Title: Are 'healthy cohorts' real-world relevant? Comparing the National Child Development Study (NCDS) with the ONS Longitudinal Study (LS)

Keywords: census, representativeness, attrition, loss to follow-up, cohort study

Dr Gemma Archer[1,2], Ms Wei W Xun[1], Ms Rachel Stuchbury[1], Dr Owen Nicholas[3], Professor Nicola Shelton[1]

1. Department of Epidemiology and Public Health Care, UCL, London, UK
2. Department of Psychological Medicine, KCL, London, UK
3. Department of Statistical Science, UCL, London, UK

Corresponding author: Gemma Archer

gemma.archer@kcl.ac.uk

Word count: 3711 (excluding abstract, tables, and references); 5492 (including abstract, tables, and references).

Key Messages (3-4): 100 characters

- This study compared the NCDS with the ONS LS – a 'gold-standard' national reference population
- Participants from the most recent NCDS 55-year survey were mostly unrepresentative of age-matched LS respondents
- Despite differences in sample characteristics, longitudinal associations were similar in the NCDS and LS samples

Funding details:

This work was supported by the ESRC ref ES/R00823X/1 under Grant number ES/R00823X/1.

Data availability statement:

The authors take responsibility for the integrity of the data and the accuracy of the analysis. NCDS data can be accessed through the UK Data Service. ONS LS data can be accessed through the Centre for Longitudinal Study Information and User Support (CeLSIUS) at UCL, London, UK.

Conflict of interest statement:

The authors declare that there is no conflict of interest

Acknowledgements:

The permission of the Office for National Statistics to use the Longitudinal Study is gratefully acknowledged as is the support of CeLSIUS, the Centre for Longitudinal Study Information and User Support. This work contains statistical data which is Crown Copyright. The use of the ONS statistical data does not imply the endorsement of the ONS in relation to its interpretation or analysis. The authors alone are responsible for the interpretation of the data. This work uses research datasets which may not exactly reproduce National Statistics aggregates. The authors are also grateful to Professor Jenny Head for statistical advice.

ABSTRACT

Comparisons between cohort studies and nationally representative 'real-world' samples are limited. The NCDS (1958 British birth cohort) follows those born in Britain in a single week in March 1958 (n=18,558); and the ONS Longitudinal Study (LS) contains linked census data and life events for a 1% sample of the population of England and Wales (> 1 million records; allowing for sub-samples by age, ethnicity, or other socio-demographic factors). Common country and age-matched socio-demographic variables were extracted from the closest corresponding time-points, NCDS 55-year survey in 2013 (n=8107) and LS respondents aged 55 in 2011 (n=7052). Longitudinal associations between socio-demographic exposures (from the NCDS 46-survey in 2003 and LS respondents aged 45 in 2001) and long-term limiting illness (from NCDS 2013 and LS respondents 2011, aged 55) were assessed using logistic regression. The NCDS 55-year sample had similar characteristics to LS respondents aged 55 for sex and marital status, but the NCDS sample had lower levels of long-term limiting illness (19.7% vs 22.8%), non-white ethnicity (2.1% vs 11.7%) and living in South England (46.9% vs 50.1%), and higher levels of full-time employment (61.2% vs 55.2%), working in professional/higher managerial occupations (35.7% vs 29.2%), and living with a spouse (69.1% vs 64.9%), all $p<0.001$. Nevertheless, longitudinal associations between socio-demographic exposures and long-term limiting illness were similar in the NCDS and LS samples (all tests of between-study heterogeneity in mutually adjusted models $p>0.09$) suggesting these NCDS findings are largely generalisable to the population of England and Wales.

Introduction

Cohort studies are frequently used to assess how exposure to different social, demographic and economic factors impact later health outcomes. Over time however; cohort studies are vulnerable to attrition bias e.g. through emigration, inability to trace participants, and withdrawal, where those in lower socio-economic positions and in poorer health become increasingly under-represented (Atherton et al., 2008; Ferrie et al., 2009; Stafford et al., 2013) – known as the 'healthy cohort effect'. Moreover, increasing levels of global mobility mean that studies spanning several decades may be no longer be ethnically representative. For example, in the UK, the foreign-born population has tripled from 4.2% in 1951 (2.1 million) to 13.0% in 2011 (7.5 million) (Jefferies, 2005; ONS, 2002). Cohort attrition and immigration are potentially problematic as the quality of evidence for public health applications is largely dependent on how well samples reflect the 'real-world' population (Rothman et al., 2013).

At its inception the National Childhood Development Study (NCDS; also known as the 1958 British birth cohort) was a nationally representative sample including all babies born in England, Scotland and Wales in one week in March 1958. The cohort has since been followed-up ten times with a new data collection currently underway, but almost half the original sample has now been lost through death (8.9%), emigration (6.9%), and refusal or ineligibility (34.9%). The representativeness of the NCDS has been examined at several sweeps (Atherton et al., 2008; Hawkes and Plewis, 2006; Nathan, 1999; Plewis et al., 2004), most recently at the 45-year survey in 2002 (Atherton et al., 2008); however, these studies have tended to focus on attrition i.e. comparing various NCDS samples to the original baseline survey, with little (Atherton et al., 2008) or no (Hawkes and Plewis, 2006; Nathan, 1999; Plewis et al., 2004) comparison to nationally representative data. Likewise, to the authors' knowledge, no study has examined whether exposure-outcome associations in the NCDS, or other long-running birth cohort studies, are comparable to associations in nationally representative census data. It is plausible that exposure-outcome associations in the NCDS will be unbiased given that there is an array of missing data mechanisms in which missingness depends jointly on outcome and exposure(s) but for which the effect estimate is still estimated without bias (Bartlett et al., 2015). Likewise, other studies have demonstrated that exposure-outcome associations are comparable to the source

populations in e.g. occupational (Batty et al., 2014), simulated (Pizzi et al., 2011) and web-based cohorts (Pizzi et al., 2012).

This study aims to examine to what extent the NCDS sample remains representative of, or its findings generalisable to, the national population. Key socio-demographic characteristics and their associations with a general health outcome will be compared in the NCDS and the Office for National Statistics Longitudinal Study (LS); the LS contains linked census and vital events data for England and Wales, which is not reliant on voluntary surveys and therefore represents a 'gold-standard' national reference population. First, we will compare the prevalence of key socio-demographic factors and longstanding limiting illness in the most recent NCDS survey with an equivalent LS sample. Second, we will assess to what extent longitudinal associations between socio-demographic factors and long-term limiting illness in the NCDS are comparable to associations in the LS.

METHODS

Data

The National Child Development Study

The National Child Development Study (NCDS; the 1958 British birth Cohort study) includes all children born in England, Scotland and Wales during one week in March 1958 (n=17,638) and 920 immigrants with the same birth week recruited up to age 16 (Power and Elliott, 2006) (n=18,558). Of 11,553 invited, 9137 participated in the 55 year survey – 49% of all those ever enrolled; study attrition occurred due to death (n=1659), emigration (n=1286), and permanent refusal or ineligibility (e.g. uncontactable, unproductive interview) (n=6476). From 2000, ethical approval was given by the London Multi-centre Research Ethics Committee; informed consent was obtained from participants at various sweeps.

The ONS Longitudinal Study

The ONS Longitudinal Study (LS) contains census and life events data (e.g. births, deaths, cancer registrations) for an approximate 1% sample of the population of England and Wales. Records have been linked across five successive decennial censuses, beginning in 1971, for all those born on one of four selected dates in a calendar year. The sample is updated at each census, most recently in 2011, with new LS members entering the study through birth and immigration. The LS includes records for over 1.2 million individuals and with a 94% response rate for the 2011 Census, is considered nationally representative. Over 580,000 individuals were enumerated in the 2011 sample (Shelton et al., 2018).

Study variables

Study variables were obtained from LS in 2011 and NCDS in 2013 (55-year survey), and LS in 2001 and NCDS in 2004 (46-year survey), which were the closest corresponding time-points (see figure 1). Variables were selected if they were available in both studies, and had identical, or near identical wording (see appendix 1 for exact wording of questions).

LS data was obtained from census questionnaires in 2001 and 2011, and NCDS data was obtained by computer assisted telephone interviewing (CATI) in 2004, and by CATI and computer assisted web interviewing (CAWI) in 2013.

FIGURE 1 HERE

Figure 1. Timing of NCDS and LS data collections (NCDS sample restricted to those resident in England and Wales; LS sample includes those aged 55 in 2011)

Socio-demographic factors

Key socio-economic factors were identified as those known to be associated with health status (Kuh and Ben-Shlomo, 2004; Lantz et al., 1998). Socio-demographic factors common to the NCDS and LS surveys included sex, ethnicity (five categories collapsed into 'white' and 'non-white', due to low proportion of non-white in NCDS); region ('South': South West, South East, East of England, East Midlands, West Midlands; 'North': North East, North West, Yorkshire and Humberside; 'Wales'; and 'Scotland'); socio-economic classification (defined by the National Statistics Socio-economic Classification (NS-SEC) and collapsed into three groups: 'higher managerial, administrative and professional occupations', 'intermediate occupations', 'routine and manual occupations', and 'other' – including never worked, long-term unemployed, not working, and unclassifiable); employment status ('full time' – 30 hours or more; 'part-time' – under 30 hours; 'unemployed and seeking work', 'long-term sick or disabled', 'looking after home or family', and 'other' – including full-time education, government training scheme, retired, temporarily sick or disabled); current marital status ('married', 'divorced, separated, or widowed', 'single and never married' – including civil partnership equivalent); living arrangements ('no partner', 'spouse', 'cohabiting' – available in the NCDS 2013 and LS 2011 surveys only); and, housing tenure ('own outright', 'own with mortgage', 'renting or other arrangement' – available in the NCDS 2004 survey and LS 2001 surveys only).

Exposures

Socio-economic factors available in the NCDS 2004 survey and LS 2001 surveys (listed previously) were used as exposure variables in longitudinal analysis.

Outcome

Long-term limiting illness was used as the main outcome variable and was available in the NCDS 2013 and LS 2011 surveys only. LS respondents in 2011 were asked 'Are your day-to-day activities limited because of a health problem or disability which has lasted, or is expected to last, at least 12 months?' ('yes limited a lot', 'yes limited a little', and 'no'). NCDS participants in 2013 were asked 'Do you have any physical or mental health conditions or illnesses lasting or expected to last 12 months or more? And if yes, 'Do any of your conditions or illnesses reduce your ability to carry out day-to-day activities?' ('yes, a lot', 'yes, a little', 'not at all'). For both studies, long-term limiting illness was defined by those who answered that they were limited 'a lot' or 'a little'.

Study samples

For the NCDS, eligible participants included all those who responded to the 55-year survey in 2013, and who were resident in England or Wales at the time of data collection (n=8107); participants not resident in England and Wales were excluded to match the LS England and Wales census data. For longitudinal analyses, participants were included if they had complete data on the outcome and socio-demographic exposure variables; 155 (1.9%) participants had missing data on long-standing illness; missingness across exposure variables ranged from n=0 to n=187 (0-2.3%).

The LS sample included all those who were enumerated in the 2011 sample (n=585,900) and who were aged 55 years at the time of data collection (n=7052). Participants were included in longitudinal analyses if they had complete data on long-term limiting illness and socio-demographic exposure variables. 141 (2.1%) respondents had missing data on long-standing illness; missingness across exposure variables ranged from n=0 to n=151 (0-2.1%).

Statistical analysis

To examine whether the NCDS was representative of the population of England and Wales, we compared the prevalence of socio-demographic factors and longstanding limiting illness in the most recent NCDS 55-year survey in 2013 with LS respondents aged 55 in 2011. The prevalence of socio-demographic factors in the NCDS 46-year survey in 2004 and LS respondents aged 45 in 2001 were also compared which represented the exposure variables used in longitudinal analyses.

To assess the generalisability of NCDS findings, we compared longitudinal associations between socio-economic exposures and long-term limiting illness in the NCDS and LS studies. In the NCDS, long-term limiting illness in 2013 was regressed against exposures in 2004; and in the LS, long-term limiting illness in 2011 was regressed against exposures in 2001 (when respondents were aged 55 and 45, respectively). For each study, we first conducted univariable logistic regression to assess the relationship between each exposure and long-term limiting illness. Second, models were mutually adjusted for all exposures to examine the extent to which univariable associations were independent; due to multicollinearity between economic activity and NS-SEC, models were run including and excluding these variables in turn.

Longitudinal associations were assessed using logistic regression, and chi-squared tests were used to examine whether between-sample differences were statistically significant. To maintain adequate power for statistical analyses, region, economic activity, ethnicity and NS-SEC were collapsed into fewer categories. Sex interactions with socio-demographic exposures were assessed using likelihood ratios tests.

Sensitivity analyses included excluding all immigrants arriving in the UK after age 16 from the LS sample to assess the extent to which between-study differences could be explained by immigration. To examine whether restricting the NCDS sample to those resident in England and Wales altered associations, we repeated analyses using an NCDS sample including all possible participants, as would be typically used by researchers.

To assess whether imputation could produce a more representative sample, multiple imputation using chained equations was used to impute missing NCDS data for all those who participated in the 2013 survey and resident in England and Wales (n=8107) (White et al., 2011). Imputation models were run across ten datasets, and included variables shown to predict non-response in the NCDS (Atherton et al., 2008). To examine the representativeness of the longitudinal sample, we also compared descriptive characteristics of the of the complete-case and imputed NCDS sample, with LS data in 2011. The imputed results were similar to those using original values so the former are presented in supplementary tables 1 and 2.

All analyses were conducted using Stata 14 (StataCorp LP, 2014).

RESULTS

TABLE 1 HERE

Table 1 compares sample characteristics for the NCDS and LS samples. NCDS participants aged 55 in 2013 had similar sex and marital status profiles compared to LS respondents aged 55 in 2011. There was evidence of between-sample differences for long-term limiting illness, ethnicity, region, employment status, social class, and living arrangements (all $p<0.001$). LS respondents had a higher prevalence of long-term limiting illness, non-white ethnicity, residency in the South of England, and working in routine and manual, or 'other' occupations. A larger proportion of the NCDS sample were in full-time employment, working in professional or higher managerial occupations, and living with their spouse.

Similar patterns were observed when comparing NCDS participants at the earlier 46-year survey in 2004 with LS respondents aged 45 in 2001. There was also evidence of between-sample differences for housing tenure, where NCDS participants aged 46 in 2004 were more likely to be married, and less likely to be living in rented accommodation compared to LS respondents aged 45 in 2001 ($p<0.001$; housing tenure was not available in the later NCDS 55-year survey).

TABLE 2 HERE

Table 2 compares unadjusted longitudinal associations between socio-demographic factors and long-term limiting illness in the NCDS and LS samples. Associations for sex were almost identical between the NCDS and the age-matched LS samples. Odds ratios for region, economic activity, NS-SEC, marital status, and housing tenure (renting/other arrangement) were larger in the LS sample; however, between study differences were not statistically significant (all $p>0.07$). Sex interactions with socio-demographic exposures were similar in the NCDS and LS (not shown). Associations for NS-SEC and economic activity were slightly stronger in males (both $p<0.004$) with all other exposures demonstrating weak evidence of interaction in the NCDS ($p=0.3$-0.9) and LS ($p=0.2$-0.7) samples.

TABLE 3 HERE

Table 3 shows mutually adjusted longitudinal associations between socio-demographic factors and long-term limiting illness in the NCDS and LS samples. After mutual adjustment

for socio-demographic exposures, the NCDS and LS samples demonstrated a similar pattern of attenuation. Odds ratios were slightly larger in the LS sample; although these differences were not statistically significant (all p>0.09). A similar pattern of results were found for mutually adjusted models including NS-SEC (all p≥0.18; supplementary table 3).

Sensitivity analysis showed that excluding LS respondents who arrived in the UK after age 16 minimised between-study differences in sample characteristics for ethnicity and region (supplementary table 4); for example, the proportion of white respondents aged 55 in 2011 increased from 88.3% to 95.8% (p<0.001) and those resident in the south of England decreased from 50.1 to 47.4% (p=0.005); however, characteristics for sex, employment status, NS-SEC, marital status, living arrangements, and housing tenure were largely unaffected (all p>0.12). Excluding immigrants from the LS sample did little to alter unadjusted (table 2) or mutually adjusted (table 3) odds ratios for socio-demographic variables and long-term limiting illness.

Supplementary table 5 shows that including all possible participants (i.e. including those not resident in England or Wales) in the NCDS sample did not change associations between socio-demographic factors and long-term limiting illness (tests of between-sample heterogeneity all p ≥ 0.31).

The complete-case and imputed sample characteristics were mostly similar (supplementary table 1); and between-study differences for associations between socio-demographic factors and long-term limiting illness were largely unaffected when using imputed NCDS data (supplementary table 2).

DISCUSSION

This study compared sample characteristics and longitudinal associations between socio-demographic factors and long-term limiting illness in the NCDS and LS – a 'gold-standard' reference population for England and Wales with unparalleled coverage and sample size. We showed two important findings: first, participants from the most recent NCDS 55-year sample were mostly unrepresentative of age-matched LS respondents: characteristics for sex and marital status were similar, but NCDS participants demonstrated lower levels of long-term limiting illness and non-white ethnicity, and higher levels of full-time employment, working in professional or higher managerial occupations, and living with their spouse. Second, we found that despite differences in sample characteristics, longitudinal associations between socio-demographic factors and long-term limiting illness were broadly similar in the NCDS and LS samples. Associations tended to be slightly larger in the LS compared to the NCDS; however, these differences were not statistically significant.

Our study is the most comprehensive examination of national representativeness and generalisability of a British birth cohort study to date; we build on existing studies by examining a full range of socio-demographic exposures using country and age-matched samples. Consistent with the 'healthy cohort effect', we found a lower prevalence of socio-economic disadvantage and poor health in the NCDS compared to the LS. These findings are largely in keeping with earlier studies which have contrasted birth cohort data with nationally representative samples (Atherton et al., 2008; Stafford et al., 2013; Wadsworth et al., 2003). For example, Atherton et al., (2008) compared the proportion of non-white ethnicity, paid employment, marriage, and home ownership in the NCDS 45-year survey with 45-49 year olds from the England and Wales census. Between-sample differences were similar to our study, except for home-ownership, which was 1.3% lower in NCDS participants compared to census respondents. It is possible this discrepancy may be partially explained by Atherton's et al., (2008) use of a relatively older census reference population, which spanned from age 45-49 years. More extensive comparisons have been conducted using data from the National Survey of Health and Development (NSHD; also known as the 1946 British birth cohort) (Stafford et al., 2013; Wadsworth et al., 2003). Most recently, Stafford et al., (2013) compared a range of socio-demographic factors from the 60-64 year survey between 2006-2012 with an aged-matched sample of 60-64 year olds from the England

census in 2001. Statistical tests of between-sample differences were not reported, although our results were consistent in direction across all variables presented; namely, long-term limiting illness, sex, employment activity, NS-SEC, housing tenure, and marital status.

To our knowledge, this study is the first to contrast exposure-outcome associations between a birth cohort study and nationally representative data; however, we identified several related studies, which demonstrated similar findings. Batty et al., (2014) compared associations between classic risk-factors for coronary heart disease (CHD) and CHD events in an occupational cohort study (Whitehall II) and a population-based study (British Regional Heart study) – both of which were reliant on voluntary participation. Nevertheless, our findings were remarkably similar; Batty et al., (2014) reported that the occupational study had a substantially lower prevalence of CHD risk factors (i.e. was 'healthier') than the population-based study, yet the findings for risk factor-CHD associations were in close agreement between samples. Likewise, Pizzi et al., (2011) used Monte Carlo simulations to investigate whether using a restricted source population affects the validity of effect estimates. The simulations demonstrated that, under a range of realistic scenarios, a restricted source population produced only weak bias in estimates of the exposure–outcome association. Pizzi et al., (2012) repeated these findings in real-world data by comparing effect estimates in an Italian web-based birth cohort and the wider source population – obtained from birth registry data. The authors found that associations between maternal characteristics and two outcomes (low birth weight and birth by caesarean section) were not biased by sample selection.

Bartlett et al., (2015) demonstrated that complete-case odds ratios can be estimated without bias under an array of different missing data mechanisms, which could explain similarities in effect estimates between restricted and source populations. For example, when missingness occurs in the outcome, exposure(s), or potentially both, complete-case estimates are unbiased provided the probability of being a complete-case is independent of the outcome, conditional on the exposure. This is in keeping with our findings that longitudinal associations between socio-demographic factors and long-term limiting illness were largely unaffected when using imputed NCDS data.

Markedly, there was little evidence that between-sample differences were explained by immigration. Excluding immigrants who arrived in the UK after age 16 from the LS sample

reduced the proportion of non-white respondents and those living in the South of England; but did not appear to alter profiles for long-term limiting illness, sex, employment status, social class, marital status, living arrangements or housing tenure. Similarly, excluding immigrants from the LS sample did little to alter associations between socio-demographic factors and long-term limiting illness. These findings suggest that observed differences could be explained by other factors such as attrition bias, or limitations in study design. For example, the results could be skewed by discrepancies in the mode of data collection (Bowling, 2005) (e.g. nurse interview vs. self-report census) and variations in question wording (e.g. for long-term limiting illness) or coding (i.e. NS-SEC was manually coded); however, the majority of measures were considered identical (e.g. sex, ethnicity, region, marital status, living arrangements, and housing tenure) – see appendix 1 for exact question wording. Moreover, it is possible that some differences could in part be attributed to the robustness of the logistic regression model.

Other methodological considerations included low power to feasibly examine between-study differences for smaller groups; for example, those who were 'non-white' or 'unemployed' accounted for less than 3% of the NCDS sample. Likewise, confidence intervals for several associations (e.g. for 'other' economic activity and 'other' NS-SEC) were relatively wide, increasing vulnerability to type II error. Larger samples would be required to explore these associations in more detail which could be achieved in the LS through pooled years of age (e.g. a 40-49 year subsample would have n=65,600 White ethnic grouping, n=491 Mixed, n=2,317 Indian, n=1,561 Pakistani and Bangladeshi combined, n=1,370 Black and n=1,284 Other). Methodological strengths of the study include very high levels of census enumeration in the LS in 2001 and 2011 (both 94%). There were few missing data on socio-demographic factors and long-term limiting illness in the NCDS and LS samples (0-2.1% and 0-2.3%, respectively), and unlike previous studies (Atherton et al., 2008; Stafford et al., 2013; Wadsworth et al., 2003), we matched the NCDS and reference samples by country and age.

In conclusion, we have shown that NCDS participants had greater socio-demographic and health advantage compared to LS respondents; however, despite these differences, longitudinal associations between socio-demographic exposures and long-term limiting illness were similar in the NCDS and LS – suggesting that the NCDS findings shown are

largely generalisable to the population of England and Wales. Excluding immigrants from the LS sample did little to alter results, which implied that any between-study differences were likely attributable to sample and attrition bias, or other methodological factors associated with study design. Further research is required to better understand which of these factors best explain sample differences in the NCDS, and moreover, whether our results hold in other exposure-outcome relationships in different cohort studies. Pioneering techniques, such as using administrative data to create longitudinal weights (Douglas et al., 2018) for cohort data could be explored to help generate estimates of maximum relevance to public health policy.

References

Atherton, K., Fuller, E., Shepherd, P., Strachan, D. P., & Power, C. (2008) 'Loss and representativeness in a biomedical survey at age 45 years: 1958 British birth cohort', *Journal of Epidemiology and Community Health*, 62(3): 216–223.

Bartlett, J.W., Harel, O., Carpenter, J.R., 2015 'Asymptotically Unbiased Estimation of Exposure Odds Ratios in Complete Records Logistic Regression', *Am. J. Epidemiol.* 182: 730–736.

Batty, G. D., Shipley, M., Tabák, A., Singh-Manoux, A., Brunner, E., Britton, A., & Kivimäki, M. (2014) 'Generalizability of occupational cohort study findings', *Epidemiology*, 25(6): 932–933.

Bowling, A. (2005) 'Mode of questionnaire administration can have serious effects on data quality', *Journal of Public Health*, 27(3): 281–291.

Douglas, E., Rutherford, A., & Bell, D. (2018) 'Pilot study protocol to inform a future longitudinal study of ageing using linked administrative data: Healthy AGeing in Scotland (HAGIS)', *BMJ Open*, 8(1): e018802.

Ferrie, J. E., Kivimäki, M., Singh-Manoux, A., Shortt, A., Martikainen, P., Head, J., … Shipley, M. J. (2009) 'Non-response to baseline, non-response to follow-up and mortality in the Whitehall II cohort', *International Journal of Epidemiology*, 38(3): 831–837.

Hawkes, D., & Plewis, I. (2006) 'Modelling non-response in the national child development study', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3): 479–491.

Jefferies, J. (2005) 'The UK population: past, present and future', in Chappell R. (eds), *Focus on People and Migration*, Palgrave Macmillan, London, pp. 1–17.

Kuh, D., & Ben-Shlomo, Y. (2004) *A life course approach to chronic disease epidemiology (2nd ed.)*, Oxford, United Kingdom: Oxford University Press.

Lantz, P. M., House, J. S., Lepkowski, J. M., Williams, D. R., Mero, R. P., & Chen, J. (1998) 'Socioeconomic Factors, Health Behaviors, and Mortality: Results From a Nationally Representative Prospective Study of US Adults', *JAMA*, 279(21): 1703–1708.

Nathan, G. (1999) *A review of sample attrition and representativeness in three longitudinal surveys*, Government Statistical Service Methodology Series No. 13.

ONS. (2002), *International Migrants in England and Wales: 2011,* Office for National Statistics.

Pizzi, C., Stavola, B.D., Merletti, F., Bellocco, R., Silva, I. dos S., Pearce, N., Richiardi, L., (2011), 'Sample selection and validity of exposure–disease association estimates in cohort studies', *J. Epidemiol. Community Health*, 65: 407–411.

Pizzi, C., Stavola, B.L.D., Pearce, N., Lazzarato, F., Ghiotti, P., Merletti, F., Richiardi, L., (2012) 'Selection bias and patterns of confounding in cohort studies: the case of the NINFEA web-based birth cohort', *J Epidemiol Community Health*, 66: 976–981.

Plewis, I., Calderwood, L., Hawkes, D., & Nathan, G. (2004), *National Child Development Study and 1970 British Cohort Study Technical Report. 1st Edition*, London: Institute of Education.

Power, C., & Elliott, J. (2006), 'Cohort profile: 1958 British birth cohort (National Child Development Study)', *International Journal of Epidemiology*, 35(1): 34–41.

Rothman, K. J., Gallacher, J. E., & Hatch, E. E. (2013) 'Why representativeness should be avoided', *International Journal of Epidemiology*, 42(4): 1012–1014.

Shelton, N., Marshall, C. E., Stuchbury, R., Grundy, E., Dennett, A., Tomlinson, J., … Xun, W. (2018), 'Cohort Profile: the Office for National Statistics Longitudinal Study (The LS)', *International Journal of Epidemiology.*

Stafford, M., Black, S., Shah, I., Hardy, R., Pierce, M., Richards, M., … Kuh, D. (2013) 'Using a birth cohort to study ageing: representativeness and response rates in the National Survey of Health and Development', *European Journal of Ageing*, 10(2): 145–157.

StataCorp LP. (2014), *Stata Statistical Software: Release 14*, College Station, TX: StataCorp LP.

Wadsworth, M. E. J., Butterworth, S. L., Hardy, R. J., Kuh, D. J., Richards, M., Langenberg, C., … Connor, M. (2003), 'The life course prospective design: an example of benefits and problems associated with study longevity', *Social Science & Medicine*, 57(11): 2193–2205.

White, I.R., Royston, P., Wood, A.M., (2011), 'Multiple imputation using chained equations: Issues and guidance for practice', *Stat. Med*, 30: 377–399.

Table 1. Comparison of sample characteristics between NCDS participants aged 46 in 2001 and LS respondents aged 45 in 2001; and NCDS participants age 55 in 2013 and LS respondents aged 55 in 2011

	NCDS 2004 (age 46) n=8689	**LS 2001 (age 45)** n=7157	**P[c]**	**NCDS 2013 (age 55)** n=8107	**LS 2011 (age 55)** n=7052	**P[c]**
	%	%		%	%	
Long-term limiting illness						
Yes		14.9		19.7	22.8	<0.001
No		85.1		80.3	77.2	
Missing (n)		*141*		*115*	*155*	
Sex						
Male	48.7	49.4	0.37	48.5	49.3	0.32
Female	51.3	50.6		51.5	50.7	
Missing (n)	*0*	*0*		*0*	*0*	
Ethnicity						
White	98.0	90.3	<0.001	97.9	88.3	<0.001
Non-white	2.0	9.7		2.1	11.7	
Missing (n)	*0*	*113*		*0*	*116*	
Region						
South	47.9	49.4	0.06	46.0	50.1	<0.001
North	46.1	45.3		48.1	44.6	
Wales	6.0	5.3		6.0	5.3	
Missing (n)	*3*	*2*		*0*	*0*	
Employment status						
Full-time	69.0	61.1	<0.001	61.2	55.2	<0.001
Part-time	18.4	17.7		20.2	19.0	
Unemployed	1.7	3.2		2.9	4.3	
Long-term sick/disabled	4.0	6.3		5.2	9.2	
Looking after home/family	5.4	7.3		6.2	5.1	
Other[a]	1.7	4.4		4.3	7.1	
Missing (n)	*0*	*3*		*120*	*151*	
Social class NS-SEC						
Professional/higher management	41.9	33.9	<0.001	35.7	29.2	<0.001
Intermediate	19.8	18.4		23.1	20.1	
Routine and manual	25.7	28.0		20.9	25.1	
Other[b]	12.7	19.7		20.3	25.6	
Missing (n)	*27*	*3*		*120*	*0*	
Marital status						
Married	71.1	68.7	0.004	71.5	70.0	0.10
Divorced/separated/ widowed	17.5	18.7		18.6	19.4	
Single	11.5	12.7		9.9	10.6	
Missing (n)	*18*	*17*		*5*	*55*	
Living arrangements						
No partner		22.9		21.0	26.9	<0.001
Spouse		68.1		69.1	64.9	

Co-habiting		9.0		10.0	8.2
Missing (n)		*47*		*0*	*49*
Housing tenure					
Own – outright	14.3	16.2	<0.001		34.0
Own - mortgage	71.5	63.5			43.4
Rent/other	14.2	20.3			22.6
Missing (n)	*39*	*193*			*101*

NCDS sample restricted to those resident in England and Wales

Blank fields mean variables were not available in the NCDS at equivalent time-points

a: Full-time education, government training scheme, retired, temporarily sick or disabled

b: Never worked, long-term unemployed, not working, unclassifiable

c: p value for between sample heterogeneity

Source: Data ONS LS and NCDS; analysis conducted by the authors

Table 2. Unadjusted longitudinal associations between socio-demographic exposures and long-standing limiting illness in NCDS and LS studies

		Long-term limiting illness, OR (95% CI)				
Exposure		**NCDS[a]**	**LS[b]**	**p[c]**	**LS[d] (excl. immigration)**	**p[c]**
Sex						
	Male	ref	ref	0.82	ref	0.67
	Female	1.28 (1.14,1.45)	1.26 (1.11,1.42)		1.24 (1.09,1.41)	
Region						
	South	ref		0.38		0.33
	North	1.24 (1.10,1.40)	1.39 (1.22,1.57)		1.38 (1.21,1.58)	
	Wales	1.41 (1.10,1.81)	1.59 (1.23,2.06)		1.68 (1.30,2.19)	
Economic activity						
	Full-time	ref	ref	0.01	ref	0.004
	Part-time	1.22 (1.04,1.43)	1.44 (1.20,1.71)		1.48 (1.23,1.78)	
	Other[e]	4.99 (4.26,5.85)	6.75 (5.82,7.82)		6.98 (5.96,8.18)	
NS-SEC						
	Professional/higher managerial	ref	ref	0.14	ref	0.18
	Intermediate	1.09 (0.91,1.30)	1.13 (0.92,1.39)		1.07 (0.86,1.33)	
	Routine and manual	1.39 (1.18,1.62)	1.53 (1.29,1.82)		1.47 (1.23,1.77)	
	Other[f]	5.36 (4.51,6.36)	7.02 (5.91,8.35)		7.07 (5.89,8.49)	
Marital status						
	Married	ref	ref	0.07	ref	0.04
	Divorced, separated, widowed	1.39 (1.19,1.63)	1.68 (1.45,1.96)		1.71 (1.46,2.01)	
	Single	1.66 (1.39,1.98)	2.01 (1.68,2.39)		2.08 (1.73,2.51)	
Housing tenure						
	Own – outright	ref	ref	0.09	ref	0.09
	Own – mortgage/loan	0.84 (0.71,0.99)	0.87 (0.73,1.04)		0.85 (0.70,1.03)	
	Rent/other	2.26 (1.84,2.79)	3.13 (2.55,3.83)		3.16 (2.55,3.92)	

NCDS sample restricted to those resident in England and Wales
a: Long-term limiting illness in 2013 regressed against socio-demographic exposures in 2004; analytical samples range between n=7007-7038
b: Long-term limiting illness in 2011 regressed against socio-demographic exposures in 2001; analytical samples range between n=5888-6017
c: p value for heterogeneity between the NCDS and LS
d: Long-term limiting illness in 2011 regressed against socio-demographic exposures in 2001, excluding immigrants who arrived in the UK after age 16; analytical samples range between n=5370-5475
e: Looking after home or family, full-time education, government training scheme, retired, temporarily sick or disabled
f: Never worked, long-term unemployed, not working, unclassifiable
Source: Data ONS LS and NCDS; analysis conducted by the authors

Table 3. Mutually adjusted longitudinal associations between socio-demographic exposures and long-term limiting illness in the NCDS and LS

	Long-term limiting illness, OR (95% CI)				
Exposure	**NCDS[a]**	**LS[b]**	**p[c]**	**LS (excl. immigration)[d]**	**p[c]**
Sex					
Male	ref	ref	0.29	ref	0.28
Female	1.04 (0.91,1.20)	0.93 (0.80,1.09)		0.93 (0.79,1.09)	
Region					
South	ref		0.23		0.34
North	1.23 (1.08,1.40)	1.41 (1.23,1.62)		1.36 (1.17,1.57)	
Wales	1.34 (1.03,1.75)	1.62 (1.22,2.16)		1.65 (1.24,2.20)	
Economic activity					
Full-time	ref	ref	0.09	ref	0.07
Part-time	1.26 (1.05,1.51)	1.53 (1.25,1.86)		1.58 (1.28,1.95)	
Other[e]	4.38 (3.68,5.20)	5.41 (4.57,6.40)		5.39 (4.51,6.45)	
Marital status					
Married	ref	ref	0.63	ref	0.61
Divorced, separated, widowed	1.16 (0.98,1.38)	1.30 (1.09,1.55)		1.31 (1.09,1.57)	
Single	1.44 (1.19,1.74)	1.50 (1.22,1.84)		1.51 (1.22,1.87)	
Housing tenure					
Own – outright	ref	ref	0.48	ref	0.58
Own – mortgage/loan	1.01 (0.84,1.20)	1.12 (0.92,1.36)		1.09 (0.89,1.34)	
Rent/other	1.86 (1.49,2.33)	2.15 (1.73,2.69)		2.15 (1.70,2.71)	

NCDS sample restricted to those resident in England and Wales

NS-SEC excluded due to multi-collinearity with economic activity; for model including NS-SEC see supplementary table 1

a: Long-term limiting illness in 2013 regressed against socio-demographic exposures in 2004; analytical sample n=7003

b: Long-term limiting illness in 2011 regressed against socio-demographic exposures in 2001; analytical sample n=5871

c: p value for heterogeneity between the NCDS and LS

d: Long-term limiting illness in 2011 regressed against socio-demographic exposures in 2001, excluding immigrants who arrived in the UK after age 16; analytical sample n=5355

e: Looking after home or family, full-time education, government training scheme, retired, temporarily sick or disabled
Source: Data ONS LS and NCDS; analysis conducted by the authors

Supplementary table 1. Comparison of sample characteristics between original, imputed, and complete-case NCDS samples, and LS respondents aged 55 in 2011

	NCDS 2013 n=8107	**Imputed NCDS[a]** n=8107	**Complete-case NCDS[b]** n=7003	**LS 2011** n=7052	**p[c]**
	%	%	%	%	
Long-term limiting illness					
Yes	19.7	19.7	19.2	22.8	<0.001
No	80.3	80.3	80.8	77.2	
Missing (n)	*115*	-	*0*	*155*	
Sex					
Male	48.5	48.5	48.4	49.3	0.27
Female	51.5	51.5	51.6	50.7	
Missing (n)	*0*	-	*0*	*0*	
Ethnicity					
White	97.9	97.9	98.1	88.3	<0.001
Non-white	2.1	2.1	1.9	11.7	
Missing (n)	*0*	-	*0*	*116*	
Region					
South	46.0	46.0	48.1	50.1	0.04
North	48.1	48.1	46.0	44.6	
Wales	6.0	6.0	5.9	5.3	
Missing (n)	*0*	-	*0*	*0*	
Employment status					
Full-time	61.2	61.0	61.9	55.2	<0.001
Part-time	20.2	20.2	20.6	19.0	
Unemployed	2.9	2.9	2.5	4.3	
Long-term sick/disabled	5.2	5.2	4.8	9.2	
Looking after home/family	6.2	6.3	5.8	5.1	
Other[d]	4.3	4.4	4.2	7.1	
Missing (n)	*120*	-	*80*	*151*	
Social class NS-SEC					
Professional/higher management	35.7	35.8	37.2	29.2	<0.001
Intermediate	23.1	23.1	23.4	20.1	
Routine and manual	20.9	20.8	20.7	25.1	
Other[e]	20.3	20.3	18.7	25.6	
Missing (n)	*120*	-	*128*	*0*	
Marital status					
Married	71.5	71.5	72.6	70.0	0.003
Divorced/separated/ widowed	18.6	18.6	18.0	19.4	
Single	9.9	9.9	9.5	10.6	
Missing (n)	*5*	-	*3*	*55*	
Living arrangements					
No partner	21.0	21.0	19.7	26.9	<0.001
Spouse	69.1	69.1	70.4	64.9	
Co-habiting	10.0	10.0	9.9	8.2	

Missing (n)	*0*	*-*	*0*	*49*

NCDS samples restricted to those resident in England and Wales

a: Includes all those who responded in 2013 survey; based on ten imputations

b: Includes all those with complete exposure data in 2004 and outcome data in 2013 (longitudinal sample)

c: NCDS complete-case vs. LS, p value for between-sample heterogeneity

d: Full-time education, government training scheme, retired, temporarily sick or disabled

e: Never worked, long-term unemployed, not working, unclassifiable

Source: Data ONS LS and NCDS; analysis conducted by the authors

Supplementary table 2. Mutually adjusted longitudinal associations between socio-demographic exposures and long-term limiting illness in the NCDS (original and imputed) and LS

		Long-term limiting illness, OR (95% CI)				
Exposure		**NCDS[a]**	**NCDS (imputed)[b]**	**p[c]**	**LS[d]**	**p[e]**
Sex						
	Male	ref	ref	0.87	ref	0.35
	Female	1.04 (0.91,1.20)	1.03 (0.90,1.17)		0.93 (0.80,1.09)	
Region						
	South	ref		0.98		0.29
	North	1.23 (1.08,1.40)	1.25 (1.11,1.42)		1.41 (1.23,1.62)	
	Wales	1.34 (1.03,1.75)	1.34 (1.04,1.72)		1.62 (1.22,2.16)	
Economic activity						
	Full-time	ref	ref	0.98	ref	0.09
	Part-time	1.26 (1.05,1.51)	1.28 (1.05,1.55)		1.53 (1.25,1.86)	
	Other[f]	4.38 (3.68,5.20)	4.35 (3.68,5.14)		5.41 (4.57,6.40)	
Marital status						
	Married	ref	ref	0.94	ref	0.82
	Divorced, separated, widowed	1.16 (0.98,1.38)	1.21 (1.02,1.44)		1.30 (1.09,1.55)	
	Single	1.44 (1.19,1.74)	1.44 (1.21,1.72)		1.50 (1.22,1.84)	
Housing tenure						
	Own – outright	ref	ref	0.97	ref	0.59
	Own – mortgage/loan	1.01 (0.84,1.20)	1.04 (0.87,1.24)		1.12 (0.92,1.36)	
	Rent/other	1.86 (1.49,2.33)	1.88 (1.50,2.36)		2.15 (1.73,2.69)	

NCDS sample restricted to those resident in England and Wales

NS-SEC excluded due to multi-collinearity with economic activity

a: Long-term limiting illness in 2013 regressed against socio-demographic exposures in 2004; analytical sample n=7003

b: Long-term limiting illness in 2013 regressed against socio-demographic exposures in 2004; analytical sample based on ten imputations, n=8107

c: p value for heterogeneity between the original and imputed NCDS sample and LS

d: Long-term limiting illness in 2011 regressed against socio-demographic exposures in 2001; analytical sample n=5871

e: p value for heterogeneity between the imputed NCDS sample and LS

f: Looking after home or family, full-time education, government training scheme, retired, temporarily sick or disabled
Source: Data ONS LS and NCDS; analysis conducted by the authors

Supplementary table 3. Mutually adjusted longitudinal associations between socio-demographic exposures and long-term limiting illness in the NCDS and LS (including NS-SEC)

		Long-term limiting illness, OR (95% CI)				
Exposure		**NCDS[a]**	**LS[b]**	**p[c]**	**LS (excl. immigration)[d]**	**p[c]**
Sex						
	Male	ref	ref	0.60	ref	0.66
	Female	1.13 (0.99,1.28)	1.07 (0.93,1.23)		1.08 (0.93,1.25)	
Region						
	South	ref		0.18		0.30
	North	1.20 (1.06,1.37)	1.40 (1.22,1.61)		1.36 (1.17,1.57)	
	Wales	1.31 (1.01,1.71)	1.60 (1.20,2.13)		1.63 (1.22,2.18)	
NS-SEC						
	Professional/higher managerial	ref	ref	0.74	ref	0.91
	Intermediate	1.04 (0.87,1.25)	1.08 (0.88,1.34)		1.03 (0.82,1.28)	
	Routine and manual	1.27 (1.08,1.49)	1.32 (1.10,1.58)		1.26 (1.04,1.52)	
	Other[e]	4.46 (3.72,5.35)	5.13 (4.25,6.20)		4.92 (4.03,6.01)	
Marital status						
	Married	ref	ref	0.57	ref	0.59
	Divorced, separated, widowed	1.14 (0.96,1.35)	1.29 (1.09,1.54)		1.29 (1.08,1.55)	
	Single	1.42 (1.18,1.72)	1.48 (1.21,1.81)		1.49 (1.20,1.84)	
Housing tenure						
	Own – outright	ref	ref	0.38	ref	0.49
	Own – mortgage/loan	1.00 (0.84,1.20)	1.07 (0.89,1.30)		1.04 (0.85,1.27)	
	Rent/other	1.80 (1.44,2.25)	2.21 (1.77,2.75)		2.18 (1.73,2.76)	

NCDS sample restricted to those resident in England and Wales

Economic activity excluded due to multi-collinearity with NS-SEC; for model including economic activity see table 3

a: Long-term limiting illness in 2013 regressed against socio-demographic exposures in 2004; analytical sample n=6987

b: Long-term limiting illness in 2011 regressed against socio-demographic exposures in 2001; analytical sample n=5873

c: p value for heterogeneity between the NCDS and LS

d: Long-term limiting illness in 2011 regressed against socio-demographic exposures in 2001, excluding immigrants who arrived in the UK after age 16; analytical sample n=5356

e: Never worked, long-term unemployed, not working, unclassifiable

Source: Data ONS LS and NCDS; analysis conducted by the authors

Supplementary table 4. Comparison of sample characteristics from the LS in 2001 and 2011, including and excluding immigrants who entered the UK prior to age 16

	LS 2001 (age 45) n=7157	**LS 2001 (age 45) excl. immigration** n=6393	**p[c]**	**LS 2011 (age 55)** n=7052	**LS 2011 (age 55) excl. immigration** n=6170	**p[c]**
	%	%		%	%	
Long-term limiting illness						
Yes	14.9	15.0		22.8	22.5	0.65
No	85.1	85.0		77.2	77.5	
Missing (n)	*141*	*99*		*115*	*127*	
Sex						
Male	49.4	49.9	0.58	49.3	49.9	0.50
Female	50.6	50.1		50.7	50.1	
Missing (n)	*0*	*0*		*0*	*0*	
Ethnicity						
White	90.3	96.9	<0.001	88.3	95.8	<0.001
Non-white	9.7	3.1		11.7	4.2	
Missing (n)	*113*	*113*		*116*	*95*	
Region						
South	49.4	47.2	0.03	50.1	47.4	0.005
North	45.3	47.0		44.6	46.7	
Wales	5.3	5.8		5.3	5.9	
Missing (n)	*2*	*2*		*0*	*0*	
Employment status						
Full-time	61.1	62.4	0.12	55.2	56.0	0.44
Part-time	17.7	18.0		19.0	19.5	
Unemployed	3.2	3.1		4.3	4.3	
Long-term sick/disabled	6.3	6.4		9.2	9.1	
Looking after home/family	7.3	6.2		5.1	4.5	
Other[a]	4.4	4.0		7.1	6.6	
Missing (n)	*3*	*2*		*151*	*126*	
Social class NS-SEC						
Professional/higher management	33.9	34.8	0.27	29.2	30.0	0.32
Intermediate	18.4	18.6		20.1	20.7	
Routine and manual	28.0	28.3		25.1	25.0	
Other[b]	19.7	18.3		25.6	24.3	
Missing (n)	*3*	*2*		*0*	*0*	
Marital status						
Married	68.7	67.8	0.55	70.0	69.3	0.50
Divorced/separated/ widowed	18.7	19.1		19.4	19.5	
Single	12.7	13.1		10.6	11.2	
Missing (n)	*17*	*14*		*55*	*43*	
Living arrangements						
No partner	22.9	22.9		26.9	26.2	0.45
Spouse	68.1	67.7		64.9	65.1	
Co-habiting	9.0	9.4		8.2	8.7	

Missing (n)	*47*	*41*		*49*	*36*
Housing tenure					
Own - outright	16.2	15.7	0.14	34.0	35.3
Own - mortgage	63.5	65.1		43.4	43.9
Rent/other	20.3	19.2		22.6	20.8
Missing (n)	*193*	*140*		*101*	*76*

a: Full-time education, government training scheme, retired, temporarily sick or disabled

b: Never worked, long-term unemployed, not working, unclassifiable

c: p value for heterogeneity between LS samples including and excluding immigration

Source: Data ONS LS and NCDS; analysis conducted by the authors

Supplementary table 5. Unadjusted longitudinal associations between socio-demographic exposures and long-term limiting illness in the NCDS, excluding and including those not resident in England and Wales.

		Long-term limiting illness, OR (95% CI)	
Exposure		**NCDS[a]**	**NCDS (all)[b]**
Sex			
	Male	ref	ref
	Female	1.28 (1.14,1.45)	1.29 (1.15,1.45)
Region			
	South	ref	ref
	North	1.24 (1.10,1.40)	1.24 (1.10,1.40)
	Wales	1.41 (1.10,1.81)	1.42 (1.11,1.83)
	Scotland	-	1.22 (0.99,1.49)
Employment status			
	Full-time	ref	ref
	Part-time	1.22 (1.04,1.43)	1.18 (1.01,1.38)
	Other[c]	4.99 (4.26,5.85)	4.82 (4.14,5.60)
NS-SEC			
	Professional/higher managerial	ref	ref
	Intermediate	1.09 (0.91,1.30)	1.12 (0.95,1.32)
	Routine and manual	1.39 (1.18,1.62)	1.32 (1.13,1.53)
	Other[d]	5.36 (4.51,6.36)	5.16 (4.39,6.07)
Marital status			
	Married	ref	ref
	Divorced, separated, widowed	1.39 (1.19,1.63)	1.34 (1.15,1.56)
	Single	1.66 (1.39,1.98)	1.55 (1.31,1.83)
Housing tenure			
	Own – outright	ref	ref
	Own – mortgage/loan	0.84 (0.71,0.99)	0.82 (0.69,0.96)
	Rent/other	2.26 (1.84,2.79)	2.23 (1.83,2.72)

All p-values for between-sample heterogeneity $\geq$ 0.31.

a: NCDS sample restricted to those resident in England and Wales; analytical samples range between n=7007-7038

b: NCDS sample including all possible cohort members; analytical samples range between n=7768-7800

c: Includes looking after home or family, full-time education, government training scheme, retired, temporarily sick or disabled

d: Includes never worked, long-term unemployed, not working, unclassifiable

Source: Data ONS LS and NCDS; analysis conducted by the authors