Internal Medicine Journal 2005; 35: 178–187

REVIEW ARTICLE

Beauty is in the eye of the examiner: reaching agreement about physical signs and their value

A. M. JOSHUA,[1,2] D. S. CELERMAJER[2,3] and M. R. STOCKLER[1,2,4]

[1]*Department of Medical Oncology, Sydney Cancer Centre,* [2]*Central Clinical School,* [3]*Department of Cardiology, Royal Prince Alfred Hospital and* [4]*National Health and Medical Research Council Clinical Trials Centre, School of Public Health, University of Sydney, Sydney, New South Wales, Australia*

Abstract

Despite advances in other areas, evidence-based medicine is yet to make substantial inroads on the standard medical physical examination. We have reviewed the evidence about the accuracy and reliability of the physical examination and common clinical signs. The physical examination includes many signs of marginal accuracy and reproducibility. These may not be appreciated by clinicians and could adversely affect decisions about treatment and investigations or the teaching and examination of students and doctors-in-training. We provide a selected summary of the reliability and accuracy as well as important messages of key findings in the physical examination. (Intern Med J 2005; 35: 178–187)

Key words: physical examination, accuracy, reliability, Kappa, sensitivity, specificity.

INTRODUCTION

Forty years ago, up to 88% of all primary care diagnoses were made on history and clinical examination,[1] and even 20 years ago up to 75% of all diagnoses in a general medicine clinic were made using these tools.[2] Although these percentages may be even lower in recent years, the physical examination will always retain its importance as the most common diagnostic test used by doctors and an essential tool of modern practice. However, its accuracy and reliability have not been scrutinized with the same rigorous standards applied to other diagnostic modalities (which usually rely on calibrated machines and technology rather than on 'artful' physicians). Reliability and accuracy are two different measures – the findings of two doctors might agree (be reliable) yet be wrong (inaccurate) when objectively assessed. By formally reviewing various aspects of the physical examination, we hope to give clinicians a greater degree of appreciation of its real value, allowing them to refine their examination technique and, therefore, their requesting and prioritizing of appropriate investigations. As previous reviewers have found, the evidence in this area varies and there are relatively few studies of high quality.[3]

Correspondence to: Dr Anthony Joshua, Princess Margaret Hospital, 610 University Avenue, Toronto, M5G 2M9, Canada.
Email: Anthony.Joshua@doctor.com

Received 12 February 2004; accepted 14 July 2004.

Funding: None

Potential conflicts of interest: No author declares any competing interest in the publication of this article.

METHODS

We searched MEDLINE by using an iterative strategy of combining the keywords 'Kappa', 'sensitivity', 'specificity' or 'likelihood ratio' with the MESH subject heading 'physical examination', as well as the keywords of the various physical examination techniques – 'inspection', 'palpation', 'auscultation' or 'percussion' – along with various physical examination findings such as 'heart sounds', 'ascites', 'hepatomegaly', 'palsy' and 'paraesthesia'. Reference lists were scanned from previous meta-analyses, systematic reviews and major textbooks of the physical examination and of internal medical subspecialties.[4–6] Articles were limited to those in the English language.

ANALYTIC TERMINOLOGY

Kappa is an index that describes the level of agreement beyond that expected by chance alone and can be thought of as the chance-corrected proportional agreement. Possible values range from +1 (perfect agreement) via 0 (no agreement above that expected by chance) to –1 (complete disagreement). As a guide, the level of agreement reflected by a Kappa value of 0.0–0.2 is 'slight', 0.21–0.40 is 'fair', 0.41–0.60 is 'moderate', 0.61–0.80 is 'substantial' and 0.81–1.00 is 'almost perfect'.[7] The need for this statistic arises because of the substantial rate of agreement that arises by chance alone. For example, if two physicians each consider half the cases they see abnormal, then they will agree 25% of the time by chance alone. The drawback of Kappa is that it varies with prevalence; the level of agreement expected by chance varies according to the proportion of cases considered abnormal across observers. Thus, a low

Kappa in a sample with a low prevalence (e.g. 10%) does not reflect the same lack of agreement as the same Kappa in a sample with a moderate prevalence (e.g. 50%).[8] The number of possible response categories of a test also influences Kappa – a dichotomy (present or not) will give a higher Kappa value than a scale with more than two levels.[9] Thus, comparisons of Kappas across studies must be interpreted carefully. Nevertheless, most of the medical literature on interobserver reliability over the last two decades has been reported in terms of Kappa rates and it remains a useful summary measure of agreement.

Likelihood ratio

This is the probability that a test is positive in those with a disorder divided by the probability the test is positive in those without the disorder. A likelihood ratio (LR) greater than 1 gives a post-test probability that is higher than the pretest probability. A LR less than 1 produces a post-test probability that is lower than the pretest probability. When the pretest probability lies between 30 and 70%, test results with a high LR (say, greater than 10) make the presence of disease very likely and test results with a low LR (say, less than 0.1) make the presence of the disease very unlikely.

If a test is either positive or negative, one can calculate the LR for a positive test result (positive likelihood ratio) and the LR for a negative test result (negative likelihood ratio) from the sensitivity and specificity;

$$LR+ = \text{Sensitivity}/(1 - \text{Specificity})$$
$$LR- = (1 - \text{Sensitivity}/\text{Specificity})$$

Specificity

This is the probability that a finding is absent in people who do not have the disease. Highly specific tests, when positive, are useful for ruling a disease in (the mnemonic is **Spin**: **sp**ecific rules **in**).

Sensitivity

This is the probability that a finding is present in people that have the disease. Highly sensitive tests, when negative, are useful for ruling a disease out (the mnemonic is **Snout**: **sen**sitive rules **out**).

CARDIOVASCULAR EXAMINATION

Clubbing

The old adage that compares clubbing to pregnancy – 'Decide if it's present or not – there's no such thing as early clubbing' – is incorrect: a recent review has identified three variables, that is, profile angle, hyponychial angle, and phalangeal depth ratio, which can be used as quantitative indices to identify clubbing.[10] There are no angles that define clubbing, only its absence. Kappa values of 0.39–0.90[11–13] suggest that this sign will remain in clinical practice.

Atrial fibrillation

The sensitivity and specificity of an 'irregularly irregular' pulse for atrial fibrillation has never been formally assessed. However, Rawles and Rowland. looked at the R-R intervals and pulse volumes with Doppler in 74 patients with atrial fibrillation and found there were periods of pulse rate regularity in 30% and pulse volume regularity in over 50%.[14]

Blood pressure

There are minute-to-minute physiologic variations of 4 mmHg systolic and 6–8 mmHg diastolic within patients.[15,16] With respect to the examiners as the source of variability, differences of 8–10 mmHg have been reported frequently for both physicians and nurses;[17,18] this is the same order of magnitude achieved by several commonly used antihypertensive agents.[19]

The jugular venous pressure/central venous pressure

The first challenge in assessing jugular venous waveform is finding it and there is only sparse information on how well this is done. In one study, examiners 'found' a jugular venous pressure (JVP) in only 20% of critically ill patients.[20] Prediction of the central venous pressure (CVP) by assessing the JVP has been more extensively studied. One study found physicians correctly predicted the CVP only 55% of the time in an intensive care setting.[21] In another study, medical students, residents and attending physicians were asked to examine the same 50 intensive care unit patients and estimate their CVP. Agreement between the students and residents was surprisingly high (Kappa of 0.65), moderate between students and physicians (Kappa of 0.56), and lowest between residents and staff physicians (Kappa of 0.30).[22] There were also substantial interobserver and intraobserver variations of up to 7 cm in estimations of the CVP.

In another study of 62 patients undergoing right heart catheterization, various medical staff predicted whether four variables, including the CVP, were low, normal, high, or very high. The sensitivity of the clinical examination for identifying a low (< 0 mmHg), normal (0–7 mmHg), or high (> 7 mmHg) CVP was 33, 33 and 49%, respectively. The specificity of the clinical examination for identifying low, normal, or high CVP was 73, 62, and 76%, respectively. Interestingly, accuracy was no better for cases in which agreement among examiners was high.[23]

The carotid pulse

It is easy to agree on the presence of a carotid bruit (Kappa = 0.67) but not its character (Kappa < 0.40).[24] The North American Symptomatic Carotid Endarterectomy Trial showed that over a third of high-grade carotid stenoses (70–99%) had no detectable bruits and a focal ipsilateral carotid bruit had a sensitivity of only 63% and a specificity of 61% for high-grade stenosis. These unhelpfully moderate values therefore give equally unhelpful LR: the odds of high-grade stenosis are only doubled by the presence of a carotid bruit, and only halved by the absence of a carotid bruit – not nearly enough to confidently rule in or out this important pathology.[25,26]

Systolic murmurs

Clinicians agree more often when they listen to systolic murmurs on audiotapes (Kappa of 0.48)[27] than when they listen to them 'live' in a clinical environment (Kappa of 0.30). This even applies to finding loud systolic murmurs and late-peaking murmurs: Kappa of 0.74 for audiotapes, but only 0.29 on the wards.[28,29]

As a guide to the examination of systolic murmurs, a small but important study used two clinicians to assess 50 patients with systolic murmurs and the findings are summarized (Table 1).[30]

Aortic stenosis

Examples of some of the various characteristics of aortic stenosis in two high-quality studies and their LR are summarized (Table 2).[31–33] These studies were of dramatically different populations: 781 elderly nursing home patients versus 231 cardiology patients referred for the catherization. The large LR noted in the first study might be related to a large number of asymptomatic patients with no audible murmur and thus might be more reflective of the community rather than hospitalized populations.

Other systolic murmurs

Tricuspid regurgitation and mitral regurgitation have positive LR of 10.1 and 3.6–3.9, respectively, and negative LR of 0.41 and 0.12–0.34 for their underlying conditions when compared to echocardiography.[33,34]

Diastolic murmurs

Just hearing aortic incompetence increases the likelihood that it is haemodynamically significant[35,36] but other associated signs such as an Austin-Flint murmur, Corrigan's pulse or widened pulse pressure are not related to severity.[37–41]

Mitral stenosis is notoriously difficult to hear; less than 10% of residents and medical students correctly identified a mid-diastolic murmur of mitral stenosis on an audiotape,[42] whereas 43% of medical residents identified the mid-diastolic murmur of mitral stenosis using a patient simulator.[43] As for the signs of severity – the loudness of the pulmonic closure sound (P2) is more closely related to how thin the patient is than to pulmonary artery pressure. There is also little correlation between the loudness of the first heart sound (S1) and the severity of mitral stenosis, presumably because of mitral valve calcification.[44]

Lower limbs

General findings are summarized (Table 3)[45–48]. There are also several other physical signs that are supposed to help localize vascular disease; for example, an abnormal femoral pulse predicted aortoiliac disease with 100% specificity, but only 38% sensitivity.[49] Put another way,

Table 1 Sensitivity and specificity of various manoeuvres for systolic murmurs†

Murmur	Sensitivity	Specificity
RSM – Augmentation with inspiration	100	88
RSM – Diminution with expiration	100	88
HOCM – Increase with Valsalva manoeuvre	65	96
HOCM – Increase with squatting-to-standing action	95	84
HOCM – Decrease in intensity with standing-to-squatting action	95	85
HOCM – Decrease in intensity with passive leg elevation	85	91
HOCM – Decrease in intensity with handgrip	85	75
MR/VSD – Augmentation with handgrip	68	92
MR/VSD – Augmentation with transient arterial occlusion	78	100

†All murmurs were confirmed by either cardiac catheterization (41 patients, 82%) or unequivocal anatomical echocardiographic findings (nine patients, 18%). HOCM, hypertrophic obstructive cardiomyopathy; MR, mitral regurgitation; RSM, right-sided murmurs; VSD, ventricular septal defect.

Table 2 Likelihood ratios for physical signs associated with aortic stenosis[31,32]

Murmur	Likelihood ratio positive	Likelihood ratio negative
Slow rate of rise of carotid pulse	130 (33–560)	0.62 (0.51–0.75)
	2.8 (2.1–3.7)	0.18 (0.11–0.30)
Late peak murmur intensity	101 (25–410)	0.31 (0.22–0.44)
Decreased intensity or absent second heart sound	50 (24–100)	0.45 (0.34–0.58)
	3.1 (2.1–4.3)	0.36 (0.26–0.49)
Fourth heart sound	2.5 (2.1–3.0)	0.26 (0.14–0.49)
Reduced carotid volume	2.3 (1.7–3.0)	0.31 (0.21–0.46)
Presence of any murmur	2.4 (2.2–2.7)	0 (0.00–0.13)
Radiation to right carotid	1.4 (1.3–1.5)	0.1 (0.13–0.40)
	1.5 (1.3–1.7)	0.05 (0.01–0.2)

Table 3 Findings in examination of the vascular system for peripheral vascular disease

Sign	Gold standard	Likelihood ratio positive	Likelihood ratio negative	Reference
Abnormal foot pulses	AAI of (a) < 0.9 (b) < 0.5	(a) 45.0 (b) 3.5	(a) 0.4 (b) 0.1	45
Femoral arterial bruits	AAI < 0.8–0.9	4.7–5.7	0.7–0.8	46,47
Prolonged capillary refill time > 5 s	AAI < 0.5	1.6–1.9	0.8–0.9	48
Trophic changes	AAI < 0.5	1.4–1.6	0.7–0.8	48

AAI, Ankle–arm index

Table 4 Kappa rates for the respiratory examination

Sign	Kappa	Reference
Dullness to percussion	0.52	12
Wheezes	0.43–0.93	52–55
Chest expansion	0.38	12
Bronchial breath sounds	0.32	12
Crackles	0.30–0.63	52–55
Cough	0.29	51
Tachypnoea	0.25	50
Breath sound intensity	0.23–0.46	52–55
Tactile fremitus	0.01	12

Table 5 Clinical findings in Ascites[70,71]

Finding	Sensitivity (%)	Specificity (%)
Flank dullness	94	29
Shifting dullness	85	56
Bulging flanks	75	40–70
Fluid wave	50–53	82–90

all people without aortoiliac disease had a normal femoral pulse (100% specificity) but only 38% with aortoiliac disease had an abnormal femoral pulse.

RESPIRATORY EXAMINATION

General findings in the respiratory examination are summarized in Table 4.[50–55]

Deep venous thrombosis

The clinical signs of pain, tenderness, oedema, Homan's sign, swelling and erythema have sensitivities of 60–88% and specificities of 30–72% in well-designed studies, using venography as the reference standard.[56] Studies of Homan's sign suggest it is positive from 8 to 56% of people with proven deep venous thrombosis (DVT), but also positive in more than 50% of symptomatic people without DVT.[57–60]

GASTROENTEROLOGICAL EXAMINATION

Hepatomegaly

In studies using reference standards, the ability to palpate a liver is not closely correlated with either liver size or volume.[61–64] In fact, the chance that a palpable liver meets reference standards for enlargement is 46%.[65] The LR for hepatomegaly, given a palpable liver is a modest 2.5, and the LR for hepatomegaly in the absence of a palpable liver is 0.45.[66]

Theodossi et al. had five observers perform a structured history and physical examination on 20 jaundiced patients and reported only moderate agreement in detecting the presence or absence of hepatomegaly (Kappa of 0.30).[67] Perhaps this Kappa rate is related to the findings of Meyhoff et al. who examined 23 patients (and had all measurements carried out from a set midclavicular line) and found the average maximum interobserver difference in measurement from the costal margin was 6.1 cm (standard deviation 2.7 cm).[68] The ability to detect more subtle characteristics of the liver, such as its consistency, nodularity and tenderness, shows considerable variability: Kappas of 0.11, 0.26 and 0.49, respectively.[67,69]

Ascites

In a set of 20 jaundiced patients, the Kappa value for ascites is 0.63; however, this would undoubtedly be lower in a less selected group.[67] The sensitivities and specificities of various signs are summarized (Table 5).[70.71]

Splenomegaly

Percussing for the spleen (via the Castell method – in the supine position, in the lowest intercostal space in the left anterior axillary line) compared to the gold standard of scintigraphy had an impressive sensitivity of 82% and specificity of 83% in one study,[72] but disappointing Kappa rates in another study (0.19–0.41).[73] As for palpation, if it is part of a routine examination, then the sensitivity was reported at an even more disappointing 27%, but specificity was impressively high at 98% with Kappa values ranging from 0.56 to 0.70.[74,75] An elegant study by Barkun *et al.* demonstrated that palpation is a significantly better discriminator among patients in whom percussion was already positive because when percussion dullness was present, palpation ascertained splenomegaly correctly 87% of the time. However, if percussion dullness was absent, palpation was little better than tossing a coin.[75] This implies that if percussion note is resonant over the spleen, then it is not worth palpating for splenomegaly.

NEUROLOGICAL EXAMINATION

Clinical examination of the nervous system is challenging to perform, difficult to master, and apparently almost impossible to replicate. The 'Emperor's clothes syndrome' is a phrase coined to reflect the fact that doctors may be influenced to report a sign as present on the basis of previous information.[76] This phenomenon is alive and well in neurology. Van Gijn *et al.* showed that the interpretation of equivocal Babinski responses was significantly affected by the history presented with films of the reflex.[77] It is not surprising then that Kappa values for the Babinski response range from 0.17 to 0.59.[78,79]

Interobserver variation has also been studied extensively for other neurological signs. Shinar *et al.* had six neurologists examine 17 patients and found that abnormal extraocular movements were the signs they agreed on most (Kappa of 0.77).[80] Surprisingly, they agreed only a little more often about peripheral motor abnormalities (Kappa of 0.36–0.67) than about sensory abnormalities (Kappa of 0.28–0.60). Hansen *et al.* had two senior neurologists and two trainees examine 202 consecutive unselected inpatients for eight different neurological signs. They found Kappa coefficients for neurologists ranged from 0.40 to 0.67 and from 0.22 to 0.81 for trainees. The neurologists had higher Kappa values than the trainees for five of the eight signs but the difference was only statistically significant for jerky eye movements. The highest agreement was for facial palsy (Kappa of 0.71) and the lowest was for grading of knee jerks (Kappa of 0.32).[78] Maschot *et al.* evaluated two graded scales for the assessment of four common tendon reflexes, with two or three physicians judging four common reflexes in two groups of 50 patients. They found that the highest Kappa was only 0.35 and concluded that numerical grading of tendon reflexes is not useful, and recommended a simplified verbal classification.[81] This was in contrast to a previous study, which found better reliability with Kappa values of 0.43–0.8.[82] This latter study found no improvement in Kappas after attempted standardization of the examination technique.

DISCUSSION

The physical examination should remain a mainstay of clinical assessment. However, our objective analysis raises doubts about the use of much of the clinical examination. Doctors need to elicit relevant signs reliably, accurately, and with an explicit understanding of their implications. Guidelines emphasize the need to understand the contexts of individual clinical skills,[83] but the scientific properties and basis of the physical examination seem to be notions that are rarely taught or acknowledged.[84] Certainly, an understanding of the examination is at least as important as experiencing it – as Dr Tinsley Harrison used to say, 'Clinical Experience is like military experience' and then quote Napoleon, 'My mule has more military experience than any general I have faced, yet still is unfit to lead an army'.[6]

Need for more research of the clinical examination

Our findings suggest considerable variation in reported estimates of sensitivity, specificity and reproducibility of many of the time-honoured physical signs. More research is needed to better define those signs that demonstrate high levels of reliability and accuracy. Sackett identified five reasons for the paucity of studies about the physical examination: (i) the difficulties in designing and doing studies of the physical examination, (ii) the difficulty of analysing a single sign when a diagnosis is made up of constellations of symptoms and signs, (iii) academic staff showing little inclination to investigate the physical examination as they spend little time at the bedside, (iv) the realities and pressures of modern medicine discouraging a careful history and physical examination and (v) the unpopularity of research when it challenges authority and the 'art of medicine'.[85]

Decline in bedside teaching of clinical examination

The wide variations in interobserver agreement for different signs probably reflect a combination of factors including the complexity of manoeuvres, difficulty in perceiving signs, differences in training and experience, and increasing reliance on diagnostic tests. However, evidence suggests that more and better teaching would help. One study found that in some US teaching hospitals, only 11% of available time on rounds was spent at the bedside with 63% spent in the conference room.[86] La Combe reported that bedside teaching comprised 75% of teaching 30 years ago, but had decreased to 16% by 1978.[87] Mangione and Nieman studied postgraduate trainees and found that they recognized less than half of all respiratory auscultatory pathologies and 20% of all cardiac auscultatory findings with little or no improvement according to years of training.[88,89] Previous work by the same investigators found that only 10% of US medical residence programmes offered structured teaching of pulmonary auscultation[90] and only 25% offered structured teaching of cardiac auscultation.[91] This diminishing emphasis on bedside teaching is evident when intern and resident staff are observed carrying out the physical examination. Wray and Friedland found observation of junior staff by attending physicians showed error rates of about 15% with incorrect findings in 4% and missed findings in 10%.[92]

Improving performance by standardizing techniques and examining more frequently

These errors reflect the combination of variations within and among clinicians, patients, and circumstances. There are only two ways to minimize these errors: reduce the variation or increase the number of observations. The key to reducing variation is greater standardization of the most robust techniques and a better understanding of the examination technique and its failings. How best to increase the number of observations depends on the nature of the variation. For example, within patient variation is the major source of error in diagnosing hypertension, and this is overcome by having the same clinician measure blood pressure on several occasions.[93]

Implications for junior doctor assessment

Our observations and summary of accuracy and reliability (Tables 6–8)[23,27,29,31,39,46–49,52,67,70,71,74,80,94–113] have important implications for the assessment of junior doctors.

For example, the emphasis in many 'short-case' board and membership examinations is on eliciting a sign or performing a type of examination whereas in 'long-case' examinations candidates are often not given the opportunity to integrate the patient's history and clinical findings with investigations before discussing the case with examiners. The interobserver error inherent in eliciting signs, their variable relationship to disease and the fact that clinical diagnosis rarely relies on isolated

Table 6 Comparisons of Kappa values for common clinical signs

Sign	Kappa value	References
Heberden's nodes	0.78	94
Abnormality of extra-ocular movements	0.77	80
Size of goitre by examination	0.74	95
Forced expiratory time	0.70	52
Presence of goitre by inspection	0.65	95
Signs of liver disease (e.g. jaundice, Dupytren's contracture, spider naevi)	0.65	96
Palpation of the posterior tibial pulse	0.60	97
Presence of a displaced apical impulse	0.53–0.73	98
Palpation of dorsalis pedis pulse	0.51	97
Tender liver edge	0.49	67
Hearing a systolic murmur	0.30–0.48	27, 29
Detecting oedema in cardiac failure	0.27–0.64	98
Leuconychia	0.27	96
Clinical breast examination for cancer	0.22–0.59	99
Third heart sound	0.10–0.50	100
Neck stiffness	–0.01	80

Table 7 Examples of sensitivities and specificities for common clinical signs

Sign	Underlying Condition	Sensitivity	Specificity	Reference
An apical impulse lateral to the mid-clavicular line	Left ventricular enlargement	100	18	101
Systolic murmur radiating to neck	Severe aortic stenosis†	90–98	22–36	31,102
Shifting dullness	Ascites	85	50	70,71
S3	Ejection fraction < 30%	78	88	103
Sustained apical movement	Detecting severe aortic stenosis in patients with aortic flow murmurs	78	81	104
Sustained left lower parasternal movement	Detecting right ventricular peak pressure of 50 mmHg or more	71	80	105
Examination of goitre	Presence of goitre	70	82	95
Heberden nodes	Generalized osteoarthritis	68	52	94
Abnormal foot pulses	Peripheral vascular disease	63–95	73–99	46
Palpable spleen‡	Splenomegaly	58	92	74
Pulse pressure > 80 mmHg	Moderate to severe aortic regurgitation§	57	95	39
Clinical breast examination	Detecting breast cancer	54	94	99
S3	Ejection fraction < 50%	51	90	103
Paradoxic splitting of heart sounds	Significant aortic stenosis	50	79	31
Murphy's sign	Cholecystitis	50–97	80	106–108
Trophic skin changes	Peripheral vascular disease	43–50	70	49
Phalens test	Carpal tunnel syndrome	40–90	80	109
Initial impression	COPD	25	95	52
Prolonged capillary refill	Peripheral Vascular Disease	25–28	85	49
Tinel's sign	Carpal Tunnel Syndrome	25–75	75–90	110–112
Hepatojugular reflux	Congestive cardiac failure	24–33	95	23
Femoral arterial bruit	Peripheral vascular disease	20–29	95	47,48
Kernigs sign	Meningitis	5	95	113

†Peak gradiant > 50 mmHg. ‡Specifically examined for. §Using angiography as the gold standard.

Table 8 Summary of key points

DO
Cardiovascular
· Do try to elicit hepatojugular reflux. It is useful sign of cardiac failure when other signs may be equivocal, that is, highly specific
· Do listen for an S3 to confirm your suspicion of cardiac failure – also highly specific
· Do ventilatory manoeuvres (such as held inspiration and expiration) to confirm your suspicions about the nature of cardiac murmurs
Gastrointestinal
· Do percuss the splenic bed before palpating for the splenomegaly
· Do be reasonably confident that a patient with a positive Murphy's sign has cholecystitis
Respiratory
· Do count the respiratory rate carefully because clinicians often disagree about tachypneoa
DON'T
Cardiovascular
· Don't rely on clinical assessment of carotid pulse character or carotid bruits. Don't hesitate to order carotid imaging for symptoms in the absence of signs
· Don't rely on the commonly quoted signs of severity for aortic incompetence.
Gastrointestinal
· Don't waste your time screening for ascites with shifting dullness; looking at the abdomen for bulging flanks is just as accurate
· Don't decide a liver is enlarged on the basis of palpation alone
Respiratory
· Don't depend on tactile fremitus – your colleagues are unlikely to agree
Other
· Don't hesitate to repeat your examination or get a colleague to repeat it if you are unsure

findings argue against this piecemeal approach to assessment. However, it is reassuring, at least for some, that physicians with greater training and experience often,[114,115] but not always,[116] agree more frequently than physicians with less training and experience in a particular task.

Future directions

Where to from here? Ramani *et al.* conducted focus groups from the Boston University School of Medicine's faculty to address the diminishing time spent teaching the physical exam. They suggested four solutions: (i) improve bedside teaching skills through faculty training in clinical skills and teaching methods, (ii) reassure clinical faculty that they possess more than adequate bedside skills to educate trainees, (iii) establish a learning environment that allows teachers to admit their limitations and (iv) address the undervaluing of teaching on a department level with adequate recognition and rewards.[117]

At a personal level, acknowledging the uncertainty inherent in the physical examination is the first step to improving it. Medical staff can optimize their examination skills by being well rested,[118] by examining in an appropriate environment, by trying to avoid being influenced by their expectations, by examining patients on more than one occasion to verify findings,[119] and by documenting findings and progress. There should be no shame in asking a colleague to verify physical findings. Large multinational studies that will provide high-quality data about the reliability of the physical examination are underway and eagerly awaited.[120,121] A greater appreciation of the complexities of patient assessment and a clearer understanding of its limitations should lead to a more rational and cost-effective approach to diagnosis, investigation and decision-making.

FURTHER READING

The most complete evidence-based review of the physical examination is *Evidence-Based Physical Diagnosis* by Steven McGee (W B Saunders; 1st edition) but perhaps the most entertaining overview is *Sapira's Art and Science of Bedside Diagnosis* (Lippincott Williams & Wilkins; 2nd edition). Alternatively readers are directed to the excellent and ongoing JAMA series, *The Rational Clinical Examination.*

McGee S (2001). *Evidence-Based Physical Diagnosis.* WB Saunders, Philadelphia.

Orient JM (2000). *Sapira's Art and Science of Bedside Diagnosis*, 2nd edn. Lippincott Williams & Wilkins, Philadelphia.

JAMA – The Journal of the American Medical Association (from 1992 continuing). *The Rational Clinical Examination.* JAMA, Chicago.

REFERENCES

1 Crombie DL. Diagnostic process. J Coll Gen Pract 1962; 6: 579–89.
2 Sandler G. The importance of the history in the medical clinic and the cost of unnecessary tests. Am Heart J 1980; 100: 928–31.
3 Simel D, Drummond R. The clinical examination: an agenda to make it more rational. JAMA 1997; 277: 572–4.
4 Talley N, O'Connor S. Examination Medicine – A Guide to Physician Training, 4th edn. Sydney: MacLennan, Petty; 2001.
5 Yusuf S, Cairns JA, Camm AJ, Fallen EL, Gersh BJ, eds. Evidence-Based Cardiology, 2nd edn. Philadelphia: BMJ Books; 2003.
6 Sapira JD. The Art and Science of Bedside Diagnosis, 1st edn. Baltimore, MD: Williams & Wilkins; 1990.
7 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159–74.
8 Hansen M, Sindrup SH, Christensen PB, Olsen NK, Kristensen O, Friis ML. Interobserver variation in the evaluation of neurological signs: oberserver dependent factors. Acta Neurol Scand 1994; 90: 145–9.

9 Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall; 1996.
10 Myers KA, Farquhar DR. The rational clinical examination. Does this patient have clubbing? JAMA 2001; 286: 341–7.
11 Pyke DA. Finger clubbing. Lancet 1954; 2: 352–4.
12 Spiteri MA, Cook DG, Clarke SW. Reliability of eliciting physical signs in examination of the chest. Lancet 1988; 1: 873–5.
13 Smyllie HC, Blendis LM, Armitage P. Observer disagreement in physical signs of the respiratory system. Lancet 1965; 2: 412–13.
14 Rawles JM, Rowland E. Is the pulse in atrial fibrillation irregularly irregular? Br Heart J 1986; 56: 4–11.
15 Armitage P, Rose GA. The variability of measurements of casual blood pressure, I. a laboratory study. Clin Sci 1966; 30: 325–35.
16 Hypertension Detection and Follow-up Program. Variability of blood pressure and the results of screening in the Hypertension Detection and Follow-up Program. J Chron Dis 1978; 31: 651–67.
17 Scherwitz LW, Evans LA, Hennrikus DJ, Vallbona C. Procedures and discrepancies of blood pressure measurements in two community health centers. Med Care 1982; 20: 727–38.
18 Neufeld PD, Johnson DL. Observer error in blood pressure measurement. Can Med Assoc J 1986; 135: 633–7.
19 Smith D, Neutel J, Lacourcière Y, Kempthorne-Rawson J. Prospective, randomized, open-label, blinded-endpoint (PROBE) designed trials yield the same results as double-blind, placebo-controlled trials with respect to ABPM Measurements. J Hypertens 2003; 21: 1291–8.
20 Davison R, Cannon R. Estimation of central venous pressure by examination of jugular veins. Am Heart J 1974; 87: 279–82.
21 Eisenberg PR, Jaffe AS, Schuster DP. Clinical evaluation compared to pulmonary artery catheterization in the hemodynamic assessment of critically ill patients. Crit Care Med 1984; 12: 549–53.
22 Cook DJ. The clinical assessment of central venous pressure. Am J Med Sci 1995; 299: 175–8.
23 Connors AF, McCaffree DR, Gray BA. Evaluation of right heart catheterization in the critically ill patient without acute myocardial infarction. N Engl J Med 1983; 308: 263–7.
24 Chambers BR, Norris JW. Clinical significance of asymptomatic neck bruits. Neurology 1985; 35: 742–5.
25 Sauve JS, Thorpe KE, Sackett DL, Taylor W, Barnett HJM, Haynes RB *et al.* Can bruits distinguish high-grade from moderate symptomatic carotid stenosis? Ann Intern Med 1994; 120: 633–7.
26 Hankey GJ, Warlow CP. Symptomatic carotid ischaemic events: safest and most cost effective way of selecting patients for angiography, before carotid endarterectomy. BMJ 1990; 300: 1485–91.
27 Dobrow RJ, Calatayud JB, Abraham S, Caceres CA. A study of physician variation in heart sound interpretation. Med Ann Dist Columbia 1964; 33: 305–8.
28 Forssell G, Jonasson R, Orinius E. Identifying severe aortic valvular stenosis by bedside examination. Acta Med Scand 1985; 218: 397–400.
29 Taranta A, Spagnuolo M, Snyder R, Gerbarg DS, Hofler JJ., Enslein K. Data Acquisition and Processing in Biology and Medicine, Vol. 3. New York: Macmillan Publishing Co Inc; 1964: 23–52.
30 Lembo NJ, Dell'Italia LJ, Crawford MH, O'Rourke RA. Bedside diagnosis of systolic murmurs. N Engl J Med 1988; 318: 1572–8.
31 Hoagland PM, Cook EF, Wynee J, Goldman L. Value of noninvasive testing in adults with suspected aortic stenosis. Am J Med 1986; 80: 1041–50.
32 Aronow WS, Kronzon I. Correlation of prevalence and severity of aortic regurgitation detected by pulsed doppler echocardiography with the murmur of aortic regurgitation in elderly patients in a long-term health care facility. Am J Cardiol 1989; 63: 128–9.
33 Etchells E, Bell C. Does this patient have an abnormal systolic murmur. JAMA 1997; 277: 564–71.
34 Rahko PS. Prevalence of regurgitant murmurs in patients with valvular regurgitation detected by Doppler echocardiography. Ann Intern Med 1989; 111: 466–72.
35 Meyers DG, McCall D, Sears TD, Olson TS, Felix GL. Duplex pulsed Doppler echocardiography in mitral regurgitation. J Clin Ultrasound 1986; 14: 117–21.
36 Grayburn PA, Smith MD, Handshoe R, Friedman BJ, De Marie AN. Detection of aortic insufficiency by standard echocardiography, pulsed Doppler echocardiography, and auscultation. Ann Intern Med 1986; 104: 599–605.
37 Bloomfield DA, Sinclair-Smith BC. Aortic insuffficency: a physiological and clinical appraisal. South Med J 1973; 66: 55–65.
38 Cohn LH, Mason DT, Ross J, Morrow AG, Braunwald E. Preoperative assessment of aortic regurgitation in patients with mitral valve disease. Am J Cardiol 1967; 19: 177–82.
39 Frank MJ, Casanegra P, Migliori AJ, Levinson GE. The clinical evaluation of aortic regurgitation. Arch Intern Med 1965; 116: 357–65.
40 Spagnuolo M, Kloth H, Taranta A, Doyle E, Pasternack B. Natural history of aortic regurgitation: criteria predictive of death, congestive heart failure, and angina in young patients. Circulation 1971; 44: 368–80.
41 Smith HJ, Neutze JM, Roche HG, Agnew TM, Barratt-Boyes BG. The natural history of rheumatic aortic regurgitation and the indications for surgery. Br Heart J 1976; 38: 147–54.
42 Mangoine S, Nieman LZ. Cardiac auscultatory skills of internal medicine and family practice trainees. JAMA 1997; 278: 717–22.
43 St Clair EW, Oddone EZ, Waugh RA, Corey GR, Feussner JR. Assessing housestaff diagnostic skills using a cardiology patient simulator. Ann Intern Med 1992; 117: 751–6.
44 Surawicz B, Mercer C, Chlebus H, Reeves J, Spencer F. Role of the phonocardiogram in evaluation of the severity of mitral stenosis and detection of associated valvular lesions. Circulation 1966; 34: 795–806.
45 McGee SR, Boyko EJ. Physical examination and chronic lower-extremity ischaemia. A Critical review. Arch Intern Med 1998; 158: 1357–64.
46 Christensen JH, Freundlich M, Jacobsen BA, Falstie-Jensen N. Clinical relevance of pedal pulse palpation in patients suspected of peripheral arterial insufficiency. J Intern Med 1989; 226: 95–9.
47 Criqui MH, Fronek A, Klauber MR, Barrett-Connor E, Gabriel S. The sensitivity, specificity and predictive value of traditional clinical evaluation of peripheral arterial disease: results from noninvasive testing in a defined population. Circulation 1985; 71: 516–22.
48 Stoffers HEJH, Kester ADM, Saiser V, Rindens PELM, Knottnerus JA. Diagnostic value of signs and symptoms associated with peripheral arterial occlusive disease seen in general practice: a multivariable approach. Med Decis Making 1997; 17: 61–70.
49 Boyko EJ, Ahroni JH, Davignon D, Stensel V, Prigeon RL, Smith DG. Diagnostic utility of the history and physical examination for peripheral vascular disease among patients with diabetes mellitus. J Clin Epidem 1997; 50: 659–68.
50 Metlay JP, Kapoor WN, Fine MJ. Does this patient have community acquired pneumonia? Diagnosing pneumonia by history and physical examination. JAMA 1997; 278: 1440–45.
51 Holleman DR Jr,, Simel DL, Goldberg JS. Diagnosis of obstructive airways disease from the clinical examination. J Gen Intern Med 1993; 8: 63–8.
52 Schapira RM, Schapira MM, Funahashi A, McAuliffe TL, Varkey B. The value of forced expiratory time in the physical diagnosis of obstructive airways disease. JAMA 1993; 270: 731–6.
53 Badgett RG, Tanaka DJ, Hunt DK, Jelley MJ, Feinberg LE, Steiner JF *et al.* Can moderate chronic obstructive pulmonary

disease be diagnosed by historical and physical findings alone? Am J Med 1993; 94: 188–96.
54 Nairn JR, Turner-Warwick M. Breath sounds in emphysema. Br J Dis Chest 1969; 63: 29–37.
55 Mulrow CD, Dolmatch BL, Delong ER, Feussner JR, Benyunes MC, Dietz JL *et al.* Observer variability in the pulmonary examination. J Gen Intern Med 1986; 1: 364–7.
56 Anand SS, Wells PS, Hunt D, Brill-Edwards P, Cook D, Ginsberg JS. Does this patients have deep vein thrombosis? JAMA 1998; 279: 1094–9.
57 Hirsh J, Hull RD, Raskob GE. Clinical features and diagnosis of venous thrombosis. J Am Coll Cardiol 1986; 8 (6 Suppl. B): 114B–27B.
58 Wheeler HB. Diagnosis of deep vein thrombosis. Review of clinical evaluation and impedance plethysmography. Am J Surg 1985; 150: 7–13.
59 Johnson WC. Evaluation of newer techniques for the diagnosis of venous thrombosis. J Surg Res 1974; 16: 473–81.
60 O'Donnell TF, Abbott WM, Athanasoulis CA, Millan VG, Callow AD. Diagnosis of deep venous thrombosis in the out patient by venography. Surg Gynaecol Obstet 1980; 150: 69–74.
61 Naffalis J, Leevy CM. Clinical examination of liver size. Am J Dig Dis 1963; 8: 236–43.
62 Ralphs DNL, Venn G, Kan O, Palmer JG, Cameron DE, Hobsley M. Is the undeniably palpable liver ever 'normal'? Ann R Coll Surg Engl 1983; 65: 159–60.
63 Sullivan S, Krasner N, Williams R. The clinical estimation of liver size: a comparison of techniques and an analysis of the source of the error. Br Med J 1976; 2: 1042–3.
64 Halpern S, Coel M, Ashburn W, Alazraki N, Littenburg R, Hurwitz S *et al.* Correlation of liver and spleen size: determinations by nuclear medicine studies and physical examination. Arch Intern Med 1974; 134: 123–4.
65 Rosenfeld AT, Laufer I, Schneider PB. The significance of a palpable liver. Am J Roentgenol Radiat Ther Nucl Med 1974; 122: 313–17.
66 Naylor CD. Physical examination of the liver. JAMA 1994; 271: 1859–65.
67 Theodossi A, Knill-Jones RP, Skene A, Lindberg G, Bjerregaard B, Holst-Christensen J *et al.* Inter-observer variation of symptoms and signs in jaundice. Liver 1981; 1: 21–32.
68 Meyhoff HH, Roder O, Anderson B. Palpatory estimation of liver size. Acta Chir Scand 1979; 145: 479–81.
69 Espinoza P, Ducot B, Pelletier G, Attali P, Buffet C, David B *et al.* Interobserver agreement in the physical diagnosis of alcoholic liver disease. Dig Dis Sci 1987; 32: 244–7.
70 Cummings S, Papadakis M, Melnick J, Gooding GA, Tierney LM Jr. The predictive value of physical examination for ascites. West J Med 1985; 142: 633–6.
71 Cattau EL, Benjamin SB, Knuff TE, Castell DO. The accuracy of physical examination in the diagnosis of suspected ascites. JAMA 1982; 247: 1184–6.
72 Sullivan S, Williams R. Reliability of clinical techniques for detecting splenic enlargement. BMJ 1976; 2: 1043–4.
73 Barkun AN, Camus M, Meagher T, Green L, Coupal L, De Stempel J *et al.* Splenic enlargment and Traube's space: how useful is percussion? Am J Med 1989; 87: 562–6.
74 Grover SA, Barkun A, Sackett DL. Does this patient have splenomegaly? JAMA 1993; 270: 2218–21.
75 Barkun AN, Camus M, Green L, Meagher T, Coupal L, De Stempel J *et al.* The bedside assessment of splenic enlargement. Am J Med 1991; 91: 512–18.
76 Gross F. The Emporer's Clothes syndrome. N Engl J Med 1971; 285: 863.
77 Van Gijn J, Bonke B. Interpretation of plantar reflexes: biasing effect of other signs and symptoms. J Neurol Neurosurg Psychiatry 1977; 40: 787–9.
78 Hansen M, Sindrup SH, Christensen PB, Olsen NK, Kristensen O, Friis ML. Interobserver variation in the evaluation of neurological signs: observer dependent factors. Acta Neurol Scand 1994; 90: 145–9.
79 Maher J, Reilly M, Hutchinson M. Planter Power: reproducibility of the planter response. BMJ 1992; 304: 482.
80 Shinar D, Gross CR, Mohr JP, Caplan LR, Price TR, Wolf PA *et al.* Interobserver variability in the Assessment of Neurologic History and Examination in the Stroke Data Bank. Arch Neurol 1985; 42: 557–65.
81 Manschot S, van Paseel L, Buskens E, Algra A, van Gijn J. Mayo and NINDS scales for assessment of tendon reflexes: between observer agreement and implications for communication. J Neurol Neurosurg Psychiatry 1998; 64: 253–5.
82 Litvan I, Mangone CA, Werden W, Bueri JA, Estol CJ, Garcea DO *et al.* Reliability of the NINDS Myotactic Reflex Scale. Neurology 1996; 47: 969–72.
83 George JH, Doto FX. A simple five-step method for teaching clinical skills. Fam Med 2001; 33: 577–8.
84 Koran LM. The reliability of clinical methods data, judgments (second of two parts). N Engl J Med 1975; 293: 695–701.
85 Sackett DL, Renniwe D. The science of the art of clinical examination. JAMA 1992; 267: 2650–2.
86 Miller M, Johnson B, Greene HL, Baier M, Nowlin S. An observational study of attending rounds. J Gen Intern Med 1992; 7: 646–8.
87 LaCombe M. On bedside teaching. Ann Intern Med 1997; 126: 217–20.
88 Mangoine S, Nieman LZ. Pulmonary auscultatory skills during training in internal medicine and family practice. Am J Respir Crit Care Med 1999; 159: 1119–24.
89 Mangione S, Nieman LZ. Cardiac auscultatory skills of internal medicine and family practice trainees: a comparison of diagnostic proficiency. JAMA 1997; 278: 717–22.
90 Mangione S, Loudon RG, Fiel SB. Lung auscultation during internal medicine and pulmonary training: a nationwide survey. Chest 1993; 104: 70S.
91 Mangione S, Nieman LZ, Gracely E, Kaye D. The teaching and practice of cardiac auscultation during internal medicine and cardiology training: a nationwide survey. Ann Intern Med 1993; 119: 47–54.
92 Wray NP, Freidland JA. Detection and correction of house staff error in physical diagnosis. JAMA 1987; 249: 1035–7.
93 Perloff D, Grim C, Flack J, Frolich ED, Hill M, McDonald M *et al.* Human blood pressure determination by sphygmomanometry. Circulation 1993; 88: 2460–67.
94 Cicuttini FM, Baker J, Hart DJ, Spector TD. Relation between Heberden's nodes and distal interphalangeal joint osteophytes and their role as markers of generalized disease. Ann Rheum Dis 1998; 57: 246–8.
95 Siminoski K. Does this patient have a goiter? JAMA 1995; 273: 813–37.
96 Espinoza P, Ducot B, Pelletier G, Attali P, Buffet C, David B *et al.* Interobserver agreement in the physical diagnosis of alcoholic liver disease. Dig Dis Sci 1987; 32: 244–7.
97 Meade TW, Gardner MJ, Cannon P, Richardson PC. Observer variability in recording the peripheral pulses. Br Heart J 1968; 30: 661–5.
98 Gadsboll N, Hoilund-Carlsen PF, Nielsen GG, Berning J, Bruun NE, Hein E. Interobserver agreement and accuracy of bedside estimation of right and left ventricular ejection fraction in acute myocardial infarction. Am J Cardiol 1989; 63: 1301–7.
99 Berton MB, Harris R, Fletcher SW. The rational clinical examination: does this patient have breast cancer? The screening clinical breast examination: should it be done? How? JAMA 1999; 282: 1270–80.

100 Ishmail AA, Wing S, Ferguson J, Hutchinson TA, Magder S, Flegel KM. Interobserver agreement by auscultation in the presence of the third heart sound in patients with congestive heart failure. Chest 1987; 91: 870–73.
101 Ellen SD, Crawford MH, O'Rourke RA. Accuracy of precordial palpation for detecting increased left ventricular volume. Ann Intern Med 1983; 99: 628–30.
102 Hoagland PM, Cook EF, Wynne J, Goldman L. Value of noninvasive testing in adults with suspected aortic stenosis. Am J Med 1986; 80: 1041–50.
103 Patel R, Bushnell DL, Sobotka PA. Implications of an audible third heart sound in evaluating cardiac function. West J Med 1993; 158: 606–9.
104 Forssell G, Jonassen R, Orinius E. Identifying severe aortic valvular stenosis by bedside examination. Acta Medica Scandinavica 1985; 218: 397–400.
105 Gillam PMS, Deliyannis AA, Mounsey JPD. The left parasternal impulse. Br Heart J 1964; 26: 726–36.
106 Adedeji OA, McAdam WA. Murphy's sign, acute cholecystitis and elderly people. J R Coll Surg Edinb 1996; 41: 88–9.
107 Singer AJ, McCracken G, Henry MC, Thode HC Jr, Cabahug CJ. Correlation among clinical, laboratory and hepatobiliary scanning findings in patients with suspected acute cholecystitis. Ann Emerg Med 1996; 28: 267–72.
108 Singh A, Bagga SP, Jindal VP, Singh K, Rato SS. Gall bladder disease: an analytical report of 250 cases. J Indian Med Assoc 1989; 87: 253–6.
109 Kuschner SH, Ebramzadeh E, Johnson D, Brien WW. Sherman Tinel's sign and Phalen's test in carpal tunnel syndrome. Orthopaedics 1992; 15: 1297–1302.
110 Stewart JD, Eisen A. Tinel's sign and the carpal tunnel syndrome. BMJ 1978; 2: 1125–6.
111 Katz JN, Larson MG, Sabra A, Krarup C, Stirrat CR, Sethi R *et al.* The carpal tunnel syndrome: diagnostic utility and physical examination findings. Ann Intern Med 1990; 112: 321–7.
112 Gelmers HJ. The significance of Tinnel's sign in the diagnosis of carpal tunnel syndrome. Acta Neurochirurgica (Wien) 1979; 49: 225–58.
113 Thomas KE, Hasbun R, Jekel J, Quagliarello VJ. The diagnostic accuracy of Kernig's sign, Brudzinski's sign and nuchal rigidity in adults with suspected meningitis. Clin Infect Dis 2002; 35: 46–52.
114 Raftery EB, Holland WW. Examination of the Heart: an investigation into variation. Am J Epidemiol 1967; 85: 438–44.
115 Tomasello F, Mariani F, Fieschi C, Argentino C, Bono G, De Zanche L *et al.* Assessment of interobserver differences in the Italian multicenter study on reversible cerebral ischaemia. Stroke 1982; 13: 32–5.
116 Vogel H-P. Influence of additional information on interrater reliability in the neurologic examination. Neurology 1992; 42: 2076–81.
117 Ramani S, Orlander JD, Strunin L, Barber TW. Whither bedside teaching? A focus-group study of clinical teachers. Acad Med 2003; 78: 384–90.
118 Friedman RC, Bigger JT, Koranfield DS. The intern and sleep loss. N Engl J Med 1971; 285: 201–3.
119 Stockler M, Tannock I. Comment on Angulo *et al.*: 'A Man's Got To Know His Limitations'. Urol Oncol 1995; 1: 206–7.
120 McAlister FA, Straus SE, Sackett DL. Why we need large, simple studies of the clinical examination: the problem and a proposed solution. CARE-COAD1 group. Clinical Assessment of the Reliability of the Examination-Chronic Obstructive Airways Disease Group. Lancet 1999; 354: 1721–4.
121 Straus SE, McAlister FA on behalf of the CARE Group. Collaborative Studies of the Accuracy and Precision of the Clinical Examination. Toronto: University of Toronto [updated Dec 2004; cited 24 Jul 2003]. Available from: http://www.carestudy.com