LETTERS TO THE EDITOR

Assessment of the SF-36 version 2 in the United Kingdom

EDITOR,—I read with interest the recent article detailing changes to the format of the SF-36.[1] The authors present data regarding the psychometric properties and epidemiological characteristics of the SF-36 version 2. The authors present the results from a large data sample of people aged 18–64. The analysis reveals that the questionnaire has good internal consistency and construct validity. The layout of the new questionnaire is certainly improved and in this respect I think that participants will find it easier to complete. However, I believe that many of the problems that were inherent in the original version have not been resolved. The validity and reliability of the questionnaire relies in part upon users completing it accurately. Any change in the questionnaire's format should be designed to improve the accuracy of users responses, which will in turn improve the psychometric qualities of the questionnaire. The authors concede that the present data are only based upon people of working age and so it remains unclear how suitable this measure is for older age groups. They suggest that further research is needed to determine how applicable the SF-36 is for this age group.

In my personal experience I would suggest that the SF-36 is not a suitable measure to use with older age groups. The main shortcoming with the questionnaire is not the layout but rather the language of the questions. I would be grateful for an opportunity to draw your attention to my experience of using this tool as an outcome measure with a large group of surgical patients. I have used the SF-36 with approximately 200 patients who were recruited to examine the effects of different vascular surgery procedures on quality of life and cognitive function. Patients were assessed before their operation and six months later. Quality of life was assessed using the SF-36 and the Hospital Anxiety and Depression Scale (HAD). The HAD scale is widely referred to in the psychiatric literature (reported sensitivity = 72–88% and specificity = 68–94%).[2-4] The patients in the first study were undergoing carotid endarterectomy (CEA), which is a prophylactic procedure carried out to reduce the risk of stroke. The second study examined the effects abdominal aortic aneurysm repair (AAA) on quality of life. The average age of patients in the two studies was 69 and 73 respectively.

It became evident very quickly that some patients failed to understand the questionnaire and completed it incorrectly. Patients were supervised when they completed the form. If I believed that a patient had misinterpreted a question then I would stop them and re-read the question to them and also read out each response option (for example, all of the time, most of the time, etc). If the patient then said that they understood the question they would be left to complete the rest of it. I noted down any occasion when a patient changed their mind regarding their response after I had re-read the question.

Problems were most commonly encountered with the following questions.

Question 3 The following questions are about activities that you might do during a typical day. Does your health limit you in these activities? If so, how much?

Yes, limited a lot   Yes, limited a little
No, not limited at all

Some patients misunderstood the concept of health limiting their physical activity. This may be because they misread the question or simply didn't understand it. Commonly, relatively elderly patients would read "3a Vigorous activities such as running ..." and would tick "No, not limited at all". When I re-read the question to the patient they would typically respond "No, no I can't do that sort of thing". In this example 23% of patients went on to change their mind when the question was re-read. If left unchecked I suspect that many of these patients would have gone on to complete all 10 parts of question 3 incorrectly.

Question 5 During the past four weeks, have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)?

After answering 14 consecutive questions regarding physical activity, many patients appeared to find it difficult to switch to thinking about emotional problems in question 5. Commonly patients reported that they thought questions 4 and 5 were the same and so they responded with the same answers. When the question was re-read, 19% of patients felt that their first response to the question was incorrect. Clearly it is difficult to judge given the nature of the question whether people were completing it correctly. Therefore the scores from this question were compared with patients' scores from the HAD scale. The scores from question 5 form the basis of the role functioning-emotional scale (RE). Patients' scores from this scale had a low correlation with the measures of anxiety and depression from the HAD scale (anxiety $r = -0.27$, depression $r = -0.26$). A strong association between scores on the RE scale and the mental health scale (MH) ($r = 0.60$) has been reported.[5] In the total sample reported here (n=208) this correlation was much lower ($r=0.35$).

Question 9 The following questions are about how you feel and about how things have been with you during the past month

Some patients also had difficulties interpreting question 9 and in particular they appeared to find the Likert scale hard to use. (I note that Likert scales are used more frequently in the revised version). Some patients appeared to re-code the scale as "bad" to "good" rather than "All of the time" to "None of the time". For example a patient may respond "None of the time" to questions 9b and c regarding feelings of anxiety and depression. Further down the questionnaire they would respond "None of the time" again to 9h "Have you been a happy person?" while saying "Yes I am a happy person". When the question was re-read, many patients changed their original response and felt that they had made a mistake.

The observations reported here are little more than anecdotal. My method of recording errors was arbitrary. Indeed my method of re-reading questions to patients when I considered that they had made a mistake could be criticised for biasing patients' responses. However, when I started using the SF-36 I quickly became convinced that many patients were failing to understand the questions. I believe that version 2 has not resolved the shortcomings that were inherent in the original version. As it stands I believe that the SF-36 should not be used as an assessment of quality of life in older patients. Investigators should also be cautious about using the tool with any patients who have evidence of head injuries, cognitive impairments or communication problems.

ANDREW LLOYD
Oxford Centre for Health Care Research and Development, Oxford Brookes University

1 Jenkinson C, Stewart-Brown S, Petersen S, *et al*. Assessment of the SF-36 version 2 in the United Kingdom. *J Epidemiol Community Health* 1999;**53**:46–50.

2 Barczak P, Kane N, Andrews S, *et al.* Patterns of psychiatric morbidity in a genito-urinary clinic: A validation of the Hospital Anxiety Depression scale (HAD). *Br J Psychiatry* 1988;**152**: 698–700.

3 Lewis G, Wessely S. Comparison of the General Health Questionnaire and the Hospital Anxiety and Depression Scale. *Br J Psychiatry* 1990; **157**:860–4.

4 Hamer D, Sanjeev D, Butterworth E, *et al*. Using the Hospital Anxiety and Depression Scale to screen for psychiatric disorders in people presenting with deliberate self-harm. *Br J Psychiatry* 1991;**158**:782–4.

5 Ware JE, Kosinski M, Keller SD. *SF-36 Physical and Mental Health Summary Scales: A user's manual, 1994.* Boston: New England Medical Centre, 1994.

Reply

We have some sympathy with the views expressed by Dr Lloyd, and have indeed written regularly on the issues of questionnaire selection.[1,2] Measures should be selected for inclusion in studies because evidence is available to support their use, or because the study is designed to assess a given measure with a specific population. We would agree that caution is required when using any version of the SF-36 in older age groups because of the fact that an, albeit fairly limited, body of evidence has emerged that suggests the measure may be inappropriate. However, while caution should be the watch word when it comes to selection and interpretation of measures, we would also advise readers not to readily dismiss the SF-36 for older adults. Dr Lloyd's letter raises the issue in a rather speculative and, as he himself remarks, a rather anecdotal fashion. Two critcisms could be raised against his concerns, one related to the manner in which his research was conducted and another concerning the properties of scales, rather than items.

Firstly, Dr Lloyd remarks that he has used the SF-36 in studies in which he has re-read parts of the form to elderly patients because he believed they had misunderstood or misinterpreted questions. As he himself points out such treatment of patients could influence their responses, (for example they may change their answers because they thought they had got the question wrong in some way). He remarks that while some people may criticise him for biasing patient responses he only did so because he quickly became aware that many questions were misunderstood by patients. Such a view should be supported by evidence. Many people have offered such criticisms of the SF-36 but have produced scarce scientific proof to support their claims. Claims that the measure is inappropriate for the elderly are more often than not based upon little more than anecdotes, rather than rigorously conducted qualitative studies.

Secondly, Dr Lloyd suggests that there will be errors in the answers provided by older respondents to the questions on the SF36. This is not particularly surprising and is to be expected with all age groups. All questionnaire items consist of true measurement plus an error term. The trick is to reduce the error term as much as is possible. This is why health status measurement has for the most part adopted multi-item scales. If we take more than one item to measure the same underlying attribute then the summed score of all the items will be more reliable than a single question. This is because all true measurement from each item will be summed, while error terms on all the items will be random and, effectively, non-additive (the logic here is that for every person who scores a little high on a given item there will be someone who scores a little low, and so on). This, of course, assumes that items have been selected carefully and are neither unrelated or too closely related; an assumption that is implicitly built into the SF-36.

Recent data report on the successful use of the SF-36 in older patients in a large scale survey. Normative data are available.[3] This evidence would seem to suggest the SF-36 is useful in this patient group, but specific research must investigate this issue. In a world that now embraces evidence based medicine it might be wise to adopt a similarly rigorous approach to questionnaire selection and application.

CRISPIN JENKINSON
SARAH STEWART-BROWN
Health Services Research Unit, University of Oxford, Institute of Health Sciences, Oxford OX3 7LF

1 Jenkinson C, Peto V, Coulter A. Making sense of ambiguity: evaluation of internal reliability and face validity of the SF-36 in patients presenting with menorrhagia. *Quality in Health Care*1996;**5**:9–12.

2 Jenkinson C, McGee H. *Health status measurement: a brief but critical introduction.* Oxford: Radcliffe Medical Press, 1998.

3 Bowling A, Bond M, Jenkinson C, *et al.* Short Form 36 (SF-36) Health Survey questionnaire: which normative data should be used? Comparisons between the norms provided by the Omnibus Survey in Britain, the Health Survey for England and the Oxford Health and Lifestyle Survey. *J Public Health Med* (in press).

Mortality in poorer areas

Editor,—Law and Morris state that "about 85% of the overall excess mortality with deprivation was attributable to heavier smoking" in their study of deaths in England and Wales in 1992.[1] They correctly state that strengths of their study include allowing for the generally higher tar yield and number of cigarettes smoked by lower socioeconomic smokers, and the generally younger age of starting smoking for lower socioeconomic people. They also offer the plausible argument that cohort studies may be biased against finding a substantial role for smoking as an intermediary between lower socioeconomic status and mortality, because people recorded as non-smokers in a cohort study may actually have recently stopped smoking because of the early symptoms of smoking related disease. An ecological study will avoid this latter bias in part.

However, there are problems with the ecological study of Law and Morris that suggest the figure of 85% is likely to be a substantial overestimate. Firstly, the median local authority district size of 102 000 is large for a study that is attempting to "ecologically infer" the relation of deprivation and smoking with mortality. Greenland and colleagues have shown that the larger the size of the study unit in ecological studies, the more likely that cross level bias (the "ecological fallacy") will cause error in the inferred relations at the individual level.[2–5] The direction and magnitude of the cross level bias is impossible to predict from the ecological data alone, but is often biased away from the null. Secondly, and a component of the previous reason, both the predicted and observed relative risks used by Morris and Law will be confounded by other lifestyle factors, resulting in an overestimate of the contribution of smoking. Thirdly, the external source of the relative risk data for smoking,[6] while a highly reputable study, was based on a cohort of male doctors and may not be generalisable to the total population of England and Wales. This lack of generalisability would arise if, as would be expected, non-smoking doctors had a lower mortality rate than non-smoking members of the population generally because of a favourable profile of other risk factors. This in turn could result in higher relative risks of smoking being observed among doctors than non-doctors.

Yes, smoking is undoubtedly an intermediate variable between deprivation and mortality. But I doubt that if, in the counterfactual, none of the population alive in England and Wales at 1992 had ever smoked (or even just that there was no variation by deprivation in smoking) that as much as 85% of the inequality in mortality by deprivation would have been removed.

TONY BLAKELY
Department of Public Health, Wellington School of Medicine, PO Box 7343, Wellington, New Zealand

1 Law M, Morris J. Why is mortality higher in poorer areas and in more northern areas of England and Wales? *J Epidemiol Community Health* 1998;**52**:344–52.

2 Greenland S, Robins J. Invited commentary: ecologic studies - biases, misconceptions, and counterexamples. *Am J Epidemiol* 1994;**139**: 747–60.

3 Greenland S. Divergent biases in ecologic and individual-level studies. *Stat Med* 1992;**11**: 1209–23.

4 Greenland S, Morgernstern H. Ecological bias, confounding, and effect modification. *Int J Epidemiol* 1989;**18**:269–74.

5 Morgenstern H. Ecologic studies in epidemiology: concepts, principles, and methods. *Annu Rev Public Health* 1995;**16**:61–81.

6 Doll R, Peto R, Hall E, *et al.* Mortality in relation to smoking: 40 years' observations on male British doctors. *BMJ* 1994;**309**:900–11.

Reply

We concluded in our paper that all cause mortality was 15% higher in the most deprived compared with the least deprived districts, and that heavier smoking accounted for most (about 85%) of this excess mortality.[1] We disagree with Blakely that the figure of 85% is likely to be a substantial overestimate. Statistical calculations are not necessary to see that smoking accounts for most of the excess mortality in the more deprived districts. One need only consider the specific causes of death that are more common in deprived districts (table 2 in our paper[1]); almost all of them are smoking related. Three diseases that are strongly smoking related (lung cancer, chronic bronchitis and emphysema, and ischaemic heart disease) accounted for two thirds of the excess mortality, and other smoking related cancers and circulatory diseases accounted for a further sixth of the excess. Diseases reflecting other behavioural differences (cirrhosis of the liver, AIDS), or differences in medical care, accounted for little of the total excess mortality, while two important aetiological factors in circulatory diseases, serum cholesterol and blood pressure, show little difference between deprived and affluent districts (see references 37–39 in our paper[1]).

Blakely has three concerns about our smoking analysis. We do not think the "ecological fallacy" of Greenland and colleagues (which may produce a bias in either direction) is a material problem in this context, particularly as we are not inferring relations at the individual level. Exaggeration of relations between smoking and diseases through confounding is unlikely. Asbestos and other occupational exposures that cause lung cancer may be more common in smokers, but these exposures cause relatively few lung cancer cases in relatively few districts. Associations between smoking and other heart disease risk factors tend to be weak, and as stated above, blood pressure and serum cholesterol show little variation between affluent and deprived districts. Blakely suggests that relative risk estimates from the British Doctors Study are not generalisable. The results of the British Doctors Study in relation to smoking have in general been supported quantitatively by other large cohort studies, and we confirmed this for ischaemic heart disease.[2] Moreover one would expect estimates of relative risk to be generalisable: the proportionate increase in risk in smokers should be the same in populations where smoking is relatively common or uncommon or where, for reasons other than smoking, the disease is relatively common or uncommon.

M R LAW
J K MORRIS
Department of Environmental and Preventive Medicine, Wolfson Institute of Preventive Medicine, Medical College of St Bartholomew's Hospital, Charterhouse Square, London EC1M 6BQ

1 Law MR, Morris JK. Why is mortality higher in poorer areas and in more northern areas of England and Wales? *J Epidemiol Community Health* 1998;**52**:344–52.

2 Law MR, Morris JK, Wald NJ. Environmental tobacco smoke exposure and ischaemic heart disease: an evaluation of the evidence. *BMJ* 1997;**315**:973–88.

Bayesian analysis

Editor,—We are delighted to see your journal publish an excellent paper showing by example how a statistical analysis that has run into difficulties can be converted into a Bayesian analysis and thus rescued.[1]

Burton *et al*[1] state that a 95% confidence interval can be interpreted as a 95% Bayesian credible interval (also known as a posterior probability interval), thus allowing the interpretation that the true hypothesis is 95% certain to lie within the interval, provided that the design admits "a uniform prior distribution for the main outcome measure". Lindley[2] is cited as the theoretical justification for this assertion.