Note to authors of letters: Please note that all letters submitted for publication should be typed with *double spacing*. Failure to comply with this may lead to delay in publication.

Patients in fundholding and non-fundholding practices

Sir,

To investigate the charge that the fundholding scheme channels resources to those areas least in need[1] a study was undertaken in 1994 to compare health need indicators for patients of fundholding and non-fundholding practices in the former East Anglian Regional Health Authority.

Copies of the part of patient registers containing details of age, sex, postcode and general practitioner registration were obtained from Cambridgeshire, Norfolk and Suffolk family health services authorities. The proportions of patients aged 0–4 years and 75 years and over were calculated for each practice. The distribution of each practice's patients in census wards was found by matching address postcodes to wards,[2] and 1991 census values for wards were converted into estimates for practices by multiplying ward values by the number of practice patients living in the ward, adding the products and dividing by the number of patients in the practice.[3] The health needs indicators derived from the 1991 census were long-term illness, the Townsend deprivation index,[4] and the Jarman underprivileged area score.[5] The prevalence of long-term illness was assessed in the 1991 census with the question: 'Does the person have any long-term illness, health problem or handicap which limits his/her activities or the work he/she can do?' The replies of people aged 75 years and over were excluded, and age was indirectly standardized using national rates in 10-year age bands. Both the Townsend and Jarman indices were calculated from East Anglian region ward means and standard deviations. In addition mortality data were obtained for 1989 to 1991 from the Office of Population Censuses and Surveys (Table VS4).[6] Ward standardized mortality ratios were calculated from national age-specific rates for people aged under 75 years, and these were also converted into estimates for practices. After excluding practices with fewer than 500 patients, practices were divided into fundholding (first to third wave) and non-fundholding practices.

There were 51 fundholding and 241 non-fundholding practices. Fundholding practices had significantly more patients than non-fundholding practices (Student's t test, $P<0.001$) (Table 1). Fundholding practices had a slightly higher proportion of patients aged 75 years and over compared with non-fundholding practices, but this difference was not statistically significant. Standardized illness and mortality ratios were low compared with the national average of 100, but these and the deprivation indices were similar between fundholding and non-fundholding practices.

Because the Townsend and Jarman indices were calculated from East Anglia ward data, the resulting indices are comparable only within the region and not with national values. The difference in size between fundholding and non-fundholding practices was to be expected since a minimum list size of 11 000 and subsequently 7000 patients was a condition for fundholding status. In East Anglia the patients of fundholding and non-fundholding practices have similar health needs.

Table 1. Health needs indicators for fundholding and non-fundholding practices

Indicator	Mean (standard deviation) in Fundholding practices		Non-fundholding practices	
No. of patients	10 580	(4005)	6282	(3467)
% aged 0-4 years	*6.2*	*(0.9)*	*6.3*	*(1.0)*
% aged 75+ years	*8.2*	*(2.2)*	*7.6*	*(2.1)*
Standardized illness ratio				
Males	83.9	(11.1)	82.1	(15.2)
Females	85.4	(10.6)	84.5	(13.6)
Standardized mortality ratio				
Males	85.8	(11.3)	85.4	(13.4)
Females	87.5	(12.3)	85.5	(12.9)
Townsend deprivation index	0.88	(2.04)	0.96	(2.80)
Jarman underprivileged area score	6.2	(10.3)	5.8	(13.7)

ROBIN HAYNES
ANDREW LOVETT
SUSAN GALE
GRAHAM BENTHAM

School of Environmental Sciences
University of East Anglia
Norwich NR4 7TJ

References

1. Coulter A. Fundholding general practices. *BMJ* 1992; **304:** 397-398.
2. Raper J, Rhind D, Shepherd J. *Postcodes: the new geography.* Harlow: Longman, 1992.
3. Majeed FA, Cook DG, Anderson HR, *et al.* Using patient and general practice characteristics to explain variations in cervical smear uptake rates. *BMJ* 1994; **308:** 1272-1276.
4. Townsend P, Phillimore P, Beattie A. *Health and deprivation: inequality and the north.* London: Croom Helm, 1988.
5. Jarman B. Underprivileged areas: validation and distribution of scores. *BMJ* 1984; **289:** 1587-1592.
6. Office of Population Censuses and Surveys. *Key population and vital statistics: series VS. Numbers 16,17,18.* London: HMSO, 1990, 1991, 1992.

Surveys of GPs: methodological considerations

Sir,

Sibbald and colleagues rightly emphasize that low response rates are a particular problem in postal surveys of general practitioners in the United Kingdom (July *Journal*, p.297). They reported a mean response rate in 26 surveys published in the *Journal* between January 1991 and June 1993 to be 61%.

Research shows that response rates are influenced by questionnaire length, use of a different method for a subsequent approach, and the investigating agency.[1,2] These factors have been taken into consideration when designing a study and our experiences may be useful for other researchers.

The study, funded by South Thames Project Grants Scheme, is a confidential postal survey of the general practitioner management of cervical chlamydia infec-

tion in 500 general practices in England and Wales. The questionnaire is short — one side of A4 with 10 questions — and takes approximately two minutes to complete. The introductory letter, signed by the researchers, was addressed to the practice manager requesting him or her to give it to the general practitioner most involved in cervical smear testing. For the follow up of non-respondents, the practice manager was asked to give the questionnaire to the general practitioner most likely to complete it; the follow-up letter to the general practitioner was signed by the head of department. The response rate after the first questionnaire was 50%, and after the second questionnaire the response rate had increased to 76%.

Although Sibbald and colleagues gained a good response from the telephone questionnaire, the costs of obtaining a completed questionnaire by telephone were approximately four times higher than by post. In addition, telephone surveys of general practitioners can be frustrating and time consuming for the researcher who finds the general practitioner's telephone number engaged, on answerphone or that the general practitioner is unavailable in surgery or out on visits.

DIANA MASON
PIPPA OAKESHOTT

Division of General Practice and Primary Care
St George's Hospital Medical School
Hunter Wing
Cranmer Terrace
London SW17 0RE

References

1. Heberlein TA, Baumgartner R. Factors affecting response rates to mailed questionnaires: a quantitative analysis of the published literature. *Am Soc Rev* 1978; **4:** 447-462.
2. Goyder JC. Further evidence on factors affecting response rates to mailed questionnaires. *Am Soc Rev* 1982; **47:** 550-553.

Leicester assessment package

Sir,

I am concerned that the authors of two papers[1,2] conclude that the Leicester assessment package can be recommended for assessment of consultation competence in general practice. However, the results presented lead to the conclusion that the Leicester assessment package cannot (yet) be recommended.

Any procedure passing or failing individuals will suffer from misclassification. That is, a proportion of individuals who should pass will be failed, and a proportion of individuals who should fail will pass. The question of how to establish exactly who should pass and fail is difficult — hence the need for assessment packages. A gold standard assessment might be approximated by a number of experienced assessors (say four) each assessing a large number of consultations (say 20) several times over a period of months (say three), and deciding finally in conference whether to pass or fail. Less intensive methods, such as the Leicester package, should be validated against such a gold standard (criterion validation) to establish sensitivity and specificity — a process which clearly requires detailed specification of how the pass/fail decision is to be made. Deciding how high sensitivity and specificity need to be, taking into account the different costs associated with low sensitivity (prejudicial to careers) and low specificity (prejudicial to the health of the public) is essentially a political decision, and requires an open debate. The face validation reported[1] in no way guarantees that the Leicester assessment package will be able to approximate the results of a gold standard satisfactorily.

Variability owing to temporal factors (that is repeatability) was not assessed, and could affect both trainees and assessors. If temporal variability is large (good days and bad days occur frequently and have a considerable effect on performance) achieving adequate reliability will require assessments to be made on several well-spaced occasions.

The authors attempted to assess the variability due to differing assessors, patients and doctors. The estimates of components of variance due to these sources were not presented, however, and are likely to be imprecise owing to the small sample sizes. Brennan emphasizes that the 'study design be as large as possible. Doing so helps ensure that the resulting estimated variance components will be as stable as possible'.[3] Moreover, samples of five or six are unlikely to be representative of parent distributions, and very unlikely to include significant representation from the tails of the distributions. But this is where likely causes for concern are: difficult patients, borderline doctors, eccentric assessors. As a result, there can be little confidence that the results obtained will be stable (that is, a repeat study might produce very different estimates).

The authors used (Cronbach's) alpha coefficients to assess internal consistency. Unfortunately, although often called a measure of internal consistency, alpha increases not only with consistency, but also with number of items. A matrix of correlation coefficients gives a much better idea of consistency. For a given value of alpha, and a given number of items, it is possible to estimate the mean correlation coefficient. For an assessor with alpha = 0.8 (intra-assessor consistency) the mean correlation between scores given for pairs of patients will be about 0.4.

One assessor was recognized as 'inconsistent' (with an estimated mean correlation coefficient of 0.04 corresponding to alpha = 0.22), and it was suggested that 'all assessors should be trained and calibrated before being sanctioned to assess real candidates...'. This may have the desired effect, but is hardly a firm basis for recommending the Leicester package. Evidence that training of assessors does improve consistency (and hence overall reliability) is needed.

The correct interpretation of alpha, and the generalizability coefficient, is as an indicator of the internal reliability of the mean of several items. One definition of alpha is the ratio of the true (with perfect assessment) subject (in this case, doctor) variance to the expected observed (with variation due to assessors and cases) subject variance. If the observed variance σ_o^2 is taken as equal to the true variance σ_t^2 plus a random error variance σ_e^2, then it is simple to see that $\sigma_e^2/\sigma_t^2 = (1/\alpha) - 1$ so that $\sigma_e/\sigma_t = \sqrt{((1/\alpha)-1)}$. Suppose that the Leicester assessment package were actually measuring exactly the same qualities as an idealized gold standard assessment (this needs establishing by criterion validation) — but with random measurement error added. If alpha = 0.8 then from above we have $\sigma_e/\sigma_t = 0.5$ that is the (random) error standard deviation is half the true subject standard deviation. Sensitivity and/or specificity will be relatively low. Suppose the true cut off (for pass/fail) were two true subject standard deviations below the true mean — so that the worst 2.3% of doctors should fail (assuming perfect assessment, and normality). A particular doctor whose true score is 0.5 standard deviation below the pass mark (that is in the worst 0.6%) will have a 16% chance of passing. Fairly difficult calculations show that the sensitivity will be 98%, but the specificity only 73% (that is more than one in four of those who should fail will in fact pass). The negative predictive value would be just 45% (that is only 45% of those who failed would actually deserve to fail).

This is why Nunnally[4] suggests that while for basic research purposes a reliability of 0.7 or higher is sufficient, for 'many applied settings a reliability of 0.8 is not nearly high enough... In those applied settings where important decisions are made with respect to specific test scores, a reliability of 0.90 is the min-