International Journal of Population Data Science

Journal Website: www.ijpds.org

Adjusting for confounding in population administrative data when confounders are only measured in a linked cohort

Richard J. Silverwood[1*], Gergő Baranyi[1], Lisa Calderwood[1], Bianca L. De Stavola[2], George B. Ploubidis[1], Ian R. White[3], and Katie Harron[2]

Submission History

Submitted:	23/05/2025
Accepted:	01/12/2025
Published:	06/03/2026

[1] Centre for Longitudinal Studies, UCL Social Research Institute, University College London, 20 Bedford Way, London WC1H 0AL, United Kingdom

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

[3] MRC Clinical Trials Unit at UCL, University College London, 90 High Holborn, London, WC1V 6LJ, United Kingdom

Abstract

Introduction
Analyses of population administrative data can often only be minimally adjusted due to a lack of control variables, potentially leading to bias due to residual confounding.

Objectives
We aimed to use linked cohort data to help address residual confounding in analyses of population administrative data. One particular aim was to explore strategies for when linked cohort and population administrative data cannot be accessed together in the same environment (are "siloed").

Methods
We propose a multiple imputation-based approach, introduced through application to simulated data in three different scenarios related to the structure of the datasets. We then apply this approach to a real-world example – examining the association between pupil mobility (changing schools at non-standard times) and Key Stage 2 (age 11) attainment using data from the UK National Pupil Database (NPD). The limited control variables available in the NPD are supplemented by multiple measures of socioeconomic deprivation captured in linked Millennium Cohort Study (MCS) data.

Results
In our real-world example, we included 509,670 individuals in the population NPD data, of whom 7,768 (1.5%) were MCS cohort members. The unadjusted estimate of -1.86 (95% CI -1.92, -1.81) for the association between pupil mobility and Key Stage 2 attainment was attenuated to -0.92 (95% CI -0.97, -0.88) through adjustment for the NPD control variables, and further attenuated to -0.76 (95% CI -0.86, -0.67) through adjustment for the MCS control variables.

Conclusions
Linked cohort data can be used to address residual confounding in analyses of population administrative data, and our proposed approach performed well across a range of simulated and real-world scenarios. The underlying principles are widely applicable: any analysis of administrative data could potentially be strengthened by linking a subset of individuals into richer cohort data. More research is required to understand how these methods can be applied more broadly.

Keywords
population administrative data; cohort data; confounding; multiple imputation; methodology

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

https://doi.org/10.23889/ijpds.v11i1.3015

Introduction

Over recent decades, the increasing availability of administrative data created by government and public bodies across the UK has led to an expansion of research utilising these resources [1]. Such administrative data tend to be population-representative and large in size, but as they are not collected specifically for research purposes, they often lack information in key domains [2]. For example, administrative data typically contain area-level measures socioeconomic deprivation such as the Index of Multiple Deprivation [3], which is a suboptimal indicator of socioeconomic status and can lead to incorrect inferences [4]. The UK also has a long and successful history of national longitudinal cohort studies [5]. These cohort studies aim to be population-representative at initiation and collect rich information across a wide multidisciplinary range of research areas. For example, cohort data many include several different measures of socioeconomic deprivation (e.g. occupational social class, income, wealth, education), at different levels (e.g. individual, household, parental) and repeatedly over a potentially long period of time. However, cohort studies do not have population coverage and can therefore be insufficiently powered to address some research questions, e.g. on rare conditions or specific subgroups of individuals, and can also be affected by attrition during follow-up, limiting representativeness.

Using these different sources of data in ways that complement and add value to each other has the potential to lead to more valuable research insights than using each separately. One way in which this is done is through linkage of cohort and administrative data, which has become increasingly common over recent years [2]. These linkages can provide more detailed information on cohort members that is not available in administrative data [2], which can be used for substantive research or for other purposes, such as quality assurance/data validation, or improved handling of missing cohort data [6]. There are many instances where analyses of administrative data can only be minimally adjusted due to the unavailability of a full set of control variables, which can lead to (potentially substantial) bias due to residual confounding [7]. Combining the well powered analysis using population administrative data with rich information on confounders in the cohort study provides an opportunity to utilise the strengths of both data sources. The present work focuses on how the more granular information available in cohort data, even if only available for a subset of individuals within a given administrative dataset, can be used to help to reduce confounding.

For various logistical and governance reasons, it is not always possible to access linked cohort and population administrative data together in the same environment. One particular aim of the present work was therefore to explore whether strategies for using linked cohort data to help handle confounding in analyses of population administrative data are effective when these data are stored separately (are "siloed"). We approach this by framing the unobserved confounding variables in the population administrative data as a missing data problem. In particular, we propose the use of multiple imputation (MI). In MI, the analyst specifies an appropriate imputation model, from which a series of imputed datasets are created. Each imputed data set is analysed using the substantive model of interest and the results are combined using standard rules [8], resulting in standard errors that incorporate the variability in results between the imputed data sets. In this way, uncertainty about the missing data is appropriately accounted for in the inference. Over recent years, MI has been widely adopted because it is practical for applied researchers in a wide range of settings and can be undertaken using standard statistical software [9]. Use of MI in the present context allows us to utilise existing theory, methods and software implementations to provide an intuitive yet rigorously justified solution that can be straightforwardly applied using standard statistical software. We use a simple simulated example and a real-world example to illustrate this process.

Simple simulated example

Outline

We introduce the proposed approach through a simulated example dataset where interest is in estimating the association between a binary exposure, X, and a binary outcome, Y, in a population administrative dataset. In this simulation, we suppose we have access to an administrative dataset containing records for a population of 100,000 individuals. We also have access to data from a cohort study of 10,000 individuals who form a representative subsample of those in the population administrative dataset and all of whom have been linked to their administrative data. There are two normally distributed confounders of the X-Y association (common causes of X and Y): C, which is recorded in the population administrative dataset and hence available for everyone, and U, which is recorded in the cohort study but not in the population administrative dataset. U and C are themselves correlated, as would often be the case with confounders. Further specifics of the data generating process are provided in Methods S1 (Supplementary Material).

An analysis of the X-Y association for all 100,000 individuals in the population administrative dataset will only allow us to account for confounding by C (e.g. through regression adjustment) as U is unobserved. As there is additional confounding by U (i.e. having already accounted for confounding by C), we would expect the resulting estimate of the X-Y association to therefore be biased.

An analysis of the X-Y association for the 10,000 individuals in the cohort study with linked administrative data records will allow us to account for confounding by both C and U, as within this subsample both are observed. We would therefore expect this estimate of the X-Y association to be unbiased, though corresponding to a less well powered analysis relative to the inclusion of all 100,000 individuals in the population administrative dataset.

Our proposed methodology (see below) allows us to analyse all 100,000 individuals in the population administrative dataset and account for confounding by both C and U, while appropriately incorporating the uncertainty inherent in an analysis with such a high proportion of unobserved information. We proceed by framing the problem as one of missing data, allowing us to utilise existing theory, methods and software implementations.

The overall approach is to leverage observed information on the associations between all variables (here C, X, Y and U) in the linked cohort data to impute values of the variables only available in the cohort study (here U) for non-cohort members. We use MI to appropriately account for uncertainty. The key assumption is that U is missing at random (MAR) given the observed data, meaning that missingness in U does not depend on unobserved variables. In this simple simulated example, we know that the MAR assumption holds since the individuals in the cohort study are essentially a random subset of those in the population administrative data. In general, the plausibility of the MAR assumption will require careful consideration.

Scenarios and methods

We consider three different scenarios related to the structure of the datasets, proposing alternative versions of the methodological approach in each.

Scenario 1

In this scenario, the linked cohort data and population administrative data can be analysed together in the same location and the cohort members' identities within the population administrative dataset are known. The datasets can therefore be combined by merging in the cohort data for cohort members and the combined dataset of 100,000 individuals can be rearranged (for ease of illustration) so that the first 10,000 rows relate to cohort members (with all of X, Y, C and U observed) and the following 90,000 rows relate to individuals in the population administrative dataset but not in the cohort study (with X, Y and C observed but U unobserved) (Table 1).

It is then straightforward to use MI to impute the values of U for the individuals not in the cohort study. We specify the imputation model as a linear regression of U on X, Y and C and create 50 imputed datasets. A variety of different MI approaches are available. Here only a single variable is being imputed, but to allow the imputation of multiple variables simultaneously we would suggest using multiple imputation by chained equations (MICE) [10, 11]. Once the values of U are imputed, each imputed dataset can be analysed using the model that would have been used had all variables been fully observed in the population administrative dataset (here a logistic regression of Y on X controlling for C and U). The estimates are then combined across imputed datasets using Rubin's rules [8]. Implementation was via the "mi" suite of commands in Stata [12]. The resultant estimate of the X-Y association uses data on all 100,000 individuals in the population administrative dataset and appropriately accounts for confounding by both C and U. An overview of the proposed approach is provided in Fig. 1.

Scenario 2

In this scenario, the linked cohort data and population administrative data can similarly be analysed together in the same location, but within the population administrative data we do not know which individuals are in the cohort. Indeed, it could be the case that some or all of the individuals in the cohort are not contained within the population administrative dataset at all. The datasets can be combined by appending the linked cohort data to the population administrative dataset, creating an augmented dataset of size 110,000 in which the first 100,000 rows relate to individuals in the population administrative dataset who may or may not be cohort members (with X, Y and C observed but U unobserved) and the following 10,000 rows relate to cohort members (with all of X, Y, C and U observed) (Table 2). In this scenario, we may therefore have duplicated observations for the 10,000 cohort members (one observation has linked cohort data, the other does not).

Similarly to scenario 1, we can use MI to impute the values of U, but now we are doing so for all 100,000 individuals in the population administrative data (i.e. including those who may actually be cohort members). MI would proceed in exactly the same way with the imputation model similarly specified. The difference is that once the values of U are imputed, we remove from the dataset the appended cohort data, leaving just the 100,000 individuals in the population administrative dataset, all of whom have imputed values of U. Each imputed dataset can then be analysed and the estimates combined across imputed datasets similarly to scenario 1. Implementation was via the "mi" suite of commands in Stata [12]. Again, the resultant estimate of the X-Y association will use data on all 100,000 individuals in the population administrative dataset and appropriately account for confounding by both C and U. While this estimate should be unbiased, it is known that in situations where only a subset of the records used to fit the imputation model is used to fit the analysis model, the variance may not be correctly estimated using Rubin's rules [13, 14]. The variance estimator of Robins and Wang [13] could possibly be used in this setting. Further work is required

Table 1: Example dataset for scenario 1

id	cohort	x	y	c	u
1	1	1	0	0.9592	0.804864
2	1	0	0	−0.00181	0.768874
⋮	⋮	⋮	⋮	⋮	⋮
10,000	1	1	1	−0.39435	−0.47173
10,001	0	1	1	0.215663	.
⋮	⋮	⋮	⋮	⋮	⋮
100,000	0	0	0	1.014853	.

Figure 1: Overview of proposed approach under scenario 1

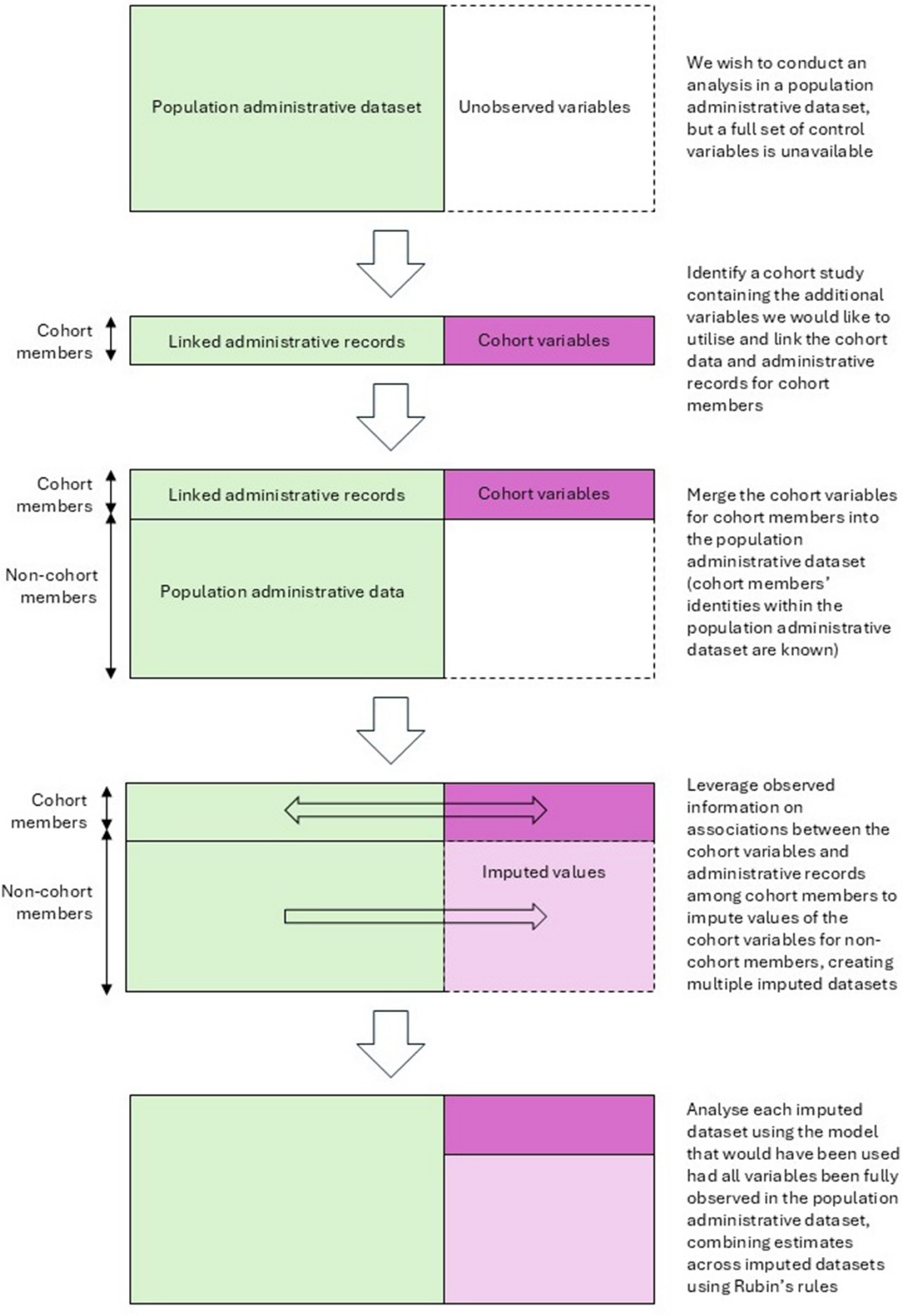


Table 2: Example dataset for scenario 2

id	cohort	x	y	c	u
1	?	1	0	0.9592	.
2	?	0	0	−0.00181	.
⋮	⋮	⋮	⋮	⋮	⋮
100,000	?	0	0	1.014853	.
100,001	1	1	0	0.9592	0.804864
⋮	⋮	⋮	⋮	⋮	⋮
110,000	1	1	1	−0.39435	−0.47173

to examine the potential implications of this in the present context. An overview of the proposed approach is provided in Fig. S1, Supplementary Material.

Scenario 3

In this scenario, the linked cohort data and population administrative data can no longer be analysed together in the same location. That is, while we can link cohort members to their administrative data, we cannot access administrative data for non-cohort members in the same environment. The cohort members may or may not be contained within the population administrative dataset, but given the separation of their cohort data from the population administrative dataset, even if they were, their cohort data will not be available within the population administrative dataset (Table 3). This scenario is becoming increasingly common with different data holders requiring analysis within their own secure data environments (SDEs), with limited (or no) possibility of moving datasets between SDEs.

This scenario is more challenging to address as the standard implementation of MICE necessitates application within a single dataset containing all complete and incomplete variables. To overcome this challenge, we propose a novel implementation of the MICE procedure in which the processes for the development of the imputation model and its application are separated. More specifically, we develop the imputation model (here a linear regression of U on X, Y and C) in the linked cohort dataset in the usual way and save (perturbed) imputation model regression coefficients for each imputation. We then apply each set of coefficients in the population administrative dataset to create 50 imputed datasets. We achieve this through modification of the Stata "ice" package for the implementation of MICE [15]. The imputed datasets are then analysed as previously described. Similarly to scenario 2, U is again imputed for everyone in the population administrative dataset. This estimate should be unbiased, but again the variance may not be correctly estimated using Rubin's rules [13, 14] with the variance estimator of Robins and Wang [13] possibly providing a solution. Further work is required to examine the potential implications of this in the present context. An overview of the proposed approach is provided in Fig. S2, Supplementary Material.

Additional simulated features

We explored a number of additional simulated features:

a) Confounder U being binary rather than continuous. The approaches described for scenarios 1-3 are unchanged, with the exception that U is imputed using logistic regression rather than linear regression. This is straightforward to undertake using standard implementations of MICE in scenarios 1 and 2, but again requires a small modification to the Stata "ice" code in scenario 3.

b) Confounder U being categorical rather than continuous. The approaches described for scenarios 1-3 are unchanged, with the exception that U is imputed using multinomial logistic regression rather than linear regression. This is straightforward to undertake using standard implementations of MICE in scenarios 1 and 2, but again requires a small modification to the Stata "ice" code in scenario 3.

c) Multiple normally distributed partially observed confounders (U_1 and U_2) and multiple normally distributed fully observed confounders (C_1 and C_2). The approaches described for scenarios 1-3 are essentially unchanged, but two imputation models are now specified: a linear regression of U_1 on X, Y, C_1, C_2 and U_2 and a linear regression of U_2 on X, Y, C_1, C_2 and U_1. Both standard implementations of MICE and our modified version using Stata "ice" are designed to handle missingness in multiple variables, so no further modification is required.

All analyses were conducted using Stata version 18 [16]. Stata code for the simulation of these example datasets and the analyses corresponding to scenarios 1, 2 and 3 is available at https://osf.io/mqnzx/.

Results

Estimated X-Y associations in all scenarios are presented in Fig. 2 and Table S1 (Supplementary Material). The first estimate (model 1) used data for all 100,000 individuals in the population administrative dataset and fully adjusted for both C and U (i.e. prior to U being recoded to missing for the

Table 3: Example dataset for scenario 3

Location 1					
id1	**cohort**	**x**	**y**	**c**	**u**
1	?	1	0	0.9592	.
2	?	0	0	−0.00181	.
⋮	⋮	⋮	⋮	⋮	⋮
100,000	?	0	0	1.014853	.
Location 2					
id2	**cohort**	**x**	**y**	**c**	**u**
1	1	1	0	0.9592	0.804864
⋮	⋮	⋮	⋮	⋮	⋮
10,000	1	1	1	−0.39435	−0.47173

non-cohort members), so forms the "known truth" to which the other results using the population administrative dataset should be compared.

The next set of results (models 2-4) use the data on the 10,000 cohort members who have both C and U observed. Given the extent of confounding simulated in the dataset, the unadjusted estimate (model 2) exhibits considerable bias. Some of this is overcome through adjustment for C (model 3). Adjustment for both C and U (model 4) results in an estimate close to the corresponding estimate ("known truth") in the population administrative dataset, albeit with much wider confidence intervals (CIs) due to the smaller sample size.

In the population administrative dataset, the unadjusted estimate (model 5) and the estimate adjusting for C only (model 6) are similar to the corresponding estimates using the 10,000 cohort members, though with greater precision, as would be expected.

Additional adjustment for U using the proposed MI-based approaches results in estimates very close to the known truth in all of scenarios 1, 2 and 3 (models 7, 8 and 9, respectively). They are also only slightly less precisely estimated than the estimates in the population administrative dataset adjusting for C only (95% CI widths of 0.067 (scenario 1), 0.071 (scenario 2) and 0.069 (scenario 3) vs. 0.056 when adjusting for C only), even though information on the confounder U has been imputed for 90% (scenario 1) or 100% (scenarios 2 and 3) of individuals contributing to the analysis.

Additional simulated features

The proposed approaches worked well when considering binary, categorical or multiple normally distributed confounders (Fig. S3-S5, Supplementary Material).

Real-world example

In addition to examining the performance of the proposed approach in a simulated dataset with known properties, it is also informative to see how this would work in a real-world context.

Methods

Data

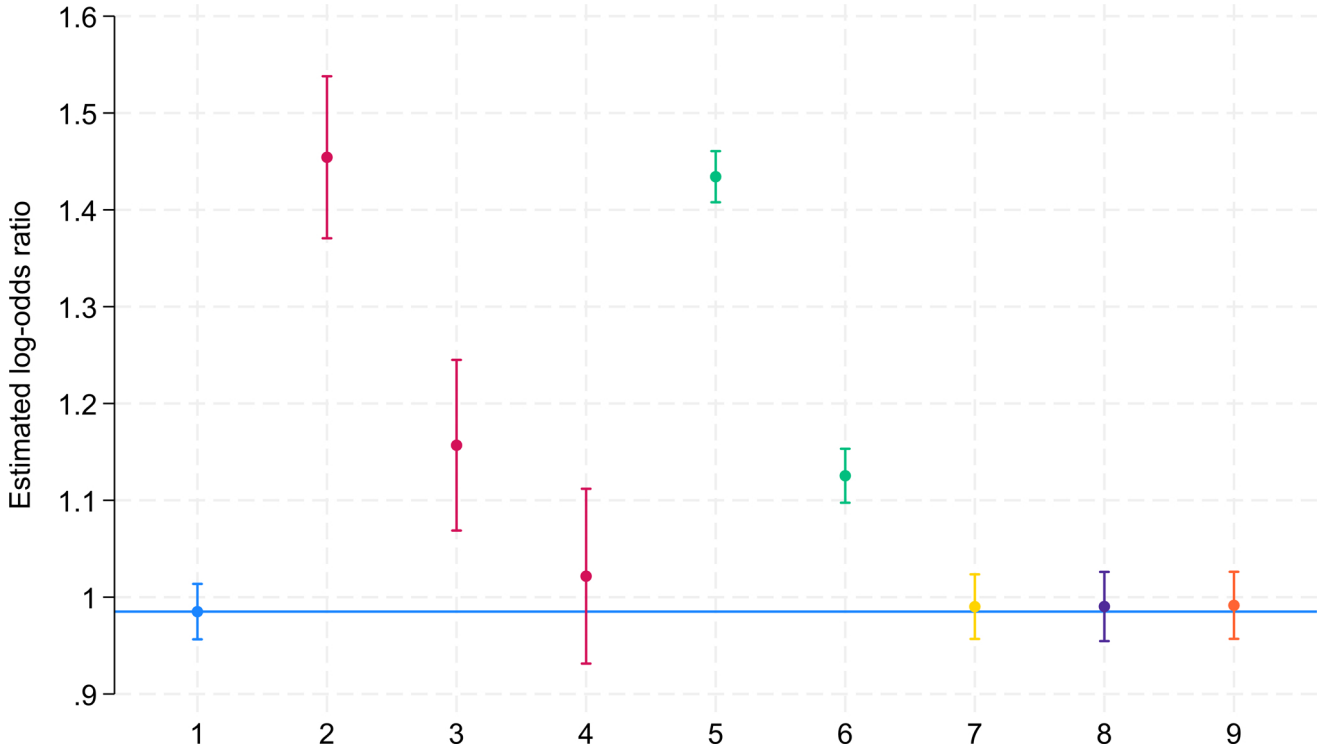
We use data from the Millennium Cohort Study (MCS) [17, 18] and National Pupil Database (NPD) [19], including both linked MCS-NPD data [20, 21] and NPD data for the whole population with birth years aligned to those in MCS.

The MCS is an ongoing birth cohort of around 19,000 individuals born across the UK in 2000-2002 [17, 18]. It has a clustered design, with oversampling on certain geographic, socioeconomic and ethnic criteria. Data have so far been collected at ages 9 months and 3, 5, 7, 11, 14, 17 and 23 years. It is a highly multidisciplinary study, designed to capture the influence of early family context on child development and outcomes through to adulthood.

The NPD is a record-level administrative data resource curated by the Department for Education which contains child-level and school-level data on all pupils in state schools in England [19, 22]. Information in NPD includes child characteristics, school enrolment, alternative provision, exam attainment, absence, and exclusions. Key Stage 1 (KS1), KS2 and KS4 tests are undertaken in all state-maintained schools in England, at ages 7, 11 and 16 respectively.

Information held on MCS cohort members has already been linked to NPD records on the basis of consents from parents/carers at MCS Sweep 4 (age 7), which were obtained for 8489 (93.8%) of the 9047 MCS Sweep four participants who were resident in England during that period (and who would therefore be eligible for inclusion in the NPD). Full information on the linkage is available elsewhere [21]. Briefly, in September 2018 the Department of Education linked all consenting MCS participants who were resident in England during any of MCS Sweeps 3, 4 or 5 (that is, when the participants were of school age) to their records in the NPD. The linked records provide access to KS1-KS4 data, as well as Pupil Level Annual School Census data, and absence data. Deterministic linkage was undertaken on the basis of the cohort member's name, sex, date of birth, most recent postcode, and postcode at the three most recent sweeps of data collection (ages 7, 11 and 14). A total of 8438 consenting MCS participants were successfully matched to records in one

Figure 2: Estimated X-Y associations in the simple simulated example with normally distributed fully observed confounder C and normally distributed cohort-only confounder U



Points are estimates; bars are 95% confidence intervals.

1. Estimate using all 100,000 individuals in the population administrative dataset, fully adjusted for both C and U (i.e. prior to U being recoded to missing for the non-cohort members).
2. Estimate using the 10,000 cohort members only; unadjusted.
3. Estimate using the 10,000 cohort members only; adjusted for C only.
4. Estimate using the 10,000 cohort members only; adjusted for C and U only.
5. Estimate using all 100,000 individuals in the population administrative dataset; unadjusted.
6. Estimate using all 100,000 individuals in the population administrative dataset; adjusted for C only.
7. Estimate using all 100,000 individuals in the population administrative dataset; adjusted for C and with adjustment for U as described in "scenario 1".
8. Estimate using all 100,000 individuals in the population administrative dataset; adjusted for C and with adjustment for U as described in "scenario 2".
9. Estimate using all 100,000 individuals in the population administrative dataset; adjusted for C and with adjustment for U as described in "scenario 3".

or more of the (many) annual NPD datasets, corresponding to an overall linkage rate of 99.4%.

To illustrate our proposed approach in the NPD data we will address the question of the extent to which pupil mobility (that is, changing schools at non-standard times) is associated with KS2 attainment (age 11) after accounting for other important factors. We conduct the analysis within an NPD whole population dataset with birth years aligned to those in MCS (2000-2001 for births in England). Our target population is therefore children in England born in these years only. However, we want to leverage the MCS-NPD linkage to adjust for MCS only variables. We emphasise that the substantive findings should not be overinterpreted – the analysis is intended more as a proof of concept and exemplar.

The variables that were included in the analysis were as follows:

Exposure (NPD): Pupil mobility (whether pupil joined school after 12 September 2010 (in year 5); binary).

Outcome (NPD): KS2 attainment (average point score (fine grade scores) for English and Mathematics combined; continuous).

NPD control variables: month of birth (4 categories); gender (binary); ethnicity (5 categories); geographical region (9 categories); English as an additional language (EAL) status (binary); special educational needs (SEN) status (3 categories); free school meal (FSM) eligibility

(binary); Income Disadvantage Affecting Children Index (IDACI) (5 categories).

MCS control variables (all observed at MCS Sweep 4, when cohort members were age 7): household income (5 categories); housing tenure (3 categories); parental occupational social class (5 categories); parental education (7 categories).

We restricted our analysis sample to individuals with complete data on all the NPD analysis variables. Missing data in these variables could naturally be handled using MI as part of the MI-based approach outlined above. However, for this exemplar analysis we wished to focus on the use of our proposed approach for handling residual confounding – if we were simultaneously dealing with missing data in other variables then it would not be straightforward to disentangle the impact of the two types of imputation on the findings.

Statistical analysis

We used linear regression to estimate the association between pupil mobility and KS2 attainment. Five different models were fitted:

1. Unadjusted.
2. Adjusted for sociodemographic covariates only (month of birth, gender, ethnicity, EAL status, geographical region).
3. Additionally adjusted for SEN status.
4. Additionally adjusted for socioeconomic status (FSM eligibility, IDACI).
5. Additionally adjusted for cohort socioeconomic variables (household income, housing tenure occupational social class, parental education).

All five models were fitted in the linked MCS-NPD sample for comparison (using only those MCS cohort members in the linked data sample with complete data on all analysis variables). These analyses accounted for the complex survey structure (clustering and oversampling) in the initial MCS sample to make them more representative of the underlying population. In the population NPD data, Models 1-4 were fitted using observed data and Model 5 was fitted using the MI-based approach outlined above. As the MCS cohort members are identified within the population NPD data, the structure corresponds to "scenario 1". However, to provide a full demonstration of the proposed method, we also conducted analyses of the data under scenario 2 (by ignoring the identification of the MCS cohort members within the population NPD data) and scenario 3 (by analysing the linked MCS-NPD separately to the population NPD data). In each case, the cohort socioeconomic variables (household income, housing tenure occupational social class, parental education) were imputed using MICE. The imputation models for each variable were specified as multinomial logistic regression models, each included as explanatory variables all the variables included in the analysis, and we used 50 imputed datasets (chosen to avoid excessive run time in this large dataset). The MCS complex survey structure was not accounted for during imputation or analysis of the imputed datasets.

All analyses were conducted using Stata version 18 [16].

Results

The analysis sample (those with complete data on all NPD analysis variables) included 509,670 individuals in the population NPD data, of whom 7,768 (1.5%) were MCS cohort members. Descriptive statistics are presented for the analysis sample in Table 4.

The estimated associations between pupil mobility and average KS2 points score for the linked MCS-NPD sample and the population NPD data are presented in Fig. 3 and Table S2 (Supplementary Material). In the linked MCS-NPD data, the analysis sample included 7,256 MCS cohort members with complete data on all analysis variables (including the cohort socioeconomic variables, which explains the deviation from the 7,768 reported above). The unadjusted estimate of −1.70 (95% CI −2.24, −1.15) was attenuated to -0.80 (95% CI −1.28, −0.32) through adjustment for the NPD control variables. It was further attenuated to −0.67 (95% CI −1.16, −0.17) through adjustment for the MCS control variables.

In the population NPD data analysis, the unadjusted estimate of −1.86 (95% CI −1.92, −1.81) was attenuated to -0.92 (95% CI −0.97, −0.88) through adjustment for the NPD control variables. It is worth emphasising that this is that point at which the analysis would generally have to conclude, with the analysis adjusting for the NPD control variables perhaps labelled as "fully adjusted" given the lack of information on any further control variables. Through application of our MI-based approach we can go further and additionally adjust for the control variables only observed in MCS. This further attenuated the point estimate to −0.76, though the level of precision differed depending on the assumed scenario, with 95% CIs of (−0.86, −0.67), (−0.89, −0.63) and (−0.90, −0.61) for scenarios 1, 2 and 3, respectively.

Discussion

We have demonstrated an MI-based strategy that allows us to use linked cohort data to help handle confounding in analyses of population administrative data, using simulated and real-world examples.

While performance in a single dataset should not be overinterpreted, the fact that we were able to obtain estimates in our simple simulated example very close to the known truth in each scenario demonstrates the promise of the proposed approach.

Although these are simulated data, it is worth considering how these results would be presented if this were a real-world analysis of population administrative data: the (clearly biased) estimate adjusting for C only would likely be reported as "fully adjusted" given that confounder U was not observed and therefore could not be controlled for. Access to linked cohort data and application of our proposed approach has allowed us to go further in obtaining estimates that represent the true magnitude of the fully adjusted association.

Table 4: Descriptive statistics for analysis of National Pupil Database records linked to Millennium Cohort Study data and population National Pupil Database data (analysis sample)

	Linked MCS-NPD sample (N = 7,256)[A]			Population NPD data (N = 509,670)[B]	
	n	%[C]	%[W]	n	%
NPD variables					
Pupil mobility: pupil joined school after 12 September 2010 (in year 5)					
No	6,887	94.9	95.1	477,573	93.7
Yes	369	5.1	4.9	32,097	6.3
Month of birth					
Sep-Nov	1,864	25.7	25.9	129,330	25.4
Dec-Feb	1,780	24.5	24.3	124,428	24.4
Mar-May	1,830	25.2	25.1	126,814	24.9
Jun-Aug	1,782	24.6	24.7	129,098	25.3
Gender					
Female	3,586	49.4	49.4	249,704	49.0
Male	3,670	50.6	50.6	259,966	51.0
Ethnicity					
Any other ethnic group/unclassified	104	1.4	1.0	10,487	2.1
Asian	958	13.2	6.0	47,306	9.3
Black	258	3.6	2.0	24,180	4.7
Mixed	280	3.9	3.2	21,319	4.2
White	5,656	78.0	87.8	406,378	79.7
English as an additional language status					
First language English	6,169	85.0	93.0	435,645	85.5
English as an additional language	1,087	15.0	7.0	74,025	14.5
Special educational needs (SEN)					
No identified SEN	5,683	78.3	79.6	386,308	75.8
SEN without a statement	1,400	19.3	18.1	107,876	21.2
SEN with a statement	173	2.4	2.4	15,486	3.0
Socioeconomic disadvantage – student level: entitlement to a free school meal (FSM)					
False	6,149	84.7	87.9	416,753	81.8
True	1,107	15.3	12.1	92,917	18.2
Socioeconomic disadvantage – neighbourhood level: Income Disadvantage Affecting Children Index (IDACI) quintile					
1 Least deprived	1,281	17.7	22.8	92,838	18.2
2	1,296	17.9	21.6	95,489	18.7
3	1,342	18.5	20.8	95,569	18.8
4	1,466	20.2	18.0	102,339	20.1
5 Most deprived	1,871	25.8	16.8	123,435	24.2
Geographical region					
East Midlands	647	8.9	9.4	44,101	8.7
East of England	883	12.2	13.4	57,493	11.3
London	979	13.5	10.2	72,590	14.3
North East	344	4.7	4.6	25,278	5.0
North West	912	12.6	12.2	71,167	14.0
South East	1,168	16.1	18.9	80,605	15.8
South West	665	9.2	11.1	48,358	9.5
West Midlands	824	11.4	9.7	57,533	11.3
Yorkshire and the Humber	834	11.5	10.5	52,545	10.3

Continued

Table 4: Continued

	Linked MCS-NPD sample (N = 7,256)[A]					Population NPD data (N = 509,670)[B]		
	n	%[C]		%[W]		n	%	
MCS variables								
Income								
£4,700	1,347	18.6		15.8				
£4,700-£15,600	1,613	22.2		19.2				
£15,600-£26,000	1,496	20.6		20.3				
£26,000-£36,400	1,313	18.1		19.9				
>£36,400	1,487	20.5		24.8				
Tenure								
Own	4,844	66.8		70.8				
Rent	2,267	31.2		27.5				
Other	145	2.0		1.7				
Occupational social class								
Managerial and professional	3,193	44.0		49.4				
Intermediate	1,017	14.0		14.3				
Small employers and self-employed	705	9.7		9.0				
Lower supervisory and technical	579	8.0		7.6				
Semi-routine and routine	1,762	24.3		19.7				
Education								
NVQ1	605	8.3		7.4				
NVQ2	2,443	33.7		33.9				
NVQ3	646	8.9		9.4				
NVQ4	2,083	28.7		31.9				
NVQ5	611	8.4		9.0				
Overseas qual only	181	2.5		1.5				
None of these	687	9.5		6.9				
		Mean[C]	SD[C]	Mean[W]	SD[W]		Mean	SD
KS2 average point score (fine grade scores)	7,256	28.7	4.6	28.9	4.5	509,670	28.3	4.8
Prior attainment: KS1 average points score (reading, writing and maths)	7,256	15.5	3.7	15.8	3.6	509,670	15.2	3.9

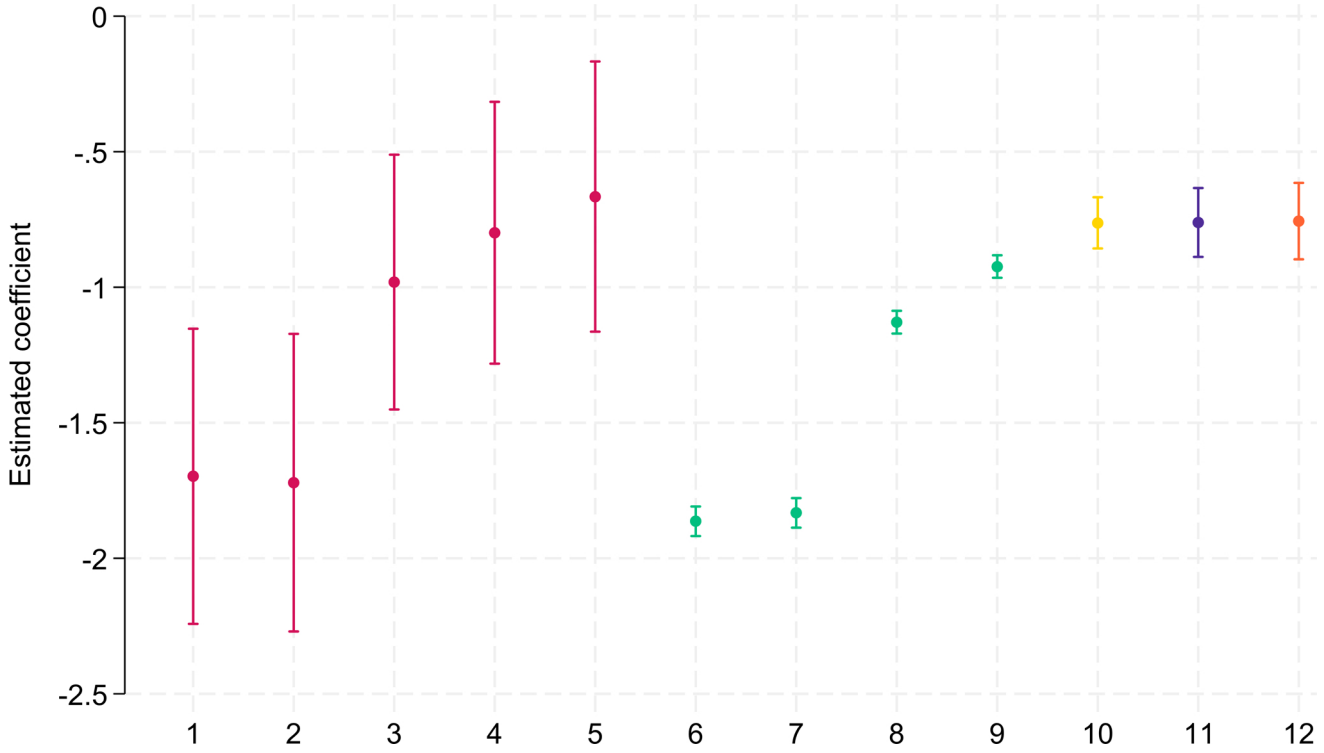
KS: Key Stage. NVQ: National Vocational Qualification.
[A] Those with data on all NPD variables and all MCS variables included in the analysis.
[B] Those with data on all NPD variables included in the analysis.
[C] Crude (does not account for survey structure).
[W] Weighted (appropriately accounts for all survey structure).

Our MCS-NPD example showed that additional adjustment for cohort-only variables resulted in further attenuation of the estimated association between pupil mobility and KS2 attainment. While the qualitative conclusions from the analyses with and without adjustment for cohort socioeconomic variables may be similar in this instance (that pupil mobility is associated with a lower average KS2 points score), the additional attenuation of 17% is not insubstantial. Comparable attenuation may well be sufficient to affect the qualitative conclusions in other settings.

A further notable feature of this analysis is the relatively wider 95% CIs when adjusting for the cohort socioeconomic variables using the MI-based approach compared to the model not additionally adjusting for these variables: approximately double the width for scenario 1 through to treble the width for scenario 3. Given the linked MCS-NPD sample constitutes only 1.5% of the population NPD data – and that therefore the MCS control variables are being imputed for 98.5% of the population dataset even under scenario 1 – it is perhaps unsurprising (and entirely appropriate) that there is a commensurate reduction in precision.

There are a few other features this analysis of population NPD data that are worth brief discussion. MCS only includes children born in the UK whereas the NPD covers all children who are in England at that point in time, some of whom will have been born outside the UK. It is unclear how important this slight disconnect is in practice, but this could be explored further by performing a sensitivity analysis in which the NPD analysis sample is restricted to children born in UK (though this would require further linkage to acquire country of birth data). Furthermore, linked data are only available for MCS cohort members who consented to linkage (and had their data

Figure 3: Estimated association between pupil mobility (joined current school in Year 5 or later) and average Key Stage 2 points score using National Pupil Database records linked to Millennium Cohort Study data (N = 7,768) and population National Pupil Database data (N = 509,670)



Points are estimates; bars are 95% confidence intervals.

1. Estimate using the 7,768 Millennium Cohort Study cohort members with linked National Pupil Database records only; unadjusted.
2. Estimate using the 7,768 Millennium Cohort Study cohort members with linked National Pupil Database records only; adjusted for sociodemographic covariates only (month of birth, gender, ethnicity, English as an additional language, geographical region).
3. Estimate using the 7,768 Millennium Cohort Study cohort members with linked National Pupil Database records only; additionally adjusted for special educational needs status.
4. Estimate using the 7,768 Millennium Cohort Study cohort members with linked National Pupil Database records only; additionally adjusted for socioeconomic status (free school meal eligibility, neighbourhood socioeconomic deprivation).
5. Estimate using the 7,768 Millennium Cohort Study cohort members with linked National Pupil Database records only; additionally adjusted for cohort socioeconomic variables (household income, housing tenure occupational social class, parental education).
6. Estimate using all 509,670 individuals in the population National Pupil Database dataset; unadjusted.
7. Estimate using all 509,670 individuals in the population National Pupil Database dataset; adjusted for sociodemographic covariates only (month of birth, gender, ethnicity, English as an additional language, geographical region).
8. Estimate using all 509,670 individuals in the population National Pupil Database dataset; additionally adjusted for special educational needs status.
9. Estimate using all 509,670 individuals in the population National Pupil Database dataset; additionally adjusted for socioeconomic status (free school meal eligibility, neighbourhood socioeconomic deprivation).
10. Estimate using all 509,670 individuals in the population National Pupil Database dataset; additionally adjusted for cohort socioeconomic variables as described in "scenario 1".
11. Estimate using all 509,670 individuals in the population National Pupil Database dataset; additionally adjusted for cohort socioeconomic variables as described in "scenario 2".
12. Estimate using all 509,670 individuals in the population National Pupil Database dataset; additionally adjusted for cohort socioeconomic variables as described in "scenario 3".

successfully linked, though the high linkage rate makes this less of a concern). If such individuals are a selected subset of all MCS cohort members, the effect on the imputation and population-level analysis should be given further consideration. However, it is worth noting that the levels of the estimated associations and the patterns of attenuation through the sequential adjustment are very similar in the linked MCS-NPD sample and the population NPD data. This gives us some confidence that the cohort data are well aligned with the population administrative data in this exemplar analysis, going some way to justifying the plausibility of the MAR assumption. However, in substantive analyses this would require further detailed consideration.

Finally, when analysing MCS survey data, the complex survey structure (clustering and oversampling of certain population subgroups) should be accounted for in order for inferences to be reflective of the study target population. We did not account for the complex survey structure during imputation or analysis of the imputed datasets. The latter seems appropriate as analysis of the population data does not rely on the MCS design. Whether and how the MCS complex survey structure might be accounted for in the imputation phase requires further consideration.

The key assumption underlying our MI-based approach is that the cohort-only variables are MAR in the population administrative data given the observed data (i.e. given the variables in the population administrative data). More precisely, this states that U is independent of S given Y, X and C, where S signifies cohort members (with U observed) vs. the non-cohort members (with U unobserved) in the population administrative data (other notation as defined in the simple simulated example). If this assumption did not hold, and there remained expected differences in U between the cohort members and the rest of the sample even after taking into account Y, X and C, then, intuitively, our approach would not provide appropriate imputed values of U.

In the presence of missing data, we cannot definitively identify the missingness mechanism from the observed data alone. However, exploratory analyses of the observed data can help assess whether the MAR assumption may be plausible. In general, recommended approaches include summaries of (near-)fully observed variables across missingness patterns and logistic regression of missingness indicators on (near-)fully observed variables [9]. In the present context, assuming relatively complete administrative records and relatively complete cohort-only variables among cohort members, the missingness pattern will largely reduce to cohort members vs. non-cohort members.

In our simple simulated example, where the individuals in the cohort study were essentially a random subset of those in the population administrative data, we knew that the MAR (and, indeed, the missing completely at random (MCAR)) assumption holds. In the analysis of NPD data, we considered whether the linked MCS-NPD sample is likely to be similarly well aligned to population NPD data. In general, it will be important to consider a variety of factors such as the alignment of the target populations underlying both the cohort study and the population administrative data and selection arising at various stages from processes including initial participation in the cohort study, attrition from the cohort study prior to linkage consent being sought, consent to linkage, eligibility for linkage and successful linkage [23]. The plausibility of the MAR assumption will always require careful consideration: this might not a trivial task and requires good knowledge of both data sources.

In practice, it is impossible to know that data truly are MAR and therefore we might wish to explore how robust our results are to the MAR assumption. A variety of sensitivity analyses have been proposed, which typically involve imputing data under a hypothesised missing not at random (MNAR) mechanism, for example using a pattern-mixture modelling approach [9].

We emphasise that in application of this approach, the usual MI considerations also apply. For example, one should consider: imputation model compatibility and specification, including non-linearities and interactions; the number of imputations required; and the inclusion of auxiliary variables [11]. Standard MI quality control measures, such as checks on model convergence, trace data and distributions of imputed values, should also be implemented.

Our proposed approach in scenario 3, when the linked cohort data and population administrative data cannot be analysed together in the same location, involves the export of perturbed regression coefficients from the SDE containing the linked cohort data and the import of those coefficients into the other SDE. Regression coefficients would typically be in scope for what can be safely (i.e. without danger of disclosure) be exported from an SDE. The fact that the regression coefficients are perturbed (as part of the MI process) should further reduce any perceived danger of disclosure.

Framing unobserved confounding variables in the population administrative data as a missing data problem and using MI allows us to utilise existing theory, methods and software implementations. Other approaches to missing data handling, such as inverse probability weighting (IPW) and Bayesian modelling, could also be considered. MI is often preferred to IPW, as it is usually more efficient [24]. In the present setting, IPW would reweight only the cohort members and would not use information from non-cohort members on the relationships between variables in the administrative data, so would likely be particularly inefficient relative to MI. Bayesian modelling could potentially obtain similar results to MI, but is typically less straightforward to implement relative to the MI implementations readily available in standard statistical software. The use of MICE, in particular, allows partially observed (i.e. cohort-only) variables to be imputed using an imputation model appropriate to their distribution (e.g. linear, logistic or multinomial logistic regression for continuous, binary or categorical variables, respectively).

We focused our attention here on linkage with cohort studies. While much of our discussion relates to any form of survey data, cohort studies have several key advantages relative to cross-sectional surveys. Longitudinal observation of the same construct in a cohort study, over a period of years or even decades, allows better characterisation of the construct than would observation at a single point in time. In particular, long-running birth cohorts allow characterisation of constructs at different stages of the life course. This is an important consideration when aiming to address (residual) confounding, as multiple stages of the life course may be

of relevance. The nature of cohort data also makes it more likely that potential confounding variables will be observed prior to an exposure of interest. In addition, it is more likely that cohort studies, often forming substantial research infrastructure with ongoing support and investment, have undertaken the necessary linkages with administrative data.

Our simple simulated example had complete data on all variables other than those only observed in the cohort study and our application using population NPD data required complete data on all the NPD analysis variables so that we could focus on the proposed method for handling residual confounding. However, missing data in other variables could naturally be handled using MI as part of the proposed MI-based approach. This constitutes a major advantage of the approach since such missingness will almost always be an additional issue to handle in this type of analysis.

Although we have framed our proposed approach as a solution to the problem of residual confounding in analyses of population administrative data, we note that the same approach could in principle also be used if the variables observed only in the cohort study were instead other variables of interest, perhaps most plausibly potential mediating variables.

Future work

In order to understand whether and how this approach can be applied more broadly, a comprehensive simulation study would be helpful. This could examine the performance of the method when varying factors such as the relative sizes of the linked cohort sample and population administrative dataset (i.e. the proportion of missing data to be imputed) and misalignment of the linked cohort sample and population administrative dataset in terms of the populations that they represent (i.e. transgressions of the MAR assumption). Sensitivity in this specific setting to factors that have already been examined in conventional contexts, such as correct model specification and the strength of association between partially observed (i.e. cohort-only) variables and fully observed (i.e. population administrative data) variables, may also be of benefit.

As previously noted, the variance may not be correctly estimated in scenarios 2 and 3 when using Rubin's rules [13, 14]. Further work is required to examine the potential implications of this in the present context.

While readily available MI implementations are sufficient to implement the proposed approach when the linked cohort data and population administrative data can be analysed in the same location (scenarios 1 and 2), further work is required to programme a generalised and user-friendly MI solution for when the two datasets must be kept separate (scenario 3). A recently developed Stata command "mi impute from" [25], based on the conditional quantile imputation approach [26], appears to be a promising alternative to the MICE implementation described in this paper. However, the restriction to quantile regression when imputing a continuous variable may not always be desirable and the current implementation only allows imputation of one variable at a time.

Conclusion

Residual confounding is likely to remain a concern for analysis of population administrative data, since these data are not collected primarily for research purposes and do not always include all of the information we might need. By utilising information from linked cohort data, our strategy allows us to handle such residual confounding. In a simple simulated example and a real-world example, the approach works well, but further research is required to understand whether and how it can be applied more broadly.

Acknowledgements

This work was supported by the Economic & Social Research Council and Administrative Data Research UK [grant number ES/V006037/1] and the Medical Research Council [grant number MC_UU_00004/07], and the UCL Centre for Longitudinal Studies is supported by the Economic & Social Research Council [grant number ES/W013142/1]. The funders 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.

The Millennium Cohort Study is only possible due to the commitment and enthusiasm of the cohort members and their families; their time and contribution is gratefully acknowledged.

Statement on conflicts of interest

None declared.

Ethics statement

Ethical approval for each sweep of Millennium Cohort Study (MCS) data collection has been obtained from an NHS Research Ethics Committee. Informed consent for participation has been obtained from parents, as well as from the children themselves as they have grown up. Ethical approval for the fourth sweep of MCS data collection, when cohort members were age 7, was obtained from the Yorkshire MREC [07/MRE03/32].

Education records up to age 16 were linked to MCS survey data on the basis of parent/carer consents collected at the fourth sweep of MCS data collection, when cohort members were age 7. Ethics approval for this linkage was obtained from the London-Hampstead NRES [REC-14/LO/0868].

Data availability statement

Stata code to generate (and analyse) the data for the simple simulated example is available at https://osf.io/mqnzx/.

Millennium Cohort Study (MCS) survey data are available via a standard End User Licence Agreement from the UK Data Service (Study Number 2000031): http://doi.org/10.5255/UKDA-Series-2000031.

Linked MCS and National Pupil Database (NPD) data are available via Secure Access from the UK Data Service (Study Number 8481): http://doi.org/10.5255/UKDA-SN-8481-3.

Whole population NPD data are made available for research purposes by the Department for Education, including via Secure Access from the UK Data Service (Study Number 2000108): http://doi.org/10.5255/UKDA-Series-2000108.

References

1. 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

2. 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.

3. Ministry of Housing Communities & Local Government. English indices of deprivation 2019. 2019 [Available from: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019. Accessed: May 2025.].

4. Brown EM, Franklin SM, Ryan JL, Canterberry M, Bowe A, Pantell MS, et al. Assessing Area-Level Deprivation as a Proxy for Individual-Level Social Risks. Am J Prev Med. 2023;65(6):1163-71. https://doi.org/10.1016/j.amepre.2023.06.006

5. Davis-Kean P, Chambers RL, Davidson LL, Kleinert C, Ren Q, Tang S. Longitudinal Studies Strategic Review: 2017 Report to the Economic and Social Research Council. ESRC; 2018.

6. Rajah N, Calderwood L, De Stavola BL, Harron K, Ploubidis GB, Silverwood RJ. Using linked administrative data to aid the handling of non-response and restore sample representativeness in cohort studies: the 1958 national child development study and hospital episode statistics data. BMC Medical Research Methodology. 2023;23(1):266. https://doi.org/10.1186/s12874-023-02099-w

7. Greenland S, Lash TL. Bias analysis. In: Rothman KJ, Greenland S, Lash TL, editors. Modern Epidemiology. Third ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

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

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

10. Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in medicine. 1999;18(6):681-94. https://doi.org/10.1002/(SICI)1097-0258(19990330)18:6%3C681::AID-SIM71%3E3.0.CO;2-R

11. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30(4):377-99. https://doi.org/10.1002/sim.4067

12. StataCorp. Stata 18 Multiple-Imputation Reference Manual. College Station, TX: Stata Press; 2023.

13. Robins JM, Wang N. Inference for Imputation Estimators. Biometrika. 2000;87(1):113-24. https://doi.org/10.1093/biomet/87.1.113

14. Hughes RA, Sterne JAC, Tilling K. Comparison of imputation variance estimators. Statistical Methods in Medical Research. 2014;25(6):2541-57. https://doi.org/10.1177/0962280214526216

15. Royston P, White IR. Multiple Imputation by Chained Equations (MICE): Implementation in Stata. Journal of Statistical Software. 2011;45(4):1 - 20. https://doi.org/10.18637/jss.v045.i04

16. StataCorp. Stata Statistical Software: Release 18. College Station, TX: StataCorp LLC; 2023.

17. Connelly R, Platt L. Cohort Profile: UK Millennium Cohort Study (MCS). International Journal of Epidemiology. 2014;43(6):1719-25. https://doi.org/10.1093/ije/dyu001

18. Joshi H, Fitzsimons E. The Millennium Cohort Study: the making of a multi-purpose resource for social science and policy. Longitudinal and Life Course Studies. 2016;7(4):409-30. https://doi.org/10.14301/llcs.v7i4.410

19. Jay MA, McGrath-Lone L, Gilbert R. Data Resource: the National Pupil Database (NPD). Int J Popul Data Sci. 2018;4(1):08. https://doi.org/10.23889/ijpds.v4i1.1101

20. University College London, UCL Institute of Education, Centre for Longitudinal Studies, Department for Education. Millennium Cohort Study: Linked Education Administrative Datasets (National Pupil Database), England: Secure Access. [data collection]. 2nd Edition. UK Data Service. SN: 8481, http://doi.org/10.5255/UKDA-SN-8481-2

21. Rihal S, Gomes D. Millennium Cohort Study: A guide to the linked education administrative datasets (2nd edition). London: UCL Centre for Longitudinal Studies; 2021.

22. Department for Education. National pupil database. 2020 [Available from: https://www.gov.uk/government/collections/national-pupil-database. Accessed 15 May 2020.].

23. Silverwood RJ, Rajah N, Calderwood L, De Stavola BL, Harron K, Ploubidis GB. 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. Int J Popul Data Sci. 2024;9(1):2137. https://doi.org/10.23889/ijpds.v9i1.2137

24. Seaman SR, White IR, Copas AJ, Li L. Combining Multiple Imputation and Inverse-Probability Weighting. Biometrics. 2012;68(1):129-37. https://doi.org/10.1111/j.1541-0420.2011.01666.x

25. Thiesmeier T, Bottai M, Orsini N. Imputing Missing Values with External Data. arXiv. 2024:arXiv:2410.02982. https://doi.org/10.48550/arXiv.2410.02982

26. Thiesmeier R, Bottai M, Orsini N. Systematically missing data in distributed data networks: multiple imputation when data cannot be pooled. Journal of Statistical Computation and Simulation. 2024;94(17):3807-25. https://doi.org/10.1080/00949655.2024.2404220

Abbreviations

EAL:	English as an additional language
FSM:	Free school meal
IDACI:	Income Disadvantage Affecting Children Index
KS:	Key Stage
MAR:	Missing at random
MCS:	Millennium Cohort Study
MI:	Multiple imputation
MICE:	Multiple imputation by chained equations
NPD:	National Pupil Database
OSF:	Open Science Framework
SDE:	Secure data environment
SEN:	Special educational needs

Supplementary material

Methods S1. Data generating process for the simple simulated example.

Here we describe the data generating process for the simple simulated example with a single continuous variable U which is recorded in the cohort study but not in the population administrative dataset. The data generating processes for the other simple simulated examples (binary U, categorical U and multiple normally distributed U) were similar. Stata code for the simulation of all the simple simulated example datasets and the analyses corresponding to scenarios 1, 2 and 3 for each is available at https://osf.io/mqnzx/.

For the 100,000 individuals in the population administrative dataset:

1. U and C were drawn from a bivariate standard normal distribution, with correlation 0.5.
2. X was simulated as a function of U and C using a logit function with parameters 0.5 and 0.5 respectively (no constant). X therefore has an expected prevalence of 0.5.
3. Y was simulated as a function of U, C and X using a logit function with parameters 0.5, 0.5 and 1 respectively, with a constant of -0.5. Y therefore has an expected prevalence of 0.5.

U was then recoded to missing for non-cohort members.

Table S1: Estimated X-Y associations in the simple simulated example with continuous fully observed confounder C and continuous cohort-only confounder U

	Log-odds ratio	
Model	**Estimate**	**95% CI**
1	0.985	0.956, 1.014
2	1.454	1.371, 1.538
3	1.157	1.069, 1.245
4	1.022	0.931, 1.112
5	1.434	1.408, 1.461
6	1.125	1.098, 1.153
7	0.990	0.957, 1.024
8	0.990	0.955, 1.026
9	0.992	0.957, 1.026

1. Estimate using all 100,000 individuals in the population administrative dataset, fully adjusted for both C and U (i.e. prior to U being recoded to missing for the non-cohort members).
2. Estimate using the 10,000 cohort members only; unadjusted.
3. Estimate using the 10,000 cohort members only; adjusted for C only.
4. Estimate using the 10,000 cohort members only; adjusted for C and U only.
5. Estimate using all 100,000 individuals in the population administrative dataset; unadjusted.
6. Estimate using all 100,000 individuals in the population administrative dataset; adjusted for C only.
7. Estimate using all 100,000 individuals in the population administrative dataset; adjusted for C and with adjustment for U as described in "scenario 1".
8. Estimate using all 100,000 individuals in the population administrative dataset; adjusted for C and with adjustment for U as described in "scenario 2".
9. Estimate using all 100,000 individuals in the population administrative dataset; adjusted for C and with adjustment for U as described in "scenario 3".

Table S2: Regression models for the association between pupil mobility (joined current school in Year 5 or later) and average Key Stage 2 point score

	Linked MCS-NPD sample (N = 7,256)		Population NPD data (N = 509,670)	
Model	**Coeff**	**95% CI**	**Coeff**	**95% CI**
1	−1.70	−2.24, −1.15	−1.86	−1.92, −1.81
2	−1.72	−2.27, −1.17	−1.83	−1.89, −1.78
3	−0.98	−1.45, −0.51	−1.13	−1.17, −1.09
4	−0.80	−1.28, −0.32	−0.92	−0.97, −0.88
5	−0.67	−1.16, −0.17	−0.76	−0.86, −0.67
6	−0.67	−1.16, −0.17	−0.76	−0.89, −0.63
7	−0.67	−1.16, −0.17	−0.76	−0.90, −0.61

CI: confidence interval; MCS: Millennium Cohort Study; NPD: National Pupil Database.

1. Unadjusted.
2. Adjusted for sociodemographic covariates only (month of birth, gender, ethnicity, geographical region, English as an additional language, geographical region).
3. Additionally adjusted for special educational needs status.
4. Additionally adjusted for socioeconomic status (free school meal eligibility, neighbourhood socioeconomic deprivation).
5. Additionally adjusted for cohort socioeconomic variables (household income, housing tenure occupational social class, parental education) (following multiple imputation for population NPD data analysis as described in "scenario 1").
6. Additionally adjusted for cohort socioeconomic variables (household income, housing tenure occupational social class, parental education) (following multiple imputation for population NPD data analysis as described in "scenario 2").
7. Additionally adjusted for cohort socioeconomic variables (household income, housing tenure occupational social class, parental education) (following multiple imputation for population NPD data analysis as described in "scenario 3").

Figure S1: Overview of proposed approach under scenario 2

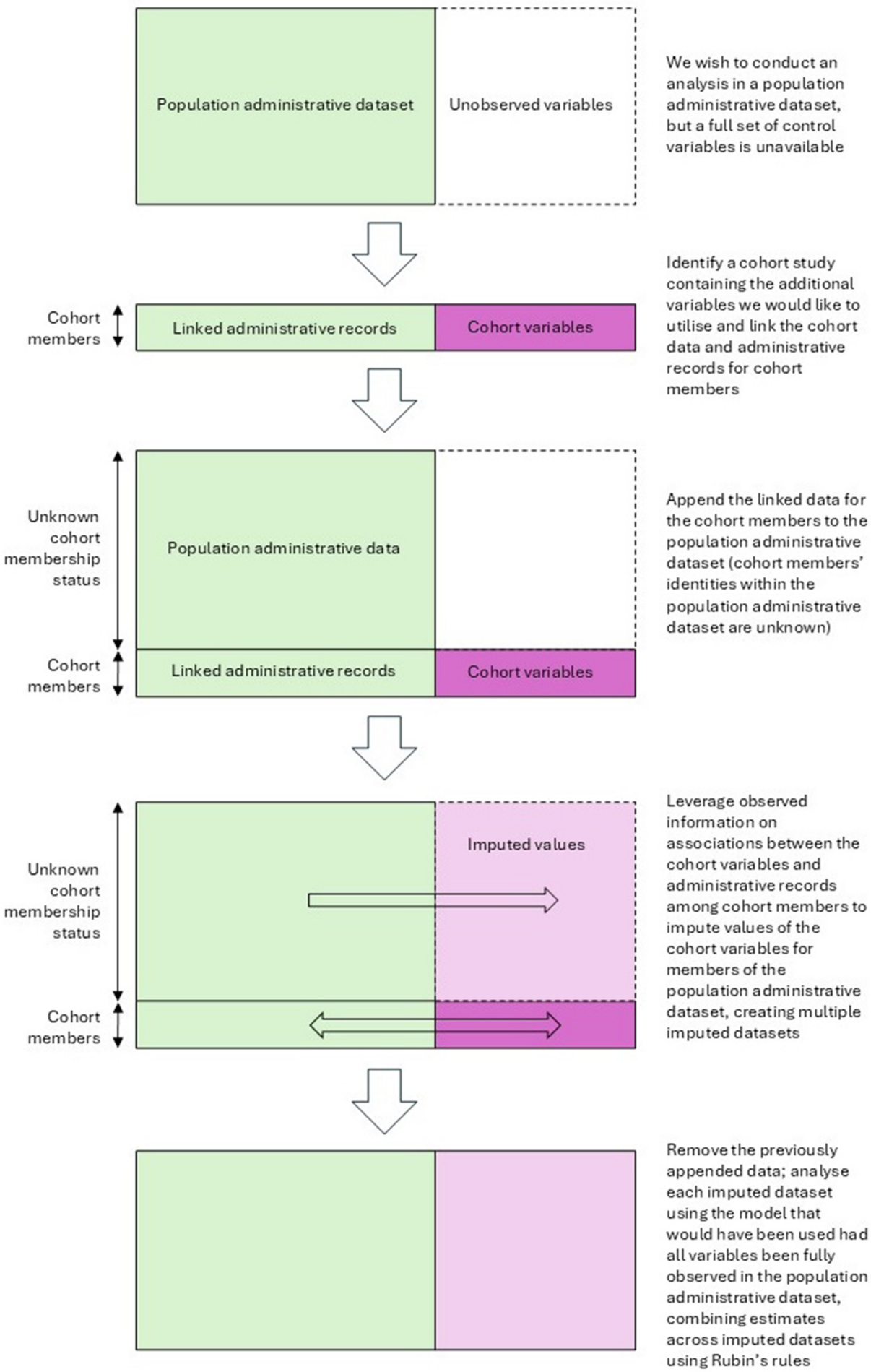


Figure S2: Overview of proposed approach under scenario 3

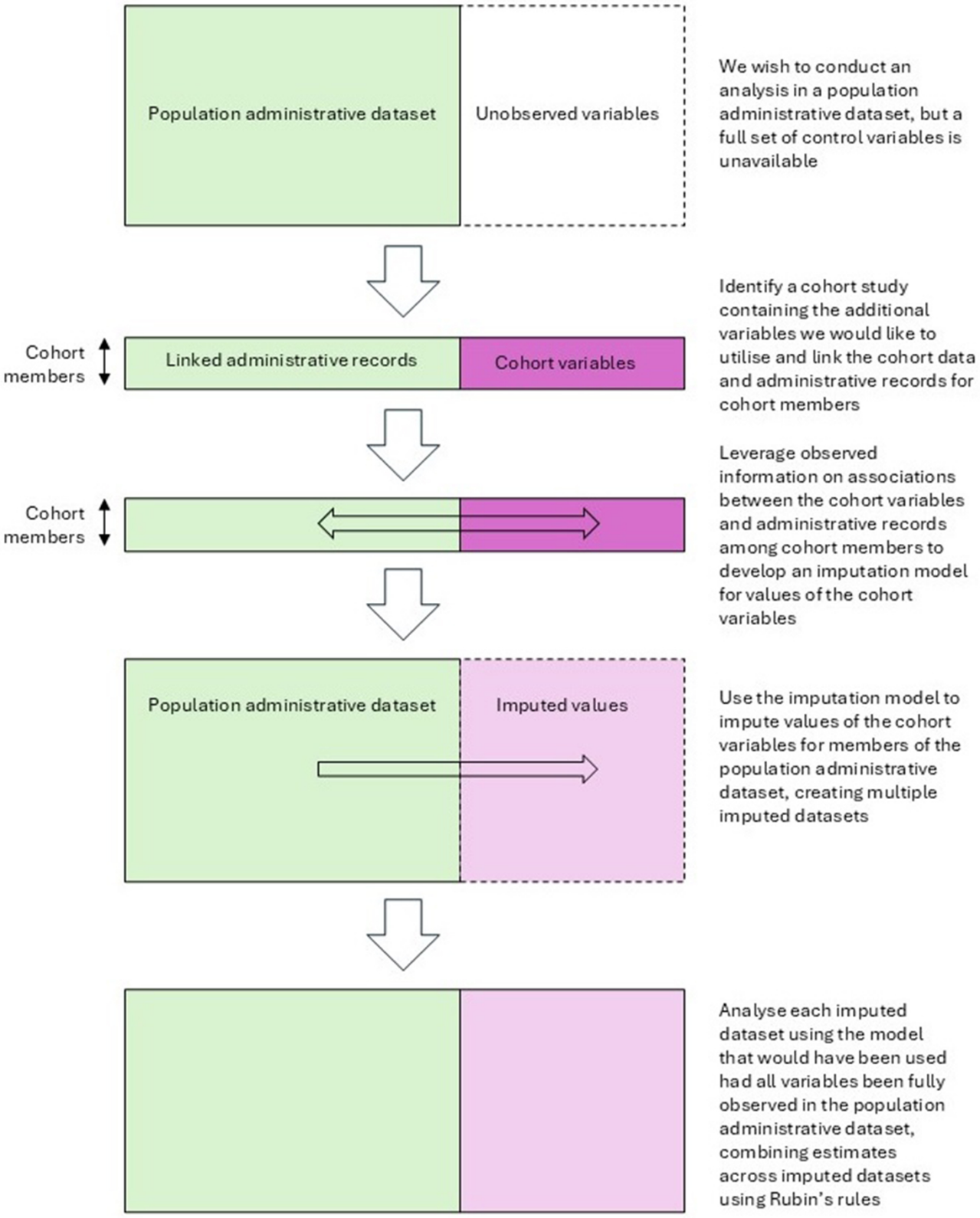
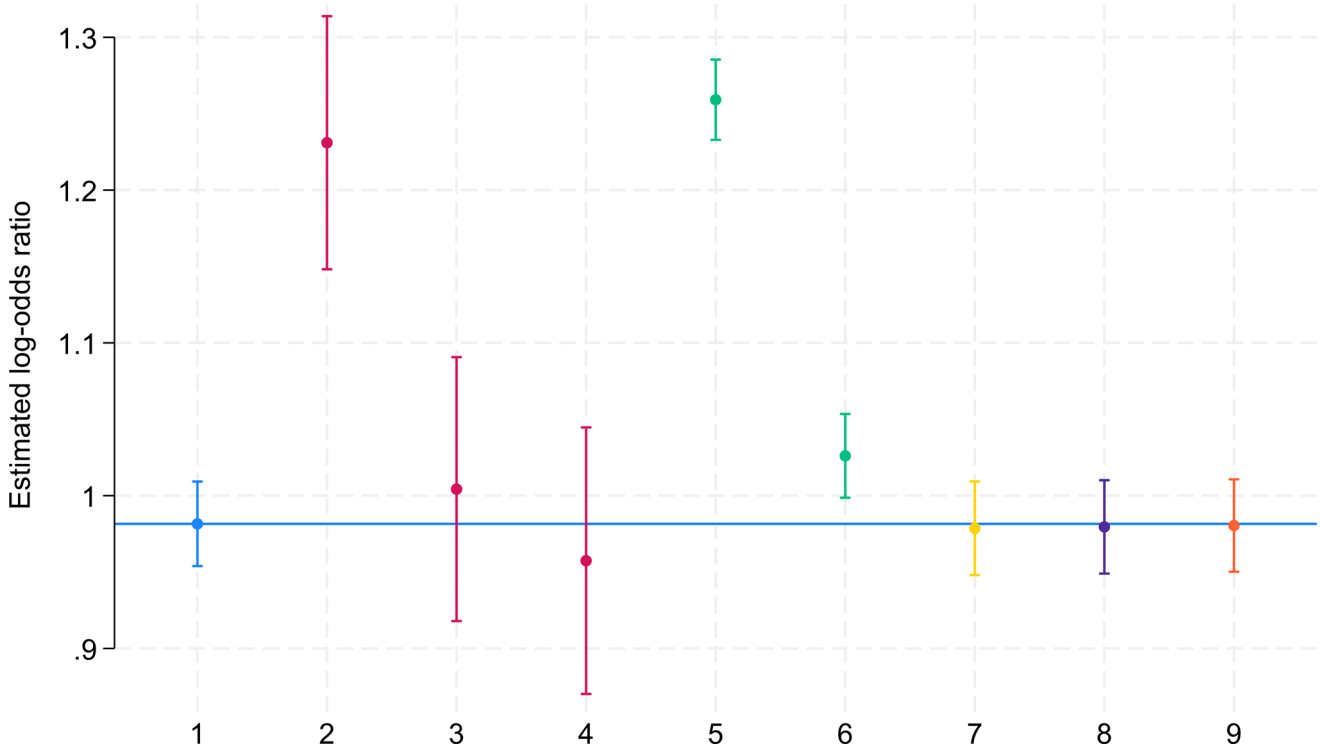


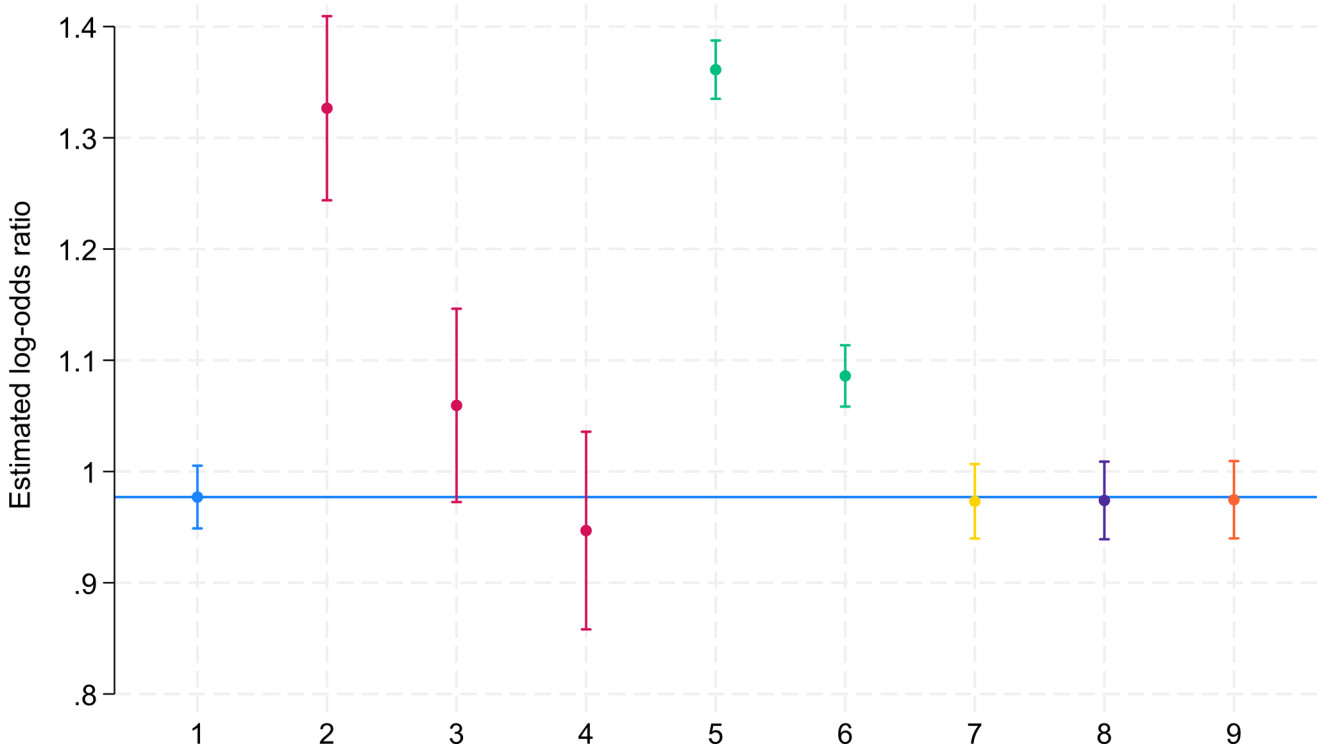
Figure S3: Estimated X-Y associations in the simple simulated example with normally distributed fully observed confounder C and binary cohort-only confounder U



Points are estimates; bars are 95% confidence intervals.

1. Estimate using all 100,000 individuals in the population administrative dataset, fully adjusted for both C and U (i.e. prior to U being recoded to missing for the non-cohort members).
2. Estimate using the 10,000 cohort members only; unadjusted.
3. Estimate using the 10,000 cohort members only; adjusted for C only.
4. Estimate using the 10,000 cohort members only; adjusted for C and U only.
5. Estimate using all 100,000 individuals in the population administrative dataset; unadjusted.
6. Estimate using all 100,000 individuals in the population administrative dataset; adjusted for C only.
7. Estimate using all 100,000 individuals in the population administrative dataset; adjusted for C and with adjustment for U as described in "scenario 1".
8. Estimate using all 100,000 individuals in the population administrative dataset; adjusted for C and with adjustment for U as described in "scenario 2".
9. Estimate using all 100,000 individuals in the population administrative dataset; adjusted for C and with adjustment for U as described in "scenario 3".

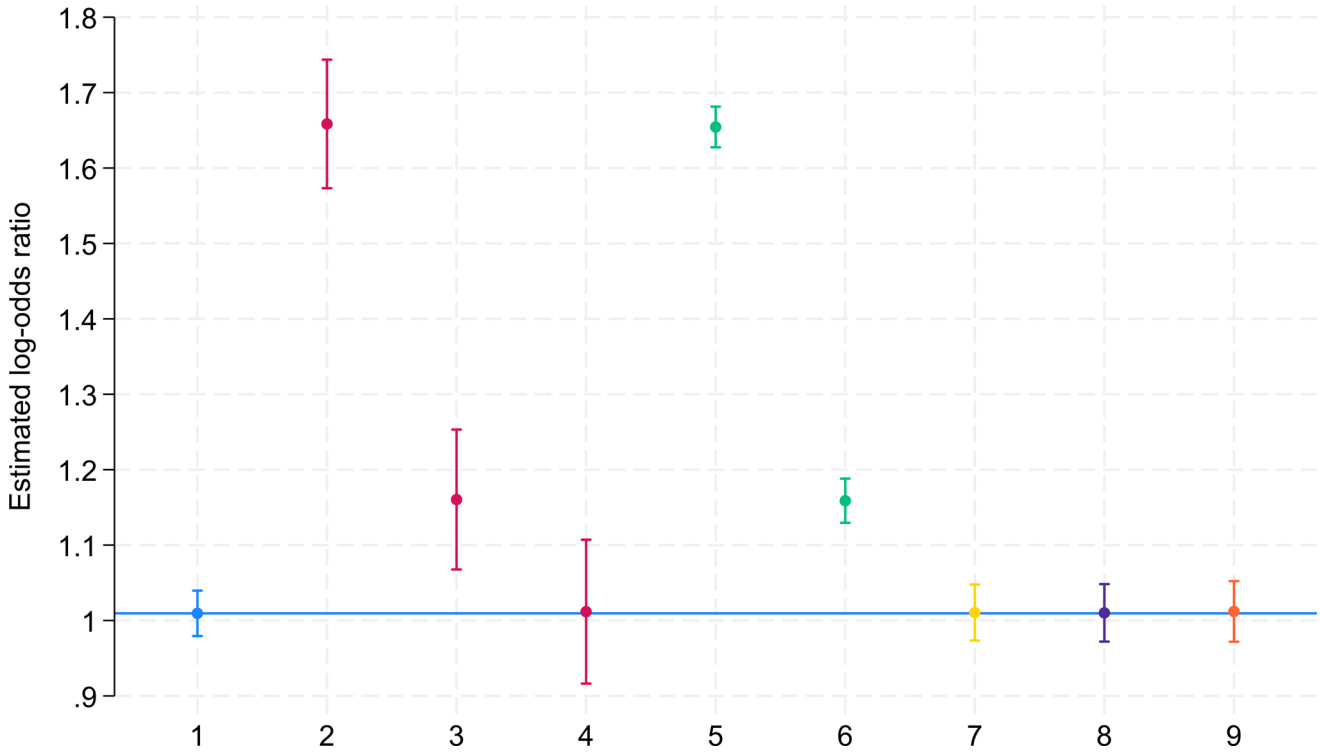
Figure S4: Estimated X-Y associations in the simple simulated example with normally distributed fully observed confounder C and categorical cohort-only confounder



Points are estimates; bars are 95% confidence intervals.

1. Estimate using all 100,000 individuals in the population administrative dataset, fully adjusted for both C and U (i.e. prior to U being recoded to missing for the non-cohort members).
2. Estimate using the 10,000 cohort members only; unadjusted.
3. Estimate using the 10,000 cohort members only; adjusted for C only.
4. Estimate using the 10,000 cohort members only; adjusted for C and U only.
5. Estimate using all 100,000 individuals in the population administrative dataset; unadjusted.
6. Estimate using all 100,000 individuals in the population administrative dataset; adjusted for C only.
7. Estimate using all 100,000 individuals in the population administrative dataset; adjusted for C and with adjustment for U as described in "scenario 1".
8. Estimate using all 100,000 individuals in the population administrative dataset; adjusted for C and with adjustment for U as described in "scenario 2".
9. Estimate using all 100,000 individuals in the population administrative dataset; adjusted for C and with adjustment for U as described in "scenario 3".

Figure S5: Estimated X-Y associations in the simple simulated example with multiple normally distributed fully observed confounders (C_1 and C_2) and multiple normally distributed cohort-only confounders (U_1 and U_2)



Points are estimates; bars are 95% confidence intervals.

1. Estimate using all 100,000 individuals in the population administrative dataset, fully adjusted for both C and U (i.e. prior to U being recoded to missing for the non-cohort members).
2. Estimate using the 10,000 cohort members only; unadjusted.
3. Estimate using the 10,000 cohort members only; adjusted for C only.
4. Estimate using the 10,000 cohort members only; adjusted for C and U only.
5. Estimate using all 100,000 individuals in the population administrative dataset; unadjusted.
6. Estimate using all 100,000 individuals in the population administrative dataset; adjusted for C only.
7. Estimate using all 100,000 individuals in the population administrative dataset; adjusted for C and with adjustment for U as described in "scenario 1".
8. Estimate using all 100,000 individuals in the population administrative dataset; adjusted for C and with adjustment for U as described in "scenario 2".
9. Estimate using all 100,000 individuals in the population administrative dataset; adjusted for C and with adjustment for U as described in "scenario 3".