Retrospective analysis of evidence base for tests used in diagnosis and monitoring of disease in respiratory medicine

Z Borrill, C Houghton, P Sestini, P J Sullivan

Department of Cardiorespiratory Medicine, Hope Hospital, Manchester M6 8HD
Z Borrill *clinical fellow*
C Houghton *clinical fellow*
P J Sullivan *consultant*

Department of Clinical Medicine and Immunological Sciences, Division of Respiratory Diseases, University of Siena, Viale Bracci 3, 53100 Siena, Italy
P Sestini *associate professor of respiratory diseases*

Correspondence to:
P J Sullivan
Paul.sullivan@srht.nhs.uk

BMJ 2003;327:1136–8

Abstract

Objectives To determine how many common clinical tests used in a respiratory medicine outpatient clinic are based on high quality evidence.
Design Retrospective review of case notes. Record of first three tests for each patient. Diagnostic tests, tests used to assess existing condition, explicit trials of therapy were included. Literature search for supporting evidence and grading of best evidence for each test.
Setting Inner city university teaching hospital in the United Kingdom.
Participants All new outpatients referred to a single respiratory medicine team over a period of three months.
Main outcome measures Proportion of tests supported by level 1a-1c evidence (scale developed by Centre for Evidence Based Medicine).
Results Only half the tests that were used to make or exclude a diagnosis and a fifth of the tests used to assess a known condition were supported by level 1a-1c evidence. There was no evidence to support trials of therapy.
Conclusions A large proportion of clinical tests in respiratory medicine are not supported by level 1a-1c evidence. None of the therapeutic trials that were used were supported by evidence.

Introduction

Clinical practice based on scientific evidence is a major goal of the clinical governance process.[1] The randomised controlled trial is regarded as the standard for the assessment of therapeutic interventions.[2] Several studies have examined how many treatments in everyday clinical practice are based on good evidence in a range of specialties and in general practice.[3–6] However, good treatment relies on accurate diagnosis and doubts have been expressed regarding the quality and breadth of the current evidence base for diagnostic tests. Criteria for appraisal of papers that assess medical tests are available,[7] just as they are for studies that look at therapeutic interventions, and in diagnostic testing poor study design has been shown to be associated with significant outcome bias.[8]

We used established criteria to assess the quality of available evidence for tests used in routine outpatient clinical practice in one respiratory medicine clinic. Previous studies of the proportion of therapeutic interventions that are evidence based have used the patient as denominator, expressing findings as the proportion of patients who received at least one evidence based intervention. Tests behave differently in that the final diagnosis may be based on a combination of test results. If an individual patient undergoes a series of tests that include high quality evidence based tests as well as inaccurate or unassessed tests the final diagnosis may be incorrect. We therefore used tests as the denominator rather than patients.

Methods

The study took place in a UK inner city teaching hospital that provides a referral service for primary care and other specialties. We examined the notes of all consecutive patients referred to the respiratory outpatient clinic in a three month period and recorded the first three eligible tests ordered for each patient. We included tests if they were performed to make a diagnosis or to assess a prediagnosed condition. We excluded tests performed as part of routine preclinical investigation and tests, such as full blood count, if they seemed to have been performed without any specific diagnosis in mind. Routine clinical examination was not included. The tests used were recorded along with the question that they were being used to answer. We used these test-question combinations as the denominator for this study—for example, "serum angiotensin converting enzyme concentration to diagnose sarcoidosis" or "serum angiotensin converting enzyme concentration to assess activity of known sarcoidosis" were considered separately.

We divided tests into three groups: group A comprised tests aimed at making a diagnosis; group B comprised tests performed to assess a previously diagnosed condition; and group C was a trial of therapy, which we included as a special type of test, when a drug was prescribed for a limited period with the explicit intention of predicting future response in an individual. A comprehensive Medline search was performed (1966-2001) for each test-question combination by two researchers experienced in searching medical databases. We used a published strategy with a sensitivity of 92%[9] followed by a freely improvised search for each test-question pair. The best evidence that we retrieved for each test-question was graded according to the scale devised by the Centre for Evidence Based Medicine, Oxford, (www.cebm.net/levels_of_evidence.asp) (table 1). Some group A tests were regarded as absolutely specific and therefore graded as level 1c. In group C we searched for evidence that the result of a short term trial could predict the usefulness of a drug for an individual in the longer term.

Results

Referrals were received for 90 patients during the three month period. Patients were seen by a consultant (PJS) or specialist registrar (or equivalent) in the same team. Not all patients had three eligible tests. A total of 165 tests were recorded, 137 in group A, 15 in group B, and 13 in group C. The tests could be represented as 38 different test-question combinations; 26 in group A, 5 in group B, and 7 in group C. Table 2 shows the best

Table 1 Levels of evidence according to criteria from Centre for Evidence Based Medicine, Oxford

1a	Systematic review of level 1 studies with homogeneity
1b	Prospective cohort study with good reference standards
1c	Absolutely specific or sensitive
2a	Systematic review of level 2 or >2 studies with homogeneity
2b	Exploratory cohort study with good reference standards
3a	Systematic review of studies from level 1, 2, or 3b
3b	Study of non-consecutive cases or without universally applied reference standard
4	Case-control study with poor or non-independent reference standard
5	Expert opinion or based on physiology, bench research, or first principles

evidence found for each test categorised and ranked according to the Centre for Evidence Based Medicine criteria. The finding of visible tumour on bronchoscopy with histological confirmation and the finding of mycobacterium tuberculosis in bronchial washings when tuberculosis was the suspected diagnosis were regarded as absolutely specific and therefore level 1c. Both investigators agreed on the level of evidence assigned to each study. In group A there was level 1a-1c evidence for half of the of test-question combinations and in group B a fifth. In group C we found no studies that examined the predictive role for five of the seven therapeutic trials. In the case of trials of oral or inhaled corticosteroids in chronic obstructive pulmonary disease we found literature that we thought did not show that these trials were predictive.

Discussion

Few, if any, diagnostic tests give unambiguous results. To deal with this we are advised to combine clinical impressions of pretest probability with test results to derive a post-test probability of disease.[10] This requires that the test be assigned a weighting, expressed formally as a likelihood ratio—that is, calculated from the results of scientific studies of the test's performance. Standards for research of diagnostic tests have

Table 2 Test-context combinations and best evidence found by literature review (or by applying rule that absolutely specific tests are level 1c)

Group*	Test	Context (for diagnosis unless stated otherwise)	Best evidence found†
A	Serum angiotensin converting enzyme	Sarcoidosis	2b
A	Serum antineutrophil cytoplasm antibody titre	Churg Strauss disease	4
A	Bronchial washings	Tuberculosis	1c
A	Bronchial washings	To exclude tuberculosis	4
A	Bronchoscopy	Lung cancer (abnormal chest x ray)	1c
A	Bronchoscopy	Lung cancer (haemoptysis)	1c
A	Bronchial biopsy	Lung cancer	1b
A	Bronchial brush for cytology	Lung cancer	1b
A	Serial chest x rays	To exclude lung cancer if stable	None
A	Computed tomography bone densitometry	Corticosteroid induced osteoporosis	4
A	Electrocardiogram	To explore a cardiac cause of breathlessness	1b
A	Eosinophil count	To help diagnose asthma	4
A	Erythrocyte sedimentation rate	To help diagnose tuberculosis	None
A	Exercise stress test	Ischaemic heart disease	1a
A	Exercise oximetry	Presence of lung disease	None
A	Flow volume loop	Upper airway obstruction	4
A	Fine needle aspiration of lung mass (transcutaneous)	Lung cancer	1b
A	Heaf test	Tuberculosis	4
A	High resolution computed tomography	Interstitial lung disease	3
A	High resolution computed tomography	Bronchiectasis	1b
A	Serum immunoglobulin E concentration	To help diagnose asthma	None
A	Overnight pulse oximetry	Sleep apnoea syndrome	1b
A	Peak flow chart	Asthma	1b
A	Pleural fluid protein concentration	To differentiate exudate/transudate	1a
A	Ventilation-perfusion scan	Pulmonary embolus	1b
A	Venogram	Deep venous thrombosis	4
B	Serum angiotensin converting enzyme concentration	To follow activity of sarcoidosis	None
B	Peak flow chart	Asthma: to step down treatment if low variability	None
B	Oral corticosteroid trial	Asthma: to define best possible lung function	None
B	Peak flow chart	COPD: to treat with inhaled steroid if variable	None
B	Computed tomography scan of thorax	Lung cancer: to assess operability	1a
C	Trial of inhaled corticosteroid	COPD	‡
C	Trial of oral corticosteroid	COPD	‡
C	Trial of inhaled β agonist	COPD	None
C	Trial of oral theophylline	COPD	None
C	Trial of inhaled long acting β agonist	COPD	None
C	Trial of nasal corticosteroid	Cough	None
C	Trial of oral proton pump inhibitor	Cough	None

COPD=chronic obstructive pulmonary disease.
*A = tests used to make or exclude diagnosis; B = tests used to assess prediagnosed condition; C = trials of therapy used explicitly to predict future response in individual patient.
† "None" indicates no studies found of test in this context.
‡Studies that were found did not support use of trials of therapy to predict future response.

What is already known on this topic

Correct interpretation of test results requires information from scientific studies of test performance

If the studies do not meet quality standards the value of the test tends to be overestimated

What this study adds

Many diagnostic tests and tests used to monitor disease are not supported by high quality evidence

been published,[7] and when these standards are not met studies have been shown to overestimate the value of tests.[11] Many of the trials of diagnostic tests that are available fall short of these standards.

In 1996-7 only 30% of studies in one survey met at least six of eight standards[11] and a similar survey in 1990-3 gave a figure of only 18%.[12] Studies that evaluate diagnostic tests are also relatively rare. In a search of four prominent journals over a period of 16 years only 112 studies gave information on sensitivity, specificity, or likelihood ratios derived from more than 10 participants.[13] It is therefore not surprising that a survey of 300 clinicians in a range of different specialties found that only 4% used formal methods to assess the accuracy of tests and 1% utilised likelihood ratios.[14] Only half of the common tests we identified were supported by level 1a-1c evidence. We have also shown that there is little evidence to support tests that were used to assess previously diagnosed chronic diseases. The use of therapeutic trials to predict long term efficacy from short term response was similarly unsupported.

Our study reflects the practice in a single unit and the proportion of evidence based tests used elsewhere may be higher. Nevertheless, there is a clear need for further high quality research into medical tests, at least in the specialty that we have studied. There is also a need for an evidence base for the use of trials of therapy.

Contributors: PS had the original idea for the study. PJS and PS designed the study. PJS and ZB surveyed case notes, performed literature searches, and graded evidence. CH surveyed case notes. All authors commented on drafts. PJS is guarantor and can provide further details of the evidence found.

Funding: None.

Competing interests: None declared.

1 McSherry R, Haddock J. Evidence-based health care: its place within clinical governance. *Br J Nurs* 1999;8:113-7.
2 Guyatt GH, Sackett DL, Cook DJ. Users'guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? Evidence-based medicine working group. *JAMA* 1993;270:2598-601.
3 Ellis J, Mulligan I, Rowe J, Sackett DL. Inpatient general medicine is evidence based. A-Team, Nuffield Department of Clinical Medicine. *Lancet* 1995;346:407-10.
4 Gill P, Dowell AC, Neal RD, Smith N, Heywood P, Wilson AE. Evidence based general practice: a retrospective study of interventions in one training practice. *BMJ* 1996;312:819-21.
5 Geddes JR, Game D Jenkins NE Peterson LA Pottinger GR Sackett DL. What proportion of primary psychiatric interventions are based on evidence from randomised controlled trials? *Qual Health Care* 1996;5:215-7.
6 Howes N, Chagla L, Thorpe-M, and McCulloch-P. Surgical practice is evidence based. *Br J Surg* 1997;84:1220-3.
7 Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-based medicine working group. *JAMA* 1994;271:389-91.
8 Ransohoff D, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;299:926-30.
9 Haynes RB, Wilczynski NL, McKibbon MA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in Medline. *J Am Med Inform Assoc* 1994;1:447-58.
10 Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? Evidence-based medicine working group. *JAMA* 1994;271:703-7.
11 Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JHP, et al. Empirical evidence of design related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-6.
12 Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research: getting better but still not good. *JAMA* 1995;274:645-51.
13 Reid C, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. *JAMA* 1995;274:645-51.
14 Reid C, Lane DA, Feinstein AR. Academic calculation versus clinical judgements: practicing physicians' use of quantitative measures of test accuracy. *Am J Med* 1998;104:374-80.

(Accepted 4 September 2003)

ELPS

This is an abridged version; the full version is on bmj.com

Effects on pregnancy outcome of changing partner between first two births: prospective population study

Lars J Vatten, Rolv Skjærven

Department of Public Health and General Practice, Norwegian University of Science and Technology, NO-7489 Trondheim, Norway
Lars J Vatten *professor*

continued over

BMJ 2003;327:1138–41

Abstract

Objective To compare the effects on pregnancy outcomes of changing partner between the first two births with having the same partner for both births.
Design Prospective population study.
Setting Norway.
Participants 31 683 women who changed partner between their first two births and 456 458 women with the same partner for both births.
Results After adjustment for maternal age and education, interval between births, and decade of birth, the risk of adverse pregnancy outcomes was higher for the second birth for women who changed partner between their first two births compared with those who had the same partner for both births: preterm birth (< 37 weeks; relative risk 2.0, 95% confidence interval 1.9 to 2.1), low birth weight (< 2500 g; 2.5, 2.3 to 2.6), and infant mortality (1.8, 1.6 to 2.1). For the first birth, the risk of these adverse pregnancy outcomes was only slightly higher for mothers who subsequently had a second birth with another partner.
Conclusion Women who change partner between their first two births are at an increased risk of delivering a preterm, low birthweight baby with an increased risk of infant mortality compared with women who have the same partner for both births.

Introduction

A consequence of the increase in remarriages in Western societies is that a growing number of women are having children with different partners. The risk of