# Can we spot deleterious ageing in two waves of data? The Lothian Birth Cohort 1936 from ages 70 to 73

**Wendy Johnson,[1,2,4] Alan J Gow,[1,4] Janie Corley,[4] Paul Redmond,[4] Ross Henderson[5], Catherine Murray,[4] John Starr,[1,3,4] and Ian J Deary[1,4]**

[1]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, UK

[2]Department of Psychology, University of Minnesota – Twin Cities, USA

[3]Department of Geriatric Medicine, University of Edinburgh, UK

[4]Department of Psychology, University of Edinburgh, UK

[5]Edinburgh Medical School, Edinburgh, UK

wendy.johnson@ed.ac.uk

(Received April 2012 Revised August 2012)

## Abstract

*'Younger' old age (the late 60s through early 70s) is, for many, a period of stability of lifestyle and considerable freedom to pursue leisure activities. Despite the stability that many enjoy, the mortality rate is about 2% per year in western nations. This increases to about 5% by age 80. It would be useful to know if those most vulnerable can be identified through patterns of deleterious ageing, and especially if this could be accomplished with just two waves of data. The Lothian Birth Cohort 1936 was surveyed on a host of individual difference variables including cognition, personality, biomarkers of physical health, and activities at ages 70 and 73 years. Overall, the group showed the expected basic stability in mean levels for these variables, but some individuals had died and others did show substantial changes that could be considered statistically reliable. These presumably reliable changes were at least as likely to be positive (reflecting improved condition/ability) as negative (reflecting decline/ageing). Moreover, limitations in the estimated reliabilities of the measures meant that most of the observed changes could not be considered reliable. The changes clustered only weakly around general health to predict death over the next approximately two years. We concluded that two waves of longitudinal data were not sufficient to assess meaningful patterns of ageing, despite often being used to do so.*

**Keywords**: cognitive ability, health, longitudinal data, ageing, mortality

## Introduction

'Younger' old age (the late 60s through early 70s) is, for many in developed economies, a period of stability of lifestyle and general condition. Though far from universally true, many are in good health, experience few restrictions in activities, enjoy the newfound freedoms of recent occupational retirement and absence of financial responsibility for offspring, and are only beginning to tap their old-age financial resources. Despite this stability and generally favourable picture, however, at age 70 the mortality rate in western nations is about 2% per year, with male mortality running about 50% higher than female (http://www.mortality-trends.org/, 16 January, 2012, based on World Health Organisation data). The rate increases to about 5% per year by age 80. This means that about 22% of the age-70 population dies between ages 70 and 80.

Unfortunately, few in western nations are fortunate enough to experience good health until just prior to death; many go through some period of illness and/or increasing disability in the years prior to death, a period that often involves substantial reduction in quality of life, inability to live independently, and extensive medical care with its associated costs. Given these facts, the ability to identify those most likely to die within the next 5-10 years in an apparently healthy group in this younger range of old age could be very useful in developing methods to minimize these periods of illness and disability and perhaps even to extend longevity.

This idea has long been discussed with respect to cognitive ageing, or the normative declines in many cognitive functions that accompany old age. Beginning with observations in the late 1950s and early 1960s, that those who survived and agreed to be retested at some later point in time had generally showed higher original cognitive test scores than those who either refused testing or died in the interim, gerontologists have speculated that perhaps the observed declines in mean function with age result primarily from the presence in elderly samples of individuals who do not survive long beyond assessment (Jarvik & Falek, 1963; Lieberman, 1966; Rabbitt et al., 2011; Riegel & Riegel, 1972). That is, it is possible that decline in function is minimal until some overt disruption takes place, after which decline is quite rapid and ends in death. Even samples screened for clinical manifestations of impairment would inevitably include some who had entered this period but remained undiagnosed, and their numbers could be expected to increase with age, which could create the declines in mean function with age that studies consistently show. As longitudinal data samples have proliferated and statistical analytical techniques have improved, research efforts have been directed toward describing individual trajectories of decline (e.g., Finkel et al., 2005; McGue & Christensen, 2002) and/or the specific interval before death at which some period of 'terminal decline' begins (e.g., Rabbitt et al., 2011; Sliwinski et al., 2006; Terrera et al., 2011; Wilson et al., 2003). These studies have produced widely varying results about the extent of change, its rate of acceleration if considered, and the length of any terminal decline interval.

There are likely many features of study design that contribute to variation in results, including differences in sample selectivity, differences in measures assessed and the rates of normative rates of change to which they may be subject, unacknowledged constraints on results imposed by the models used, and the large age ranges sampled in many studies. In addition, three inter-related realities complicate even the most optimal study design. These apply to all measures of aging, whether cognitive, physiological, or behavioral. First, at any age, there are large individual differences in function. They stem from diverse sources, but many of the most salient ones show considerable rank-order stability throughout the lifespan. For example, those within an ageing cohort who perform relatively well on cognitive tasks or show relatively high levels of physical fitness, would have tended to do so in youth as well, had they been assessed then. This is borne out by data from the few studies for which such information is available. For example, the correlation between IQ scores at ages 11 and 79 in the Lothian Birth Cohort 1921 was .66; with correction for restriction of range in the sample, it rose to .73 (Deary et al., 2004). Without clear recognition in study design of the stability of inter-individual variability, this stability can easily be mistaken for intra-individual change, especially in cross-sectional studies or longitudinal studies with large sample age ranges in relation to the follow-up periods (Sliwinski, Hoffman, & Hofer, 2010). The second reality is that, over the timespans of most longitudinal studies, average rates of decline in function are small in relation to lifespan-stable individual differences in level of function. And third, individual differences in rates of decline are also small in relation to individual differences in lifespan-stable level. These latter realities act to keep statistical power low to detect rates of change accurately.

The problem of estimating rates of change or intervals of terminal decline is further complicated by the strong likelihood that, in addition to individual differences in linear rates of decline, there is also meaningful variability in rates of acceleration of decline with age and/or intervals of terminal decline. Most modeling of terminal decline intervals has been based on the assumption that there is one uniform interval of decline for all (Sliwinski et al., 2006), and violations of this assumption may especially bias estimates of individual differences in rates of change before and after the beginning of the supposedly uniform terminal decline interval. One very likely

possibility is that any terminal decline intervals vary at least with the pathology finally involved in death. Evidence for this was provided recently (Rabbitt et al., 2011).

## Links among physical, cognitive, and psychosocial characteristics in ageing

Gerontological research over the past 10 years or so has increasingly suggested that ageing is a rather general systemic process. That is, degrees of well-being in many aspects of life; including physiological processes and biomarkers, health as measured both by self-assessment, and more objective criteria such as clinical diagnoses and medication, cognitive function, and psychosocial affect; all tend to be correlated, in level and possibly also rate of change (e.g., Backman & MacDonald, 2006; Deary et al., 2011; Dixon, 2011; Dolcos et al., 2012; Johnson et al., 2009; Li & Lindenberger, 2002). This repeated observation alone suggests that, examination of a wide range of potential risk and protective factors in the same group of individuals, may be worthwhile in helping to identify individuals entering final stages of ageing that portend some period of acute disability or impairment ending in death.

Theoretical considerations also point to the value of examining a wide range of potential factors involved in ageing. There are several reasons that many otherwise disparate aspects of function may be linked, particularly in old age. First, many chronic physical illnesses common in old age, including cardiovascular disease and diabetes (e.g., Cosway et al., 2001; Hassing et al., 2004; Rafnsson et al., 2007; Schram et al., 2007) also undermine cognitive function, possibly because they increase proinflammatory and oxidative stress markers and impede vascular function. These conditions are often aggravated by failure to adhere to somewhat detailed treatment regimens, and such failure is more common when cognitive function is impaired (Deary et al., 2009). As well, the chronic nature of these conditions and associated disabilities can contribute to reduction in quality of life, leading to depression (Fiske, Wetherell, & Gatz, 2009; Kendler et al., 2009). Second, lifetime-stable cognitive ability can contribute to the development and maintenance of lifestyle and habits such as nutrition, exercise, smoking, and drinking that support or undermine physical health. Evidence for this comes from studies (e.g. Adler & Snibbe, 2003; Hart et al., 2003) that have found that lower IQ scores in childhood were associated with smoking and other unhealthy habits in later life as well as with greater morbidity (Batty, Deary & Macintyre, 2007; Batty et al., 2007; Deary, Weiss & Batty, 2010). Also, lower IQ scores have also been robustly associated with poorer living circumstances that can contribute to poor psychological wellbeing and depression (Adler & Snibbe, 2003; Gallo & Matthews, 2003). Finally, there may be individual differences in some form of biological or constitutional 'integrity' and/or ageing processes that contribute to both physical and cognitive function as well as ability to sustain psychological wellbeing (Christensen et al., 2001; Gale et al., 2009; Li & Lindenberger., 2002). Clearly, these possibilities are not mutually exclusive.

## Measuring change in evaluating ageing processes in two data waves

The simplest way to evaluate ageing processes is through analysis of differences among individuals of different ages. Early studies of ageing made it clear that this is of limited value, however, due to the large likelihood of sampling differences and cohort effects among the ages (e.g., Riegel & Riegel, 1972). This realisation led to the development of longitudinal studies. Efficiency and convenience in data collection is always a consideration, so many studies have been designed to sample individuals from a wide age range, for example 20 years, only twice, for example 3 years apart. This makes it possible in principle to address change over the whole 23-year period. Though many samples with this design remain in use in recent publications (e.g., Dolcos et al., 2012; Gayman, Turner & Cui, 2008; Gerstorf, Rocke & Lachman, 2011; Hanson et al., 2011; Kooij & Van De Voork, 2011; Lapi et al., 2009; Mather et al., 2010, Menezes et al., 2011; Ramsden et al., 2011; Schelleman-Offermans, Kuntsche & Knibbe, 2011; Whitehead et al., 2011), the limitations of two waves of data for understanding change have been well documented (e.g., Rogosa, 1995; Rogosa, Brandt, & Zimoski, 1982), as have the additional complications introduced by large age ranges within samples (e.g., Sliwinski, Hoftman, & Hofer, 2010). Essentially, the limitations of two waves of data surround the fact that, with only two waves of data, it is not possible to distinguish true between-individual differences in overall level from error of measurement in the estimation of indvidual change trajectories. Additional complications are introduced in samples with wide age ranges because it is necessary to assume that the individual change trajectories depend only on

the ages of the indviduals, and not on when the individuals attained those ages. There is long-standing and substantial evidence that this assumption is often violated (e.g., Kuhlen, 1940; Schaie, 1965). In addition, it is necessary to assume that each age group within the sample represents the underlying population to the same degree. This assumption too is generally violated due to the underlying associations between survival and ability to participate at any given age and overall function noted above.

### The purpose of this study

Despite these limitations and complications, pressure from funding agencies, publication goals, and sheer scientific curiousity provide strong temptation to wring some information from two waves of data, while study administrators await the availability of, and/or justify requests for, funding of additional data waves. Though limited, some information *can* be gleaned from such an approach, and the greater the volume of data available about each participant, the greater the amount of information it should be possible to extract, especially if the complications associated with wide sample age ranges can be avoided. At 2% mortality rate per year, some proportion of participants should be either in or entering such periods, though the specific proportion would depend on the length of the terminal decline period. The purpose of this study was to explore the potential capacity to use two data waves to identify individuals who, in the period from ages 70 to 73 years, might be experiencing negative changes in many areas consistent with terminal decline. We did this through examination of three questions: 1) How and to what extent did individuals change during this period? 2) Were there correlates or predictors of these changes? and 3) Did changes tend to cluster in ways that could distinguish healthy ageing from terminal decline? Our analysis made use of the Lothian Birth Cohort 1936 (LBC1936; Deary et al., 2007; Deary et al., in press), a sample of 1,091 initially healthy 70-year-olds living in the Edinburgh area of Scotland, all of whom were born in 1936 and who completed a broad assessment of both cognitive and physical function. We thus avoided the analytical complications associated with samples with large age ranges and had access to a wealth of information about these individuals.

## Method

### Participants

The LBC1936 study was designed to take advantage of the Scottish Mental Survey 1947 (Scottish Council for Research in Education 1949). On 4 June 1947, almost all children born in 1936 and attending school in Scotland on that day completed a valid cognitive ability test. LBC1936 recruited 1091 of these individuals who were living independently in the area of Edinburgh, Scotland when they were mean age 70 between 2004 and 2007, with the intent of following them through old age. Recruitment was limited to the Edinburgh area for practical reasons of ease of access to the clinical research facility; within the recruitment catchment area the goal was to recruit at least 1,000. The only eligibility criteria were birth in 1936, enrolment in school in Scotland at age 11, and current ability to get to the clinic to participate in about 4 hours of psychological and medical testing. Taxi transportation was provided if needed. Recruitment was accomplished with the assistance of the Lothian Health Board and through advertisements. The Lothian Health Board wrote to 3,810 individuals on the Lothian Community Health Index who were born in 1936 and thus might have taken part in the Scottish Mental Survey 1947, of whom 3,686 were invited to hear about the study. Of these, 1,703 (46.2%) responded, 1,226 (72.0% of 1,703) were interested and considered themselves eligible, and 1,091 (89.0% of 1,226) participated, with some small supplementation from advertisements. See Deary et al., (2007 in press) for further details on participant recruitment and assessment. Participants included 548 males and 543 females, aged 67.7 to 71.3 years at time of first assessment in old age (mean=69.6, SD =0.80). The presence in the sample of fewer females than males suggests that the female participants may have been less representative of the overall population in this age group than the males, as the population sex ratio favours females in this age group due to longer female longevity. Ethical approval for the study was granted by the Multi-Centre Research Ethics Committee for Scotland and by Lothian Research Ethics Committee. The study was carried out in compliance with the Helsinki Declaration.

The sample was assessed in essentially the same way a second time, approximately 3 years later (mean=3.0, SD=.3). Of the original sample, 866 (79%) returned (448 males, 82% of original; 418

females, 77% of original). The primary reason for failure to return was death, or self-assessment of inability to participate. Compared to returning participants, non-returners were poorer at the first assessment in immune and inflammation indicators; lung function; walk and visual search speed; memory span; and performance on logical memory, matrix reasoning, block design, and reaction time task performance, with standardised mean score differences (Cohen's *d*s) ranging from .25 to .35. Non-returning participants also had lower cognitive ability scores at age 11 from the Scottish Mental Survey (*d*=.22), fewer years of education (*d*=.20), earlier retirement age (*d*=.19), and lower current social class (*d*=.20), suggesting that their lower level of at least cognitive function may have been long-term rather than some indication of greater failing health or proximity to death.

## Measures

Participants were interviewed and tested individually during a single session in each testing wave, by a trained psychology researcher and a research nurse at the Wellcome Trust Clinical Research Facility at the Western General Hospital in Edinburgh. The assessment was broken by two periods of at least 15 minutes for rest and refreshments. It began with orientation to the study and collection of informed written consent to participate, followed by provision of basic demographic and medical information through structured interview. This included educational attainment, primary occupation during working life (and that of spouse for married women), age of retirement, history of medical diagnoses and current prescription medications, smoking history and current status, and current pattern of alcohol consumption. For this study, we made use of the current demographic variables and numbers of current medical diagnoses and prescription medications from this interview (see Deary et al., 2007 for further details of the assessment).

***Hospital anxiety and depression scale*** (Zigmond & Snaith 1983). Participants completed this written questionnaire, which consists of 14 items, half of which reflect anxiety and half depression. Maximum score on each scale is 21, with probable clinical levels at scores of at least 11.

***Tests of current cognitive function***. Participants completed a battery of cognitive tasks intended to measure various aspects of cognitive function. The tests used here, their content, and sources are listed in Table 1.

***Test of childhood cognitive function***. Most participants in LBC1936 had participated in the Scottish Mental Survey 1947, which took place on 4 June 1947, when participants were age about 11 years. The primary focus of this survey was administration of Moray House Test No. 12, a well-validated predominantly verbal reasoning test. Scores on this test were obtained from the survey records for the purposes of LBC1936. LBC1936 participants completed the same test again at age 70.

***Physical examination and interview***. This included measurement of height and weight, time in seconds to walk 6 metres, demi-span in cm, responses to a 9-item activities of daily living scale (Townsend, 1979), sitting and standing systolic and diastolic blood pressure, forced expiratory volume from lungs in 1 sec. (FEV1; best of 3), grip strength in the right and left hands, and corrected and uncorrected vision in right and left eyes. Participants provided blood samples used to assess haemoglobin, white cell and platelet counts, prothrombin time, activated partial thromboplastin time (APTT), fibrinogen, serum folate, albumin, calcium, cholesterol, HDL cholesterol, glycated haemoglobin (HbA1C), C-reactive protein levels, and estimated glomerular filtration rate. To avoid distortions, we did not make use of prothrombin time for any participant taking the medication warfarin.

***LBC1936 Study questionnaires***. Participants were distributed questionnaires and stamped return envelopes, with instructions for completing the questionnaires at home and returning them. For this study, we made use of the personality questionnaire that was measured at both time points. This was the International Personality Item Pool inventory of 50 items, 10 measuring each of the so-called Big Five personality traits of Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Intellect**,** which can be freely downloaded at http://ipip.ori.org/. The word 'I' was added to each of the fragments making up these items, to make them more closely match wording in other questionnaires used. Participants rated how well they believed the items described them on a 5-point scale (very accurate to very inaccurate). At Time 1, 87% of participants returned these questionnaires. At Time 2, 99% did so.

**Table 1. Cognitive tests administered at ages 70 and 73 to Lothian Birth Cohort 1936 participants**

| Test | General Description | Source (Citations below) |
|---|---|---|
| Logical Memory | Verbal declarative memory | Wechsler Memory Scale-III$^{UK}$ |
| Spatial Span | Non-verbal memory | Wechsler Memory Scale-III$^{UK}$ |
| Verbal Paired Associates | Verbal learning and memory | Wechsler Memory Scale-III$^{UK}$ |
| Symbol Search | Speed of information processing | Wechsler Adult Intelligence Scale-III$^{UK}$ |
| Digit Symbol | Speed of information processing | Wechsler Adult Intelligence Scale-III$^{UK}$ |
| Matrix Reasoning | Pictorial pattern inference | Wechsler Adult Intelligence Scale-III$^{UK}$ |
| Letter-Number Sequencing | Working memory | Wechsler Adult Intelligence Scale-III$^{UK}$ |
| Digit Span Backwards | Manipulation of memory | Wechsler Memory Scale-III$^{UK}$ |
| Block Design | Constructional ability | Wechsler Adult Intelligence Scale-III$^{UK}$ |
| Simple Reaction Time | Mean response to simple stimulus | |
| Choice Reaction Time | Mean defined response to specific stimulus among 4 alternatives | |
| Inspection Time | Visual discrimination of briefly presented stimulus | |
| Verbal Fluency | Attention focus; association flexibility | |
| National Adult Reading Test | Reading vocabulary | |
| Wechsler Test of Adult Reading | Reading vocabulary | |
| Mini-Mental State Exam | Space/time orientation; dementia screen | |

*Notes. Wechsler Memory Scale-III$^{UK}$: Wechsler, D. (1998). WMS-IIIUK Administration and Scoring Manual. London: Psychological Corporation. Wechsler Adult Intelligence Scale-III$^{UK}$: Wechsler, D. (1998). WAIS-IIIUK Admini stration and Scoring Manual. London: Psychological Corporation. Reaction times: Deary, I. J., Der, G., & Ford, G.(2001). Reaction times and intelligence differences: A population-based cohort study. Intelligence, 29, 389-399. Inspection Time: Deary, I. J., Simonotto, E., Meyer, M., Marshall, A., Marshall, I., Goddard, N., & Wardlaw, J. M. 2004). The functional anatomy of inspection time: an event-related fMRI study. NeuroImage, 22, 1466-1479. Verbal fluency: Lezak, M. (2004). Neuropsychological Testing. Oxford: Oxford University Press. National Adult Reading Test: Nelson, H. E., & Willison, J. R. (1991). National Adult Reading Test (NART) Test Manual (Part II). Folstein, S. E., & McHugh, P. R. (1975). Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research, 12, 189-198.*

***Death information***. The study receives monthly reports of participant deaths from the General Register Office for Scotland, part of the National Records for Scotland. We made use of reports through 31 December, 2011. At that date, there had been 30 deaths among participants who had completed Time 2 assessments, the majority in 2009.

## Data treatment

We followed several steps in preparing the data for analysis. First, we regressed the effects of sex and height from the measures of 6-metre walk time, demi-span, FEV, and grip strength. We then standardised all variables to place them on the same scale, trimming isolated outliers by observation, separately by variable. We consolidated height and weight by calculating body mass index (BMI) as weight in kg/height in metres$^2$. Several other variables tapped similar constructs, suggesting the formation of summary or composite variables. We combined the systolic and diastolic blood pressure readings to estimate mean arterial pressure, using the formula (mean sitting diastolic reading) + (mean sitting systolic reading – mean sitting diastolic reading)/3. We averaged grip strength readings for right and left hands; scores from the Digits Backwards and Letter-Number Sequencing tests to form a Memory Span variable; scores from the two scores from the two reading tests to form a Word Reading variable; corrected vision in right and left eyes; standardised prothrombin time and APTT to form an indicator of blood clotting function; standardised folate, albumin, and cholesterol levels (reversed as appropriate) to form a variable indicating nutritive status; standardised platelet counts, fibrinogen levels, and C-reactive

protein levels to measure inflammation; and standardised HDL cholesterol and glycated haemoglobin (reversed as appropriate) to form an indicator of metabolic vascular risk. This resulted in 36 variables for study. We adjusted Time 2 standardised variables to reflect their mean standardised differences from Time 1 standardised variables, while preserving variance differences from Time 1. Finally, we subtracted Time 1 values from Time 2 values to produce change scores.

The decision to make use of difference scores as the measure of change requires some explanation, as statistical advice against this is not uncommon (e.g., Campbell & Kenny, 1999; Cronbach & Furby, 1970). Two common observations have led to this advice. First, difference scores are often negatively correlated with Time 1 scores, leading to a perception that the difference score is a negatively biased estimate of change. The reality is the opposite: the difference score is an unbiased estimate of true change, but the observed correlation is a negatively biased estimate of the correlation between true initial status and true change (Rogosa, Brandt, & Zimowski, 1982), due to the presence of the error of measurement associated with the observation of initial status in both that observation and the observed change. Second, it has been common to assume that variance remains constant from one measurement occasion to another. Under this assumption, as the correlation between the two measurement occasions increases, the reliability of the difference score decreases, yielding the impression that greater reliability of measures leads to lower reliability of difference scores. In reality, however, it is very common in developmental situations for variances to change over time (e.g., McCardle & Woodcock, 1997). The reliability of the difference score is sensitive to these changes (Nesselroade & Cable, 1974). The greater the changes are, the more reliable is the difference score. Because variance changes are common, the difference score is commonly quite reliable. The most commonly used alternative, the residual from regressing Time 2 values on Time 1 values, is imprecise and often considerably biased (Rogosa, Brandt, & Zimoski, 1982). Perhaps most important, however, the regression residual does not address the simple question of change. Instead, it addresses the question of what the expected change for an individual would have been, had that individual been at the mean level at Time 1. Appropriate interpretation of any answer to this question is far from clear.

## Results

### Basic change statistics

Table 2 shows descriptive statistics for the raw study variables at the two assessments. Most variables showed large variation among individuals, but the means of some gave clear indications of the overall health and wellbeing of participants. For example, at Time 1, participants had on average (SD) 3.8 (1.9) diagnosed medical conditions, for which they took an average of 3.0 (2.5) prescription medications. By Time 2, these averages had grown to 4.4 (2.0) and 4.0 (2.3). Despite this, average difficulties with activities of daily living were very small at both time points, with most participants reporting no difficulties at all. Average BMIs were 27.82 (4.56) and 27.92 (4.43), respectively at Times 1 and 2, which would generally be considered overweight but not obese. Hospital Anxiety and Depression averages were low at both time points, indicating generally good levels of well-being. Mini-Mental State Exam average scores were 28.8 (1.4) and 28.8 (1.4) at Times 1 and 2 respectively. At Time 1, 21 had scores less than 25; at Time 2, 18, and the lowest score at both time points was 22. Thus, most participants suffered some health impairments, but not sufficiently to have undermined basic cognitive function and overall well-being.

The variables all showed both substantial correlations between Time 1 and Time 2 indicating considerable stability and substantial individual differences in extent of change. The smallest correlation was .47 for the MMSE, likely because of restriction of range due to a strong ceiling effect. The largest correlation was .96 for Word Reading, and the overall mean correlation was .71. Across all the variables, the mean largest individual participant increase was 3.49 SDs, and the mean largest decrease was 3.91 SDs. The distributions of change variables were generally close to normal, with average skewness at -.002. This suggested two things: first, positive change was basically as likely as negative change. Second, it would likely be difficult to identify participants who were experiencing terminal decline, which would be evidenced by substantially skewed distributions in which most values hovered around 0. The mean change, adjusted to reflect the direction of each measure that indicated decline in function, was -.01 SD, indicating small overall absolute decline. Acknowledging the multiple tests run, 26 of the 36 variables showed no significant mean change. There was considerable heterogeneity in direction even among those variables showing significant mean

change, and cognitive functions were largely stable at the mean level. Lung function, walk speed, reading and search abilities, nutritive status, and grip strength declined significantly and participants were taking significantly more medications. Inflammatory and metabolic vascular risk markers, and corrected vision improved significantly. The variables that showed significant declines more likely indicated ageing, while the variables that showed overall improvement more likely reflected better health care. Full details of this information are shown in Table 3.

## Reliability of changes

Another way to evaluate changes is to measure the extent to which the changes observed could be considered reliable. To do this, we made use of the Reliable Change Index (Christensen & Mendoza, 1986; Hsu, 1989), which indicates the magnitude of change that can be considered reliable after accounting for measurement error and regression to the mean. The full formula is

$$\frac{x_2 - x_1}{\sqrt{2SD_1^2(1-r)}}$$

where the subscripts refer to the time points, *x* to a data point, *SD* to standard deviation, and *r* to test-retest reliability. The denominator is the standard error of the difference between the two test scores, and describes the expected variability in change scores if no actual change occurred. If the variables are normally distributed, the index will be too, and there will be 95% probability that change did occur if the index is greater than 1.96. Conceptually, the situation is analogous to inferring that mean differences are significant when their 95% confidence intervals do not overlap: here the differences between measures at two time points are significantly different when the intervals reflecting their standard errors of measurement do not overlap.

Implementing this formula involved some judgment, as the formula requires short-term test-retest reliability of the measures and this information was not available for most of our measures. Psychometricians tend to think of test reliabilities in the range of .75-.85 (often inappropriately assessed with a measure of internal consistency rather than short-term test-retest correlations) as strong. Medical practitioners, however, tend to think of a clinical measure as reliable when a short-term reassessment would likely generate a deviation of no more than 10% from the first observation, and many of our variables were clinical/medical. To understand the medical perspective in psychometric terms, consider IQ scores and the T-scale. Most IQ tests are scaled with mean 100 and SD 15, and the T-scale has mean 50 and SD 10. Medical practitioners, then, might expect a short-term retest of an IQ-scale observation of 100 to generate a score between 90 and 110, and a T-scale observation of 50 to generate a score between 45 and 55. Translation to psychometric perspective can be modeled by adding uniformly-distributed random values within .67 SD to the IQ-scale scores and within .5 SD to the T-scale scores from any sample, and correlating these scores with the original scores. This generates correlations in the range of .92-.97 for .67 SD and .96-.98 for .5 SD. To produce correlations in the range generally considered to indicate reliability by psychometricians, it is necessary to add random uniformly-distributed values within 1.0-1.5 SD, or 15-22 points on the IQ scale and 10-15 points on the T-scale. The reliability guidelines used by psychometricians thus allow very considerable individual variation, perhaps more than most researchers realise.

This was reflected in application of the Reliable Change Index to our data. We lacked short-term test-retest reliability data for our measures in general, but could reasonably assume that average reliability lay in the .75-.85 range. We thus estimated the numbers of variables on which each participant showed reliable changes if test-retest reliability was .75 and if test-retest reliability was .85. Assuming .85 reliability, on average, participants showed reliable changes in 10.4 (SD=6.0) of the 36 variables, with a range of 0-24. The distribution of reliable changes was slightly skewed at .70. Assuming .75 relability, on average participants showed reliable changes in 7.9 (SD=4.8, skewness=.43) variables, with a range of 0-22. Regardless of level of test reliability, about half the reliable changes represented improvements rather than declines in function. We also separated the reliable changes for each participant into those that represented decline and improvement. Numbers of reliable changes indicating improvements were correlated .53 ($p$<.001) with numbers of reliable changes indicating declines in function assuming .85 reliablity, and .52 ($p$<.001) assuming .75 reliability, indicating a substantial tendency for the same participants to show changes reflecting both improvements and declines in function. This is summarised in Table 4.

### Associations with pervasiveness of change

Number of reliable changes was not significantly correlated with age-11 IQ (.00 [-.04], assuming .85 [.75] reliability, nor with age-70 IQ (.04 [.01]) (all $p$>.15). Years of education and social class status were similarly uncorrelated with reliable changes. With no adjustment for multiple testing and assuming .75 reliability (thus a liberal reading), numbers of reliable changes were correlated with changes in numbers of drugs (.12), BMI (-.08), six-metre walk time (.24), ADLs (.10), mean arterial pressure (-.08), FEV (-.07), Logical Memory (-.13), Search Speed (-.13), simple reaction time (.07), inspection time (-.08), haemoglobin (-.10), and metabolic vascular risk (-.11), generally indicating that greater decline in function was associated with greater numbers of reliable changes. This is summarised in Table 4. Most Time 1 variables were not correlated with number of reliable changes, but there were more significant correlations between better function and number of changes indicating declines in function, than changes indicating improvements in function. These observations thus likely reflected the negative bias in correlations between observed Time 1 scores and observed difference scores, which would make them largely artefactual; one hint that this might be the case was that the strongest such associations with number of reliable changes were -.15 with IPIP Conscientiousness and -.12 with IPIP Agreeableness.

### Clustering of change variables

Significance of the correlation between numbers of changes indicating improvements and declines in function suggested some tendency for changes to cluster, though it also suggested that the changes may be random. We nevertheless ran a factor analysis of the change variables. Parallel analysis indicated 4 factors, but most variables did not show substantial loadings on any of them. Table 4 shows the results, which did support some clustering of sources of change. The first factor appeared to indicate declines in memory and speed of information processing; the second suggested declines in emotional well-being and positive personality function; the third grouped increases in numbers of diseases and numbers of drugs taken with decreases in BMI, mean arterial pressure, and haemoglobin (which can be negative health indicators in this age group, and are increasingly so with greater age beyond 73); and the fourth grouped increases in numbers of diseases with decreases in, haemoglobin, white cell count, and inflammation. The factors were basically independent, with all correlations being between .05 and .06. We labelled the factors Memory/Speed, Personality, Metabolism, and Physical Robustness. Full results are shown in Table 5.

### Deaths since Time 2

As noted, we observed 30 deaths between the Time 2 assessment and 31 December, 2011. It is not possible to make precise calculations of how much power we had to detect whether participants were in or entering periods of terminal decline, without indications of the expected effect sizes and lengths of periods over which terminal decline might operate. Based on the overall 2% per year death rate for this age group, however, we would expect about 200 deaths from the full sample over the 10-year period from ages 70 to 80. Given a not-uncommon estimate in the literature of periods of terminal decline on the order of 5-7 years, we would expect at least that number to have been in or entered such a period between ages 70 and 73. Based on this, we had over 80% power to detect changes on the order of .2 standard deviations, noticeably smaller than we could actually measure with any reliability.

We regressed death status as of 31 December, 2011 on the factor scores to assess their potential as markers of terminal decline. Together, the 4 variables explained 11.3% of variance, but the overall regression did not reach significance ($p$=.103). Entered singly, declines in Metabolism and Memory/Speed were each significant predictors of mortality, but Memory/Speed was not significant when entered with Metabolism. Their odds ratios were 1.58 (95% confidence interval 1.09-2.29 ) for Metabolism and 1.66 (95% confidence interval 1.10-2.54) for Memory/Speed when entered singly. At the same time, numbers of reliable change indicating declines in function at 75% reliability predicted death with an odds ratio of 1.27 (95% confidence interval 1.14-1.41, $p$<.001). But so did numbers of reliable change indicating improvement in function at 75% reliability (odds ratio = 1.13, 95% confidence interval 1.03-1.32, $p$=.031). This is summarised in Table 6.

**Table 2. Descriptive statistics of raw study variables**

| | All age 70 | | | | | Age 70, returning at 73 | | | | | Age 73 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | Min. | Max. | Mean | SD | N | Min. | Max. | Mean | SD | Min. | Max. | Mean | SD |
| Number of diseases | 1091 | 0 | 11 | 3.8 | 1.9 | 866 | 0 | 11 | 3.7 | 1.9 | 0 | 13 | 4.4 | 1.9 |
| Number of drugs taken | 1091 | 0 | 8 | 3.0 | 2.5 | 762 | 0 | 8 | 2.9 | 2.5 | 1 | 8 | 4.0 | 2.3 |
| HADS anxiety score | 1089 | 0 | 17 | 4.9 | 3.2 | 865 | 0 | 16 | 4.8 | 3.1 | 0 | 18 | 4.5 | 3.1 |
| HADS depression score | 1086 | 0 | 13 | 2.8 | 2.2 | 865 | 0 | 13 | 2.7 | 2.1 | 0 | 13 | 2.6 | 2.2 |
| BMI | 1089 | 16.02 | 72.00 | 27.82 | 4.56 | 866 | 16.02 | 48.52 | 27.80 | 4.37 | 16.67 | 48.50 | 27.92 | 4.43 |
| 6 metre walk time | 1085 | 2.0 | 11.0 | 3.85 | 1.13 | 860 | 2.0 | 11.0 | 3.78 | 1.05 | 2.30 | 12.30 | 4.34 | 1.24 |
| Demi-span in cm | 1088 | 65 | 91 | 77.8 | 4.8 | 864 | 65 | 91 | 78.1 | 4.8 | 66 | 91 | 78.2 | 4.7 |
| Activities of daily living | 1089 | 0 | 14 | 1.0 | 2.0 | 865 | 0 | 13 | .9 | 1.9 | 0 | 14 | 1.0 | 2.1 |
| Mean arterial pressure | 1088 | 72.00 | 145.00 | 104.15 | 11.87 | 866 | 72.00 | 145.00 | 104.02 | 11.88 | 63.78 | 140.00 | 101.64 | 11.33 |
| Forced expiratory vol. | 1085 | .74 | 4.34 | 2.36 | .68 | 856 | .74 | 4.34 | 2.41 | .68 | .78 | 4.25 | 2.30 | 0.67 |
| Average grip strength | 1086 | 5.00 | 55.50 | 28.03 | 9.93 | 865 | 5.00 | 55.50 | 28.49 | 9.78 | 5.00 | 53.25 | 28.11 | 9.26 |
| Logical memory | 1087 | 16 | 117 | 71.4 | 17.9 | 864 | 16 | 117 | 72.56 | 17.26 | 20 | 116 | 74.3 | 17.8 |
| Spatial span | 1084 | 5 | 24 | 14.7 | 2.8 | 861 | 6 | 24 | 14.8 | 2.8 | 7 | 23 | 14.7 | 2.8 |
| Verbal paired assoc. | 1050 | 0 | 40 | 26.4 | 9.1 | 843 | 1 | 40 | 26.9 | 9.0 | 0 | 16 | 9.1 | 3.8 |
| Symbol search | 1086 | 2 | 49 | 24.7 | 6.4 | 862 | 2 | 49 | 25.0 | 6.4 | 3 | 45 | 24.6 | 6.2 |
| Digit symbol | 1086 | 25 | 98 | 56.6 | 12.9 | 862 | 25 | 98 | 57.5 | 12.7 | 22 | 94 | 56.4 | 12.3 |
| Matrix reasoning | 1086 | 4 | 24 | 13.5 | 5.1 | 863 | 4 | 24 | 13.9 | 5.1 | 4 | 25 | 13.2 | 5.0 |
| Letter-number seq. | 1079 | 1 | 21 | 10.9 | 3.2 | 863 | 1 | 21 | 11.1 | 3.1 | 1 | 20 | 10.9 | 3.1 |
| Digit Span Backwards | 1090 | 2 | 14 | 7.7 | 2.3 | 866 | 2 | 14 | 7.8 | 2.3 | 2 | 14 | 7.8 | 2.3 |
| Block design | 1085 | 10 | 65 | 33.8 | 10.3 | 864 | 11 | 65 | 34.5 | 10.1 | 10 | 66 | 33.6 | 10.1 |
| Log simple RT | 1085 | .16 | .51 | .24 | .04 | 865 | .16 | .47 | .24 | .04 | .17 | .44 | .24 | .04 |
| Choice reaction time | 1084 | .45 | 1.13 | .64 | .09 | 865 | .45 | 1.13 | .64 | .08 | .46 | .97 | .65 | .09 |
| Inspection time | 1041 | 70 | 140 | 112.2 | 10.9 | 838 | 70 | 140 | 112.7 | 10.7 | 67 | 137 | 111.3 | 11.4 |
| Verbal fluency | 1087 | 10 | 83 | 42.4 | 12.5 | 865 | 10 | 83 | 42.8 | 12.7 | 8 | 85 | 43.2 | 12.9 |

*(Table 2 cont'd)*

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Natl adult reading test | 1089 | 10 | 50 | 34.5 | 8.1 | 864 | 10 | 50 | 34.8 | 8.1 | 9 | 50 | 34.4 | 8.1 |
| Wechsler T Adult Reading | 1089 | 14 | 50 | 41.0 | 7.2 | 864 | 14 | 50 | 41.3 | 7.1 | 16 | 50 | 41.0 | 6.9 |
| Mini-mental state exam | 1090 | 22 | 30 | 28.8 | 1.4 | 865 | 22 | 30 | 28.8 | 1.4 | 22 | 30 | 28.8 | 1.4 |
| Average corr. vision | 775 | -.10 | .80 | .10 | .16 | 607 | -.10 | .80 | .10 | .16 | -.10 | .90 | .15 | .17 |
| Haemoglobin | 1063 | 101 | 181 | 145.2 | 13.0 | 825 | 101 | 181 | 145.1 | 13.1 | 101 | 180 | 140.1 | 13.3 |
| White cell count | 1062 | 2.5 | 27.0 | 7.05 | 2.22 | 824 | 3.0 | 27.0 | 6.96 | 2.24 | 2.6 | 27.0 | 6.95 | 2.20 |
| Platelet count | 1061 | 105 | 508 | 274.1 | 64.6 | 820 | 105 | 508 | 271.1 | 61.8 | 73 | 460 | 245.0 | 58.2 |
| Prothrombin time | 1051 | 9 | 13 | 9.7 | .6 | 820 | 9 | 13 | 9.7 | .6 | 9 | 22 | 11.6 | 1.0 |
| APTT | 1051 | 21 | 41 | 28.6 | 3.1 | 820 | 21 | 41 | 28.6 | 3.1 | 21 | 50 | 31.0 | 3.9 |
| Fibrinogen | 1051 | 1.6 | 5.5 | 3.3 | .6 | 819 | 1.6 | 5.5 | 3.25 | .62 | 1.8 | 5.3 | 3.32 | 0.59 |
| Serum folate | 911 | 3.3 | 25.0 | 12.84 | 6.29 | 716 | 3.6 | 25.0 | 13.03 | 6.30 | 2.5 | 25.0 | 11.66 | 6.11 |
| Albumin | 1058 | 37 | 54 | 44.7 | 3.0 | 831 | 37 | 54 | 44.7 | 3.0 | 35 | 51 | 43.8 | 2.9 |
| Cholesterol | 1054 | 2.7 | 8.9 | 5.45 | 1.15 | 832 | 2.7 | 8.9 | 5.45 | 1.14 | 2.4 | 9.3 | 5.15 | 1.15 |
| HDL cholesterol | 969 | .53 | 3.32 | 1.52 | .44 | 832 | .68 | 3.32 | 1.53 | 0.44 | .57 | 3.00 | 1.46 | 0.43 |
| HbA1C | 1061 | 4.5 | 9.5 | 5.93 | .71 | 826 | 4.5 | 9.5 | 5.93 | 0.74 | 4.4 | 8.9 | 5.75 | 0.65 |
| C-reactive protein | 1053 | 1.5 | 45.0 | 5.19 | 6.03 | 830 | 1.5 | 45.0 | 4.98 | 5.69 | 1.5 | 45.0 | 4.80 | 5.80 |
| Glomerular filtration rate | 1060 | 41 | 160 | 81.4 | 18.2 | 833 | 41 | 160 | 81.5 | 18.1 | 30 | 157 | 77.8 | 19.2 |
| IPIP extraversion | 954 | 1 | 40 | 21.3 | 7.1 | 854 | 1 | 40 | 21.3 | 7.0 | 2 | 40 | 21.6 | 7.2 |
| IPIP agreeableness | 952 | 11 | 40 | 31.1 | 5.4 | 854 | 12 | 40 | 30.9 | 5.5 | 10 | 40 | 30.8 | 5.6 |
| IPIP conscientiousness | 952 | 9 | 40 | 28.2 | 6.0 | 854 | 9 | 40 | 28.1 | 6.1 | 5 | 40 | 27.7 | 6.1 |
| IPIP emotional stability | 950 | 1 | 40 | 24.6 | 7.7 | 853 | 1 | 40 | 24.9 | 7.7 | 2 | 40 | 25.0 | 7.7 |
| IPIP intellect | 948 | 5 | 40 | 23.8 | 5.7 | 852 | 5 | 40 | 23.9 | 5.7 | 5 | 40 | 23.7 | 5.9 |

**Table 3. Descriptive statistics of individual-level changes from ages 70 to 73**

| | Mean paired difference | Standard deviation | Standard error of mean | Skewness | Kurtosis | t | df | Prob. (2-tailed) | T1-T2 correlation |
|---|---|---|---|---|---|---|---|---|---|
| Number of diseases | .040 | .655 | .022 | .36 | .60 | 1.78 | 865 | .08 | .78 |
| Number of drugs Ttaken | -.096 | .684 | .025 | .35 | 2.29 | -3.88 | 761 | <.001* | .75 |
| HADS anxiety | .021 | .768 | .026 | .21 | 1.90 | .79 | 862 | .43 | .70 |
| HADS depression | .040 | .820 | .028 | .48 | 3.28 | 1.42 | 860 | .16 | .65 |
| BMI | .006 | .323 | .011 | -.45 | 2.71 | .52 | 864 | .60 | .95 |
| 6-Metre walk time | .510 | .875 | .030 | 1.49 | 9.59 | 17.05 | 856 | <.0001* | .63 |
| Demi-span | .062 | .962 | .033 | -.32 | 3.04 | 1.89 | 863 | .06 | .53 |
| Activities of daily living | .042 | .699 | .024 | .89 | 9.25 | 1.78 | 864 | .08 | .75 |
| Mean arterial pressure | .011 | 1.014 | .034 | -.30 | 1.46 | .31 | 865 | .76 | .49 |
| Forced expiratory vol. | -.152 | .514 | .018 | .29 | 5.60 | -8.63 | 852 | <.001* | .86 |
| Average grip strength | -.056 | .404 | .014 | -.44 | 2.22 | -4.07 | 863 | <.001* | .93 |
| Logical memory | -.062 | .758 | .026 | -.09 | .41 | -2.41 | 861 | .02 | .70 |
| Memory span | -.058 | .751 | .026 | .17 | .48 | -2.26 | 865 | .02 | .72 |
| Verbal paired assoc. | -.053 | .773 | .027 | .02 | .42 | -1.98 | 827 | .05 | .69 |
| Search speed | -.073 | .575 | .020 | -.58 | 3.48 | -3.75 | 862 | .000* | .83 |
| Matrix Reasoning | -.070 | .836 | .028 | -.05 | .26 | -2.45 | 861 | .01 | .65 |
| Block design | -.062 | .689 | .023 | -.15 | .84 | -2.64 | 860 | .01 | .76 |
| Verbal fluency | -.028 | .619 | .021 | .02 | .08 | -1.32 | 863 | .19 | .81 |
| Log simple reaction time | .070 | .906 | .031 | -.25 | 5.00 | 2.26 | 863 | .02 | .56 |
| Choice reaction time | .066 | .682 | .023 | -.30 | 4.42 | 2.86 | 863 | .00 | .76 |
| Inspection time | -.053 | .895 | .031 | -.20 | 3.00 | -1.69 | 821 | .09 | .59 |
| Word reading | -.044 | .282 | .010 | -.32 | 1.13 | -4.58 | 863 | <.001* | .96 |
| Mini-mental state exam | -.035 | 1.015 | .035 | .16 | 1.58 | -1.03 | 864 | .30 | .47 |
| Average corr. vision | .054 | .148 | .006 | -.11 | 1.38 | 8.61 | 551 | <.001* | .58 |

*(Table 3 cont'd)*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| White cell count | -.046 | .629 | .026 | .06 | 2.85 | -2.07 | 810 | .04 | .73 |
| Haemoglobin | -.063 | .711 | .025 | -.02 | 2.28 | -2.52 | 812 | .01 | .76 |
| Clot | -.062 | .531 | .019 | -.03 | .29 | -3.23 | 754 | .001 | .50 |
| Nutritive status | -.461 | .896 | .031 | .29 | .21 | -14.70 | 816 | <.001* | .57 |
| Inflammation | -.348 | .917 | .032 | .09 | 5.72 | -10.90 | 823 | <.001* | .52 |
| Metabolic vascular risk | -.569 | .709 | .025 | -.29 | 4.61 | -23.01 | 822 | <.001* | .74 |
| Glomerular filtration rate | -.004 | .579 | .020 | -.08 | 1.19 | -.21 | 817 | .83 | .83 |
| IPIP extraversion | .005 | .576 | .021 | -.02 | .31 | .22 | 776 | .82 | .84 |
| IPIP agreeableness | .039 | .748 | .027 | -.47 | 2.01 | 1.46 | 775 | .14 | .72 |
| IPIP conscientiousness | .057 | .676 | .024 | .03 | .33 | 2.36 | 774 | .02 | .77 |
| IPIP emotional stability | -.030 | .701 | .025 | -.12 | .81 | 1.20 | 771 | .23 | .76 |
| IPIP intellect | .017 | .699 | .025 | -.22 | .86 | .67 | 771 | .51 | .76 |

*Note. Variables were standardised to Time 1 level, so the mean difference was effect size relative to that level. With adjustment for multiple testing, only probability levels of .001 or less should be considered significant (*). Change was Time 2 less Time 1, so negative differences indicate declines in scores.*

**Table 4. Observations regarding numbers of reliable changes**

| | 85% Reliability | 75% Reliability |
|---|---|---|
| Number of reliable changes - mean (sd) | 10.4(6.0) | 7.9(4.8) |
| Correlation between reliable improvements and declines | .53 | .52 |
| Correlation of number of reliable changes with: | | |
| age-11 IQ | ns | ns |
| age-70 IQ | ns | ns |
| years of education | ns | ns |
| social class | ns | ns |
| number of drugs taken | --- | .12 |
| BMI | --- | -.08 |
| 6-metre walk time | --- | .24 |
| activities of daily living | --- | .10 |
| mean arterial pressure | --- | -.08 |
| forced expiratory volume | | -.07 |
| logical memory | --- | -.13 |
| search speed | --- | -.13 |
| simple reaction time | --- | .07 |
| inspection time | --- | -.08 |
| haemoglobin | --- | -.10 |
| metabolic vascular risk | --- | -.11 |

*Note. Only significant (with no adjustment for multiple testing) correlations with change are shown. 'ns' is 'not significant.' '---' is 'not calculated.' Number of reliable changes refers to number of changes per person that were reliable among 36 variables assessed at about ages 70 and 73. 85/75% reliability refer to assumed test-retest reliability of the measures. See text for further explanation.*

**Table 5. Factor analysis of change variables**

| | Memory/ Speed | Personality | Metabolism | Physical robustness |
|---|---|---|---|---|
| Number of diseases | -.02 | .05 | **.18** | **.17** |
| Number of drugs taken | -.08 | .05 | **.32** | .06 |
| HADS anxiety | -.06 | **.33** | .10 | .00 |
| HADS depression | -.03 | **.32** | .07 | .04 |
| BMI | .03 | -.07 | **-.37** | .08 |
| 6-metre walk time | -.14 | .03 | .06 | .02 |
| Demi-span | -.02 | -.01 | -.08 | .02 |
| Activities of daily living | -.05 | .01 | .04 | .08 |
| Mean arterial pressure | -.06 | .02 | **-.65** | -.01 |
| Forced expiratory vol. | -.01 | -.04 | .07 | .04 |
| Average grip strength | .08 | -.01 | -.13 | -.08 |
| Logical memory | **.57** | .01 | -.01 | .01 |
| Memory span | **.18** | .00 | -.06 | .01 |
| Verbal paired assoc. | **.40** | .04 | -.02 | .09 |
| Search speed | .14 | -.03 | -.07 | .07 |
| Matrix reasoning | .06 | .04 | -.01 | -.01 |
| Block design | .13 | -.02 | .09 | -.04 |
| Verbal fluency | **.21** | -.03 | .03 | .07 |
| Simple reaction time | **-.33** | .02 | .00 | .04 |
| Choice reaction time | **-.39** | .11 | -.01 | .01 |
| Inspection time | .07 | .06 | -.04 | -.05 |
| Word reading | **.18** | .05 | .02 | .02 |
| Mini-mental state exam | **.18** | .02 | .05 | .07 |
| Average corr. vision | .03 | -.04 | -.07 | -.01 |
| White cell count | .01 | .05 | -.01 | **-.67** |
| Haemoglobin | .01 | .05 | **-.27** | **-.19** |
| Clot | -.07 | .03 | .03 | -.01 |
| Nutritive status | .00 | .02 | .14 | -.11 |
| Inflammation | -.04 | -.06 | .04 | **-.49** |
| Metabolic vascular risk | .05 | -.10 | -.10 | -.10 |
| Glomerular filtration rate | .03 | .06 | .14 | .14 |
| IPIP extraversion | -.09 | **-.37** | -.01 | -.01 |
| IPIP agreeableness | .05 | **-.49** | -.04 | -.04 |
| IPIP conscientiousness | -.02 | **-.33** | .04 | .04 |
| IPIP emotional stability | -.05 | **-.34** | .09 | .10 |
| IPIP intellect | -.02 | **-.38** | .03 | .03 |

*Note. Factor loadings greater than .15 in absolute value are in bold.*

**Table 6. Change variables as individual predictors of death since Time 2**

| | Odds ratio | 95% Confidence interval |
|---|---|---|
| Decline in memory/speed | 1.67 | 1.10-2.54 |
| Decline in personality traits | ns | --- |
| Decline in metabolism | 1.58 | 1.09-2.29 |
| Decline in physical robustness | ns | --- |
| Number of reliable declines | 1.27 | 1.14-1.41 |
| Number of reliable improvements | 1.13 | 1.03-1.32 |

*Note. 'ns' is 'not significant.' Logistic regression including all factors was not significant.*

## Discussion

In this study, we explored the potential capacity to use two longitudinal data waves to distinguish 'normal' from disadvantageous and even terminal ageing patterns in the LBC1936 between ages 70 and 73. To do this, we examined 36 variables indicating aspects of cognitive, emotional, and physical function. In the process, we addressed 3 questions: 1) How and to what extent did individuals change during this period? 2) Were there correlates or predictors of these changes? and 3) Did changes tend to cluster in ways that could distinguish healthy ageing from terminal decline? Overall, our measures tended to show mean changes indicating declines in function, as would be expected. Most of these mean differences were not significant after adjustment for multiple testing, however, despite our good-sized sample. The lack of statistical significance of most of the mean changes indicated the gradual nature of the overall ageing process, especially since some of the significant mean differences indicated improvement in average function, including number of drugs taken, average corrected vision, and inflammation.

Within each variable, we observed substantial individual differences in change, but most could not be considered reliable. Moreover, the changes that could be considered reliable were at least as likely to indicate improvements in function as declines and changes indicating improvements and declines in function were substantially correlated. This suggests that ageing is far from a uniform process, but it also suggests that increasing variability in 'measurability' may be an important indicator of its progress, a topic receiving increasing attention in the ageing literature (e.g., Ram et al., 2011). Of course practice effects on some variables could have accounted for improvements as well. In sum, we observed substantial change, in the aggregate indicating ageing, but most of the individual observations could not be considered reliable. The lack of reliability of measures of change based on two waves of data is well known, and certainly results from the inability to distinguish error of measurement of individual level from error of measurement of individual change. It is rare, however, to see it so clearly documented as was possible here, given the large number of variables available. Even if more measurement occasions made it possible to minimise error of measurement, it is possible that a proportion of individual variation in ageing trajectories reflects a random walk process, wherein differences between adjacent measurement occasions are at least partially completely random. Such random walk processes are common throughout nature and society, and their presence and importance in understanding developmental progressions is increasingly being recognised, for example, in economics (Kuljanin, Braun, & DeShon, 2011).

### Study limitations

Before discussing our observations in further detail, we note the primary limitations of our study. The most important of these was the relatively select nature of our sample, which was of somewhat higher childhood mental ability than the overall population. The average age-11 IQ score in the LBC1936 was 0.78 standard deviation higher

than the overall average for the full Scottish Mental Survey 1947, and the variance was restricted by 44%. In general, range restriction tends to reduce the magnitudes of associations involving the variable on which range has been restricted, but this is not always the case. Any such reduction would have been small in this case (.01-.03 at most), due to the small magnitudes of the correlations observed. The sample selectivity at least partly reflected general mortality patterns, as IQ is associated with greater longevity (Batty, Deary, & Gottfredson, 2007; Calvin et al., 2011). Our sample was likely of higher educational attainment and social class than the overall population as well, though we could not quantify the degree to which this was the case. We did not have short-term test-retest reliability data for most of our measures, and thus were forced to make assumptions about their likely values. We did this somewhat crudely, making two overall assumptions for all variables. Within this, however, our conclusions were very similar for the two levels assumed.

### Correlates or predictors of change and clustering of change variables

We found no associations between number of reliable changes and age-11 IQ, age-70 IQ, number of years of education, or current social class, suggesting that change was relatively evenly distributed throughout the sample. Personality may have contributed to lower reliability of measurement, as lower IPIP Conscientiousness and Agreeableness at age 70 had the largest negative correlations with number of reliable changes. This question deserves greater research attention. To the extent we were able to pick up leading indicators of ageing, they appeared to be increases in numbers of drugs, six-metre walk time, ADLs, and clot function, and decreases in BMI, mean arterial pressure, FEV, Logical Memory, Search Speed, haemoglobin, and metabolic vascular risk.

One intriguing observation was that the individual age-70 variable with the strongest correlation with number of reliable changes was IPIP Conscientiousness, with IPIP Agreeableness second. Both correlations were negative (-.15 and -.12, respectively), indicating that those with lower Conscientiousness and Agreeableness at age 70 tended to show larger changes that could be considered statistically reliable. At the same time, however, given the variables in question, it seems likely that those changes, though large enough to be considered reliable statistically, were in fact not particularly reliable at all. That is, participants scoring lower in Conscientiousness and/or Agreeableness may have used less care in completing all measures over which they had some overt control, and may have varied more even on the physiological measures due to less routine in daily dietary, sleep, and other habits. This should be pursued in future research.

Evidence of clustering among the change variables was weak. The average absolute value of correlation was less than .05, and most variables did not load on any of the four factors that might reasonably be considered substantial within the data. In general, however, the variables that did load on these factors clustered around constructs of memory and attention, personality, perhaps appetite status, and overall health. These factors basically reflected the possible leading-indicator variables just noted. Indicators of failing overall health, particularly failing appetite in the form of increases in BMI, mean arterial pressure, glomerular filtration rate, and total number of reliable declines in function appeared to be the strongest candidates as markers of terminal decline, in the sense that they predicted rather imminent death. Many would consider these to be obvious, and effects were sufficiently weak that recovery from any particular state was clearly possible.

## Conclusions

The limitations of two waves of longitudinal data to explore change have been well documented, but studies claiming to have uncovered important associations with change based on only two waves continue to be published regularly. We undertook this study both to provide an empirical demonstration of the degree of uncertainty of such estimates of change and to explore whether such change estimates might aggregate in more meaningful ways. Within levels of short-term test-retest reliability considered acceptable in the field of psychology, this study showed that most change observed over a 3-year period was not reliable. Moreover, levels of correlation considered by psychologists to indicate substantial stability actually allow a level of movement, whether random or systematic, that many psychologists will likely find surprising. Even change that could be considered reliable often represented improvement (some of which might

have been test familiarity) rather than decline in function in this ageing sample in which overall decline might be expected. This suggests strongly that ageing proceeds in fits and starts, and it is possible that individual trajectories of any one variable include a considerable component best described as a random walk. Though it increases the assessment burden considerably, we urge researchers planning longitudinal projects to anticipate the need for more than two waves of data in order to draw meaningful conclusions about change, and we strongly suggest that journals publishing two-wave studies require inclusion of information about the extent to which the change measures could be considered reliable.

## Acknowledgements

We thank the LBC1936 participants; Caroline Brett, Alison Pattie, Michelle Taylor and Caroline Cameron for data collection; the LBC1936 Study Secretary, Paula Davies; and the nurses and other staff at the Wellcome Trust Clinical Research Facility, Edinburgh (http://www.wtcrf.ed.ac.uk). The LBC1936 sample was funded by Research Into Ageing (Wave 1), and continues to be funded by Age UK as part of the Disconnected Mind project (Wave 2). Wendy Johnson, Alan Gow, John Starr, and Ian Deary are members of the University of Edinburgh Centre for Cognitive Ageing and Cognitive Epidemiology (http://www.ccace.ed.ac.uk), which is supported by the BBSRC, EPSRC, ESRC, and MRC as part of the cross-council Lifelong Health and Wellbeing Initiative (Grant #G0700704/84698).

## References

Adler, N. E., & Snibbe, A. C. (2003). The role of psychosocial processes in explaining the gradient between socioeconomic status and health. *Current Directions in Psychological Science,* 12, 119-123.

Backman, L., & MacDonald, S. W. (2006). Death and cognition: Synthesis and outlook. *European Psychologist,* 11, 224-235.

Batty, G. D., Deary, I. J., & Gottfredson, L. S. (2007). Premorbid (early life) IQ and later mortality risk: Systematic review. *Annals of Epidemiology,* 17, 278-288.

Batty, G. D., Deary, I. J. & Macintyre S. (2007). Childhood IQ in relation to physiological and behavioural risk factors for premature mortality in middle-aged persons: the Aberdeen Children of the 1950s Study. Journal of Epidemiology and Community Health, 61, 241-247.

Batty G.D., Deary, I.J., Schoon, I., & Gale, C. R. (2007). Childhood mental ability in relation to food intake and physical activity in adulthood: the 1970 British Cohort Study. Pediatrics, 119, e38-e45.

Calvin, C. M., Deary, I, J., Fenton, C., Roberts, B., Der, G., Leckenby, N., & Batty, G. D. (2011). Intelligence in youth and all-cause mortality: systematic review with meta-analysis. *International Journal of Epidemiology*, 40, 626-644.

Campbell, D. T., & Kenny, D. A. (1999). *A Primer on Regression Artifacts.* New York: Guilford Press.

Christensen, H., Mackinnon, A. J., Korten, A., & Jorm, A. F. (2001). The 'common cause' hypothesis of cognitive ageing: Evidence not only a common factor but also specific associations of age with vision and grip strength in a cross-sectional analysis. *Psychology and Aging,* 16, 588-599.

Christensen, L., & Mendoza, J. L. (1986). A method of assessing change in a single subject: an alteration of the RC index. *Behavior Therapy,* 17, 305-308.

Cosway, R., Strachan, M. W., Dougall, A., Frier, B. M., & Deary, I. J. (2001). Cognitive function and information processing in Type 2 diabetes. *Diabetes Medicine,* 18, 803-810.

Cronbach, L. J., & Furby, L. (1970). How should we measure change? Or should we? *Psychological Bulletin,* 74, 68-80.

Deary, I. J., Gale, C. R., Stewart, M. C. W., Fowkes, F. G. R., Murray, G. D., Batty, G. D.,& Price, J. F. (2009). Intelligence and persisting with medication for two years: analysis in a randomised controlled trial. *Intelligence*, 37, 607-612.

Deary. I. J., Gow, A. J., Pattie, A., & Starr, J.M. (in press). Cohort profile: The Lothian Birth Cohorts of 1921 and 1936. *International Journal of Epidemiology*.

Deary, I. J., Gow, A. J., Taylor, M. D., Corley, J., Brett, C., Wilson, V., Campbell, C., Whalley, L. J., Visscher, P. M., Porteus, D. J. & Starr, J. M. (2007). The Lothian birth Cohort 1936: A study to examine influences on cognitive ageing from age 11 to age 70 and beyond. *BMC Geriatrics,* 7, doi:10.1186/1471-2318-7-28.

Deary, I. J., Johnson, W., Gow, A. J., Pattie, A., Brett, C. E., Bates, T. C., & Starr, J. M. (2011). Losing one's grip: A bivariate growth curve model of grip strength and nonverbal reasoning from age 79 to 87 years in the Lothian Birth Cohort 1921. *Journals of Gerontology, Series B: Psychological Sciences,* 66, 699-707.

Deary, I. J., Weiss, A., & Batty, G. D. (2010). Intelligence and personality as predictors of illness and death: How researchers in differential psychology and chronic disease epidemiology are collaborating to understand and address health inequalities. *Psychological Science in the Public Interest*, 11, 53-79.

Deary, I. J., Whiteman, M. C., Starr, J. M., Whalley, L. J. & Fox, H. C. (2004). The impact of childhood intelligence on later life: Following up the Scottish Mental Surveys of 1932 and 1947. *Journal of Personality and Social Psychology,* 86, 130-147.

Dixon, R. A. (2011). Enduring theoretical themes in psychological aging: Derivations, functions, perspectives, and opportunities. In K. W.Schaie amd S. L. Willis (eds.), *Handbook of the Psychology of Aging, 7th Ed.* (Pp 3-23). San Diego, CA: Academic Press.

Dolcos, S., MacDonald, S. W., Braslavsky, A., Camicioli, R., & Dixon, R. A. (2012). Mild cognitive impairment is associated with selected functional markers: Integrating concurrent, longitudinal, and stability effects. *Neuropsychology,* 26, doi:10.1037/a0026760.

Finkel, D., Reynolds, C. A., McCardle, J. J., & Pedersen, N. L. (2005). The longitudinal relationship between processing speed and cognitive ability: Genetic and environmental influences. *Behavior Genetics,* 35, 535-549.

Fiske, A., Wetherell, J. L., & Gatz, M. (2009). Depression in older adults. *Annual Review of Clinical Psychology,* 5, 363-389.

Gale, C. R., Batty, G. D., Cooper, C., & Deary, I. J. (2009). Psychomotor coordination and intelligence in childhood and heatlh in adulthood: Testing the 'system integrity' hypothesis. *Psychosomatic Medicine,* 71, 675-681.

Gallo, L. C., & Matthews, K. A. (2003). Understanding the association between socioeconomic status and physical health: Do negative emotions play a role. *Psychological Bulletin,* 129, 10-51.

Gayman, M. D., Turner, R. J., & Cui, M. (2008). Physical limitations and depressive symptoms: Exploring the nature of the association. *Journals of Gerontology Series B-Psychological Sciences and Social Sciences,* 63, S219-S228.

Gerstorf, D., Rocke, C., & Lachman, M. E. (2011). Antecedent-Consequent Relations of Perceived Control to Health and Social Support: Longitudinal Evidence for Between-Domain Associations Across Adulthood. *Journals of Gerontology Series B-Psychological Sciences and Social Sciences,* 66, 61-71.

Hanson, L. L. M., Akerstedt, T., Naswall, K., Leineweber, C., Theorell, T., & Westerlund, H. (2011). Cross-Lagged Relationships Between Workplace Demands, Control, Support, and Sleep Problems. *Sleep,* 34, 1403-U147.

Hart, C. L., Taylor, M. D., Davey-Smith, G., Whalley, L. J., Starr, J. M., Hole, D. J., & Deary, I. J. (2003). Childhood IQ, social class, deprivation, and their relationships with mortality and morbidity in later life: Prospective observational study linking the Scottish Mental Survey 1932 and the Midspan studies. *Psychosomatic Medicine,* 65, 877-883.

Hassing, L. B., Grant, M. D., Hofer, S. M., Pedersen, N. L., Nilsson, S. E., & Berg, S. (2004). Type 2 diabetes mellitus contributes to cognitive decline in old age: A longitudinal, population-based study. *Journal of the International Neurological Society,* 10, 599-607.

Hsu, L. M. (1989). Reliable changes in psychotherapy: Taking into account regression toward the mean. *Behavioral Assessment,* 11, 459-467.

Jarvik, L. F., & Falek, A. (1963). Intellectual stability and survival in the aged. *Journal of Gerontology,* 18, 173-176.

Johnson, W., Deary, I. J., McGue, M., & Christensen, K. (2009). Genetic and environmental transactions linking cognitive ability, physical fitness, and education lin late life. *Psychology and Aging,* 24, 48-62.

Kendler, K. S., Gardner, C., Fiske, A., & Gatz, M. (2009). Major depression and coronary artery disease in the Swedish Twin Registry: Phenotypic, genetic, and environmental sources of comorbidity. *Archives of General Psychiatry,* 66, 857-863.

Kuhlen, R. G. (1940). Social change: A neglected factor in psychological studies of the life span. *School and Society,* 52, 14-16.

King, L. A., King, D. W., McArdle, J. J., Saxe, G. N., Doron-LaMarca, S., & Orazem, R. J. (2006). Latent difference score approach to longitudinal traums research. *Journal of Traumatic Stress,* 19, 771-785.

Kooij, D., & Van De Voorde. (2011). How changes in subjective general health predict future time perspective, and development and generativity motives over the lifespan. *Journal of Occupational and Organizational Psychology,* 84, 228-247.

Kuljanin, G., Braun, M. T. & DeShon, R. P. (2011). A cautionary note on modeling growth trends in longitudinal data. *Psychological Methods,* 16, 249-264.

Lapi, F., Pozzi, C., Mazzaglia, G., Ungar, A., Fumagalli, S., Marchionni, N., Geppetti, P., Mugelli, A., & Di Bari, M. (2009). Epidemiology of Suboptimal Prescribing in Older, Community Dwellers A Two-Wave, Population-Based Survey in Dicomano, Italy. *Drugs & Aging*, 26, 1029-1038.

Li, K. Z., & Lindenberger, U. (2002). Relations between ageing sensory/sensorimotor and cognitive functions. *Neuroscience Biobehavioural Review,* 26, 777-783.
Lieberman, M. A. (1966). Observations on death and dying. *Journal of Gerontology,* 6, 70-72.
Mather, K. A., Jorm, A. F., Anstey, K. J., Milburn, P. J., Eastel, S., & Christensen, H. (2010). Cognitive performance and leukocyte telomere length in two narrow age-range cohorts: a population study. *BMC Geriatrics*, 10, 62.
McCardle, J. J., & Woodcock, R. W. (1997). Expanding test-retest designs to include developmental time-lag components. *Psychological Methods,* 2, 403-435.
McGue, M., & Christensen, K. (2002). The heritability of level and rate-of-change in cognitive function in Danish twins age 70 years and older. *Experimental Aging Research,* 28, 435-451.
Menezes Costa, L. D., Maher, C. G., McAuley, J. H., Hancock, M. J., & Smeets, R. J. E. M. (2011). Self-efficacy is more important than fear of movement in mediating the relationship between pain and disability in chronic low back pain. *European Journal of Pain,* 15, 213-219.
Nesselroade, J. R., & Cable, D. E. (1974). "Sometimes it's okay to factor difference scores" -- The separation of state and trait anxiety. *Multivariate Behavioral Research,* 9, 273-284.
Rabbitt, P., Lunn, M., Pendleton, N., & Yardafagar, G. (2011). Terminal pathologies affect rates fo decline to different extents and age accelerates the effects of terminal pathology on cognitive decline. *Journals of Gerontology Series B - Psychological Sciences and Social Sciences,* 66, 325-334.
Rafnsson, S. B., Deary, I. J., Smith, F. B., Whiteman, M. C., Rumley, A., & Lowe, G. D. (2007). Cognitive decline and markers of inflammation and hemostasis: The Edinburgh Artery Study. *Journal of the American Geriatric Society,* 55, 700-707.
Ram, N., Gerstorf, D., Lindenberger, U., & Smith, J. (2011). Developmental change and intraindividual variability: Relating cognitive aging to cognitive plasticity, cardiovascular lability, and emotional diversity. *Psychology and Aging*, 26, 363-371.
Ramsden, S., Richardson, F. M., Josse, G., Thomas, M. S. C., Ellis, C., Shakesshaft, C., Seghier, M. L. & Price, C. J. (2011). Verbal and non-verbal intelligence changes in the teenage brain. *Nature***,** 479, 113-116.
Riegel, K. F., & Riegel, R. M. (1972). Development, drop, and death. *Developmental Psychology,* 6, 306-319.
Rogosa, D. (1995). Myths and methods: "Myths about longitudinal research" plus supplemental questions. In J. L. Gottman(Ed.), *The Analysis of Change* (pp 1-55). Mahwah, NJ: Lawrence Erlbaum Associates.
Rogosa, D., Brandt, D., & Zimoski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin,* 92, 726-748.
Schaie, K. W. (1965). A general model for the study of developmental problems. *Psychological Bulletin,* 64, 92-107.
Schelleman-Offermans, K., Kuntsche, E., & Knibbe, R. A. (2011). Associations between drinking motives and changes in adolescents' alcohol consumption: a full cross-lagged panel study. (2011) *Addiction,* 106, 1270-1278.
Schram, M. T., Euser, S. M., de Craen, A. J., Witteman, J. C., Frolich, M., & Hofman, A. (2007). Systemic markers of inflammation and cognitive decline in old age. *Journal of the American Geriatric Society,* 55, 708-716.
Scottish Council for Research in Education. (1949). *The Trend of Scottish Intelligence: A Comparison of the 1947 and 1932 Surveys of the intelligence of Eleven-Year-Old Pupils.* London:University of London Press.
Sliwinski, M. J., Stawski, R. S., Hall, C. B., Katz, M., Verghese, J., & Lipton, R. (2006). Distinguishing preterminal and terminal cognitive decline. *European Psychologist,* 11, 172-181.
Sliwinski, M., Hoftman, L., & Hofer, S. M. (2010). Evaluating convergence of within-person change and between-person age differences in age-heterogeneous longitudinal studies. *Research in Human Development,* 7, 45-60.
Terrera, G. M., van den Hout, A., & Matthews, F. E. (2011). Random change point models: Investigating cognitive decline in the presence of missing data. *Journal of Applied Statistics,* 38, 705-716.
Townsend, P. (1979). *Poverty in the United Kingdom: A Survey of Household Resources and Standards of Living.* Harmondsworth: Penguin.
Whitehead, B. P., Dixon, R. A., Hultsch, D. F., & MacDonald, S. W. S. (2011). Are neurocognitive speed and inconsistency similarly affected in type 2 diabetes? *Journal of Clinical and Experimental Neuropsychology,* 33, 647-657.
Wilson, R. S., Beckett, L. A., Bienias, J. L., Evans, D. A., & Bennett, D. A. (2003). Terminal decline in cognitve function. *Neurology,* 60, 1782-1787.
Zigmond, A. S., & Snaith, R. P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavia,* 67, 361-370.