Examining the quality and sample representativeness of linked survey and administrative data

Linking the 1958 National Child Development Study to Hospital Episode Statistics data

CLS working paper number 2022/5

Richard J Silverwood[1], Nasir Rajah[1], Lisa Calderwood[1], Bianca L De Stavola[2], Katie Harron[2], George B Ploubidis[1]

1 Centre for Longitudinal Studies, UCL Social Research Institute, 20 Bedford Way, London WC1H 0AL

2 Population, Policy & Practice Research and Teaching Department, UCL Great Ormond Street Institute of Child Health, 30 Guilford Street, London WC1N 1EH

CENTRE FOR LONGITUDINAL STUDIES

UKRI Economic and Social Research Council

Contact the author

Richard Silverwood

UCL Centre for Longitudinal Studies

r.silverwood@ucl.ac.uk

This working paper was first published in May 2022 by the UCL Centre for Longitudinal Studies.

UCL Social Research Institute
University College London
20 Bedford Way
London WC1H 0AL
www.cls.ucl.ac.uk

The UCL Centre for Longitudinal Studies (CLS) is an Economic and Social Research Council (ESRC) Resource Centre based at the UCL Social Research Institute, University College London. It manages four internationally-renowned cohort studies: the 1958 National Child Development Study, the 1970 British Cohort Study, Next Steps, and the Millennium Cohort Study. For more information, visit www.cls.ucl.ac.uk.

This document is available in alternative formats. Please contact the Centre for Longitudinal Studies.

Tel: +44 (0)20 7612 6875
Email: clsfeedback@ucl.ac.uk

Disclaimer

This working paper has not been subject to peer review.

CLS working papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Any opinions expressed here are those of the author(s) and not those of the UCL Centre for Longitudinal Studies (CLS), the UCL Social Research Institute, University College London, or the Economic and Social Research Council.

How to cite this paper

Silverwood, R., Rajah, N., Calderwood, L., De Stavola, B.L., Harron, K., Ploubidis, G.B. (2022) *Examining the quality and sample representativeness of linked survey and administrative data: linking the 1958 National Child Development Study to Hospital Episode Statistics data.* CLS Working Paper 2022/5. London: UCL Centre for Longitudinal Studies.

Abstract

Recent years have seen an increase in linkages between cohort and administrative data. It is important to evaluate the quality of such data linkages to discern the likely reliability of research using the linked data resource. In this paper we consider a recent linkage between the 1958 National Child Development Study (NCDS), a cohort following the lives of an initial 17,415 people born in Great Britain in a single week of 1958, and Hospital Episode Statistics (HES) databases, which contain details of all admissions, accident and emergency attendances and outpatient appointments at NHS hospitals in England. We examine the quality of the linkage in terms of the associations between key cohort member sociodemographic characteristics and successful linkage, and compare the levels of successful linkage within strata of NCDS variables which may be expected to be associated with hospital attendance, and hence with successful HES linkage (self-reported hospital attendance, self-rated general health, self-reported long-term illness). We additionally evaluate the population representativeness of the linked sample using external data (hospital admission rates in the general population). Our findings suggest that the linkage quality of the NCDS-HES data is high and that the linked sample maintains an excellent level of population representativeness. We hope that these analyses will both improve the quality and transparency of research using this linked data resource and encourage providers and users of other linked data resources to undertake and publish similarly thorough evaluations.

Key Words

Administrative data; Cohort studies; Data linkage; Hospital Episode Statistics; Linkage quality; National Child Development Study; Representativeness

Introduction

Over recent decades, the increasing availability of administrative data, generally derived from the operation of administrative systems, typically by public sector agencies (1), has led to an expansion of research utilising these resources (2). Administrative data afford exciting new opportunities for health (3) and social science research (1), in particular to answer questions that require additional information, large sample sizes or involve hard-to-reach populations (4). There has been a corresponding increase in the linkage of surveys with administrative data, with the primary motivation being to enhance the survey data in order to provide greater opportunities for research (5). Linkages between surveys and administrative data provide the opportunity to harness the richness of the self-reported survey data alongside the scale and (often) detail of the administrative data, resulting in a resource with greater potential for research than either data source in isolation.

It is important to evaluate the quality of such data linkages to discern the likely reliability of research using the linked data resource and assess whether additional methods may need to be employed to try to address potential biases. Linkage is generally undertaken only for survey participants who have given explicit consent. Since consenters may not be representative of the broader sample (6), there is the potential for selection bias even in the presence of perfect linkage of administrative data for consenters. Such perfect linkage, however, is unlikely in practice, raising further possibility of bias. ‘Linkage error’ describes missed links between records that relate to the same person (‘missed matches’) or false links between unrelated records (‘false matches’) (7). Despite advances in linkage methods and improvements in data quality over time, some degree of linkage error or uncertainty remains almost always inevitable for linkages involving administrative data (4). It is important to examine how linkage errors differ with respect to variables of interest. Differential linkage error can lead to substantial bias, even when overall error rates are low (4), and there is evidence that key participant characteristics are often unevenly distributed in matched and unmatched records (8), suggesting the potential presence of differential linkage error.

Recommended methods for evaluating linkage quality include using a ‘gold standard’ dataset to quantify false matches and missed matches, comparing characteristics of linked and unlinked data to identify potential sources of bias, and sensitivity analyses to evaluate how sensitive results are to changes in linkage procedure (9). However, particularly in settings with a separation of processes for linkage and analysis to help preserve privacy, not all these options may be available to the researcher. Use of gold standard data, where the true match status of each pair of records is known, generally requires the involvement of the data linker. This may also be the case for sensitivity analyses where the linkage procedure is varied, though if this approach is enacted at the time of the initial linkage then researchers may have access to matching meta-data which allow sensitivity analyses to be conducted. Comparisons of characteristics of linked and unlinked data may therefore be the most viable option, though in settings where not all records within a dataset are expected to link to records in the other dataset, for example linking hospital records into a general population master dataset, interpretation is not straightforward (9).

In linkages of survey and administrative data, if there exist data collected on the survey participants not via the linkage in question which capture similar information to that contained in the linked data, this can provide an opportunity to examine linkage quality. By comparing the corresponding survey and administrative variables at the individual level it is possible to assess to what extent the two data sources are in agreement across the linked sample. Interpretation of any discrepancies must consider whether each data source can be assumed to provide a valid measure of the intended underlying construct or whether measurement error is also likely to be a contributory factor.

Comparisons of linked data with external representative data sources can also help identify whether the linked records are broadly representative, or whether linkage errors might have contributed to observed differences (4). Target population representativeness of linked survey-administrative data may be compromised for one of several reasons, including: selective consent to data linkage, differential linkage error (5), selection into the survey, and, for longitudinal studies, selective attrition prior to linkage consent being sought. Using external data, comparison must necessarily be at the sample (or sub-sample), rather than individual, level: is the distribution of the linked administrative variable across the linked sample comparable to the distribution of the corresponding external variable across the population? In order for the assessment of the representativeness to be meaningful, it is important that the data are comparable in terms of scope, timeframe and demography.

In this paper we consider a recent linkage between the 1958 National Child Development Study (NCDS; a long-running British cohort study) and Hospital Episode Statistics (HES; database of English hospital admissions, attendances and appointments) data (10). Due to the separation of processes for linkage and analysis, available approaches for the evaluation of linkage quality do not include use of a gold standard dataset to quantify false matches and missed matches or sensitivity analyses to evaluate how sensitive results are to changes in linkage procedure (9). Instead, we utilise several different approaches using additional variables from within NCDS and published population-level HES data. As NCDS is a longitudinal cohort study, we are able to utilise variables collected at different time points. In addition to providing an examination of the quality and representativeness of the NCDS-HES linkage specifically, these analyses form a demonstration of generalisable methods for evaluating linkage quality which could be utilised in similar settings.

We examine the quality of the linkage in terms of the associations between key cohort member sociodemographic characteristics and successful linkage, and compare the levels of successful linkage within strata of NCDS variables which may be expected, to a greater or lesser extent, to be associated with hospital attendance, and hence with successful HES linkage (self-reported hospital attendance, self-rated general health, self-reported long-term illness). We additionally evaluate the population representativeness of the linked sample using external data (hospital admission rates in the general population). One feature of linkage to HES data is that cohort members may legitimately not have a HES record (i.e. if they have not attended an NHS hospital in England over the period being considered) – the links can be ‘meaningfully interpreted’ (7). In the absence of additional contextual information, such cases are not distinguishable from cohort members who did have

HES records but were not successfully linked (missed matches), but this is an important difference with consequences for potential bias in subsequent analyses. Additional consideration will be given to this issue.

Our aim is that by providing a thorough assessment of the linkage quality and sample representativeness of the NCDS-HES linkage we will both improve the quality and transparency of research using this linked data resource (9) and encourage providers and users of other linked data resources to undertake and publish similarly detailed evaluations.

Methods

Data

NCDS

The NCDS follows the lives of an initial 17,415 people born in Great Britain in a single week of 1958 (11). Since the birth sweep, NCDS cohort members have been followed up many times, with the most recent completed conventional sweep undertaken in 2013 when cohort members were 55, three waves of COVID-19-specific surveys undertaken between May 2020 and March 2021, and a further conventional sweep currently underway (as of 2022). The study includes information on cohort members' physical and educational development, economic circumstances, employment, family life, health behaviour, wellbeing, social participation, and attitudes.

HES

HES is a collection of databases containing details of all admissions (Admitted Patient Care (APC) and Critical Care (CC)), Accident and Emergency (A&E) attendances and Outpatient (OP) appointments at NHS hospitals in England, maintained by NHS Digital (12). The period of data availability differs by dataset, from 1997 for APC, from 2007 for A&E, from 2009 for CC and from 2003 for OP. This paper focusses on data obtained from the HES APC dataset, which provides, for each hospital episode, information on admission and discharge dates, diagnoses, procedures, patient demographics, and hospital characteristics (13) and from the HES OP dataset, which includes, for each outpatient appointment, information on dates, diagnoses, procedures and patient demographics.

Linked NCDS-HES data

Linkage between NCDS and all four HES datasets has recently been undertaken, on the basis of consent at sweep 8 (age 50) (10), with data available via secure access through the UK Data Service (14). Matching was conducted using deterministic linkage based on combinations of the participant's name, sex, date of birth and postcode. Linked HES data are currently available from the start of data availability (see above) until 2017, with data due to be periodically refreshed.

Population statistics

HES statistics relating to the whole population of England were extracted from published Health & Social Care Information Centre and NHS Digital reports (15, 16). Population data for England were obtained from Office for National Statistics mid-year population estimates (17).

Variables

NCDS

In order to explore the extent of selection in terms of response, linkage consent and successful linkage, we considered cohort member's sex, their father's social class and the number of persons per room in their home (a marker of socioeconomic circumstances), all recorded at the birth sweep of data collection. These variables were chosen due to their previously observed associations with response at sweep 8 of NCDS (18). We restricted our attention to key sociodemographic variables observed at birth to avoid issues due to attrition at subsequent sweeps of data collection. These variables were examined in terms of their associations with: i) response at sweep 8 (age 50), when the linkage consents were sought; ii) linkage consent; and iii) linkage to HES data. The social class of the cohort member's father was coded "I/II", "III non-manual", "III manual" or "IV/V" and the number of persons per room in their home was coded "≤ 1", ">1 to 1.5" or "> 1.5".

We explored the quality of the linkage by calculating the percentage of cohort members with linked HES data within strata of NCDS variables which would be expected to be associated with linkage. As NCDS is a long-running study with rich data on cohort members, we were able to consider two types of variable: i) those that are directly comparable to the HES data, where we would expect close correspondence with HES linkage and ii) those that are indirectly comparable to the HES data ("proxy" measures), where we would expect less close correspondence with HES linkage, but where findings nevertheless provide additional evidence with regards to linkage quality.

The directly comparable NCDS variables we considered related to day patient or in-patient attendance and out-patient attendance, both reported at sweep 8 (2008; age 50). Day patient or in-patient attendance was obtained through the question "Since [date of last interview/1 January 2000], have you been in a hospital or clinic as a day patient or in-patient, overnight or longer? Do not include visits for routine, ante-natal or maternity care" and out-patient attendance was obtained through the question "Since [date of last interview/1 January 2000], approximately how many times have you attended a hospital or clinic as an out-patient? Do not include visits to Accident and Emergency. Do not include visits for routine ante-natal or maternity care" (19). The former question was used in its original binary (yes/no) coding and the latter was recoded to a binary (yes/no) variable for analysis. These survey variables conceptually relate closely to the information recorded in HES, so, under the assumption that the NCDS variables capture the intended constructs, if linkage quality is high we would anticipate close correspondence with the HES variables (described below). However, measurement/misclassification error, in particular due to errors in recall, may affect the reliability of the survey data. A further consideration is the exclusion of maternity care from the survey question, as delivery episodes are included in HES APC. Although this is likely to occur only for a small number of participants in this age group, we analyse males and females separately to allow us to examine this potential issue.

The indirectly comparable NCDS variables we considered related to self-rated general health and long-term conditions or illnesses, both observed at sweep 9

(2013; age 55) (similar variables from earlier sweeps could also be considered but using variables from this later sweep maximises the overlap with HES data). Self-rated general health was obtained through the question “In general, would you say your health is excellent, very good, good, fair or poor?" and long-term conditions or illnesses were assessed using the question “Do you have any physical or mental health conditions or illnesses lasting or expected to last 12 months or more?” (20). The original 5-point and binary (yes/no) variables, respectively, were used in the analyses. These survey variables may be less directly related to the information recorded in HES, but we would still expect individuals with lower self-rated general health and those with long-term conditions or illnesses to be more likely to have hospital admissions and outpatient appointments. We would therefore not anticipate such close correspondence with the HES variables, even in the presence of high linkage quality, but moderate correspondence would nevertheless provide additional evidence with regards to linkage quality.

Linked NCDS-HES data

To examine the quality of the linkage we derived variables using the linked HES data which were designed to correspond as closely as possible to the above NCDS survey variables. To compare to the sweep 8 day patient/in-patient and outpatient survey variables we derived binary variables using, respectively, HES APC and OP data which indicated whether cohort members had any HES records over the period between the date of last interview prior to the sweep 8 interview or 1 January 2000 (whichever was later) and the date of sweep 8 interview. Therefore, the precise period under consideration generally differed between individuals.

The sweep 9 self-rated general heath and long-term illness variables are essentially “current” measures rather than relating to a specific retrospective period, though given the nature of the constructs can be reasonably assumed to relate to a period prior to the point of response. Whilst there is therefore no predetermined period over which to derive the corresponding linked HES variables for comparison, we have chosen to do so over the five-year period preceding sweep 9. Specifically, we derived binary variables using, separately, HES APC and OP data which indicated whether cohort members had any HES records over the five-year period prior to the date of sweep 9 interview. The period could therefore again differ between individuals.

We also considered finished admission episodes (FAEs) within the linked HES data for comparison with population data equivalents (see below). Each record in HES APC is a ‘hospital episode’ relating to a period of care for a patient under a single consultant within one hospital provider. A stay in hospital from admission to discharge is called a ‘spell’ and can be made up of one or more episodes of care (16). FAEs are the first episode in a spell of care. The total number of FAEs across all cohort members (noting that each cohort member could potentially contribute more than one FAE) within the linked NCDS-HES APC dataset was identified within each financial year between 1997-1998 and 2015-2016. Data for the financial year 2016-17 were excluded as complete HES data are not yet available in the NCDS-HES linkage.

Population statistics

The number of FAEs are available for 5-year age bands from published reports (15, 16). The number of FAEs within the 5-year age band corresponding to the current age of the NCDS cohort was extracted for each financial year (2004-2005 to 2015-16; unavailable for earlier years) For example, in the financial year 2004-2005 the NCDS participants were age 46 years, so the age 45-49 FAE data were extracted. Office for National Statistics mid-year population estimates for England by single year of age were extracted for the relevant years and aggregated to the same age bands (17). For example, in the financial year 2004-2005 the age 45, 46, 47, 48 and 49 2004 mid-year population estimates were aggregated to obtain the estimated population for the 45-49 year age band. FAE rates per 1000 population for each financial year were then calculated as the ratio of the number of FAEs and the aggregated population in each age band.

Statistical analysis

Linkage rates

To be eligible for linkage we required cohort members to have lived in England at one or more sweeps between sweep 6 (2000, age 42) and sweep 9 (2013, age 55) to align with HES data availability. Although these NCDS sweeps do not cover the entire period of HES data availability (1997-2017), this is as close as can be achieved given NCDS data availability. The place of residence at the date of interview was used, meaning any cohort members who lived in England only in periods between sweeps would not have been deemed eligible for linkage.

Among cohort members eligible for linkage we explored the sequence of events leading up to successful linkage: response at sweep 8 (when health data linkage consents were sought), linkage consent being given (calculating the consent rate) and successful linkage of HES data (calculating the linkage rate).

Under this definition of linkage eligibility there could be a concern that cohort members who lived in England for some but not all of sweeps 6-9 may be less likely to have all their hospital episode data successfully linked (due to the unavailability within HES of hospital episode data from outside England). To explore this issue we first calculated the proportion of sweep between sweep 6 and sweep 9 that each cohort member reported living in England, with sweeps with missing information excluded from the calculations given uncertainly over status. We then calculated the linkage rate for each different proportion among cohort members providing consent for linkage.

Associations of baseline characteristics with response at sweep 8 (age 50), linkage consent and linkage to HES data

Two different analyses were undertaken to explore the extent of selection in terms of response, linkage consent and successful linkage. The first used a sequential approach to consider associations between baseline characteristics (cohort member's sex, the social class of their father and the number of persons per room in their home) and i) response at sweep 8 among cohort members eligible for HES linkage and in the sweep 8 target population (still alive and living in UK at age 50), ii) consent to health record linkage among respondents at sweep 8, and iii) HES linkage among cohort members who had consented. The second analysis used an overall approach to consider separate associations between baseline characteristics and i) response at sweep 8, ii) consent to health record linkage, and iii) HES linkage, all among cohort members eligible for HES linkage and in the sweep 8 target population. Modified Poisson regression (21), returning risk ratios for ease of interpretation and avoiding issues related to the non-collapsibility of the odds ratio, was used to model the associations in both analyses. Unadjusted univariable models are presented because the interest is in simple descriptions of the extent of selection.

Examining linkage quality using internal data

We cross-tabulated i) self-reported day patient/in-patient attendance at age 50 against HES APC linkage, ii) out-patient attendance at age 50 against HES OP linkage, iii) self-rated general heath against both HES APC and OP linkage, and iv) long-term illness against both HES APC and OP linkage. Because there is potential uncertainty about whether linkage consenters without linked HES records truly had no HES record over this period or in fact did have one or more HES records but were missed matches, the cross-tabulations are presented separately within:

a) Individuals with any linked HES APC record ever (n = 4,846) or any linked HES OP record ever (n = 5,783), depending on the analysis.
b) Individuals with any linked HES record ever (i.e. in A&E, APC, CC or OP) (n = 6,119).
c) All HES linkage consenters (n = 6,593).

In a) the analysis is restricted to cohort members with linked HES APC or OP records (as appropriate). However, we would expect that some cohort members will truly not have had such HES records over this period and their exclusion will distort the findings. This analysis sample definition is therefore unlikely to be appropriate, but it is included for comparison. Given that the matching approach was the same across all HES datasets, it may be reasonable to assume that a cohort member with no matched record in (say) HES APC but a matched record in at least one of the other HES datasets truly did not have any records in HES APC (rather than this potentially being a missed match). This assumption corresponds to b), considering all individuals with any linked HES record ever. An alternative is to additionally assume that cohort members with no matched HES records across any HES dataset truly had no records in any HES dataset (rather than this potentially being a missed match in one or more datasets). This would correspond to considering all cohort members who consented to HES linkage, as in c). Cross-tabulations are presented by males and females separately and combined.

Examining linkage representativeness using external data

In the linked NCDS-HES data the rate of FAEs per 1000 individuals in each financial year was calculated using three different denominators, corresponding to the assumptions discussed above:

a) Individuals with any linked HES APC record ever (n = 4,846).
b) Individuals with any linked HES record ever (i.e. in APC, CC, A&E or OP) (n = 6,119).
c) All HES linkage consenters (n = 6,593).

These calculated FAE rates were plotted against financial year alongside the population FAE rates for the corresponding 5-year age band, as described above. A comparison of the three different NCDS-HES rates with the population FAE rate may be suggestive of which of the above assumptions is more plausible.

Results

Linkage rates

The flow of data, from the full sample of NCDS cohort members to the linked samples for each HES dataset, is shown in the data flow diagram in Fig. 1. Of the 10,535 cohort members meeting our definition of linkage eligibility (living in England at one or more sweeps between sweep 6 and sweep 9), 8,403 responded at sweep 8, with 6,593 providing consent for linkage, giving a consent rate of 78.5%. Among these linkage consenters, 6,119 had linked data from one or more of the HES datasets, giving a linkage rate of 92.8%.

Of the 6,953 cohort members who were considered eligible and who gave consent for linkage, 6,450 (92.8%) lived in England for all the sweeps between sweep 6 and sweep 9 at which information was available (Table S1, Supplementary Material). There was a clear pattern of increasing linkage rates with increasing proportion of sweeps in which cohort members lived in England, from 48.0% (12 out of 25) in those living in England at only one of the four sweeps to 93.3% (6,020 out of 6,450) in those living in England at all sweeps (Table S1, Supplementary Material).

Associations of baseline characteristics with response at sweep 8 (age 50), linkage consent and linkage to HES data

In the sequential analysis, there was some evidence that females were more likely to respond at sweep 8 and were more likely to have successfully linked HES data conditional on having consented, but there was no evidence that they were more likely to consent given that they had responded at sweep 8 (Table 1). This resulted in females being more likely to have linked HES data overall (risk ratio 1.03; 95% confidence interval 1.00, 1.06), though there was no association with consent overall (1.00; 0.97, 1.03).

A higher social class of the cohort member’s father (I/II or III non-manual) was associated with higher response at sweep 8 and a lower likelihood of successful linkage given consent, but wasn’t associated with consent conditional on response at sweep 8. This meant that, overall, there was imbalance in terms of consent, with both I/II and III non-manual 8% more likely to give consent relative to IV/V (1.08; 1.03, 1.13 and 1.08; 1.02, 1.15, respectively), but this was somewhat lower in terms of successful linkage (1.04; 0.99, 1.10 and 1.06; 0.99, 1.12).

Number of persons per room followed a similar pattern, with fewer people per room associated with higher response at sweep 8 and a slightly lower likelihood of successful linkage conditional on consent, but less consistent evidence of an association with consent given response at sweep 8. Overall, there were higher consent rates among those with ≤ 1 or > 1 to 1.5 people per room relative to > 1.5 (1.12; 1.06, 1.19 and 1.12; 1.05, 1.19, respectively), and similarly for successful linkage (1.09; 1.03, 1.16 and 1.10; 1.03, 1.18).

Examining linkage quality using internal data

Table 2 shows the cross-tabulations of linked HES APC data and self-reported day patient or in-patient attendance at age 50 in males and females combined. There was a high level of correspondence between the two measures – for example, among all linkage consenters, 86.0% of cohort members who reported no day patient or in-patient attendance had no linked HES APC data and 76.3% of those who reported having day patient or in-patient attendance did have linked HES APC data over the corresponding period. The level of agreement differed somewhat depending on the sample used, with the no-no correspondence highest (86.0%) when considering all linkage consenters and the yes-yes correspondence highest (82.3%) when considering individuals with linked HES APC data only. It should be noted that these patterns (though not the magnitudes) are to be expected: as we move from those with linked HES APC data through those with any linked HES data to all linkage consenters, we are adding exclusively individuals who did not have linked HES APC data, meaning that no-no correspondence must increase and yes-yes correspondence must decrease. The levels of correspondence were similar when males and females were considered separately (Tables S2 and S3, Supplementary Material).

The level of correspondence between linked HES OP data and self-reported out-patient attendance at age 50 was similarly high in cohort members who reported outpatient attendance (for example, 77.6% among all linkage consenters), but lower in those who did not report outpatient attendance (70.5% in the same group) (Table 3). The level of correspondence again differed somewhat depending on the sample used, with the no-no correspondence highest (70.5%) when considering all linkage consenters and the yes-yes correspondence highest when considering individuals with linked HES OP data only (82.3%). There was some evidence of differences in the levels of correspondence between males and females (Tables S4 and S5, Supplementary Material) – for example, among all linkage consenters, 72.5% of males vs. 68.1% of females who reported no out-patient attendance had no linked HES OP data and 73.5% of males vs. 81.0% of females who reported having out-patient attendance did have linked HES OP data over the corresponding period.

Linked HES APC data showed a clear gradient across age 55 self-rated general health groups, from 25.6% in the excellent health group to 73.4% in the poor health group among all linkage consenters (Table 4). The corresponding results for linked HES OP data increased from 52.0% (excellent health) to 92.0% (poor health) (Table 5). Figures were somewhat higher in females than males (Tables S6-S9, Supplementary Material).

Differences in linked HES data were present but less pronounced when considering long-term illness at age 55. Among all linkage consenters, those reporting long-term illness were more likely to have both linked HES APC data (53.6 % vs. 31.7%; Table 6) and linked HES OP data (80.9% vs. 58.9%; Table 7) than those not reporting long-term illness. Figures were again somewhat higher in females than in males (Tables S10-S13, Supplementary Material).

Examining linkage representativeness using external data

The obtained FAE rates in the linked NCDS-HES and population statistics are reported in Table S14 (Supplementary Material) and presented graphically in Fig. 2. FAE rates increased over the time period under consideration in both data sources. The pattern of increase in the linked NCDS-HES data was similar to the population statistics, with the NCDS-HES rates calculated using individuals with any linked HES record ever or, to a lesser extent, using all HES linkage consenters, both corresponding closely to the population rate.

Discussion

In this paper we have provided a thorough evaluation of linkage quality and sample representativeness of the recent NCDS-HES linkage. We observed a clear pattern of increasing linkage rates with increasing proportion of sweeps in which cohort members lived in England. Under an assumption that interaction with hospital services is independent of the proportion of sweeps in which cohort members lived in England, this finding is suggestive that cohort members who lived outside England for part of the period may well have had additional hospital interactions outside England which would not be observed within HES and therefore would be unknown to a researcher using the NCDS-HES data. The implications of this will likely differ on an analysis-by-analysis basis but are unlikely to be serious in most cases given that the vast majority (92.8%) of cohort members who were considered eligible for linkage in fact lived in England for all the sweeps between sweep 6 and sweep 9 at which information was available. If it were of vital importance for an analysis that HES data for the entirety of the period had to be observed for each individual, researchers might consider restricting analyses to cohort members who lived in England for all the sweeps between sweep 6 and sweep 9.

We found that females, those whose father was of a higher social class and those with fewer people per room in their home were associated with a somewhat higher likelihood of successful linkage (though all <10% greater than the respective reference categories). Covariate imbalance (i.e. lack of representativeness with respect to the original sample) in linked cohort-administrative data may be due to one of several reasons, including: selective attrition of the cohort prior to linkage consent being sought, selective consent to data linkage, and selection in the linkage itself (i.e. subpopulations having differential propensity for missed matches or false matches) (5). The NCDS-HES linkage relies on consent given in 2008 (age 50) and previous work in NCDS has identified predictors of response at age 50 (18). The correlates of successful linkage identified in the present analysis are consistent with this previous work, and if there was a similar likelihood of linkage across these groups within consenters, we would expect groups with higher consent rates to have higher linkage rates. Future analyses could consider whether a wider variety of cohort member characteristics are associated with successful linkage. Given these findings, if an analysis of the linked data was intended to be fully representative of the original cohort sample, then researchers may wish to consider additional analytic approaches. For example, they could model the probability of being included in the linked dataset, either within the original NCDS sample or relative to a known population distribution, in order to derive weights to use in inverse probability weighted analyses (22) or use similar variables within a multiple imputation approach (23).

When examining linkage quality using directly comparable survey data we found high levels of correspondence between linked HES APC data and self-reported day patient or in-patient attendance at age 50 and between linked HES OP data and self-reported out-patient attendance at age 50. Linked HES APC and OP data also showed clear gradients across age 55 self-rated general health groups and between those reporting long-term illness and not. Differences between the survey-based measures and HES linkage may be due to linkage errors (missed matches or false matches), but alternative context-specific factors should be considered. It is possible that the scope of the HES APC and OP datasets and the survey questions (or, more specifically, the cohort members' interpretation of them) may not be fully aligned. In particular, the relatively higher percentage of females who reported no out-patient attendance but who did have linked HES OP data (31.9% among all linkage consenters) could be at least partially explained by the exclusion of maternity related appointments in the survey question; this hypothesis is supported by the relatively lower equivalent value among males (27.5%). The self-reported nature of the survey data means that misclassification, particularly when the recall period extends to between four and eight years, is a distinct possibility. To improve comparability, we restricted HES linkages to the period over which the survey questions were asked insofar as this was possible: for sweep 8 variables there was a clearly defined period (from the date of last interview prior to the sweep 8 interview or 1 January 2000 (whichever was later) until the date of sweep 8 interview), but for sweep 9 variables this was an approximation (the five-year period prior to the date of sweep 9 interview), which may have affected the findings. Given these concerns, we believe that the observed high levels of correspondence between HES linkage and the highly comparable survey measures of day patient or in-patient and outpatient attendance are suggestive of high levels of linkage quality. Moreover, the findings for self-rated general health and self-reported long-term illness, whilst not so directly comparably with the levels of HES linkage, provide additional evidence with regards to linkage quality.

We found the rates of FAEs across time to be similar in linked NCDS-HES data to population statistics, with the NCDS-HES rates calculated using individuals with any linked HES record ever or using all HES linkage consenters corresponding closely to the population rate. The FAE population statistics are in 5-year age bands, so in financial years that correspond to NCDS ages which are towards the edge of an age band differences are likely to be greater. However, years towards the middle of age bands (2005, 2010, 2015) are likely to provide a fairer comparison in this respect. Given the number of factors which could potentially impact on the population representativeness of the linked sample, it is encouraging that such high levels of correspondence are observed, indicating a high level of population representativeness.

Although our analyses do not provide a definitive answer to the question of how to handle cohort members who consented to linkage but do not have linked HES records, on the balance of evidence we would tentatively suggest that they should be assumed to truly not have HES records (regardless of whether or not they had matched records in other HES datasets). Both correspondence of HES linkage with directly comparable survey variables and population representativeness remained high under this assumption. Alternative assumptions could be employed in sensitivity analyses. An additional consideration with assuming that a cohort member with no matched record in one HES dataset but a matched record in at least one of the other HES datasets truly did not have any records in the first HES dataset is that, due to the different periods covered by the HES datasets (APC 1997-2017, CC 2009-2017, A&E 2007-2017, OP 2003-2017), the assumption could not be applied in the same way in all datasets. For example, a cohort member who had no matched record in HES CC, A&E and OP but an APC record in 2005 would be assumed to truly have no HES records in CC, A&E and OP, but a cohort member who had no matched record in HES APC, CC and OP but had an A&E visit in 2005 (prior to HES A&E records being available) would be excluded from analyses. This would not be a desirable property.

There are several strengths to this analysis. We were able to identify comparable cohort data and population-representative data with which to compare the linked NCDS-HES data. We utilised a number of different variables and approaches in order to undertake a thorough examination of linkage quality and sample representativeness. We have demonstrated generalisable methods for evaluating linkage quality in the absence of access to the linkage identifiers and in settings with a separation of processes for linkage and analysis.

There are also a number of limitations to the analysis. The sequential analyses of baseline characteristics with linkage consent and successful linkage could possibly be subject to a form of index event (collider) bias due to selection into the analysis sample: linkage consent analyses were only conducted among respondents at sweep 8 and successful linkage analyses were only conducted among cohort members who had consented, so unaccounted common causes of response and consent or of consent and linkage could lead to bias. These findings should therefore be interpreted with some caution. In the analyses considering self-rated general health and long-term illness at age 55, the 5-year look-back period in the HES data is somewhat arbitrary. Only a single external statistic (FAEs per financial year) was used to assess population representativeness due to difficulties in identifying additional directly comparable population-representative external data. In particular, we were only able to compare HES APC data (i.e. not CC, A&E or OP) data to external population-representative data, though linkage quality would be expected to be similar across HES datasets since all linkages were undertaken as part of the same process. This paper has focused on examining the linkage quality and sample representativeness of the linked NCDS-HES data – the potential consequences of linkage error and approaches to examine or address these are beyond the scope of the present paper but have been described elsewhere (4, 7).

Our findings suggest that the linkage quality of the NCDS-HES data is high and that the linked sample maintains an excellent level of population representativeness.

However, we have only investigated a limited characterisation of the linked data and it therefore remains possible that the observed levels of linkage quality and population representativeness are not replicated in other features of the data. Further analyses could be undertaken, though identification of additional comparable cohort variables or population-level information is challenging.

We hope that these analyses will both improve the quality and transparency of research using this linked data resource and encourage providers and users of other linked data resources to undertake and publish similarly thorough evaluations.

Funding

This work was funded by the Economic & Social Research Council (ES/V006037/1) and the Centre for Longitudinal Studies is supported by the Economic & Social Research Council (ES/M001660/1). The funder played no role in study design, in the collection, analysis and interpretation of data, in the writing of the report, or in the decision to submit the article for publication.

References

1. Connelly R, Playford CJ, Gayle V, Dibben C. The role of administrative data in the big data revolution in social science research. Social Science Research. 2016;59:1-12.

2. Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML, et al. Challenges in administrative data linkage for research. Big Data & Society. 2017;4(2):2053951717745678.

3. Gavrielov-Yusim N, Friger M. Use of administrative medical databases in population-based research. Journal of Epidemiology and Community Health. 2014;68(3):283.

4. Harron K, Doidge JC, Goldstein H. Assessing data linkage quality in cohort studies. Annals of Human Biology. 2020;47(2):218-26.

5. Calderwood L, Lessof C. Enhancing Longitudinal Surveys by Linking to Administrative Data. In: Lynn P, editor. Methodology of Longitudinal Surveys. Chichester: Wiley; 2009. p. 55-72.

6. Peycheva D, Ploubidis G, Calderwood L. Determinants of Consent to Administrative Records Linkage in Longitudinal Surveys: Evidence from Next Steps. In: Lynn P, editor. Advances in Longitudinal Survey Methodology. Chichester: John Wiley & Sons; 2021.

7. Doidge JC, Harron KL. Reflections on modern methods: linkage error bias. Int J Epidemiol. 2019;48(6):2050-60.

8. Bohensky MA, Jolley D, Sundararajan V, Evans S, Pilcher DV, Scott I, et al. Data Linkage: A powerful research tool with potential problems. BMC Health Services Research. 2010;10(1):346.

9. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. International Journal of Epidemiology. 2017;46(5):1699-710.

10. Kerry-Barnard S, Gomes D. National Child Development Study: A guide to the linked health administrative datasets – Hospital Episode Statistics (HES). London: UCL Centre for Longitudinal Studies; 2020.

11. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study). Int J Epidemiol. 2006;35(1):34-41.

12. NHS Digital. Hospital Episode Statistics (HES). 2020 [Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics. Accessed 15 May 2020.].

13. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). International Journal of Epidemiology. 2017;46(4):1093-i.

14. University College London, UCL Institute of Education, Centre for Longitudinal Studies, NHS Digital. National Child Development Study: Linked Health Administrative Datasets (Hospital Episode Statistics), England, 1997-2017: Secure Access. [data collection]. UK Data Service. SN: 8697, DOI: 10.5255/UKDA-SN-8697-1. 2021.

15. Health & Social Care Information Centre. Hospital Episode Statistics: Admitted Patient Care, England – 2014-15. [Table 9: Finished Admission Episodes by patient age-group, 2004-05 to 2014-15.]. 2015.

16. NHS Digital. Hospital Admitted Patient Care Activity: 2015-16. [Table 6: FAEs and England population by five year age bands, 2005-06 to 2015-16.]. 2016.

17. Office for National Statistics. Population estimates for the UK and constituent countries by sex and age; historical time series. [Table 11: Population estimates for England, by sex and single year of age, mid-1971 to mid-2019.]. 2019.

18. Mostafa T, Narayanan M, Pongiglione B, Dodgeon B, Goodman A, Silverwood RJ, et al. Missing at random assumption made more plausible: evidence from the 1958 British birth cohort. J Clin Epidemiol. 2021;136:44-54.

19. University of London, Institute of Education, Centre for Longitudinal Studies. National Child Development Study: Age 50, Sweep 8, 2008-2009. [data collection]. 3rd Edition. UK Data Service. SN: 6137, DOI: 10.5255/UKDA-SN-6137-2. 2020.

20. University of London, Institute of Education, Centre for Longitudinal Studies. National Child Development Study: Age 55, Sweep 89 2013. [data collection]. UK Data Service. SN: 7669, DOI: 10.5255/UKDA-SN-7669-1. 2020.

21. Zou G. A modified poisson regression approach to prospective studies with binary data. Am J Epidemiol. 2004;159(7):702-6.

22. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res. 2013;22(3):278-95.

23. Carpenter JR, Kenward MG. Multiple Imputation and its Application. Chichester, UK: John Wiley & Sons, Ltd; 2013.

Tables & Figures

Table 1. Estimated unadjusted associations with response at age 50, consent to health record linkage and Hospital Episode Statistics (HES) linkage among cohort members eligible for HES linkage (lived in England at one or more waves between wave 6 and wave 9) and in the wave 8 target population (still alive and living in UK at age 50) (n = 10,355).

		Sequential									Overall					
		Response at wave 8 (age 50)			Consent to linkage			Linked HES data			Consent to linkage			Linked HES data		
	N (%)	n (%)	RR	95% CI	n (%)	RR	95% CI	n (%)	RR	95% CI	n (%)	RR	95% CI	n (%)	RR	95% CI
Sex (N = 10,355)																
Male	5,137 (49.6)	4,136 (80.5)	1.00	(ref)	3,270 (79.1)	1.00	(ref)	2,990 (91.4)	1.00	(ref)	3,270 (63.7)	1.00	(ref)	2,990 (58.2)	1.00	(ref)
Female	5,218 (50.4)	4,267 (81.8)	1.02	1.00, 1.03	3,323 (77.9)	0.99	0.96, 1.01	3,129 (94.2)	1.03	1.02, 1.04	3,323 (63.7)	1.00	0.97, 1.03	3,129 (60.0)	1.03	1.00, 1.06
Social class of father (N = 9,276)																
I/II	1,739 (18.8)	1,485 (85.4)	1.08	1.05, 1.11	1,164 (78.4)	1.00	0.96, 1.04	1,058 (90.9)	0.96	0.94, 0.99	1,164 (66.9)	1.08	1.03, 1.13	1,058 (60.8)	1.04	0.99, 1.10
III non-manual	964 (10.4)	799 (82.9)	1.05	1.01, 1.09	648 (81.1)	1.04	0.99, 1.08	595 (91.8)	0.97	0.95, 1.00	648 (67.2)	1.08	1.02, 1.15	595 (61.7)	1.06	0.99, 1.12
III manual	4,711 (50.8)	3,789 (80.4)	1.02	0.99, 1.04	2,967 (78.3)	1.00	0.97, 1.03	2,783 (93.8)	1.00	0.98, 1.01	2,967 (63.0)	1.02	0.97, 1.06	2,783 (59.1)	1.02	0.97, 1.06
IV/V	1,862 (20.1)	1,475 (79.2)	1.00	(ref)	1,155 (78.3)	1.00	(ref)	1,088 (94.2)	1.00	(ref)	1,155 (62.0)	1.00	(ref)	1,088 (58.4)	1.00	(ref)

Number of persons per room (N = 9,486)																
≤ 1	6,894 (72.7)	5,698 (82.7)	1.10	1.06, 1.14	4,464 (78.3)	1.02	0.98, 1.06	4,129 (92.5)	0.97	0.95, 0.99	4,464 (64.8)	1.12	1.06, 1.19	4,129 (59.9)	1.09	1.03, 1.16
> 1 to 1.5	1,554 (16.4)	1,230 (79.2)	1.06	1.01, 1.10	999 (81.2)	1.06	1.01, 1.11	937 (93.8)	0.98	0.96, 1.01	999 (64.3)	1.12	1.05, 1.19	937 (60.3)	1.10	1.03, 1.18
> 1.5	1,038 (10.9)	778 (75.0)	1.00	(ref)	598 (76.9)	1.00	(ref)	570 (95.3)	1.00	(ref)	598 (57.6)	1.00	(ref)	570 (54.9)	1.00	(ref)

Table 2. Linked Hospital Episode Statistics (HES) Admitted Patient Care (APC) data between date of last interview/2000 and date of wave 8 (age 50) interview against self-reported day patient or in-patient attendance at wave 8 (age 50) in the National Child Development Study (NCDS): males and females combined.

		Linked HES APC data between date of last interview/2000 and date of wave 8 (age 50) interview								
		Individuals with linked APC data			Individuals with at least one linked HES dataset			All linkage consenters		
		No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total
Age 50 day patient or in-patient attendance	No	2,441 (78.8)	657 (21.2)	3,098	3,615 (84.6)	657 (15.4)	4,272	4,050 (86.0)	657 (14.0)	4,707
	Yes	309 (17.7)	1,438 (82.3)	1,747	408 (22.1)	1,438 (77.9)	1,846	447 (23.7)	1,438 (76.3)	1,885
	Total	2,750 (56.8)	2,095 (43.2)	4,845	4,023 (65.8)	2,095 (34.2)	6,118	4,497 (68.2)	2,095 (31.8)	6,592

Table 3. Linked Hospital Episode Statistics (HES) Outpatient (OP) data between date of last interview/2000 and date of wave 8 (age 50) interview against self-reported out-patient attendance at wave 8 (age 50) in the National Child Development Study (NCDS): males and females combined.

		Linked HES OP data between date of last interview/2000 and date of wave 8 (age 50) interview								
		Individuals with linked OP data			Individuals with at least one linked HES dataset			All linkage consenters		
		No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total
Age 50 out-patient attendance	No	1,422 (62.8)	843 (37.2)	2,265	1,663 (66.4)	843 (33.6)	2,506	2,018 (70.5)	843 (29.5)	2,861
	Yes	622 (17.7)	2,896 (82.3)	3,518	717 (19.9)	2,896 (80.1)	3,613	836 (22.4)	2,896 (77.6)	3,732
	Total	2,044 (35.3)	3,739 (64.7)	5,783	2,380 (38.9)	3,739 (61.1)	6,119	2,854 (43.3)	3,739 (56.7)	6,593

Table 4. Linked Hospital Episode Statistics (HES) Admitted Patient Care (APC) data for the 5 years prior to the date of the wave 9 (age 55) interview against self-rated general health at wave 9 (age 55) in the National Child Development Study (NCDS): males and females combined.

		Linked HES APC data for the 5 years prior to the date of the wave 9 (age 55) interview								
		Individuals with linked APC data			Individuals with at least one linked HES dataset			All linkage consenters		
		No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total
Age 55 self-rated general health	Excellent	282	195	477	482	195	677	566	195	761
		(59.1)	(40.9)		(71.2)	(28.8)		(74.4)	(25.6)	
	Very good	736	599	1,335	1,202	509	1,801	1,383	509	1,982
		(55.1)	(44.9)		(66.7)	(33.3)		(69.8)	(30.2)	
	Good	625	757	1,382	989	757	1,746	1,102	757	1,859
		(45.2)	(54.8)		(56.6)	(43.4)		(59.3)	(40.7)	
	Fair	247	441	688	336	441	777	363	441	804
		(35.9)	(64.1)		(43.2)	(56.8)		(45.2)	(54.8)	
	Poor	68	248	316	84	248	332	90	248	338
		(21.5)	(78.5)		(25.3)	(74.7)		(26.6)	(73.4)	
	Total	1,958	2,240	4,198	3,093	2,240	5,333	3,504	2,240	5,744
		(46.6)	(53.4)		(58.0)	(42.0)		(61.0)	(39.0)	

Table 5. Linked Hospital Episode Statistics (HES) Outpatient (OP) data for the 5 years prior to the date of the wave 9 (age 55) interview against self-rated general health at wave 9 (age 55) in the National Child Development Study (NCDS): males and females combined.

		Linked HES OP data for the 5 years prior to the date of the wave 9 (age 55) interview								
		Individuals with linked OP data			Individuals with at least one linked HES dataset			All linkage consenters		
		No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total
Age 55 self-rated general health	Excellent	211	396	607	281	396	677	365	396	
		(34.8)	(65.2)		(41.5)	(58.5)		(48.0)	(52.0)	
	Very good	523	1,163	1,686	638	1,163	1,801	819	1,163	
		(31.0)	(69.0)		(35.4)	(64.6)		(41.3)	(58.7)	
	Good	380	1,276	1,656	470	1,276	1,746	583	1,276	
		(23.0)	(77.0)		(26.9)	(73.1)		(31.4)	(68.6)	
	Fair	103	659	762	118	659	777	145	659	
		(13.5)	(86.5)		(15.2)	(84.8)		(18.0)	(82.0)	
	Poor	20	311	331	21	311	332	27	311	
		(6.0)	(94.0)		(6.3)	(93.7)		(8.0)	(92.0)	
	Total	1,237	3,805	5,042	1,528	3,805	5,333	1,939	3,805	
		(24.5)	(75.5)		(28.7)	(71.3)		(33.8)	(66.2)	

Table 6. Linked Hospital Episode Statistics (HES) Admitted Patient Care (APC) data for the 5 years prior to the date of the wave 9 (age 55) interview against self-reported long-term illness at wave 9 (age 55) in the National Child Development Study (NCDS): males and females combined.

		Linked HES APC data for the 5 years prior to the date of the wave 9 (age 55) interview								
		Individuals with linked APC data			Individuals with at least one linked HES dataset			All linkage consenters		
		No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total
Age 55 long-term illness	No	1,382	1,217	2,599	2,264	1,217	3,481	2,616	1,217	3,833
		(53.2)	(46.8)		(65.0)	(35.0)		(68.3)	(31.7)	
	Yes	572	1,018	1,590	822	1,018	1,840	881	1,018	1,899
		(36.0)	(64.0)		(44.7)	(55.3)		(46.4)	(53.6)	
	Total	1,954	2,235	4,189	3,086	2,235	5,321	3,497	2,235	5,732
		(46.7)	(53.3)		(58.0)	(42.0)		(61.0)	(39.0)	

Table 7. Linked Hospital Episode Statistics (HES) Outpatient (OP) data for the 5 years prior to the date of the wave 9 (age 55) interview against self-reported long-term illness at wave 9 (age 55) in the National Child Development Study (NCDS): males and females combined.

		Linked HES OP data for the 5 years prior to the date of the wave 9 (age 55) interview								
		Individuals with linked OP data			Individuals with at least one linked HES dataset			All linkage consenters		
		No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total
Age 55 long-term illness	No	972 (30.1)	2,258 (69.9)	3,230	1,223 (35.1)	2,258 (64.9)	3,481	1,575 (41.1)	2,258 (58.9)	3,833
	Yes	265 (14.7)	1,536 (85.3)	1,801	304 (16.5)	1,536 (83.5)	1,840	363 (19.1)	1,536 (80.9)	1,899
	Total	1,237 (24.6)	3,794 (75.4)	5,031	1,527 (28.7)	3,794 (71.3)	5,321	1,938 (33.8)	3,794 (66.2)	5,732

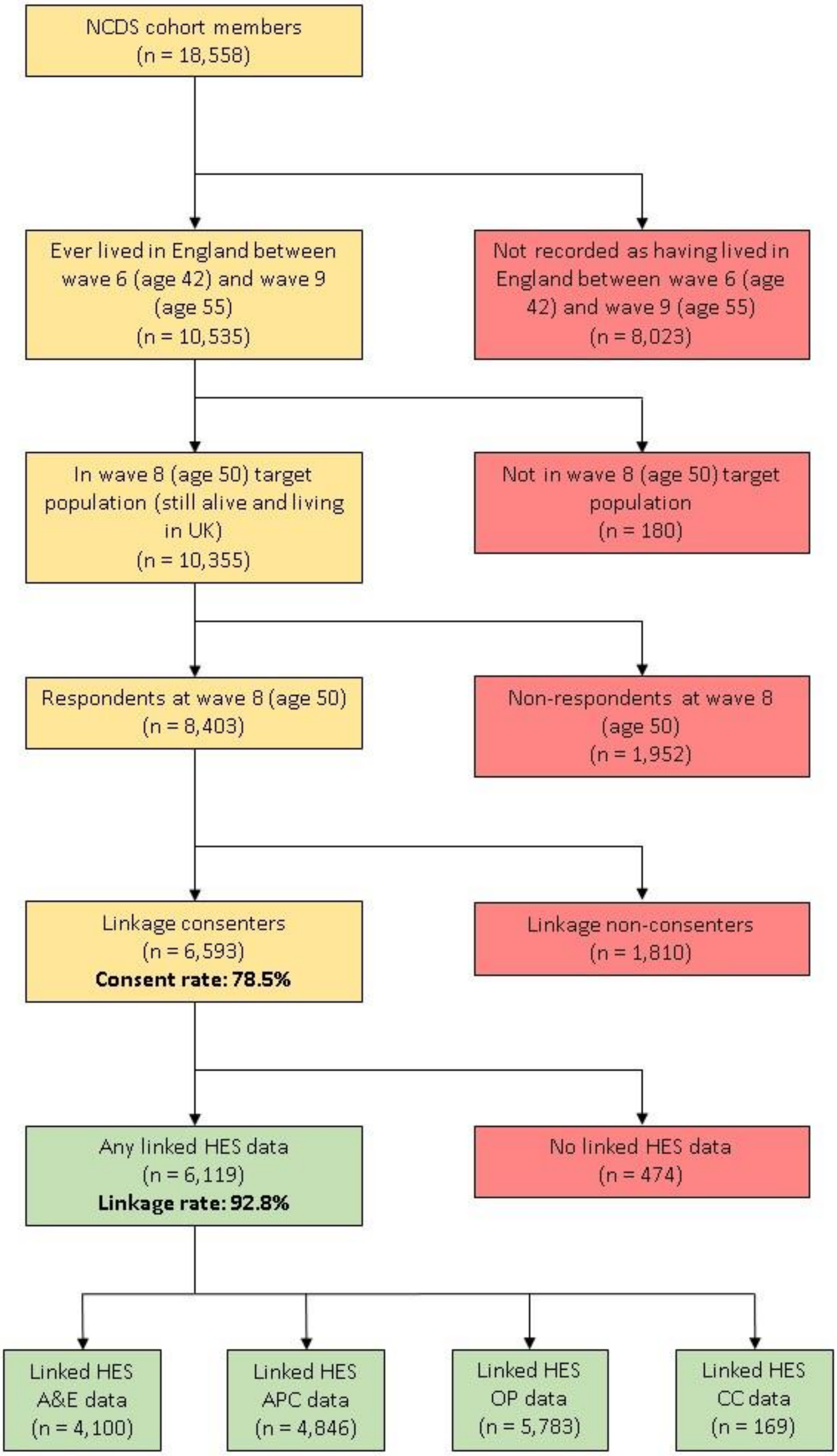


Fig. 1. Flow diagram showing National Child Development Study (NCDS)-Hospital Episode Statistics (HES) data linkage and data availability.

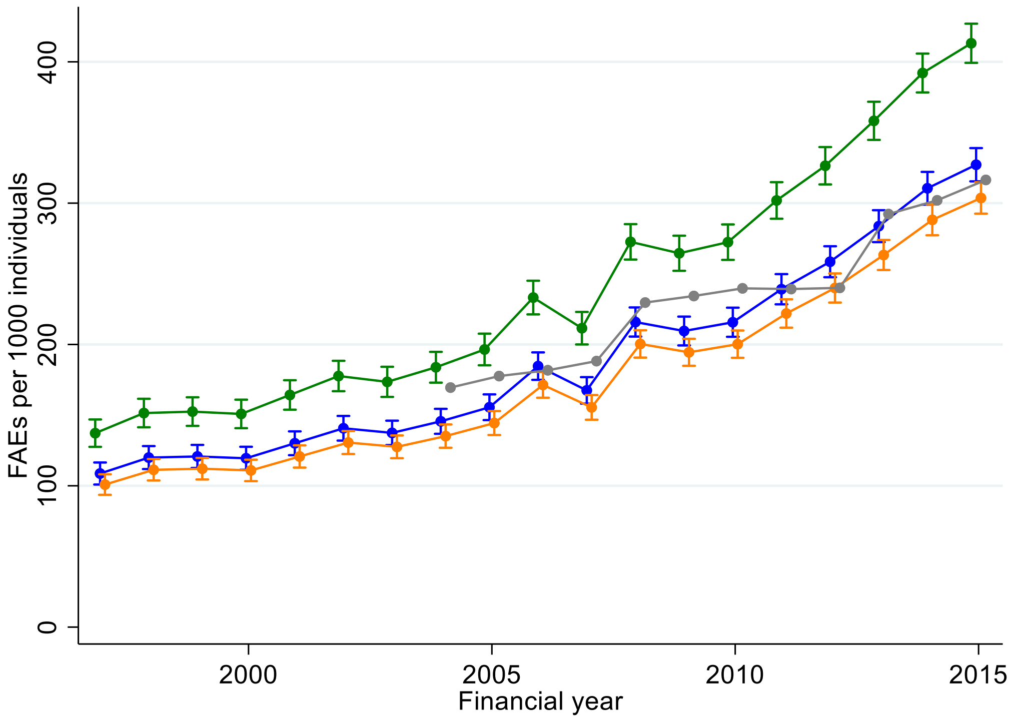


Fig. 2. Linked National Child Development Study (NCDS)-Hospital Episode Statistics (HES) Admitted Patient Care (APC) and population (HES APC) finished admission episode (FAE) rates and 95% confidence intervals by financial year. Green: rate using linked NCDS-HES data among those with any linked HES APC record ever; blue: rate using linked NCDS-HES data among those with any linked HES record ever; orange: rate using linked NCDS-HES data among all HES linkage consenters; grey: rate using whole population HES data.

Supplementary Material

Contents

Table S1. Number (%) of 1958 National Child Development Study (NCDS) cohort members with linked Hospital Episode Statistics (HES) data by proportion of waves between 6 and 9 that they lived in England. Waves with missing data on residency are excluded from the proportion calculation [A]. Analysis restricted to cohort members providing consent for linkage.

	Proportion of waves between 6 and 9 living in England						
	1/4	1/3	1/2	2/3	3/4	1	Total
No linked HES data	13	4	13	3	11	430	474
	(52.0)	(40.0)	(36.1)	(27.3)	(18.0)	(6.7)	(7.2)
Linked HES data	12	6	23	8	50	6,020	6,119
	(48.0)	(60.0)	(63.9)	(72.7)	(82.0)	(93.3)	(92.8)
Total	25	10	36	11	61	6,450	6,593

[A] So, for example, a proportion of "1/3" means 1 wave living in England, 2 waves not living in England and 1 wave with unknown residency and a proportion of "2/4" means either 2 waves living in England and 2 waves not living in England or 1 wave living in England, 1 wave not living in England and 2 waves with unknown residency.

Table S2. Linked Hospital Episode Statistics (HES) Admitted Patient Care (APC) data between date of last interview/2000 and date of wave 8 (age 50) interview against self-reported day patient or in-patient attendance at wave 8 (age 50) in the National Child Development Study (NCDS): males.

		Linked HES APC data between date of last interview/2000 and date of wave 8 (age 50) interview								
		Individuals with linked APC data			Individuals with at least one linked HES dataset			All linkage consenters		
		No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total
Age 50 day patient or in-patient attendance	No	1,189	322	1,511	1,853	322	2,175	2,111	322	2,433
		(78.7)	(21.3)		(85.2)	(14.8)		(86.8)	(13.2)	
	Yes	143	623	766	192	623	815	214	623	837
		(18.7)	(81.3)		(23.6)	(76.4)		(25.6)	(74.4)	
	Total	1,332	945	2,277	2,045	945	2,990	2,325	945	3,270
		(58.5)	(41.5)		(68.4)	(31.6)		(71.1)	(28.9)	

Table S3. Linked Hospital Episode Statistics (HES) Admitted Patient Care (APC) data between date of last interview/2000 and date of wave 8 (age 50) interview against self-reported day patient or in-patient attendance at wave 8 (age 50) in the National Child Development Study (NCDS): females.

		Linked HES APC data between date of last interview/2000 and date of wave 8 (age 50) interview								
		Individuals with linked APC data			Individuals with at least one linked HES dataset			All linkage consenters		
		No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total
Age 50 day patient or in-patient attendance	No	1,252	335	1,587	1,762	335	2,097	1,939	335	2,274
		(78.9)	(21.1)		(84.0)	(16.0)		(85.3)	(14.7)	
	Yes	166	815	981	216	815	1,031	233	815	1,048
		(16.9)	(83.1)		(21.0)	(79.0)		(22.2)	(77.8)	
	Total	1,418	1,150	2,568	1,978	1,150	3,128	2,172	1,150	3,322
		(55.2)	(44.8)		(63.2)	(36.8)		(65.4)	(34.6)	

Table S4. Linked Hospital Episode Statistics (HES) Outpatient (OP) data between date of last interview/2000 and date of wave 8 (age 50) interview against self-reported out-patient attendance at wave 8 (age 50) in the National Child Development Study (NCDS): males.

		Linked HES OP data between date of last interview/2000 and date of wave 8 (age 50) interview								
		Individuals with linked OP data			Individuals with at least one linked HES dataset			All linkage consenters		
		No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total
Age 50 out-patient attendance	No	771 (64.0)	434 (36.0)	1,205	921 (68.0)	434 (32.0)	1,355	1,146 (72.5)	434 (27.5)	1,580
	Yes	337 (21.3)	1,242 (78.7)	1,579	393 (24.0)	1,242 (76.0)	1,635	448 (26.5)	1,242 (73.5)	1,690
	Total	1,108 (39.8)	1,676 (60.2)	2,784	1,314 (44.0)	1,676 (56.0)	2,990	1,594 (48.8)	1,676 (51.2)	3,270

Table S5. Linked Hospital Episode Statistics (HES) Outpatient (OP) data between date of last interview/2000 and date of wave 8 (age 50) interview against self-reported out-patient attendance at wave 8 (age 50) in the National Child Development Study (NCDS): females.

		Linked HES OP data between date of last interview/2000 and date of wave 8 (age 50) interview								
		Individuals with linked OP data			Individuals with at least one linked HES dataset			All linkage consenters		
		No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total
Age 50 out-patient attendance	No	651 (61.4)	409 (38.6)	1,060	742 (64.5)	409 (35.5)	1,151	872 (68.1)	409 (31.9)	1,281
	Yes	285 (14.7)	1,654 (85.3)	1,939	324 (16.4)	1,654 (83.6)	1,978	388 (19.0)	1,654 (81.0)	2,042
	Total	936 (31.2)	2,063 (68.8)	2,999	1,066 (34.1)	2,063 (65.9)	3,129	1,260 (37.9)	2,063 (62.1)	3,323

Table S6. Linked Hospital Episode Statistics (HES) Admitted Patient Care (APC) data for the 5 years prior to the date of the wave 9 (age 55) interview against self-rated general health at age 55 in the National Child Development Study (NCDS): males.

		Linked HES APC data for the 5 years prior to the date of the wave 9 (age 55) interview								
		Individuals with linked APC data			Individuals with at least one linked HES dataset			All linkage consenters		
		No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total
Age 55 self-rated general health	Excellent	124	90	214	237	90	327	293	90	383
		(57.9)	(42.1)		(72.5)	(27.5)		(76.5)	(23.5)	
	Very good	346	271	617	602	271	873	706	271	977
		(56.1)	(43.9)		(69.0)	(31.0)		(72.3)	(27.7)	
	Good	297	359	656	490	359	849	549	359	908
		(45.3)	(54.7)		(57.7)	(42.3)		(60.5)	(39.5)	
	Fair	127	201	328	180	201	381	196	201	397
		(38.7)	(61.3)		(47.2)	(52.8)		(49.4)	(50.6)	
	Poor	32	105	137	44	105	149	47	105	152
		(23.4)	(76.6)		(29.5)	(70.5)		(30.9)	(69.1)	
	Total	926	1,026	1,952	1,553	1,026	2,579	1,791	1,026	2,817
		(47.4)	(52.6)		(60.2)	(39.8)		(63.6)	(36.4)	

Table S7. Linked Hospital Episode Statistics (HES) Admitted Patient Care (APC) data for the 5 years prior to the date of the wave 9 (age 55) interview against self-rated general health at wave 9 (age 55) in the National Child Development Study (NCDS): females.

		Linked HES APC data for the 5 years prior to the date of the wave 9 (age 55) interview								
		Individuals with linked APC data			Individuals with at least one linked HES dataset			All linkage consenters		
		No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total
Age 55 self-rated general health	Excellent	158	105	263	245	105	350	273	105	378
		(60.1)	(39.9)		(70.0)	(30.0)		(72.2)	(27.8)	
	Very good	390	328	718	600	328	928	677	328	1,005
		(54.3)	(45.7)		(64.7)	(35.3)		(67.4)	(32.6)	
	Good	328	398	726	499	398	897	553	398	951
		(45.2)	(54.8)		(55.6)	(44.4)		(58.2)	(41.8)	
	Fair	120	240	360	156	240	396	167	240	407
		(33.3)	(66.7)		(39.4)	(60.6)		(41.0)	(59.0)	
	Poor	36	143	179	40	143	183	43	143	186
		(20.1)	(79.9)		(21.9)	(78.1)		(23.1)	(76.9)	
	Total	1,032	1,214	2,246	1,540	1,214	2,754	1,713	1,214	2,927
		(46.0)	(54.0)		(55.9)	(44.1)		(58.5)	(41.5)	

Table S8. Linked Hospital Episode Statistics (HES) Outpatient (OP) data for the 5 years prior to the date of the wave 9 (age 55) interview against self-rated general health at wave 9 (age 55) in the National Child Development Study (NCDS): males.

		Linked HES OP data for the 5 years prior to the date of the wave 9 (age 55) interview								
		Individuals with linked OP data			Individuals with at least one linked HES dataset			All linkage consenters		
		No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total
Age 55 self-rated general health	Excellent	103 (35.9)	184 (64.1)	287	143 (43.7)	184 (56.3)	327	199 (52.0)	184 (48.0)	383
	Very good	284 (35.2)	523 (64.8)	807	350 (40.1)	523 (59.9)	873	454 (46.5)	523 (53.5)	977
	Good	210 (26.5)	583 (73.5)	793	266 (31.3)	583 (68.7)	849	325 (35.8)	583 (64.2)	908
	Fair	62 (16.8)	308 (83.2)	370	73 (19.2)	308 (80.8)	381	89 (22.4)	308 (77.6)	397
	Poor	12 (8.1)	136 (91.9)	148	13 (8.7)	136 (91.3)	149	16 (10.5)	136 (89.5)	152
	Total	671 (27.9)	1,734 (72.1)	2,405	845 (32.8)	1,734 (67.2)	2,579	1,083 (38.5)	1,734 (61.5)	2,817

Table S9. Linked Hospital Episode Statistics (HES) Outpatient (OP) data for the 5 years prior to the date of the wave 9 (age 55) interview against self-rated general health at wave 9 (age 55) in the National Child Development Study (NCDS): females.

		Linked HES OP data for the 5 years prior to the date of the wave 9 (age 55) interview								
		Individuals with linked OP data			Individuals with at least one linked HES dataset			All linkage consenters		
		No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total
Age 55 self-rated general health	Excellent	108	212	320	138	212	350	166	212	378
		(33.8)	(66.2)		(39.4)	(60.6)		(43.9)	(56.1)	
	Very good	239	640	879	288	640	928	365	640	1,005
		(27.2)	(72.8)		(31.0)	(69.0)		(36.3)	(63.7)	
	Good	170	693	863	204	693	897	258	693	951
		(19.7)	(80.3)		(22.7)	(77.3)		(27.1)	(72.9)	
	Fair	41	351	392	45	351	396	56	351	407
		(10.5)	(89.5)		(11.4)	(88.6)		(13.8)	(86.2)	
	Poor	8	175	183	8	175	183	11	175	186
		(4.4)	(95.6)		(4.4)	(95.6)		(5.9)	(94.1)	
	Total	566	2,071	2,637	683	2,071	2,754	856	2,071	2,927
		(21.5)	(78.5)		(24.8)	(75.2)		(29.2)	(70.8)	

Table S10. Linked Hospital Episode Statistics (HES) Admitted Patient Care (APC) data for the 5 years prior to the date of the wave 9 (age 55) interview against self-reported long-term illness at wave 9 (age 55) in the National Child Development Study (NCDS): males.

		Linked HES APC data for the 5 years prior to the date of the wave 9 (age 55) interview								
		Individuals with linked APC data			Individuals with at least one linked HES dataset			All linkage consenters		
		No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total
Age 55 long-term illness	No	642 (52.4)	583 (47.6)	1,225	1,124 (65.9)	583 (34.1)	1,707	1,332 (69.6)	583 (30.4)	1,915
	Yes	282 (39.1)	439 (60.9)	721	425 (49.2)	439 (50.8)	864	455 (50.9)	439 (49.1)	894
	Total	924 (47.5)	1,022 (52.5)	1,946	1,549 (60.3)	1,022 (39.7)	2,571	1,787 (63.6)	1,022 (36.4)	2,809

Table S11. Linked Hospital Episode Statistics (HES) Admitted Patient Care (APC) data for the 5 years prior to the date of the wave 9 (age 55) interview against self-reported long-term illness at wave 9 (age 55) in the National Child Development Study (NCDS): females.

		Linked HES APC data for the 5 years prior to the date of the wave 9 (age 55) interview								
		Individuals with linked APC data			Individuals with at least one linked HES dataset			All linkage consenters		
		No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total
Age 55 long-term illness	No	740 (53.9)	634 (46.1)	1,374	1,140 (64.3)	634 (35.7)	1,774	1,284 (66.9)	634 (33.1)	1,918
	Yes	290 (33.4)	579 (66.6)	869	397 (40.7)	579 (59.3)	976	426 (42.4)	579 (57.6)	1,005
	Total	1,030 (45.9)	1,213 (54.1)	2,243	1,537 (55.9)	1,213 (44.1)	2,750	1,710 (58.5)	1,213 (41.5)	2,923

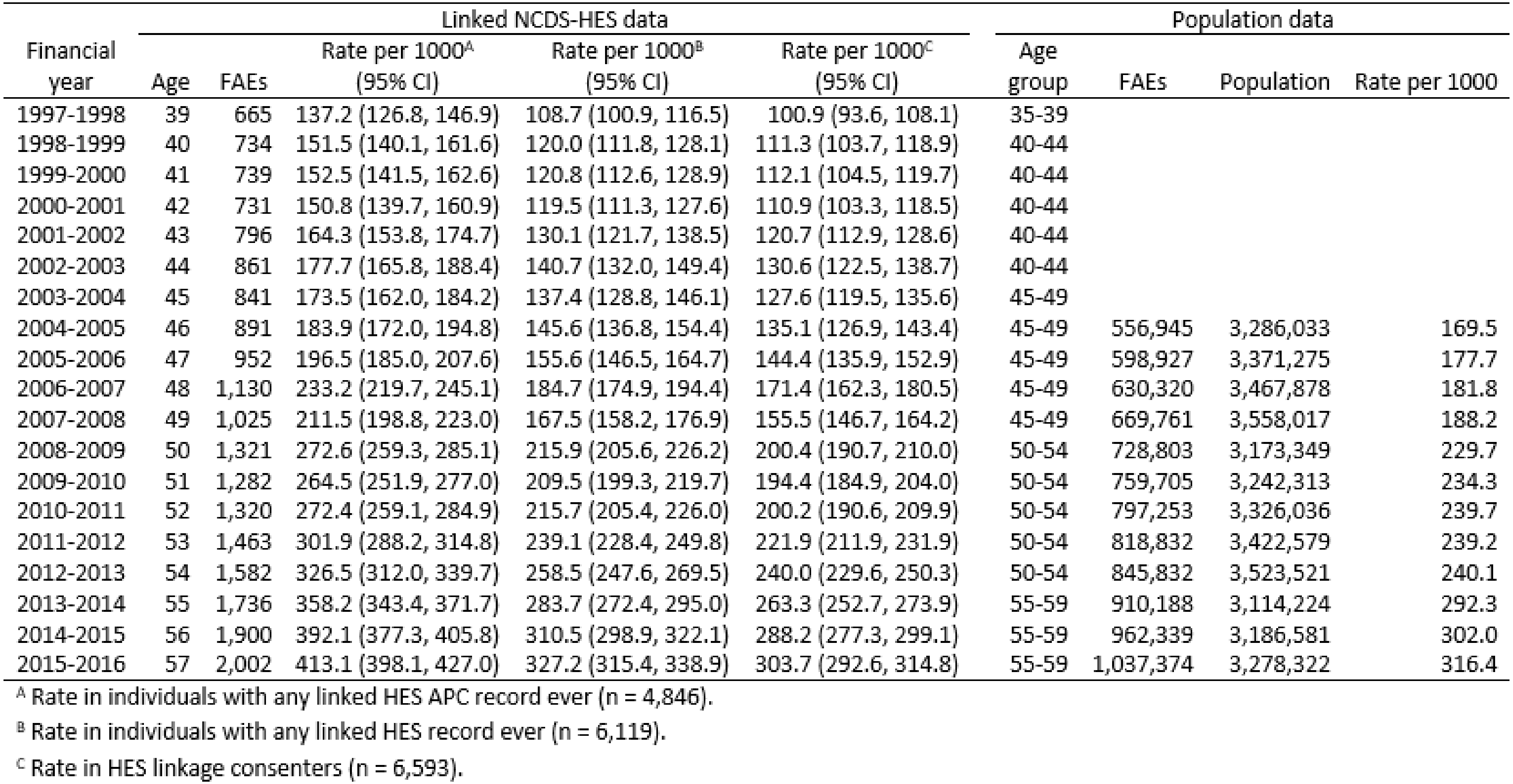
Table S12. Linked Hospital Episode Statistics (HES) Outpatient (OP) data for the 5 years prior to the date of the wave 9 (age 55) interview against self-reported long-term illness at wave 9 (age 55) in the National Child Development Study (NCDS): males.

		Linked HES OP data for the 5 years prior to the date of the wave 9 (age 55) interview								
		Individuals with linked OP data			Individuals with at least one linked HES dataset			All linkage consenters		
		No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total
Age 55 long-term illness	No	526 (33.8)	1,031 (66.2)	1,557	676 (39.6)	1,031 (60.4)	1,707	884 (46.2)	1,031 (53.8)	1,915
	Yes	145 (17.3)	695 (82.7)	840	169 (19.6)	695 (80.4)	864	199 (22.3)	695 (77.7)	894
	Total	671 (28.0)	1,726 (72.0)	2,397	845 (32.9)	1,726 (67.1)	2,571	1,083 (38.6)	1,726 (61.4)	2,809

Table S13. Linked Hospital Episode Statistics (HES) Outpatient (OP) data for the 5 years prior to the date of the wave 9 (age 55) interview against self-reported long-term illness at wave 9 (age 55) in the National Child Development Study (NCDS): females.

		Linked HES OP data for the 5 years prior to the date of the wave 9 (age 55) interview								
		Individuals with linked OP data			Individuals with at least one linked HES dataset			All linkage consenters		
		No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total	No linked HES data	Linked HES data	Total
Age 55 long-term illness	No	446 (26.7)	1,227 (73.3)	1,673	547 (30.8)	1,227 (69.2)	1,774	691 (36.0)	1,227 (64.0)	1,918
	Yes	120 (12.5)	841 (87.5)	961	135 (13.8)	841 (86.2)	976	164 (16.3)	841 (83.7)	1,005
	Total	566 (21.5)	2,068 (78.5)	2,634	682 (24.8)	2,068 (75.2)	2,750	855 (29.3)	2,068 (70.7)	2,923

Table S14. Linked National Child Development Study (NCDS)-Hospital Episode Statistics (HES) Admitted Patient Care (APC) and population (HES APC) finished admission episode (FAE) data by financial year.



	Linked NCDS-HES data					Population data			
Financial year	Age	FAEs	Rate per 1000[A] (95% CI)	Rate per 1000[B] (95% CI)	Rate per 1000[C] (95% CI)	Age group	FAEs	Population	Rate per 1000
1997-1998	39	665	137.2 (126.8, 146.9)	108.7 (100.9, 116.5)	100.9 (93.6, 108.1)	35-39			
1998-1999	40	734	151.5 (140.1, 161.6)	120.0 (111.8, 128.1)	111.3 (103.7, 118.9)	40-44			
1999-2000	41	739	152.5 (141.5, 162.6)	120.8 (112.6, 128.9)	112.1 (104.5, 119.7)	40-44			
2000-2001	42	731	150.8 (139.7, 160.9)	119.5 (111.3, 127.6)	110.9 (103.3, 118.5)	40-44			
2001-2002	43	796	164.3 (153.8, 174.7)	130.1 (121.7, 138.5)	120.7 (112.9, 128.6)	40-44			
2002-2003	44	861	177.7 (165.8, 188.4)	140.7 (132.0, 149.4)	130.6 (122.5, 138.7)	40-44			
2003-2004	45	841	173.5 (162.0, 184.2)	137.4 (128.8, 146.1)	127.6 (119.5, 135.6)	45-49			
2004-2005	46	891	183.9 (172.0, 194.8)	145.6 (136.8, 154.4)	135.1 (126.9, 143.4)	45-49	556,945	3,286,033	169.5
2005-2006	47	952	196.5 (185.0, 207.6)	155.6 (146.5, 164.7)	144.4 (135.9, 152.9)	45-49	598,927	3,371,275	177.7
2006-2007	48	1,130	233.2 (219.7, 245.1)	184.7 (174.9, 194.4)	171.4 (162.3, 180.5)	45-49	630,320	3,467,878	181.8
2007-2008	49	1,025	211.5 (198.8, 223.0)	167.5 (158.2, 176.9)	155.5 (146.7, 164.2)	45-49	669,761	3,558,017	188.2
2008-2009	50	1,321	272.6 (259.3, 285.1)	215.9 (205.6, 226.2)	200.4 (190.7, 210.0)	50-54	728,803	3,173,349	229.7
2009-2010	51	1,282	264.5 (251.9, 277.0)	209.5 (199.3, 219.7)	194.4 (184.9, 204.0)	50-54	759,705	3,242,313	234.3
2010-2011	52	1,320	272.4 (259.1, 284.9)	215.7 (205.4, 226.0)	200.2 (190.6, 209.9)	50-54	797,253	3,326,036	239.7
2011-2012	53	1,463	301.9 (288.2, 314.8)	239.1 (228.4, 249.8)	221.9 (211.9, 231.9)	50-54	818,832	3,422,579	239.2
2012-2013	54	1,582	326.5 (312.0, 339.7)	258.5 (247.6, 269.5)	240.0 (229.6, 250.3)	50-54	845,832	3,523,521	240.1
2013-2014	55	1,736	358.2 (343.4, 371.7)	283.7 (272.4, 295.0)	263.3 (252.7, 273.9)	55-59	910,188	3,114,224	292.3
2014-2015	56	1,900	392.1 (377.3, 405.8)	310.5 (298.9, 322.1)	288.2 (277.3, 299.1)	55-59	962,339	3,186,581	302.0
2015-2016	57	2,002	413.1 (398.1, 427.0)	327.2 (315.4, 338.9)	303.7 (292.6, 314.8)	55-59	1,037,374	3,278,322	316.4

[A] Rate in individuals with any linked HES APC record ever (n = 4,846).

[B] Rate in individuals with any linked HES record ever (n = 6,119).

[C] Rate in HES linkage consenters (n = 6,593).