International Journal of Population Data Science

Journal Website: www.ijpds.org

Examining the quality and population representativeness of linked survey and administrative data: guidance and illustration using linked 1958 National Child Development Study and Hospital Episode Statistics data

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

Submission History

Submitted:	12/01/2023
Accepted:	31/10/2023
Published:	xx/xx/20xx

[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

Abstract

Introduction

Recent years have seen an increase in linkages between survey and administrative data. It is important to evaluate the quality of such data linkages to discern the likely reliability of ensuing research. Evaluation of linkage quality and bias can be conducted using different approaches, but many of these are not possible when there is a separation of processes for linkage and analysis to help preserve privacy, as is typically the case in the UK (and elsewhere).

Objectives

We aimed to describe a suite of generalisable methods to evaluate linkage quality and population representativeness of linked survey and administrative data which remain tractable when users of the linked data are not party to the linkage process itself. We emphasise issues particular to longitudinal survey data throughout.

Methods

Our proposed approaches cover several areas: i) Linkage rates, ii) Selection into response, linkage consent and successful linkage, iii) Linkage quality, and iv) Linked data population representativeness. We illustrate these methods using a recent linkage between the 1958 National Child Development Study (NCDS; a cohort following an initial 17,415 people born in Great Britain in a single week of 1958) and Hospital Episode Statistics (HES) databases (containing important information regarding admissions, accident and emergency attendances and outpatient appointments at NHS hospitals in England).

Results

Our illustrative analyses suggest that the linkage quality of the NCDS-HES data is high and that the linked sample maintains an excellent level of population representativeness with respect to the single dimension we assessed.

Conclusions

Through this work we hope to encourage providers and users of linked data resources to undertake and publish thorough evaluations. We further hope that providing illustrative analyses using linked NCDS-HES data will improve the quality and transparency of research using this particular linked data resource.

Keywords

administrative data; cohort studies; data linkage; Hospital Episode Statistics; National Child Development Study; data quality; data representativeness

*Corresponding Author:
Email Address: r.silverwood@ucl.ac.uk (Richard J. Silverwood)

https://doi.org/10.23889/ijpds.v6i1.2137

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 in the UK and beyond [2–5]. Administrative data afford exciting new opportunities for health [6] and social science research [1], in particular to answer questions that require additional information, large sample sizes or involve hard-to-reach populations [7]. Data linkage is a tool for enhancing administrative data, and refers to the process by which information about the same entity that is recorded in disparate data sources is brought together [8–11]. Use of data linkage has increased over recent years, and 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 [12]. 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.

There are two main challenges associated with linked survey-administrative data. First, linkage is generally undertaken only for survey participants who have given explicit consent (i.e. have provided permission, usually as part of the survey data collection, for their administrative data to be linked to their survey data). Consenters may not be representative of the broader sample, with evidence that age, sex, ethnicity, education and income are often related to the likelihood of consent [13]. There is therefore the potential for selection bias even in the presence of perfect linkage of administrative data for consenters. Second, perfect linkage of administrative data 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') [14]. These errors can occur when there is a lack of a unique identifier (e.g., if the National Health Service number is unavailable or not well completed in the survey data), or when partial identifiers that are collected such as date of birth, postcode or sex are recorded incorrectly, change over time, or include missing values. This is often the case with administrative data in the UK. The problem of a lack of well-completed, unique identifiers can be exacerbated by the simple deterministic linkage methods most frequently used by linkage bodies in England [15]: more sophisticated approaches such as probabilistic linkage are not yet readily implemented [16].

Despite advances in linkage methods and improvements in data quality over time [17], some degree of linkage error or uncertainty therefore remains almost always inevitable for linkages involving administrative data. Linkage error can manifest as missing data, misclassification or measurement error, or erroneous inclusion or exclusion of people from an analysis [18]. Misclassification and measurement error may cause information bias, whereas erroneous inclusion or exclusion can lead to selection bias. Differential linkage error, where some groups of individuals are more likely to experience linkage errors than others, can lead to substantial bias even when overall error rates are low [7]. There is a large body of evidence that key participant characteristics are often unevenly distributed in matched and unmatched records [19], suggesting the potential presence of differential linkage error. It is therefore important to examine how linkage errors differ with respect to variables or characteristics of interest, in order to understand the likely effect of bias on results, and to assess whether additional methods may need to be employed to try to address any potential biases. The details of such methods are beyond the scope of the present paper, but they include statistical adjustments based on estimated error rates and distributions (quantitative bias analysis [20]) and probabilistic techniques involving imputation or weighting [7, 18]. A simple example of a quantitative bias analysis would be to consider the best- and worst-case scenarios for linkage error, demonstrating the sensitivity of results to a range of plausible assumptions [7].

Evaluation of linkage quality and linkage bias can be conducted using a number of different approaches [21]. For example, a 'gold standard' or training dataset can be used to quantify false matches and missed matches, comparisons of characteristics of linked and unlinked data can be used to identify potential sources of bias, and sensitivity analyses can be conducted to evaluate how sensitive results are to changes in linkage procedure. However, since linkage of administrative data in the UK (and other countries) typically requires a separation of processes for linkage and analysis [22] to help preserve privacy (as researchers do not have access to the identifiers), it can be difficult for researchers to conduct a comprehensive evaluation of linkage quality. For example, high quality gold-standard data are rarely available, and sensitivity analyses varying the linkage procedure requires the involvement and cooperation of the data linkers, which is not always possible. Information on the confidence in each link, for example the match rank, could be requested from the data linkers and used to conduct sensitivity analyses in which records with linkage below a certain confidence threshold are excluded, but this information may not always be available. Comparisons of characteristics of linked and unlinked data may therefore be the most viable option, though interpretation is not straightforward in settings where all records within a dataset are not necessarily expected to link to records in the other dataset, for example when linking hospital records into a general population master dataset [23].

Linkage of survey and administrative data provides an alternative approach to evaluating linkage quality. If there exist survey data collected on the survey participants which capture similar information to that contained in the linked administrative data, comparison of these variables at the individual level allows on assessment of the extent to which 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.

Representativeness of the linked data sample is also an important issue to consider. This can be considered in two different ways: whether the individuals within the linked data sample are typical of individuals in the wider survey sample, and whether they are typical of individuals in the broader population that the administrative data represent. Consideration of survey data relating to the wider survey sample and whole-population administrative data can help us address these questions of representativeness. A lack of representativeness may suggest the presence of linkage errors [7]. However, there may be other sources of discrepancy, including: selective consent to data linkage (i.e. those consenting being systematically different to those not), selection into initial survey participation (i.e. those included in the survey at initiation being systematically different to those not), and, for longitudinal studies, selective attrition prior to linkage consent being sought (i.e. those remaining in the survey being systematically different to those not). Using external population-level data, in which the survey members would not be identifiable, comparison must necessarily be at the sample (or sub-sample), rather than individual, level: is the distribution of the administrative variable in the linked sample comparable to the distribution of the same variable in the population?

In this paper we describe a suite of generalisable methods to evaluate linkage quality and population representativeness of linked survey and administrative data, emphasising issues particular to longitudinal survey data. The proposed approaches are particularly valuable in the setting where users of the linked data are not party to the linkage process itself. We illustrate these approaches using a recent linkage between the 1958 National Child Development Study (NCDS; a long-running British cohort study) [24] and Hospital Episode Statistics (HES; a database of English hospital admissions, attendances and appointments) data [25, 26]. This linkage was conducted by NHS Digital using a deterministic linkage method [27]. We utilise several different approaches for evaluating linkage quality, 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.

We examine the quality of the linkage in terms of the associations between key cohort member sociodemographic characteristics and successful linkage (i.e. having at least one record linked to the survey data of a given cohort member). We 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. We additionally evaluate the population representativeness of the linked sample using external population-representative data. 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’ [18]. 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.

By describing a suite of generalisable methods to evaluate linkage quality and population representativeness, our aim is to encourage providers and users of linked data resources to undertake and publish detailed evaluations. We further aim to improve the quality and transparency of research using linked NCDS-HES data by providing detailed illustrative analyses using this particular linked data resource [28] .

Methods

Illustrative data linkage

NCDS

The NCDS follows the lives of an initial 17,415 people born in Great Britain in a single week of 1958 [24]. Corresponding non-Great Britain-born immigrants traced through schools joined the survey during the childhood sweeps (up to age 16). Since the initial data collection shortly following birth, NCDS cohort members have been followed up at many sweeps (rounds of data collection). The most recent completed conventional sweep (sweep 9) was undertaken in 2013 when cohort members were 55, three waves of COVID-19-specific surveys were undertaken between May 2020 and March 2021, and a further conventional sweep is currently underway (as of 2023). The study includes information on cohort members’ physical and educational development, economic circumstances, employment, family life, health behaviour, wellbeing, social participation, and attitudes. As part of sweep 8 (2008, age 50), consents for administrative data linkages, including for health data, were sought.

HES

HES is a collection of databases containing details of all admissions (Admitted Patient Care (APC) and Critical Care (CC; which always correspond to a parent APC record)), Accident and Emergency (A&E) attendances and Outpatient (OP) appointments at NHS hospitals in England, maintained by NHS Digital [26]. The four HES databases are not mutually exclusive – an individual can appear in none, one or more than one of them. 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. Here we focus 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 [29].

Linked NCDS-HES data

Linkage between NCDS and all four HES datasets has recently been undertaken, on the basis of consent obtained at sweep 8 (age 50) [25], with data available via secure access through the UK Data Service [30]. Matching was conducted by NHS Digital using deterministic linkage based on combinations of the participant’s name, sex, date of birth and postcode (all fully observed for NCDS cohort members). Linked HES data are currently available from the start of data availability (see above) until 2017, with HES data due to be periodically refreshed. The data linkage itself is not the focus of this paper; here we are concerned with what researchers can do to evaluate the quality of the linkage once it has been performed.

Methods for the evaluation of survey and administrative data linkages

We describe a suite of generalisable methods for the evaluation of survey and administrative data linkages across the following areas: i) Linkage rates, ii) Selection into response, linkage consent and successful linkage, iii) Linkage quality, and iv) Linked data population representativeness. The approaches are particularly relevant when users of the linked data are not party to the linkage process itself. Throughout, there is a focus on nuances specific to the longitudinal survey setting. We illustrate these methods through their application in the linked NCDS-HES dataset.

Linkage rates

An important measure of linkage quality is the linkage rate (the proportion of eligible survey members with linked administrative data). Before the linkage rate can be calculated it is essential to have an understanding of which survey members should be considered eligible for linkage. If linkage is being undertaken on the basis of consent, then clearly only those survey members providing linkage consent should be considered eligible. However, there may be further considerations. If the survey and administrative data do not completely align in terms of age, calendar time, geography, or other factors, then it may be that only a subset of the survey members are eligible for linkage. Moreover, even within the subset of survey members considered eligible for linkage, differential propensities for linkage may be apparent, particularly in the case of longitudinal surveys where survey members may enter and leave eligibility. For example, if an administrative dataset covering a specific geographic area was linked into a longitudinal survey that covered a broader geographical area, then survey members may: i) always live outside the administrative data area; ii) always live inside the administrative data area; or iii) live sometimes inside and sometimes outside the administrative data area. Analyses should be undertaken to explore these issues so that the consequences of different definitions of linkage eligibility can be understood. The sensitivity of subsequent substantive analyses to the definition of linkage eligibility could also be examined through use of alternative definitions in sensitivity analyses.

The above example is precisely the situation we observe in the NCDS-HES linkage: NCDS includes individuals from across Great Britain, whereas HES only contains data from English hospitals. To be considered 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.

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, we calculated the proportion of sweeps between sweep 6 and sweep 9 that each cohort member reported living in England, then calculated the linkage rate for each different proportion.

However, interpretation of linkage rates calculated in the context of the NCDS-HES linkage is not straightforward. Even among NCDS members eligible for linkage, there may be individuals who had no interactions with English NHS hospitals over the period covered by the HES datasets, meaning such individuals would have no HES records to link to. The subsample of NCDS members without linked HES data is therefore potentially made up of both individuals who truly had no HES records (appropriate non-matches) as well as individuals who did in fact have HES records (missed matches).

The calculation of linkage rates, both overall and within population subgroups, is very widely implemented, though the analysis of strata with assumed different linkage likelihood due to structural considerations provides relative novelty in this example.

Selection into response, linkage consent and successful linkage

For linkage rates below 100%, it is important to consider whether certain population subgroups are disproportionately unlikely to have linked data (or, equivalently, to compare characteristics of linked and unlinked data [21]). Such differential linkage rates, signifying selection into linkage, may lead to bias in subsequent substantive analyses [7]. Since linkage between survey and administrative data is generally conducted on the basis of explicit consent from survey members, a further (upstream) potential source of bias is differential linkage consent rates across population subgroups (selection into linkage consent). Another issue, specific to the context of longitudinal surveys, is that if linkage consent was only sought at a more recent sweep of data collection, then attrition from the survey prior to this point may already make respondents at the consent sweep non-representative of the initial survey sample (selection into response). It may therefore be helpful to also examine selection into consent and selection into response at the consent sweep to allow a more comprehensive understanding of the mechanisms underlying any imbalances in the linkage rates. Such analyses typically compare the distributions of previously observed survey variables between groups (i.e. between those with and without successfully linkage, those who did and did not consent to linkage, and those who did and did not respond at the consent sweep). A strength of longitudinal surveys in this regard is that they usually include important sociodemographic variables captured at or close to survey initiation in the vast majority of survey members. Such variables would naturally be important targets of investigations into selection due to their importance in many subsequent substantive analyses of the linked data. A complication is introduced if variables collected in later sweeps are to be considered: these will almost certainly be affected by attrition from the survey. Inclusion of such variables would therefore require careful handling of missing data to ensure that this attrition was not conflated with the selection mechanisms under consideration.

We illustrate this in the NCDS-HES linkage by exploring the sequence of events leading up to successful linkage: i) response at sweep 8 (when health data linkage consents were sought); ii) linkage consent being given; and iii) successful linkage of HES data. We considered a small number of important sociodemographic variables, all recorded at the birth sweep of data collection: cohort member's sex, their father's social class (a marker of childhood socioeconomic status) and the number of persons per room in their home (a marker of socioeconomic circumstances) (for full variable derivation details see Table 1).

Two different analyses were undertaken. The first used a sequential approach to consider associations between the baseline characteristics 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 the UK at age 50), ii) consent to health record linkage among respondents at sweep 8, and iii) successful HES linkage among those 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) successful HES linkage, all among all cohort members eligible for HES linkage and in the sweep 8 target population. Whilst other analytic approaches (e.g. logistic regression) could be used, we applied modified Poisson regression (i.e. using a robust standard error estimator) to model the associations. This method returns risk ratios for ease of interpretation and avoids issues related to the non-collapsibility of the odds ratio. Unadjusted univariable models are presented because the interest is in simple descriptions of the extent of selection.

Similar methods to those proposed here have been used elsewhere (e.g. [35–37]), though the sequential analytic approach is more novel.

Linkage quality

It is also important to examine the quality of the linkage among individuals for whom successful linkage was possible. To do so, we can use survey data collected on the survey participants which capture similar information to that contained in the linked administrative data. 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. 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. For example, self-reported information within either type of data source may be subject to recall issues or social desirability bias, while coding omissions in administrative data may erroneously be interpreted as an individual lacking a certain characteristic. The potential for measurement error should be considered within the context of the existing literature relating to the measurement properties of that variable. Although the proposed approach is not specific to the context of longitudinal surveys, such studies will tend to have collected more information across a longer time period, increasing the potential for the identification of survey information which is well aligned with that in the administrative data.

We explored the quality of the linkage in the NCDS-HES linkage by calculating the percentage of cohort members with linked HES data (i.e. with a relevant hospital admission) within strata of NCDS variables which would be expected to be associated with this. We considered examples of two types of NCDS 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 variable we considered relates to day patient or in-patient attendance reported at sweep 8 (2008; age 50) [38] (Table 1). This survey variable conceptually relates closely to the information recorded in HES, so, under the assumption that the NCDS variable captures the intended constructs, if linkage quality is high we would anticipate close correspondence with the presence of a HES APC record over the same period. However, measurement/misclassification error, in particular due to errors in recall, may affect the reliability of the survey data. The indirectly comparable NCDS variable we considered relates to self-rated general health observed at sweep 9 (2013; age 55) [39]. This survey variable may be less directly related to the presence of a HES APC record, but we would still expect individuals with lower self-rated general health to be more likely to have hospital admissions.

We cross-tabulated each survey variable against the derived HES APC variable. 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. This is an important issue that will potentially affect any linkage setting with meaningfully interpretable links – and the way in which such individuals are considered may impact on the analysis undertaken and the conclusions drawn. Although there is no approach by which such individuals can be definitively characterised, performing descriptive analyses under different assumptions may help to better understand the issue. To this end, we consider the cross-tabulations within different subsets of the linked data sample, corresponding to different assumptions. The cross-tabulations are presented separately within:

a) Individuals with any linked HES APC record ever (n = 4,846).

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).

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. The analysis a) 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 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 being a missed match): this assumption corresponds to analysis b). 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 HES datasets): this assumption corresponds to analysis c). Cross-tabulations are presented by males and females separately and combined.

Table 1: Variable derivation details

Analysis	Dataset	Variable	Coding
Linkage rates	NCDS	Lived in England at one or more sweeps between sweep 6 and sweep 9	Yes, no
		Proportion of sweeps between sweep 6 and sweep 9 living in England	Proportion; sweeps with missing information excluded from calculations given uncertainty over status
Selection into response, linkage consent and successful linkage	NCDS	Sex	Male, female
		Father's social class at birth of cohort member	Registrar General's Social Class[1]: I (professional)/II (managerial and technical), III (non-manual skilled), III (manual skilled), IV (partly-skilled)/V (unskilled)
		Number of persons per room in their home at birth	$\leq 1, > 1$ to 1.5, >1.5
Linkage quality	NCDS	Day patient or in-patient attendance at sweep 8	"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": Yes, no
		Self-rated general health at sweep 9	"In general, would you say your health is excellent, very good, good, fair or poor?": Excellent, very good, good, fair or poor
	Linked NCDS-HES data (APC)	Day patient or in-patient attendance (corresponding to NCDS sweep 8)	Binary variable indicating whether a cohort member had any HES APC record(s) 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 the sweep 8 interview[2].
		Day patient or in-patient attendance (corresponding to NCDS sweep 9)	Binary variable indicating whether cohort members had any HES APC record(s) over the five-year period prior to the date of the sweep 9 interview[2].
Linked data population representativeness	Linked NCDS-HES data (APC)	FAE rate	The total number of FAEs across all cohort members (noting that each cohort member could potentially contribute more than one FAE) was identified within each financial year between 1997-98 and 2015-16 and the rate per 1000 cohort members calculated. For example, if there were 1320 FAEs in a given financial year across 6593 eligible cohort members then the FAE rate would be 200 per 1000. Data for the financial year 2016-17 were excluded as complete HES data are not yet available in the NCDS-HES linkage.
	Population HES APC data	FAE rate	The number of FAEs across the entire population of England are available for 5-year age bands from published reports [31, 32]. 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 (2004–05 to 2015-16; unavailable for earlier years). For example, in the financial year 2004-05 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 [33]. For example, in the financial year 2004-05 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. For example, if there were 797,253 FAEs in a given financial year across an age band-specific population of 3,326,036 then the FAE rate would be 240 per 1000.

APC: Admitted Patient Care; FAE: finished admission episode; HES: Hospital Episode Statistics; NCDS: National Child Development Survey; OP: Outpatients.

[1] Registrar General's Social Class – also referred to as Social Class based on Occupation – is an official scheme of social class designation used in British surveys and censuses for much of the twentieth century [34].

[2] The precise period under consideration was allowed to differ between individuals.

The analytic methods applied here are relatively straightforward, but they are facilitated by the overlap in observed substantive information between the two data sources. This is something that is more likely to be present in the context of linked survey and administrative data that in other linkage settings.

Linked data population representativeness

Our earlier consideration of selection into successful linkage aimed to examine whether the linked dataset remained representative of the broader survey sample. Whole-population administrative data will often allow us to also explore the representativeness of the linked data with respect to the broader population of the administrative data. In situations where the survey is intended to be representative of the same population as that represented by the administrative data, then such a comparison may again be informative about selection into the linked data. Even if this is not the case, exploration of this question will help determine the extent to which analyses of the linked data are likely to be representative of the broader population represented by the administrative data [7]. This can be achieved by comparing distributions of variables within the linked administrative dataset with distributions of the same variables in the population administrative dataset – or to distributions within a relevant subpopulation. For example, if a regional survey was linked to a national administrative dataset, then comparison would be made with the distribution of the variable within the administrative dataset restricted to individuals in the relevant region. In order for the assessment of the population representativeness to be meaningful, it is important that the data are comparable in terms of scope, timeframe and demography. Where relevant aggregated population data already exist, it may not be necessary to access individual-level population data to make such comparisons. Complete representativeness of the linked dataset – i.e. these individuals being typical of the population in *every* conceivable way – is impossible to assess. In practice, it will only ever be possible to address representativeness in a relatively more limited way using a finite set of variables. These should be chosen so as to evaluate population representativeness with respect to features that will be important for subsequent substantive research.

To illustrate this in in the NCDS-HES linkage, we considered ‘finished admission episodes’ (FAEs). 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 [32]. FAEs are the first episode in a spell of care and a count of FAEs therefore equates to the total number of spells of care. We derived the FAE rate per 1000 NCDS cohort members for each financial year between 1997-1998 and 2015-2016 (Table 1). FAE rates were calculated using three different denominators, corresponding to the assumptions discussed above (i.e. a), b) and c)).

Comparable population FAE rates per 1000 individuals for each financial year were calculated using FAE data from published reports [31, 32] and Office for National Statistics mid-year population estimates for England [33] (Table 1). Linked data and population FAE rates were plotted against financial year for comparison.

Similar approaches have been used in linkage evaluations (e.g. [40]), though looking at population representativeness over many individual years and further using this as a tool to compare between different assumptions around meaningfully interpretable linkage is not common in the literature.

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 Figure 1. Of the 10,355 cohort members meeting our definition of linkage eligibility (living in England at one or more sweeps between sweep 6 and sweep 9) and remaining in the target population (alive and living in the UK), 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 (Supplementary Table 1). 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 (Supplementary Table 1).

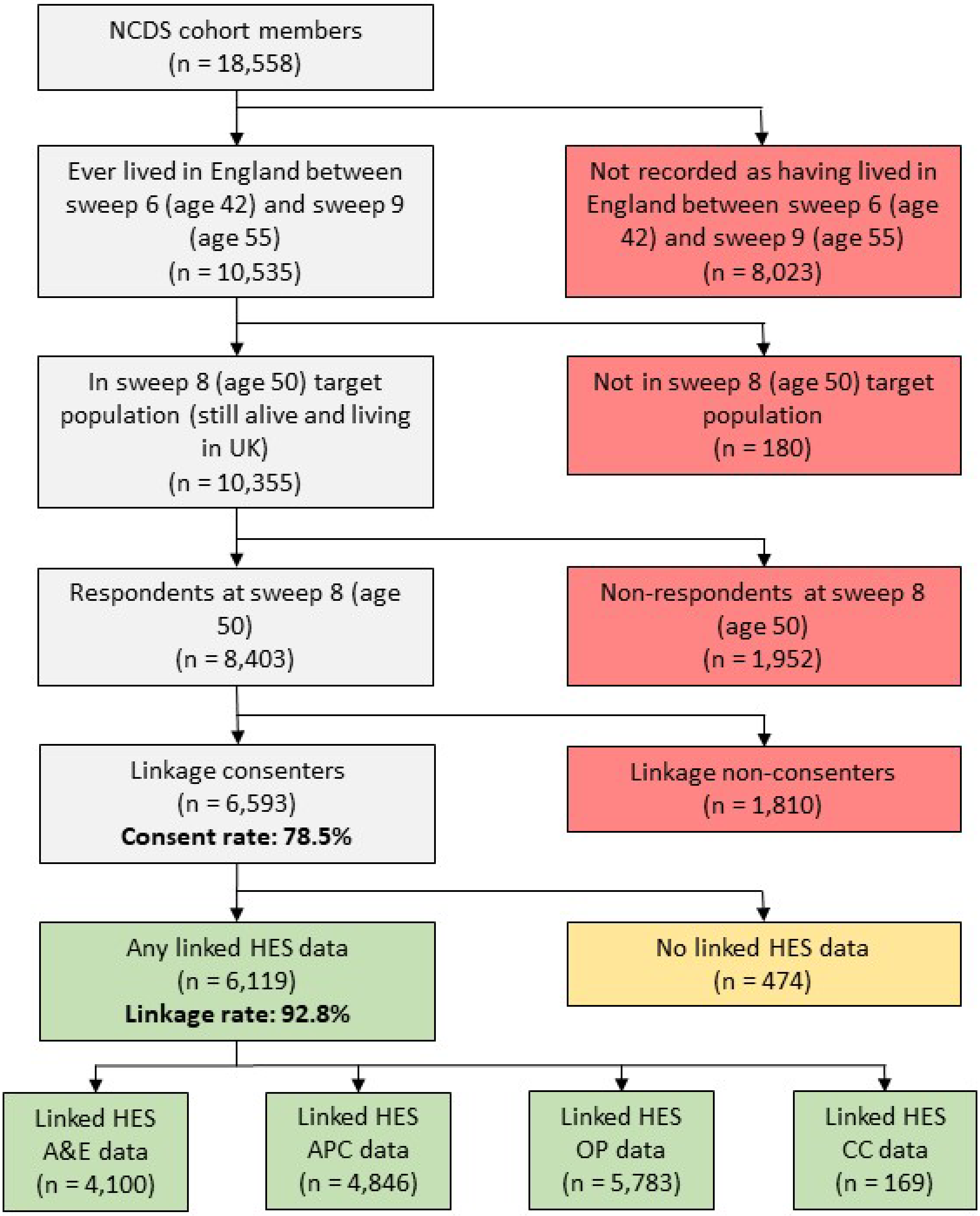
Selection into response, linkage consent and successful linkage

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 2). 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,

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



APC: admitted patient care; CC: critical care; A&E: accident and emergency; OP: outpatients.

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).

Linkage quality

Table 3 shows the cross-tabulations of HES APC linkage and self-reported day patient or in-patient attendance at wave 8 (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:

Table 2: Estimated unadjusted associations with response at sweep 8 (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[1]								
		Response at wave 8 (age 50)			**Consent to linkage**			**Linked HES data**		
	N (%)	**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)
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
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
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
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
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)
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
>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
>1.5	1,038 (10.9)	778 (75.0)	1.00	(ref)	598 (76.9)	1.00	(ref)	570 (95.3)	1.00	(ref)

		Overall[2]								
		Response at wave 8 (age 50)			**Consent to linkage**			**Linked HES data**		
	N (%)	**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 (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 (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 (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 (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 (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 (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 (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 (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 (57.6)	1.00	(ref)	570 (54.9)	1.00	(ref)

RR: Risk ratio.
Risk ratios estimated using modified Poisson regression [41].

[1] In sequential analyses, associations are considered conditional on the outcome at the previous stage, i.e. i) response at sweep 8 among cohort members eligible for HES linkage (lived in England at one or more waves between wave 6 and wave 9) 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) successful HES linkage among cohort members who had consented.

[2] In overall analyses, each association is considered among all cohort members eligible for HES linkage and in the wave 8 target population.

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 (Supplementary Tables 2, 3).

HES APC linkage showed a clear gradient across wave 9 (age 55) categories of self-rated general health (an indirectly comparable "proxy" measure). Linkage ranged from 25.6% in the excellent health group to 73.4% in the poor health group among all linkage consenters (Table 4). Figures were somewhat higher in females than males (Supplementary Tables 4, 5).

Linked data population representativeness

The calculated FAE rates in the linked NCDS-HES data and population data are reported in Supplementary Table 6 and presented graphically in Figure 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

Table 3: 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 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 (59.1)	195 (40.9)	477	482 (71.2)	195 (28.8)	677	566 (74.4)	195 (25.6)	761
	Very good	736 (55.1)	599 (44.9)	1,335	1,202 (66.7)	509 (33.3)	1,801	1,383 (69.8)	509 (30.2)	1,982
	Good	625 (45.2)	757 (54.8)	1,382	989 (56.6)	757 (43.4)	1,746	1,102 (59.3)	757 (40.7)	1,859
	Fair	247 (35.9)	441 (64.1)	688	336 (43.2)	441 (56.8)	777	363 (45.2)	441 (54.8)	804
	Poor	68 (21.5)	248 (78.5)	316	84 (25.3)	248 (74.7)	332	90 (26.6)	248 (73.4)	338
	Total	1,958 (46.6)	2,240 (53.4)	4,198	3,093 (58.0)	2,240 (42.0)	5,333	3,504 (61.0)	2,240 (39.0)	5,744

extent, using all HES linkage consenters, both corresponding closely to the population rate.

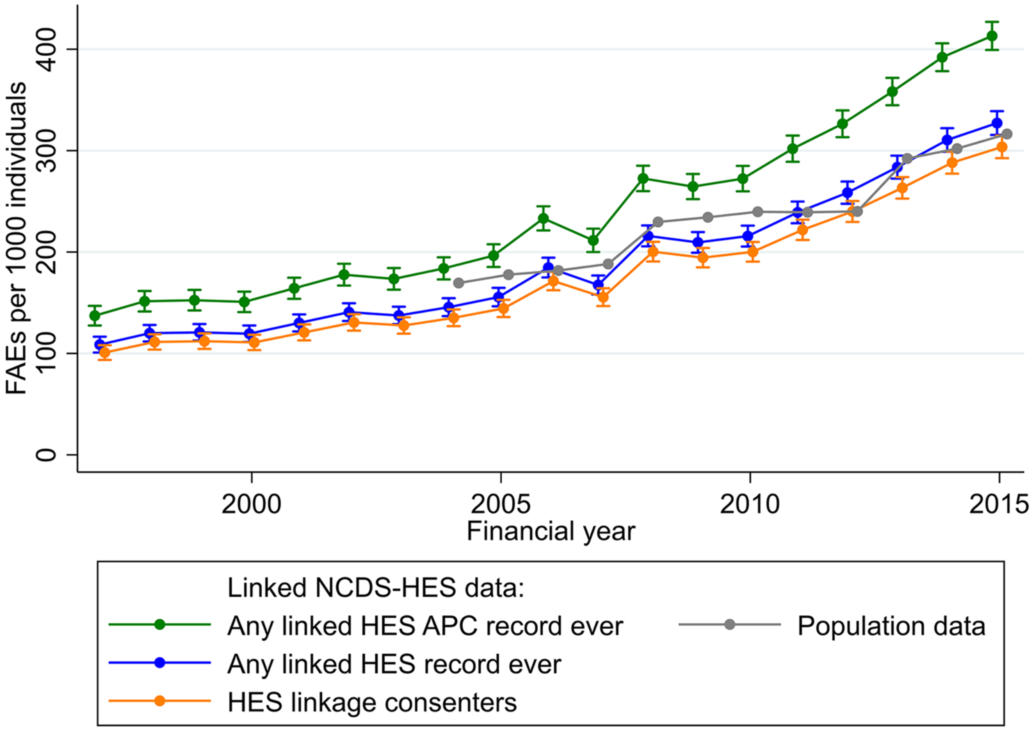
Discussion

In this paper we have described a suite of generalisable methods to evaluate linkage quality and population representativeness of linked survey and administrative data, particularly in the setting where users of the linked data are not party to the linkage process itself. Throughout, there has been a focus on nuances specific to the longitudinal survey setting, an area which has been relatively neglected in the existing literature.

We have illustrated the application of the proposed methods in the NCDS-HES linkage. We observed a clear pattern of increasing linkage rates with increasing proportion of sweeps in which cohort members lived in England. Whilst perhaps unsurprising, 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. Such hospital interactions 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. In other settings where cohort members may enter and leave linkage eligibility over time, the disparity of linkage rates may be greater, meaning that further consideration should be given to this issue. In such situations, the inclusion of a substantial number of cohort members with only partial linked administrative data may lead to bias, particularly if the extent of linkage eligibility is associated with analysis variables of interest. Researchers might therefore consider restricting analyses to cohort members who reach a certain threshold for linkage eligibility (e.g. requiring them to have lived in England for all the sweeps between sweep 6 and sweep 9 in the case of the NCDS-HES linkage) or conducting a series of sensitivity analyses at different threshold values.

We found that females, those whose father was of a higher social class at birth and those with fewer people per room in their home at birth were associated with a somewhat higher likelihood of HES 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 longitudinal survey-administrative data may be due to one of several reasons, including: selective attrition of the sample prior to linkage consent being sought, selective consent to data linkage, and selection into the linkage itself

Figure 2: Linked National Child Development Study (NCDS)-Hospital Episode Statistics (HES) Admitted Patient Care (APC) and whole-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.

(i.e. subpopulations having differential propensity for missed matches) [12]. The NCDS-HES linkage relies on consent given at sweep 8 (age 50) and previous work in NCDS has identified predictors of response at this sweep, including more advantaged socioeconomic background in childhood, better mental health and higher cognitive ability in early life, and greater civic and social participation in adulthood [42]. 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 when considering the NCDS sample as a whole. Females were more likely to have successfully linked HES data conditional on having consented to linkage, which is in broad agreement with the observation that females do generally have slightly more admissions recorded in HES than males [29]. We identified lower levels of linkage among cohort members of higher social class. This finding is counter to the evidence in the literature, where it has been noted that deprivation tends to be positively associated with the probability of a missed match [21]. However, not all NCDS cohort members will have had an interaction with an NHS hospital over this period and therefore will not have a HES record to match to. If the pattern of linkage is in line with that observed in the literature, this finding could therefore be explained by those in a higher social class having lower levels of hospital interactions due to their acknowledged tendency towards better health [43]. This observation therefore feeds into the discussion of how best of consider NCDS cohort members with no linked HES data, suggesting that this group includes many who truly have no hospital interactions. Future analyses could consider whether a wider variety of cohort member characteristics are associated with successful linkage, beyond those considered here for illustrative purposes. Given these findings, if an analysis of the linked data was intended to be fully representative of the original survey 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 [44] or use similar variables within a multiple imputation approach [45]. However, researchers may first wish to undertake further exploratory analyses to examine the plausibility of the missing at random assumption underlying such methods [46].

When examining linkage quality using directly comparable survey data we found high levels of agreement between linked HES APC data and self-reported day patient or in-patient attendance at sweep 8 (age 50). Linked HES APC data also showed a clear gradient across age 55 self-rated general health groups. 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 dataset and the survey questions (or, more specifically, the cohort members' interpretation of them) may not be fully aligned. The self-reported nature of the survey data means that misclassification is a possibility: poor health or hospital attendance may be under-reported due to stigma, social desirability, or other mechanisms; non-differential misclassification may be exacerbated by the recall period extending over many years. To improve comparability, we restricted HES linkages to the period over which the survey questions were asked insofar as this was possible. Given these concerns, we believe that the observed high levels of correspondence between HES linkage and the highly comparable survey measures are suggestive of high levels of linkage quality. Moreover, the findings for self-rated general health, 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. There are a number of previously discussed factors which could potentially impact on the population representativeness of the linked sample, but it is also worth reflecting on the alignment between the whole population data and the NCDS target population. In particular, the NCDS sample includes only individuals born in the designated week in 1958 (the initial sample) plus corresponding non-Great Britain-born immigrants who were traced through schools and joined the survey during the childhood sweeps (up to age 16), whereas the whole population data will contain all non-Great Britain-born immigrants. Given all this, 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. In similar settings with meaningfully interpretable links, the question of how best to handle individuals with no linked records remains a difficult one. We suggest the application of the above-described methods to provide some evidence with which to make a more informed decision, which may differ on an analysis-specific basis. As it is not possible to make this decision with certainty, we would encourage the use of sensitivity analyses which explore the extent to which conclusions differ depending on the way in which such individuals are handled.

Overall, 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 with respect to the single dimension we were able to assess. However, we have only investigated a relatively 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 survey variables or aggregated population-level information is challenging.

There are several strengths to this analysis. We demonstrated generalisable methods for evaluating linkage quality which can be employed even in the absence of access to linkage identifiers and in settings with a separation of processes for linkage and analysis. With regards to the illustrative NCDS-HES analyses, we were able to identify comparable survey data and population-representative data with which to compare the linked data. We utilised a number of different variables and approaches in order to undertake a thorough examination of linkage quality and population representativeness.

There are also a number of limitations to the analysis. 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: if linkage consent analyses are only conducted among respondents at the consent sweep and successful linkage analyses are only conducted among cohort members who had consented, then unaccounted common causes of response and consent or of consent and linkage could lead to bias. Such analyses should therefore be interpreted with some caution. Alternative analytic approaches, such as inverse probability weighting or simultaneous estimation of the sequential models, could be considered. This paper has focused on exploring potential linkage errors and population representativeness. Whilst we have briefly described the potential consequences of linkage error and some approaches to examine or address these, further details are beyond the scope of the present paper but have been elucidated elsewhere [7, 18]. With regards to the illustrative NCDS-HES analyses, 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.

Conclusion

In this paper we have described a suite of generalisable methods to evaluate linkage quality and population representativeness of linked survey and administrative data, particularly in the setting where users of the linked data are not party to the linkage process itself. This is particularly important as the use of data safe havens for working with linked administrative data becomes more common in the UK, leading to further separation of processes and meaning that researchers are less likely to access raw data. Throughout, there has been a focus on nuances specific to the longitudinal survey setting, an area which has been relatively neglected in the existing literature. Through this work we hope to encourage providers and users of such linked data resources to undertake and publish thorough evaluations. We further hope that providing detailed illustrative analyses using linked NCDS-HES data will improve the quality and transparency of research using this particular linked data resource.

Acknowledgements

This work was supported by the Economic & Social Research Council and Administrative Data Research UK [grant number ES/V006037/1] and the Centre for Longitudinal Studies is supported by the Economic & Social Research Council [grant number ES/W013142/1]. 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.

Statement on conflicts of interest

The authors have no conflicts of interest to declare.

Ethics statement

NCDS was approved by the National Health Service Research Ethics Committee and all participants have given informed consent.

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. https://doi.org/10.1016/j.ssresearch.2016.04.015

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. https://doi.org/10.1177/2053951717745678

3. Ali MS, Ichihara MY, Lopes LC, Barbosa GCG, Pita R, Carreiro RP, et al. Administrative Data Linkage in Brazil: Potentials for Health Technology Assessment. Frontiers in Pharmacology. 2019;10. https://doi.org/10.3389/fphar.2019.00984

4. Chiu M, Lebenbaum M, Lam K, Chong N, Azimaee M, Iron K, et al. Describing the linkages of the immigration, refugees and citizenship Canada permanent resident data and vital statistics death registry to Ontario's administrative health database. BMC Medical Informatics and Decision Making. 2016;16(1):135. https://doi.org/10.1186/s12911-016-0375-3

5. Hand DJ. Statistical challenges of administrative and transaction data. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2018;181(3):555–605. https://doi.org/10.1111/rssa.12315

6. Gavrielov-Yusim N, Friger M. Use of administrative medical databases in population-based research. Journal of Epidemiology and Community Health. 2014;68(3):283. https://doi.org/10.1136/jech-2013-202744

7. Harron K, Doidge JC, Goldstein H. Assessing data linkage quality in cohort studies. Annals of Human Biology. 2020;47(2):218–26. https://doi.org/10.1080/03014460.2020.1742379

8. Dunn H. Record linkage. Am J Public Health. 1946;36(12):1412–16.

9. Fellegi I, Sunter A. A theory for record linkage. J Am Stat Assoc. 1969;64(328):1183–210. https://doi.org/10.1080/01621459.1969.10501049

10. Newcombe H. Handbook of record linkage: methods for health and statistical studies, administration, and business. New York, NY, USA Oxford University Press, Inc; 1988.

11. Jutte DP, Roos LL, Brownell MD. Administrative Record Linkage as a Tool for Public Health Research. Annual Review of Public Health. 2011;32(1):91–108. https://doi.org/10.1146/annurev-publhealth-031210-100700

12. 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.

13. 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.

14. Herzog TN, Scheuren FJ, Winkler WE. Data Quality and Record Linkage Techniques. New York, NY: Springer; 2007.

15. Zhu Y, Matsuyama Y, Ohashi Y, Setoguchi S. When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. Journal of Biomedical Informatics. 2015;56:80–6. https://doi.org/10.1016/j.jbi.2015.05.012

16. Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic record linkage. International Journal of Epidemiology. 2015;45(3):954–64. https://doi.org/10.1093/ije/dyv322

17. Christen P, Ranbaduge T, Schnell R. Linking Sensitive Data: Methods and Techniques for Practical Privacy-Preserving Information Sharing: Springer; 2020.

18. Doidge JC, Harron KL. Reflections on modern methods: linkage error bias. Int J Epidemiol. 2019;48(6):2050–60. https://doi.org/10.1093/ije/dyz203

19. 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. https://doi.org/10.1186/1472-6963-10-346

20. Fox MP, MacLehose RF, Lash TL. Applying quantitative bias analysis to epidemiologic data. Second edition. Switzerland: Springer Nature; 2021.

21. 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. https://doi.org/10.1093/ije/dyx177

22. Kelman CW, Bass AJ, Holman CD. Research use of linked health data–a best practice protocol. Aust N Z J Public Health. 2002;26(3):251-5. https://doi.org/10.1111/j.1467-842x.2002.tb00682.x

23. Pratt NL, Mack CD, Meyer AM, Davis KJ, Hammill BG, Hampp C, et al. Data linkage in pharmacoepidemiology: A call for rigorous evaluation and reporting. Pharmacoepidemiology and Drug Safety. 2020;29(1):9–17. https://doi.org/10.1002/pds.4924

24. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study). Int J Epidemiol. 2006;35(1):34–41. https://doi.org/10.1093/ije/dyi183

25. 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.

26. 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 19 July 2022.].

27. Harper G. Linkage of Maternity Hospital Episode Statistics data to birth registration and notification records for births in England 2005–2014: Quality assurance of linkage of routine data for singleton and multiple births. BMJ Open. 2018;8(3):e017898. https://doi.org/10.1136/bmjopen-2017-017898

28. Gilbert R, Lafferty R, Hagger-Johnson G, Harron K, Zhang LC, Smith P, et al. GUILD: GUidance for Information about Linking Data sets. J Public Health (Oxf). 2018;40(1):191–8. https://doi.org/10.1093/pubmed/fdx037

29. 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. https://doi.org/10.1093/ije/dyx015

30. 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. https://doi.org/10.5255/UKDA-SN-8697-1. 2021.

31. 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.

32. 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.

33. 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.

34. Scott J, Marshall G. Registrar General's Classification. A Dictionary of Sociology. New York: Oxford University Press; 2009.

35. Libuy N, Harron K, Gilbert R, Caulton R, Cameron E, Blackburn R. Linking education and hospital data in England: linkage process and quality. Int J Popul Data Sci. 2021;6(1):1671. https://doi.org/10.23889/ijpds.v6i1.1671

36. Raffray M, Bayat S, Lassalle M, Couchoud C. Linking disease registries and nationwide healthcare administrative databases: the French renal epidemiology and information network (REIN) insight. BMC Nephrology. 2020;21(1):25. https://doi.org/10.1186/s12882-020-1692-4

37. Ford JB, Roberts CL, Taylor LK. Characteristics of unmatched maternal and baby records in linked birth records and hospital discharge data. Paediatric and perinatal epidemiology. 2006;20(4):329–37. https://doi.org/10.1111/j.1365-3016.2006.00715.x

38. 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. https://doi.org/10.5255/UKDA-SN-6137-2. 2020.

39. University of London, Institute of Education, Centre for Longitudinal Studies. National Child Development Study: Age 55, Sweep 9 2013. [data collection]. UK Data Service. SN: 7669. https://doi.org/10.5255/UKDA-SN-7669-1. 2020.

40. Harron K, Gilbert R, Cromwell D, van der Meulen J. Linking Data for Mothers and Babies in De-Identified Electronic Health Data. PLoS One. 2016;11(10):e0164667. https://doi.org/10.1371/journal.pone.0164667

41. Zou G. A modified poisson regression approach to prospective studies with binary data. Am J Epidemiol. 2004;159(7):702–6. https://doi.org/10.1093/aje/kwh090

42. 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. https://doi.org/10.1016/j.jclinepi.2021.02.019

43. Marmot M. Fair society, healthy lives: the Marmot Review; strategic review of health inequalities in England post-2010. London: The Marmot Review; 2010.

44. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res. 2013;22(3):278–95. https://doi.org/10.1177/0962280210395740

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

46. Little RJA, Rubin DB. Statistical Analysis with Missing Data. Third Edition. Hoboken, NJ: Wiley; 2020.

Abbreviations

A&E:	Accident and Emergency
APC:	Admitted Patient Care
CC:	Critical Care
FAE:	Finished admission episodes
HES:	Hospital Episode Statistics
NCDS:	National Child Development Study
OP:	Outpatient

Supplementary Material

Supplementary Table 1: 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 (52.0)	4 (40.0)	13 (36.1)	3 (27.3)	11 (18.0)	430 (6.7)	474 (7.2)
Linked HES data	12 (48.0)	6 (60.0)	23 (63.9)	8 (72.7)	50 (82.0)	6,020 (93.3)	6,119 (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.

Supplementary 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

		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 (78.7)	322 (21.3)	1,511	1,853 (85.2)	322 (14.8)	2,175	2,111 (86.8)	322 (13.2)	2,433
	Yes	143 (18.7)	623 (81.3)	766	192 (23.6)	623 (76.4)	815	214 (25.6)	623 (74.4)	837
	Total	1,332 (58.5)	945 (41.5)	2,277	2,045 (68.4)	945 (31.6)	2,990	2,325 (71.1)	945 (28.9)	3,270

Supplementary Table 3: 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 (78.9)	335 (21.1)	1,587	1,762 (84.0)	335 (16.0)	2,097	1,939 (85.3)	335 (14.7)	2,274
	Yes	166 (16.9)	815 (83.1)	981	216 (21.0)	815 (79.0)	1,031	233 (22.2)	815 (77.8)	1,048
	Total	1,418 (55.2)	1,150 (44.8)	2,568	1,978 (63.2)	1,150 (36.8)	3,128	2,172 (65.4)	1,150 (34.6)	3,322

Supplementary 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 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 (57.9)	90 (42.1)	214	237 (72.5)	90 (27.5)	327	293 (76.5)	90 (23.5)	383
	Very good	346 (56.1)	271 (43.9)	617	602 (69.0)	271 (31.0)	873	706 (72.3)	271 (27.7)	977
	Good	297 (45.3)	359 (54.7)	656	490 (57.7)	359 (42.3)	849	549 (60.5)	359 (39.5)	908
	Fair	127 (38.7)	201 (61.3)	328	180 (47.2)	201 (52.8)	381	196 (49.4)	201 (50.6)	397
	Poor	32 (23.4)	105 (76.6)	137	44 (29.5)	105 (70.5)	149	47 (30.9)	105 (69.1)	152
	Total	926 (47.4)	1,026 (52.6)	1,952	1,553 (60.2)	1,026 (39.8)	2,579	1,791 (63.6)	1,026 (36.4)	2,817

Supplementary Table 5: 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 (60.1)	105 (39.9)	263	245 (70.0)	105 (30.0)	350	273 (72.2)	105 (27.8)	378
	Very good	390 (54.3)	328 (45.7)	718	600 (64.7)	328 (35.3)	928	677 (67.4)	328 (32.6)	1,005
	Good	328 (45.2)	398 (54.8)	726	499 (55.6)	398 (44.4)	897	553 (58.2)	398 (41.8)	951
	Fair	120 (33.3)	240 (66.7)	360	156 (39.4)	240 (60.6)	396	167 (41.0)	240 (59.0)	407
	Poor	36 (20.1)	143 (79.9)	179	40 (21.9)	143 (78.1)	183	43 (23.1)	143 (76.9)	186
	Total	1,032 (46.0)	1,214 (54.0)	2,246	1,540 (55.9)	1,214 (44.1)	2,754	1,713 (58.5)	1,214 (41.5)	2,927

Supplementary Table 6: 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 ($n = 6,593$).