INTRODUCTION TO TEST EVALUATION RESEARCH

Patrick MM Bossuyt

Overview

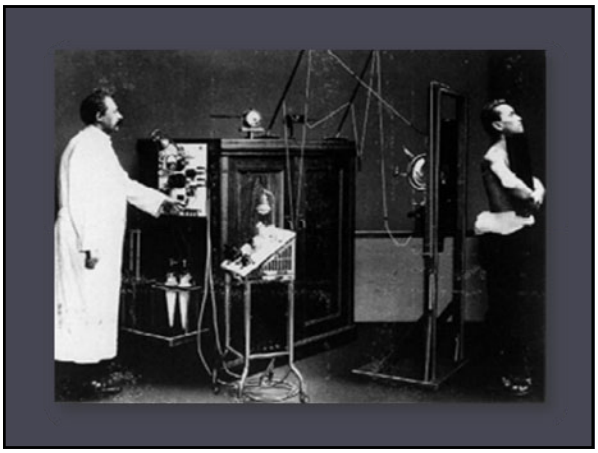
1. Tests
2. Evaluation of Tests – RCT
3. Target Condition
4. Test Accuracy Studies
5. Systematic Reviews of Test Accuracy
6. Coda

1. Tests

Why use tests?

- Diagnosis
- Monitoring course disease
- Selecting therapy
- Following effects of therapy
- Determining drug levels or drug effects
- Evaluate Health or Fitness
- Screening
- Case Finding

2. Evaluate Tests



Medical Test → *Information*

Can I trust the information from this test?

Medical Test → *Information*

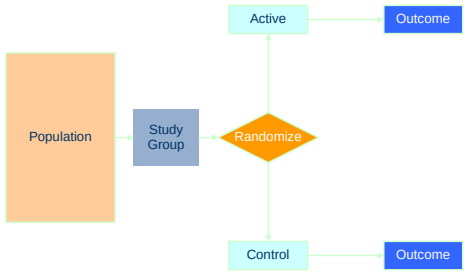
Can I trust that this information tells me something about the patients' health?

Medical Test → *Information*

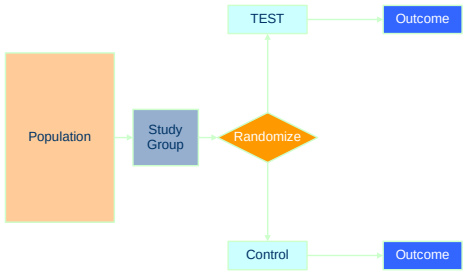
Can I trust that this information *improves* patients' health?

Patient Outcome

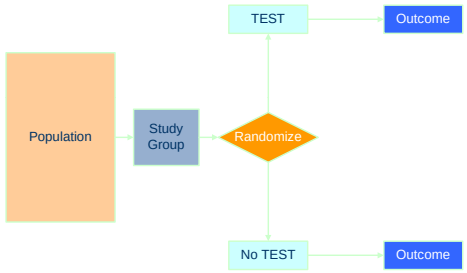
Randomized Clinical Trial



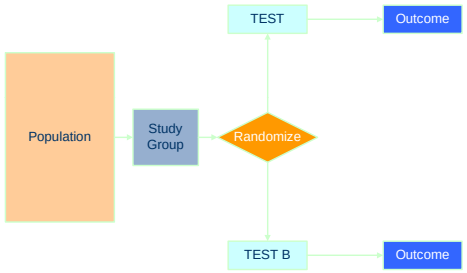
Medical Test RCT



Medical Test RCT



Medical Test RCT



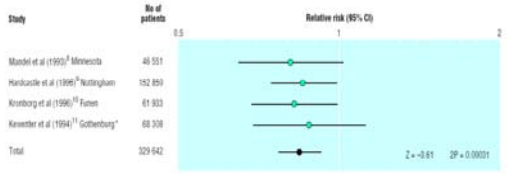
The Cochrane Library — Evidence for healthcare decision-making

Combination antimicrobial susceptibility testing for acute exacerbations in chronic infection of Pseudomonas aeruginosa in cystic fibrosis
Valerie Waters, Felix Ratjen
Year: 2008

Screening for colorectal cancer using the faecal occult blood test, Hemoccult
P Hewitson, P Glasziou, L Irwig, B Towler, E Watson
Year: 2007

Urodynamic investigations for management of urinary incontinence in children and adults
CMA Glazener, MC Lapitan
Year: 2002

Meta-Analysis Screening RCT



Relative Risk CRC Mortality: 0.84

(95% CI: 0.77 to 0.93)

BMJ 1998;317:559–65

THE LANCET

Randomised controlled trial of faecal-occult-blood screening for colorectal cancer

Jack D Hardcastle, Jocelyn O Chamberlain, Michael H E Robinson, Susan M Moss, Satya S Amar, Tom W Balfour, Peter D James, Christine M Mangham

Summary

Background There is growing evidence that faecal-occult-blood (FOB) screening may reduce colorectal cancer (CRC) mortality, but this reduction in CRC mortality has not been shown in an unselected population-based randomised controlled trial. The aim of this study was to assess the effect of FOB screening on CRC mortality in such a setting.

Methods Between February, 1981, and January, 1991, 152 850 people aged 45–74 years who lived in the Nottingham area of the UK were recruited to our study. Participants were randomly allocated FOB screening (76 466) or no screening (controls; 76 384). Controls were not told about the study and received no intervention. Screening-group participants were sent a Haemoccult FOB test kit with instructions from their family doctor. FOB tests were not rehydrated and dietary restrictions were imposed only for retesting borderline results. Individuals with negative FOB tests at the first screening, together with those who tested positive but in whom no neoplasia was found on colonoscopy, were invited to take part in further screening every 2 years. Screening was stopped in February, 1995, by which time screening-group participants people died from CRC in the screening group compared with 420 in the control group—a 15% reduction in cumulative CRC mortality in the screening group (odds ratio=0·85 [95% CI 0·74–0·98], p=0·026).

Interpretation Our findings together with evidence from other trials suggest that consideration should be given to a national programme of FOB screening to reduce CRC mortality in the general population.

Lancet 1996; **348:** 1472–77

See Commentary page 1463

Introduction

Colorectal cancer (CRC) is the second commonest cause of death from malignant disease in England and Wales, and resulted in about 16 000 deaths in 1993.[1] Although there have been advances in the management of symptomatic CRC, there has been little overall reduction in CRC mortality during the past 30 years. Tumour stage is an important determinant of outcome; 24–28% of patients have metastatic disease at presentation and the tumour is confined to the bowel wall in only 6–10% (Dukes' stage A).[2,3] Early diagnosis before the

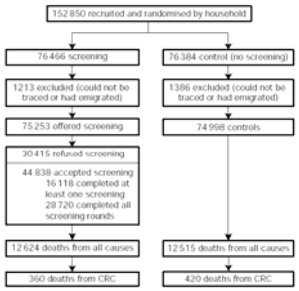


Figure 1: **Trial profile**

Lancet 1996;348:1473

UK RCT: CRC incidence & mortality

	Rate (/1000 pyrs)		RR
	Screening	Control	
CRC	1.49	1.44	1.04
CRC mortality	0.59	0.67	0.88
Total mortality	21.1	21.0	1.01 (0.98-1.03)

Median follow-up 7·8 years

Lancet 1996;348:1475

THE LANCET

Articles

Randomised study of screening for colorectal cancer with faecal-occult-blood test

Ole Kronborg, Claus Fenger, Jørn Olsen, Ole Dan Jørgensen, Ole Søndergaard

Summary

Background Case-control studies and a voluntary-based follow-up study have suggested that repeated screening with faecal-occult-blood (FOB) tests can lead to a reduction in mortality from colorectal cancer (CRC). The aim of this randomised study was to compare mortality rates after FOB tests every 2 years during a 10-year period with those of unscreened similar controls.

Methods 140 000 people aged 45–75 years lived in Funen, Denmark, in August, 1985, and were considered for inclusion in our study. Before randomisation we excluded individuals who had CRC or precursor adenomas and those who had taken part in a previous pilot study. Randomisation of 137 485 people in blocks of 14 allocated three per 14 to the screening group (30 967), three per 14 to the control group (30 966), and eight not to be enrolled in the study (75 552). Controls were not told about the study and continued to use health-care facilities as normal. Hemoccult-II blood tests (with dietary restrictions but the effect of the removal of more precursor adenomas in the screening-group participants than in controls on CRC incidence.

Lancet 1996; **348:** 1467–71

See Commentary page 1463

Introduction

Denmark has high incidence and mortality rates for colorectal cancer (CRC)—the incidence of colonic cancer is currently increasing and mortality remains constant, whereas incidence and mortality of rectal cancer are declining. Early stages of CRC are commonly found in only 10–15% of patients with symptoms. However, there is evidence that earlier diagnosis and treatment of CRC in symptom-free patients may reduce mortality. Two case-control studies of screening with faecal-occult-blood (FOB) tests reported reductions in CRC mortality rates of 31%[1] and 57%[2] (the latter reduction was found only in women). Similarly, Winawer and colleagues' non-randomised study[3] showed that annual rigid

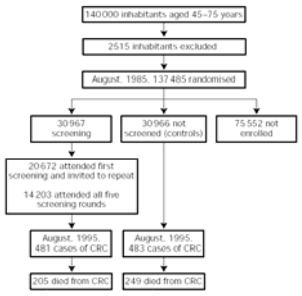


Figure 1: **Study profile**

Lancet 1996;348:1468

Urodynamic investigations for management of urinary incontinence in children and adults

...

Objectives

The objective of this review was to discover if treatment according to a urodynamic-based diagnosis led to clinical improvements in urinary incontinence outcomes, compared to treatment based on history and examination.

...

Three small trials were found, which included 184 people, although information was only available for 128 participants. There was not enough evidence to determine whether these tests lead to better outcomes. There was some evidence that urodynamic testing increased the number of people prescribed drug treatments or treated by surgery, but it was not known whether this resulted in less incontinence or a better quality of life.

More research is needed, in which people are randomised to having treatment decisions based on either their symptoms and examination alone, or the extra information provided by urodynamic tests.

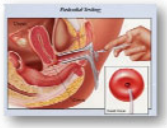
Effectiveness of the postcoital test: randomised controlled trial

S Guid Oei, Frans M Helmerhorst, Kirry W M Bloemenkamp, Frederieke A M Hollanders, Debbie E M Meerpoel, Marc J N C Keirse

Abstract

Objectives: to investigate the impact of the postcoital test on the pregnancy rate among subfertile couples and on the number of other diagnostic tests and treatments.

Design: Randomised controlled study.



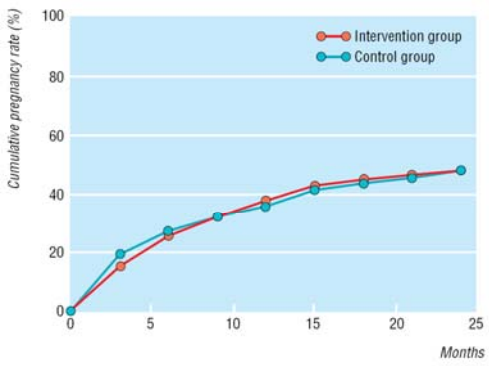
BMJ 1998;317:502–5

Table 2 Characteristics of couples participating in study

Characteristics	Intervention group (n=227)	Control group (n=217)
Mean (range) age of woman (years)	30.8 (18.8-41.0)	30.4 (19.4-43.6)
Mean (range) age of man (years)	33.5 (22.8-56.4)	32.6 (22.5-57.7)
Mean (range) duration of infertility (months)	26.7 (12-480)	24.8 (12-144)
No (%) of couples with history of sexual problems	35 (15)	27 (12)
No (%) of partners with sperm disorder	72 (32)	72 (33)
No (%) of women with ovulatory disorder	50 (22)	38 (18)
No (%) of women with tubal disorder	20 (9)	27 (12)

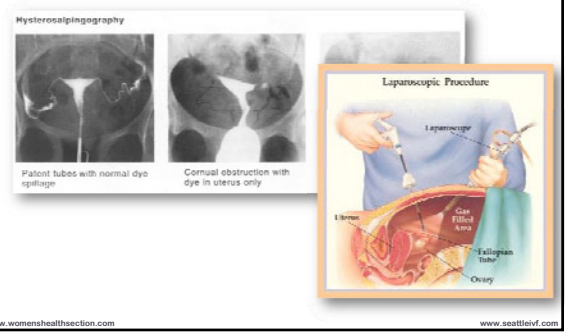
Table 3 Number (percentage) of infertility investigations performed in intervention group and control group

Investigation	Intervention group (n=227)	Control group (n=217)
Postcoital test	146 (64)	3 (1)
In vitro cervical mucus test	17 (7)	1 (0.4)
Semen analysis	219 (96)	201 (93)
Hysterosalpingography	114 (50)	117 (54)
Laparoscopy	32 (14)	44 (20)



Cumulative pregnancy rates for 227 couples in intervention group (which included postcoital test) and 217 couples in control group (which excluded the test)

Tubal integrity testing



Human Reproduction

doi:10.1093/humrep/deh478

Routine use of hysterosalpingography prior to laparoscopy in the fertility workup: a multicentre randomized controlled trial

D.A.M.Perquin[1,4], P.J.Dörr[1], A.J.M.de Craen[2] and F.M.Helmerhorst[3]

[1]Department of Obstetrics and Gynaecology, Medical Centre Haaglanden, The Hague, [2]Department of Gerontology and Geriatrics and [3]Department of Gynaecology, Division of Reproductive Medicine, Leiden University Medical Centre, Leiden, The Netherlands

[4]To whom correspondence should be addressed at: Department of Obstetrics and Gynaecology, Medical Centre Haaglanden, PO Box 432, 2501 CK, The Hague, The Netherlands. E-mail: dperquin@haaglanden.nl

BACKGROUND: A multicentre randomized controlled trial with or without hysterosalpingography (HSG) was conducted to assess the usefulness of HSG as a routine investigation in the fertility workup prior to laparoscopy and dye. **METHODS:** From 1 April 1997 to 1 April 2002, subfertile women were allocated by a computer-based 1 : 1 ratio randomization procedure, either for an HSG followed by laparoscopy and dye (the intervention group) or for laparoscopy and dye only (the control group) as a part of their fertility workup. Cumulative pregnancy rate (CPR) within 18 months after randomization was the primary outcome of interest. **RESULTS:** 344 women were randomized to the intervention group ($n = 169$) and the control group ($n = 175$). There was no significant difference in CPR at 18 months in the intervention group (49.1%) [95% confidence interval (CI) 41.6 to 56.6] and the control group (50.3%) (95% CI 42.8 to 57.8), a difference of −1.2% (95% CI −11.8% to 9.5%). **CONCLUSION.** The routine use of HSG at an early stage in the fertility workup prior to laparoscopy and dye does not influence CPR, compared with the routine use of laparoscopy and dye without HSG.

Key words: hysterosalpingography/laparoscopy and dye/pregnancy rate/randomized controlled trial

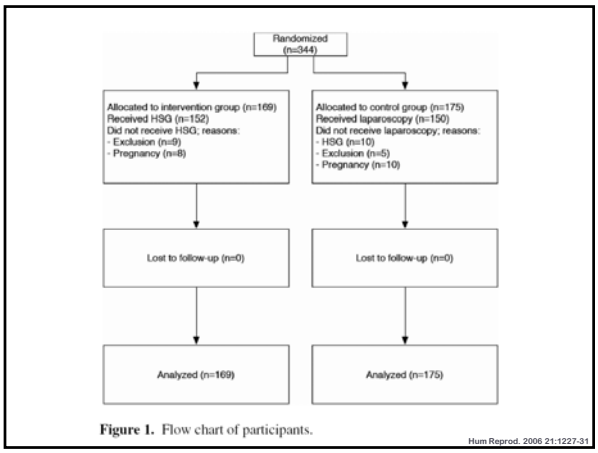
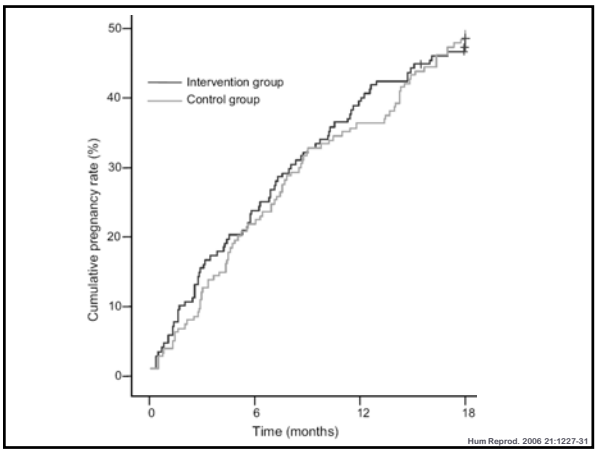


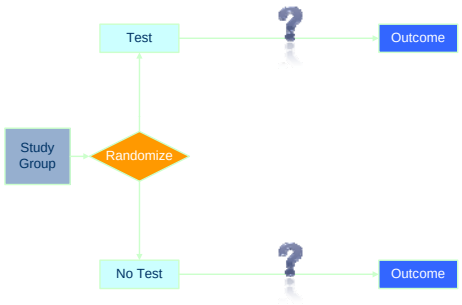
Figure 1. Flow chart of participants.

Hum Reprod. 2006 21:1227-31

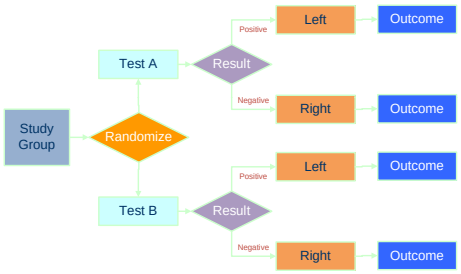


Hum Reprod. 2006 21:1227-31

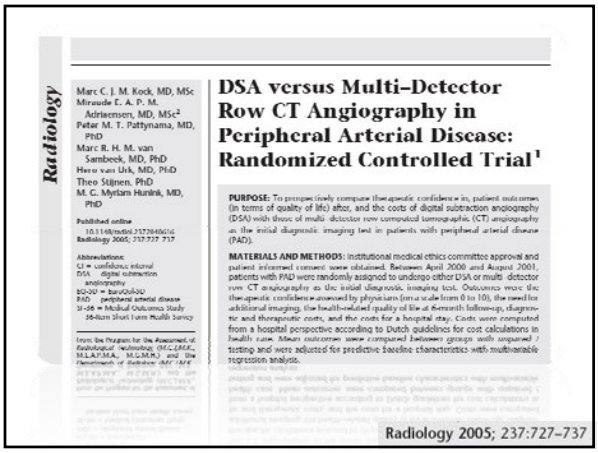
RCT Medical Test



RCT Medical Test



Bossuyt et al; Lancet 2000



Radiology

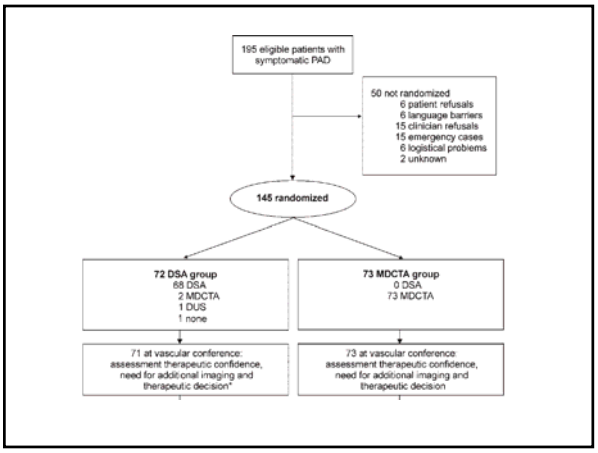
Marc C. J. M. Kock, MD, MSc
Miraude E. A. P. M. Adriaensen, MD, MSc
Peter M. T. Pattynama, MD, PhD
Marc R. H. M. van Sambeek, MD, PhD
Hero van Urk, MD, PhD
Theo Stijnen, PhD
M. G. Myriam Hunink, MD, PhD

DSA versus Multi–Detector Row CT Angiography in Peripheral Arterial Disease: Randomized Controlled Trial[1]

PURPOSE: To prospectively compare therapeutic confidence in, patient outcomes (in terms of quality of life) after, and the costs of digital subtraction angiography (DSA) with those of multi–detector row computed tomographic (CT) angiography as the initial diagnostic imaging test in patients with peripheral arterial disease (PAD).

MATERIALS AND METHODS: Institutional medical ethics committee approval and patient informed consent were obtained. Between April 2000 and August 2001, patients with PAD were randomly assigned to undergo either DSA or multi–detector row CT angiography as the initial diagnostic imaging test. Outcomes were the therapeutic confidence assessed by physicians (on a scale from 0 to 10), the need for additional imaging, the health-related quality of life at 6-month follow-up, diagnostic and therapeutic costs, and the costs for a hospital stay. Costs were computed from a hospital perspective according to Dutch guidelines for cost calculations in health care. Mean outcomes were compared between groups with unpaired *t* testing and were adjusted for predictive baseline characteristics with multivariable regression analysis.

Radiology 2005; 237:727–737

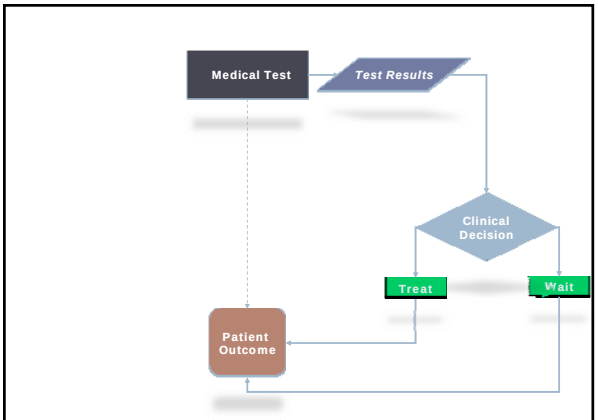


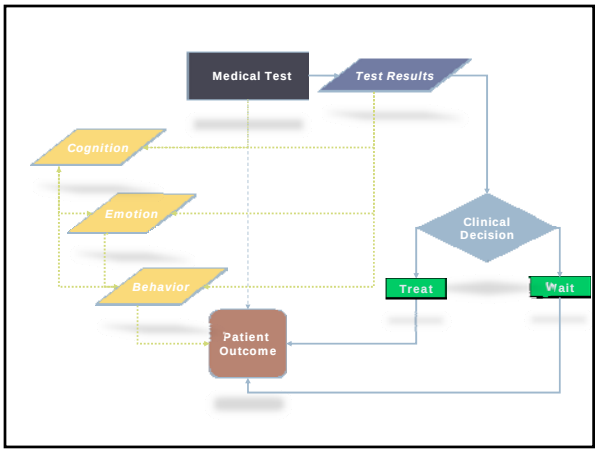
angiography group. There were 47 men in the DSA group and 58 men in the CT angiography group. Physician confidence in making a correct therapeutic choice was significantly higher at DSA (mean confidence score, 8.2) than at CT angiography (mean score, 7.2; $P < .001$). During 6 month follow up, 14% less additional

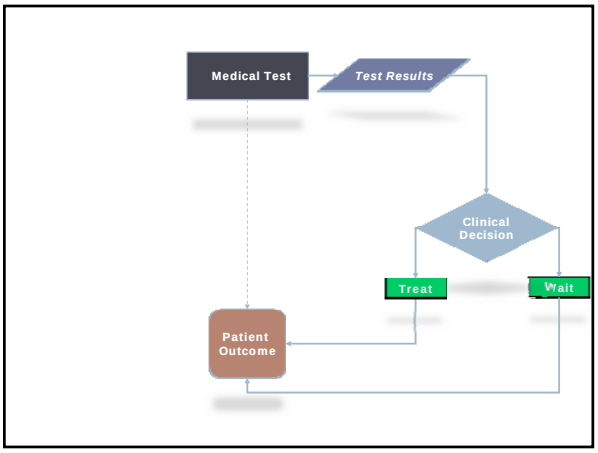
RCT of Testing

- Best evidence of effectiveness
- Rare
- Usually need large sample sizes
- Need protocol
- Need patient outcomes that matter

3. Target Condition

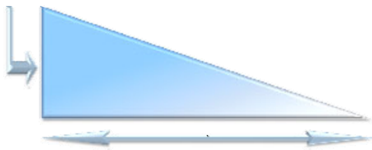






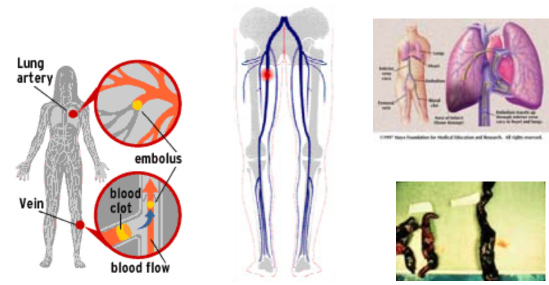
Spectrum of disease

Probability of Positive Test Result
Benefits of Treatment



Spectrum of Disease

Venous Thromboembolism



Centennial Dissertation

Honoring Percy Brown, MD and Frederick H. Baetjer, MD

Percy Brown
12th President of ARRS
1911–1912

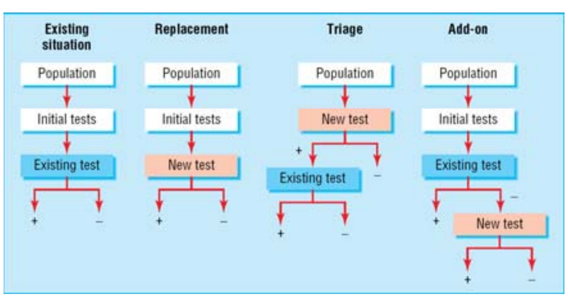
Pulmonary Embolism: What's Wrong with This Diagnosis?

Tony P. Smith[1]

The first description of pulmonary embolism is attributed to Laennec [1] in 1819 and what probably represents the first case description to Helie [2] in 1837. However, it was von Virchow [3] in 1846 who first described the connection between venous thrombosis and pulmonary embolic disease, even using the term "embolis." The first radiographic description of pulmonary embolism is that of Wharton and Pierson [4] of a chest radiograph in 1922. Since that time, virtually all areas of radiology—most notably chest radi-

published incidence figures are not surprising. Certainly, pulmonary embolism remains a clinical problem. Silverstein et al. [8] recently published a retrospective review of records from a population-based study and found the incidence of pulmonary embolism to be 69 in 100,000. Although studies such as this provide a better understanding of the scope of the problem, in the final analysis these figures are reliable only when a confident and sure diagnosis of pulmonary embolism can be reached. The current diagnostic measures for acute pulmonary embolism

Roles of tests



Bossuyt et al; BMJ 2006

4. Test Accuracy

Bladder Tumor Markers (BTM)

- To optimize monitoring of tumor recurrence and progression, without incurring more invasive and expensive medical tests

Diagnostic Accuracy

Index Test

↓

Reference Standard

Diagnostic Accuracy Study

Series of Patients

↓

Index Test

↓

Reference standard

↓

Cross-classification

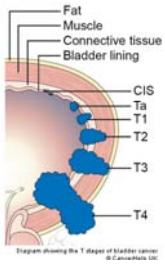
The Results

Reference Standard

Index Test	Target Condition	Other Condition
Positive		
Negative		

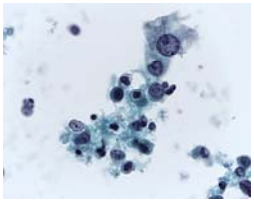
Bladder Cancer

close follow-up due to the significant risk of tumor recurrence.



Cytology: central test

- Efficient
- Non-invasive
- Inexpensive
- But imperfect



- Cystoscopy

Bladder Tumor Markers (BTM)

- To optimize monitoring of tumor recurrence and progression,
  without incurring more invasive and expensive medical tests

Example

TOPIC PAPER

The role of BTA *stat* Test in follow-up of patients with bladder cancer: results from FinnBladder studies

Miika-P. Raitanen · The FinnBladder Group

Received: 11 October 2007 / Accepted: 15 November 2007 / Published online: 8 January 2008

Abstract
Objectives To summarize the results of the FinnBladder studies of the BTA stat Test in follow up of bladder cancer and more importantly to provide guidelines for daily clinical practice.
urine infection, in those having received BCG, or in those with present instillation of any type. In case of positive test result but negative cystoscopy, urine cytology should be obtained as the first line examination. Positive cytology is the indication for further examinations, whereas patients

Diagnostic Accuracy Study

501 patients
↓
BTA *stat*
↓
Cystoscopy
↓
Cross-classification

The Results

Cystoscopy (columns) × BTA *stat* (rows)

	BCR	No BCR
Positive	71	*96*
Negative	*62*	272

Measures of Diagnostic Test Accuracy

- Sensitivity & Specificity 54 % 74 %
- PPV & NPV 43 % 81 %
- Likelihood Ratios 2.0 .63
- Diagnostic Odds Ratio 3.2

	BCR	no BCR
Positive	71	96
Negative	62	272

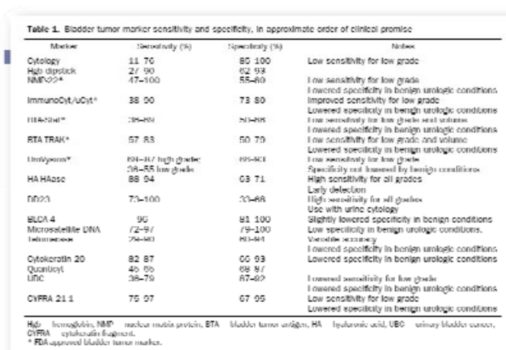


Table 1. Bladder tumor marker sensitivity and specificity, in approximate order of clinical promise

Marker	Sensitivity (%)	Specificity (%)	Notes
Cytology	11–76	80–100	Low sensitivity for low grade
Hgb dipstick	27–90	62–93	
NMP-22*	47–100	55–80	Low sensitivity for low grade Lowered specificity in benign urologic conditions
ImmunoCyt/uCyt*	38–90	73–80	Improved sensitivity for low grade Lowered specificity in benign urologic conditions
BTA-Stat*	[illegible]	[illegible]	Low sensitivity for low grade and volume Lowered specificity in benign urologic conditions
BTA TRAK*	57–83	50–79	Low sensitivity for low grade and volume Lowered specificity in benign urologic conditions
UroVysion*	[illegible] high grade; 38–55 low grade	[illegible]	Low sensitivity for low grade Specificity not lowered by benign conditions
HA HAase	88–94	63–71	High sensitivity for all grades Early detection
DD23	73–100	33–68	High sensitivity for all grades Use with urine cytology
BLCA-4	96	81–100	Slightly lowered specificity in benign conditions
Microsatellite DNA	72–97	79–100	Low specificity in benign urologic conditions.
Telomerase	[illegible]	[illegible]	Variable accuracy Lowered specificity in benign urologic conditions
Cytokeratin 20	82–87	66–93	Lowered specificity in benign urologic conditions
Quanticyt	45–65	68–87	
UBC	38–79	87–92	Lowered sensitivity for low grade Lowered specificity in benign urologic conditions
CYFRA 21-1	75–97	67–95	Low sensitivity for low grade Lowered specificity in benign urologic conditions

Hgb hemoglobin, NMP nuclear matrix protein, BTA bladder tumor antigen, HA hyaluronic acid, UBC urinary bladder cancer, CYFRA cytokeratin fragment.
* FDA approved bladder tumor marker.

Message

- Measures of Diagnostic Test Accuracy

  are

  test characteristics,

  NOT fixed test properties

5. Systematic Reviews

Why systematic reviews

- Extensive/exhaustive search
- Critical appraisal
- Meta-analysis
  - Increased precision
  - Explore variability

1. Focused question

Systematic Review: Question Elements

What is the diagnostic accuracy of

<Index test>

[versus <comparator>]
for detecting <target condition>
in <patient description>

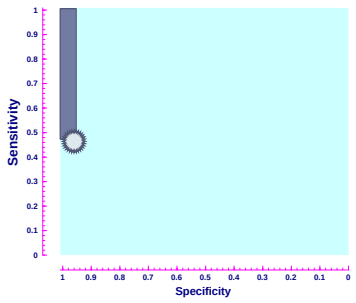
Systematic Review: Question Elements

What is the diagnostic accuracy of

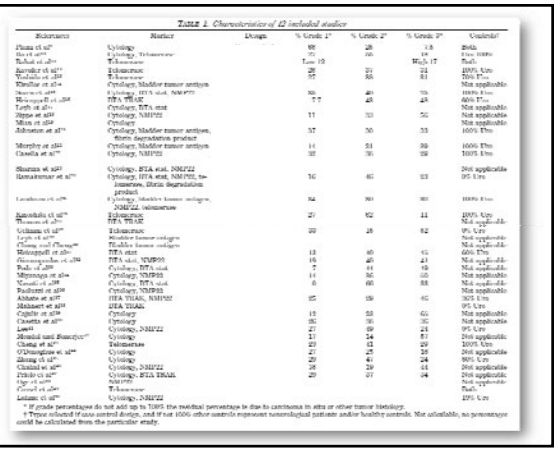
BTA stat test

versus cytology
for detecting recurrent disease
in patients with bladder cancer

ROC space



2. Identification & selection studies



3. Quality Assessment

Differential Verification Bias

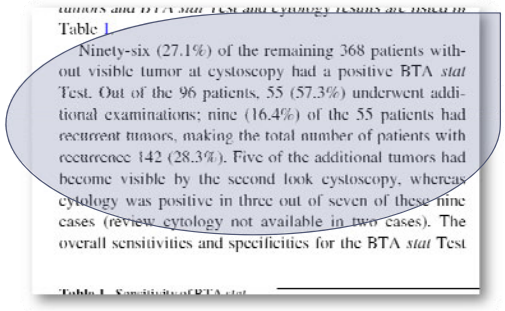
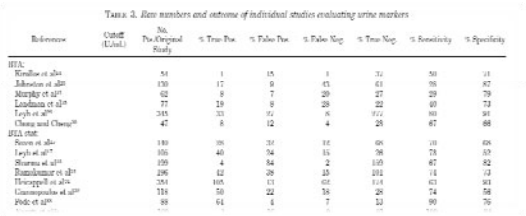


Table I

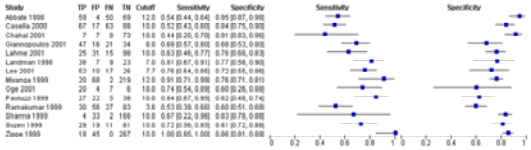
Ninety-six (27.1%) of the remaining 368 patients without visible tumor at cystoscopy had a positive BTA *stat* Test. Out of the 96 patients, 55 (57.3%) underwent additional examinations; nine (16.4%) of the 55 patients had recurrent tumors, making the total number of patients with recurrence 142 (28.3%). Five of the additional tumors had become visible by the second look cystoscopy, whereas cytology was positive in three out of seven of these nine cases (review cytology not available in two cases). The overall sensitivities and specificities for the BTA *stat* Test

4. Meta-analysis

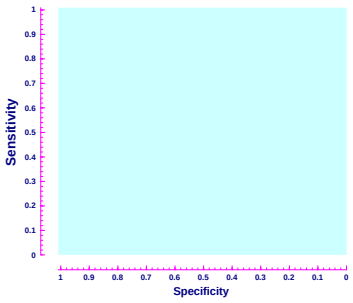
Study results



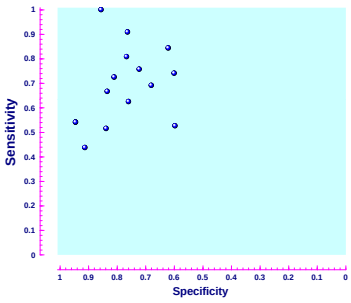
Paired Forest Plots



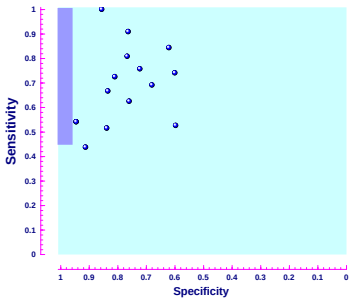
ROC space



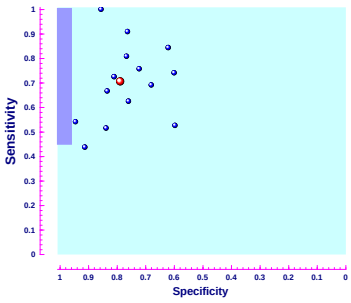
ROC space



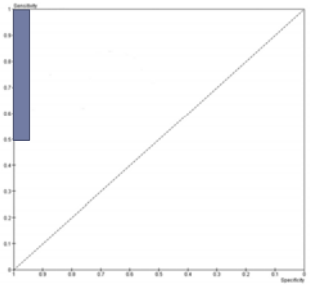
ROC space



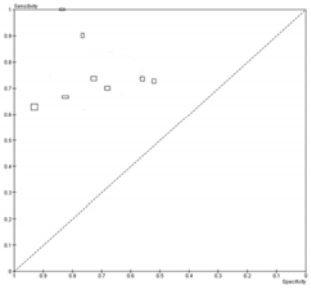
ROC space



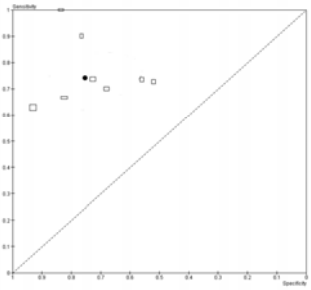
Summary Point in ROC space



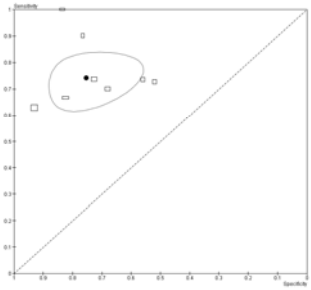
Summary Point in ROC space



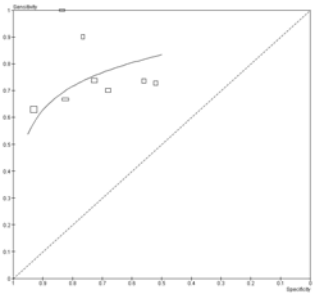
Summary Point in ROC space



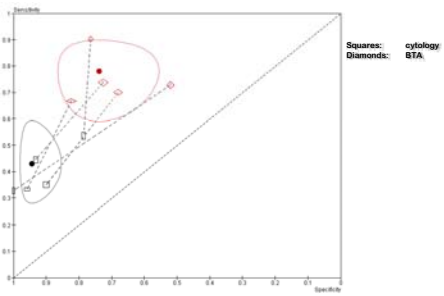
Summary Point in ROC space



Summary Curve in ROC space



Paired studies: cytology & BTA



5. Interpretation

The Journal of Urology®
Copyright © 2003 by American Urological Association

Printed in U.S.A.

Review Article

TUMOR MARKERS IN THE DIAGNOSIS OF PRIMARY BLADDER CANCER. A SYSTEMATIC REVIEW

AFINA S. GLAS, DAPHNE ROOS, MARIJE DEUTEKOM, AEILKO H. ZWINDERMAN, PATRICK M. M. BOSSUYT AND KARL H. KURTH

From the Departments of Clinical Epidemiology and Biostatistics, and Urology, Academic Medical Center, and the Department of General Surgery, The Netherlands

ABSTRACT

Purpose: We systematically reviewed the available evidence, and obtained and compared summary estimates of the sensitivity and specificity of cytology and the urine based markers bladder tumor antigen, BTA stat (Polymedco, Redmond, Washington), BTA TRAK (Polymedco), NMP22 (Matritech, Cambridge, Massachusetts), telomerase and fibrin degradation product in detecting primary bladder cancer.

Materials and Methods: Studies on the diagnosis of primary bladder cancer published from 1990 through November 2001 in English and German were retrieved from MEDLINE and EMBASE data bases. In our research we included studies that evaluated 1 or more of the markers, used cystoscopy as the reference standard and allowed the construction of a 2×2 contingency table for a per patient analysis. The data plus items on study and clinical characteristics were extracted by 2 observers. Sensitivity and specificity for each marker were estimated using a bivariate random effect meta-analysis. A multivariable analysis was performed to explain study variation.

Results: A total of 42 studies were included in our review. Only 2 studies were available on fibrin degradation product, hence a meta analysis was not possible. Cytology had the best specificity at 94% (95% CI: 90% to 96%). This figure was significantly better than that of the other markers except for telomerase (specificity 86% [71% to 94%]). Telomerase had the best sensitivity (75% [71% to 79%]) but it was not significantly better than that of BTA stat (70% [66% to 74%]). Case control designs yielded lower values for sensitivity for the tumor markers cytology, bladder tumor antigen and BTA stat.

Conclusions: Cytology has the best specificity and telomerase the best sensitivity. However, none of the markers studied here is sensitive enough to be recommended for daily routine.

KEY WORDS: sensitivity and specificity, urine, biological markers, bladder neoplasms

Museum Boijmans van Beuningen, Rottrdam