Is the SF-36 a valid measure of change in population health? Results from the Whitehall II study

Harry Hemingway, Mai Stafford, Stephen Stansfeld, Martin Shipley, Michael Marmot

Abstract

Objective: To measure within-person change in scores on the short form general health survey (SF-36) by age, sex, employment grade, and disease status.

Design: Longitudinal study with a mean of 36 months (range 23-59 months) follow up, with screening examination and questionnaire to detect physical and psychiatric morbidity.

Setting: 20 civil service departments originally located in London.

Participants: 5070 male and 2197 female office based civil servants aged 39-63 years.

Main outcome measures: Change in the eight scales of the SF-36 (adjusted for baseline score and length of follow up) and effect sizes (adjusted change/standard deviation of differences).

Results: Within-person declines (worsening health) with age were greater than estimated by cross sectional data alone. General mental health showed greater declines among younger participants (P for linear trend <0.001). Employment grade was inversely related to change; lower grades had greater deteriorations than higher grades ($P<0.001$ for each scale in men; $P<0.05$ for each scale in women except general health perceptions and role limitations due to physical problems). The greatest declines were seen among participants with disease at baseline, with the effects of physical and psychiatric morbidity being additive. Effect sizes ranged from 0.20 to 0.65 in participants with both physical and psychiatric morbidity.

Conclusions: Health functioning, as measured by the SF-36, changed in hypothesised directions with age, employment grade, and disease status. These changes occurred within a short follow up period, in an occupational, high functioning cohort which has not been the subject of intervention, suggesting that the SF-36 is sensitive to changes in health in general populations.

Introduction

Measuring changes in population health is important to evaluate interventions and to predict the need for health and social care. The traditional measures of mortality and morbidity, although useful, have limitations: showing changes in mortality requires prolonged periods of observation or large numbers of events, or both, and changes in morbidity are more expensive to measure and do not take account of the functional impact on a patient's life. A given level of objectively assessed morbidity may have widely differing impacts on individuals' physical, psychological, and social functioning.[1] Since levels of functioning are important in predicting demand for services, changes in such health related quality of life outcomes might complement mortality and morbidity measures. Although changes in quality of life are increasingly used as outcome measures in clinical trials,[2 3] they have rarely been studied in populations other than patients.

The short form 36 health survey (SF-36)[4-6] is a 36 item questionnaire which measures health functioning on eight scales and is among the most widely used measure of quality of life in studies of patients[7 8] and the general population.[9-19] Cross sectional data from population studies have shown that the SF-36 is reliable and able to detect differences between groups defined by age, sex, socioeconomic status, geographical region, and clinical conditions. The SF-36 may therefore be a useful tool for monitoring changes in health in the population.

There are, however, no reports of using repeated measures of the SF-36 in population studies, so it is not known whether it is sensitive in detecting changes within individuals over time. We report here individual changes in SF-36 scores in the Whitehall II study of British civil servants. On the basis of our previous cross sectional data[19] we hypothesised that over a three year follow up period a decline in scale scores (worsening health) would be associated with age (directly for physical functioning, inversely for general mental health); socioeconomic status (inversely); and the presence of chronic, progressive, or recurrent disease at baseline.

Methods

Participants

All non-industrial civil servants aged 35 to 55 years working in the London offices of 20 departments were invited to participate in this study. The final cohort consisted of 10 308, with an overall response rate of 73%, although the true response rate was probably higher as 4% of those on the list of employees had moved before the start of the study and were therefore not eligible for inclusion. Employment grade within the civil service was used as a measure of socioeconomic status. On the basis of salary the civil service identifies 12 non-industrial grades. To obtain sufficient numbers for meaningful analysis we combined the top six groups into grade 1 and the bottom two groups into grade 6, thus producing six grade categories. The salaries ranged from £6483-11 917 (grade 6) to £28 904-£87 620 (grade 1) in 1992.

Measures

Baseline SF-36 scores (UK standard version) were measured at the third phase of the study, between August 1991 and May 1993 on 8349 participants (5763 men and 2586 women). At phase 4, between April 1995 and June 1996, an identical version of the SF-36 was completed by 7949 participants (5467 men and 2482 women). For the purposes of this paper, phase 3 measurements are referred to as baseline and phase 4 measurements as follow up. Sixty seven

International Centre for Health and Society, Department of Epidemiology and Public Health, University College London Medical School, London WC1E 6BT
Mai Stafford, *statistician*
Stephen Stansfeld, *senior lecturer in community psychiatry*
Martin Shipley, *senior lecturer in medical statistics*
Michael Marmot, *professor of epidemiology and public health*

Department of Public Health, Kensington & Chelsea and Westminster Health Authority, London W2 6LX
Harry Hemingway, *senior lecturer in epidemiology*

Correspondence to Dr H Hemingway, International Centre for Health and Society, Department of Epidemiology and Public Health, University College London Medical School, London WC1E 6BT
h.hemingway@public-health.ucl.ac.uk

BMJ 1997;315:1273–9

participants died between the end of phase 3 and the beginning of phase 4.

The SF-36 consists of 36 items scored in eight scales: general health perceptions (5 items), physical functioning (10), role limitations due to physical functioning (4), bodily pain (2), general mental health (5), role limitations due to emotional problems (3), vitality (4), and social functioning (2). The remaining item, relating to change in health, is not scored as a separate dimension. As an example of scale content, the physical functioning scale covers limitations during a typical day ("a lot," "a little," "none") in vigorous activities (strenuous sports, running, etc), moderate activities (housework, playing golf, etc), lifting and carrying, climbing stairs, bending, kneeling, and walking. Scores for all scales were calculated using the medical outcomes study (MOS) scoring system[20] and ranged from 0 (lowest wellbeing) to 100 (highest wellbeing). These scales had high internal consistency at baseline (Cronbach's α 0.76-0.86). The mean percentage of items missing across all scales was related to sex (0.38% in men and 0.50% in women; $P = 0.02$), age (0.65% in those 55 years and older and 0.14% in those 44 years and younger; $P < 0.001$), and grade (0.60% in the lowest and 0.25% in the highest grade; $P < 0.001$).

Participants were categorised into four mutually exclusive groups according to their disease status at baseline: healthy (free of the following conditions), physical disease only, minor psychiatric disorder only, and both physical disease and minor psychiatric disorder. Physical diseases (chosen on a priori grounds as likely to affect physical functioning) were defined as one or more of the following: angina (n = 450),[21] self report of doctor diagnosed heart attack or angina (n = 150), probable or possible ischaemia on resting electrocardiogram (Minnesota codes 1-1 to 1-3, 4-1 to 4-4, 5-1 to 5-3, and 7-1-1) (n = 707), hypertension (> 160/90 or taking antihypertensive drugs) (n = 1554), claudication (n = 125),[21] diabetes (self report or oral glucose tolerance test) (n = 222),[22] chronic bronchitis (n = 914),[23] musculoskeletal disorders (self report) (n = 1257), and cancer (OPCS registration or self report) (n = 128). Minor psychiatric disorder, principally anxiety and depression, was defined as a score of ≥5 on the 30 item general health questionnaire (n = 1489).[24]

Statistical analysis

Changes in SF-36 were examined by age, employment grade, and disease status separately for men and women. A negative change reflects a decline in scores and, if valid, a deterioration in health. As expected, participants who had high scores at baseline had lower scores at follow up and vice versa, a common phenomenon known as regression to the mean. Analysing such data using simple differences is problematic as the magnitude of the change would depend on the level at baseline.[25 26] Furthermore, the changes in scores, unlike single measurements, are normally distributed. Therefore, we used regression models separately for men and women for each scale of the form:

follow up score − baseline score =
baseline score + covariate 1 + covariate 2 … etc

These models give a change score adjusted for the potential bias of regression to the mean. Longitudinal estimates of change per year were obtained using these models from the coefficients for length of follow up in which the intercept term was constrained to be zero. Adjustment was made for the potential confounding of

Table 1 Absolute change in SF-36 scores between baseline and follow up (mean 36 months)

	Men			Women		
	Mean (SD) at baseline	**Mean at follow up**	**Mean difference (95% CI)**	**Mean (SD) at baseline**	**Mean at follow up**	**Mean difference (95% CI)**
Minium No of observations	5737	5384	5070	2570	2384	2197
General health perceptions	72.5 (17.6)	70.7	−2.0 (−2.4 to −1.6)	71.8 (19.1)	70.0	−1.9 (−2.5 to −1.3)
Physical functioning	91.9 (11.9)	89.7	−2.1 (−2.5 to −1.8)	83.7 (19.2)	80.3	−3.0 (−3.7 to −2.3)
Physical role limitation	91.9 (21.8)	86.0	−5.9 (−6.8 to −5.1)	84.4 (30.4)	77.1	−7.3 (−8.8 to −5.8)
Bodily pain	87.6 (16.5)	83.8	−3.7 (−4.3 to −3.2)	78.8 (21.9)	75.8	−2.5 (−3.4 to −1.6)
General mental health	77.0 (14.7)	75.6	−1.5 (−1.9 to −1.1)	73.6 (16.0)	72.0	−1.5 (−2.2 to −0.9)
Emotional role limitation	89.7 (24.7)	86.1	−3.8 (−4.6 to −2.9)	86.0 (29.1)	80.9	−4.5 (−6.0 to −3.1)
Vitality	63.8 (17.5)	61.5	−2.6 (−3.0 to −2.1)	57.8 (20.1)	55.9	−1.7 (−2.4 to −1.0)
Social functioning	91.3 (17.0)	87.3	−4.1 (−4.7 to −3.5)	82.0 (21.3)	81.4	−4.5 (−5.5 to −3.6)

P <0.05 (paired *t* test) for each difference.

Table 2 Mean change in SF-36 scores per year of follow up

	Men		Women	
	Cross sectional estimate using baseline data (95% CI)	**Longitudinal estimate* (95% CI)**	**Cross sectional estimate using baseline data (95% CI)**	**Longitudinal estimate* (95% CI)**
Minimum No of observations	5737	5070	2570	2197
General health perceptions	0.03 (−0.05 to 0.10)	−0.38 (−0.97 to −0.21)	−0.02 (−0.14 to −0.10)	−0.69 (−1.67 to 0.29)
Physical functioning	−0.34 (−0.39 to −0.29)	−0.51 (−1.02 to 0.00)	−0.72 (−0.84 to −0.60)	−0.69 (−1.73 to 0.35)
Physical role limitation	−0.09 (−0.18 to −0.00)	−1.93 (−3.16 to 0.70)	−0.16 (−0.37 to 0.03)	−0.39 (−2.60 to 1.82)
Bodily pain	−0.09 (−0.16 to −0.02)	−0.88 (−3.16 to 0.70)	−0.11 (−0.24 to 0.04)	−1.00 (−2.29 to 0.29)
General mental health	0.39 (0.32 to 0.45)	0.29 (−0.28 to 0.86)	0.35 (0.25 to 0.45)	−0.89 (−1.85 to 0.07)
Emotional role limitation	0.28 (0.18 to 0.39)	0.20 (−1.00 to 1.40)	0.40 (0.22 to 0.58)	0.62 (−1.44 to 2.68)
Vitality	0.38 (0.30 to 0.45)	−0.17 (−0.82 to 0.48)	0.37 (0.24 to 0.50)	−0.42 (−1.54 to 0.70)
Social functioning	0.11 (0.04 to 0.18)	−0.37 (−1.19 to 0.45)	0.20 (0.07 to 0.33)	−0.82 (−2.23 to 0.59)

* Adjusted for baseline score.

Table 3 Mean change (adjusted for baseline score) in SF-36 scores per year of follow up within age groups

	Men (age in years)					Women (age in years)				
	39-44	45-49	50-54	⩾55	Test for trend	39-44	45-49	50-54	⩾55	Test for trend
Minimum No of observations	1368	1394	1006	1302		503	556	446	686	
General health perceptions	−2.8	−1.6	−1.6	−1.8	P=0.06	−1.3	−2.4	−3.2	−1.2	P=0.7
Physical functioning	−1.0	−1.6	−3.1	−3.1	P<0.001	−0.8	−2.5	−4.4	−4.1	P<0.001
Physical role limitation	−5.1	−3.8	−7.7	−7.8	P<0.01	−4.5	−8.1	−8.8	−7.6	P=0.2
Bodily pain	−3.4	−3.0	−4.1	−4.5	P<0.05	−0.8	−3.1	−3.6	−2.6	P=0.2
General mental health	−4.1	−2.8	−0.6	1.9	P<0.001	−2.4	−3.3	−1.6	0.6	P<0.001
Emotional role limitation	−6.1	−4.2	−3.8	−0.8	P<0.001	−3.6	−6.8	−5.8	−2.5	P=0.3
Vitality	−5.3	−3.6	−2.0	1.1	P<0.001	−3.2	−3.1	−2.5	1.1	P<0.001
Social functioning	−6.3	−3.5	−4.4	−2.3	P<0.001	−3.7	−6.8	−4.6	−3.2	P=0.2

Table 4 Change (adjusted for baseline score, length of follow up, and age) in SF-36 scores by employment grade

	1 (Highest)	2	3	4	5	6 (Lowest)	Test for trend	Effect size (95% CI) (grade 6 *v* grade 1)	
								Change	Cross sectional
Men									
Minimum No of observations	1169	1349	856	885	813	290			
General health perceptions	−1.0	−1.9	−2.0	−1.8	−2.8	−5.0	P<0.001	0.28 (0.15 to 0.41)	0.28 (0.17 to 0.39)
Physical functioning	−0.6	−2.0	−1.8	−2.2	−3.3	−6.8	p<0.001	0.49 (0.36 to 0.62)	0.57 (0.46 to 0.68)
Physical role limitation	−3.7	−6.1	−4.3	−7.5	−7.9	−10.4	P<0.001	0.21 (0.08 to 0.34)	0.18 (0.07 to 0.29)
Bodily pain	−1.8	−3.2	−3.5	−4.1	−6.2	−8.3	P<0.001	0.33 (0.20 to 0.46)	0.35 (0.24 to 0.46)
General mental health	−0.8	−0.7	−0.9	−2.1	−3.1	−4.6	P<0.001	0.27 (0.14 to 0.40)	0.19 (0.08 to 0.30)
Emotional role limitation	−1.7	−2.8	−2.2	−6.2	−6.6	−8.5	P<0.001	0.22 (0.09 to 0.35)	0.08 (−0.03 to 0.19)
Vitality	−1.5	−2.2	−1.5	−3.8	−4.3	−4.4	P<0.001	0.19 (0.06 to 0.32)	0.00 (−0.11 to 0.11)
Social functioning	−1.9	−3.0	−3.2	−5.4	−8.1	−10.0	P<0.001	0.39 (0.26 to 0.52)	0.37 (0.26 to 0.48)
Women									
Minimum No of observations	142	223	184	340	506	799			
General health perceptions	−0.7	−1.8	−2.1	−1.2	−1.6	−2.6	P=0.2	0.12 (−0.06 to 0.30)	0.09 (−0.08 to 0.26)
Physical functioning	−0.1	0.4	−0.1	−3.4	−2.8	−5.0	P<0.001	0.29 (0.11 to 0.47)	0.45 (0.28 to 0.62)
Physical role limitation	−4.1	−5.1	−9.5	−6.8	−5.2	−9.6	P=0.07	0.15 (−0.03 to 0.33)	0.09 (−0.08 to 0.26)
Bodily pain	3.7	0.1	−1.1	−2.9	−1.6	−5.1	P<0.001	0.40 (0.23 to 0.57)	0.26 (0.09 to 0.43)
General mental health	1.7	−0.5	−0.7	−2.3	−1.6	−2.2	P<0.01	0.24 (0.06 to 0.42)	0.01 (−0.16 to 0.18)
Emotional role limitation	−1.6	1.8	−3.7	−5.7	−6.5	−5.9	P<0.001	0.27 (0.09 to 0.45)	0.04 (−0.13 to 0.21)
Vitality	2.7	−1.3	−1.3	−2.3	−1.8	−2.4	P<0.01	0.28 (0.10 to 0.46)	0.09 (−0.08 to 0.26)
Social functioning	−1.7	−3.1	−4.1	−4.5	−3.9	−6.0	P<0.05	0.18 (0.01 to 0.35)	0.08 (−0.09 to 0.25)

differing lengths of follow up in all models, even though age, grade, and disease status had only small effects on length of follow up. Two tailed tests were used throughout. Effect sizes for cross sectional data comparing two groups were calculated by dividing the difference between the means of the two groups by the sex specific standard deviation for that scale. Effect sizes for change were calculated by dividing the adjusted change by the sex specific standard deviation of the differences for that scale. A higher effect size indicates greater sensitivity to change. All analyses were performed using the SAS statistical package (SAS Institute, Cary, NC).

Results

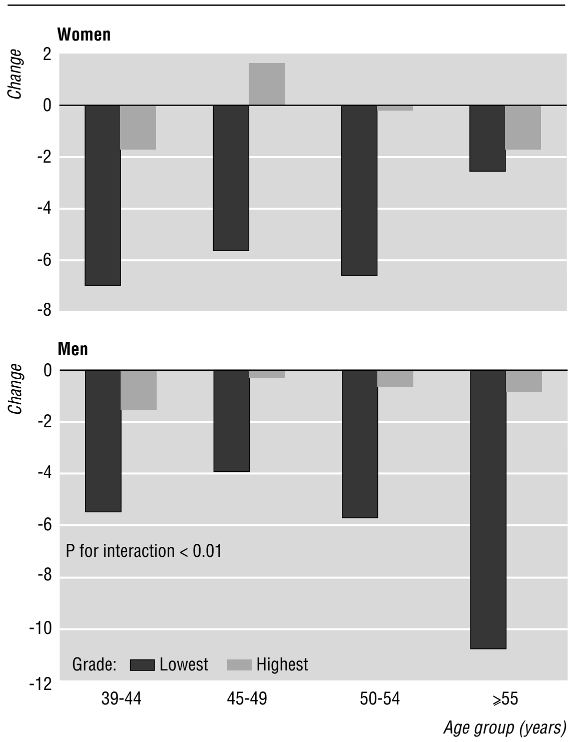
The median follow up was 36 months (range 23-59 months). The median age at follow up was 52 (42-65) years for men and 53 (42-65) years for women. Participants who did not respond at follow up had lower mean scores at baseline on all scales except the vitality scale for women (data not shown).

Unadjusted mean scores on all scales were lower at follow up (table 1). Table 2 shows a comparison of cross sectional and longitudinal estimates of change in scale score per year. The cross sectional data tended to underestimate the within-person decline with increasing age in general health perceptions, physical role limitation, and bodily pain. While the cross sectional data suggested improvement with age in general mental health, emotional role limitation, vitality, and social functioning, the longitudinal estimates showed that only emotional role limitations (men and women) and general mental health (men only) improved with age.

Table 3 shows the adjusted change within five year age groups at baseline. Younger men had greater declines in general mental health, emotional role limitation, vitality, and social functioning than older men (P for linear trend <0.001). Among women, a similar relation was seen in the vitality and general mental health scales ($P<0.01$). Indeed, adjusted change was positive on general mental health and vitality in the oldest age group. Older participants showed greater declines in physical functioning than younger participants ($P<0.001$).

Men in the lower employment grades had greater declines on all scales than men in higher grades (P for linear trend <0.001), and women in the lower grades had significantly greater declines on all scales except physical role limitation and general health perceptions (table 4). Women in the high grades showed positive change for bodily pain, vitality, emotional role limitation, and general mental health.

The figure shows the effect of civil service employment grade on age related declines in physical functioning. For men there was evidence of an



Employment grade and age differences in change (adjusted for baseline score and length of follow up) in physical functioning measured by SF-36 questionnaire

interaction (P < 0.01), with the men in the highest grade less likely than men in the lowest grade to show declines with age.

Participants with both physical and psychiatric morbidity had worse health at baseline than those with either physical or psychiatric morbidity alone, with effect sizes ranging from 0.99 to 1.31 (table 5). Table 6 shows the effect of baseline disease category on change in SF-36 scores adjusting for baseline score, length of follow up, and age. For both cross sectional and longitudinal data, there was some evidence that physical diseases predominantly affected physical scales (physical functioning, physical role limitation, and pain) and psychiatric morbidity predominantly affected psychological and social functioning. Participants who had either physical disease or were defined as cases on the general health questionnaire had greater declines than those with no medical conditions (P for heterogeneity < 0.01). The presence of physical disease or minor psychiatric morbidity was associated with a similar magnitude of effect for all scales except emotional role limitations, for which the decline was larger among the cases. The effects of physical disease and minor psychiatric morbidity were approximately additive; participants who had both of these experienced greater decline on all scales than participants who had either of these. Effect sizes in this group ranged from 0.20 to 0.65.

Discussion

This is the first report of the ability of the SF-36, a simple and inexpensive measure of health outcomes, to detect change in health in a general population. In the high functioning, occupational cohort of the Whitehall II study, each of the eight dimensions of health measured by the SF-36 showed a mean decline over three years of follow up. As hypothesised, employment grade was inversely related to decline, with lower grades experiencing greater deteriorations than higher grades. The greatest declines were seen among subjects with disease (considered to be chronic, recurrent, or progressive) at baseline, with the effects of physical and psychiatric morbidity being additive.

Changes in health

Measuring health on continuous scales, rather than dichotomising individuals as diseased or disabled or not, allows elucidation of trajectories of change. Decline in health and functioning has been assumed to be the biologically inevitable consequence of aging, although this concept is increasingly questioned[27] by the heterogenous patterns of change that are found in populations aged over 65.[28–30] Cross sectional results in the Whitehall II study (participants aged 39-63) tended to underestimate the longitudinal decline with age in general perceptions of health, physical role limitation, and bodily pain. Furthermore, while these cross sectional data[19] suggested improvement with age in general mental health, emotional role limitations, vitality, and social functioning, the longitudinal data supported this improvement only for emotional role

Table 5 Cross sectional mean SF-36 scores (adjusted for age) by disease category at baseline

	Men					Women				
	Healthy*	Physical disease only†	Minor psychiatric morbidity only‡	Physical disease and minor psychiatric morbidity	Effect size (95% CI) for physical disease and minor psychiatric morbidity	Healthy*	Physical disease only†	Minor psychiatric morbidity only ‡	Physical disease and minor psychiatric morbidity	Effect size (95% CI) for physical disease and minor psychiatric morbidity
Minimum No of observations	3120	597	1132	221		1227	261	569	140	
General health perceptions	76.7	71.4	65.1	56.2	1.16 (1.02 to 1.30)	77.4	69.3	65.8	53.5	1.25 (1.08 to 1.42)
Physical functioning	94.1	87.5	90.4	81.4	1.07 (0.93 to 1.21)	88.2	75.5	82.2	66.3	1.14 (0.97 to 1.31)
Physical role limitation	95.4	90.7	87.0	74.3	0.97 (0.83 to 1.11)	92.1	78.2	76.5	61.8	1.00 (0.83 to 1.17)
Bodily pain	90.9	81.4	85.6	70.1	1.26 (1.12 to 1.40)	84.8	68.3	75.7	59.2	1.17 (1.00 to 1.34)
General mental health	82.1	82.1	63.6	62.8	1.31 (1.17 to 1.45)	80.3	79.0	60.3	60.0	1.27 (1.10 to 1.44)
Emotional role limitation	96.4	95.8	72.4	72.0	0.99 (0.85 to 1.13)	95.4	93.1	67.8	64.7	1.05 (0.88 to 1.22)
Vitality	68.7	66.9	51.8	49.2	1.11 (0.97 to 1.25)	65.1	59.1	45.4	42.1	1.14 (0.97 to 1.31)
Social functioning	95.9	93.2	80.9	77.4	1.09 (0.95 to 1.23)	93.0	87.8	75.4	69.0	1.13 (0.96 to 1.30)

P<0.01 for difference between disease categories for all scales (ANOVA).
*No physical disease or psychiatric morbidity.
†Angina, self report of doctor diagnosed ischaemia, probable or possible ischaemia on resting electrocardiogram, hypertension, claudication, diabetes, chronic bronchitis, musculoskeletal disorders, cancer.
‡≥5 on the 30 item general health questionnaire.

Table 6 Change (adjusted for baseline score, length of follow up, and age) in SF-36 scores by presence of disease at baseline

	Men					Women				
	Healthy*	Physical disease only†	Minor psychiatric morbidity only‡	Physical disease and minor psychiatric morbidity	Effect size (95% CI) for physical disease and minor psychiatric morbidity	Healthy*	Physical disease only†	Minor psychiatric morbidity only‡	Physical disease and minor psychiatric morbidity	Effect size (95% CI) for physical disease and minor psychiatric morbidity
Minimum No of observations	3120	597	1132	221		1227	261	569	140	
General health perceptions	−1.0	−4.3	−2.8	−5.4	0.38 (0.24 to 0.52)	−1.0	−4.0	−2.3	−4.4	0.29 (0.12 to 0.46)
Physical functioning	−1.4	−3.4	−2.8	−5.7	0.49 (0.35 to 0.63)	−1.5	−7.1	−3.1	−8.1	0.49 (0.32 to 0.66)
Physical role limitation	−3.0	−10.8	−8.9	−18.5	0.59 (0.45 to 0.73)	−2.4	−14.3	−11.3	−21.0	0.58 (0.41 to 0.75)
Bodily pain	−1.9	−7.2	−5.4	−11.5	0.59 (0.45 to 0.73)	−0.5	−5.9	−3.8	−8.6	0.39 (0.22 to 0.56)
General mental health	−0.3	−1.7	−3.7	−6.5	0.46 (0.32 to 0.60)	−0.3	−1.8	−3.5	−3.9	0.24 (0.07 to 0.41)
Emotional role limitation	−1.1	−1.7	−10.0	−15.5	0.50 (0.36 to 0.64)	−0.7	−5.1	−10.8	−11.7	0.33 (0.16 to 0.50)
Vitality	−1.7	−4.3	−3.1	−6.6	0.42 (0.28 to 0.56)	−0.4	−4.2	−2.9	−3.7	0.20 (0.03 to 0.37)
Social functioning	−2.0	−5.1	−7.5	−13.7	0.65 (0.51 to 0.79)	−2.1	−6.3	−7.2	−12.2	0.51 (0.34 to 0.68)

P<0.01 for difference between disease categories for all scales (ANOVA).

*No physical disease or psychiatric morbidity.

†Angina, self report of doctor diagnosed ischaemia, probable or possible ischaemia on resting electrocardiogram, hypertension, claudication, diabetes, chronic bronchitis, musculoskeletal disorders, cancer.

‡≥5 on the 30 item general health questionnaire.

limitations (men and women) and general mental health (men). However, general mental health and vitality showed greater declines among younger participants. This is evidence against a simple age effect and suggests the existence of cohort or period effects, or both.[31] Although further measures of the SF-36 are required to distinguish between these, it is plausible that increasing privatisation in the civil service may constitute a relevant period effect.[32] Sex differences also showed a different pattern in cross sectional and longitudinal results. In cross sectional results,[10] women consistently had lower SF-36 scores than men whereas in longitudinal results there were no such sex differences. Existing studies measuring change in health functioning—all in elderly people—have either used a simple measure of global health status or have concentrated on physical functioning and disability.[33–36]

Socioeconomic status is inversely associated with both the risk of developing disease and the risk of people with disease experiencing complications; furthermore, there is a growing recognition that these risks are mediated via mechanisms which may change through the life course.[37] Few studies have examined the effects of socioeconomic status on changes in health in individuals. Socioeconomic status was inversely associated with baseline SF-36 and with change in score after adjustment for baseline score. The effect sizes for high versus low civil service employment grade on change in health functioning tended to be greater than for cross sectional effects at baseline, particularly among women. Among men, there was an interaction between age and grade in changes in physical functioning, such that men in the lowest grade showed greater declines with age than men in the highest grade. This is consistent with the hypothesis of environmental determinants of "successful aging."[27]

The impact (in terms of effect sizes) of physical and psychiatric morbidity on baseline scores was similar to that reported in other cross sectional studies in patient populations.[38] Effect sizes for change in score ranged from 0.20 to 0.65 in participants with both physical and psychiatric morbidity, and such changes are of comparable magnitude to short term changes after clinical interventions.[39 40] It should not be assumed that effect sizes of clinical and public health significance are equivalent[41]; indeed, the latter may be smaller.[42] The disease groups were deliberately chosen to reflect morbidity which was chronic, progressive, or recurrent. However, in the absence of independent measures of disease severity over time, it is not possible to say which of these, or other, effects were responsible for the observed changes.

The ability of the SF-36 to detect change in less healthy general populations is likely to be greater than that observed among Whitehall II participants. The Whitehall II cohort is high functioning by virtue of the comparatively young age of participants (none was older than 65 years) and the fact that all were employed as civil servants in non-industrial grades at the start of the study (in terms of mortality the Whitehall II participants enjoy a healthy worker effect of 0.5). The relatively short period of follow up (mean 36 months) and the absence of any systematic intervention further suggest the potential for the SF-36 to detect changes in health.

Potential limitations of study

Potential limitations of this study should be considered. Non-response at follow up is likely to have biased the effects conservatively, since non-responders tended to be from lower employment grades and have lower SF-36 scores at baseline. Since both the SF-36 and some of the diseases were based on self reported measures, it is possible that a reporting bias could arise. However, the use of objectively defined disease categories as well as the similarity of effect sizes with other studies which have used doctors' diagnoses[38] suggests that this is unlikely to be important.

There is no gold standard measure of change in health functioning. Changes in SF-36 scores, if valid, may be expected to be associated quantitatively with use of health services, sickness absence, morbidity, and mortality and qualitatively with individuals' own accounts of their health. Quantitative studies are required for interpreting the significance of a given level of change in SF-36 score; it can not be assumed that a small effect size (arbitrarily defined as 0.2[43]) is necessarily unimportant. Furthermore it is likely that the sensitivity in detecting true change differs between the scales of the SF-36. Future analyses within the

Key messages

- The SF-36, an inexpensive measure of health outcomes, is capable of detecting change in health in a general population
- Health and functioning do not decline uniformly with age; general mental health shows greater declines among younger participants
- Socioeconomic status is associated inversely with baseline functioning and, independently, with decline in health
- The greatest declines were seen among subjects with physical and psychiatric morbidity at baseline

Whitehall II study will examine this and, by using further repeated measures of the SF-36, the trajectories and predictors of changes in health.

Implications

In the primary care led NHS, general practitioner and health authority commissioners have responsibility for evaluating the need for and effectiveness of health care in populations defined by a general practitioner's list or residence in a health authority. Comparisons of health outcome between differing patterns of primary care (for example, between fundholders and non-fundholders) have been limited by the lack of an outcome measure sensitive to change and the inability to adjust for differences in case mix. The SF-36 may offer a partial solution. Assessing changes in health functioning has the advantage of reflecting the impact of all causes of morbidity[44] on different dimensions of health. Statistical adjustment of change in health functioning for baseline values offers the potential of a simple method for taking account of case mix, although the validity of this approach needs testing.

Changes in health functioning in hypothesised directions with age, employment grade, and disease status were observed in this young, high functioning population. These results provide sufficient support for the validity of the SF-36 in measuring change in health in populations to recommend this use in other studies, with the caveat that the validity of change continues to be tested.

We thank all participating civil service departments and their welfare, personnel, and establishment officers; the Civil Service Occupational Health Service (Dr George Sorrie, Dr Adrian Semmence, and Dr Elizabeth McCloy); the Civil Service Central Monitoring Service and Dr Frank O'Hara; the Council of Civil Service Unions and all participating civil servants. We thank Jenny Head for comments on an earlier draft.

Funding: Grants from the Agency for Health Care Policy and Research (5 RO1 HS06516); New England Medical Centre-Division of Health Improvement; National Heart Lung and Blood Institute (2RO1 HL36310); National Institute on Aging (R01 AG13196-02); John D and Catherine T MacArthur Foundation Research Network on Successful Midlife Development; Institute for Work and Health, Ontario, Canada; Volvo Research Foundation, Sweden; Medical Research Council; Health and Safety Executive; and British Heart Foundation. MM is supported by an MRC research professorship.

Conflict of interest: None.

1 Ware JE. Measuring patient function and well-being: some lessons from the medical outcomes study. In: Heithoff KA, Lohr KN, eds. *Effectiveness and outcomes in health care.* Washington, DC: National Academy Press, 1990;107-19.
2 Guyatt GH. Measuring quality of life in clinical trials: a taxonomy and review. *Can Med Assoc J* 1989;140:1441-8.
3 Fitzpatrick R. Quality of life measures in health care. I. Applications and issues in assessment. *BMJ* 1992;305:1074-7.
4 Ware JE, Donald Shebourne C. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992;30:473-83.
5 McHorney CA, Ware JE, Raczek AE. The MOS 36-item short-form health status survey (SF-36). II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Med Care* 1993;31: 247-63.
6 McHorney CA, Ware JE, Lu JFR, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). III. Test of data quality, scaling assumptions, and reliability across diverse patient groups. *Med Care* 1994;32: 40-66.
7 Greenfield S, Rogers W, Mangotich M, Carney MF, Tarlov AR. Outcomes of patients with hypertension and non-insulin dependent diabetes mellitus treated by different systems and specialities. *JAMA* 1995;274:1436-44.
8 Jenkinson C, Lawrence K, McWhinnie D, Gordon J. Sensitivity to change of health status measures in a randomized controlled trial: comparison of the COOP charts and the SF-36. *Qual Lif Res* 1995;4:47-52.
9 Brazier JE, Harper R, Jones NMB, O'Cathain A, Thomas KJ, Usherwood T, et al. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *BMJ* 1992;305:160-4.
10 Jenkinson C, Coulter A, Wright L. Short form 36 (SF-36) health survey questionnaire: normative data for adults of working age. *BMJ* 1993;306:1437-40.
11 Garratt AM, Ruta D, Abdall M, Buckingham J, Russell I. The SF-36 health survey questionnaire: an outcome measure suitable for routine use within the NHS? *BMJ* 1993;306:1440-4.
12 Lyons RA, Lo SV, Littlepage BNC. Comparative health status of patients with 11 common illnesses in Wales. *J Epidemiol Community Health* 1994;48:388-90.
13 Lyons RA, Fielder H, Littlepage BN. Measuring health status with the SF-36: the need for regional norms. *J Public Health Med* 1995;17:46-50.
14 Bullinger M. German translation and psychometric testing of the SF-36 health survey: preliminary results from the IQOLA project, international quality of life assessment. *Soc Sci Med* 1995;41:1359-66.
15 Sullivan M, Karlsson J, Ware JE. The Swedish SF-36 health survey. I. Evaluation of data quality, scaling assumptions, reliability and construct validity across general populations in Sweden. *Soc Sci Med* 1995;41: 1349-58.
16 Alonso J, Prieto L, Anto JM. The Spanish version of the SF-36 health survey (the SF-36 health questionnaire): an instrument for measuring clinical results. *Med Clin Barc* 1995;104:771-6.
17 Perneger TV, Leplege A, Etter JF, Rougemont A. Validation of a French language version of the MOS 36 item short form health survey (SF-36) in young healthy adults. *J Clin Epidemiol* 1995;48:1051-60.
18 McCallum J. The SF-36 in an Australian sample: validating a new generic health status measure. *Aust J Public Health* 1995;19:160-6.
19 Hemingway H, Nicholson A, Stafford M, Roberts R, Marmot M. The impact of socioeconomic status on health functioning as assessed by the SF-36 questionnaire: the Whitehall II study. *Am J Public Health* 1997;87:1484-90
20 Ware JE, Snow KK, Kosinski M, Gandek B. *SF-36 health survey manual and interpretation guide.* Boston: New England Medical Center, 1993.
21 Rose G, Blackburn H, Gillum RF, Prineas RJ. *Cardiovascular survey methods.* Geneva: World Health Organisation, 1982.
22 WHO Study Group. *Diabetes mellitus: report of a WHO Study Group.* Geneva: World Health Organization, 1985.
23 Medical Research Council. *Questionnaire on respiratory symptoms and instruction for interviewers.* London: MRC, 1976.
24 Goldberg DP. *The detection of psychiatric illness by questionnaire.* Oxford: Oxford University Press, 1972.
25 Glynn RJ, Rosner B, Silbert JE. Changes in cholesterol and triglyceride as predictors of ischemic heart disease in men. *Circulation* 1982;66:724-31.
26 Cain KC, Kronmal RA, Kosinski AS. Analysing the relationship between change in a risk factor and risk of disease. *Statistics in Medicine* 1992;11:783-97.
27 Rowe JW, Kahn RL. Human ageing: usual and successful. *Science* 1987;237:143-9.
28 Seeman TE, Charpentier PA, Berkman LF, Tinetti ME, Guralnik JM, Albert M, et al. Predicting changes in physical performance in a high functioning elderly cohort: MacArthur studies on successful aging. *J Gerontol* 1994;49:M97-108.
29 Seeman TE, Berkman LF, Charpentier PA, Blazer DG, Albert MS, Tinetti ME. Behavioural and psychosocial predictors of physical performance: MacArthur studies of successful aging. *J Gerontol Med Sci* 1995;4: M177-83.
30 Beckett LA, Brock DB, Lemke JH, Mendes de Leion CF, Guralnik JM, Fillenbaum GG, et al. Analysis of change in self-reported physical function among older person in four population studies. *Am J Epidemiol* 1996;143:766-78.
31 Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic research: principles and quantitative methods.* New York: Van Nostrand Reinhold, 1982.
32 Ferrie JE, Shipley MJ, Marmot MG, Stansfeld S, Davey Smith G. Health effects of anticipation of job change and non-employment: longitudinal data from the Whitehall II study. *BMJ* 1995;311:1264-9.
33 Markides KS, Lee DJ. Predictors of well-being and functioning in older Mexican Americans and Anglos: an eight-year follow-up. *Gerontology* 1990;45:S69-73.

34 Strawbridge WJ, Camacho TC, Cohen RD, Kaplan GA. Gender differences in factors associated with change in physical functioning in old age: a 6 year longitudinal study. *Gerontologist* 1993;33:603-9.
35 Palmore EB. Predictors of function among the old-old: a 10-year follow-up. *J Gerontol* 1985;40:244-50.
36 Harris T, Kovar M, Suzman R, Kleinman JC, Feldman JJ. Longitudinal study of physical ability in the oldest-old. *Am J Public Health* 1989;79:698-702.
37 Kuh D, Ben-Shlomo Y, eds. *A life course approach to chronic disease epidemiology.* Oxford: Oxford Medical Publications, 1997.
38 Spitzer R, Kroenke K, Linze M, Hahn SR, Williams JBW, DeGruy FV III, et al. Health related quality of life in primary care patients with mental disorders: results from the PRIME-ME 1000 study. *JAMA* 1995;274:1511-7.
39 Gliklich RE, Hilinski JM. Longitudinal sensitivity of generic and specific health measures in chronic sinusitis. *Qual Lif Res* 1995;4:27-32.
40 Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH. Comparative measurement of sensitivity of short and longer health status instruments. *Med Care* 1992;30:917-25.
41 Lydick E, Epstein RS. Interpretations of quality of life changes. *Quality of Life Research* 1993;2:2216.
42 Rose G. Sick individuals and sick populations. *Int J Epidemiol* 1985;14:32-8.
43 Cohen J. *Statistical power analysis for the behavioural sciences.* New York: Academic Press, 1977.
44 Ettinger WH, Fried LP, Harris T, Shemanski L, Schulz R, Robbins L. Self-reported causes of physical disability in older people: the cardiovascular health study. CHS Collaborative Research Group. *J Am Geriatr Soc* 1994;41:1035-44.
45 Barsky AJ. The paradox of health. *N Engl J Med* 1988;318:414-8.

(Accepted 7 July 1997)

Deficient colour vision and interpretation of histopathology slides: cross sectional study

C J M Poole, D J Hill, J L Christie, J Birch

Abstract

Objective: To determine whether histopathologists with deficient colour vision make more errors in slide interpretation than those with normal colour vision.
Design: Examination of projected transparencies of histopathological slides under standardised conditions by subjects whose colour discriminating ability was accurately assessed.
Setting: Departments of histopathology in 45 hospitals in the United Kingdom.
Subjects: 270 male histopathologists and medical laboratory scientific officers.
Main outcome measures: Number of slides correctly identified by subjects whose colour vision was measured on the Ishihara, City University, and Farnsworth-Munsell 100 hue tests.
Results: Mean (SD) scores (out of 10) for doctors with colour deficient vision were 9.4 (0.7) *v* 9.9 (0.4) for controls ($P < 0.01$) and 7.5 (1.6) *v* 9.4 (0.7) for scientific officers ($P < 0.001$). When subjects with colour deficient vision were categorised into severe, moderate, or mild, there was a significant trend towards those with severe deficiency making more mistakes ($P < 0.001$).
Conclusions: Histopathologists and medical laboratory scientific officers should have their colour vision tested; if they are found to have a severe protan or deutan deficiency, they should be advised to adopt a safe system of working.

Introduction

Histopathologists use a variety of coloured stains as an aid to diagnosis. In some cases the additional information provided by colour is superfluous and a diagnosis can be made by pattern recognition in a clinical context. Many of the stains use combinations of pigments which may be difficult for people with congenital red-green colour deficiency (dyschromatopsia) to distinguish, and there is evidence that doctors with colour deficient vision have problems with histopathological diagnoses.[1-3] There are also anecdotal accounts of trainee histopathologists leaving the specialty because of difficulties with colour recognition.

If accurate colour discrimination is relevant to the diagnostic process it is important to guide doctors with colour deficient vision in the use, or avoidance, of particular stains and techniques that might cause difficulty. Normal colour vision may also be important to medical laboratory scientific officers who prepare histological specimens for diagnosis. As the prevalence of congenital red-green colour deficiency is much greater in males (8%) than in females (0.5%), we compared the performance of male histopathologists and medical laboratory scientific officers who had colour deficient vision with that of their colleagues who had normal colour vision in answering questions about stained histopathological slides.

Methods

A total of 270 male histopathologists and medical laboratory scientific officers in 45 hospitals across the United Kingdom took part in the study. Letters and a protocol were sent to the heads of departments of histopathology explaining the project and asking for volunteers who had spent more than a year in histopathology to take part. Subjects were asked by DH to answer questions on 20 projected (Kindermann) 35 mm transparencies of histopathological slides. The slides were chosen because people with colour deficient vision might confuse the colours in the stains. Transparencies of specific features taken from slides rather than the slides themselves were used so as to standardise the process of observation and ensure that each subject had the opportunity of seeing the same features.

The slides (see table) were in pairs and consisted of a true and false answer. Each correctly identified slide was worth half a mark, and subjects were marked out of 10.

After subjects answered questions about the transparencies, an optometrist (DJH) assessed their colour vision using the 38 plate, 1989 edition of the Ishihara plates; the City University test, second edition;

Department of Occupational Health, Dudley Priority Health NHS Trust, Central Clinic, Dudley DY2 7BX
C J M Poole, *consultant occupational physician*

Department of Histopathology, Dudley Group of Hospitals NHS Trust, Russells Hall Hospital, Dudley DY1 2HQ
J L Christie, *consultant histopathologist*

Department of Optometry and Visual Science, City University, London EC1 0HB
D J Hill, *optometrist*
J Birch, *senior lecturer*

Correspondence to: Dr Poole

BMJ 1997;315:1279–81