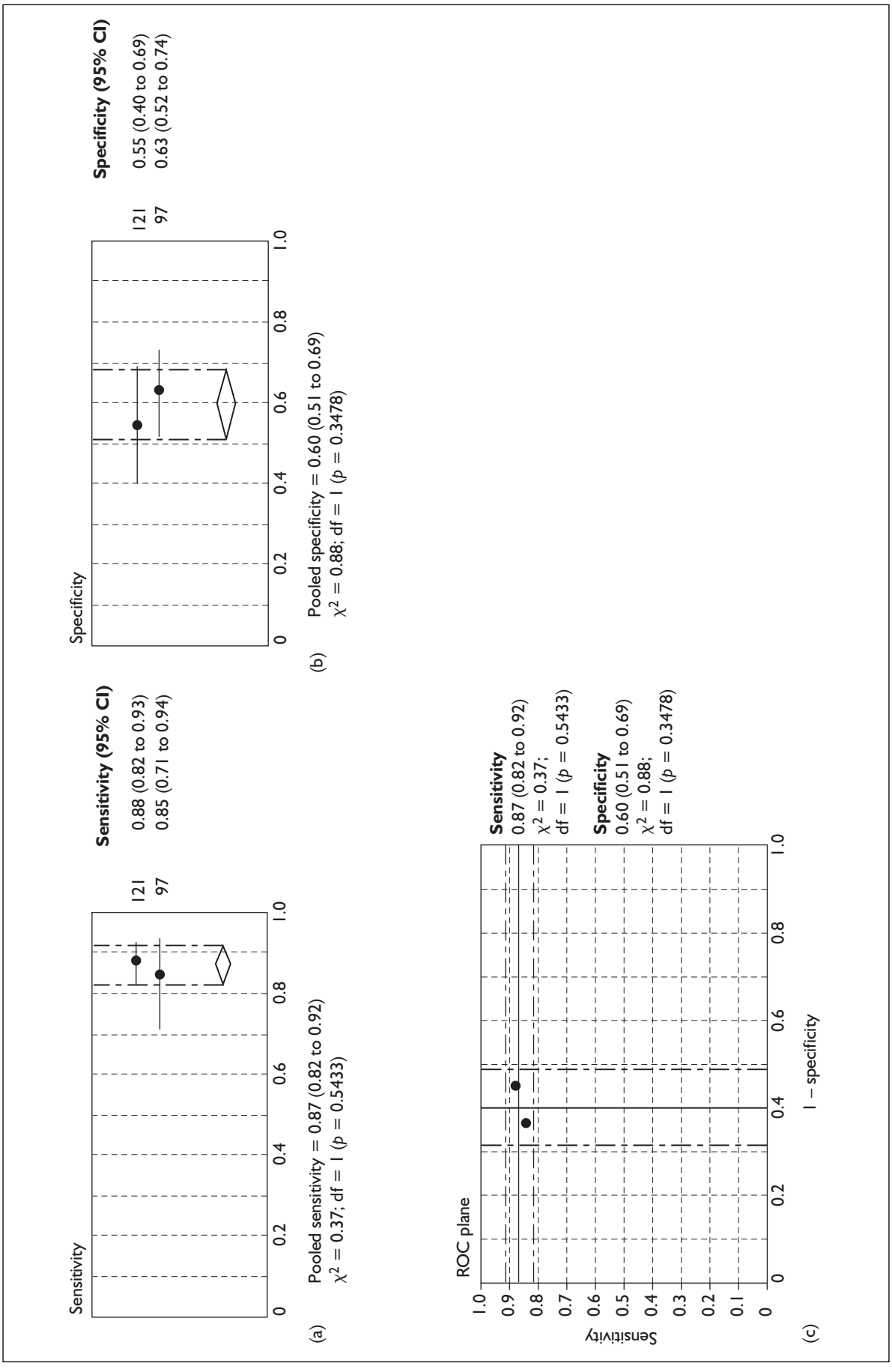


FIGURE 6 *Pooled random effect results: validated scale versus MCU for diagnosis of USI in women. (a) Independently pooled sensitivity; (b) independently pooled specificity; (c) sensitivity and specificity for each study and pooled estimates plotted in ROC space.*

designed to highlight either USI or DO. This study reports an optimum sensitivity of 0.60 and specificity of 0.77 for the diagnosis of USI. One paper[177] studied the ability of the Gaudenz incontinence questionnaire to diagnose USI and DO; this consists of 26 questions and also allows grading of severity of the type of incontinence. The paper reports sensitivities of 0.56 and 0.62 and specificities of 0.45 and 0.56 for the diagnosis of USI and DO, respectively.

The ability of the American Urological Association (AUA) symptom index to diagnose DO in male patients was studied by two papers.[101,179] Both papers compared the scores on the seven-question AUA symptom index with diagnosis using multichannel urodynamics. Neither paper presented data in a format that allowed summary statistics of diagnostic accuracy to be calculated. One paper[179] found no difference in AUA symptom score between DO and non-DO groups; however, the other found that those patients with DO had significantly higher irritative scores on the AUA.[101]

One paper studied the correlation between urodynamic diagnosis and score on a quality of life questionnaire (SEAPI QMM incontinence classification system) in women with confirmed USI.[106] This study found no statistically significant correlation between the two methods.

Pad test compared with clinical history

Six studies compared a pad test with clinical history for the assessment of urinary incontinence (*Table 16*). One study included both male and female patients,[180] the other five only females. Four studies were performed in secondary care, one in primary care and one did not specify where it was performed.

Three types of pad test were studied. Two studies investigated the use of the 48-hour pad test; one reported a sensitivity of 0.73 and specificity of 1.00 for the prediction of patient-reported incontinence status.[180] One paper assigned patients to three severity groups according to their self-reported urine loss and found significant differences in mean urinary loss between the three groups as measured by the 48-hour pad test.[112]

Two papers studied the 24-hour pad test: one study[115] comparing the mean pad weight gain between self-reported incontinent and continent patient groups found no significant differences between the two groups. The other,[112] however, found significant differences in mean pad weight gain between three groups of patients grouped according to the self-perceived severity of their symptoms.

Four papers compared a short-term pad test with patient history. Presenting individual patient data, one study reported an optimum sensitivity of 0.87 and specificity of 0.64 for the rapid exercise pad test for predicting self-reported incontinence status.[111] For the same test a second study reported a sensitivity of 0.90 and specificity of 1.00; however, as the raw data were not presented in this paper it was not possible to pool these results.[181] A third study reported correlations of between $r = 0.31$ and 0.67 between the 1-hour pad test and various history questions, with the largest correlation being between the pad test and self-reported number of incontinent episodes.[99] In the fourth study when the ICS 1-hour pad test was compared with self-reported grade of incontinence severity, significant differences between mean pad weight gain were found across the three groups.[112]

Pad test compared with urodynamics

Seven studies were identified that compared the use of a pad test with urodynamics (*Table 17*). All studies used only female patients and were performed in a secondary care setting, apart from one study that was conducted in mixed care settings.[105]

Two studies presented data in a cross-tabulated format that allowed sensitivity and specificity to be calculated. One study found the ICS 1-hour pad test to be 0.94 sensitive and 0.45 specific for diagnosing any leakage compared with multichannel urodynamics;[134] the other found the 48-hour pad test to be 0.92 sensitive and 0.72 specific for diagnosing USI.[135]

Four other papers studied the use of short-term pad tests for diagnosing USI compared with multichannel urodynamics. One study found a rapid exercise pad test to be 0.86 sensitive in diagnosing patients with a urodynamic diagnosis of USI.[105] A second study compared three different pad tests: unknown volume, 250 ml and 1 hour, also in urodynamically positive patients, and found sensitivities ranging from 0.79 to 0.94.[182] A third study reported a correlation between the rapid exercise pad test and multichannel urodynamics of 0.59.[183] Finally, in a fourth study significantly higher results of the ICS 1-hour test, 24-hour and 48-hour pad tests were found in urodynamically confirmed USI compared with asymptomatic controls.

TABLE 16 *Pad test compared with clinical history*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Hellstrom[180]	37	Mixed	Primary	Symptoms of UI (all 85 years old)	USI	History (severity of UI)	48-hour pad test	Individual patient data	Sensitivity = 0.73 Specificity = 1.00 (optimum)
Hahn[111]	50	F	Secondary	Symptoms of USI	USI	History (severity of UI)	Exercise pad test	Individual patient data	Sensitivity = 0.87 Specificity = 0.64 (optimum)
Papa Petros[181]	113	F	Not specified	Symptoms of UI + controls	USI/DO/mixed	History	Rapid exercise pad test	Sensitivity and specificity	Sensitivity = 0.90 Specificity = 1.00
Jackson[99]	85	F	Secondary	Symptoms of UI	USI/DO/mixed	History (various measures of leakage)	1-hour pad test	Correlation	r = 0.31–0.67
Mouritsen[112]	72	F	Secondary	Symptoms of UI	USI/DO/mixed	History (three severity groups)	ICS: 1-hour 24 hour 48 hour	Difference in pad gain between groups	Significant differences for all pad tests
Ryhammer[115]	144	F	Secondary	Clinical trial for UI	Any leakage	History (incontinent/ continent)	24-hour pad test	Difference in pad gain between groups	No significant difference

TABLE 17 *Pad test compared with urodynamics*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Jorgensen[134]	49	F	Secondary	Symptoms of UI	Any leakage	Multichannel urodynamics	ICS: 1-hour pad test	Full contingency table	Sensitivity = 0.94 Specificity = 0.44
Versi[135]	105	F	Secondary	LUTS	USI	Multichannel urodynamics	48-hour home pad test	Full contingency table	Sensitivity = 0.92 Specificity = 0.72
Versi[105]	99	F	Mixed	USI (all positive urodynamics)	USI	Multichannel urodynamics	Rapid exercise pad test	Sensitivity	Sensitivity = 0.86
Mayne[182]	33	F	Secondary	USI (all positive urodynamics)	USI	Multichannel urodynamics	Three pad tests: unknown bladder 250 ml 1 hour	Sensitivity	Unknown volume = 0.79 250 ml = 0.91 1 hour = 0.94
Mouritsen[112]	97	F	Secondary	Symptoms of UI + asymptomatic controls (25)	USI/DO/mixed	Multichannel urodynamics	ICS: 1 hour 24 hour 48 hour	Mean values	Significant differences between patients and controls
Siltberg[183]	15	F	Secondary	USI (all positive urodynamics)	USI	Multichannel urodynamics	Rapid exercise pad test	Correlation	$r = 0.59$
Berglund[100]	45	F	Secondary	USI (all positive urodynamics)	USI	2 × 2-hour pad test 1 minimal exercise 1 maximal exercise	Multichannel urodynamics	Sensitivity	Sensitivity = 0.86

TABLE 18 *Validated scale compared with clinical history*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Miller[184]	51	F	Clinical trial for UI	Symptoms of UI (all >60 years)	USI	History (frequency of leakage)	6-day diary (frequency of leakage)	Correlation	$r = 0.33$
Jackson[99]	85	F	Secondary	Symptoms of UI	USI/DO/mixed	History (frequency of leakage)	Diary (unspecified) (frequency of leakage)	Kappa	$\kappa = 0.62$
Elser[19]	265	F	Clinical trial for UI	Symptoms of UI	USI/DO/mixed	History (frequency of leakage)	7-day diary (frequency of leakage)	Correlation	$r = 0.63$

One paper compared multichannel urodynamics with an exercise pad test, with the pad test result taken to be the gold standard.[100] Multichannel urodynamics were reported to be 0.86 sensitive in diagnosing patients with a positive pad-test result.

Urinary diary compared with clinical history

Three studies compared clinical history with a urinary diary for the measurement of frequency of leakage (*Table 18*). None of these papers presented data in a form that enabled sensitivity or specificity to be calculated with either a correlation coefficient or kappa statistic being used. There was a high level of variance between the levels of agreement demonstrated by the three papers. One paper reported a correlation of 0.33,[184] one a correlation of 0.63[19] and one a kappa of 0.62.[99]

Urinary diary compared with urinary diary

Two studies performed a comparison of two different urinary diaries (*Table 19*). One study compared a 7-day diary with different types of instructions: extensive and minimal for different symptoms of incontinence in women with a urodynamic diagnosis.[185] The correlation between the two methods ranged from 0.67 to 0.78.

One study compared the first 3 days of a 7-day diary with the last 4 days in elderly male patients.[186] The correlation between the mean number of incontinent episodes for this period was $r = 0.84$.

Urinary diary compared with urodynamics

Four papers studied the use of a urinary diary compared with urodynamics (*Table 20*). However, the data from three studies were not presented in a form suitable for inclusion in any analysis and attempts to contact the authors were unsuccessful.[108,109,187]

One study compared the 24-hour diary with multichannel urodynamics for the diagnosis of USI, DO and mixed incontinence in female patients.[187] This paper reported significant differences between diagnostic groups for various diary parameters. Mean voided volume showed the highest differentiating power between the three diagnostic groups, but statistically significant differences were also found for total voided volume, mean voided volume, largest single voided volume and smallest single voided volume.

One study compared the use of a 7-day diary with multichannel urodynamics for the diagnosis of USI in women with symptoms of pure stress leakage.[108] Data from patients with a normal urinary diary only were presented and therefore neither sensitivity nor specificity could be calculated. However, out of 555 women with a negative diary, incontinence (USI, DO or mixed incontinence) was confirmed in 81%.

One study investigated the ability of a urinary diary differentially to diagnose USI and DO in a female population with urodynamically confirmed urinary incontinence.[109] Data were not presented in a format that would allow sensitivity or specificity to be calculated. Based on logistic regression analysis, the parameters of a urinary diary that resulted in the best differentiation between USI and DO were frequency of micturition and mean voided volume.

One study aimed to validate the Bladder Instability Discriminant Index (BIDI), a score derived from a 7-day urinary diary for the non-invasive diagnosis of DO.[136] A score was developed based on parameters including weekly averages of diurnal micturition, nocturnal micturition, and mean, lowest and highest daily micturition volume. By using a cut-off point of below –0.554 to identify a positive result when compared with urodynamic diagnosis a sensitivity of 0.88 and specificity of 0.83 were obtained.

Paper towel test compared with clinical history

One study compared a simple paper towel test with patient history of incontinence (*Table 21*). No significant correlation was found between patient perception of amount of leakage and the results of the paper towel test.

Physical examination compared with clinical history

One paper studied the relationship between the pelvic muscle rating scale and patient history (*Table 22*).[188] The scale was found to have a sensitivity of 0.68 and specificity of 0.71.

Physical examination compared with electromyography

Two studies compared the use of a pelvic muscle rating scale for the measurement of pelvic muscle strength compared with surface electromyography (sEMG) measurements (*Table 23*). Although these papers do not deal specifically with the diagnosis of urinary incontinence, pelvic muscle strength is a crucial part of any evaluation of urinary symptoms.

TABLE 19 *Urinary diary compared with urinary diary*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Robinson[185]	278	F	Clinical trial for UI	All positive urodynamics	Any leakage	7-day diary (detailed instructions)	7-day diary (minimal instructions)	Correlation (various symptoms of incontinence)	$r = 0.67–0.78$ (range)
Robb[186]	44	M	Not stated	Symptoms of UI (elderly)	Any leakage	7-day diary (first 3 days)	7-day diary (last 4 days)	Correlation (incontinent episodes per day)	$r = 0.84$

TABLE 20 *Urinary diary compared with urodynamics*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Contreras Ortiz[136]	271	F	Secondary	Symptoms of UI	DO	Multichannel urodynamics	BIDI derived from urinary diary	Full contingency table	Sensitivity = 88% Specificity = 83%
Fink[187]	132	F	Secondary	Symptoms of UI	USI/DO/mixed	Multichannel urodynamics	24-hour FVC (various parameters)	Difference between diagnostic groups	Significant differences found between diagnostic groups
James[108]	555	F	Secondary	Symptoms of UI	USI (all positive urodynamics)	Multichannel urodynamics	7-day diary	Partial contingency table	Agreement between diary and urodynamics = 108/555
Larsson[109]	142	F	Secondary	LUTS	USI/DO	Multichannel urodynamics	48-hour diary	Difference between diagnostic groups	Significant differences found between diagnostic groups

FVC, frequency volume chart.

TABLE 21 *Paper towel test compared with clinical history*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Miller[184]	51	F	Clinical trial for UI	Symptoms of UI (all ≥ 60 years)	USI	History	Paper towel test	Correlation	No statistically significant correlation

TABLE 22 *Physical examination compared with clinical history*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Romanzi[188]	57	F	Clinical trial	Volunteers (incontinent/ continent)	Pelvic muscle strength	History	Pelvic muscle rating scale	Full contingency table	Sensitivity = 0.68 Specificity = 0.71

TABLE 23 *Physical examination compared with sEMG*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Romanzi[188]	57	F	Clinical trial	Volunteers (incontinent/ continent)	Pelvic muscle strength	Surface EMG	Pelvic muscle rating scale	Correlation	r = 0.46–0.57
Brink[189]	208	F	Clinical trial	Symptoms of UI	Pelvic muscle strength	Surface EMG	Pelvic muscle rating scale	Correlation	r = 0.37–0.63

Both papers report a moderate association between the two measures, with correlations of r = 0.46–0.57[188] and r = 0.37–0.63.[189]

Physical examination compared with battery of tests

One paper compared the diagnosis of USI by a battery of tests with that by physical examination, specifically genital prolapse (*Table 24*). A sensitivity of 0.72 and specificity of 0.46 are reported.

Q-tip test compared with urodynamics

Four papers investigated the use of the Q-tip test compared with urodynamics (*Table 25*). Two papers studied the ability of the Q-tip test, measuring straining angle, to diagnose USI compared with multichannel urodynamics.[137,138] Both papers presented data in a form that allowed sensitivity and specificity to be calculated; however, different cut-off points were used to classify a positive result and therefore the data cannot be combined. A cut-off point of 35 degrees or greater resulted in a sensitivity of 0.75 and specificity of 0.58,[137] and a cut-off of 30 degrees or greater in a sensitivity of 0.53 and specificity of 0.53.[138]

A further two studies also compared the Q-tip with multichannel urodynamics.[191,192] These studies did not present data in a form suitable for calculating summary measures of diagnostic accuracy; however, they both report significantly higher mean straining angles in the USI-confirmed group than in asymptomatic controls.

Algorithm compared with urodynamics

Three studies researched the accuracy of algorithm diagnostic tools compared with multichannel urodynamics in elderly women (*Table 26*). One study investigated the Resident Assessment Protocol (RAP), a non-urodynamic algorithm.[193] They report the RAP to have a sensitivity of 0.76 and specificity of 0.97 for the diagnosis of USI, and a sensitivity of 0.76 and specificity of 0.71 for the diagnosis of DO.

Two studies investigated the ability, retrospectively, of an algorithm method to predict diagnosis of USI, DO and mixed incontinence by multichannel urodynamics.[194,195] They reported that treatment based on the algorithm method would have been correct in 85%[194] and 95% of cases.[195]

Battery of tests compared with clinical history

One paper[190] studied the association between diagnosis of USI using a battery of tests compared with a clinical history (*Table 27*). The battery of tests consisted of a physical examination, cystometry and a stress test. A patient's history of their symptoms was found to be 0.52 sensitive and 0.85 specific in predicting diagnosis based on the battery.

Battery of tests compared with urodynamics

Two papers studied the use of a battery of tests compared with multichannel urodynamics (*Table 28*). One study compared a diagnosis based on a Q-tip test, cough test and patients' symptoms with multichannel urodynamics.[196] Good agreement was found between the two methods, with a sensitivity and specificity of 0.94 and 0.84 for the diagnosis of USI or mixed incontinence and 0.71 and 0.96 for the diagnosis of DO.

One study compared the combination of a pad test and patient history for the diagnosis of DO only;[197] this reports a sensitivity of 0.88 compared with diagnosis made by multichannel urodynamics.

Conductance measurement compared with multichannel urodynamics

One paper[198] studied the measurement of distal urethral conductance (DUEC) for the diagnosis of USI compared with multichannel urodynamics (*Table 29*) and reported a sensitivity of 0.64 and specificity of 0.86.

Urodynamics compared with ultrasound

Nine studies compared the use of ultrasound imaging with urodynamic investigations (*Table 30*). Unfortunately, data from two papers were not presented in a form suitable for inclusion in the formal analysis.[98,199] Attempts to contact the authors for further information resulted in one reply with the full, individual patient data requested.[98]

All nine studies included only female patients and all were conducted in a secondary care setting. Two papers investigated the use of translabial colour Doppler ultrasound.[139,141] This was compared as an alternative to fluoroscopy for the detection of urinary leakage during urodynamic investigation for the diagnosis of USI, DO and mixed incontinence.

Two papers studied the use of transrectal ultrasound for the evaluation of the bladder base and urethrovesical junction compared with the ICS-defined diagnosis of USI by urodynamic investigation.[143,144] A urethrovesical junction drop during stress of at least 1 cm was defined as the cut-off for USI.

TABLE 24 *Physical examination compared with battery of tests*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Statistical tests	Main findings
Fischer-Rasmussen[190]	212	F	Secondary	Symptoms of UI	USI	History/pelvic floor examination/cystometry/ stress test	Physical examination (genital prolapse)	Full contingency table	Sensitivity = 0.72 Specificity = 0.46

TABLE 25 *Q-tip test compared with urodynamics*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Bergman[137]	115	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	Q-tip	Full contingency table	Sensitivity = 0.75 Specificity = 0.58
Montz[138]	100	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	Q-tip	Full contingency table	Sensitivity = 0.53 Specificity = 0.53
Karram[191]	63	F	Secondary	Symptoms of UI + controls	USI	Multichannel urodynamics	Q-tip	Difference between diagnostic groups	Mean straining angle significantly higher in USI group
Walters[192]	48	F	Secondary	Symptoms of UI + controls	USI	Multichannel urodynamics	Q-tip	Difference between diagnostic groups	Mean straining angle significantly higher in USI group

TABLE 26 *Algorithm compared with urodynamics*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Resnick[193]	102	F	Secondary	Symptoms of UI	USI/DO	Multichannel urodynamics	Algorithm (RAP)	Sensitivity and specificity	USI: Sensitivity = 0.76 Specificity = 0.71 DO: Sensitivity = 0.76 Specificity = 0.97
Eastwood[194]	65	F	Secondary	Referred for urodynamics	USI/DO/mixed	Multichannel urodynamics	Algorithm	Retrospective comparison between diagnostic pathways	Algorithm would have resulted in correct treatment in 85% of cases
Hilton[195]	100	F	Secondary	Referred for urodynamics	USI/DO/mixed	Multichannel urodynamics	Algorithm	Retrospective comparison between diagnostic pathways	Algorithm would have resulted in correct treatment in 95% of cases

TABLE 27 *Battery of tests compared with clinical history*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Statistical tests	Main findings
Fischer Rasmussen[190]	212	F	Secondary	Symptoms of UI	USI	History/pelvic floor examination/cystometry/ stress test	History	Full contingency table	Sensitivity = 0.52 Specificity = 0.85

TABLE 28 *Battery of tests compared with urodynamics*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Summitt[196]	90	F	Secondary	Symptoms of UI	USI/DO/mixed	Multichannel urodynamics	Q-tip/symptoms/ cough test	Full contingency table	USI/Mixed: Sensitivity = 0.94 Specificity = 0.84 DO: Sensitivity = 0.71 Specificity = 0.96
Griffiths[197]	100	F	Secondary	Symptoms of UI	DO	Multichannel urodynamics	Pad/history	Sensitivity	Sensitivity = 0.88

TABLE 29 *Conductance measurement compared with urodynamics*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Statistical tests	Main findings
Creighton[198]	F	F	Secondary	Symptoms of UI	USI/mixed	Multichannel urodynamics	DUEC	Full contingency table	Sensitivity = 0.64 Specificity = 0.86

TABLE 30 *Urodynamics compared with ultrasound*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Dietz[139]	37	F	Secondary	Referred for urodynamics	USI/DO/mixed	Multichannel urodynamics	Translabial ultrasound (visible leakage)	Full contingency table	Sensitivity = 77% Specificity = 75%
Dietz[140]	117	F	Secondary	Symptoms of UI	USI/DO/mixed	Multichannel urodynamics	Transperineal (opening of bladder neck and mid-urethra)	Full contingency table	Sensitivity = 83% Specificity = 76%
Chen[98]	102	F	Secondary	USI (positive urodynamics) continent controls	USI	Multichannel urodynamics	Perineal (rotational angle and BND)	Full contingency table	Sensitivity = 73% Specificity = 77%
Kiilholma[102]	38	F	Secondary	USI (positive urodynamics)	USI	Multichannel urodynamics	Perineal (BND)	Sensitivity	Sensitivity = 72%
Bergman[143]	91	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	Transrectal (drop of the urethrovesical junction)	Full contingency table	Sensitivity = 86% Specificity = 92%
Bergman[144]	32	F	Secondary	USI controls	USI (DO = control)	Multichannel urodynamics	Transrectal (drop of the urethrovesical junction)	Full contingency table	Sensitivity = 94% Specificity = 89%
Dietz[141]	52	F	Secondary	Referred for urodynamics	USI/DO	Multichannel urodynamics	Translabial (visible leakage)	Full contingency table	Sensitivity = 94% Specificity = 93%
Quinn[142]	124	F	Secondary	Not specified	USI	Multichannel urodynamics	Vaginal (opening of bladder neck/proximal urethral with leakage during cough)	Full contingency table	Sensitivity = 96% Specificity = 82%
Kolbi[199]	32	F	Secondary	USI	USI	Multichannel urodynamics	Perineal (urethrovesical angle)	Difference between diagnostic methods	No significant difference

Three studies compared ultrasound (vaginal[142] and transperineal[98,140]) with fluoroscopy during videourodynamics. The imaging of bladder neck descent (BND) and rotation of the proximal urethra were recorded using both methods. Simple funnelling or opening of the proximal urethra during valsalva was taken to be the measure of USI.

Imaging techniques compared with multichannel urodynamics

When imaging the lower urinary tract during investigation of urinary incontinence two anatomical features are commonly used: observation of leakage from the bladder and descent of the bladder neck. Two methods for directly observing leakage from the bladder are reported: X-ray imaging performed during urodynamics (*Table 31*) and ultrasound (as described in the previous section).

Four studies report the accuracy of observed leakage using ultrasound for the diagnosis of USI compared with multichannel urodynamics (*Figure 7*). The data from these studies were combined to provide a pooled sensitivity of 0.89 (95% CI 0.84 to 0.93) and specificity of 0.82 (95% CI 0.73 to 0.89). The positive likelihood ratio associated with the pooled sensitivity and specificity is 4.94 (95% CI 3.88 to 6.01), and the AUC for the ROC curve corresponding to the pooled DOR is 0.90 (95% CI 0.84 to 0.96) (*Figure 7*). Two studies used X-ray imaging for the detection of leakage;[145,146] when combined, these studies provide a sensitivity of 0.60 (95% CI 0.52 to 0.68) and specificity of 0.74 (95% CI 0.68 to 0.81) for the diagnosis of USI compared with multichannel urodynamics. The positive likelihood ratio associated with the pooled sensitivity and specificity is 2.31 (95% CI 1.62 to 3.00) (*Figure 8*).

Three studies used ultrasound imagining of BND during stress for the diagnosis of USI in women compared with multichannel urodynamics.[98,143,144] The data from these studies were combined to provide a pooled sensitivity of 0.84 (95% CI 0.76 to 0.90) and specificity of 0.86 (95% CI 0.79 to 0.91). The positive likelihood ratio associated with the pooled sensitivity and specificity is 6 (95% CI 4.72 to 7.28) and the AUC for the ROC curve corresponding to the pooled DOR is 0.94 (95% CI 0.84 to 1.00) (*Figure 9*).

Two studies used X-ray to image BND.[147,148] The data from these studies resulted in a pooled sensitivity of 0.79 (95% CI 0.67 to 0.88) and specificity of 0.55 (95% CI 0.43 to 0.66). The positive likelihood ratio associated with the pooled sensitivity and specificity is 1.76 (95% CI 0.90 to 2.61) (*Figure 10*).

Stress test compared with multichannel urodynamics

Six studies were identified that compared the use of a stress test with multichannel urodynamics (*Table 32*).

All of the studies included only female patients and were performed in a secondary care setting. One study included only nursing home residents, meaning that their sample consisted entirely of elderly women.[156]

Each of the six papers dealt with the diagnosis of USI. In all cases a positive stress test was defined as leakage occurring coinciding with cough or valsalva.

Two papers used the supine stress test, one with the bladder filled with 200 ml saline,[149] the other with an empty bladder.[201] Two papers used a standing stress test, both with a full bladder (>200 ml).[156,158]

One paper performed the stress test in both the supine and standing position with a full bladder.[150] One paper performed a cough stress test with the patient sitting in the erect position; however, the diagnosis was also dependent on the result of single-channel urodynamics.[17]

The quality of reporting of the studies in this group was high. All six papers presented full contingency tables. One paper only provided data for patients who were positive on multichannel urodynamics; therefore, for this study only sensitivity could be calculated.[158]

Based on advice from the clinical members of the investigation team, data from three papers were combined to provide a pooled sensitivity of 0.85 (95% CI 0.78 to 0.91) and specificity of 0.83 (95% CI 0.74 to 0.90) for the diagnosis of USI in women using the supine clinical stress test compared with multichannel urodynamics (*Figure 11*). The positive likelihood ratio associated with the pooled sensitivity and specificity is 5.00 (95% CI 3.79 to 6.21) and the AUC for the ROC curve corresponding to the pooled DOR is 0.87 (95% CI 0.69 to 1.00).

Single-channel cystometry compared with multichannel urodynamics

Eight studies were identified that compared the use of single-channel urodynamics with

TABLE 31 *X-ray imaging compared with multichannel urodynamics*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Statistical tests	Main findings
Pelsang[145]	159	F	Secondary	LUTS	USI	Multichannel urodynamics	Bead chain urethrocystography (observed leakage)	Full contingency table	Sensitivity = 0.61 Specificity = 0.70
Grischke[147]	84	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	Bead chain urethrocystography (BND)	Full contingency table	Sensitivity = 0.59 Specificity = 0.60
Bergman[148]	59	F	Secondary	Symptoms of SUI + controls	USI	Multichannel urodynamics	Bead chain urethrocystography (urethrovesical angle)	Full contingency table	Sensitivity = 0.94 Specificity = 0.56 (optimum)
Scotti[146]	204	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	Urethroscopy (opening of the urethrovesical junction)	Full contingency table	Sensitivity = 0.60 Specificity = 0.79
Rose[200]	1584	M	Secondary	LUTS	BOO	Multichannel urodynamics	Micturating cystourethrography (trabeculated bladder)	Full contingency table	Sensitivity = 0.91 Specificity = 0.91

BOO, bladder outlet obstruction.

TABLE 32 *Stress test compared with multichannel urodynamics*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Hsu[149]	41	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	Supine cough stress test	Full contingency table	Sensitivity = 0.94 Specificity = 0.90
Resnick[156]	97	F	Secondary	Symptoms of UI (nursing home)	USI	Multichannel urodynamics	Clinical stress test	Full contingency table	Sensitivity = 0.67 Specificity = 0.98
Lobel[201]	304	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	Empty supine stress test	Full contingency table	Sensitivity = 0.49 Specificity = 0.95
Kadar[150]	37	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	Cough stress test	Full contingency table	Sensitivity = 0.78 Specificity = 0.74
Scotti[17]	145	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	Cough stress test + single-channel urodynamics	Full contingency table	Sensitivity = 0.49 Specificity = 0.95
Swift[158]	108	F	Secondary	LUTS	USI	Multichannel urodynamics	Cough stress test	Sensitivity only	Sensitivity = 0.91

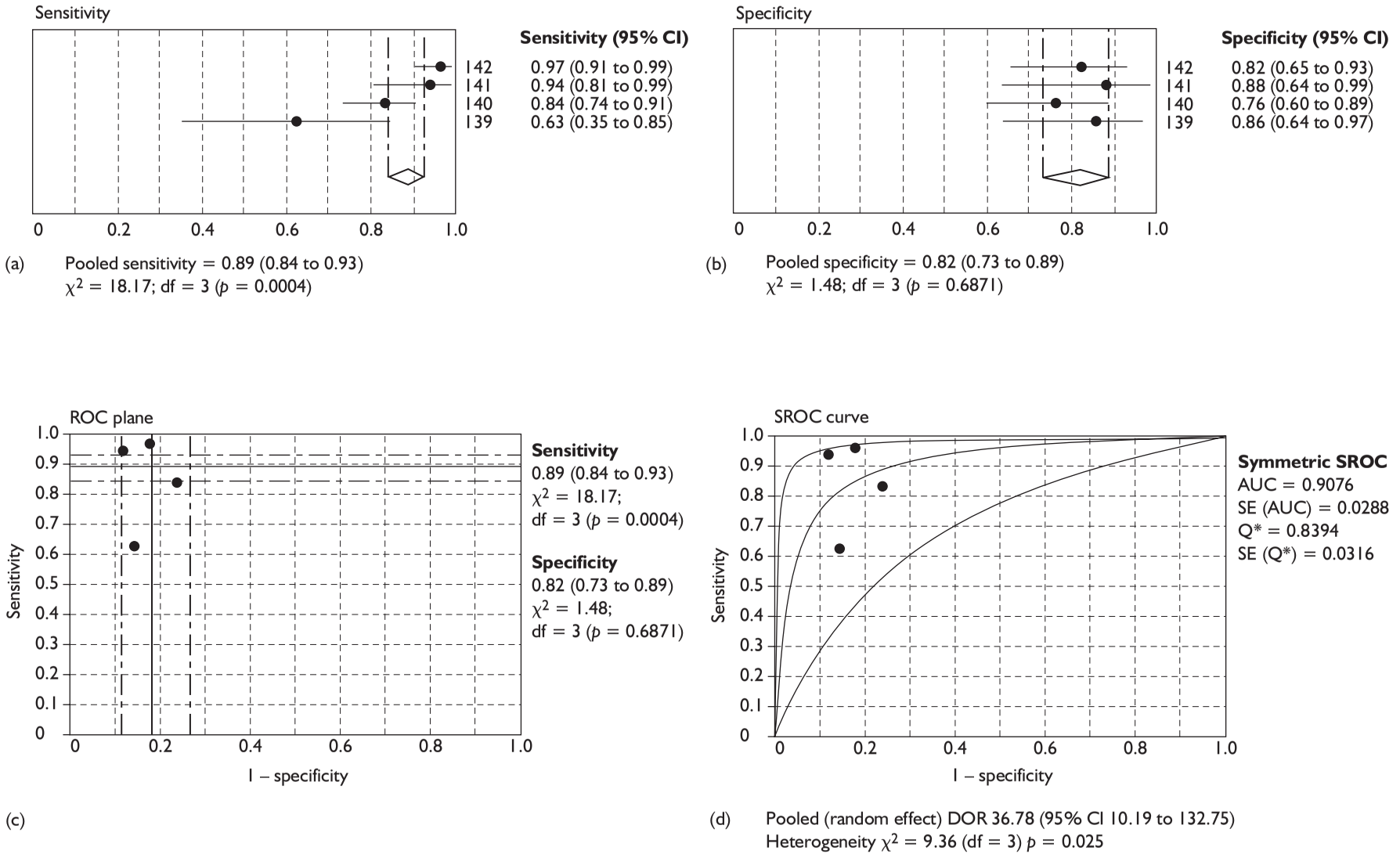


FIGURE 7 *Pooled random effect results: imaging of observed leakage using ultrasound versus MCU for diagnosis of USI in women. (a) Independently pooled sensitivity; (b) independently pooled specificity; (c) sensitivity and specificity for each study and pooled estimates plotted in ROC space; (d) pooled DOR (random effect) plotted in ROC space.*

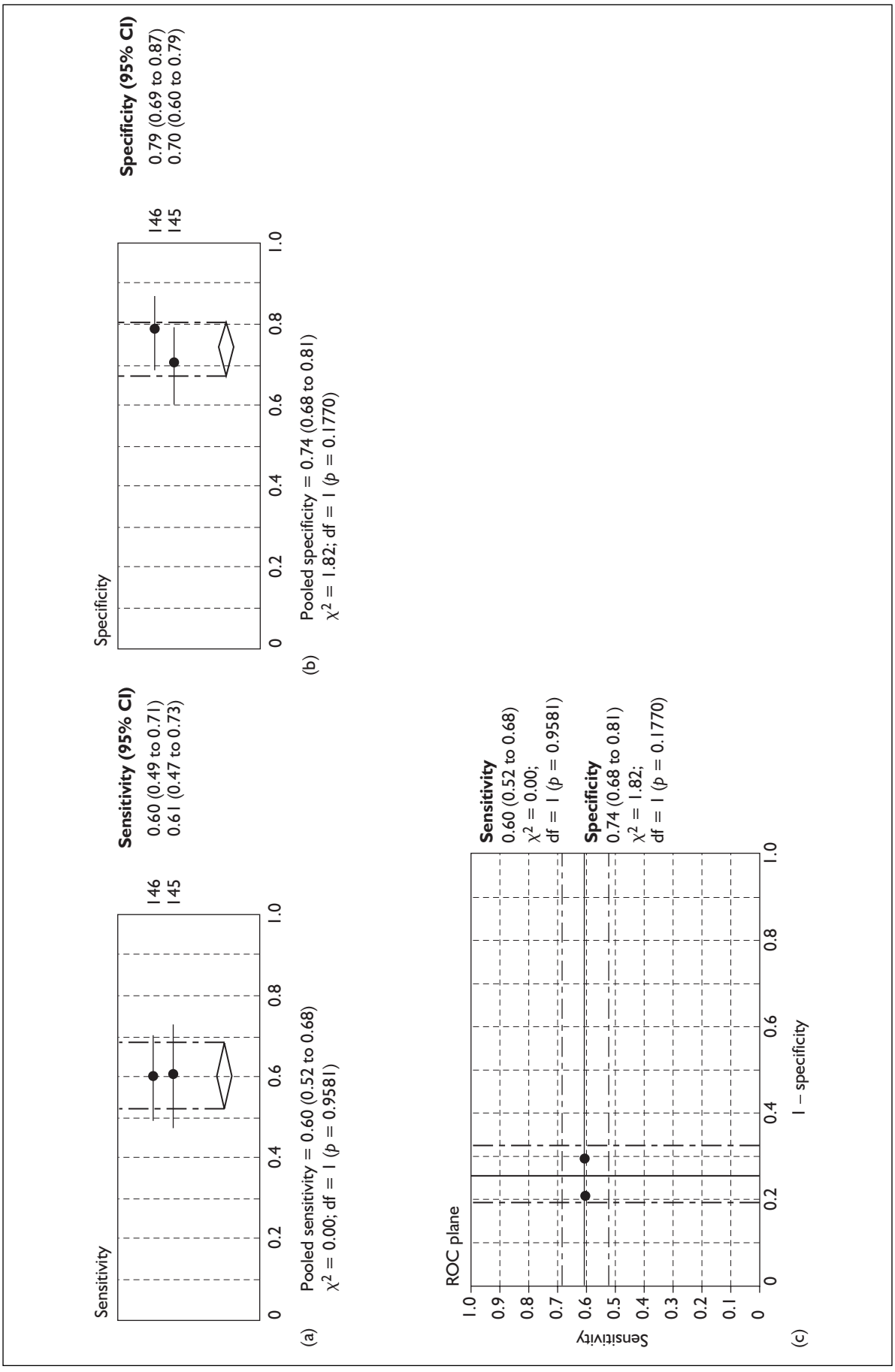


FIGURE 8 *Pooled random effect results: imaging of observed leakage using X-ray versus MCU for diagnosis of USI in women. (a) Independently pooled sensitivity; (b) independently pooled specificity; (c) sensitivity and specificity for each study and pooled estimates plotted in ROC space.*

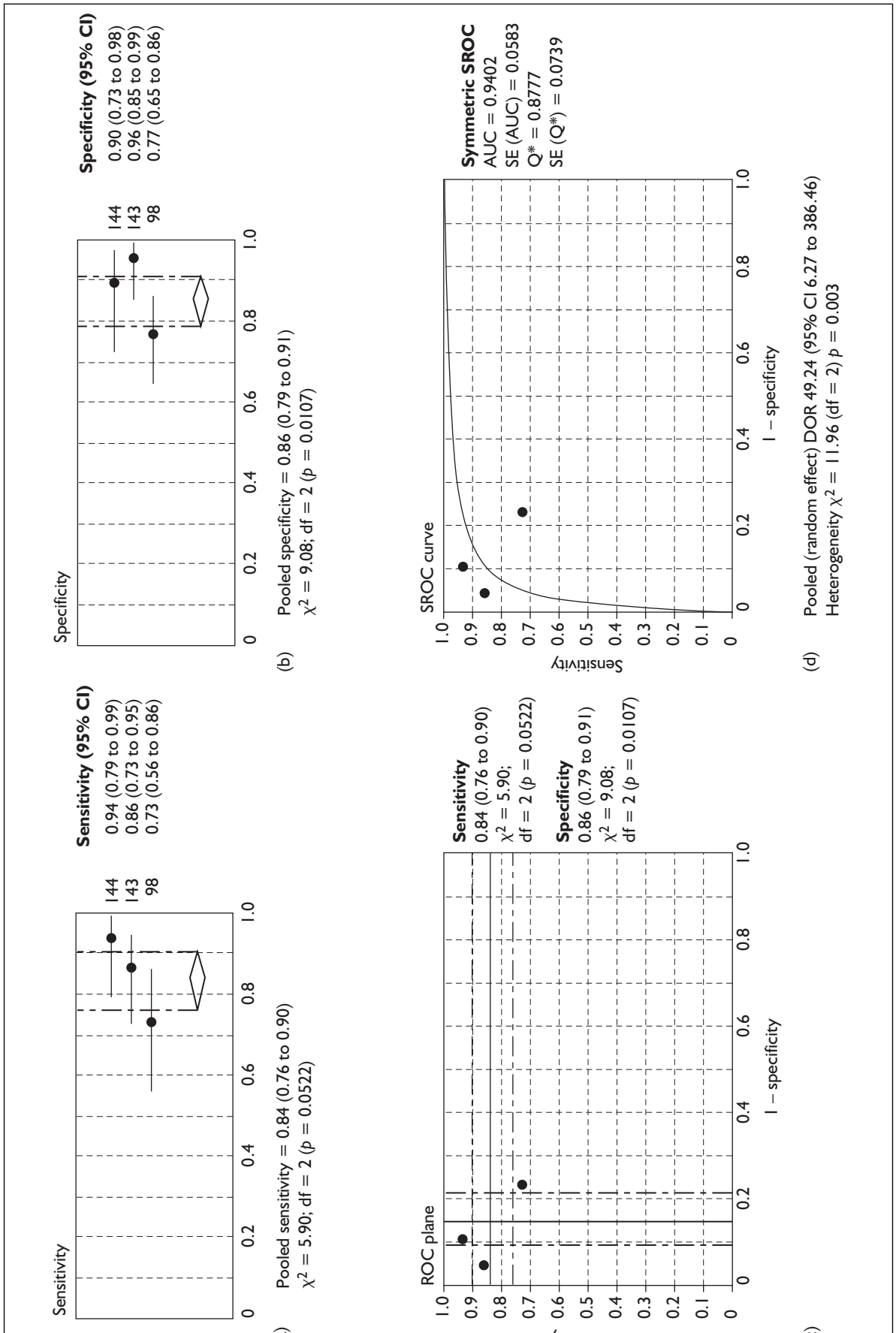


FIGURE 9 *Pooled random effect results: imaging of bladder neck descent using ultrasound versus MCU for diagnosis of USI in women. (a) Independently pooled sensitivity; (b) independently pooled specificity; (c) sensitivity and specificity for each study and pooled estimates plotted in ROC space; (d) pooled DOR (random effect) plotted in ROC space.*

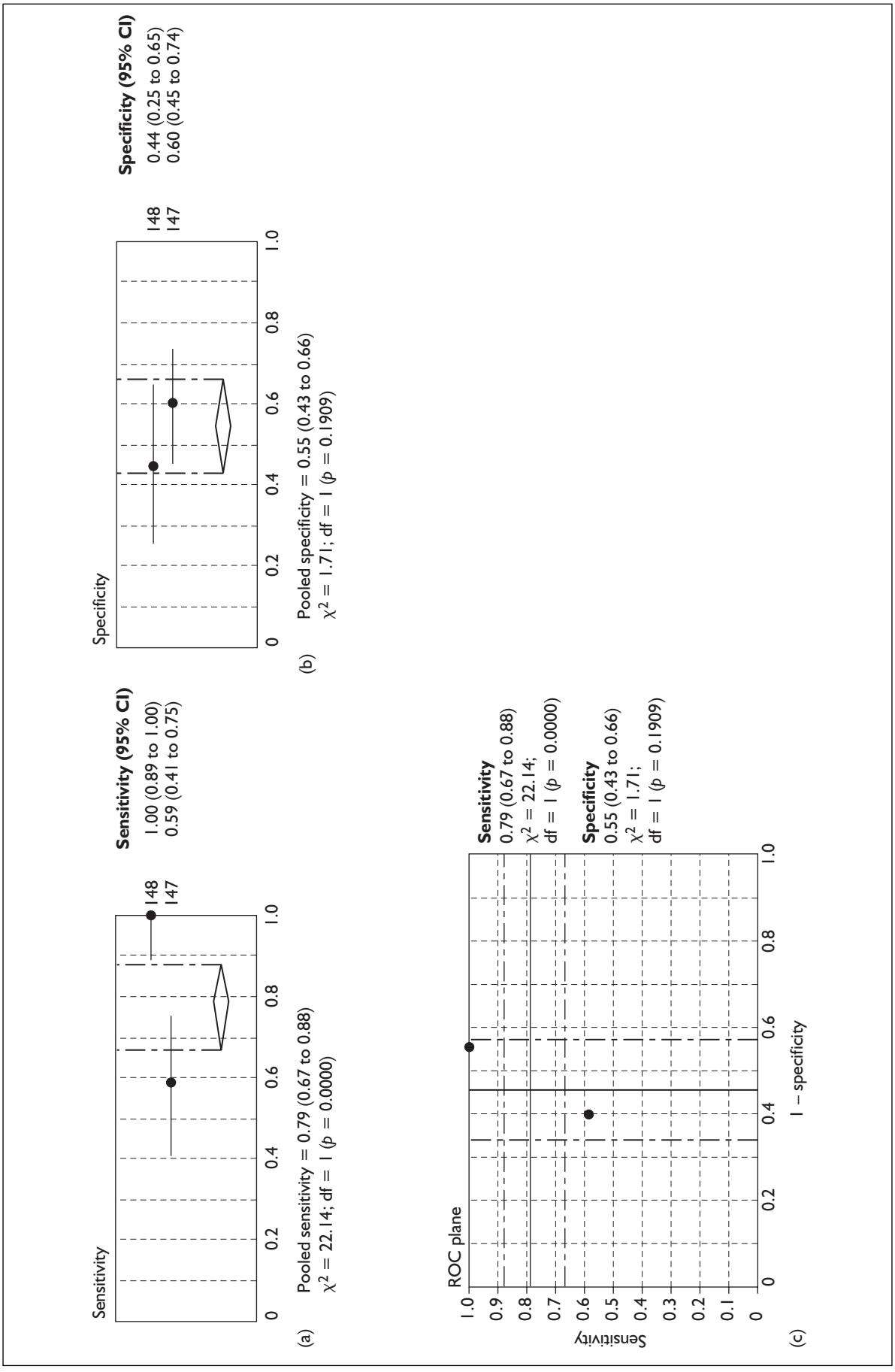


FIGURE 10 *Pooled random effect results: imaging of bladder neck descent using X-ray versus MCU for diagnosis of USI in women. (a) Independently pooled sensitivity; (b) independently pooled specificity; (c) sensitivity and specificity for each study and pooled estimates plotted in ROC space.*

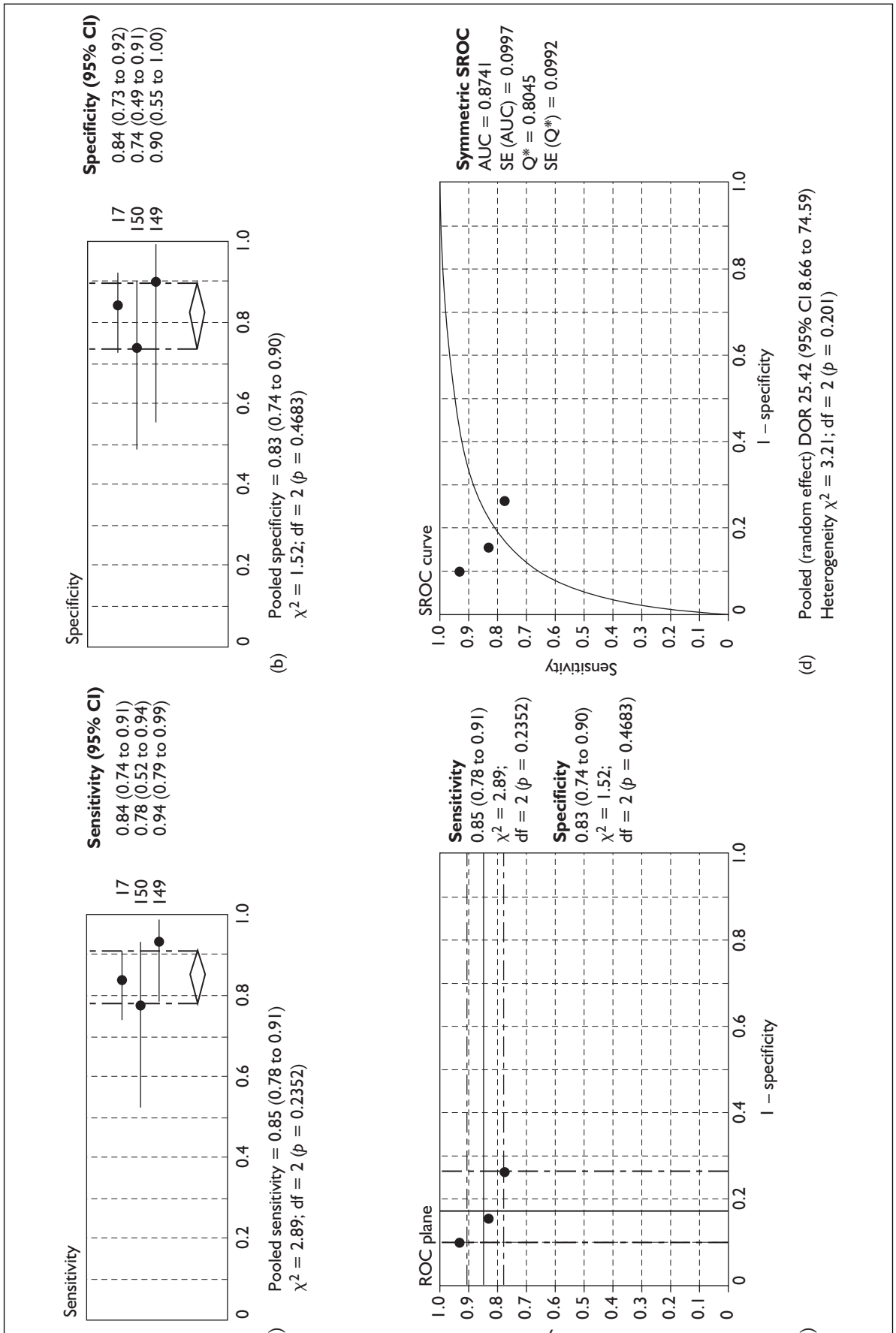


FIGURE 11 *Pooled random effect results: clinical stress test versus MCU for diagnosis of USI in women. (a) Independently pooled sensitivity; (b) independently pooled specificity; (c) sensitivity and specificity for each study and pooled estimates plotted in ROC space; (d) pooled DOR (random effect) plotted in ROC space.*

multichannel urodynamics (*Table 33*). Six of the studies used only female patients,[17,151–153,156,202] whereas two studies used both male and female patients.[154,155] All studies were conducted in a secondary care setting. Three studies investigated elderly populations (older than 70, 60 and 65 years, respectively).[154–156]

Six studies were concerned only with the diagnosis of DO[151–155,202] and two studies with USI.[17,156] The criterion standard used in each of the eight studies was standard multichannel urodynamics. In addition, one study used videoimaging as part of the multichannel urodynamic procedure.[156]

Full contingency tables were provided for all papers, allowing pooling of data. One study used only urodynamically confirmed patients and therefore only sensitivity could be calculated.[202]

After clinical advice, data from two papers were combined to provide a pooled sensitivity of 0.74 (95% CI 0.65 to 0.82) and specificity of 0.77 (95% CI 0.66 to 0.86) for the diagnosis of DO in elderly women using supine single-channel cystometry (*Figure 12*). The positive likelihood ratio associated with the pooled sensitivity and specificity is 12 (95% CI 10.58 to 13.42) and the AUC for the ROC curve corresponding to the pooled DOR is 0.92 (95% CI 0.80 to 1.00). Data from the same two papers were combined to provide a pooled sensitivity of 0.89 (95% CI 0.76 to 0.96) and specificity of 0.92 (95% CI 0.62 to 1.00) for the diagnosis of DO in elderly men using supine single-channel cystometry. The positive likelihood ratio associated with the pooled sensitivity and specificity is 18.2 (95% CI 12.62 to 23.78) (*Figure 13*).

Data from three papers were combined to provide a pooled sensitivity of 0.63 (95% CI 0.55 to 0.71) and specificity of 0.88 (95% CI 0.84 to 0.92) for the diagnosis of DO in women using standing single-channel cystometry (*Figure 14*).

Ambulatory urodynamics compared with multichannel urodynamics

Ambulatory urodynamic monitoring is the monitoring of leakage, flow recordings and pressure in the bladder and abdomen, with or without pressure in the urethra, in an ambulatory setting.[113]

Six studies compared the use of ambulatory urodynamics with standard multichannel urodynamics (*Table 34*). One paper was concerned with the diagnosis of USI in women,[157] one with the diagnosis of BOO in males.[203] The other four were concerned with the diagnosis of DO: one in female patients,[110] one in male patients[204] and two in mixed populations.[103,205]

Owing to the variability in this group it is not possible to combine the data from any of these studies. The sensitivities and specificities demonstrated by these studies are heterogeneous. It is not possible, therefore, to draw any conclusions about the efficacy of ambulatory urodynamics.

There is an issue with ambulatory urodynamics, in that it is thought by some experts to be more sensitive than standard multichannel urodynamics and should be the true gold standard for the diagnosis of urinary incontinence. However, the view of the ICS is that ambulatory urodynamics is overly sensitive but not very specific in detecting urinary leakage. Ambulatory urodynamics has not been standardised by the ICS and therefore cannot be recommended for routine clinical practice. The International Consultation on Incontinence group on urodynamics in 2002 concluded that further study of the place and advantages of ambulatory monitoring was necessary.[9]

Urethral pressure profile compared with multichannel urodynamics

Five studies investigated the use of the urethral pressure profile (UPP) for the diagnosis of USI (*Table 35*). Each study included female patients and was carried out in a secondary care setting.

The data from two studies were combined to provide a pooled sensitivity of 0.62 (95% CI 0.52 to 0.72) and specificity of 0.70 (95% CI 0.61 to 0.77) for the diagnosis of USI in women by UPP in the sitting position (*Figure 15*).

Flow-rate acceleration compared with multichannel urodynamics

One paper compared the use of flow-rate acceleration for the diagnosis of DO with multichannel urodynamics (*Table 36*). Forty female patients with symptoms of urinary incontinence were studied. Flow-rate acceleration was found to be 0.75 sensitive and specific for the diagnosis of DO.

Cystometry by foetal monitoring compared with multichannel urodynamics

Two studies investigated the accuracy of cystometry using the intrauterine pressure channel of a foetal monitor compared with multichannel

TABLE 33 *Single-channel urodynamics compared with multichannel urodynamics*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Scotti[17]	145	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	Single-channel cystometry (sitting) + cough stress test	Full contingency table	Sensitivity = 0.49 Specificity = 0.95
Resnick[156]	97	F	Secondary	Symptoms of UI (nursing home)	USI	Multichannel urodynamics	Single-channel cystometry (supine)	Full contingency table	Sensitivity = 0.86 Specificity = 0.86
Sand[151]	100	F	Secondary	Symptoms of UI	DO	Multichannel urodynamics	Single-channel cystometry (standing)	Full contingency table	Sensitivity = 0.84 Specificity = 0.69
Sand[152]	218	F	Secondary	LUTS	DO	Multichannel urodynamics	Single channel cystometry (supine and standing)	Full contingency table	Supine: Sensitivity = 0.25 Specificity = 0.94 Standing: Sensitivity = 0.59 Specificity = 0.82
Sutherst[153]	100	F	Secondary	Symptoms of UI	DO	Multichannel urodynamics	Single-channel cystometry (supine and standing)	Full contingency table	Sensitivity = 1.00 Specificity = 0.83
Fonda[154]	70	Mixed	Secondary	Symptoms of UI (> 60 years)	DO	Multichannel urodynamics	Single-channel cystometry (supine)	Full contingency table	M: Sensitivity = 0.95 Specificity = 1.00 F: Sensitivity = 0.81 Specificity = 0.68
Ouslander[155]	264	Mixed	Secondary	Symptoms of UI (>65 years)	DO	Multichannel urodynamics	Single-channel cystometry (supine)	Full contingency table	M: Sensitivity = 0.84 Specificity = 0.83 F: Sensitivity = 0.73 Specificity = 0.81
Hebert[202]	47	F	Secondary	DO (positive urodynamics)	DO	Multichannel urodynamics	Single-channel cystometry (supine)	Sensitivity only	Sensitivity = 0.74

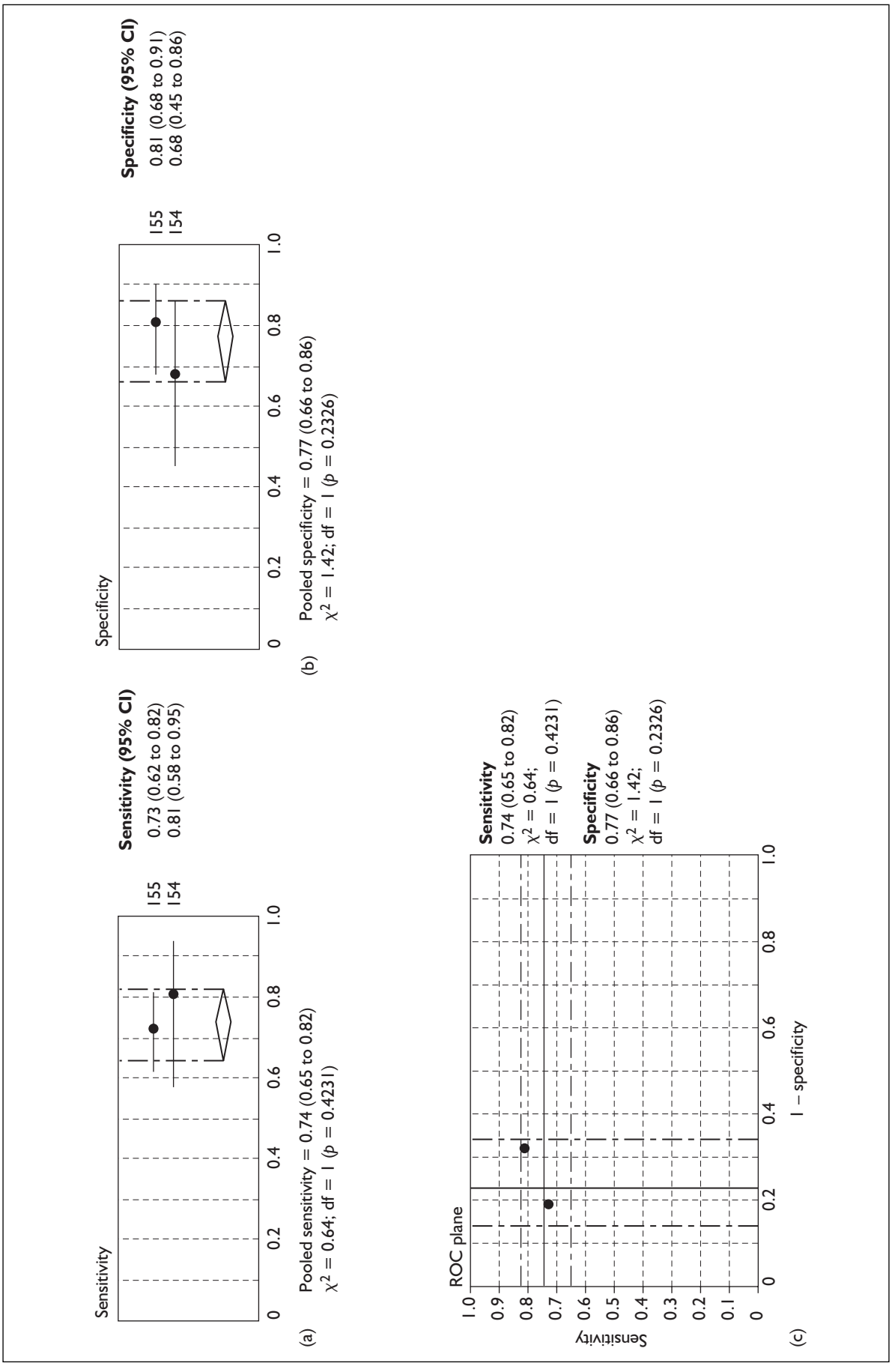


FIGURE 12 *Pooled random effect results: supine single-channel urodynamics (SCU) versus MCU for diagnosis of DO in women over 60. (a) Independently pooled sensitivity; (b) independently pooled specificity; (c) sensitivity and specificity for each study and pooled estimates plotted in ROC space.*

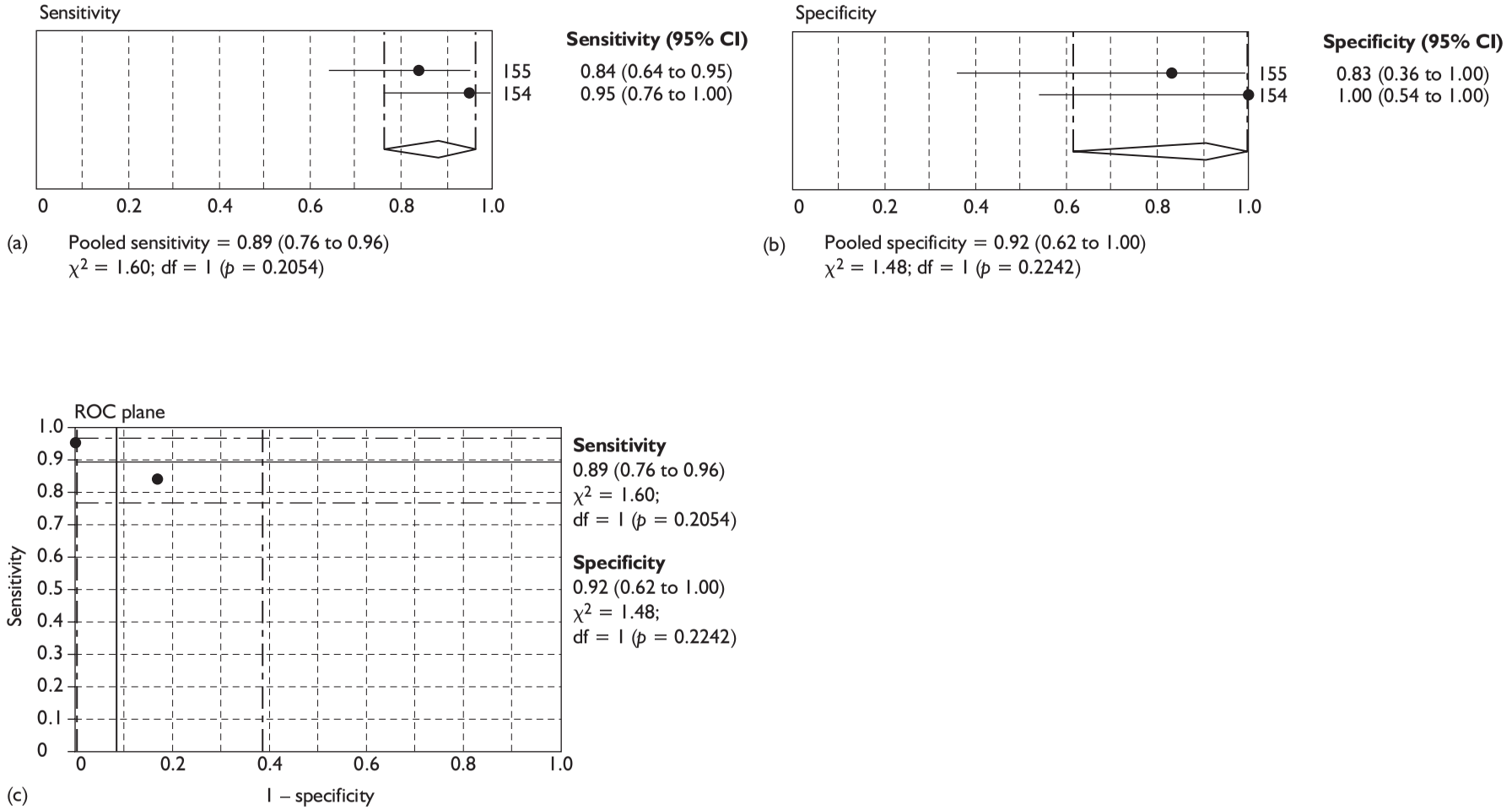


FIGURE 13 *Pooled random effect results: supine SCU versus MCU for diagnosis of DO in men over 60. (a) Independently pooled sensitivity; (b) independently pooled specificity; (c) sensitivity and specificity for each study and pooled estimates plotted in ROC space.*

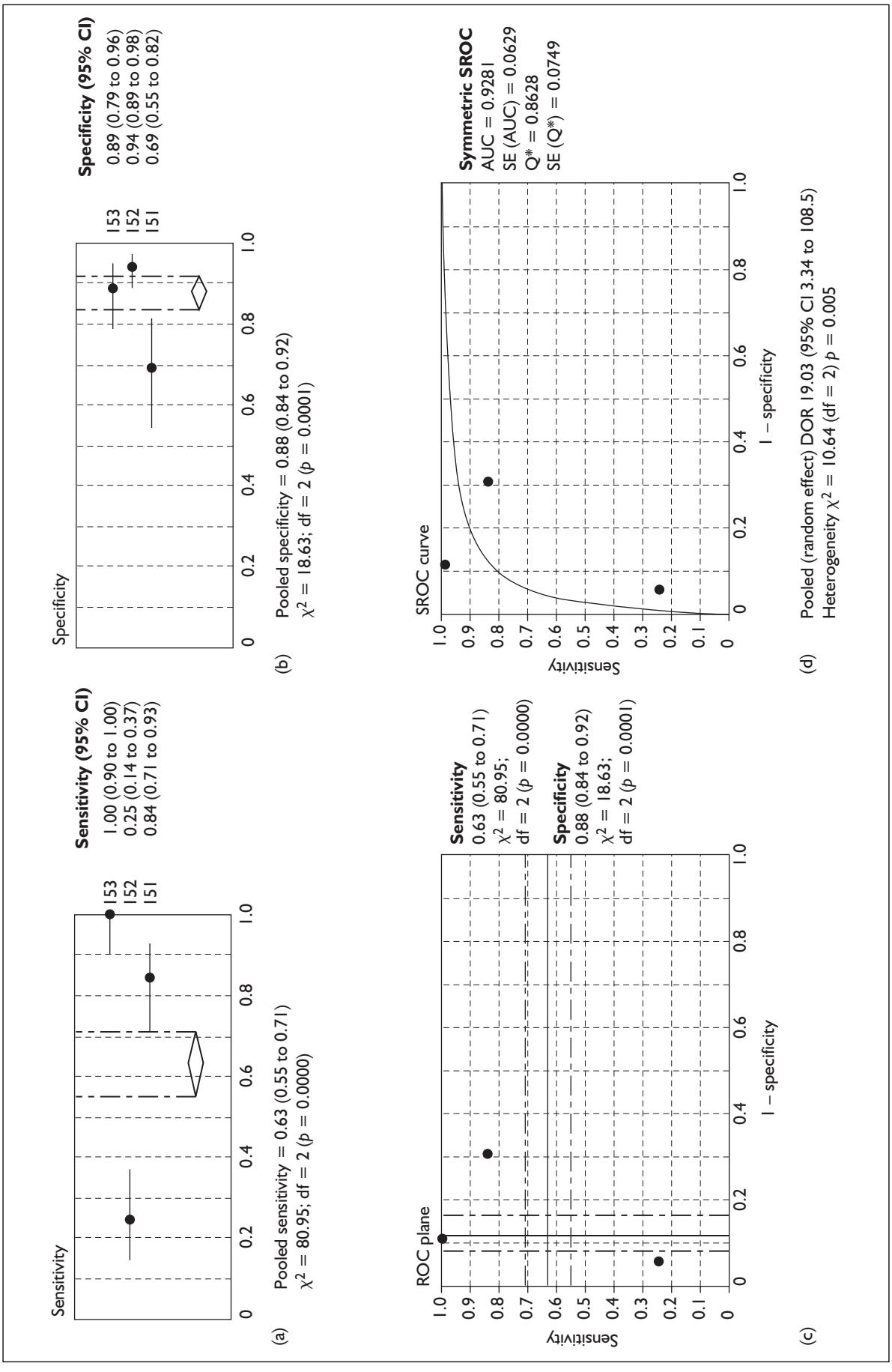


FIGURE 14 *Pooled random effect results: standing SCU versus MCU for diagnosis of DO in women. (a) Independently pooled sensitivity; (b) independently pooled specificity; (c) sensitivity and specificity for each study and pooled estimates plotted in ROC space; (d) pooled DOR (random effect) plotted in ROC space.*

TABLE 34 *Ambulatory urodynamics compared with multichannel urodynamics*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Davis[157]	60	F	Military	Symptoms of UI + controls	USI	Multichannel urodynamics	Ambulatory urodynamics	Full contingency table	Sensitivity = 0.78 Specificity = 0.07
Rosario[203]	63	M	Secondary	Borderline obstruction	BOO	Multichannel urodynamics	Ambulatory urodynamics	Full contingency table	Sensitivity = 0.25 Specificity = 0.88
McInerney[103]	20	Mixed	Secondary	DO symptoms	DO	Multichannel urodynamics	Ambulatory urodynamics	Full contingency table	Sensitivity = 1.00 Specificity = 0.58
Bhatia[204]	26	M	Secondary	LUTS	DO	Multichannel urodynamics	Ambulatory urodynamics	Full contingency table	Sensitivity = 0.43 Specificity = 0.58
Webb[205]	52	Mixed	Secondary	LUTS (all negative urodynamics)	DO	Multichannel urodynamics	Ambulatory urodynamics	Specificity only	Specificity = 0.60
Davila[110]	27	F	Secondary	Symptoms of DO	DO	Multichannel urodynamics	Ambulatory urodynamics	Agreement between two methods	More cases identified by ambulatory urodynamics

TABLE 35 *Urethral pressure profile compared with multichannel urodynamics*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Swift[158]	108	F	Secondary	LUTS	USI	Multichannel urodynamics	UPP (sitting)	Full contingency table	Sensitivity = 0.49 Specificity = 0.98
Richardson[159]	144	F	Secondary	LUTS	USI	Multichannel urodynamics	UPP (sitting)	Full contingency table	Supine: Sensitivity = 0.32 Specificity = 0.93 Standing: Sensitivity = 0.41 Specificity = 0.92
Versi[160]	172	F	Secondary	Symptoms of USI	USI	Multichannel urodynamics	UPP (supine)	Full contingency table	Sensitivity = 0.77 Specificity = 0.79
Pajoncini[206]	119	F	Secondary	USI (positive urodynamics)	USI	Multichannel urodynamics	VLPP MUCP	Sensitivity and specificity only	VLPP: Sensitivity = 0.84 Specificity = 0.60 MUCP: Sensitivity = 0.63 Specificity = 0.52
Versi[207]	303	F	Secondary	Symptoms of UI	USI	Multichannel urodynamics	UPP (supine)	Sensitivity and specificity only	Sensitivity = 0.48 Specificity = 0.84

VLPP, vesical leak point pressure; MUCP, maximum urethral closure pressure.

TABLE 36 *Flow-rate acceleration measurement compared with multichannel urodynamics*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Cucchi[104]	40	F	Secondary	Symptoms of UI	DO	Multichannel urodynamics	Flow-rate acceleration	Individual patient data	Sensitivity = 0.75 Specificity = 0.75

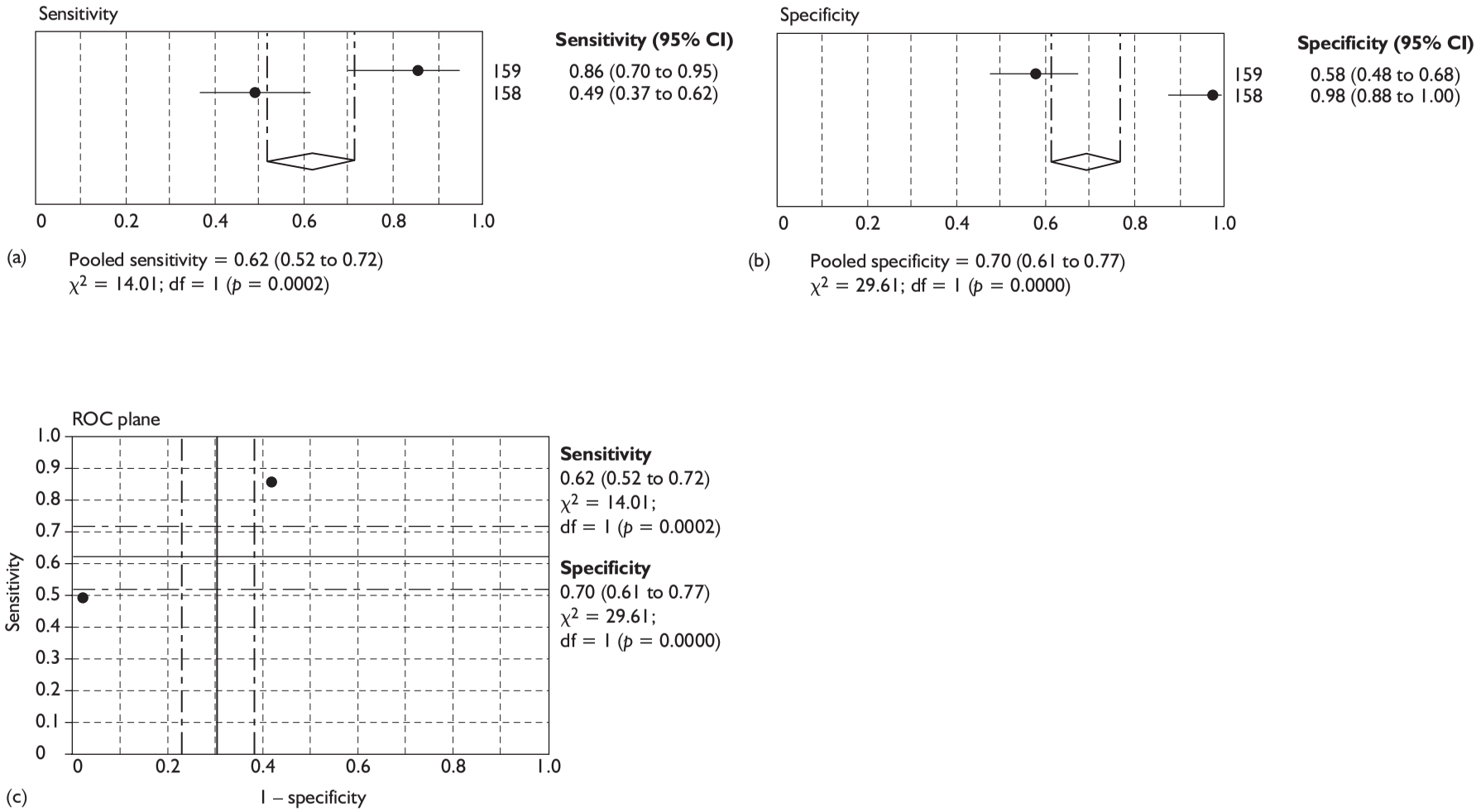


FIGURE 15 *Pooled random effect results: UPP versus MCU for diagnosis of USI in women. (a) Independently pooled sensitivity; (b) independently pooled specificity; (c) sensitivity and specificity for each study and pooled estimates plotted in ROC space.*

urodynamics (*Table 37*). Both studies were concerned with the diagnosis of DO in women in secondary care. Because of the form in which the data were presented in these studies and the homogeneous nature, the results were combined to provide a pooled sensitivity of 0.92 (95% CI 0.76 to 0.98) and specificity of 0.89 (95% CI 0.78 to 0.94).

Ice-water test compared with multichannel urodynamics

One paper studied the use of the ice-water test for the diagnosis of detrusor overactivity, specifically with regard to distinguishing this condition from detrusor hyperflexia (*Table 38*). The ice-water test was found to have a sensitivity of 0.85 and a specificity of 0.65 when diagnosing DO. This study was performed in a very specific population where 82% of the sample had a neurological disease; therefore, the applicability of the results may be restricted.

Fluid-bridge test compared with standard cystometry

One study compared the use of the fluid-bridge test for the diagnosis of USI in women compared with standard cystometry (*Table 39*). A sensitivity of 0.86 and specificity of 0.42 were demonstrated when the test was performed in the supine position, and 1.00 (sensitivity) and 0.24 (specificity) when in the erect position. The fact that there is only one paper studying this test and that this was published in 1981 indicates that this is not a test of great relevance to clinicians.

Urethral closure pressure profile compared with the clinical stress test

One paper studied the ability of a UPP to diagnose USI in women compared with the clinical stress test (*Table 40*). Measurement of UPP was found to have a sensitivity of 0.93 and specificity of 0.83; however, this test was not compared with the recognised gold standard of multichannel urodynamics.

Stop test compared with single-channel cystometry

One study compared the use of the stop test with single-channel cystometry for the diagnosis of DO in women (*Table 41*). This test was found to be 0.95 sensitive and 0.66 specific.

TABLE 37 *Cystometry by foetal monitor compared with multichannel urodynamics*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Data available	Main findings
Swift[208]	66	F	Secondary	Symptoms of UI	DO	Multichannel urodynamics	Cystometry by foetal monitor	Full contingency table	Sensitivity = 0.91 Specificity = 0.86
Bergman[209]	35	F	Secondary	Symptoms of UI	DO	Multichannel urodynamics	Cystometry by foetal monitor	Full contingency table	Sensitivity = 1.00 Specificity = 0.96

TABLE 38 *Ice-water test compared with multichannel urodynamics*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Statistical tests	Main findings
Petersen[210]	130	Mixed	Secondary	Symptoms of UI	DO	Multichannel urodynamics	Ice-water test	Full contingency table	Sensitivity = 0.85 Specificity = 0.65

TABLE 39 *The fluid-bridge test compared with standard cystometry*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Statistical tests	Main findings
Sutherst[211]	127	F	Secondary	Symptoms of UI	USI	Cystometry	Fluid-bridge test (supine and erect)	Full contingency table	Supine: Sensitivity = 0.86 Specificity = 0.42 Erect: Sensitivity = 1.00 Specificity = 0.24

TABLE 40 *Urethral closure pressure profile compared with the clinical stress test*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Statistical tests	Main findings
Hanzal[212]	981	F	Secondary	Symptoms of UI	USI	Clinical stress test	UPP	Full contingency table	Sensitivity = 0.93 Specificity = 0.83

TABLE 41 *Stop test compared with single-channel cystometry*

Reference	*n*	Gender	Care setting	Population	Type of incontinence	Gold standard	Index test	Statistical tests	Main findings
Frigerio[213]	112	F	Secondary	Symptoms of UI	DO	Single-channel cystometry	Stop test	Full contingency table	Sensitivity = 0.95 Specificity = 0.66

Chapter 4
Economic modelling

Introduction

Diagnostic techniques for urinary symptoms, like the majority of healthcare interventions, have a potential to consume healthcare resources. These resources would otherwise be available for alternative forms of healthcare. If the use of a diagnostic test is to be justified then the benefits received need to exceed the costs incurred in carrying out this test. This study aimed to examine the cost-effectiveness of diagnostic techniques for urinary symptoms from a primary care perspective, as this is where most clinical/nursing assessments are undertaken. These tests are likely to have resource implications, as there are costs, such as primary care practitioner time, in carrying them out. In addition, the results of these tests, both positive and negative, are likely to have consequences in terms of other care received.

The framework within which any primary care-based diagnostic test would be used is outlined in *Figure 1*. As can be seen from this diagram, there is no simple relationship where individuals receive diagnostic tests and actions are taken on the basis of the results of these tests. Treatment and testing are linked in this framework as individuals under primary care management may receive treatment from their primary care practitioner and may only be referred to specialist assessment and care if there is no improvement with primary treatment. Ideally, an economic model in this framework would consider all tests and treatments received as a common part of the process of improving health. All resources used would be costed and the outcome measure would be health related, for example quality-adjusted life-years (QALYs). This would enable comparison with a wide range of other healthcare situations. However, a number of problems precluded this approach. These all related to the availability of data and the original remit of the project (which did not consider the results of treatment, only of diagnostic tests). No sufficiently reliable data were found to enable evaluation of the effectiveness of all treatments that could potentially be received by individuals on a common framework. In addition, information was not available on the QALY gains obtained from successful treatment of urinary symptoms or of the QALY changes due to changes in urinary symptoms caused by successful treatment. One possible solution would have been to use expert opinion as to the QALY change caused by successful treatment. However, it was felt this would be unlikely to generate feasible values because of the uncertainty involved. The expert would need an opinion on the type of treatments likely to be carried out in primary care. They would then need to form an opinion of the effectiveness of these treatments in reducing symptoms and the QALY change caused by these symptoms. The final level of uncertainty would be that they are giving a proxy value of the QALY change, that is, what they believe would be the value that an individual would put on a change in their urinary symptoms. Because of these factors it was not felt that this approach would be appropriate or credible. Finally, there were insufficient data to estimate the proportions of individuals who would have any particular test or treatment and who would be referred to and from primary care and GP/specialist care. For these reasons this type of model was always considered outside the scope of the current project.

Therefore, a limited approach to the economic evaluation was taken. A cost-effectiveness study was conducted where the measure of effectiveness was limited to how well the test detected any of the underlying urinary conditions that an individual may have. It was also assumed that positives from these diagnostic techniques could then be referred to secondary specialist assessment. By this means, an attempt was made to isolate the diagnosis of urinary conditions from the rest of the treatment pathway. This enabled judgements to be made about the accuracy and cost-effectiveness of different diagnostic techniques in diagnosing urinary conditions.

Methods

Population groups considered

Although these tests can be used in the diagnosis of USI and DO in men and women, the evidence from the systematic review related to their use in women. Therefore, the models constructed were specific for women and not men. An inclusion criterion for the review was adults only; in

addition, studies were excluded if they studied a purely elderly population.

Alternative diagnostic test strategies

Four alternative diagnostic test strategies on which some data were available were considered. These were history-taking, history and a 48-hour pad test, history and validated scales, and history and urinary diary. As all individuals were assumed to have a history taken, the additional costs and accuracy of the 48-hour pad test, validated scale and urinary diary compared with history alone were evaluated. Evidence from the systematic review showed that history could be used to diagnose both USI and DO. There was also evidence that both a 48-hour pad test and validated scales could be used to diagnose USI. As there was no evidence on the effectiveness of these tests in detecting DO, the assumption was made that they would only be used to diagnose USI. Similarly, there was evidence that a urinary diary was useful for diagnosing DO, but no evidence regarding its use for diagnosing USI. Therefore, the urinary diary was only considered as a diagnostic tool for DO.

The economic model

The model of the cost-effectiveness of primary care diagnostic tests is set in the context of primary care management, where much of the assessment, diagnosis and treatment of urinary conditions is undertaken. The viewpoint of the model is from the perspective of a healthcare provider. The diagnostic strategies evaluated are outlined in *Figure 16*. Diagnosis can be made by the primary care practitioner using history only. In addition to this, other diagnostic strategies are available. These include any of the following: 48-hour pad test, validated scales and urinary diary. For each of these strategies the model has the structure outlined in *Figure 17*. The individual has a true condition, which is unknown by the primary care practitioner who carries out the diagnostic tests; in this model the condition may be USI, DO or both (here referred to as mixed). In addition, an individual may have neither of these conditions. In all cases the model structure is the same; only the probabilities of entering any branch, and the payoffs at the end nodes will change. Therefore, only the model structure if the individuals true condition is USI is shown in *Figure 17*. Regardless of an individual's true condition, primary care tests can declare they have USI, DO, mixed or no condition. If an individual has any of these diagnoses they may then be referred for a specialist secondary assessment.

Parameters used in the model

Cost variables

The cost variables used are given in *Table 42*. In addition to the mean values, the distribution parameters assigned to each cost variable are given; these represent the uncertainty involved in estimation for each parameter. For cost variables log-normal distributions were used as these only return positive values. Furthermore, these distributions are often used for cost data as they have a skewed distribution; this reflects the fact that cost data often have a positive skewed distribution, with a small number of high cost estimates giving distributions a long tail.[214] *Table 42* also provides details of the derivation of each parameter. The cost of carrying out pad tests, validated scales and diaries included consumables costs and any extra time required from the practitioner. This information was obtained from two experts in providing these forms of nursing services. The experts were asked to provide lists of all consumables required to carry out tests. They were also asked for estimates of any extra time that would be required to perform tests. Cost estimates for the tests were

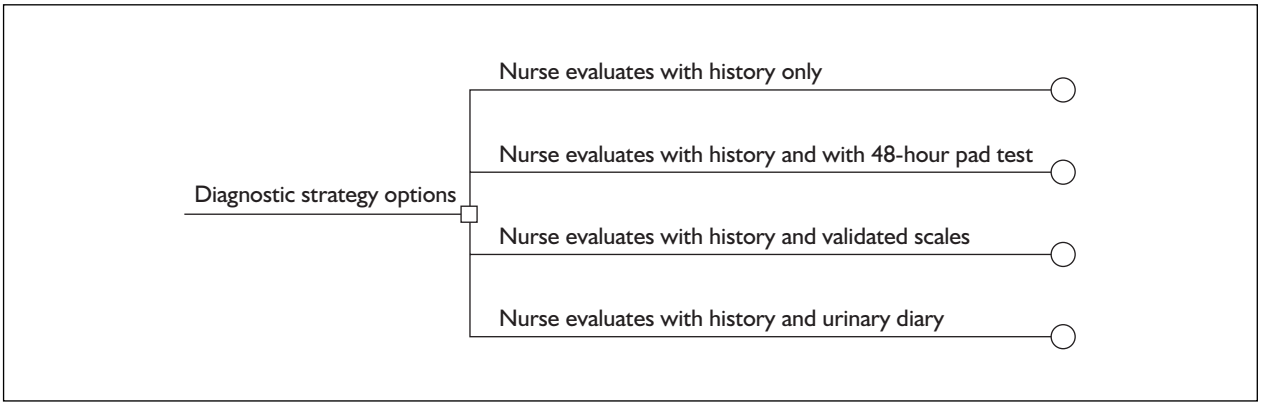


FIGURE 16 *Treatment options considered*

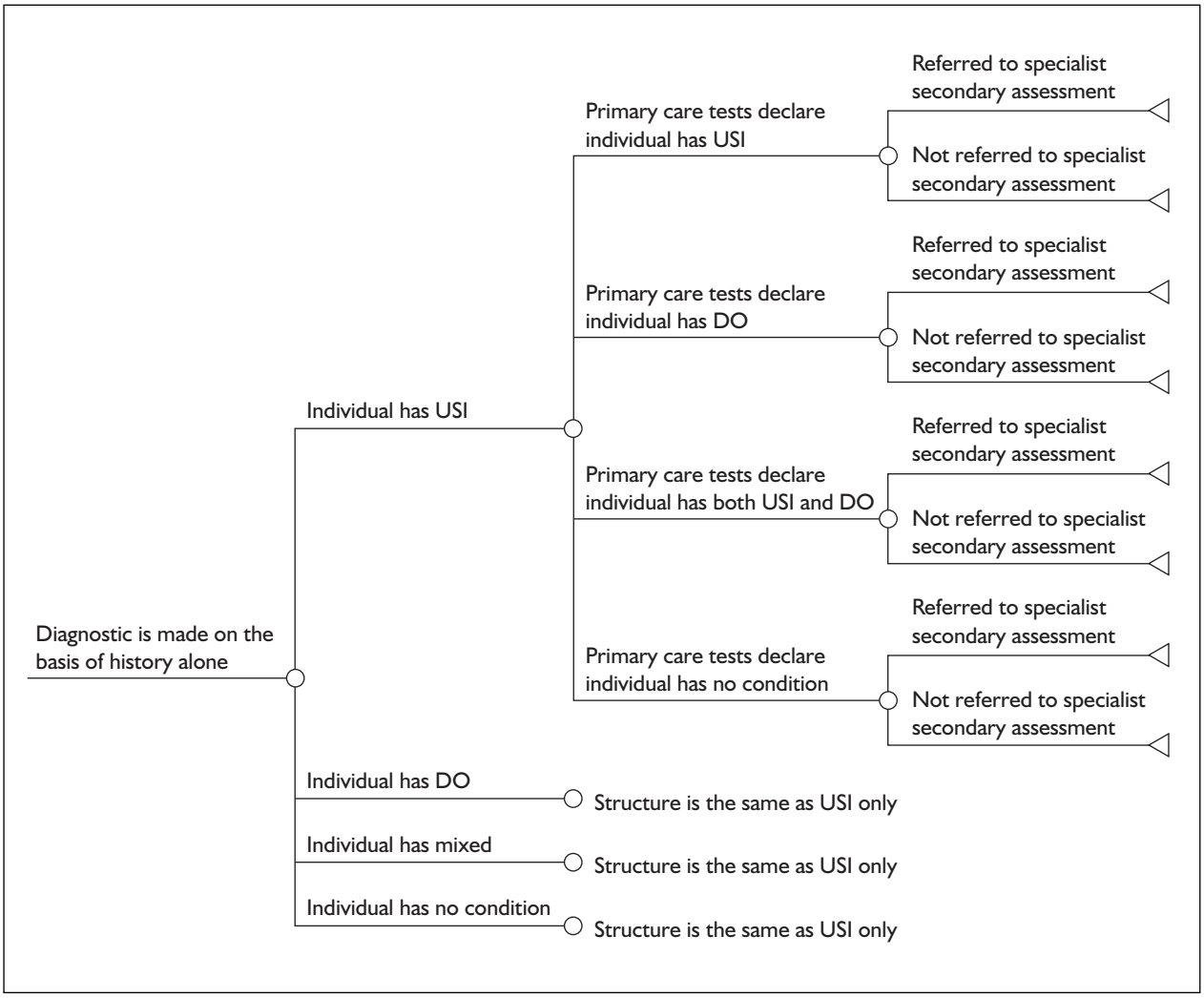


FIGURE 17 *Model structure when history only is used as a diagnostic tool*

then derived from these estimates; further information is given in *Table 42*. All cost variables are in UK pounds for 2002. For all strategies an individual is assumed to have a nurse consultation and the time taken to take a history and evaluate it is included as part of the duration of this consultation.

Outcome variables

The aim was to compare how well primary care tests performed in detecting the underlying conditions causing urinary symptoms. The measure of effectiveness was therefore the number of individuals who had at least one of their conditions successfully detected by a primary care test. The outcomes considered are outlined in *Table 43*.

Prevalence

The measure of the prevalence of urinary conditions was taken from an investigation of the relationship between symptoms reported in a urinary questionnaire and urodynamic diagnosis by Matharu and colleagues.[13] In this study, individuals who reported symptoms in a postal questionnaire were invited to attend a randomised clinical trial comparing a nurse-led continence service with GP management. At the end of this trial individuals who had not responded to treatment were invited to attend urodynamics. The individuals who had urodynamics were therefore either the more severe cases or those whose condition was least responsive to treatment. This may mean that the numbers reported by Matharu are not representative of the proportions of individuals with each condition that would be found in a primary care setting. However, urodynamics would not be routinely used on this type of population, so these data are unavailable. As Matharu and colleagues considered individuals who were appropriate for primary care treatment, these were felt to be the best available data. The following prevalences were reported: USI 0.336

TABLE 42 *Cost variables used in analysis (2002 UK pounds sterling)*

	Mean cost (SE)	Distribution	Derivation
Cost of validated scales	3.75 (0.34)	Log normal	Time taken from expert opinion. Cost of time taken from midpoint of grade F salary scale with on-costs[a]. Distribution taken from assumption that high and low point of estimates given approximate to 95% CI
Cost of diary	3.75 (0.34)	Log normal	Time taken from expert opinion. Cost of time taken from midpoint of grade F salary scale with on-costs[a]. Distribution taken from assumption that high and low point of estimates given approximate to 95% CI
Costs of nurse consultation	18.17 (0.09)	Log normal	Average length of time taken from a database obtained from a locally conducted trial of a continence nurse practitioner-led service.[116] Nurse pay was taken as midpoint of grade F[a]. Overhead rate of 37% applied
Costs of pad test	4.06 (0.56)	Log normal	Time taken from expert opinion. Cost of time taken from midpoint of grade F salary scale with on-costs[a]. Distribution taken from assumption that high and low point of estimates given approximate to 95% CI. Cost of consumables obtained from local service providers
Cost of first referral to urology department	56.22 (6.89)	Log normal	Obtained from NHS reference cost[215]
Cost of urodynamics	125.10 (16.71)	Log normal	Obtained from NHS reference cost.[215] This value used for sensitivity analysis

[a] Costs of nursing time are increased to take into account face-to-face contact only. This involved calculating the proportion of all time that involved face-to-face contact by means of nurse-completed diaries. This information was then used to generate a multiplier and the average cost per minute for all nurse time was increased by this multiplier.

TABLE 43 *Payoffs from diagnosis*

Underlying condition	Diagnosis from primary care tests (history and any additional tests carried out)	Outcome
USI	UI	1
	DO	0
	Mixed	1
	None	0
DO	USI	0
	DO	1
	Mixed	1
	None	0
Mixed	USI	1
	DO	1
	Mixed	1
	None	0
None	USI	0
	DO	0
	Mixed	0
	None[a]	0

[a] In the case of no condition, where the diagnosis of no condition is made, a payoff of zero is recorded even though the correct diagnosis is made. This is because the measure of effectiveness is individuals with any condition correctly diagnosed. An individual cannot have a condition correctly diagnosed if they have no condition to be diagnosed.

TABLE 44 *Performance of primary care tests*

Variable	Mean value (95% CI)	Distribution (α and β parameters)	Derivation
Sensitivity of history for USI	0.92 (0.91 to 0.93)	β (2600 and 226)	Systematic review
Specificity of history for USI	0.56 (0.53 to 0.60)	β (432 and 340)	Systematic review
Sensitivity of history for DO	0.61 (0.57 to 0.65)	β (348 and 222)	Systematic review
Specificity of history for DO	0.87 (0.85 to 0.89)	β (944 and 141)	Systematic review
Sensitivity of pad test for USI	0.92 (0.82 to 0.97)	β (45.3 and 3.9)	Systematic review
Specificity of pad test for USI	0.72 (0.57 to 0.83)	β (32.3 and 12.6)	Systematic review
Sensitivity of diary for DO	0.88 (0.71 to 0.96)	β (22.0 and 3.0)	Systematic review
Specificity of diary for DO	0.83 (0.77 to 0.87)	β (179.1 and 36.7)	Systematic review
Sensitivity of scales for USI	0.87 (0.82 to 0.92)	β (150 and 22.5)	Systematic review
Specificity of scales for USI	0.6 (0.51 to 0.69)	β (68 and 45.1)	Systematic review

(95% CI 0.294 to 0.378), DO 0.291 (95% CI 0.251 to 0.331), mixed 0.207 (95% CI 0.171 to 0.243) and no condition in 0.166 (95% CI 0.133 to 0.199).[13] Mean values were reported in Matharu[13] and the confidence intervals were obtained from one of the authors (Matthews R, University of Leicester: personal communication, 2003). For each sample the probability of these four parameters had to add up to 1, as they were mutually exclusive events, one of which always had to occur. For this reason the four probabilities were varied randomly around their means (using beta-distributions as these distributions are bounded between 0 and 1), but the sum of these four variables was re-based always to equal 1.

Effectiveness of primary care diagnostic tests

The estimates of performance of primary care tests in detecting USI and DO were obtained from the current systematic review. These are detailed in *Table 44*. Again, these variables were assumed to have beta-distributions. Where two tests are used together, for example history and pad test, the results are assumed to be independent, that is, the probability of 48-hour pad tests correctly diagnosing an individual is unrelated to the probability that history detected this condition.

Interpretation of test results

Where more than one test is used there may be situations where different tests give contradictory results. For example, history may declare an individual positive for USI while an individual tests negative using a 48-hour pad test. For these situations a decision rule was needed as to the results of the combination of the two tests. In this analysis the assumption was that if either test was positive then that individual was considered to have tested positive for that condition. For example, if an individual tested positive for USI using one test but negative using another they would still be considered positive for USI. If they tested positive for USI using one test and positive for DO on a different test they would be considered as positive for mixed.

Referral to specialist care

One important consequence of diagnostic tests in primary care is likely to be referrals to specialist secondary care assessment. The authors had no access to data that indicated the proportions of individuals in primary care services who would have referrals after positive results from a primary care test. Estimating these proportions by means of expert opinion proved problematic, as referral to specialist secondary assessment would depend on individual characteristics in each case. Two specialists in clinical care were asked their opinion on the proportions referred to secondary care. However, neither expert felt able to give referral rates because they felt that this would be so dependent on individual circumstances. There is a further complication here as the ability for primary care practitioners to refer individuals to secondary care may be influenced by supply constraints in secondary care, for example a limited capacity to carry out urodynamics. Therefore, what constitutes sufficient grounds to refer may differ from one location to another depending on local capacity.

Because of these factors no data were available that indicated the probability of an individual being referred to specialist secondary assessment given different primary care diagnostic test results. For this reason analyses were carried out using two extreme cases. In the first case none of the individuals with a positive diagnosis for USI or DO or both would be referred to specialist

secondary assessment; in the second case all individuals with a positive diagnosis for any of these conditions would be referred. A midpoint analysis was also evaluated; in this case 50% of all positive diagnoses would be referred to specialist secondary assessment. It was assumed that individuals who tested negative on all tests used would not be referred to specialist secondary assessment.

Model evaluation

The primary analysis was carried out using second order Monte Carlo simulation. This approach assigns a distribution to model parameters. Random values from those distributions are taken for each sample of the Monte Carlo simulation and a cost-effectiveness result is generated based on these values. The model was evaluated using 10,000 samples for each simulation. The model was constructed and evaluated using Microsoft Excel. This probabilistic analysis allowed confidence intervals around costs and effects to be generated. As history should be taken in all cases the research question was the additional costs and effects of further tests in addition to history. Therefore, an incremental analysis was calculated. This gives the extra costs generated by strategies involving additional tests compared with the costs of history-taking alone. The extra proportion of individuals who have any of their conditions correctly diagnosed was also calculated. Finally, the extra costs were calculated per extra individual with any condition correctly diagnosed for strategies involving history and another test compared with history alone. The results of these analyses are presented in cost-effectiveness acceptability curves that track the changing percentage of samples that are cost-effective given different values for detecting cases of urinary disorder.

Probabilistic sensitivity analysis

In addition to the above analysis, an additional probabilistic analysis was performed where it was assumed that individuals who were referred to specialist secondary assessment would also receive urodynamics. This would have two effects. First, it would increase the costs associated with individual diagnoses; and second, referral to urodynamics would also result in more cases being correctly diagnosed, because urodynamics is assumed to be a reference standard. It was therefore assumed that any individual referred to urodynamics, even on the basis of an incorrect diagnosis in primary care, would then be correctly diagnosed.

Deterministic sensitivity analysis

In addition to the probabilistic sensitivity analysis, a series of analyses are presented that involve single parameters being varied through a range of values to estimate the effect that different parameters have on the results of the model. These analyses are evaluated deterministically, that is, only mean values are used to parameterise the model. These will be referred to here as one-way sensitivity analyses.

Results

The probability of individuals being detected as having each diagnosis and their underlying condition is given in *Tables 45–48*. As can be seen from *Table 45*, history accurately detects USI, with 80% of individuals correctly diagnosed as having USI. It performs much less effectively in detecting DO and in identifying individuals who are mixed compared with USI. History also diagnoses correctly only about 50% of individuals who have no condition. It can be seen from *Tables 46* and *48* that history, in combination with pad tests or validated scales, performs better than history alone in terms of diagnosing USI; less than 1% of individuals with USI are diagnosed as having DO or no condition. However, because there are two tests working in combination fewer individuals with no condition are now diagnosed as such. The performance of diary in addition to history can be seen in *Table 47*. This performs less well in terms of USI diagnosed. However, it performs much better in diagnosing DO, with 95% of individuals diagnosed as either DO or mixed. Again, fewer individuals with no condition are diagnosed as such. In general, using additional tests generates more positive and less negative results (using a decision rule of a positive from any test being taken as a positive diagnosis).

As stated earlier, the outcome used in this analysis was cost per individual who has at least one condition correctly diagnosed. The costs and units of effectiveness from the probabilistic model are given in *Table 49*. In all cases the values given are incremental compared with history, that is, they are the extra costs and extra units of effectiveness generated by carrying out history and an additional test when compared with history alone. The results are presented in this way as it was assumed that history would always be performed. *Table 49* gives the result of the two extreme analyses where 0% and 100% of individuals who have a positive diagnosis are referred to secondary specialist assessment. *Table 49* also presents a midpoint analysis where 50% of all individuals declared positive are referred to specialist secondary assessment. *Table 49* shows that all costs

TABLE 45 *Results of history*

Condition and diagnosis	Total for all individuals (95% percentile)	Mean value as a percentage of the total for each condition
Individual has USI, history declares USI	0.269 (0.24 to 0.298)	80.0%
Individual has USI, history declares DO	0.003 (0.003 to 0.004)	1.0%
Individual has USI, history declares mixed	0.04 (0.033 to 0.048)	12.0%
Individual has USI, history declares no condition	0.023 (0.02 to 0.027)	7.0%
Total USI	0.336 (0.301 to 0.371)	100.0%
Individual has DO, history declares USI	0.05 (0.042 to 0.059)	17.2%
Individual has DO, history declares DO	0.099 (0.085 to 0.115)	34.2%
Individual has DO, history declares mixed	0.078 (0.066 to 0.091)	26.8%
Individual has DO, history declares no condition	0.064 (0.053 to 0.075)	21.8%
Total DO	0.291 (0.257 to 0.325)	100.0%
Individual has mixed, history declares USI	0.074 (0.061 to 0.089)	35.9%
Individual has mixed, history declares DO	0.01 (0.008 to 0.012)	4.9%
Individual has mixed, history declares mixed	0.116 (0.097 to 0.137)	56.1%
Individual has mixed, history declares no condition	0.006 (0.005 to 0.008)	3.1%
Total mixed	0.207 (0.175 to 0.240)	100.0%
Individual has no condition, history declares USI	0.064 (0.052 to 0.077)	38.3%
Individual has no condition, history declares DO	0.012 (0.009 to 0.015)	7.3%
Individual has no condition, history declares mixed	0.009 (0.007 to 0.012)	5.7%
Individual has no condition, history declares no condition	0.081 (0.066 to 0.097)	48.7%
Total for no condition	0.166 (0.137 to 0.197)	100.0%

TABLE 46 *Results of history and pad test*

Condition and diagnosis	Total for all individuals (95% percentile)	Mean value as a percentage of the total for each condition
Individual has USI, combination of tests declares USI	0.290 (0.26 to 0.322)	86.4%
Individual has USI, combination of tests declares DO	0.0003 (0.0001 to 0.0006)	0.1%
Individual has USI, combination of tests declares mixed	0.043 (0.036 to 0.052)	12.9%
Individual has USI, combination of tests declares no condition	0.002 (0.001 to 0.004)	0.6%
Total USI	0.336 (0.301 to 0.371)	100.0%
Individual has DO, combination of tests declares USI	0.068 (0.055 to 0.082)	23.3%
Individual has DO, combination of tests declares DO	0.072 (0.055 to 0.089)	24.6%
Individual has DO, combination of tests declares mixed	0.106 (0.088 to 0.127)	36.4%
Individual has DO, combination of tests declares no condition	0.046 (0.035 to 0.058)	15.7%
Total DO	0.291 (0.257 to 0.325)	100.0%
Individual has mixed, combination of tests declares USI	0.080 (0.066 to 0.096)	38.8%
Individual has mixed, combination of tests declares DO	0.001 (0.000 to 0.002)	0.4%
Individual has mixed, combination of tests declares mixed	0.125 (0.105 to 0.147)	60.6%
Individual has mixed, combination of tests declares no condition	0.0005 (0.0001 to 0.0011)	0.2%
Total mixed	0.207 (0.175 to 0.240)	100.0%
Individual has no condition, combination of tests declares USI	0.086 (0.068 to 0.107)	51.9%
Individual has no condition, combination of tests declares DO	0.009 (0.006 to 0.012)	5.2%
Individual has no condition, combination of tests declares mixed	0.013 (0.010 to 0.017)	7.8%
Individual has no condition, combination of tests declares no condition	0.058 (0.043 to 0.074)	35.1%
Total for no condition	0.166 (0.137 to 0.197)	100.0%

TABLE 47 *Results of history and diary*

Condition and diagnosis	Total for all individuals (95% percentile)	Mean value as a percentage of the total for each condition
Individual has USI, combination of tests declares USI	0.223 (0.197 to 0.251)	66.4%
Individual has USI, combination of tests declares DO	0.007 (0.006 to 0.009)	2.2%
Individual has USI, combination of tests declares mixed	0.086 (0.07 to 0.104)	25.6%
Individual has USI, combination of tests declares no condition	0.019 (0.016 to 0.023)	5.8%
Total USI	0.336 (0.301 to 0.371)	100.0%
Individual has DO, combination of tests declares USI	0.006 (0.001 to 0.013)	2.1%
Individual has DO, combination of tests declares DO	0.155 (0.133 to 0.179)	53.4%
Individual has DO, combination of tests declares mixed	0.122 (0.104 to 0.141)	41.9%
Individual has DO, combination of tests declares no condition	0.008 (0.002 to 0.017)	2.6%
Total DO	0.291 (0.257 to 0.325)	100.0%
Individual has mixed, combination of tests declares USI	0.009 (0.002 to 0.020)	4.3%
Individual has mixed, combination of tests declares DO	0.016 (0.013 to 0.019)	7.6%
Individual has mixed, combination of tests declares mixed	0.182 (0.153 to 0.212)	87.7%
Individual has mixed, combination of tests declares no condition	0.001 (0.000 to 0.002)	0.4%
Total mixed	0.207 (0.175 to 0.240)	100.0%
Individual has no condition, combination of tests declares USI	0.053 (0.042 to 0.064)	31.8%
Individual has no condition, combination of tests declares DO	0.026 (0.020 to 0.033)	15.6%
Individual has no condition, combination of tests declares mixed	0.020 (0.015 to 0.026)	12.2%
Individual has no condition, combination of tests declares no condition	0.067 (0.054 to 0.081)	40.4%
Total for no condition	0.166 (0.137 to 0.197)	100.0%

TABLE 48 *Results of history and validated scales*

Condition and diagnosis	Total for all individuals (95% percentile)	Mean value as a percentage of the total for each condition
Individual has USI, combination of tests declares USI	0.289 (0.259 to 0.32)	86.1%
Individual has USI, combination of tests declares DO	0.0005 (0.0003 to 0.0007)	0.1%
Individual has USI, combination of tests declares mixed	0.043 (0.036 to 0.052)	12.9%
Individual has USI, combination of tests declares no condition	0.003 (0.002 to 0.004)	0.9%
Total USI	0.336 (0.301 to 0.371)	100.0%
Individual has DO, combination of tests declares USI	0.075 (0.063 to 0.089)	25.9%
Individual has DO, combination of tests declares DO	0.060 (0.048 to 0.073)	20.5%
Individual has DO, combination of tests declares mixed	0.118 (0.100 to 0.137)	40.5%
Individual has DO, combination of tests declares no condition	0.038 (0.030 to 0.047)	13.1%
Total DO	0.291 (0.257 to 0.325)	100.0%
Individual has mixed, combination of tests declares USI	0.080 (0.066 to 0.095)	38.6%
Individual has mixed, combination of tests declares DO	0.0013 (0.0008 to 0.0019)	0.6%
Individual has mixed, combination of tests declares mixed	0.125 (0.105 to 0.147)	60.4%
Individual has mixed, combination of tests declares no condition	0.0008 (0.0005 to 0.0013)	0.4%
Total mixed	0.207 (0.175 to 0.240)	100.0%
Individual has no condition, combination of tests declares USI	0.096 (0.078 to 0.116)	57.8%
Individual has no condition, combination of tests declares DO	0.007 (0.005 to 0.009)	4.4%
Individual has no condition, combination of tests declares mixed	0.014 (0.011 to 0.018)	8.6%
Individual has no condition, combination of tests declares no condition	0.049 (0.038 to 0.061)	29.2%
Total for no condition	0.167 (0.137 to 0.197)	100.0%

TABLE 49 *Results of cost-effectiveness analyses (probabilistic values)*

	Referral to specialist secondary assessment	Incremental costs (95% percentile) (£)	Incremental effectiveness (95% percentile)	Incremental cost-effectiveness (£)
0% Referred	Pad test	4.06 (3.07 to 5.25)	0.0307 (0.0255 to 0.0361)	132
	Diary	3.75 (3.12 to 4.46)	0.1057 (0.0830 to 0.1276)	35
	Scale	3.74 (3.14 to 4.45)	0.0290 (0.0246 to 0.0339)	129
50% Referred	Pad test	5.97 (4.44 to 8.09)	0.0307 (0.0256 to 0.0361)	195
	Diary	5.98 (4.64 to 8.58)	0.1055 (0.0782 to 0.1640)	57
	Scale	6.09 (4.64 to 8.32)	0.0290 (0.0246 to 0.0337)	210
100% Referred	Pad test	7.82 (5.43 to 11.48)	0.0307 (0.0255 to 0.036)	255
	Diary	8.16 (5.67 to 12.06)	0.1054 (0.0837 to 0.1266)	77
	Scale	8.42 (5.81 to 12.63)	0.029 (0.0245 to 0.0339)	290

are positive since all strategies that involve an additional diagnostic test involve greater cost than history alone. They are also more effective than history alone in detecting cases. The incremental cost-effectiveness shows the additional costs incurred per additional case detected. The incremental cost-effectiveness ratio indicates differences between the tests. The additional cost per extra case detected was generally highest for scales, varying between £129 and £290. Next highest was the pad test, which varied from £129 to £255. Diary had the most favourable cost-effectiveness ratios, varying between £35 and £77 per extra unit of effectiveness.

These results are also presented in *Figure 18* as a cost-effectiveness acceptability curve. Shown here are the curves for the two extreme cases, 0% referred and 100% referred. These curves show the probability that each strategy is cost-effective, given different values placed on a case detected. The higher the value of a case detected, the more likely it is that a strategy detecting additional cases will be considered worthwhile. The curves are incremental; history alone is compared to the other three strategies and each of these strategies is compared to history. For very low values given to a case detected, history alone is the preferred strategy as it has the lowest cost. However, as the value given to a case detected rises so does the probability that any of the other strategies are cost-effective. Increasing the proportion referred increases the value of a case detected that is needed for the joint test strategies to be preferred to history alone.

Probabilistic sensitivity analysis

Table 50 and *Figure 19* show the results of a probabilistic model where individuals are referred for urodynamics as well as specialist secondary assessment. It can be seen from *Table 50* that referral to urodynamics dramatically increases the incremental cost per individual with any condition diagnosed compared with history alone. This is because more individuals are being referred in the joint test strategies and referral is more expensive because it includes urodynamics. However, including urodynamics also increases the number of individuals with any condition diagnosed, as urodynamics is effective in detecting cases. Although there are extra cases detected, the incremental costs per additional unit of effect increase as there are large additional costs (the urodynamic referral) but only small extra numbers of individuals with any condition diagnosed. For example, with 100% referral, diary in addition to history costs an extra £275 per person with any condition diagnosed, compared with history alone.

One-way sensitivity analysis

Table 51 presents a series of one-way sensitivity analyses that were carried out on a deterministic model. The sensitivity analysis was carried out on a model where 50% of individuals were referred to specialist secondary assessment. In all cases the

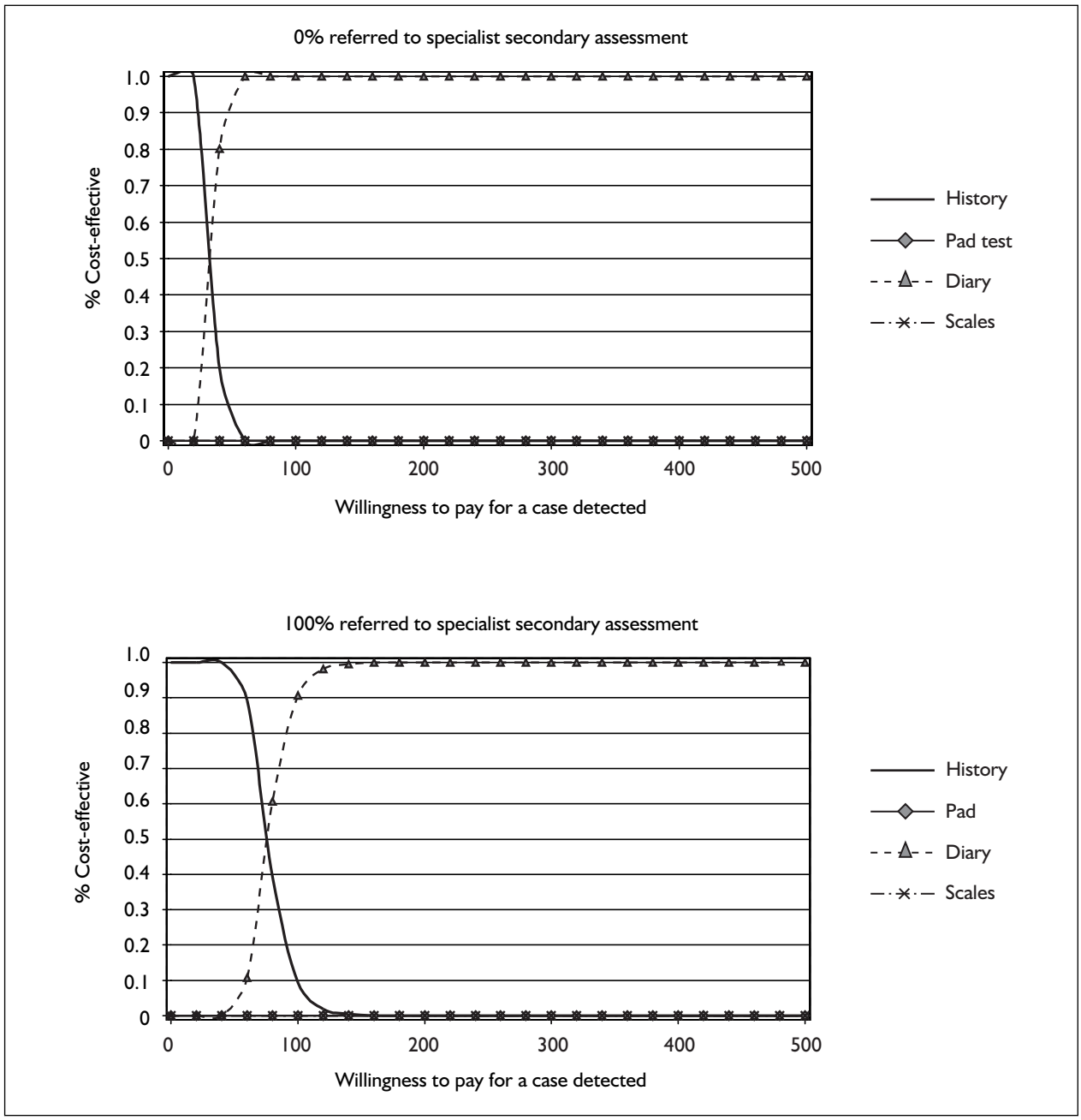


FIGURE 18 *Cost-effectiveness acceptability curves for referral to specialist secondary assessment*

values given are the incremental cost per extra unit of effect generated compared with history alone. In the first part of *Table 51* the proportion of individuals who had no condition was varied from 0 to 1. The more individuals have USI, DO or mixed, the lower the cost-effectiveness ratios. If 80% of the sample have no condition the cost-effectiveness ratios for pad tests and scales are approximately £1000 per unit of effect. In the second part of *Table 51* the performance of the various tests is varied between the upper and lower points of their 95% confidence intervals. Of particular importance is the sensitivity of history for USI, as the higher the sensitivity of history, the fewer cases remain for additional tests to detect. Also given are the effects of varying the sensitivity and specificity of pad test, diary and scales. As expected, as sensitivity and specificity increase, the cost-effectiveness ratios become more favourable. The final part of *Table 51* shows the effect of varying cost estimates. As the cost of carrying out tests and referrals increases so does the incremental cost-effectiveness ratio. However, the model seems less sensitive within the range of the confidence intervals for costs than for other variables such as sensitivity and specificity.

TABLE 50 *Results of model with positives referred to specialist secondary assessment and urodynamics*

	Referral to specialist secondary assessment	Incremental costs (95% percentile) (£)	Incremental effectiveness (95% percentile)	Incremental cost-effectiveness (£)
50% Referred	Pad test	10.23 (6.99 to 15.16)	0.038 (0.0317 to 0.0448)	269
	Diary	10.93 (7.55 to 16.15)	0.0855 (0.0682 to 0.1023)	128
	Scale	11.34 (7.84 to 16.83)	0.0402 (0.0350 to 0.0459)	282
100% Referred	Pad test	16.28 (10.27 to 25.94)	0.0452 (0.0359 to 0.0556)	360
	Diary	18.05 (11.47 to 28.54)	0.0655 (0.0522 to 0.0785)	275
	Scale	18.84 (11.90 to 29.71)	0.0513 (0.0436 to 0.0597)	367

The case where referral is 0% is not shown as this is equivalent to values for 0% in *Table 49*.

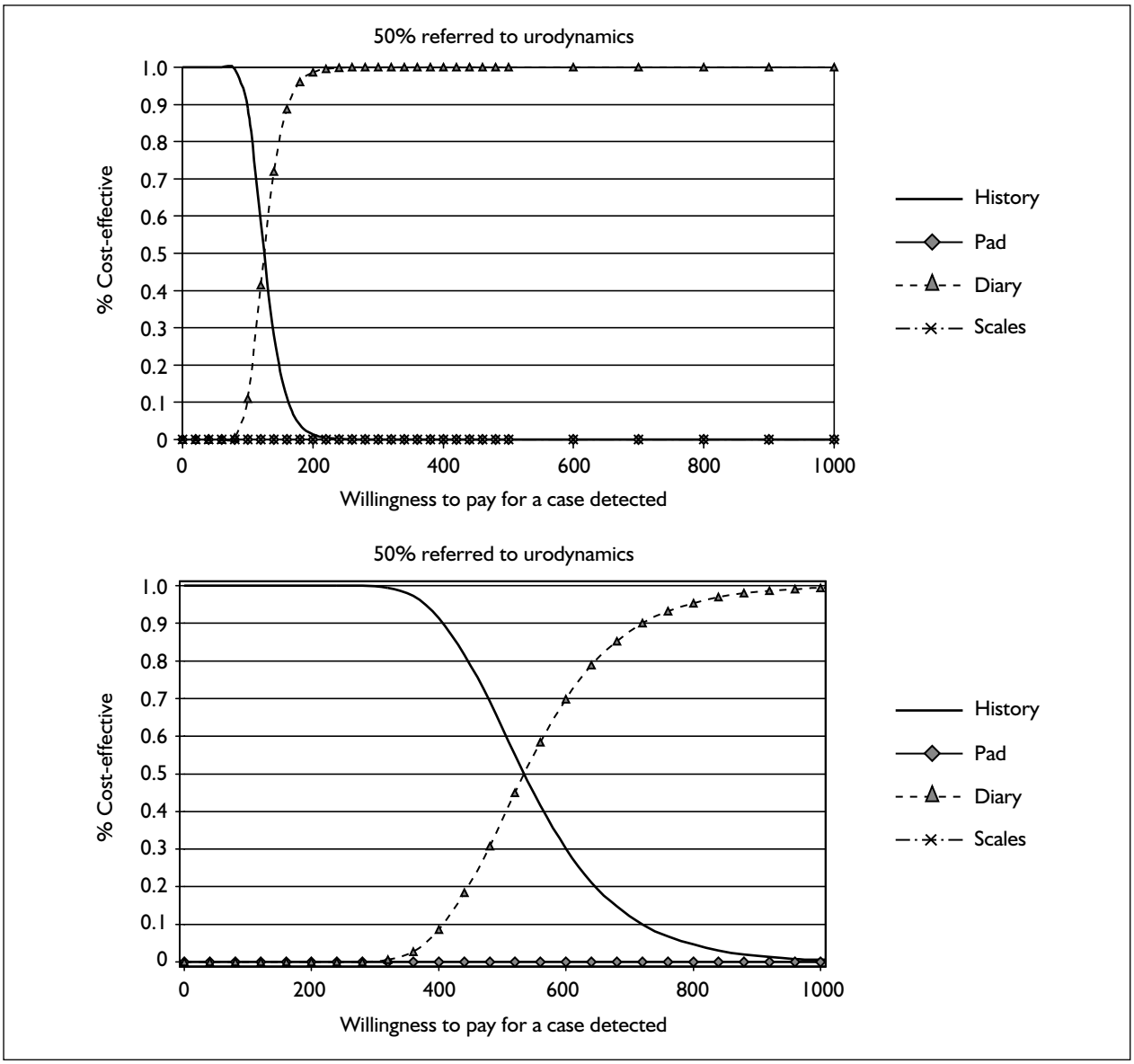


FIGURE 19 *Cost-effectiveness acceptability curves for sensitivity analysis on the effect of referral to urodynamics*

TABLE 51 *One-way sensitivity analyses on the probabilities used in the model*

Probability (0 to 1)		**0**	**0.2**	**0.4**	**0.6**	**0.8**	**1**
Proportion of individuals who have no condition (base case 0.166)	Pad test	152	205	295	474	1010	NA
	Diary	47	59	79	119	239	NA
	Scales	157	224	334	555	1218	NA
Probability (range of 95% CI)				**Lower 95% CI**		**Upper 95% CI**	
Sensitivity of history for USI	Pad test			176		219	
	Diary			57		57	
	Scales			190		237	
Specificity of history for USI	Pad test			193		197	
	Diary			56		58	
	Scales			207		214	
Sensitivity of history for DO	Pad test			193		196	
	Diary			53		61	
	Scales			209		211	
Specificity of history for DO	Pad test			194		195	
	Diary			56		57	
	Scales			209		211	
Sensitivity of pad test for USI	Pad test			215		186	
Specificity of pad test for USI	Pad test			214		180	
Sensitivity of diary test	Diary			66		53	
Specificity of diary test for USI	Diary			58		55	
Sensitivity of scales	Scales			222		200	
Specificity of scales	Scales			223		198	
Cost variables				**Lower 95% CI**		**Upper 95% CI**	
Pad test cost	Pad test			159		230	
Diary cost	Diary			50		63	
Scales cost	Scales			187		233	

NA, not applicable.

Chapter 5
Discussion

This is the first systematic review of methods for diagnosing urinary incontinence, meta-analysing the data, where possible, from different studies to generate conclusions about the diagnostic performance of commonly used diagnostic methods in both primary and secondary care. The objectives of the review were to identify, appraise and summarise the published evidence, quantitatively synthesise the extracted evidence (where possible) and construct an economic model to examine the cost-effectiveness of simple, commonly used primary care tests.

Appraisal of the systematic review

Research methodology

Search strategy

A systematic literature search was undertaken using three databases. There was an overlap between the databases, particularly MEDLINE and EMBASE: 45% of the studies identified by EMBASE were also identified by MEDLINE. CINAHL contributed the lowest papers to the review (seven). The search strategy was based on the Cochrane and NHS CRD strategies for identifying studies of diagnostic performance, which is well validated. It is important, for consistency and accuracy, for systematic reviews of diagnostic methods to use these strategies.

Keywords were added to the generic search strategies for identifying diagnostic studies to identify all possible tests used for the diagnosis of urinary incontinence, including terms for potential permutations of their names. However, it is possible that relevant studies may have been missed that use unusual or obscure diagnostic tests.

The development of online bibliographic databases in recent years means that handsearching of journals has become less important.[216] As urinary incontinence and diagnostic performance are well-established medical subheadings it was felt that using a detailed search strategy would identify a high proportion of relevant studies and that handsearching would not identify a significant number of additional studies.

A large number of papers was identified from the search (6009), of which 121 were deemed relevant for inclusion in the review. All papers compared two or more assessment/diagnostic techniques. A two-stage exclusion process was applied and decisions on relevance were checked in a random selection of 20% of cases; it is acknowledged that some papers of relevance may have been excluded unintentionally. There was diversity across the papers in diagnostic methods studied, methodology, analysis of the data and quality of reporting.

Inclusion/exclusion criteria

The extent to which the questions within a systematic review can be answered depends on the nature and quality of primary studies available. The inclusion criteria in this study were broad: studies that presented any quantitative comparison between two or more methods of assessing urinary incontinence. The study excluded case reports, letters, non-primary research and research involving only children. All studies presented in a non-English language were also excluded, as time and financial constraints did not allow for the translation of such papers. However, it is possible that this may have excluded important studies.[217,218]

Assessment of relevance

A critical part of classifying the papers included in the systematic review was to determine what tests were compared. The development of the cross-tabulation table enabled this to be clearly recorded and all similar studies to be grouped together, aiding the quality assessment and data extraction processes.

Quality assessment

It is important to assess the quality of studies included in any systematic review in terms of internal validity, external validity, and the quality of data analysis and reporting. The QUADAS tool[13,219] that was devised for this purpose is an important development. However, the relatively low levels of agreement between the investigators assessing the same papers using the tool suggest that it has limitations and that additional instructions need to be added according to the topic area of the individual review.

The most significant problem associated with study quality was in the reporting of results, with only a small proportion of relevant studies presenting data in a way that allowed inclusion in a meta-analysis. It was noticed that the quality of reporting was significantly higher in the more recent studies, indicating that standards are improving and this will be furthered by developments such as the Standards for Reporting of Diagnostic Accuracy (STARD) initiative.[220]

Data extraction

This is a potential source of error in any systematic review. The method of extracting data within a meeting of at least two study investigators was designed to minimise this, as studies could be discussed at length, reducing the chance of data being missed or incorrectly interpreted. A predefined form was used to record all relevant data during the data extraction process.

Data synthesis

The number of studies suitable for data synthesis was small. Another major problem was that studies that appeared to be comparing the same diagnostic tests were in fact comparing very different variations of the same test. For example, within pad tests, there were three different types: 1 hour, 24 hour and 48 hour. Both the paucity of evidence and the heterogeneous reporting of those studies that were identified severely limited the ability to undertake meaningful meta-analyses.

In addition, the heterogeneous nature of the studies identified, in terms of the precise diagnostic methods used or the patient population to which they were applied, meant that those meta-analyses that could be performed only included a small number of studies.

Specific methodological issues that were identified during the systematic review included the issue of indirect comparisons, classification of patients into more than two diagnostic categories, e.g. USI, DO or mixed, and the reporting of both raw data in terms of an ROC curve/table and summary data, for example a single estimate of sensitivity and specificity. This parallels the situation found in other areas in which some studies report individual patient, while others report only summary data.[221]

Economic modelling

It was assumed that it would always be good practice to take a history. The relevant question is, therefore, is it worth carrying out other tests in primary care in addition to taking a history? Therefore the extra costs and numbers of individuals with any condition diagnosed compared with history alone were examined. On this basis, the urinary diary performs well as it generates extra cases detected for the lowest extra cost. This is because the diary has been taken as a test for DO and the sensitivity of history for detecting DO is much lower than for detecting USI. In other words, far more cases of DO are not detected by history and therefore there is more scope for an additional test to detect additional 'missing' cases. However, a number of things should be considered when evaluating these results. It is important to consider that these tests are only evaluated in terms of their ability to diagnose urinary conditions and do not consider any other benefits that the information they generate have in treating individuals, for example if considerations of severity of leakage had an impact on the likelihood of receiving surgery. It should also be noted that the unit of effectiveness considers the value of a case of DO, USI and mixed found to be of equal importance. If it was considered more important to diagnose USI than DO then the relative values of tests for DO and tests for USI may change. Finally, the measures of the performance of these tests are generally based on single studies, so there is likely to be considerable uncertainty over the values of these estimates.

The estimates of prevalence used in the model come from urodynamics carried out on a group of individuals referred from a primary care setting. These are likely to be the more serious or intractable cases. The prevalence of these conditions in the more general group, who present to primary care, may be lower. Sensitivity analysis shows that the cost-effectiveness of these tests is sensitive to the prevalence; the likely occurrence of these conditions is therefore an important consideration in their implementation.

It is clear from this analysis that the decisions taken after the use of these tests have implications for their cost-effectiveness. There is likely to be wide variation in referral patterns among primary care practitioners. It is important to consider that in this simple model the analysis ends at secondary care referral, when in reality there may be a series of secondary care services received, and benefits obtained, from these services.

An important consideration in the interpretation of this work is the value placed on an individual with any condition detected. It is clear from the cost-effectiveness acceptability curves (*Figures 18* and *19*) that as the values of this outcome change,

then so do the conclusions for optimum management. If detecting an individual's condition is not highly valued then strategies where only history and no further tests are carried out would be the optimum ones. As the value placed on this outcome increases then strategies that involve extra costs but generate extra benefits will be optimum. The value of detecting an individual's urinary condition would depend on a number of factors not explicitly tested here. This would be expected to include the burden of a condition on an individual, and the cost and effectiveness of available treatments and therapies.

Implications of the findings

The literature dealing with the diagnosis of urinary incontinence is highly fragmented. Within primary care there are so many types of each test that it is almost impossible to find two studies that compare the same tests. There is no real agreement among clinical experts on what the 'gold standard' is for diagnosing urinary incontinence, whether it is urodynamics and, if so, what methods should be used. This review used the ICS-defined criterion that multichannel urodynamics is the gold standard test for diagnosing USI or DO. Owing to the large number of comparisons between a lot of different diagnostic tests, only the areas of high clinical interest will be discussed; namely, the most popular, simple and advanced investigations compared with multichannel urodynamics. Within each group there is a lack of literature dealing with the diagnosis of urinary incontinence or BOO in men, and for this reason the discussion of results will concentrate on diagnosis in women.

It is critical to make a distinction between tests and assessment methods that can be undertaken in primary care and those that can only be undertaken in secondary care. The majority of diagnostic and assessment processes can be undertaken in primary care and comprise clinical history-taking, the use of scales, physical examination, and simple tests such as diaries and pad tests. These tests are simple, are low in cost and carry low risks. The results of assessments and tests are used to identify a presumed diagnosis on which an appropriate management/treatment plan can be instigated.

Clinical history

The recording of a clinical history is critical in determining a symptomatic diagnosis in primary care. A large number of studies comparing the use of clinical history and urodynamics in female patients was identified. Pooled sensitivity and specificity values for diagnosis of USI in women suggest that a clinical history is highly sensitive (0.92, 95% CI 0.91 to 0.93), but less specific (0.56, 95% CI 0.53 to 0.60) in diagnosing USI. These findings suggest that a large proportion of women with USI can be correctly diagnosed in primary care and that initiating low-risk, low-cost behavioural treatment at this stage may be appropriate. The lower specificity suggests that women without USI may be incorrectly diagnosed; however, behavioural therapy should not have any detrimental effects and may result in some alleviation of symptoms.

With regard to the diagnosis of DO by clinical history-taking, sensitivity was found to be lower than for USI (0.61, 95% CI 0.57 to 0.65), but specificity was found to be high (0.87, 95% CI 0.85 to 0.89). This indicates that history-taking may correctly exclude those women who do not have DO, but that further investigations may be required for those who present with DO symptoms to confirm their status before any treatment is initiated. The next stage for those whose history suggests DO may be a further simple, non-invasive test, such as a urinary diary.

Simple investigations

Validated scales

The studies in this group highlight the fragmented nature of the overall literature. Seven different scales are compared and there is currently no consensus on the most effective scales to use in clinical practice. The most commonly researched scale was the UDI. Combining data from two studies resulted in a sensitivity of 0.87 (95% CI 0.82 to 0.92) and specificity of 0.60 (95% CI 0.51 to 0.69) for the diagnosis of USI in women based on one question from the UDI. The diagnostic value of this scale is comparable to taking just a clinical history, indicating that this scale may not add anything to the diagnostic procedure.

Little evidence was found on scales that seek to diagnose DO. One study reported the Gaudenz incontinence questionnaire to be 0.45 sensitive and 0.56 specific, less accurate than clinical history-taking.

There needs to be consensus about the most appropriate scale for the diagnosis of urinary incontinence. Efforts should be concentrated on developing and amending one or two scales, rather than continually developing new scales, unless based on specific clinical need.

Pad tests

Because of the many different pad tests used to investigate urinary incontinence it is difficult to draw any firm conclusions about diagnostic accuracy. The majority of literature in this area was concerned with the diagnosis of USI. Although high sensitivity and specificity values were reported in some studies, there were insufficient studies that compared the same pad tests and presented the data appropriately, and therefore no formal pooling of data could be carried out.

Urinary diary

A number of different urinary diaries was studied. Four studies compared a urinary diary with urodynamics and each study used a different type of diary. The only study to present data in a format that allowed sensitivity and specificity to be calculated reported values of 0.88 (95% CI 0.71 to 0.96) and 0.83 (95% CI 0.77 to 0.87), respectively. This indicates that this type, an index derived from various variables of a urinary diary, may be effective for the diagnosis of DO. The economic modelling suggests that the urinary diary performs well in combination with a clinical history for the diagnosis of DO. As the review has shown a clinical history to have a relatively low sensitivity for diagnosing DO (0.56) there is more scope for an additional test to detect additional cases. These conclusions should be treated with some caution as they were drawn from the results of a single study.

A recent symposium at the International Continence Society 2003 Annual Conference found that 59% of clinicians prefer to use a urinary diary for the initial evaluation of patients, suggesting that this is the non-invasive test of choice.[222] This opinion contrasts with the amount of literature available on the urinary diary.

Other simple investigations

A small number of studies investigated the diagnosis of urinary incontinence by an algorithm method or a battery of tests. This appears to be a sensible approach, particularly in primary care, and arguably the most similar to real-life clinical practice. Although the number of studies in these groups was small and pooling of the data was not possible, the agreement between the results of these tests and multichannel urodynamics indicates that future research may be of significant interest.

Advanced investigations

Imaging by ultrasound and X-ray

A large amount of literature was identified that dealt with imaging the lower urinary tract for the diagnosis of USI through ultrasound and X-ray methods. Ultrasound was found to be the most effective method of imaging the two anatomical features used for the diagnosis of USI: the observance of leakage from the bladder and movement of the bladder neck during provocation. This method resulted in higher sensitivities (0.89, 95% CI 0.84 to 0.93, and 0.84, 95% CI 0.76 to 0.90) and specificities (0.82, 95% CI 0.73 to 0.89, and 0.86, 95% CI 0.79 to 0.91) for these landmarks than X-ray imaging. This suggests that ultrasound is a valuable diagnostic tool that could be used in secondary care as an alternative to multichannel urodynamics, owing to likely lower risks, costs and discomfort for the patient, although few studies reported these patient-based outcomes.

Urodynamics

The review identified literature on a number of different urodynamic tests compared with the gold standard of multichannel urodynamics. It is arguable, however, whether such tests are less unpleasant, expensive or of less risk to perform, and whether it would be better just to perform the gold-standard test.

A number of papers compared the clinical stress test with multichannel urodynamics for the diagnosis of USI, resulting in a high sensitivity of 0.85 (95% CI 0.78 to 0.91) and specificity of 0.83 (95% CI 0.74 to 0.90). These studies performed the clinical stress test with an artificially filled bladder, which increases the invasiveness of the test. If the test could be performed with a naturally full bladder, with no significant detriment to diagnostic accuracy, then this would be a very useful non-invasive diagnostic test that could be used in primary and secondary care. Research into such a test would be of great clinical interest.

Within the review, far fewer studies were undertaken in primary care than in secondary care settings. This has important implications for interpretation of the findings. The studies undertaken in secondary care are mainly undertaken on referred patients attending as outpatients. They are very different to undifferentiated patients presenting in primary care. It is likely that referred patients have already undergone some form of diagnostic process and, therefore, using various diagnostic assessment tools with this population may produce greater levels of sensitivity and specificity than in a mainly unreferred, undifferentiated population.

Chapter 6

Conclusions, implications and recommendations

Conclusions

- This is the first systematic review of methods of assessing urinary incontinence.
- In total, 6009 papers were identified from the search, of which a final 121 were deemed relevant for inclusion in the review. These papers compared two or more assessment/diagnostic techniques.
- A large number of different tests is used in the diagnosis of urinary incontinence, generating a great number of possible comparisons. The extent of heterogeneity between studies meant that few papers actually compared the same assessment/diagnostic tests. A matrix was constructed so that each relevant paper could be assigned to a cell in the matrix. However, even when a cell contained ten papers comparing, for example, scales with urodynamics, within the cell seven different scales had been used, making actual comparison impossible.
- Reporting in the primary studies was generally poor. Both the clinical heterogeneity and poor reporting meant that it was often impossible to synthesise results, although studies reported in recent years generally reported better than older studies.
- Clinical interpretation was often difficult because few studies could actually be synthesised and conclusions drawn. The following information could be deduced from the available data:
  - A large proportion of women with USI can be correctly diagnosed in primary care from clinical history alone.
  - The value of validated scales or pad tests could not be determined from the available data owing to the wide range of different types of instrument used.
  - On the basis of diagnosis the diary appears to be the most cost-effective of the three primary care tests (diary, pad test and validated scales) when used in addition to clinical history.
  - Ultrasound imaging may offer a valuable alternative to urodynamic investigation.
  - The clinical stress test is effective in the diagnosis of USI. Adaptation of such a test so that it could be performed in primary care with a naturally filled bladder may prove clinically useful.
  - If a patient is to undergo an invasive urodynamic procedure, multichannel urodynamics is likely to give the most accurate result in a secondary care setting.
- There is a dearth of literature on the diagnosis of urinary incontinence in men, with no studies meeting the criteria for data extraction in the diagnosis of BOO.

Implications

- Most simple diagnostic methods can be undertaken in primary or secondary care.
- A thorough and accurate clinical history is crucial.
- The use of simple investigations (e.g. pad test and diary) may offer useful information on severity which, when combined with history, may provide sufficient information to commence primary care interventions (which are low cost and low risk).
- From the data available the urinary diary is the most cost-effective simple investigation to use in combination with the clinical history.
- If urodynamic investigations are deemed necessary, multichannel urodynamics will offer the most accurate result.
- There is a lack of research in certain areas of clinical interest and a general lack of high-quality work, particularly economic studies.

Future research recommendations

- There is a need for large-scale, high-quality, primary studies evaluating the systematic use of a number of diagnostic methods in a primary care setting, so that the results of this systematic review can be verified or not. Such studies should include not only an assessment of clinical effectiveness, in this case diagnostic accuracy, but also an assessment of costs and quality of life/patient acceptance/satisfaction to inform future health policy decisions.
- There is a need for the development and standardisation of scales, pad tests and diaries

for use in the diagnosis and measurement of severity of urinary incontinence.
- Only a small number of studies investigated the diagnosis of urinary incontinence using an algorithm or a battery of tests. Such a common-sense approach, which mirrors clinical practice, warrants further investigation.
- Research on the accuracy of a stress test using a naturally filled bladder would be of clinical interest.
- In terms of economic modelling, the literature has only begun to address the cost-effectiveness of the use of diagnostic tools in urinary incontinence. There has been some work published examining the use of urodynamics before surgery.[223,224] However, there is a lack of studies that consider the use of low-cost tests such as diaries in primary care. Since these are widely used techniques and they have the potential to impact on other services in terms of referrals to secondary care and treatment received, it would be important to consider explicitly the cost-effectiveness of their use. In terms of the use of simple diagnostic tests there would be a potential for their results to be used in primary care to inform treatment options. This could lead to improvements in health.
- A full economic model, which incorporates both diagnosis and treatment, and evaluates outcomes in terms of cost per QALY, would enable more rational decisions to be made; this would represent an important focus for future work.
- Studies should be carried out and reported to a better standard. The recommendations of the STARD initiative should be followed to ensure the accuracy and completeness of reporting design and results. The flowchart for the suggested design and checklist for the reporting of a study of diagnostic accuracy developed by STARD are presented in Appendix 8.
- Given the demographics of the UK population and the recently reported prevalence of any urinary incontinence (in those aged 40 and over) of 34% for women and 14% for men,[2] there will be an increasing burden placed on primary (and secondary) care services in terms of the diagnostic assessment and appropriate treatment of incontinence. Therefore, identifying which are the most clinically and cost-effective methods is of crucial importance.

Dissemination and timescale for updating

The target audience for dissemination of these results is clinicians. It may also prove interesting to those involved in systematic reviews of diagnostic methods. Realistically, in light of the broad nature of the literature and the improvements in reporting, the updating of this review should be considered within 4–6 years.

Acknowledgements

The authors would like to thank Mary Edmunds-Otter for help in developing the search strategies and Ariadna Juarez-Garcia for assisting with the economic modelling. We are grateful to Lesley Harris for administrative support and proof-reading of the final report, Ruth Matthews and Clare Gillies for additional statistical help, and Dr Helen Dallosso for reading the final report. We would also like to thank NHS CRD at York and particularly Penny Whiting, who allowed us to pilot the QUADAS tool. David Turner is funded by the Trent Institute for Health Services Research.

We would also like to thank those authors who responded to our requests for additional information: Antonio Cucchi, Hogne Sandvik, Gary Lemack and Gin-Den Chen.

Contribution of authors

All the authors were involved in the conception and design of the study, or analysis and interpretation of the data; drafting and revising the report; and final approval of the version to be published.

Individual contributions were as follows: JL Martin (Research Fellow) undertook the day-to-day activity on the project, quality assessment, interpretation and presentation of data, drafting of the full report and making revisions to the report. KS Williams (Senior Research Fellow in Nursing) was principal investigator and was involved in the design of the study, day-to-day supervision of the study, quality assessment, interpretation of data, drafting parts of the report and revising the report. KR Abrams (Professor of Medical Statistics) was involved in the conception and design of the study, was fully involved in supervising the interpretation and presentation of data, drafted sections of the report and commented on it. D Turner (Research Fellow in Health Economics) was involved in the conception and design of the study, with particular emphasis on the health economics component, undertook analysis, drafting and revision of the economics chapter and commented on the full report. A Sutton (Senior Lecturer in Medical Statistics) was involved in the conception and design of the study, was involved in interpretation and presentation of data, drafted sections of the report and commented on it. C Chapple (Consultant Urologist) was involved in the conception and design of the study, quality assessment, clinical interpretation and commenting on the report. RP Assassa (Consultant Gynaecologist) was involved in the conception and design of the study, quality assessment, clinical interpretation and commenting on the report. C Shaw (Senior Research Fellow) was involved in the conception and design of the study, quality assessment, interpretation and commenting on the report. F Cheater (Professor of Public Health Nursing) was involved in the conception and design of the study, quality assessment, interpretation and commenting on the report.

References

1. Abrams P, Cardozo L, Fall M, Griffiths D. The standardisation of terminology in lower urinary tract function: report from the standardisation sub-committee of the International Continence Society. *Urology* 2003;**61**:37–49.
2. McGrother CM, Donaldson MMK, Wagg A, Matharu G, Williams KS, Warsame J, *et al*. Continence. In Stevens A RJMJ, editor. *Health care needs assessment: the epidemiologically based needs assessment reviews*. Abingdon: Radcliffe Medical Press; 2003. http://www.hcna.radcliffe-oxford.com/contframe.htm
3. Turner DA, McGrother CW, Dallosso HM, Shaw C, Cooper NJ, MRC Incontinence Study Team. The cost of urinary storage disorders in the UK. 2003 (submitted).
4. Valvanne J, Juva K, Erkinjuntti T, Tilvis R. Major depression in the elderly: a population study in Helsinki. *Int Psychogeriatr* 1996;**8**:437–43.
5. Berglund AL, Eisemann M, Lalos O. Personality characteristics of stress incontinent women: a pilot study. *Journal of Psychosomatic Obstetrics and Gynecology* 1994;**15**:165–70.
6. Herzog AR, Fultz N, Brock BM, Brown MB, Diokno AC. Urinary incontinence and psychological distress among older adults. *Psychol Aging* 1988;**3**:115–21.
7. Van Der Vaart CH, De Leeuw JR, Roovers JP, Heintz AP. The effect of urinary incontinence and overactive bladder symptoms on quality of life in young women. *British Journal of Urology International* 2002;**90**:544–9.
8. Abrams P, Lowry SK, Wein AJ, Bump R, Denis L, Kalache A. Assessment and treatment of urinary incontinence. *Lancet* 2000;**355**:2153–8.
9. Homma Y, Batista J, Bauer D, Griffiths P, Hilton G, Kramer G, Lose G, Rosier P. Urodynamics. In Abrams P, Cardozo L, Khoury S, Wein A, editors. *Incontinence: 2nd International Consultation on Incontinence*. 2nd ed. Plymouth: Plymbridge Distributors; 2002. pp. 317–72.
10. Schafer W, Abrams P, Liao L, Mattiasson A, Pesce F, Spangberg A, *et al*. Good urodynamics practice: uroflowmetry, filling cystometry and pressure–flow studies. *Neurourol Urodyn* 2002; **21**:261–74.
11. Gorton E, Stanton S. Women's attitudes to urodynamics: a questionnaire survey. *Br J Obstet Gynaecol* 1999;**106**:851–6.
12. Chapple CR, MacDiarmid SA. *Urodynamics made easy*. 2nd ed. London: Harcourt; 2000.
13. Matharu G, Donaldson MMK, McGrother CW, Matthews RJ. Relationship between urinary symptoms reported in a postal questionnaire and urodynamic diagnosis. *Neurourol Urodyn* 2005;**24**:100–5.
14. NHS Centre for Reviews and Dissemination. *Undertaking systematic reviews of research on effectiveness. CRD Guidelines for those carrying out commissioning reviews*. CRD Report No. 4. York: NHS Centre for Reviews and Dissemination; 2001. pp. 1–50.
15. Whiting P, Rutjes AWS, Dinnes J, Reitsma JB, Bossuyt PMM, Kleijnen J. The development and validation of methods for assessing the quality and reporting of diagnostic studies. *Health Technol Assess* 2004;**8**(25).
16. Amundsen C, Lau M, English SF, McGuire EJ. Do urinary symptoms correlate with urodynamic findings? *J Urol* 1999;**161**:1871–4.
17. Scotti RJ, Myers DL. A comparison of the cough stress test and single-channel cystometry with multichannel urodynamic evaluation in genuine stress incontinence. *Obstet Gynecol* 1993;**81**:430–3.
18. Ishiko O, Sumi T, Hirai K, Ogita S. Classification of female urinary incontinence by the scored incontinence questionnaire. *Int J Gynaecol Obstet* 2000;**69**:255–60.
19. Elser DM, Fantl JA, McClish DK. Comparison of 'subjective' and 'objective' measures of severity of urinary incontinence in women. Program for Women Research Group. *Neurourol Urodyn* 1995; **14**:311–16.
20. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PMM. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol* 2003;**56**:1129–35.
21. Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. In Egger M, Smith G, Altman D, editors. *Systematic reviews in health care, meta-analysis in context*. 2nd ed. London: BMJ Books; 2001. pp. 248–82.
22. Bergman J, Elia G. Effects of the menstrual cycle on urodynamic work-up: should we change our practice? *Int Urogynecol J* 1999;**10**:375–7.
23. Chaikin DC, Blaivas JG, Rosenthal JE, Weiss JP. Results of pubovaginal sling for stress

incontinence: a prospective comparison of 4 instruments for outcome analysis. *J Urol* 1999; **162**:1670–3.

24. Glazer HI, Romanzi L, Polaneczky M. Pelvic floor muscle surface electromyography. Reliability and clinical predictive validity. *J Reprod Med* 1999; **44**:779–82.

25. Madersbacher S, Pycha A, Klingler CH, Schatzl G, Marberger M. The International Prostate Symptom score in both sexes: a urodynamics-based comparison. *Neurourol Urodyn* 1999; **18**:173–82.

26. Aanestad O, Flink R. Urinary stress incontinence. A urodynamic and quantitative electromyographic study of the perineal muscles. *Acta Obstet Gynecol Scand* 1999;**78**:245–53.

27. Siltberg H, Larsson G, Victor A. Cough-induced leak-point pressure – a valid measure for assessing treatment in women with stress incontinence. *Acta Obstet Gynecol Scand* 1998;**77**:1000–7.

28. Watson AJ, Currie I, Curran S, Jarvis GJ. A prospective study examining the association between the symptoms of anxiety and depression and severity of urinary incontinence. *Eur J Obstet Gynecol Reprod Biol* 2000;**88**:7–9.

29. Yoon E, Swift S. A comparison of maximum cystometric bladder capacity with maximum environmental voided volumes. *Int Urogynecol J Pelvic Floor Dysfunct* 1998;**9**:78–82.

30. Kirschner-Hermanns R, Scherr PA, Branch LG, Wetle T, Resnick NM. Accuracy of survey questions for geriatric urinary incontinence. *J Urol* 1998; **159**:1903–8.

31. McLennan MT, Bent AE. Supine empty stress test as a predictor of low valsalva leak point pressure. *Neurourol Urodyn* 1998;**17**:121–7.

32. Frauscher F, Helweg G, Strasser H, Enna B, Klauser A, Knapp R, *et al*. Intraurethral ultrasound: diagnostic evaluation of the striated urethral sphincter in incontinent females. *Eur Radiol* 1998;**8**:50–3.

33. Elbadawi A, Hailemariam S, Yalla SV, Resnick NM. Structural basis of geriatric voiding dysfunction. VII. Prospective ultrastructural/urodynamic evaluation of its natural evolution. *J Urol* 1997; **157**:1814–22.

34. Elbadawi A, Hailemariam S, Yalla SV, Resnick NM. Structural basis of geriatric voiding dysfunction. VI. Validation and update of diagnostic criteria in 71 detrusor biopsies. *J Urol* 1997;**157**:1802–13.

35. Moore KH, Foote A, Siva S, King J, Burton G. The use of the bladder neck support prosthesis in combined genuine stress incontinence and detrusor instability. *Aust N Z J Obstet Gynaecol* 1997; **37**:440–5.

36. Peters S. Don't ask, don't tell. Breaking the silence surrounding female urinary incontinence [review]. *Advance for Nurse Practitioners* 1997;**5**(5):41–4.

37. Siltberg H, Larsson G, Victor A. Frequency/volume chart: the basic tool for investigating urinary symptoms. *Acta Obstet Gynecol Scand Suppl* 1997; **166**:24–7.

38. Donnellan SM, Duncan HJ, MacGregor RJ, Russell JM. Prospective assessment of incontinence after radical retropubic prostatectomy: objective and subjective analysis. *Urology* 1997;**49**:225–30.

39. van Waalwijk van Doorn ES, Ambergen AW, Janknegt RA. Detrusor activity index: quantification of detrusor overactivity by ambulatory monitoring. *J Urol* 1997;**157**:596–9.

40. Managing acute and chronic urinary incontinence. US Department of Health and human services. *Journal of the American Academy of Nurse Practitioners* 1996;**8**:390–403.

41. Theofrastous JP, Cundiff GW, Harris RL, Bump RC. The effect of vesical volume on Valsalva leak-point pressures in women with genuine stress urinary incontinence. *Obstet Gynecol* 1996; **87**:711–14.

42. Moore AA, Siu AL. Screening for common problems in ambulatory elderly: clinical confirmation of a screening instrument. *Am J Med* 1996;**100**:438–43.

43. Woodtli A. Stress incontinence: clinical identification and validation of defining characteristics. *Nursing Diagnosis* 1995;**6**:115–22.

44. Theofrastous JP, Bump RC, Elser DM, Wyman JF, McClish DK. Correlation of urodynamic measures of urethral resistance with clinical measures of incontinence severity in women with pure genuine stress incontinence. The Continence Program for Women Research Group. *Am J Obstet Gynecol* 1995; **173**:407–12; Discussion 412–14.

45. Cetinel B, Turan T, Talat Z, Yalcin V, Alici B, Solok V. Update evaluation of benign prostatic hyperplasia: when should we offer prostatectomy? *Br J Urol* 1994;**74**:566–71.

46. Ouslander JG, Simmons S, Tuico E, Nigam JG, Fingold S, Bates-Jensen B, *et al*. Use of a portable ultrasound device to measure post-void residual volume among incontinent nursing home residents. *J Am Geriatr Soc* 1994;**42**:1189–92.

47. Mayer R, Wells TJ, Brink CA, Clark P. Correlations between dynamic urethral profilometry and perivaginal pelvic muscle activity. *Neurourol Urodyn* 1994;**13**:227–35.

48. Petros PE, Ulmsten U. Natural volume handwashing urethrocystometry: a physiological technique for the objective diagnosis of the unstable detrusor. *Gynecol Obstet Invest* 1993; **36**:42–6.

49. Lalos O, Berglund AL, Bjerle P. Urodynamics in women with stress incontinence before and after surgery. *Eur J Obstetr Gynecol Reprod Biol* 1993; **48**:197–205.

50. Walter JS, Wheeler JS, Morgan C, Zaszczurynski P, Plishka M. Measurement of total urethral compliance in females with stress incontinence. *Neurourol Urodyn* 1993;**12**:273–6.

51. Vinsnes AG, Hunskaar S. Distress associated with urinary incontinence, as measured by a visual analogue scale. *Scandinavian Journal of Caring Sciences* 1991;**5**:57–61.

52. van Waalwijk van Doorn ES, Remmers A, Janknegt RA. Extramural ambulatory urodynamic monitoring during natural filling and normal daily activities: evaluation of 100 patients. *J Urol* 1991; **146**:124–31.

53. Griffiths DJ, McCracken PN, Harrison GM. Incontinence in the elderly: objective demonstration and quantitative assessment. *Br J Urol* 1991;**67**:467–71.

54. Resnick NM. Noninvasive diagnosis of the patient with complex incontinence. *Gerontology* 1990; **36** Suppl 2:8–18.

55. Brocklehurst JC. Urinary incontinence in old age: helping the general practitioner to make a diagnosis. *Gerontology* 1990;**36** Suppl 2:3–7.

56. Kong TK, Morris JA, Robinson JM, Brocklehurst JC. Predicting urodynamic dysfunction from clinical features in incontinent elderly women. *Age Ageing* 1990;**19**:257–63.

57. Frazer MI, Haylen BT. Trigonal sensitivity testing in women. *J Urol* 1989;**141**:356–8.

58. Lose G, Jorgensen 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**:211–15.

59. Varpula M, Makinen J, Kiilholma P. Cough urethrocystography: the best radiological evaluation of female stress urinary incontinence? *Eur J Radiol* 1989;**9**:191–4.

60. Walters MD, Shields LE. The diagnostic value of history, physical examination, and the Q-tip cotton swab test in women with urinary incontinence. *Am J Obstet Gynecol* 1988;**159**:145–9.

61. Khan Z, Mieza M, Bhola A. Relative usefulness of physical examination, urodynamics and roentgenography in the diagnosis of urinary stress incontinence. *Surgery, Gynecology and Obstetrics* 1988;**167**:39–44.

62. Victor A, Larsson G, Asbrink AS. A simple patient-administered test for objective quantitation of the symptom of urinary incontinence. *Scand J Urol Nephrol* 1987;**21**:277–9.

63. Goldwasser B, Rife CC, Benson RC, Furlow WL, Barrett DM. Urodynamic evaluation of patients after the Camey operation. *J Urol* 1987;**138**:832–5.

64. Glezerman M, Glasner M, Rikover M, Tauber E, Bar-Ziv J, Insler V. Evaluation of reliability of history in women complaining of urinary stress incontinence. *Eur J Obstet Gynecol Reprod Biol* 1986; **21**:159–64.

65. Vehkalahti I, Kivela SL, Seppanen J. Are cystometric and cystoscopic examinations of any value for disabled incontinent elderly? *Scandinavian Journal of Primary Health Care* 1986; **4**:243–7.

66. Kauppila A, Alavaikko P, Kujansuu E. Detrusor instability score in the evaluation of stress urinary incontinence. *Acta Obstet Gynecol Scand* 1982; **61**:137–41.

67. Walter S, Olesen KP. Urinary incontinence and genital prolapse in the female: clinical, urodynamic and radiological examinations. *British Journal of Obstetrics and Gynaecology* 1982; **89**:393–401.

68. Robinson H, Stanton SL. Detection of urinary incontinence. *British Journal of Obstetrics and Gynaecology* 1981;**88**:59–61.

69. Fantl JA, Hurt WG, Beachley MC, Bosch HA, Konerding KF, Smith PJ. Bead-chain cystourethrogram: an evaluation. *Obstet Gynecol* 1981;**58**:237–40.

70. Thuroff JW, Jonas U, Petri E, Frohneberg D. Telemetric urodynamic investigations in female incontinence. *Prog Clin Biol Res* 1981;**78**:211–22.

71. Drutz HP, Mandel F. Urodynamic analysis of urinary incontinence symptoms in women. *Am J Obstet Gynecol* 1979;**134**:789–92.

72. Awad SA, Bryniak SR, Lowe PJ, Bruce AW, Twiddy DA. Urethral pressure profile in female stress incontinence. *J Urol* 1978;**120**:475–9.

73. Drutz HP, Shapiro BJ, Mandel F. Do static cystourethrograms have a role in the investigation of female incontinence? *Am J Obstet Gynecol* 1978; **130**:516–20.

74. Susset JG, Shoukry I, Schlaeder G, Cloutier D, Dutartre D. Stress incontinence and urethral obstruction in women: value of uroflowmetry and voiding urethrography. *J Urol* 1974;**111**:504–13.

75. Diokno AC, Wells TJ, Brink CA. Comparison of self-reported voided volume with cystometric bladder capacity. *J Urol* 1987;**137**:698–700.

76. Van Venrooij GEPM, Eckhardt MD, Gisolf KWH, Boon TA. Data from frequency–volume charts versus filling cystometric estimated capacities and prevalence of instability in men with lower urinary tract symptoms suggestive of benign prostatic hyperplasia. *Neurourol Urodyn* 2002;**21**:106–11.

77. Wolters M, Methfessel HD, Goepel C, Koelbl H. Computer-assisted virtual urethral pressure profile in the assessment of female genuine stress incontinence. *Obstet Gynecol* 2002;**99**:67–74.

78. Kerschan-Schindl K, Uher E, Wiesinger G, Kaider A, Ebenbichler G, Nicolakis P, *et al*. Reliability of pelvic floor muscle strength measurement in elderly incontinent women. *Neurourol Urodyn* 2002;**21**:42–7.

79. FitzGerald MP, Butler N, Shott S, Brubaker L. Bother arising from urinary frequency in women. *Neurourol Urodyn* 2002;**21**:36–40.

80. Radley SC, Rosario DJ, Chapple CR, Farkas AG. Conventional and ambulatory urodynamic findings in women with symptoms suggestive of bladder overactivity. *J Urol* 2001;**166**:2253–8.

81. Groutz A, Samandarov A, Gold R, Pauzner D, Lessing JB, Gordon D. Role of urethrocystoscopy in the evaluation of refractory idiopathic detrusor instability. *Urology* 2001;**58**:544–6.

82. Romanzi LJ, Groutz A, Heritz DM, Blaivas JG. Involuntary detrusor contractions: correlation of urodynamic data to clinical categories. *Neurourol Urodyn* 2001;**20**:249–57.

83. Araki I, Kitahara M, Oida T, Kuno S. Voiding dysfunction and Parkinson's disease: urodynamic abnormalities and urinary symptoms. *J Urol* 2000; **164**:1640–3.

84. Moretti M, Varaldo M, Malcangi B, Cichero A, Pittaluga P, Riva D. Introital sonography and urodynamic examination in stress urinary incontinence: anatomic and functional relationships. *Urodinamica* 1998;**8**:226–31.

85. Nitahara KS, Aboseif S, Tanagho EA. Long-term results of colpocystourethropexy for persistent or recurrent stress urinary incontinence. *J Urol* 1999; **162**:138–41.

86. Kirkemo A, Peabody M, Diokno AC, Afanasyev A, Nyberg LM Jr, Landis JR, *et al*. Associations among urodynamic findings and symptoms in women enrolled in the Interstitial Cystitis Data Base (ICDB) study. *Urology* 1997;**49**(5S):76–80.

87. Yeh NH, Chen GD, Lin LY, Wu GS, Su TH. Comparison of bladder neck mobility in patients with genuine stress incontinence and continent women by perineal sonography. *Journal of Medical Ultrasound* 1996;**4**:129–33.

88. Rivas DA, Chancellor MB. Utility of the American Urological Association symptom index in the diagnosis of women with voiding dysfunction. *International Urogynecology Journal* 1994;**5**:202–7.

89. Zollner-Nielsen M, Samuelsson SM. Maximal electrical stimulation of patients with frequency, urgency and urge incontinence: report of 38 cases. *Acta Obstet Gynecol Scand* 1992;**71**:629–31.

90. Mayer R, Wells T, Brink C, Diokno A, Cockett A. Handwashing in the cystometric evaluation of detrusor instability. *Neurourol Urodyn* 1991; **10**:563–69.

91. Thind P, Gerstenberg TC. One-hour ward test vs. 24-hour home pad weighing test in the diagnosis of urinary incontinence. *Neurourol Urodyn* 1991; **10**:241–5.

92. Wilson PD, Mason MV, Herbison GP, Sutherst JR. Evaluation of the home pad test for quantifying incontinence. *British Journal of Urology* 1989; **64**:155–7.

93. Yu HJ, Kuo HC, Chen SC, Law HS, Tsai TC. Evaluation of urodynamics for the diagnosis of female stress urinary incontinence. *Journal of Surgical Association Republic of China* 1988; **21**:326–33.

94. Chou TP, Gorton E, Stanton SL, Atherton M, Baessler K, Rienhardt G. Can uroflowmetry patterns in women be reliably interpreted? *International Urogynecology Journal* 2000;**11**:142–7.

95. Yossepowitch O, Gillon G, Baniel J, Engelstein D, Livne PM. The effect of cholinergic enhancement during filling cystometry: can edrophonium chloride be used as a provocative test for overactive bladder? *J Urol* 2001;**165**:1441–5.

96. Goode PS, Locher JL, Bryant RL, Roth DL, Burgio KL. Measurement of postvoid residual urine with portable transabdominal bladder ultrasound scanner and urethral catheterization. *International Urogynecology Journal* 2000; **11**:296–300.

97. Lemack GE, Zimmern PE. Predictability of urodynamic findings based on the Urogenital Distress Inventory-6 questionnaire. *Urology* 1999; **54**:461–6.

98. Chen GD, Su TH, Lin LY. Applicability of perineal sonography in anatomical evaluation of bladder neck in women with and without genuine stress incontinence. *J Clin Ultrasound* 1997;**25**:189–94.

99. Jackson S, Donovan J, Brookes S, Eckford S, Swithinbank L, Abrams P. The Bristol Female Lower Urinary Tract Symptoms questionnaire: development and psychometric testing. *British Journal of Urology* 1996;**77**:805–12.

100. Berglund AL, Lalos O. The pre- and postsurgical nursing of women with stress incontinence. *J Adv Nurs* 1996;**23**:502–11.

101. Nitti VW, Kim Y, Combs AJ. Correlation of the AUA symptom index with urodynamics in patients with suspected benign prostatic hyperplasia. *Neurourol Urodyn* 1994;**13**:521–9.

102. Kiilholma PJ, Makinen JI, Pitkanen YA, Varpula MJ. Perineal ultrasound: an alternative for radiography for evaluating stress urinary

incontinence in females. *Ann Chir Gynaecol Suppl* 1994;**208**:43–5.

103. McInerney PD, Vanner TF, Harris SA, Stephenson TP. Ambulatory urodynamics. *British Journal of Urology* 1991;**67**:272–4.

104. Cucchi A. Acceleration of flow rate as a screening test for detrusor instability in women with stress incontinence. *British Journal of Urology* 1990; **65**:17–19.

105. Versi E, Cardozo LD. Perineal pad weighing versus videographic analysis in genuine stress incontinence. *British Journal of Obstetrics and Gynaecology* 1986;**93**:364–6.

106. Nager CW, Schulz JA, Stanton SL, Monga A. Correlation of urethral closure pressure, leak-point pressure and incontinence severity measures. *Int Urogynecol J Pelvic Floor Dysfunct* 2001;**12**:395–400.

107. Sandvik H, Seim A, Vanvik A, Hunskaar S. A severity index for epidemiological surveys of female urinary incontinence: comparison with 48-hour pad-weighing tests. *Neurourol Urodyn* 2000;**19**:137–45.

108. James M, Jackson S, Shepherd A, Abrams P. Pure stress leakage symptomatology: is it safe to discount detrusor instability? *British Journal of Obstetrics and Gynaecology* 1999;**106**:1255–8.

109. Larsson G, Blixt C, Janson G, Victor A. The frequency/volume chart as a differential diagnostic tool in female urinary incontinence. *International Urogynecology Journal* 1994;**5**:273–7.

110. Davila GW. Ambulatory urodynamics in urge incontinence evaluation. *International Urogynecology Journal* 1994;**5**:25–30.

111. Hahn I, Fall M. Objective quantification of stress urinary incontinence: a short, reproducible, provocative pad-test. *Neurourol Urodyn* 1991; **10**:475–81.

112. Mouritsen L, Berlid G, Hertz J. Comparison of different methods for quantification of urinary leakage in incontinent women. *Neurourol Urodyn* 1989;**8**:579–87.

113. Harvey MA, Kristjansson B, Griffith D, Versi E. The Incontinence Impact Questionnaire and the Urogenital Distress Inventory: a revisit of their validity in women without a urodynamic diagnosis. *Am J Obstet Gynecol* 2001;**185**:25–31.

114. Hanley J, Capewell A, Hagen S. Validity study of the severity index, a simple measure of urinary incontinence in women. *BMJ* 2001;**322**:1096–7.

115. Ryhammer AM, Laurberg S, Djurhuus JC, Hermann AP. No relationship between subjective assessment of urinary incontinence and pad test weight gain in a random population sample of menopausal women. *J Urol* 1998;**159**:800–3.

116. Williams KS, Assassa RP, Smith NKG, Jagger C, Perry S, Shaw C, *et al*. Development, implementation and evaluation of a new nurse-led continence service: a pilot study. *Journal of Clinical Nursing* 2000;**9**:566–73.

117. Cundiff GW, Harris RL, Coates KW, Bump RC. Clinical predictors of urinary incontinence in women. *Am J Obstet Gynecol* 1997;**177**:262–7.

118. De Muylder X, Claes H, Neven P, De Jaegher K. Usefulness of urodynamic investigations in female incontinence. *Eur J Obstet Gynecol Reprod Biol* 1992; **44**:205–8.

119. Diokno AC, Normolle DP, Brown MB, Herzog AR. Urodynamic tests for female geriatric urinary incontinence. *Urology* 1990;**36**:431–9.

120. Diokno AC, Wells TJ, Brink CA. Urinary incontinence in elderly women: urodynamic evaluation. *J Am Geriatr Soc* 1987;**35**:940–6.

121. FitzGerald MP, Brubaker L. Urinary incontinence symptom scores and urodynamic diagnoses. *Neurourol Urodyn* 2002;**21**:30–5.

122. Ishiko O, Hirai K, Sumi T, Nishimura S, Ogita S. The urinary incontinence score in the diagnosis of female urinary incontinence. *Int J Gynaecol Obstet* 2000;**68**:131–7.

123. Korda A, Krieger M, Hunter P, Parkin G. The value of clinical symptoms in the diagnosis of urinary incontinence in the female. *Aust N Z J Obstet Gynaecol* 1987;**27**:149–51.

124. Kujansuu E, Kauppila A. Scored urological history and urethrocystometry in the differential diagnosis of female urinary incontinence. *Ann Chir Gynaecol* 1982;**71**:197–202.

125. Lagro-Janssen AL, Debruyne FM, van Weel C. Value of the patient's case history in diagnosing urinary incontinence in general practice. *British Journal of Urology* 1991;**67**:569–72.

126. Niecestro RM, Wheeler JS, Nanninga J, Einhorn C, Goggin C. Use of stresscath for diagnosing stress incontinence. *Urology* 1992;**39**:266–9.

127. Ouslander J, Staskin D, Raz S, Su HL, Hepps K. Clinical versus urodynamic diagnosis in an incontinent geriatric female population. *J Urol* 1987;**137**:68–71.

128. Ramsay IN, Hilton P, Rice N. The symptomatic characterization of patients with detrusor instability and those with genuine stress incontinence. *International Urogynecology Journal* 1993;**4**:23–6.

129. Sand PK, Hill RC, Ostergard DR. Incontinence history as a predictor of detrusor stability. *Obstet Gynecol* 1988;**71**:257–60.

130. Sandvik H, Hunskaar S, Vanvik A, Bratt H, Seim A, Hermstad R. Diagnostic classification of female urinary incontinence: an epidemiological

survey corrected for validity. *J Clin Epidemiol* 1995; **48**:339–43.

131. Sunshine TJ, Glowacki GA. Clinical correlation of urodynamic testing in patients with urinary incontinence. *Journal of Gynecologic Surgery* 1989; **5**:93–8.

132. Cantor TJ, Bates CP. A comparative study of symptoms and objective urodynamic findings in 214 incontinent women. *British Journal of Obstetrics and Gynaecology* 1980;**87**:889–92.

133. Klovning A, Hunskaar S, Eriksen BC. Validity of a scored urological history in detecting detrusor instability in female urinary incontinence. *Acta Obstet Gynecol Scand* 1996;**75**:941–5.

134. Jorgensen L, Lose G, Andersen JT. One-hour pad-weighing test for objective assessment of female urinary incontinence. *Obstet Gynecol* 1987; **69**:39–42.

135. 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. *British Journal of Obstetrics and Gynaecology* 1996; **103**:162–7.

136. Contreras Ortiz O, Lombardo RJ, Pellicari A. Non-invasive diagnosis of bladder instability using the Bladder Instability Discriminant Index (BIDI). *Zentralbl Gynakol* 1993;**115**:446–9.

137. Bergman A, McCarthy TA, Ballard CA, Yanai J. Role of the Q-tip test in evaluating stress urinary incontinence. *J Reprod Med* 1987;**32**:273–5.

138. Montz FJ, Stanton SL. Q-tip test in female urinary incontinence. *Obstet Gynecol* 1986;**67**:258–60.

139. Dietz HP, McKnoulty L, Clarke B. Translabial color Doppler for imaging in urogynecology: a preliminary report. *Ultrasound Obstet Gynecol* 1999;**14**:144–7.

140. Dietz HP, Wilson PD. Anatomical assessment of the bladder outlet and proximal urethra using ultrasound and videocystourethrography. *Int Urogynecol J Pelvic Floor Dysfunct* 1998;**9**:365–9.

141. Dietz HP, Clarke B. Translabial color Doppler urodynamics. *Int Urogynecol J Pelvic Floor Dysfunct* 2001;**12**:304–7.

142. Quinn MJ, Fanrsworth BA, Pollard WJ, Smith PJB, Stott MA. Vaginal ultrasound in the diagnosis of stress incontinence: a prospective comparison to urodynamic investigations. *Neurourol Urodyn* 1989; **8**:302–3.

143. Bergman A, Ballard CA, Platt LD. Ultrasonic evaluation of urethrovesical junction in women with stress urinary incontinence. *J Clin Ultrasound* 1988;**16**:295–300.

144. Bergman A, McKenzie CJ, Richmond J, Ballard CA, Platt LD. Transrectal ultrasound versus cystography in the evaluation of anatomical stress urinary incontinence. *British Journal of Urology* 1988;**62**:228–34.

145. Pelsang RE, Bonney WW. Voiding cystourethrography in female stress incontinence [see comments]. *AJR Am J Roentgenol* 1996; **166**:561–5.

146. Scotti RJ, Ostergard DR, Guillaume AA, Kohatsu KE. Predictive value of urethroscopy as compared to urodynamics in the diagnosis of genuine stress incontinence. *J Reprod Med* 1990;**35**:772–6.

147. Grischke EM, Anton H, Stolz W, von Fournier D, Bastert G. Urodynamic assessment and lateral urethrocystography. A comparison of two diagnostic procedures for female urinary incontinence. *Acta Obstet Gynecol Scand* 1991; **70**:225–9.

148. Bergman A, McKenzie C, Ballard CA, Richmond J. Role of cystourethrography in the preoperative evaluation of stress urinary incontinence in women. *J Reprod Med* 1988;**33**:372–6.

149. Hsu TH, Rackley RR, Appell RA. The supine stress test: a simple method to detect intrinsic urethral sphincter dysfunction. *J Urol* 1999; **162**:460–3.

150. Kadar N. The value of bladder filling in the clinical detection of urine loss and selection of patients for urodynamic testing. *British Journal of Obstetrics and Gynaecology* 1988;**95**:698–704.

151. Sand PK, Brubaker LT, Novak T. Simple standing incremental cystometry as a screening method for detrusor instability. *Obstet Gynecol* 1991;**77**:453–7.

152. Sand PK, Hill RC, Ostergard DR. Supine urethroscopic and standing cystometry as screening methods for the detection of detrusor instability. *Obstet Gynecol* 1987;**70**:57–60.

153. Sutherst JR, Brown MC. Comparison of single and multichannel cystometry in diagnosing bladder instability. *BMJ* 1984;**288**:1720–2.

154. Fonda D, Brimage PJ, D'Astoli M. Simple screening for urinary incontinence in the elderly: comparison of simple and multichannel cystometry. *Urology* 1993;**42**:536–40.

155. Ouslander J, Leach G, Abelson S, Staskin D, Blaustein J, Raz S. Simple versus multichannel cystometry in the evaluation of bladder function in an incontinent geriatric population. *J Urol* 1988; **140**:1482–6.

156. Resnick NM, Brandeis GH, Baumann MM, DuBeau CE, Yalla SV. Misdiagnosis of urinary incontinence in nursing home women: prevalence and a proposed solution. *Neurourol Urodyn* 1996; **15**:599–618.

157. Davis G, McClure G, Sherman R, Hibbert M, Wong M, Perez R. Ambulatory urodynamics of female soldiers. *Mil Med* 1998;**163**:808–12.

158. Swift SE, Ostergard DR. Evaluation of current urodynamic testing methods in the diagnosis of genuine stress incontinence. *Obstet Gynecol* 1995; **86**:85–91.

159. Richardson DA. Value of the cough pressure profile in the evaluation of patients with stress incontinence. *Am J Obstet Gynecol* 1986;**155**:808–11.

160. Versi E. Discriminant analysis of urethral pressure profilometry data for the diagnosis of genuine stress incontinence. *British Journal of Obstetrics and Gynaecology* 1990;**97**:251–9.

161. Weidner AC, Myers ER, Visco AG, Cundiff GW, Bump RC. Which women with stress incontinence require urodynamic evaluation? *Am J Obstet Gynecol* 2001;**184**(2):20–7.

162. Clarke B. The role of urodynamic assessment in the diagnosis of lower urinary tract disorders. *Int Urogynecol J Pelvic Floor Dysfunct* 1997;**8**:196–9.

163. Bergman A, Bader K. Reliability of the patient's history in the diagnosis of urinary incontinence. *Int J Gynaecol Obstet* 1990;**32**:255–9.

164. Ng RK, Murray A. Can we afford to take short cuts in the management of stress urinary incontinence? *Singapore Med J* 1993;**34**:121–4.

165. Le Coutour X, Jung-Faerber S, Klein P, Renaud R. Female urinary incontinence: comparative value of history and urodynamic investigations. *Eur J Obstet Gynecol Reprod Biol* 1990;**37**:279–86.

166. Petros PP, Ulmsten U. Urge incontinence history is an accurate predictor of urge incontinence. *Acta Obstet Gynecol Scand* 1992;**71**:537–9.

167. Van Doorn ESCVW, Ambergen AW, Janknegt RA. Detrusor activity index: quantification of detrusor overactivity by ambulatory monitoring. *J Urol* 1997;**157**:596–9.

168. Ficazzola MA, Nitti VW. The etiology of post-radical prostatectomy incontinence and correlation of symptoms with urodynamic findings. *J Urol* 1998;**160**:1317–20.

169. Ding YY, Lieu PK, Choo PW. Is the bladder 'an unreliable witness' in elderly males with persistent lower urinary tract symptoms? *Geriatric Nephrology and Urology* 1997;**7**:17–21.

170. Hyman MJ, Groutz A, Blaivas JG. Detrusor instability in men: correlation of lower urinary tract symptoms with urodynamic findings. *J Urol* 2001;**166**:550–3.

171. Porru D, Usai E. Standard and extramural ambulatory urodynamic investigation for the diagnosis of detrusor instability-correlated incontinence and micturition disorders. *Neurourol Urodyn* 1994;**13**:237–42.

172. Gray M, McClain R, Peruggia M, Patrie J, Steers WD. A model for predicting motor urge urinary incontinence. *Nurs Res* 2001;**50**:116–22.

173. de Bolla AR, Arkell DG. Urodynamic investigation in a district general hospital. *Ann R Coll Surg Engl* 1983;**65**:173–5.

174. Robinson D, Pearce KF, Preisser JS, Dugan E, Suggs PK, Cohen SJ. Relationship between patient reports of urinary incontinence symptoms and quality of life measures. *Obstet Gynecol* 1998; **91**:224–8.

175. Uebersax JS, Wyman JF, Shumaker SA, McClish DK, Fantl JA. Short forms to assess life quality and symptom distress for urinary incontinence in women: The incontinence impact questionnaire and the urogenital distress inventory. *Neurourol Urodyn* 1995;**14**:131–9.

176. Gunthorpe W, Brown W, Redman S. The development and evaluation of an incontinence screening questionnaire for female primary care. *Neurourol Urodyn* 2000;**19**:595–607.

177. Haeusler G, Hanzal E, Joura E, Sam C, Koelbl H. Differential diagnosis of detrusor instability and stress-incontinence by patient history: the Gaudenz Incontinence Questionnaire revisited. *Acta Obstet Gynecol Scand* 1995;**74**:635–7.

178. Shumaker SA, Wyman JF, Uebersax JS, McClish D, Fantl JA. Health-related quality of life measures for women with urinary incontinence: the Incontinence Impact Questionnaire and the Urogenital Distress Inventory. Continence Program in Women (CPW) research group. *Qual Life Res* 1994;**3**:291–306.

179. Nitti VW, Kim Y, Combs AJ. Voiding dysfunction following transurethral resection of the prostate: symptoms and urodynamic findings. *J Urol* 1997; **157**:600–3.

180. Hellstrom L, Ekelund P, Larsson M, Milsom I. A comparison between experienced and objectively demonstrated urinary leakage in 85-year old men and women. *Scandinavian Journal of Caring Sciences* 1991;**5**:17–21.

181. Papa Petros PE, Ulmsten U. An analysis of rapid pad testing and the history for the diagnosis of stress incontinence. *Acta Obstet Gynecol Scand* 1992; **71**:529–36.

182. Mayne CJ, Hilton P. Short pad test: standardisation of method and comparison with 1-hour test. *Neurourol Urodyn* 1988;**7**:443–5.

183. Siltberg H, Larsson G, Hallen B, Johansson C, Ulmsten U. Validation of cough-induced leak point pressure measurement in the evaluation of pharmacological treatment of stress incontinence. *Neurourol Urodyn* 1999;**18**:591–602.

184. Miller JM, Ashton-Miller JA, Carchidi LT, DeLancey JO. On the lack of correlation between self-report and urine loss measured with standing provocation test in older stress-incontinent women. *J Womens Health* 1999;**8**:157–62.

185. Robinson D, McClish DK, Wyman JF, Bump RC, Fanti JA. Comparison between urinary diaries completed with and without intensive patient instructions. *Neurourol Urodyn* 1996;**15**:143–8.

186. Robb SS. Urinary incontinence verification in elderly men. *Nurs Res* 1985;**34**:278–82.

187. Fink D, Perucchini D, Schaer GN, Haller U. The role of the frequency–volume chart in the differential diagnostic of female urinary incontinence. *Acta Obstet Gynecol Scand* 1999; **78**:254–7.

188. Romanzi LJ, Polaneczky M, Glazer HI. Simple test of pelvic muscle contraction during pelvic examination: correlation to surface electromyography. *Neurourol Urodyn* 1999; **18**:603–12.

189. Brink CA, Wells TJ, Sampselle CM, Taillie ER, Mayer R. A digital test for pelvic muscle strength in women with urinary incontinence. *Nurs Res* 1994;**43**:352–6.

190. Fischer-Rasmussen W, Hansen RI, Stage P. Predictive values of diagnostic tests in the evaluation of female urinary stress incontinence. *Acta Obstet Gynecol Scand* 1986;**65**:291–4.

191. Karram MM, Bhatia NN. The Q-tip test: standardization of the technique and its interpretation in women with urinary incontinence. *Obstet Gynecol* 1988;**71**:807–11.

192. Walters MD, Diaz K. Q-tip test: a study of continent and incontinent women. *Obstet Gynecol* 1987;**70**:208–11.

193. Resnick NM, Brandeis GH, Baumann MM, Morris JN. Evaluating a national assessment strategy for urinary incontinence in nursing home residents: reliability of the minimum data set and validity of the resident assessment protocol. *Neurourol Urodyn* 1996;**15**:583–98.

194. Eastwood HD, Warrell R. Urinary incontinence in the elderly female: prediction in diagnosis and outcome of management. *Age Ageing* 1984; **13**:230–4.

195. Hilton P, Stanton SL. Algorithmic method for assessing urinary incontinence in elderly women. *BMJ* 1981;**282**:940–2.

196. Summitt RL, Stovall TG, Bent AE, Ostergard DR. Urinary incontinence: correlation of history and brief office evaluation with multichannel urodynamic testing. *Am J Obstet Gynecol* 1992; **166**:1835–44.

197. Griffiths DJ, McCracken PN, Harrison GM, Gormley EA. Characteristics of urinary incontinence in elderly patients studied by 24-hour monitoring and urodynamic testing. *Age Ageing* 1992;**21**:195–201.

198. Creighton SM, Plevnik S, Stanton SL. Distal urethral electrical conductance (DUEC) – a preliminary assessment of its role as a quick screening test for incontinent women. *British Journal of Obstetrics and Gynaecology* 1991;**98**:69–72.

199. Kolbl H, Bernaschek G, Wolf G. A comparative study of perineal ultrasound scanning and urethrocystography in patients with genuine stress incontinence. *Arch Gynecol Obstetr* 1988;**244**:39–45.

200. Rose DH, Eaton AC. Observations in micturating cystourethrography. *J R Soc Med* 1983;**76**:121–5.

201. Lobel RW, Sand PK. The empty supine stress test as a predictor of intrinsic urethral sphincter dysfunction. *Obstet Gynecol* 1996;**88**:128–32.

202. Hebert DB, Ostergard DR. Vesical instability: urodynamic parameters by microtip transducer catheters. *Obstet Gynecol* 1982;**60**:331–7.

203. Rosario DJ, MacDiarmid SA, Radley SC, Chapple CR. A comparison of ambulatory and conventional urodynamic studies in men with borderline outlet obstruction. *British Journal of Urology International* 1999;**83**:400–9.

204. Bhatia NN, Bradley WE, Haldeman S. Urodynamics: continuous monitoring. *J Urol* 1982; **128**:963–8.

205. Webb RJ, Ramsden PD, Neal DE. Ambulatory monitoring and electronic measurement of urinary leakage in the diagnosis of detrusor instability and incontinence. *British Journal of Urology* 1991; **68**:148–52.

206. Pajoncini C, Costantini E, Rociola W, Porena M. The maximum urethral closure pressure and the Valsalva leak point pressure in the diagnosis of intrinsic sphincter deficiency: preliminary results. *Acta Urologica Italica* 1999;**13**:231–5.

207. Versi E, Cardozo L, Cooper DJ. Urethral pressures: analysis of transmission pressure ratios. *British Journal of Urology* 1991;**68**:266–70.

208. Swift SE. The reliability of performing a screening cystometrogram using a fetal monitoring device for the detection of detrusor instability. *Obstet Gynecol* 1997;**89**:708–12.

209. Bergman A, Nguyen H, Koonings PP, Ballard CA. Use of fetal cardiotocographic monitor in the evaluation of urinary incontinence. *Israel Journal of Medical Sciences* 1988;**24**:291–4.

210. Petersen T, Chandiramani V, Fowler CJ. The ice-water test in detrusor hyper-reflexia and bladder instability. *British Journal of Urology* 1997;**79**:163–7.

211. Sutherst JR, Brown M. Detection of urethral incompetence. Erect studies using the fluid-bridge test. *British Journal of Urology* 1981;**53**:360–3.

212. Hanzal E, Berger E, Koelbl H. Reliability of the urethral closure pressure profile during stress in the diagnosis of genuine stress incontinence. *British Journal of Urology* 1991;**68**:369–71.

213. Frigerio L, Ferrari A, Candiani GB. The significance of the stop test in female urinary incontinence. *Diagnostic Gynecology and Obstetrics* 1981;**3**:301–4.

214. Briggs AH, Gray AM. Handling uncertainty when performing economic evaluation of healthcare interventions. *Health Technol Assess* 1999;**3**(2).

215. NHS Reference Costs. Department of Health, 2002. http://www.dh.gov.uk/PublicationsAndStatistics/Publications/PublicationsPolicyAndGuidance/PublicationsPolicyAndGuidanceArticle/fs/en?CONTENT_ID=4069646&chk=vzKK5z. Accessed March 2004.

216. Hopewell S, Clarke M, Lefebvre C, Scherer R. Is handsearching still worthwhile? Results of a Cochrane methodology review. *Proceedings of the XI Cochrane Colloquium*, Barcelona, Spain, 26–31 October 2003.

217. Juni P, Holenstein F, Sterne J, Bartlett C, Egger M. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. *Int J Epidemiol* 2002;**31**:115–23.

218. Egger M, Zellweger-Zahner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. *Lancet* 1997;**350**:326–9.

219. Whiting P, Rutjes AW, Dinnes J, Reitsma J, Bossuyt PM, Kleijnen J. Development and validation of methods for assessing the quality of diagnostic studies. *Health Technol Assess* 2004;**8**(25).

220. Bossuyt P, Reitsma J, Bruns D, Gatsonis CA, Glasziou PP, Irwig LM, *et al*. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003; **326**:41–4.

221. Riley R, Burchill S, Abrams K, Heney D, Lambert PC, Jones DR, *et al*. A systematic review and evaluation of the use of tumour markers in paediatric oncology: Ewing's sarcoma and neuroblastoma. *Health Technol Assess* 2003;**7**(5).

222. Chapple C, Pesce F. An interactive symposium to determine criteria for evaluating new minimally invasive treatments for SUI. *International Continence Society 33rd Annual Meeting*, Florence, Italy, 5–9 October 2003. p. 34.

223. Weber AM, Taylor RJ, Wei JT, Lemack G, Piedmonte MR, Walters MD. The cost-effectiveness of preoperative testing (basic office assessment vs urodynamics) for stress urinary incontinence in women. *British Journal of Urology International* 2002;**89**:356–63.

224. Weber AM, Walters MD. Cost-effectiveness of urodynamic testing before surgery for women with pelvic organ prolapse and stress urinary incontinence. *Am J Obstet Gynecol* 2000; **183**:1338–46.

Appendix 1

Search strategy

MEDLINE

1 exp URODYNAMICS/ or urodynamics.mp.
2 provocation stress test$.mp.
3 frequency volume chart$.mp.
4 urinanalysis.mp.
5 post-void residual volume.mp. [mp=title, abstract, registry number word, mesh subject heading]
6 (mid-stream specimen adj2 urine).mp. [mp=title, abstract, registry number word, mesh subject heading]
7 mssu.mp. [mp=title, abstract, registry number word, mesh subject heading]
8 (pad tests or pad testing or pad test).ti,ab.
9 exp URINALYSIS/ or urinalysis.mp.
10 (mid-stream sampl$ adj2 urine).ti,ab.
11 or/1-10
12 exp "Sensitivity and Specificity"/
13 sensitivity.tw.
14 specificity.tw.
15 DO.xs.
16 ri.fs.
17 du.fs.
18 or/12-17
19 exp Predictive Value of Tests/
20 Reference Values/
21 Reference Standards/
22 ROC Curve/
23 exp Diagnostic Errors/
24 ((sensitivity or specificity) adj25 (test or tests)).ti,ab.
25 (predictive value$ or predictive standard$ or predictive model$).ti,ab.
26 (roc or receiver operat$ characteristic or receiver operat$ curve$).ti,ab.
27 (likelihood ratio$ or likelihood function$).ti,ab.
28 (diagnostic error$ or (errors adj2 diagnosis) or (false adj2 reaction$)).ti,ab.
29 (false positive or false positives or false negative or false negatives).ti,ab.
30 ('gold standard'$ or reference test$ or 'gold standard'$).ti,ab.
31 (criter$ standard$ or criter$ bias or criteria test or criteria tests).ti,ab.
32 (validat$ standard or validat$ test or validat$ tests or validat$ bias).ti,ab.
33 (work-up bias or workup bias or expectation bias or verification bias).ti,ab.
34 ((observer adj2 bias) or inderminate result$).ti,ab.
35 ((observer adj25 different) or observer variation$).ti,ab.
36 Observer Variation/
37 ((interrater or intrarater or observer) adj25 reliability).ti,ab.
38 (intra adj4 reliability).ti,ab.
39 ((accuracy or reliability) adj2 (test or tests or testing or standard or standards)).ti,ab.
40 (performance adj2 (test or tests or testing or standard or standards)).ti,ab.
41 (reference value or reference values or sroc).ti,ab.
42 exp Urinary Incontinence/ or urinary incontinence.mp.
43 urge incontinence.mp.
44 stress incontinence.mp.
45 (leakage and urin$).mp. [mp=title, abstract, registry number word, mesh subject heading]
46 detrusor instability.mp.
47 or/42-46
48 or/19-41
49 48 or 11 or 18
50 47 and 49
51 limit 50 to (human and english language and all adult <19 plus years>)

EMBASE

1 exp "Sensitivity and Specificity"/
2 exp Predictive Value of Tests/
3 Reference Values/
4 Reference Standards/
5 ROC Curve/
6 exp Diagnostic Errors/
7 ((sensitivity or specificity) adj25 (test or tests)).ti,ab.
8 (predictive value$ or predictive standard$ or predictive model$).ti,ab.
9 (roc or receiver operat$ characteristic or receiver operat$ curve$).ti,ab.
10 (likelihood ratio$ or likelihood function$).ti,ab.
11 (diagnostic error$ or (errors adj2 diagnosis) or (false adj2 reaction$)).ti,ab.
12 (false positive or false positives or false negative or false negatives).ti,ab.

13 ('gold standard'$ or reference test$ or 'gold standard'$).ti,ab.
14 (criter$ standard$ or criter$ bias or criteria test or criteria tests).ti,ab.
15 (validat$ standard or validat$ test or validat$ tests or validat$ bias).ti,ab.
16 (work-up bias or workup bias or expectation bias or verification bias).ti,ab.
17 ((observer adj2 bias) or inderminate result$).ti,ab.
18 ((observer adj25 different) or observer variation$).ti,ab.
19 Observer Variation/
20 ((interrater or intrarater or observer) adj25 reliability).ti,ab.
21 (intra adj4 reliability).ti,ab.
22 ((accuracy or reliability) adj2 (test or tests or testing or standard or standards)).ti,ab.
23 (performance adj2 (test or tests or testing or standard or standards)).ti,ab.
24 (reference value or reference values or sroc).ti,ab.
25 or/1-24
26 DO.fs.
27 exp URODYNAMICS/ or urodynamics.mp.
28 exp URINALYSIS/ or urinalysis.mp.
29 (mid stream specimen adj2 urine).mp.
30 (mid stream sampl$ adj2 urine).mp.
31 pad test$.mp.
32 (validat$ adj25 scal$).mp.
33 (stress and provocation and test$).mp. [mp=title, abstract, subject headings, drug trade name, original title, device manufacturer, drug manufacturer name]
34 exp Physical Examination/ or physical examination.mp.
35 or/26-34
36 25 or 35
37 exp Urine Incontinence/ or urinary incontinence.mp.
38 exp Urge Incontinence/ or urge incontinence.mp.
39 exp Stress Incontinence/ or stress incontinence.mp.
40 exp Detrusor Dyssynergia/ or detrusor instability.mp.
41 (leak$ and urin$).mp.
42 or/37-41
43 36 and 42
44 limit 43 to (human and english language)
45 limit 44 to (adult <18 to 64 years> or aged <65+ years>)

CINAHL

1 pa.fs.
2 us.fs.
3 ra.fs.
4 DO.fs.
5 du.fs.
6 exp "Sensitivity and Specificity"/
7 sensitivity.tw.
8 specificity.tw.
9 or/1-8
10 exp Urinary Incontinence/ or urinary incontinence.mp.
11 Stress Incontinence/ or stress incontinence.mp.
12 exp Urge Incontinence/ or urge incontinence.mp.
13 detrusor instability.mp.
14 (leak$ and urin$).mp. [mp=title, cinahl subject heading, abstract, instrumentation]
15 or/10-14
16 9 and 15
17 exp Predictive Value of Tests/
18 Reference Values/ or reference values.mp.
19 roc curve.mp.
20 exp Diagnostic Errors/ or diagnostic errors.mp.
21 (predictive value$ or predictive standard$ or predictive model$).ti,ab.
22 (roc or receiver operat$ characteristic or receiver operat$ curve$).ti,ab.
23 (likelihood ratio$ or likelihood function$).ti,ab.
24 (diagnostic error$ or (errors adj2 diagnosis) or (false adj2 reaction$)).ti,ab.
25 (false positive or false positives or false negative or false negatives).ti,ab.
26 (gold standard$ or reference test$ or gold standard$).ti,ab.
27 (criter$ standard$ or criter$ bias or criteria test or criteria tests).ti,ab.
28 (validat$ standard or validat$ test or validat$ tests or validat$ bias).ti,ab.
29 (work-up bias or workup bias or expectation bias or verification bias).ti,ab.
30 ((observer adj2 bias) or inderminate result$).ti,ab.
31 ((observer adj25 different) or observer variation$).ti,ab.
32 observer variation.mp.
33 ((interrater or intrarater or observer) adj25 reliability).ti,ab.
34 (intra adj4 reliability).ti,ab.
35 ((accuracy or reliability) adj2 (test or tests or testing or standard or standards)).ti,ab.
36 (performance adj2 (test or tests or testing or standard or standards)).ti,ab.
37 (reference value or reference values or sroc).ti,ab.

38 or/17-37
39 15 and 38
40 16 or 39
41 exp URODYNAMICS/ or urodynamics.mp.
42 urinalysis.mp. [mp=title, cinahl subject heading, abstract, instrumentation]
43 (mid stream specimen adj2 urine).mp. [mp=title, cinahl subject heading, abstract, instrumentation]
44 (mid stream sampl$ adj2 urine).mp. [mp=title, cinahl subject heading, abstract, instrumentation]
45 pad test$.mp. [mp=title, cinahl subject heading, abstract, instrumentation]
46 (validat$ adj25 scale$).mp. [mp=title, cinahl subject heading, abstract, instrumentation]
47 stress provocation test$.mp. [mp=title, cinahl subject heading, abstract, instrumentation]
48 provocation stress test$.mp. [mp=title, cinahl subject heading, abstract, instrumentation]
49 physical examination.mp. [mp=title, cinahl subject heading, abstract, instrumentation]
50 or/41-49

52 40 or 51
53 limit 52 to english
54 from 53 keep 1-165

Appendix 2

Quality assessment tool

Item		Yes	No	Unclear
1.	Was the spectrum of patients representative of the patients who will receive the test in practice?			
2.	Were selection criteria clearly described?			
3.	Is the reference standard likely to correctly classify the target condition?			
4.	Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?			
5.	Did the whole sample or a random selection of the sample receive verification using a reference standard of diagnosis?			
6.	Did patients receive the same reference standard regardless of the index test result?			
7.	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?			
8a.	Was the execution of the index test described in sufficient detail to permit replication of the test?			
8b.	Was the execution of the reference standard described in sufficient detail to permit its replication?			
9a.	Were the index test results interpreted without knowledge of the results of the reference standard?			
9b.	Were the reference standard results interpreted without knowledge of the results of the index test?			
10.	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?			
11.	Were uninterpretable/intermediate test results reported?			
12.	Were withdrawals from the study explained?			

Appendix 3

Instructions for quality assessment[219]

Explanation of items included in the quality assessment tool and guide to scoring items

In addition to the quality assessment sheet, please fill in the attached sheet with the requested information on the study sample. In addition to providing useful information for categorising the paper this will also assist you with the quality assessment.

Following the pilot quality assessment some further instructions have been added to assist with the scoring of some of the items. These are included in the blue boxes after the original instructions

General Note: In the pilot study there appeared to be a reluctance to code items as 'unclear'. This is an equally valid response and should be used when appropriate. No papers will be excluded from the review on the basis of quality assessment: the coding of items as unclear is not necessarily a sign of poor quality, only a reflection of a lack of clarity in reporting. This may provide useful recommendations for the reporting of diagnostic studies in the future.

1. Was the spectrum of patients representative of the patients who will receive the test in practice?

a. What is meant by this item

Differences in demographic and clinical features between populations may produce measures of diagnostic accuracy that vary considerably; this is known as spectrum bias. Reported estimates of diagnostic accuracy may have limited clinical applicability (generalisability) if the spectrum of tested patients is not similar to the patients in whom the test will be used in practice. The spectrum of patients refers not only to the severity of the underlying target condition, but also to demographic features and to the presence of differential diagnosis and/or co-morbidity. It is therefore important that diagnostic test evaluations include an appropriate spectrum of patients for the test under investigation and that a clear definition of the characteristics of the included patients is provided.

b. Situations in which this item does not apply

This item is relevant to all studies of diagnostic accuracy and should always be included in the quality assessment tool.

c. How to score this item

Studies should score 'yes' for this item if you believe, based on the information reported or obtained from the study's authors, that the spectrum of patients included in the study was representative of those in whom the test will be used in practice. The judgement should be based on both the method of recruitment and the characteristics of those recruited. Studies which recruit a group of healthy controls and a group known to have the target disorder will be coded as 'no' on this item in nearly all circumstances. Reviewers should prespecify in the protocol of the review what spectrum of patients would be acceptable taking factors such as disease prevalence and severity, age and sex into account. If you think that the population studied does not fit into what you specified as acceptable, the study should be scored as 'no'. If there is insufficient information available to make a judgement then it should be scored as 'unclear'.

Additional instructions for Question 1:
It is not necessary for the study sample to be statistically representative of all the patients who may receive the test in practice. The study should include a sample that meets the broad remits of the study:

A sample of community-dwelling adults not exclusively consisting of patients with a related chronic disease.

Therefore, the sample does not have to consist of both men and women, to include a wide range of age groups or include a primary and secondary care population to be coded as 'yes'.

2. Were selection criteria clearly described?

a. What is meant by this item

This refers to whether studies have provided a clear definition of the criteria used as selection criteria for entry into the study.

b. Situations in which this item does not apply

This item is relevant to all studies of diagnostic accuracy and should always be included in the quality assessment tool.

c. How to score this item

If you think that all relevant information regarding how participants were selected for inclusion in the study has been provided then this item should be scored as 'yes'. If study selection criteria are not clearly reported then this item should be scored as 'no'. In situations where selection criteria are partially reported and you feel that you do not have enough information to score this item as 'yes', then it should be scored as 'unclear'.

In order for this to be coded as 'yes' the description of the sample needs to fulfil all of these criteria:
Age: either an age range or a measure of central tendency (with SD) should be presented. If a statement such as 'women over the age of 50' is the only description then this item should be coded as 'unclear'.
Gender: the proportion of male and female patients must be stated.
Location of recruitment and test: the paper should state where recruitment of patients took place and whether the tests were performed in primary or secondary care.
Sample size

3. Is the reference standard likely to correctly classify the target condition?

a. What is meant by this item

The reference standard is the method used to determine the presence or absence of the target condition. To assess the diagnostic accuracy of the index test its results are compared with the results of the reference standard; subsequently indicators of diagnostic accuracy can be calculated. The reference standard is therefore an important determinant of the diagnostic accuracy of a test. The reference standard may be obtained in many ways, including laboratory tests, imaging tests, function tests and pathology, but also clinical follow-up of participants. The decision of which reference standard to use depends on the definition of the target condition and the purpose of the study. If no single reference test is available, then careful clinical follow-up, a consensus between observers or results of two or more combined tests may be used to determine the presence or absence of the target condition. Estimates of test performance are based on the assumption that the index test is being compared to a reference standard which is 100% sensitive and specific. If there are any disagreements between the reference standard and the index test then it is assumed that the index test is incorrect. Thus, from a theoretical point of view the choice of an appropriate reference standard is very important.

b. Situations in which this item does not apply

This item is relevant to all studies of diagnostic accuracy and should always be included in the quality assessment tool. The only exception would be if a particular reference standard is specified in the inclusion criteria, i.e. to be included in the review a study may have to compare the index test to a specified reference standard.

c. How to score this item

If you believe that the reference standard is likely to correctly classify the target condition then this item should be scored 'yes'. Making a judgement as to the accuracy of the reference standard may not be straightforward. You may need experience of the topic area to know whether a test is an appropriate reference standard, or if a combination of tests is used you may have to consider carefully whether these were appropriate. If you do not think that the reference standard was likely to have correctly classified the target condition then this item should be scored as 'no'. If there is insufficient information to make a judgement then this should be scored as 'unclear'.

If urodynamics or an ICS approved pad test is used as the reference standard then this item should be coded as 'yes'.

4. Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?

a. What is meant by this item

Ideally the results of the index test and the reference standard are collected on the same patients at the same time. If this is not possible and a delay occurs, misclassification due to spontaneous recovery or a more advanced stage of disease may occur. This is known as disease progression bias. The size of the time period which may cause such bias will vary between conditions. For example, a delay of a few days is unlikely to be a problem for chronic conditions; however, for other infectious diseases a delay between performance of index and reference standard of only a few days may be important. This type of bias may occur in chronic conditions in which the reference standard involves clinical follow-up of several years.

b. Situations in which this item does not apply

This item is likely to apply in most situations.

c. How to score this item

When to score this item as 'yes' is related to the target condition. For conditions that progress rapidly even a delay of several days may be important. For such conditions this item should be scored 'yes' if the delay between the performance of the index and reference standard is very short, a matter of hours or days. However, for chronic conditions disease status is unlikely to change in a week, or a month, or even longer. In such conditions longer delays between performance of the index and reference standard may be scored as 'yes'. You will have to make judgements regarding what is considered 'short enough'. You should think about this before starting work on a review, and define what you consider to be 'short enough' for the specific topic area that you are reviewing. If you think the time period between the performance of the index test and the reference standard was sufficiently long that disease status may have changed between the performance of the two tests then this item should be scored as 'no'. If insufficient information is provided this should be scored as 'unclear'.

Some disagreement resulted from variations in the strictness of coding for this item. It is rare that time periods are explicitly presented and in some cases people made the (probably correct) assumption that the two tests were carried out at around the same time. It has been decided that no assumptions should be made when performing the quality assessment. Therefore, if there is no mention of the time period between tests then this item should always be coded as 'unclear'.

5. Did the whole sample or a random selection of the sample, receive verification using a reference standard?

a. What is meant by this item

Partial verification bias (also known as work-up bias, (primary) selection bias or sequential ordering bias) occurs when not all of the study group receive confirmation of the diagnosis by a reference standard. If the results of the index test influence the decision to perform the reference standard then biased estimates of test performance may arise. If patients are randomly selected to receive the reference standard the overall diagnostic performance of the test is, in theory, unchanged. In most cases, however, this selection is not random, possibly leading to biased estimates of the overall diagnostic accuracy.

b. Situations in which this item does not apply

Partial verification bias generally only occurs in diagnostic cohort studies in which patients are tested by the index test prior to the reference standard. If the test sequence is reversed, as it is in case–control designs, partial verification bias is generally not applicable. However, there may be exceptions to this. For example, in radiologic re-reading studies, scans are read at a later data by one or more radiologists, but the scans will usually have been obtained in regular clinical practice. If the study is limited to those with, for example, biopsy verification the index (radiological interpretations) could by influenced by the decision to biopsy or not, and verification bias may apply. In situations where the reference standard is assessed before the index test, you should first decide whether there is a possibility that verification bias could occur, and if not how to score this item. This may depend on how quality will be incorporated in the review. There are two options: either to score this item as 'yes', or to remove it from the quality assessment tool.

c. How to score this item

If it is clear from the study that all patients who received the index test went on to receive verification of their disease status using a reference standard, even if this reference standard was not the same for all patients, then this item should be scored as 'yes'. If some of the patients who received the index test did not receive verification of their true disease state then this item should be scored as 'no'. If this information is not reported by the study then it should be scored as 'unclear'.

6. Did patients receive the same reference standard regardless of the index test result?

a. What is meant by this item

Differential verification bias occurs when some of the index test results are verified by a different reference standard. This is especially a problem if these reference standards differ in their definition of the target condition, for example histopathology of the appendix and natural history for the detection of appendicitis. This usually occurs when patients testing positive on the index test receive a more accurate, often invasive, reference standard than those with negative test results. The link (correlation) between a particular (negative) test result and being verified by a less accurate reference standard will affect measures of test accuracy in a similar way as in partial verification, but less seriously.

b. Situations in which this item does not apply

Differential verification bias generally only occurs in diagnostic cohort studies in which all patients are tested by the index test prior to the reference standard. However, there may be situations in which this does not apply (see Item 3). If the test sequence is reversed, as it is in case–control designs, partial verification bias is not applicable. In situations where the reference standard is assessed before the index test, you should decide how to score this item. This may depend on how quality will be incorporated in the review. There are two options: either to score this item as 'yes', or to remove it from the quality assessment tool.

c. How to score this item

If it is clear that patients received verification of their true disease status using the same reference standard then this item should be scored as 'yes'. If some patients received verification using a different reference standard this item should be scored as 'no'. If this information is not reported by the study then it should be scored as 'unclear'.

7. Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?

a. What is meant by this item

When the result of the index test is used in establishing the final diagnosis, incorporation bias may occur. This incorporation will probably increase the amount of agreement between index test results and the outcome of the reference standard, and hence overestimate the various measures of diagnostic accuracy. It is important to note that knowledge of the results of the index test alone does not automatically mean that these results are incorporated in the reference standard. For example, a study investigating MRI for the diagnosis of multiple sclerosis could have a reference standard composed of clinical follow-up, cerebrospinal fluid analysis and MRI. In this case the index test forms part of the reference standard. If the same study used a reference standard of clinical follow-up and the results of the MRI were known when the clinical diagnosis was made but were not specifically included as part of the reference then the index test does not form part of the reference standard.

b. Situations in which this item does not apply

This item will only apply when a composite reference standard is used to verify disease status. In such cases it is essential that a full definition of how disease status is verified and which tests form part of the reference standard are provided. For studies in which a single reference standard is used this item will not be relevant and should either be scored as 'yes' or be removed from the quality assessment tool.

c. How to score this item

If it is clear from the study that the index test did not form part of the reference standard then this item should be scored as 'yes'. If it appears that the index test formed part of the reference standard then this item should be scored as 'no'. If this information is not reported by the study then it should be scored as 'unclear'.

8a. Was the execution of the index test described in sufficient detail to permit replication of the test?
8b. Was the execution of the reference standard described in sufficient detail to permit its replication?

a. What is meant by these items

A sufficient description of the execution of index test and reference standards is important for two reasons. First, variation in measures of diagnostic accuracy can sometimes be traced back to differences in the execution of index/reference standards. Second, a clear and detailed description (or references) is needed to implement a certain test in another setting. If tests are executed in different ways then this would be expected to impact on test performance. The

extent to which this would be expected to affect results would depend on the type of test being investigated.

b. Situations in which these items do not apply

These items are likely to apply in most situations.

c. How to score these items

If the study reports sufficient details to permit replication of the index test and reference standard then these items should be scored as 'yes'. In other cases these items should be scored as 'no'. In situations where details of test performance are partially reported and you feel that you do not have enough information to score this item as 'yes', then it should be scored as 'unclear'.

If the paper cites a reference for a full description of the methodology then this item should be coded as 'yes'.

For a description of urodynamics to be coded as 'yes' the following information needs to be given:
- **what type of catheter is used**
- **filling speed**
- **volume and type of medium (fluid, gas, etc.).**

9a. Were the index test results interpreted without knowledge of the results of the reference standard?
9b. Were the reference standard results interpreted without knowledge of the results of the index test?

a. What is meant by these items

This item is similar to 'blinding' in intervention studies. Interpretation of the results of the index test may be influenced by knowledge of the results of the reference standard, and vice versa. This is known as review bias, and may lead to inflated measures of diagnostic accuracy. The extent to which this may affect test results will be related to the degree of subjectiveness in the interpretation of the test result. The more subjective the interpretation the more likely that the interpreter can be influenced by the results of the index test in interpreting the reference standard, and vice versa. It is therefore important to consider the topic area that you are reviewing and to determine whether the interpretation of the index test or reference standard could be influenced by knowledge of the results of the other test.

b. Situations in which these items do not apply

If, in the topic area that you are reviewing, the index test is always performed first then interpretation of the results of the index test will usually be without knowledge of the results of the reference standard. Similarly, if the reference standard is always performed first (for example, in a diagnostic case–control study) then the results of the reference standard will be interpreted without knowledge of the index test. However, in certain situations the results of both the index test and reference standard are blinded in both directions before being interpreted. In situations where one form of review bias does not apply there are two possibilities: either score the relevant item as 'yes' or remove this item from the list. If tests are entirely objective in their interpretation then test interpretation is not susceptible to review bias. In such situations review bias may not be a problem and these items can be omitted from the quality assessment tool. Another situation in which this form of bias may not apply is when test results are interpreted in an independent laboratory. In such situations it is unlikely that the person interpreting the test results will have knowledge of the results of the other test (either index test or reference standard).

c. How to score these items

If the study clearly states that the test results (index or reference standard) were interpreted blind to the results of the other test then these items should be scored as 'yes'. If this does not appear to be the case they should be scored as 'no'. If this information is not reported by the study then it should be scored as 'unclear'.

This is also rarely explicitly mentioned although it could be assumed that when performing urodynamics some history of the patient will be known. However, no assumptions should be made and therefore the item should be coded thus:
- **If there is mention of blinding or independent interpretation – 'yes'**
- **If it is mentioned that the tests are not blinded – 'no'**
- **If blinding is not mentioned at all – 'unclear'**

10. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?

a. What is meant by this item
The availability of information on clinical data during interpretation of test results may affect estimates of test performance. In this context clinical data is defined broadly to include any information relating to the patient obtained by direct observation such as age, sex and symptoms. The knowledge of such factors can influence the diagnostic test result if the test involves an interpretative component. If clinical data will be available when the test is interpreted in practice then this should also be available when the test is evaluated. If, however, the index test is intended to replace other clinical tests then clinical data should not be available. It is therefore important to determine what information will be available when test results are interpreted in practice before assessing studies for this item.

b. Situations in which this item does not apply
If the interpretation of the index test is fully automated and involves no interpretation then this item may not be relevant and can be omitted from the quality assessment tool.

c. How to score this item
If clinical data would normally be available when the test is interpreted in practice and similar data were available when interpreting the index test in the study then this item should be scored as 'yes'. Similarly, if clinical data would not be available in practice and these data were not available when the index test results were interpreted then this item should be scored as 'yes'. If this is not the case then this item should be scored as 'no'. If this information is not reported by the study then it should be scored as 'unclear'.

11. Were uninterpretable/intermediate test results reported?

a. What is meant by this item
A diagnostic test can produce an uninterpretable/indeterminate/intermediate result with varying frequency depending on the test. These problems are often not reported in diagnostic accuracy studies, with the uninterpretable results simply removed from the analysis. This may lead to the biased assessment of the test characteristics. Whether bias will arise depends on the possible correlation between uninterpretable test results and the true disease status. If uninterpretable results occur randomly and are not related to the true disease status of the individual then, in theory, these should not have any effect on test performance. Whatever the cause of uninterpretable results it is important that these are reported so that the impact of these results on test performance can be determined.

b. Situations in which this item does not apply
This item is relevant to all studies of diagnostic accuracy and should always be included in the quality assessment tool.

c. How to score this item
If it is clear that all test results, including uninterpretable/indeterminate/intermediate, are reported then this item should be scored as 'yes'. If you think that such results occurred but have not been reported then this item should be scored as 'no'. If it is not clear whether all study results have been reported then this item should be scored as 'unclear'.

A strict approach should be used when coding this item. If there is no mention of any uninterpretable results then this should be coded as 'unclear'.

12. Were withdrawals from the study explained?

a. What is meant by this item
This occurs when patients withdraw from the study before the results of both the index test and reference standard are known. If patients lost to follow-up differ systematically from those who remain, for whatever reason, then estimates of test performance may be biased.

b. Situations in which this item does not apply
This item is relevant to all studies of diagnostic accuracy and should always be included in the quality assessment tool.

c. How to score this item
If it is clear what happened to all patients who entered the study, for example if a flow diagram of study participants is reported, then this item should be scored as 'yes'. If it appears that some of the participants who entered the study did not complete the study, i.e. did not receive both the index test and reference standard, and these patients were not accounted for then this item should be scored as 'no'. If it is not clear whether all patients who entered the study were accounted for then this item should be scored as 'unclear'.

Again a strict approach should be used when coding this item. If there is no mention of any withdrawals then this should be coded as 'unclear'.

Appendix 4

Letter to authors requesting additional data

Dear

We are currently undertaking a systematic review on the methods of diagnosing urinary incontinence. This work is funded by the Department of Health in the United Kingdom (http://www.hta.nhsweb.nhs.uk/). The results will be used to advise health care professionals on the most appropriate assessment methods when dealing with this highly prevalent condition.

We have identified your paper {*InsertReference*} as relevant for inclusion in the review as it quantitatively compares the diagnostic methods: {*insert diagnostic test 1 and diagnostic 2*}.

However, in order to be able to fully include your paper in the review and any meta-analysis we need a little further information from you. Combining data from different studies in a meta-analysis requires data in a very specific format. In order that we can include the results from all possible studies in the meta-analysis we are writing to authors for this extra information. As I am sure you are aware the very nature of systematic reviews requires as many relevant papers as possible to be included in order to provide representative results[1].

We need to know the number of patients (both with and without urinary incontinence) classified correctly and incorrectly by the index test (e.g. a 2×2 or 3×3 contingency table). This would allow us to calculate sensitivity, specificity and positive predictive value for the index test. The cut-off points used to determine a positive result for each of the diagnostic tests are also required. If your study did not define cut-off points for a positive test then it would be most useful if you could provide us with the raw data, we only require two columns of data (please see our website http://www.prw.le.ac.uk/research/hta/ for an example of what we require). To minimise the effort on your part we have attached a 'fax-back' form that you can complete by hand with the required data (which is potentially just six numbers), and our website will hopefully answer any additional queries relating to this request.

We do hope that you will be able to assist us with this request, your help will greatly improve the validity of the review and maximise its impact. You will of course be acknowledged for your assistance and sent a copy of the final report. As I am sure you can appreciate we are on a very tight timetable, therefore a response within two weeks would be greatly appreciated. However, if you are going to find this difficult please contact us.

If you require any further information about what data is required or have any questions about any aspect of the project please do not hesitate to contact us by any of the contact methods given above.

Yours sincerely

Jennifer Martin
On behalf of the study team

1. Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *BMJ* **323**(7305):157–62. www.bmj.com

Appendix 5

Blank forms sent to contacted authors

Fax

F.A.O: Dr Jennifer Martin **From:**

Fax No. +44 (0)116 252 5423 **Phone No.** +44 (0)116 252 5451

Re: Systematic Review Data

Author _____ Paper ID _____

Data Required:

		Gold Standard/Reference Test	
		+ ve	– ve
Index Test	+ ve		
	– ve		

Cut-off for a positive result on the gold standard test _____

Cut-off for a positive result on the index standard test _____

Fax

F.A.O: Dr Jennifer Martin **From:**

Fax No. +44 (0)116 252 5423 **Phone No.** +44 (0)116 252 5451

Re: Systematic Review Data

Author _____ Paper ID _____

Data Required:

	Gold Standard/Reference Test _____ _____ _____
Index Test _____	
Index Test _____	
Index Test _____	

Cut-off for a positive result on the gold standard test _____

Cut-off for a positive result on the index standard test _____

Fax

F.A.O: Dr Jennifer Martin **From:**

Fax No. +44 (0)116 252 5423 **Phone No.** +44 (0)116 252 5451

Re: Systematic Review Data

Paper ID

Data Required:

Patient No.	Gold Standard Diagnostic Test	Reference Diagnostic Test
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		

If you would rather send data in electronic form by email or send an existing data sheet by fax then please do so.

Appendix 6

Website created for contacted authors

Systematic review: methods of diagnosing urinary incontinence

Data examples

This website has been created to provide assistance to those authors that have been contacted for extra data to be included in our systematic review on methods of diagnosing urinary incontinence. This work is funded by the Department of Health. http://www.hta.nhsweb.nhs.uk

We have provided a number of examples to illustrate the form in which we require the data. These illustrate what to do in situations where cut-off points have been used to classify patients as either positive or negative on a particular test (i.e. categorical data). Examples are also given for studies where no cut-offs have been used and therefore data is in a continuous form.

In order to comply with the Data Protection Act 1998 please do not send us any unique patient identifier numbers, initials or any other information that could be used to identify individuals.

We hope that these examples will enable you to provide the data that we have asked for. However, if you have any questions whatsoever about what is required or indeed any queries about the project in general please do not hesitate to get in contact with us.

Contact email: jlm26@le.ac.uk

Example 1

This study was undertaken to determine the diagnostic accuracy of the 48-hour pad-test against the gold standard test of multichannel video urodynamics. 38 patients performed both tests. A clear cut-off point was defined for a positive result for each of the diagnostic tests. Each of the 38 patients can be assigned to one of the 4 boxes within the contingency table.

		Multichannel video urodynamics (Gold standard/reference test) + ve leakage	– ve no leakage
48 h pad test (index test)	+ ve	10	6
	– ve	4	18

Cut-off for a positive result for multichannel videourodynamics	Visualisation of leakage in absence of a detrusor contraction
Cut-off for a positive result on 48-hour pad-test	Leakage greater than 15 g

Example 2

This study was undertaken to determine the accuracy of a clinical stress test in diagnosing different types of incontinence compared with the gold standard of multichannel videourodynamics. A total of 34 patients performed both tests. Cut-off points were defined for each diagnosis. Each of the 34 patients can be assigned to one of the 9 boxes within the contingency table.

		Multichannel videourodynamics (Gold standard/reference test)		
		USI	DO	Normal
Clinical stress test (index test)	USI	17	2	1
	DO	1	8	0
	Normal	1	2	2

	USI	DO
Cut-off for a positive result on multichannel videourodynamics	Involuntary leakage during increased abdominal pressure in the absence of a detrusor contraction	Spontaneous contraction whilst the patient attempts to inhibit micturition
Cut-off for a positive result on clinical stress test	Observed leakage coincidentally with coughing or straining	Uncontrollable leakage during examination

Example 3

A total of 20 patients were studied to investigate the accuracy of using a severity index to diagnose urinary incontinence. The scale was compared with a 48-hour pad-test, which had a clear cut-off point for a positive or negative result. No cut-off point was used for the severity score therefore the raw data is given.

Patient no.	Pad test result	Severity score
1	Positive	14
2	Negative	3
3	Positive	18
4	Positive	16
5	Positive	11
6	Negative	5
7	Negative	7
8	Positive	9
9	Positive	11
10	Negative	2
11	Negative	0
12	Positive	7
13	Positive	9
14	Positive	12
15	Positive	15
16	Positive	13
17	Negative	11
18	Positive	16
19	Positive	18
20	Negative	13

Cut-off point for a positive 48-hour pad test = 15 g.

Appendix 7

Additional study information sheet

Description of Study Sample Paper No. _____

1. Age of patients*		Range/measure of central tendency
2. Gender*		% Female
3. Where sample was recruited*		Primary/2ndary/Mixed
4. Where tests were performed*		Primary/2ndary/Mixed
5. Community dwelling?		%
6. Proportion of patients with related chronic disease		%
7. Year of publication		
8. Sample size*		
9. Country of study		

* This information is required for Q2 to be coded as 'Yes'.

Currently this paper is classified as comparing the following tests.

Do you agree with this classification? (please circle) YES NO

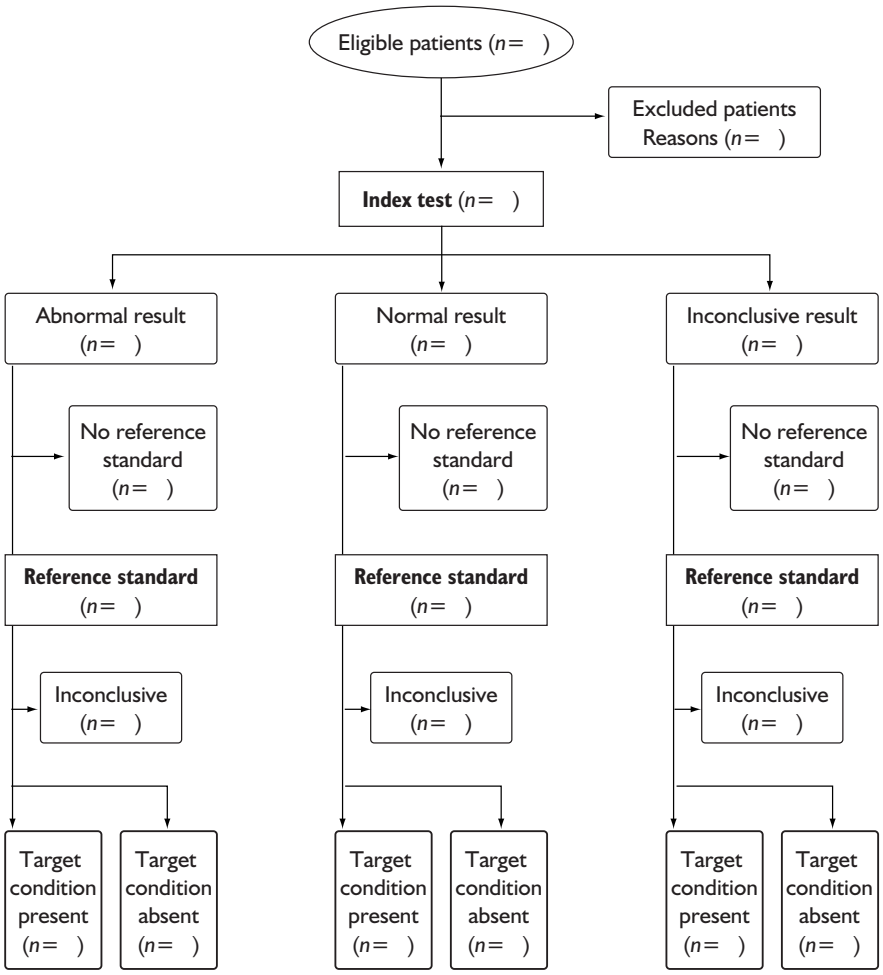
If not, how would you classify the paper?

_____

_____

Appendix 8

STARD flowchart and checklist



STARD checklist for reporting diagnostic accuracy studies

Section and topic	Item	Description
Title, abstract, and keywords	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading "sensitivity and specificity")
Introduction	2	State the research questions or aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups
Methods:		
Participants	3	Describe the study population: the inclusion and exclusion criteria and the settings and locations where the data were collected
	4	Describe participant recruitment: was this based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?
	5	Describe participant sampling: was this a consecutive series of participants defined by selection criteria in items 3 and 4? If not, specify how participants were further selected
	6	Describe data collection: was data collection planned before the index tests and reference standard were performed (prospective study) or after (retrospective study)?
Test methods	7	Describe the reference standard and its rationale
	8	Describe technical specifications of material and methods involved, including how and when measurements were taken, or cite references for index tests or reference standard, or both
	9	Describe definition of and rationale for the units, cut-off points, or categories of the results of the index tests and the reference standard
	10	Describe the number, training, and expertise of the persons executing and reading the index tests and the reference standard
	11	Were the readers of the index tests and the reference standard blind (masked) to the results of the other test? Describe any other clinical information available to the readers.
Statistical methods	12	Describe methods for calculating or comparing measures of diagnostic accuracy and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals)
	13	Describe methods for calculating test reproducibility, if done
Results:		
Participants	14	Report when study was done, including beginning and ending dates of recruitment
	15	Report clinical and demographic characteristics (e.g. age, sex, spectrum of presenting symptoms, comorbidity, current treatments, and recruitment centre)
	16	Report how many participants satisfying the criteria for inclusion did or did not undergo the index tests or the reference standard, or both; describe why participants failed to receive either test (a flow diagram is strongly recommended)
Test results	17	Report time interval from index tests to reference standard, and any treatment administered between
	18	Report distribution of severity of disease (define criteria) in those with the target condition and other diagnoses in participants without the target condition
	19	Report a cross-tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, report the distribution of the test results by the results of the reference standard
	20	Report any adverse events from performing the index test or the reference standard
Estimates	21	Report estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals)
	22	Report how indeterminate results, missing responses, and outliers of index tests were handled
	23	Report estimates of variability of diagnostic accuracy between readers, centres, or subgroups of participants, if done
	24	Report estimates of test reproducibility, if done
Discussion	25	Discuss the clinical applicability of the study findings

Health Technology Assessment Programme

Prioritisation Strategy Group

Members

HTA Commissioning Board

Members

Current and past membership details of all HTA 'committees' are available from the HTA website (www.ncchta.org)

Diagnostic Technologies & Screening Panel

Members

Chair,
Dr Ron Zimmern, Director of the Public Health Genetics Unit, Strangeways Research Laboratories, Cambridge

Ms Norma Armston, Lay Member, Bolton

Professor Max Bachmann Professor of Health Care Interfaces, Department of Health Policy and Practice, University of East Anglia

Professor Rudy Bilous Professor of Clinical Medicine & Consultant Physician, The Academic Centre, South Tees Hospitals NHS Trust

Dr Paul Cockcroft, Consultant Medical Microbiologist and Clinical Director of Pathology, Department of Clinical Microbiology, St Mary's Hospital, Portsmouth

Professor Adrian K Dixon, Professor of Radiology, University Department of Radiology, University of Cambridge Clinical School

Dr David Elliman, Consultant Paediatrician/ Hon. Senior Lecturer, Population Health Unit, Great Ormond St. Hospital, London

Professor Glyn Elwyn, Primary Medical Care Research Group, Swansea Clinical School, University of Wales Swansea

Mr Tam Fry, Honorary Chairman, Child Growth Foundation, London

Dr Jennifer J Kurinczuk, Consultant Clinical Epidemiologist, National Perinatal Epidemiology Unit, Oxford

Dr Susanne M Ludgate, Medical Director, Medicines & Healthcare Products Regulatory Agency, London

Professor William Rosenberg, Professor of Hepatology, Liver Research Group, University of Southampton

Dr Susan Schonfield, Consultant in Public Health, Specialised Services Commissioning North West London, Hillingdon Primary Care Trust

Dr Phil Shackley, Senior Lecturer in Health Economics, School of Population and Health Sciences, University of Newcastle upon Tyne

Dr Margaret Somerville, PMS Public Health Lead, Peninsula Medical School, University of Plymouth

Dr Graham Taylor, Scientific Director & Senior Lecturer, Regional DNA Laboratory, The Leeds Teaching Hospitals

Professor Lindsay Wilson Turnbull, Scientific Director, Centre for MR Investigations & YCR Professor of Radiology, University of Hull

Professor Martin J Whittle, Associate Dean for Education, Head of Department of Obstetrics and Gynaecology, University of Birmingham

Dr Dennis Wright, Consultant Biochemist & Clinical Director, Pathology & The Kennedy Galton Centre, Northwick Park & St Mark's Hospitals, Harrow

Pharmaceuticals Panel

Members

Chair,
Dr John Reynolds, Chair Division A, The John Radcliffe Hospital, Oxford Radcliffe Hospitals NHS Trust

Professor Tony Avery, Head of Division of Primary Care, School of Community Health Services, Division of General Practice, University of Nottingham

Ms Anne Baileff, Consultant Nurse in First Contact Care, Southampton City Primary Care Trust, University of Southampton

Professor Stirling Bryan, Professor of Health Economics, Health Services Management Centre, University of Birmingham

Mr Peter Cardy, Chief Executive, Macmillan Cancer Relief, London

Professor Imti Choonara, Professor in Child Health, Academic Division of Child Health, University of Nottingham

Dr Robin Ferner, Consultant Physician and Director, West Midlands Centre for Adverse Drug Reactions, City Hospital NHS Trust, Birmingham

Dr Karen A Fitzgerald, Consultant in Pharmaceutical Public Health, National Public Health Service for Wales, Cardiff

Mrs Sharon Hart, Head of DTB Publications, *Drug & Therapeutics Bulletin*, London

Dr Christine Hine, Consultant in Public Health Medicine, South Gloucestershire Primary Care Trust

Professor Stan Kaye, Cancer Research UK Professor of Medical Oncology, Section of Medicine, The Royal Marsden Hospital, Sutton

Ms Barbara Meredith, Lay Member, Epsom

Dr Andrew Prentice, Senior Lecturer and Consultant Obstetrician & Gynaecologist, Department of Obstetrics & Gynaecology, University of Cambridge

Dr Frances Rotblat, CPMP Delegate, Medicines & Healthcare Products Regulatory Agency, London

Professor Jan Scott, Professor of Psychological Treatments, Institute of Psychiatry, University of London

Mrs Katrina Simister, Assistant Director New Medicines, National Prescribing Centre, Liverpool

Dr Richard Tiner, Medical Director, Medical Department, Association of the British Pharmaceutical Industry, London

Dr Helen Williams, Consultant Microbiologist, Norfolk & Norwich University Hospital NHS Trust

Therapeutic Procedures Panel

Members

Chair,
Professor Bruce Campbell,
Consultant Vascular and General Surgeon, Department of Surgery, Royal Devon & Exeter Hospital

Dr Aileen Clarke,
Reader in Health Services Research, Public Health & Policy Research Unit, Barts & the London School of Medicine & Dentistry, London

Dr Matthew Cooke, Reader in A&E/Department of Health Advisor in A&E, Warwick Emergency Care and Rehabilitation, University of Warwick

Dr Carl E Counsell, Clinical Senior Lecturer in Neurology, Department of Medicine and Therapeutics, University of Aberdeen

Ms Amelia Curwen, Executive Director of Policy, Services and Research, Asthma UK, London

Professor Gene Feder, Professor of Primary Care R&D, Department of General Practice and Primary Care, Barts & the London, Queen Mary's School of Medicine and Dentistry, London

Professor Paul Gregg, Professor of Orthopaedic Surgical Science, Department of General Practice and Primary Care, South Tees Hospital NHS Trust, Middlesbrough

Ms Bec Hanley, Co-Director, TwoCan Associates, Hurstpierpoint

Ms Maryann L Hardy, Lecturer, Division of Radiography, University of Bradford

Professor Alan Horwich, Director of Clinical R&D, Academic Department of Radiology, The Institute of Cancer Research, London

Dr Simon de Lusignan, Senior Lecturer, Primary Care Informatics, Department of Community Health Sciences, St George's Hospital Medical School, London

Professor Neil McIntosh, Edward Clark Professor of Child Life & Health, Department of Child Life & Health, University of Edinburgh

Professor James Neilson, Professor of Obstetrics and Gynaecology, Department of Obstetrics and Gynaecology, University of Liverpool

Dr John C Pounsford, Consultant Physician, Directorate of Medical Services, North Bristol NHS Trust

Karen Roberts, Nurse Consultant, Queen Elizabeth Hospital, Gateshead

Dr Vimal Sharma, Consultant Psychiatrist/Hon. Senior Lecturer, Mental Health Resource Centre, Cheshire and Wirral Partnership NHS Trust, Wallasey

Dr L David Smith, Consultant Cardiologist, Royal Devon & Exeter Hospital

Professor Norman Waugh, Professor of Public Health, Department of Public Health, University of Aberdeen

Expert Advisory Network

Members

Professor Douglas Altman, Director of CSM & Cancer Research UK Med Stat Gp, Centre for Statistics in Medicine, University of Oxford, Institute of Health Sciences, Headington, Oxford

Professor John Bond, Director, Centre for Health Services Research, University of Newcastle upon Tyne, School of Population & Health Sciences, Newcastle upon Tyne

Mr Shaun Brogan, Chief Executive, Ridgeway Primary Care Group, Aylesbury

Mrs Stella Burnside OBE, Chief Executive, Office of the Chief Executive. Trust Headquarters, Altnagelvin Hospitals Health & Social Services Trust, Altnagelvin Area Hospital, Londonderry

Ms Tracy Bury, Project Manager, World Confederation for Physical Therapy, London

Professor Iain T Cameron, Professor of Obstetrics and Gynaecology and Head of the School of Medicine, University of Southampton

Dr Christine Clark, Medical Writer & Consultant Pharmacist, Rossendale

Professor Collette Clifford, Professor of Nursing & Head of Research, School of Health Sciences, University of Birmingham, Edgbaston, Birmingham

Professor Barry Cookson, Director, Laboratory of Healthcare Associated Infection, Health Protection Agency, London

Professor Howard Cuckle, Professor of Reproductive Epidemiology, Department of Paediatrics, Obstetrics & Gynaecology, University of Leeds

Dr Katherine Darton, Information Unit, MIND – The Mental Health Charity, London

Professor Carol Dezateux, Professor of Paediatric Epidemiology, London

Mr John Dunning, Consultant Cardiothoracic Surgeon, Cardiothoracic Surgical Unit, Papworth Hospital NHS Trust, Cambridge

Mr Jonothan Earnshaw, Consultant Vascular Surgeon, Gloucestershire Royal Hospital, Gloucester

Professor Martin Eccles, Professor of Clinical Effectiveness, Centre for Health Services Research, University of Newcastle upon Tyne

Professor Pam Enderby, Professor of Community Rehabilitation, Institute of General Practice and Primary Care, University of Sheffield

Mr Leonard R Fenwick, Chief Executive, Newcastle upon Tyne Hospitals NHS Trust

Professor David Field, Professor of Neonatal Medicine, Child Health, The Leicester Royal Infirmary NHS Trust

Mrs Gillian Fletcher, Antenatal Teacher & Tutor and President, National Childbirth Trust, Henfield

Professor Jayne Franklyn, Professor of Medicine, Department of Medicine, University of Birmingham, Queen Elizabeth Hospital, Edgbaston, Birmingham

Ms Grace Gibbs, Deputy Chief Executive, Director for Nursing, Midwifery & Clinical Support Services, West Middlesex University Hospital, Isleworth

Dr Neville Goodman, Consultant Anaesthetist, Southmead Hospital, Bristol

Professor Alastair Gray, Professor of Health Economics, Department of Public Health, University of Oxford

Professor Robert E Hawkins, CRC Professor and Director of Medical Oncology, Christie CRC Research Centre, Christie Hospital NHS Trust, Manchester

Professor Allen Hutchinson, Director of Public Health & Deputy Dean of ScHARR, Department of Public Health, University of Sheffield

Dr Duncan Keeley, General Practitioner (Dr Burch & Ptnrs), The Health Centre, Thame

Dr Donna Lamping, Research Degrees Programme Director & Reader in Psychology, Health Services Research Unit, London School of Hygiene and Tropical Medicine, London

Mr George Levvy, Chief Executive, Motor Neurone Disease Association, Northampton

Professor James Lindesay, Professor of Psychiatry for the Elderly, University of Leicester, Leicester General Hospital

Professor Julian Little, Professor of Human Genome Epidemiology, Department of Epidemiology & Community Medicine, University of Ottawa

Professor Rajan Madhok, Medical Director & Director of Public Health, Directorate of Clinical Strategy & Public Health, North & East Yorkshire & Northern Lincolnshire Health Authority, York

Professor David Mant, Professor of General Practice, Department of Primary Care, University of Oxford

Professor Alexander Markham, Director, Molecular Medicine Unit, St James's University Hospital, Leeds

Dr Chris McCall, General Practitioner, The Hadleigh Practice, Castle Mullen

Professor Alistair McGuire, Professor of Health Economics, London School of Economics

Dr Peter Moore, Freelance Science Writer, Ashtead

Dr Sue Moss, Associate Director, Cancer Screening Evaluation Unit, Institute of Cancer Research, Sutton

Mrs Julietta Patnick, Director, NHS Cancer Screening Programmes, Sheffield

Professor Tim Peters, Professor of Primary Care Health Services Research, Academic Unit of Primary Health Care, University of Bristol

Professor Chris Price, Visiting Chair – Oxford, Clinical Research, Bayer Diagnostics Europe, Cirencester

Professor Peter Sandercock, Professor of Medical Neurology, Department of Clinical Neurosciences, University of Edinburgh

Dr Eamonn Sheridan, Consultant in Clinical Genetics, Genetics Department, St James's University Hospital, Leeds

Dr Ken Stein, Senior Clinical Lecturer in Public Health, Director, Peninsula Technology Assessment Group, University of Exeter

Professor Sarah Stewart-Brown, Professor of Public Health, University of Warwick, Division of Health in the Community Warwick Medical School, LWMS, Coventry

Professor Ala Szczepura, Professor of Health Service Research, Centre for Health Services Studies, University of Warwick

Dr Ross Taylor, Senior Lecturer, Department of General Practice and Primary Care, University of Aberdeen

Mrs Joan Webster, Consumer member, HTA – Expert Advisory Network

Feedback

The HTA Programme and the authors would like to know your views about this report.

The Correspondence Page on the HTA website (http://www.ncchta.org) is a convenient way to publish your comments. If you prefer, you can send your comments to the address below, telling us whether you would like us to transfer them to the website.

We look forward to hearing from you.

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 23 8059 5639 Email: hta@soton.ac.uk
http://www.ncchta.org

ISSN 1366-5278