

Assessing trainees in the workplace: results of a pilot study

Gavin Johnson, Joe Booth, Jim Crossley and Winnie Wade

ABSTRACT – This paper outlines the development and evaluation of the utility of workplace-based assessments in higher medical training: case-based discussion (CbD); the acute care assessment tool (ACAT); audit assessment; teaching observation and patient survey (PS). The study population included trainees in higher medical training (ST3+) from physician specialties in the UK. The pilot consisted of a prospective study of the use of the new assessments using local study coordinators (LSCs) and volunteer trainees. In total, 169 LSCs were recruited and 134 trainees returned at least one assessment. The end-of-pilot questionnaire was returned by 44 assessors and 57 trainees. Questionnaire data and qualitative feedback were used to evaluate the validity, impact and feasibility of the new tools. For adequate reliability (co-efficient 0.7) a total of 12 CbDs; three ACATs and 16 PS raters are required. There was evidence for the validity and positive educational impact of all the tools. There were difficulties with the feasibility of the PS.

KEY WORDS: acute care assessment tool, audit assessment, case-based discussion, patient survey, teaching observation

Introduction

In 2007, the Joint Royal Colleges of Physicians Training Board (JRCPTB) agreed the assessment system for each physician specialty with the Postgraduate Medical Education and Training Board (PMETB). During this process a number of areas of the curricula were identified as representing a challenge to assess by existing means. The assessment methods, identified as being important in the assessment of the curricula, included established workplace-based assessments (WPBAs), such as direct observation of procedural skills (DOPS), a mini-clinical evaluation exercise (mini-CEX) and multi-source feedback (MSF). However, effective assessment of curricula areas, such as communication skills, clinical audit, practice of acute medicine and teaching, required the use of new assessment methods.

PMETB's standards¹ require assessment methods to be selected on the basis of their overall utility, made up of: validity, reliability, feasibility, cost effectiveness, opportunities for feedback and impact on learning. Validity is the degree to which a test measures what it purports to measure. The reliability of an assessment refers to the reproducibility of the scores obtained

from an assessment. Reliability is generally expressed as a coefficient ranging from 0 (not reliable) to 1 (perfect reliability). For very high stakes examinations an R coefficient of at least 0.9 is recommended, whereas 0.7 is considered sufficient for formative assessments such as WPBAs.²

A pilot project was conducted between 2008 and 2009 with the objectives of providing an evidence base for the effectiveness of modified existing WPBAs (case-based discussion (CbD), and the acute care assessment tool (ACAT)) and new WPBAs (audit assessment (AA), patient survey (PS), teaching observation (TO)) in the assessments of doctors in higher medical training:

- Case-based discussion – discussion between a trainee and assessor about a case the trainee has been involved in to explore the trainee's clinical reasoning and decision-making skills.
- Audit assessment – the quality of a clinical audit process completed by a trainee is assessed by one or more reviewers. There was no existing audit assessment tool in use in the UK.
- Teaching observation – a trainee is given qualitative feedback by an observer following a formal teaching session, which has not been subject to assessment in the past, in the context of training curricula.
- Patient survey – patients are invited to complete a brief questionnaire on the quality of a trainee's professionalism and communication skills following an outpatient clinic encounter. The questionnaires are summarised and feedback is given to the trainee by a supervisor.
- Acute care assessment tool – an assessment of a trainee during a period of practising acute medicine considering the trainee's performance in the management of the take, patient management and teamworking.

Methods

Draft formats for AA and TO were developed de novo by the project team. Existing formats for CbD, PS and ACAT were used as the starting point. The PS form was one that had previously been piloted for consultant revalidation. Draft forms were presented and circulated to specialty representatives and revisions were made in light of feedback. Most assessment forms were based on the format of existing WPBA methods, ie a combination of scored domains and free text.

The ratings for CbD, AA and ACAT mainly used a six-point scale based on 'expectations for stage of training'. For the purposes of this study, the rating scale for overall performance on the ACAT was changed from a relative scale to an absolute rating with anchor statements describing the performance expected.

Gavin Johnson, consultant gastroenterologist, University College London Hospitals; **Joe Booth**, senior education projects manager, Royal College of Physicians, London; **Jim Crossley**, senior research fellow, Academic Unit of Medical Education, University of Sheffield; **Winnie Wade**, director of education, Royal College of Physicians, London

Teaching observation was designed purely to allow formative assessment and qualitative feedback without scored ratings.

Consultants from all physician specialties were recruited to act as local study coordinators (LSC). Each LSC was responsible for finding volunteer trainees, returning completed assessment forms, and forwarding an online questionnaire towards the end of the pilot. All trainees gave consent and all their submissions to the Royal College of Physicians (RCP) were anonymous. Trainees were asked to arrange assessments and to use any of the methods appropriate to their specialty.

Data gathered by the RCP were:

- assessment forms (each one of which asked for specific statements on the assessment process itself)
- feedback forms on the PS process from trainees and their supervisors
- responses from assessors and trainees to an online questionnaire which was forwarded to participants by the LSCs.

For PS, the individual forms for each trainee were transcribed and collated by the project team and a summary of the responses was returned to the trainee's supervisor. All data from returned assessments were extracted into Excel spreadsheets and the SPSS statistics package for analysis. The reliability of the CbD, ACAT and PS were calculated using generalisability theory.³ The TO and AA data were not amenable to calculation of reliability, as explained below. Trainee and assessor feedback from the forms and questionnaire was used to determine the validity, educational impact and feasibility of each tool. A modified nominal group technique was used to reach consensus on the themes emerging from the qualitative feedback.⁴ The ACAT has been subject to a previous evaluation that highlights its educational impact and feasibility, so these data were not gathered for this study.⁵

Results

Case-based discussion

A total of 111 trainees and 140 assessors undertook at least one CbD, and a total of 231 CbDs were undertaken. The CbDs most

frequently focused on an outpatient record (97/231, 42%) or inpatient record (84/231, 36%), and discharge summaries were also used. When the time was recorded almost two thirds of CbDs took less than 20 minutes (135/215, 63%). In total, 78% of CbDs were rated as either 'above' or 'well above expectations' for overall clinical judgement and no CbD scores were 'below' or 'well below expectations'. A total of 12 CbDs was required to achieve a reliability coefficient $R=0.70$.

The assessor and trainee were invited to enter qualitative comments on each assessment form about the documentation and process of the CbD. Table 1 shows the themes emerging when the individual forms were analysed, together with the frequency with which a comment fitting with a certain 'theme' was made on a CbD form. These comments provided strong support for the educational impact of CbDs.

End-of-pilot questionnaire responses were received from 42 trainees and 36 assessors. The great majority of respondents (89–90%) either 'agreed' or 'strongly agreed' that the CbD enabled trainees to demonstrate clinical reasoning, decision making, knowledge around case and patient management. There was no consensus on whether or not it was straightforward to arrange discussions – two-thirds of trainees (67%) agreed or strongly agreed that it was straightforward but assessors' opinions were divided (38% agreed, 44% disagreed).

When considering the impact of the CbD, almost two thirds of respondents (62%) agreed or strongly agreed that the CbD usually highlighted areas for development. Almost three quarters of trainees (74%) felt that they usually learned something during the CbD and there was consensus from assessors that they were usually able to teach during a CbD (94% agreed or strongly agreed).

Acute care assessment tool

A total of 74 ACATs were completed (by 30 different trainees and 50 assessors). Almost three-quarters of ACATs took up to 20 minutes (53/74, 72%). Predominantly assessments were based on the assessor's opinion of the trainee's performance during the take period based on the post-take ward round (46/74, 62%). An assessor's direct observation of the trainee's performance on the

Table 1. Themes emerging from participants using the case-based discussion (CbD) forms (the numbers reflect the frequency with which a comment was made within the theme, from the 231 CbDs undertaken).

Positive themes	Assessors	Trainees	Negative themes	Assessors	Trainees
Useful/effective (<i>Educational impact</i>)	41	38	Not useful (<i>Educational impact</i>)	5	9
Good educational value/generates deep discussion (<i>Educational impact</i>)	24	30	Not suitable for specialty (<i>Validity</i>)	8	6
Enables feedback (<i>Educational impact</i>)	2	13	More paperwork/time-consuming (<i>Feasibility</i>)	6	3
Easy to use (<i>Feasibility</i>)	8	4	Grading difficult (<i>Reliability</i>)	9	0
Fits with current practice (<i>Feasibility</i>)	5	5	Subtlety not captured (<i>Validity</i>)	6	2
Fair (<i>Validity</i>)	8	1	Does not reflect competence (<i>Validity</i>)	2	1
Better than other WPBAs (<i>General</i>)	0	9			
Encourages reflection (<i>Educational impact</i>)	5	3			

WPBAs = workplace-based assessments.

take formed the basis of the assessment in 30% of ACATs (22/74). A total of three ACATs were required to achieve a reliability coefficient $R=0.70$, however these calculations were based on a relatively small number of assessments and require corroboration in a greater number of ACATs.

End-of-pilot questionnaire responses were received from 15 trainees and 12 assessors. The majority of responders indicated that the ACAT enabled trainees to demonstrate clinical reasoning, clinical assessment, management planning and management of critically ill patients. There was no consensus as to whether the ACAT enabled trainees to display competence in leadership and management of the take. A small majority of responders indicated that the ACAT was usually straightforward to organise. When considering the impact of the ACAT, almost half of responders agreed that ACATs highlighted areas for development. Just over half of trainees (60%) felt that they learned something during ACATs and two-thirds of assessors agreed or strongly agreed that they were able to deliver teaching during an ACAT assessment.

Audit assessment

A total of 106 AAs were completed (from 79 trainees and 88 assessors). Most assessments were based on a presentation of the audit (63/106, 59%), a review of a written report only (21/106, 20%) or both a presentation and a report (8/106, 8%). The time taken for assessors to evaluate the audit varied quite markedly from less than 10 minutes (23/106, 22%) to over half an hour (24/106, 23%), reflecting that the assessor would often watch a presentation of the audit as part of the assessment. The feedback given for the AA took less than 20 minutes for the vast majority of assessments (85/106, 80%). A total of 67% of assessments were rated as 'above' or 'well above expectation' for 'overall quality', and 32% were scored at 'meets expectation'. None of the AAs reflected a rating 'below' or 'well below expectation'. Few trainees were assessed by more than one assessor and very few assessors rated more than one audit. This meant that the audit data were not appropriate for reliability analysis.

From the questionnaire, 27 trainees and 16 assessors responded to questions on AA. The predominant view of trainees was that the tool enabled them to demonstrate competence in clinical

audit (17/106, 63% agreeing), however the view of assessors was less clear – six agreed, eight neither agreed nor disagreed. When asked about the issues related to the feasibility of the audit, trainees predominantly felt that the assessment was straightforward to organise with 18 (67%) agreeing, and only three (11%) disagreeing. Assessors indicated that the process was little extra work above what is required for trainee supervision. Table 2 shows the themes emerging when the comments on the individual AA forms were analysed, showing support for the educational impact of AA.

Teaching observation

A total of 147 TOs were undertaken by 76 different trainees and 111 assessors. The tool was designed to facilitate structured feedback, and does not request any scores from the observer, hence it is not amenable to any reliability calculations. The mean time for the observation was 43 minutes, and 12 minutes for feedback. The number of learners in the teaching session varied from two to 150 with a median of 12.

From the questionnaire, 37 trainees and 28 assessors responded to questions on TO. Only 10 trainees (25%) had experienced a formal observation of their teaching before the study. A significant number of observers had formally observed someone else's teaching sessions previously (12/28, 43%). A slight majority of trainees felt that the sessions were straightforward to organise (20/37, 54%). This was echoed by the observers who felt that the exercise was little more effort beyond attending the session. There was also consensus that the feedback following a TO led to a positive impact on future sessions, with 26 (70%) trainees and 22 (85%) observers agreeing with this suggestion. Table 3 shows the themes emerging from the comments on the individual TO forms.

Patient survey

A total of 1,258 patients returned a form for assessment of 84 trainees. The number of patient responses for each trainee varied between 1 and 31, with a mean of 15. Trainees were asked to submit 20 patient responses, based on results of a previous

Table 2. Themes emerging from participants using the audit assessment (AA) (the numbers reflect the frequency with which a comment was made within the theme, from the 106 AAs undertaken).

Positive themes	Assessors	Trainees	Negative themes	Assessors	Trainees
Useful/effective (<i>Educational impact</i>)	21	14	Difficult to use/questions not clear (<i>Feasibility</i>)	9	4
Enables feedback (<i>Educational impact</i>)	6	17	Grading difficult/should not relate to stage of training (<i>Reliability</i>)	6	2
Easy to use (<i>Feasibility</i>)	5	3	Not useful (<i>Educational impact</i>)	4	2
Fits with current practice (<i>Validity</i>)	2	2	More paperwork/time-consuming (<i>Feasibility</i>)	5	
Encouraged good audit (<i>Educational impact</i>)	1	2	Difficult to represent audit on assessment form (<i>Validity</i>)	3	2
Good educational value (<i>Educational impact</i>)		2	Does not fit with local audit practice (<i>Validity</i>)	2	
Encourages reflection (<i>Educational impact</i>)		1			

study on the patient survey conducted with consultants (unpublished). Only 20 trainees (24%) achieved this response rate. The patient ratings were overwhelmingly positive, with 97% of responders rating their overall satisfaction with the doctor as either 'very' or 'fairly' satisfied. Responses from 16 patients are required for a reliability co-efficient of 0.7.

A total of 31 trainees, and 34 of the participating trainees' supervisors, returned a feedback form commenting on their experience using PS. There was agreement between the supervisors and trainees that the PS is practical (79% supervisor, 84% trainee), fair (88%, 90%) and useful for the trainee's development (79%, 90%). Trainees indicated that the assessment provided useful information for the trainee (90% agreeing with this statement). One quarter (24%) of supervisors felt that the assessment provided information about the trainee that they did not already know despite all the other sources of supervision and assessment information.

A total of 37 trainees and 16 assessors responded to questions on PS in the pilot questionnaire. Most trainees and supervisors agreed that PS did allow trainees to demonstrate communica-

tion skills (81% supervisor, 84% trainee) and patient-centred care (88%, 75%). Thirty per cent (11/37) of responding trainees agreed that they changed their behaviour because they were being assessed. Nineteen per cent (7/37) of trainees and 24% (4/16) of supervisors agreed that the PS highlighted things to do differently in future. Over half (19/37, 51%) of responding trainees gave out the forms and the clinic nurse or administrator collected them in, although 11% of trainees indicated that they gave out and then collected the forms themselves, contrary to instructions. The themes emerging from the feedback on the PS are shown in Table 4.

Discussion

The overall utility of each method investigated was considered. The questionnaire response varied for trainees between 34% (for AA) and 50% (ACAT), and for the assessors between 18% for (AA) and 26% (CbD). Hence it could be argued that some of the questionnaire data are poorly representative, particularly for assessors.

Table 3. Themes emerging from participants using the teaching observation (TO) (the numbers reflect the frequency with which a comment was made within the theme, from the 147 TOs undertaken).

Positive themes	Assessors	Trainees	Negative themes	Assessors	Trainees
Useful/good/effective (<i>Educational impact</i>)	28	26	Scoring system/overall rating would help (<i>Reliability</i>)	15	10
Enabled feedback (<i>Educational impact</i>)	5	28	Form needs simplification/clarification (<i>Feasibility</i>)	14	8
Easy (<i>Feasibility</i>)	7	1	Not valid for different types of teaching (<i>Validity</i>)	14	4
Helped shape structure of teaching (<i>Educational impact</i>)	4	2	Time consuming (<i>Feasibility</i>)	9	8
Identified learning points (<i>Educational impact</i>)	2	4	Too many points on form, too complicated (<i>Feasibility</i>)	8	1
Encouraged reflection (<i>Educational impact</i>)	2	1	Form was constraining (<i>Validity</i>)	6	
Good to have free text (<i>General</i>)	2		Difficult/laborious (<i>Feasibility</i>)	2	2
			More training/guidance needed (<i>Feasibility</i>)	2	1
			Not useful for assessment (<i>Validity</i>)	3	

Table 4. Themes emerging from trainees and assessors using the patient survey (the numbers reflect the frequency with which a comment was made within the theme, from the 84 participating trainees and their supervisors).

Positive themes	Supervisors	Trainees	Negative themes	Supervisors	Trainees
Useful (<i>Educational impact</i>)	4	4	Not discriminating/patients too polite/uncritical (<i>Reliability</i>)	5	7
Not difficult (<i>Feasibility</i>)	3	4	Trainee can select patients (<i>Reliability</i>)	5	4
Good to get feedback (<i>Educational impact</i>)	1	5	Trainees can alter behaviour (<i>Reliability</i>)	4	4
Confirmed impression (<i>Validity</i>)	5		Time consuming (<i>Feasibility</i>)	5	3
A poor result is useful (<i>Educational impact</i>)	2	1	Difficult to do/practical problems (<i>Feasibility</i>)	3	4
			Not representative/overrates trainee (<i>Reliability</i>)	2	2
			Not useful (<i>Educational impact</i>)		2
			Cannot link negative comments with encounter (<i>Educational impact</i>)	1	

Case-based discussion

The 12 CbDs required for the process to be sufficiently reliable are comparable to the mini-CEX.^{6,7} This study shows how infrequently respondents make critical comments about trainees in WPBA; none of the assessments reflected a score 'below' or 'well below' expectation. The validity of the CbD can be inferred from the feedback from trainees and assessors that it enabled competencies in clinical reasoning, decision making and knowledge to be demonstrated. This study has shown very positive educational impact – trainees indicated that they learn during an assessment and assessors are able to deliver teaching. The time taken and comments from trainees and assessors suggest that CbD can be feasibly incorporated into training, as a trainee is frequently discussing cases with a consultant, such as a new patient in clinic or an inpatient ward referral, and these all present opportunities for CbD.

Acute care assessment tool

There is already evidence of the positive educational impact and validity of the ACAT, which was reinforced by this pilot.⁵ The ACAT showed excellent reliability in this pilot, possibly because a new absolute rating scale was used. However, the overall numbers of ACATs were relatively low, and a further calculation of the reliability of the ACAT using a much greater number of assessments from e-portfolio data is planned to corroborate the findings of this pilot.

Audit assessment

The time taken, and the infrequency with which it could be expected, suggest that AA can be feasibly incorporated into assessment systems to assess competencies in understanding and conducting audit. Only 10% of assessors took longer than 20 minutes to give feedback. This is reflected in evidence from the questionnaire with the majority of trainees indicating that the AA was straightforward to organise, and a similar proportion of assessors indicating that the AA was little extra work beyond what is required of normal audit supervision.

It was not possible to calculate the reliability of the AA based on the pilot data and it may be difficult to demonstrate reliability in future, as it would be uncommon for a trainee to be involved in more than one audit each year. Hence the value of an AA is to provide evidence of trainee involvement and to provide a framework for feedback for quality improvement, rather than as a basis for making decisions on progress.

Trainees more frequently felt that the AA allowed a demonstration of their competence in clinical audit, whereas the dominant view of assessors was to neither agree nor disagree. The predominant themes emerging from the comments provided by participants were supportive of the educational impact of the tool and its role in encouraging feedback. Many assessors and trainees felt that the AA was useful and effective, and that they appreciated the opportunity to either give or receive feedback.

Teaching observation

There was strong support from observers and trainees for the educational impact of the TO and its role in encouraging valued feedback, identifying learning points for future practice, helping to shape the structure of the session, encouraging reflective practice on the trainee's performance and boosting confidence. The mean and median times spent observing teaching, along with comments from participants, suggest that it would be more feasible to expect assessors to undertake observations if these were teaching sessions that would normally be observed.

Patient survey

The PS reliability compares similarly to other published studies of using patient ratings in the assessment of doctors.^{8–12} This pilot demonstrated a striking leniency from patient raters, such that the PS has little discriminatory use and hence has diminished validity in its current format. Possible explanations for patient leniency are: self-selection for the study biasing the cohort towards high-performing trainees; only satisfied patients responding; patients holding doctors, in general, in high regard; doctors changing behaviour during PS; doctors handing out forms only after good interaction or to selected patients.

Trainees appreciated the feedback resulting from the PS. However, supervisors indicated that it did not provide information about the trainee that was not already apparent. The feasibility of the PS can also be questioned. Almost half the trainees (46%) did not manage to achieve the 16 ratings that the calculations demonstrate are required for acceptable reliability. For this pilot the transcription and collation of individual forms was carried out centrally at the RCP. This was time-consuming and difficult and unlikely to be a practical long-term option. Concerns were expressed from participants about the ability of trainees to influence the outcome of PS by selection of patients or by changing behaviour when being assessed. There are further complications for trainees working largely outside clinic-based settings. There is a climate for patient involvement in the assessment of trainee doctors but further work into developing a more feasible approach needs to be undertaken.^{13,14}

Summary

This study has shown very positive educational impact and good opportunities for feedback for all of the formative WPBAs investigated. The study also indicates that the methods are valid, feasible and cost effective, though there are practical problems with PS. Some evidence for reliability exists but more analysis of larger data sets is needed. Overall, the utility of these methods is strong enough to recommend their use in training. For ACAT, CbD and AA assessors appeared to be generous with ratings – the great majority of trainees being rated as performing above or well above expectations. The addition of the anchor statement on the ACAT form produced a striking improvement to reliability, and appeared to encourage more realistic ratings.

A substantial study of the reliability of WPBAs when the ratings are made using an absolute scale with anchor statements is currently being undertaken, to establish whether the reliability improvement for the ACAT's new rating scale is reproduced.

References

- 1 Postgraduate Medical Education and Training Board. *Standards for curricula and assessment systems*. London: PMETB, 2008. www.pmetb.org.uk/fileadmin/user/Standards_Requirements/PMETB_Scas_July2008_Final.pdf
- 2 Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ* 2004;38:1006–12.
- 3 Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. *Med Educ* 2002;36:972–8.
- 4 Jones J, Hunter D. Quantitative research: consensus methods for medical and health services research. *BMJ* 1995;311:376–80.
- 5 Johnson GJ, Barrett J, Jones M, Wade W. The Acute Care Assessment Tool: a workplace based assessment of the performance of a physician in training on the acute medical take. *Clinical Teacher* 2009;6:105–9.
- 6 Wilkinson JR, Crossley JGM, Wragg A *et al*. Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Med Educ* 2008;42:364–73.
- 7 Norcini JJ, Blank LL, Arnold GK *et al*. The miniCEX (clinical evaluation exercise): a preliminary investigation. *Ann Intern Med* 1995;123:795–9.
- 8 Greco M, Cavanagh M, Brownlea A, McGovern J. Validation studies of the doctors' interpersonal skills questionnaire. *Educ Gen Pract* 1999;10:256–64.
- 9 Crossley J, Eiser C, Davies HA. Children and their parents assessing the doctor-patient interaction: a rating system for doctors' communication skills. *Med Educ* 2005;39:757–9.
- 10 Nelson E, Gentry M, Mook K *et al*. How many patients are needed to provide reliable evaluations of individual clinicians? *Med Care* 2004;42:259–66.
- 11 Mercer SW, Howie JGR. CQI-1 a new measure of holistic, interpersonal care in primary care consultations. *Br J Gen Pract* 2006;56:262–8.
- 12 Mercer SW, McConnachie A, Maxwell M *et al*. Relevance and performance of the Consultation and Relation Empathy (CARE) measure in general practice. *Fam Pract* 2005;22:328–34.
- 13 Department of Health. *High quality care for all: NHS Next Stage Review final report*. London: DH, 2008. www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_085828.pdf
- 14 General Medical Council. *Principle, criteria and key indicators for colleague and patient questionnaires in revalidation*. London: GMC, 2010. www.gmc-uk.org/thewayahead

Address for correspondence: W Wade, Royal College of Physicians, 11 St Andrews Place, Regent's Park, London NW1 4LE.

Email: winnie.wade@rcplondon.ac.uk

RCP books

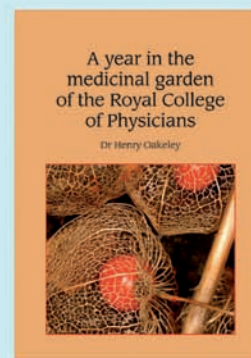
Garden diary by Dr Henry Oakley

A year in the medicinal garden of the Royal College of Physicians

The RCP medicinal garden contains nearly 1,000 different plants that are used, or have been used, in medicine during the past two millennia, and ones that commemorate physicians. This illustrated garden journal describes some of the plants as they flower, week by week, through the seasons. There is also a commentary on their uses, past and present, and space for your notes throughout the year.

Published: December 2009 **ISBN:** 978 1 86016 370 8

Price: £8.00 UK, £10.00 overseas (inc post and packing)



**Royal College
of Physicians**

10% discount for fellows and members

Quote the reference *Clinical Medicine* when making your order