International Journal of Population Data Science

Journal Website: www.ijpds.org

Administrative data linkage to Census 2021 in Wales, UK: A cross-sectional study examining completeness and representativeness for population analytics

Jane Lyons[1,†], Rhodri D. Johnson[1,†*], Michael Edwards[1], Samantha Turner[1], Richard Fry[1], Lucy J Griffiths[1], and Ronan A. Lyons[1]

Submission History

Submitted:	11/04/2025
Accepted:	03/07/2025
Published:	26/11/2025

[1] Population Data Science, Swansea University Medical School, Swansea, SA2 8PP
[†] Joint first authors

Abstract

Introduction
Measuring population representativeness is an important methodological step in public health and epidemiological studies.

Objectives
To explore the representativeness of Census 2021 data linkage when compared with the Welsh Demographic Service Dataset (WDSD) within the Secure Anonymised Information Linkage (SAIL) Databank for research on the population of Wales, UK. To understand the characteristics of individuals linked and not linked and which subgroups of the population are disproportionately represented in data linkage population-wide studies.

Methods
An observational, population-wide cross-sectional comparison study, utilising administrative demographic data and decennial survey data held in SAIL. Two data sources, the WDSD and Census 2021, were used to create and compare two cohorts of the resident population of Wales, UK, on 21st March 2021.

The two cohorts were linked to understand how many individuals from Census 2021 can be successfully linked within SAIL, in WDSD and not in Census 2021, and found across both sources. Logistic regression models analysed the variation in the linkability of the survey data within SAIL by various demographic and household characteristics.

Results
The central analytical cohort contained 2,440,191 individuals present in both data sources. WDSD contained 3,090,976 individuals with 2,965,196 individuals in Census data.

With a positively classed outcome indicating non-linkage from WDS to Census the characteristics associated with the highest odds of individuals being registered in WDS but not linked to Census (in SAIL) are male (aOR = 1.28 [95%CI 1.28,1.32]), 75+ years of age (aOR = 1.27 [95%CI 1.25,1.29]), of Asian ethnicity (aOR = 1.27 [95%CI 1.24,1.30]), a more recent migrant (arriving to UK after 2000) (aOR = 1.30 [95%CI 1.28,1.32]), a member of the LGBTQ+ community (aOR = 1.29 [95%CI 1.25,1.29]) or not disclosing LGBTQ+ status (aOR = 1.41 [95%CI 1.39,1.43]), being separated, divorced or widowed (aOR = 1.28 [95%CI 1.27,1.29]), or living in rental accommodation (aOR = 1.47 [95%CI 1.45,1.48]).

Conclusions
Results show that certain personal characteristics and sub-groups of the population of Wales are disproportionately represented when combining population estimates and utilising Census data in data linkage population-wide studies in SAIL.

Keywords
data linkage; census representativeness; administrative data

*Corresponding Author:
Email Address: r.d.johnson@swansea.ac.uk (Rhodri D. Johnson)

https://doi.org/10.23889/ijpds.v6i1.2994

Introduction

Measuring population representativeness is an important methodological step in public health and epidemiological studies as it ensures a selected and often restricted study population accurately reflects the characteristics of a wider or complete population [1]. Addressing population representativeness improves the credibility and scientific rigour of studies because it accounts for variability within the population, quantifies bias and ensures results can be meaningfully applied to the true population. Clarifying the extent to which different socio-demographic groups are represented and whether reported positive or negative outcomes can be applied to all groups of the population is an important aspect of social justice research. Such research also helps identify disparities in exposures and outcomes that can be used to inform social policy.

Access to total population data for research purposes can be limited for many reasons, including data availability and coverage, information governance and ethical considerations, and the need for privacy protection and secure storage of large data sources containing personal information [2, 3]. Research supporting infrastructures, such as secure e-research platforms and trusted research environments, overcome many of these barriers by securely housing large data repositories and allowing accredited researchers' access for approved research purposes [4, 5]. Through these technologies and data linkage analytics it is possible to access and link data from multiple sources on various populations. Linking data from different sources allows for the creation of more comprehensive datasets for analyses, therefore broadening insight on the understanding within a particular area of interest and enhancing decision-making [6, 7]. Linked data also supports longitudinal analysis as well as helping validate the accuracy and completeness of data by crosschecking information. However, linkage of datasets is rarely 100% complete. In such cases it is essential that there is understanding of how combining multiple disparate data may affect the research processes, study design choices, and interpretation of results.

The Census 2021 for England and Wales, which took place on the 21st March 2021, aims to be a complete enumeration survey, compulsory for the resident population of England and Wales to take part, and the latest iteration of the Census, completed every ten years. The purpose of the Census is to capture a total population snapshot of England and Wales to provide population-wide data and information to the UK government to develop policies, plan service needs, and allocate funds based on population demographics [8]. Census 2021 data has recently become available for research purposes within the Secure Anonymised Information Linkage (SAIL) Databank [9]. This provides a unique opportunity to identify and report any differences between the Census 2021 survey population characteristics, and the subset of data that can be linked to routinely collected electronic healthcare records (EHRs) and administrative data population characteristics held in the Welsh Demographic Service Dataset (WDSD), to understand potential biases when utilising these data for population research purposes.

The WDS contains individuals registered with a General Practitioner (GP) in Wales; within SAIL it allows creation of flexible, longitudinal, population-level cohorts to which other data sources, such as electronic health records, can be linked [10, 11]. However, there are known limitations on patient registration counts e.g., due to absence of or delayed GP registrations such as for recent migrants or transient populations [12]. Census is also not considered a full enumeration [13]; it achieved a 97% response rate in 2021, and included challenges related to under-coverage, for example in some harder to reach populations, and over-coverage due to inclusion of emigrants, and duplicated records [12]. Both data sources are important and have varied strengths and weaknesses within the research context. Additionally there are known limitations associated with data linkage and accurate assignation of linkage fields to individuals, and we describe the algorithms developed and applied in both data sources [14].

The objective of this study was therefore to explore the completeness and representativeness of Census 2021 data linkage within the SAIL Databank, a Trusted Research Environment hosted by the UK Secure eResearch Platform (SeRP) for research on the population of Wales, UK.

Methods

Study design

In this observational population-wide cross sectional comparison study, we utilised anonymised and encrypted routinely collected administrative NHS demographic data and Census 2021 data held in the SAIL Databank to create and compare two estimates of the resident population of Wales, UK, on the 21st March 2021, from these two different data sources [11, 15]. We linked these data sources and used logistic regression models to analyse the variation in the linkage of the Census 2021 data within SAIL, by various demographic and household characteristics.

Data sources

All data accessed for this study was made available through the SAIL Databank, an ISO27001 accredited trusted research environment [9]. SAIL contains billions of rows of anonymised, encrypted, individual and household-level health, demographic, social, and environmental data for the population of Wales, UK. SAIL uses a multiple encryption system in which a trusted third party, Digital Health and Care Wales (DHCW), uniquely matches personal identifiable information contained within an acquired data source to the Welsh Demographic Service Dataset (WDSD) to create an Anonymised Linkage Field (ALF) per individual and a Residential Anonymised Linkage Field (RALF) per residence before uploading data to SAIL [11, 15–17]. In SAIL, unique linkage fields (ALF) are used to anonymously link individuals and/or households between data sources.

The WDSD contains administrative information and demographic characteristics, such as age, sex, as well as residential information such as the RALF and the Lower layer Super Output Area (LSOA) of each RALF for the population of Wales registered with a Welsh General Practice which is available for all individuals within WDSD. The presence of the LSOA allows linkage to small area statistics such as the Welsh Index of Multiple Deprivation (WIMD) to include

area-level deprivation for socioeconomic categorisation in analyses [18].

The Census 2021 contains individual level information, such as demographics, protected characteristics (defined within the UK Equality Act 2010 as: age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex, sexual orientation, pregnancy and maternity), health status, household level information, such as household size and deprivation, and whether individuals resided in communal establishments for the population of Wales on the 21st March 2021 [19, 20].

Census 2021 linkage algorithm

Linkage between Census and WDS at the individual level was completed with probabilistic linkage techniques using personal identifiable information (PII) which included names (forename, middle, and surname), date of birth, and postcode. The model allowed for small variations in spelling of names, date of birth, and postcode regions, and assigned match weightings to an edge based on the likelihood of matching. An initial pre-linkage stage included identification of intra-Census de-duplicating links where constraints were placed such that links could not be created for records for individuals within a residence (based on having the same Census Questionnaire ID number). The aim of this approach was to improve linkage within residences where individuals may have similar personal information, such as non-singleton siblings, parents with same names as children, or within large communal establishments such as university halls of residence. The two sets of links were combined to produce the final links joining Census to WDS using the anonymised linking field (ALF) [11]. Records in WDS not deemed active (deceased prior to, or born after) the Census year were not considered for linkage.

We calculated linkage match rates, defined as the total number of individuals in the Census data who could be linked to WDSD using the ALF within the date range as described below.

Participants

Individuals were allocated to analysis cohorts according to their inclusion in the five permutations of the linking of individuals in WDSD and Census 2021 as shown in Box 1.

Box 1: Cohort descriptions

Cohort	Cohort description	Description
A	WDS	Individuals present in WDS
B	Census	Individuals present in Census 2021
C	WDS & Census	Individuals present in WDS and Census 2021 (linked on ALF)
D	WDS only	Individuals present in WDS only (not linked on ALF)
E	Census only	Individuals present in Census 2021 only (not linked on ALF)

All individuals alive and living in Wales on the 21st March 2021 were included. Initially, two cohorts containing the population of Wales were created using each data source (Figure 1). Cohort A (WDS only) contained all individuals within the WDSD that were alive and living in Wales on the 21st March, and registered to a Welsh General Practice (n = 3,092,152). A period of 180-days prior and 60-days post census date was used to allow for delays to address registration data within WDS, the shorter follow up period applied to restrict post census address moves. Cohort B (Census only) contained all individuals present in the Census 2021 Wales dataset, who were classified as usual residents and were not wholly imputed records (n = 2,966,372). A usual resident is defined as anyone who on the census day was in the UK and had stayed or intended to stay in the UK for a period of 12 months or more or had a permanent UK address and was outside the UK and intended to be outside the UK for less than 12 months. Wholly imputed records were excluded as these records are created to resolve survey non-response and therefore are not based on actual individuals and could never be linked to any records in SAIL.

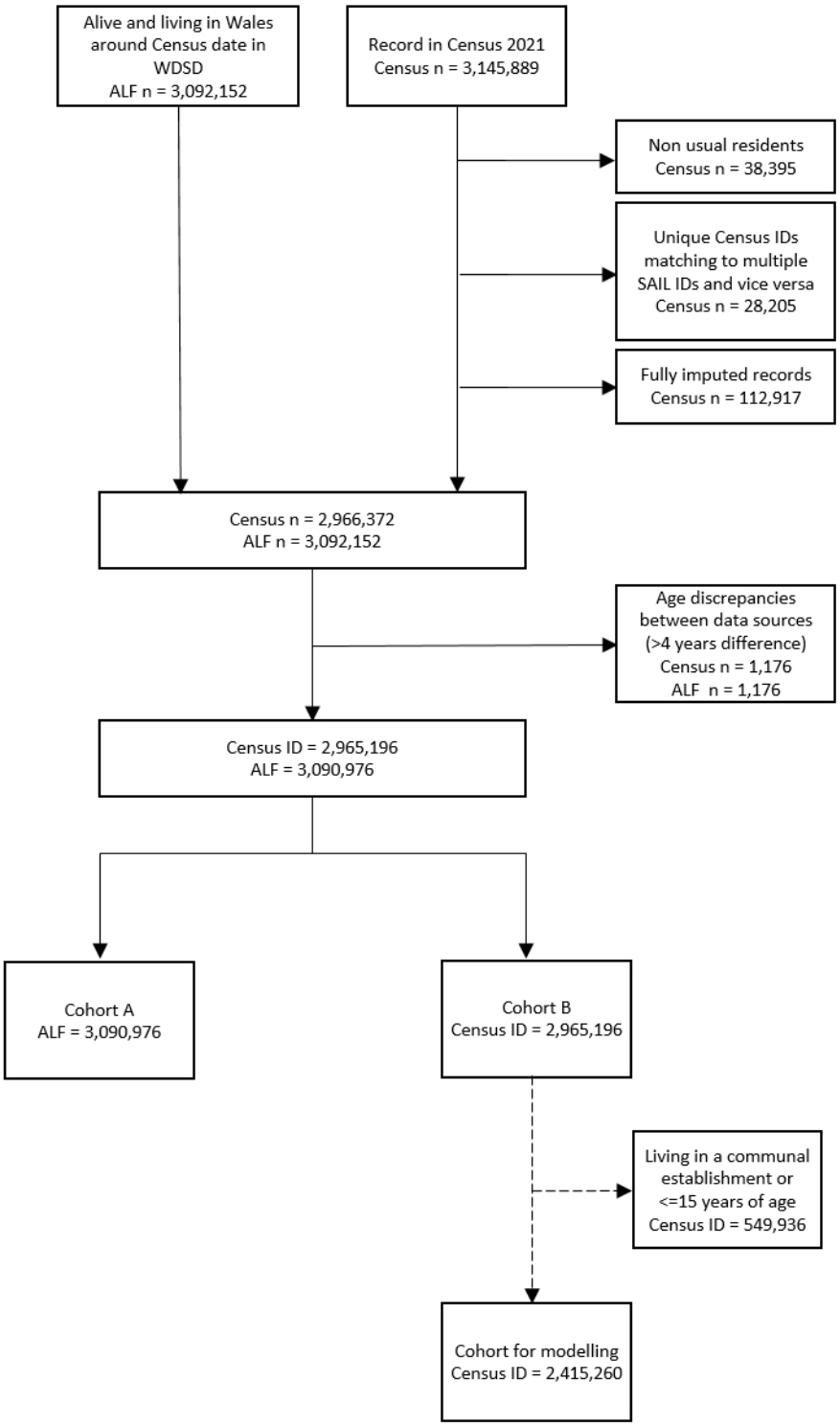
Population comparisons

To compare the completeness and representativeness of the Census 2021 data linkage, Cohort A (WDS only) and Cohort B (Census only) were linked on ALF to create Cohort C (WDS and Census) (Figures 1 & 2) ($n = 2{,}440{,}191$). Cohort C provides the details on how many individuals and their characteristics from Census 2021 can be successfully linked and demonstrates the proportion of individuals that are found across both population estimates. All remaining individuals in Cohort A that are not present in Cohort C are categorised as Cohort D ($n = 650{,}785$) (WDS only), which equates to all individuals who reside in Wales based on the WDSD but cannot be found in the Census 2021 population estimate. All remaining individuals in Cohort B that are not present in Cohort C are categorised as Cohort E ($n = 525{,}005$) (Census only), which equates to all individuals present in the Census 2021 but who cannot be linked to the WDSD population estimate. Mutual variables that are found in each cohort were used to compare the population representativeness (Table 1).

Statistical methods

Descriptive statistics were calculated for all sociodemographic and household variables per cohort comparison, including the calculation of ratios of the proportions of individuals within specific cohorts. Binary logistic regression was used to examine which sociodemographic and household characteristics influence whether individuals in the Census 2021 can be linked within SAIL and are available for longitudinal research. Due to some Census questions having an age restriction for submitting information, the main analysis focuses on individuals aged 16 years and over and not living in a communal establishment (Table 3), with full cohort comparisons available in the supplementary materials (Supplementary Table 1). The binary outcome of interest was individuals present in the Census 2021 that could be linked (an ALF could be created) and therefore used in

Figure 1: Consort diagram of study participant inclusion



longitudinal research (Cohort C - WDS and Census) compared to individuals present in the Census 2021 that could not be linked (Cohort E - Census Only). Statistical significance was set at $p < 0.05$ and the analysis was performed using the stats package in R version 4.1.3.

Census 2021 variable values were assigned as missing if they contained any of the following lookup descriptions: Missing, Failed Multi Tick, Not Answered, Unknown: Insufficient Information, Not Required. For Cohort B logistic modelling missing data values for $n = 487$ individuals were imputed using mode imputation for three variables (Activity Last Week, Highest Qualification, and Marital Status).

Results

In total, 3,615,981 individuals aged 0–105+ years where included in this cross-sectional cohort comparison study.

Figure 2: Venn diagram of the creation of cohorts

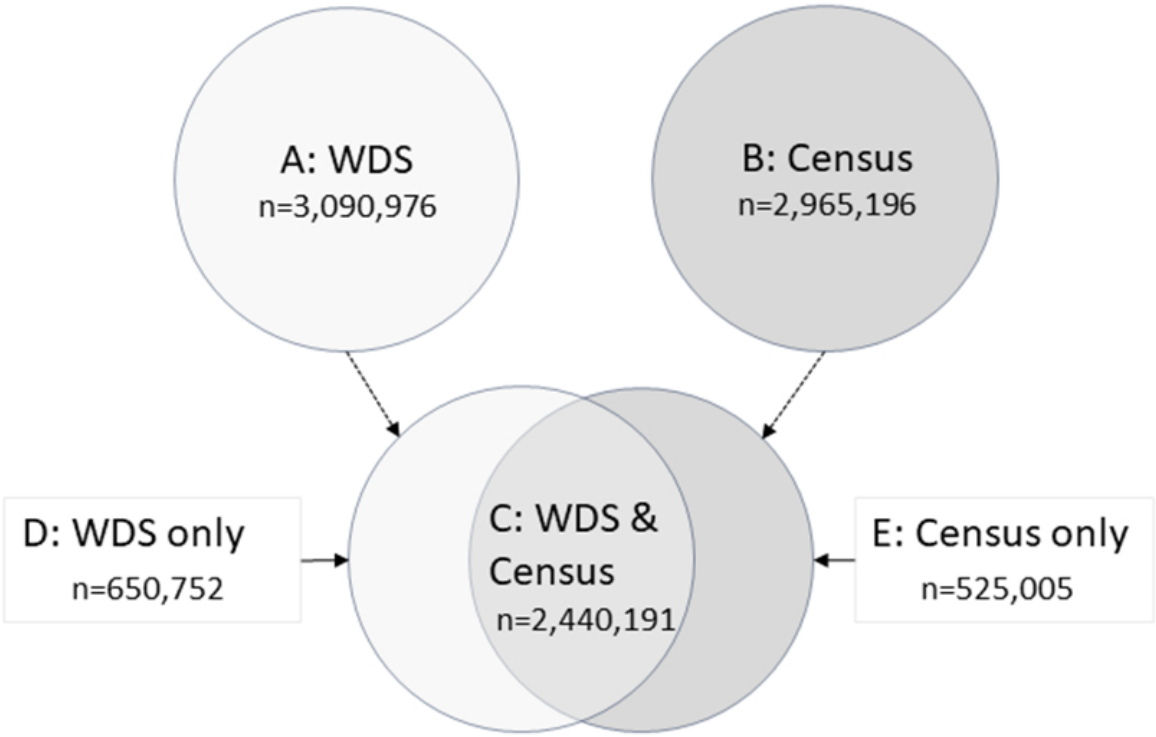


Table 1: Study cohort comparisons

Cohort comparison	Variables to include in comparison
WDS and Census (C) v WDS (A)	Age, Sex, WIMD deprivation, rurality, Local Authority
WDS and Census (C) v Census (B)	Age, Sex, and Census variables: Ethnicity, LGBTQ+ status, migrant status (year arrive to UK), marital status, religion, highest qualification, activity last week, carer status, disability, ownership of property, and household deprivation
WDS and Census (C) v WDS Only (D)	Age, Sex, WIMD deprivation, rurality, Local Authority
WDS and Census (C) v Census only (E)	Age, Sex, and Census variables: Ethnicity, LGBTQ+ status, migrant status (year arrive to UK), marital status, religion, highest qualification, activity last week, carer status, disability, ownership of property, and household deprivation
WDS only (D) v Census only (E)	Age, and Sex

Overall, cohort A (WDS) contained 3,090,976 individuals, 49.9% were male, age-group proportions ranged between 16.3-20.7% for each 15-year age-band between 0–74 year olds with the exception of 9.6% for 75+ year olds. Deprivation quintiles, at the area-level, measured using the 2019 Welsh Index of Multiple Deprivation (WIMD) [13] showed very similar distribution across the population with 20.5% of individuals living in the most deprived areas compared to 19.8% of individuals living in the least deprived areas of Wales (Table 2). The majority of individuals resided in urban areas (69.6%), with Cardiff being the most populous local authority (12%).

Cohort B (Census) included 2,965,196 individuals. 48.7% were males, and age-group proportions ranged between 17.1-20.7% for each 15-year age-band between 0-74 years (Table 2). The majority of individuals were white (94.2%), born in the UK (93.3%), with 2.6% identified as LGBTQ+, and 37.1% married or in a civil partnership. Some 46.3% of the population did not identify as being part of a religion, and the most commonly attained highest qualification level in the cohort was an intermediate level qualification (38.2%, including apprenticeships, 'A' Levels, and vocational qualifications). Working was the most common activity recorded at the time of the Census (44.6%) compared to those that were economically active or unemployed (Table 2). In addition, one in ten people reported being a carer (10.2%), 28.2% had a disability, nearly a third reported renting their accommodation (31.8%), and based on Census household deprivation, half the population (50.8%) declared being deprived in one or more dimension (deprived in either education, employment, health, and/or housing measures) (Table 2). Cohort C included 2,440,191 individuals with very similar population distributions as described about for Cohort A and B (Table 2).

The overall match rate from Census (Cohort B) to WDS (Cohort A) was 82.3%. Match rates by selected demographic factors, and for types of communal establishments are reported in Supplementary Tables (Supplementary Tables 2 and 3). Regarding cohort D (WDS only), the WDS sub-population who could not be linked to the Census 2021, this contained 650,785 individuals. 57.9% were male and proportionally more individuals were aged 16-45 years (47%) compared to

Table 2: Cohort characteristics and ratios of the proportions of individuals within specific cohorts

Characteristics		WDS (A)	Census (B)	WDS and Census (C)	WDS only (D)	Census only (E)	A:C	B:C	D:C	E:C	D:E
Sex	Male	1,543,134 (49.9%)	1,443,371 (48.7%)	1,166,262 (47.8%)	376,872 (57.9%)	277,109 (52.8%)	1.04	1.02	1.21	1.1	1.1
	Female	1,547,842 (50.1%)	1,521,825 (51.3%)	1,273,929 (52.2%)	273,913 (42.1%)	247,896 (47.2%)	0.96	0.98	0.81	0.9	0.89
Age group	0-15	525,321 (17%)	509,846 (17.2%)	438,172 (18%)	87,149 (13.4%)	71,674 (13.7%)	0.94	0.96	0.74	0.76	0.98
	16-30	549,441 (17.8%)	514,868 (17.4%)	394,175 (16.2%)	155,266 (23.9%)	120,693 (23%)	1.1	1.07	1.48	1.42	1.04
	31-45	576,178 (18.6%)	526,279 (17.7%)	425,546 (17.4%)	150,632 (23.1%)	100,733 (19.2%)	1.07	1.02	1.33	1.1	1.2
	46-60	640,987 (20.7%)	614,058 (20.7%)	517,685 (21.2%)	123,302 (18.9%)	96,373 (18.4%)	0.98	0.98	0.89	0.87	1.03
	61-74	503,530 (16.3%)	505,577 (17.1%)	423,955 (17.4%)	79,575 (12.2%)	81,622 (15.5%)	0.94	0.98	0.7	0.89	0.79
	75+	295,519 (9.6%)	294,568 (9.9%)	240,658 (9.9%)	54,861 (8.4%)	53,910 (10.3%)	0.97	1	0.85	1.04	0.82
Ethnicity	Asian, Asian British Or Asian Welsh		80,302 (2.7%)	60,053 (2.5%)		20,249 (3.9%)		1.08		1.56	
	Black, Black British, Black Welsh, Caribbean Or African		24,128 (0.8%)	17,463 (0.7%)		6,665 (1.3%)		1.14		1.86	
	Missing		0 (0%)	0 (0%)		0 (0%)					
	Mixed Or Multiple Ethnic Groups		44,745 (1.5%)	35,376 (1.4%)		9,369 (1.8%)		1.07		1.29	
	Other Ethnic Group		23,858 (0.8%)	18,213 (0.7%)		5,645 (1.1%)		1.14		1.57	
	White		2,792,163 (94.2%)	2,309,086 (94.6%)		483,077 (92%)		1		0.97	
LGBQT+ status	Missing		725,757 (24.5%)	599,676 (24.6%)		126,081 (24%)		1		0.98	
	Person Is LGBTQ+		77,430 (2.6%)	58,297 (2.4%)		19,133 (3.6%)		1.08		1.5	
	Person Is Not LGBTQ+		2,162,009 (72.9%)	1,782,218 (73%)		379,791 (72.3%)		1		0.99	
Year arrived UK	Arrived After 2000		138,179 (4.7%)	99,978 (4.1%)		38,201 (7.3%)		1.15		1.78	
	Arrived Up To 2000		59,826 (2%)	48,519 (2%)		11,307 (2.2%)		1		1.1	
	Born In UK		2,767,191 (93.3%)	2,291,694 (93.9%)		475,497 (90.6%)		0.99		0.96	
	Missing		0 (0%)	0 (0%)		0 (0%)					

Continued

Table 2: Continued

Characteristics		WDS (A)	Census (B)	WDS and Census (C)	WDS only (D)	Census only (E)	A:C	B:C	D:C	E:C	D:E
Marital status	Married/ Civil Partnership		1,101,207 (37.1%)	933,492 (38.3%)		167,715 (31.9%)		0.97		0.83	
	Missing		477,847 (16.1%)	410,086 (16.8%)		67,761 (12.9%)		0.96		0.77	
	Never Married/Civil Partnership		921,328 (31.1%)	721,373 (29.6%)		199,955 (38.1%)		1.05		1.29	
	Separated/Divorced/ Widowed		464,814 (15.7%)	375,240 (15.4%)		89,574 (17.1%)		1.02		1.11	
Religion	Christian		1,308,664 (44.1%)	1,094,057 (44.8%)		214,607 (40.9%)		0.98		0.91	
	Missing		183,653 (6.2%)	141,873 (5.8%)		41,780 (8%)		1.07		1.38	
	No Religion		1,371,452 (46.3%)	1,125,277 (46.1%)		246,175 (46.9%)		1		1.02	
	Non-Christian Religion		101,427 (3.4%)	78,984 (3.2%)		22,443 (4.3%)		1.06		1.34	
Highest qualification	Higher Level Qualifications		779,151 (26.3%)	634,356 (26%)		144,795 (27.6%)		1.01		1.06	
	Intermediate		1,133,231 (38.2%)	937,270 (38.4%)		195,961 (37.3%)		0.99		0.97	
	Lower Level Qualifications		566,072 (19.1%)	473,365 (19.4%)		92,707 (17.7%)		0.98		0.91	
	No Qualifications		486,742 (16.4%)	395,200 (16.2%)		91,542 (17.4%)		1.01		1.07	
Activity last week	Economically Inactive		1,058,298 (35.7%)	863,049 (35.4%)		195,249 (37.2%)		1.01		1.05	
	Missing		510,302 (17.2%)	438,628 (18%)		71,674 (13.7%)		0.96		0.76	
	Unemployed		73,481 (2.5%)	57,829 (2.4%)		15,652 (3%)		1.04		1.25	
	Working		1,323,115 (44.6%)	1,080,685 (44.3%)		242,430 (46.2%)		1.01		1.04	
Is carer	Carer		301,715 (10.2%)	255,425 (10.5%)		46,290 (8.8%)		0.97		0.84	
	Missing		143,099 (4.8%)	119,843 (4.9%)		23,256 (4.4%)		0.98		0.9	
	Not A Carer		2,520,382 (85%)	2,064,923 (84.6%)		455,459 (86.8%)		1		1.03	
Disability	Disability		837,547 (28.2%)	696,690 (28.6%)		140,857 (26.8%)		0.99		0.94	
	Missing		0 (0%)	0 (0%)		0 (0%)					
	No Disability		2,127,649 (71.8%)	1,743,501 (71.4%)		384,148 (73.2%)		1.01		1.03	

Continued

Table 2: Continued

Characteristics		WDS (A)	Census (B)	WDS and Census (C)	WDS only (D)	Census only (E)	A:C	B:C	D:C	E:C	D:E
Accommodation ownership	Home Owner (Owns Outright)		958,445 (32.3%)	810,728 (33.2%)		147,717 (28.1%)		0.97		0.85	
	Home Owner (With Mortgage)		1,024,398 (34.5%)	868,144 (35.6%)		156,254 (29.8%)		0.97		0.84	
	Missing		0 (0%)	0 (0%)		0 (0%)					
	Rental		941,681 (31.8%)	742,944 (30.4%)		198,737 (37.9%)		1.05		1.25	
Household deprivation	Household Deprived In One Or More Dimension		1,507,526 (50.8%)	1,249,283 (51.2%)		258,243 (49.2%)		0.99		0.96	
	Household Not Deprived In Any Dimension		1,416,998 (47.8%)	1,172,533 (48.1%)		244,465 (46.6%)		0.99		0.97	
	Missing		40,672 (1.4%)	18,375 (0.8%)		22,297 (4.2%)		1.75		5.25	
WIMD deprivation	Most deprived	633,569 (20.5%)		465,728 (19.1%)	167,841 (25.8%)		1.07		1.35		
	2	616,255 (19.9%)		479,898 (19.7%)	136,357 (21%)		1.01		1.07		
	3	618,122 (20%)		486,867 (20%)	131,255 (20.2%)		1		1.01		
	4	609,961 (19.7%)		497,589 (20.4%)	112,372 (17.3%)		0.97		0.85		
	Least Deprived	613,069 (19.8%)		510,109 (20.9%)	102,960 (15.8%)		0.95		0.76		
	Missing	0 (0%)		0 (0%)	0 (0%)						
Rurality	Rural Town And Fringe	406,910 (13.2%)		328,373 (13.5%)	78,537 (12.1%)		0.98		0.9		
	Rural Town And Fringe In A Sparse Setting	114,268 (3.7%)		90,620 (3.7%)	23,648 (3.6%)		1		0.97		
	Rural Village And Dispersed	198,983 (6.4%)		163,953 (6.7%)	35,030 (5.4%)		0.96		0.81		
	Rural Village And Dispersed In A Sparse Setting	218,025 (7.1%)		178,927 (7.3%)	39,098 (6%)		0.97		0.82		

Continued

Table 2: Continued

Characteristics		WDS (A)	Census (B)	WDS and Census (C)	WDS only (D)	Census only (E)	A:C	B:C	D:C	E:C	D:E
	Urban City And Town	2,096,840 (67.8%)		1,636,463 (67.1%)	460,377 (70.7%)		1.01		1.05		
	Urban City And Town In A Sparse Setting	55,950 (1.8%)		41,855 (1.7%)	14,095 (2.2%)		1.06		1.29		
	Missing	0 (0%)		0 (0%)	0 (0%)						
Local Authority	Blaenau Gwent	68,846 (2.2%)		54,376 (2.2%)	14,470 (2.2%)		1		1		
	Bridgend	142,533 (4.6%)		115,607 (4.7%)	26,926 (4.1%)		0.98		0.87		
	Caerphilly	176,653 (5.7%)		143,921 (5.9%)	32,732 (5%)		0.97		0.85		
	Cardiff	372,296 (12%)		275,863 (11.3%)	96,433 (14.8%)		1.06		1.31		
	Carmarthenshire	184,995 (6%)		149,788 (6.1%)	35,207 (5.4%)		0.98		0.89		
	Ceredigion	68,771 (2.2%)		52,537 (2.2%)	16,234 (2.5%)		1		1.14		
	Conwy	113,738 (3.7%)		89,826 (3.7%)	23,912 (3.7%)		1		1		
	Denbighshire	95,959 (3.1%)		75,081 (3.1%)	20,878 (3.2%)		1		1.03		
	Flintshire	155,382 (5%)		124,090 (5.1%)	31,292 (4.8%)		0.98		0.94		
	Gwynedd	106,720 (3.5%)		82,067 (3.4%)	24,653 (3.8%)		1.03		1.12		
	Isle Of Anglesey	65,317 (2.1%)		51,949 (2.1%)	13,368 (2.1%)		1		1		
	Merthyr Tydfil	61,338 (2%)		48,001 (2%)	13,337 (2%)		1		1		
	Monmouthshire	89,292 (2.9%)		73,472 (3%)	15,820 (2.4%)		0.97		0.8		
	Neath Port Talbot	141,542 (4.6%)		114,569 (4.7%)	26,973 (4.1%)		0.98		0.87		
	Newport	156,175 (5.1%)		120,898 (5%)	35,277 (5.4%)		1.02		1.08		
	Pembrokeshire	118,289 (3.8%)		96,134 (3.9%)	22,155 (3.4%)		0.97		0.87		
	Powys	123,523 (4%)		100,580 (4.1%)	22,943 (3.5%)		0.98		0.85		
	Rhondda Cynon Taf	241,299 (7.8%)		191,409 (7.8%)	49,890 (7.7%)		1		0.99		
	Swansea	244,605 (7.9%)		189,098 (7.7%)	55,507 (8.5%)		1.03		1.1		
	The Vale Of Glamorgan	132,226 (4.3%)		108,454 (4.4%)	23,772 (3.7%)		0.98		0.84		
	Torfaen	93,997 (3%)		75,954 (3.1%)	18,043 (2.8%)		0.97		0.9		
	Wrexham	137,480 (4.4%)		106,517 (4.4%)	30,963 (4.8%)		1		1.09		
	Missing	0 (0%)		0 (0%)	0 (0%)						

other age groups. WIMD Deprivation quintiles showed more individuals lived in the most deprived areas of Wales (25.8%) compared to 15.8% of individuals living in the least deprived areas of Wales (Table 2). The majority of individuals resided in urban areas (70.7%) and the highest proportion lived in the local authority of Cardiff (14.8%).

The Census only (Cohort E) sub-population comprises 525,005 individuals who could not be linked to the WDS. 52.8% were males, and the largest group of individuals were aged between 16-30 years (120,693, 23%) (Table 2). The majority of individuals were white (92%), and were born in the UK (90.6%), 3.6% identified as LGBTQ+, with never been married or in a civil partnership containing the largest group of 199,955 (38.1%) compared to those that were married or in a civil partnership or had separated, divorced or been widowed (Table 2). Some 46.9% of the population did not identify as being part of a religion, 37.3% of individuals had an intermediate level qualification compared to other qualification levels and 46.2% were working at the time of the Census compared to those that were economically active or unemployed (Table 2). Additionally, just under one in ten people stated as being a carer (8.8%), 26.8% reported a disability, nearly two in five people rented their accommodation (37.9%), and based on Census household deprivation, just under half the population (49.2%) declared being deprived in at least one deprivation dimension (education, employment, health, or housing) (Table 2).

When comparing cohorts A (WDS) and C (WDS and Census), the largest proportional differences in characteristics between populations can be seen in those aged 16-30 and 31-45 years with more individuals of that age being a member of Cohort A (A:C ratio 1.1 and 1.07 respectively) (Table 2). Conversely, more individuals aged 0-15 and 61-74 years of age are members of Cohort C (A:C ratio 0.94) (Table 2). In addition, Cohort A (WDS) individuals were more likely to live in the most deprived areas, live in Cardiff, and in urban areas (A:C ratio 1.06-1.07).

For cohorts B (Census) and C (WDS and Census), the largest proportional differences in characteristics between populations can be seen in individuals of Black, Asian, or of other ethnicities (not including mixed or white ethnicities) (B:C ratio 1.14, 1.08, and 1.14 respectively), those that are more recent migrants to the UK (B:C ratio 1.15), and, being a member of the LGBTQ+ community (B:C ratio 1.08) have larger proportional representation in cohort B (Census).

Regarding cohorts D (WDS only) and C (WDS and Census), the largest proportional differences in characteristics between populations was similar to comparing cohorts A (WDS) and C (Census). Individuals aged 16–45 years, living in the most deprived areas, living in Cardiff, and urban areas were more likely to be present in cohort D than cohort C (Table 2). Similarly, comparingCohorts E (Census only) and C identified the similar proportional differences in characteristics between populations as comparing cohort B and C; with individuals of Black, Asian, or of other ethnicities, being a more recent migrant to the UK, or member of the LGBTQ+ community identified as having a higher proportion in cohort E compare to C (Table 2).

Comparison of cohorts D (WDS only) and E (Census only) could only be compared by age and sex due to being created from two mutually exclusive data sources. Cohort D contains proportionally more males compared to cohort E (57.9% versus 52.8%) and in those aged 31-45 years (23.1% vs 19.2%), whereas those aged 61+ years are more proportionally represented in Cohort E (Table 2).

For the logistic regression component of this analysis, 2,415,260 individuals aged between 16 and 105+ years were included who were members of cohort B (Census) and separated into cohort C (WDS and Census) and E (Census only) respectively (Table 3). 1,983,977 (82.1%) of individuals were members of cohort C compared to 431,283 (17.9%) individuals in cohort E. The binary outcome was created with a positively classed outcome (1) representing non-linkage from WDS to Census, and the negatively classed outcome (0) representing those linked in both WDS and Census. Focussing on the largest differences in characteristics between these two cohorts, proportionally there were more individuals of Black, Asian, or of other ethnicities (not including mixed or white ethnicities) in the unlinked cohort E (Census only) compared to cohort C (WDS and Census) (ratio of E:C ranging between 1.43-1.83) (Table 3). Similarly, more recent migrants to the UK (arriving after 2000), being a member of the LGBT+ community or not disclosing LGBT+ status, living in rental accommodation, or undisclosed religious status were proportionately more represented in cohort E compared to cohort C (ratio of E:C 1.35-1.65) (Table 3).

Conversely, there were proportionately more individuals aged 46-60 years in cohort C (WDS and Census) compared to E (Census only). Other characteristics that were more prevalent in cohort C and had the largest difference between cohorts included being married or in a civil partnership (E:C 0.82), home-owners (E:C 0.84), and carers (E:C 0.83) (Table 3).

Table 4 reports the multivariate logistic regression analysis results to examine which sociodemographic and household characteristics influence whether individuals in the Census 2021 could be linked to WDSD within SAIL.

Demographic characteristics associated with the highest adjusted odds of not having Census linkable data in SAIL were being male (aOR = 1.28, 95% CI = 1.28-1.32), 75+ years of age (aOR = 1.27, 95% CI = 1.25-1.29), and of Asian ethnicity (aOR = 1.27, 95%CI = 1.24-1.30). Other sociodemographic characteristics with the highest adjusted odds of not having Census linkable data in SAIL were being a more recent migrant (arriving to UK after 2000) (aOR = 1.30, 95% CI = 1.28,1.32), a member of the LGBTQ+ community (aOR = 1.29, 95% CI = 1.25-1.29) or not disclosing LGBTQ+ status (aOR = 1.41, 95% CI = 1.39-1.43), being separated, divorced or widowed (aOR = 1.28, 95% CI = 1.27-1.29), or living in rental accommodation (aOR = 1.47, 95% CI = 1.45-1.48).

In terms of characteristics associated with having the highest adjusted odds of having Census linkable data in SAIL were having a long term condition or disability (aOR = 0.87, 95% CI = 0.86-0.88), having a lower level qualification (aOR = 0.86, 95% CI = 0.86-0.87), being a carer (aOR = 0.92, 95% CI = 0.91-0.93), following a non-Christian religion (aOR = 0.94, 95% CI = 0.92-0.96), being economically inactive or unemployed (aOR = 0.95, 95% CI = 0.94-0.96 and aOR = 0.94, 95% CI = 0.93-0.96 respectively), and living with any dimension of deprivation (aOR = 0.95, 95% CI = 0.94-0.96).

Table 3: Cohort characteristics and ratios of the proportions of individuals within specific cohorts for the regression analysis population

Characteristics		Census (B)	WDS and Census (C)	Census only (E)	B:C	E:C	E:B
Sex	Male	1,164,094 (48.2%)	935,127 (47.1%)	228,967 (53.1%)	1.02	1.13	1.1
	Female	1,251,166 (51.8%)	1,048,850 (52.9%)	202,316 (46.9%)	0.98	0.89	0.91
Age group	16-30	494,676 (20.5%)	388,009 (19.6%)	106,667 (24.7%)	1.05	1.26	1.2
	31-45	523,626 (21.7%)	424,589 (21.4%)	99,037 (23%)	1.01	1.07	1.06
	46-60	611,877 (25.3%)	516,588 (26%)	95,289 (22.1%)	0.97	0.85	0.87
	61-74	502,933 (20.8%)	422,257 (21.3%)	80,676 (18.7%)	0.98	0.88	0.9
	75+	282,148 (11.7%)	232,534 (11.7%)	49,614 (11.5%)	1	0.98	0.98
Ethnicity	Asian, Asian British Or Asian Welsh	58,673 (2.4%)	44,554 (2.2%)	14,119 (3.3%)	1.09	1.5	1.38
	Black, Black British, Black Welsh, Caribbean Or African	17,227 (0.7%)	12,635 (0.6%)	4,592 (1.1%)	1.17	1.83	1.57
	Mixed Or Multiple Ethnic Groups	25,659 (1.1%)	19,866 (1%)	5,793 (1.3%)	1.1	1.3	1.18
	Other Ethnic Group	17,006 (0.7%)	12,901 (0.7%)	4,105 (1%)	1	1.43	1.43
	White	2,296,695 (95.1%)	1,894,021 (95.5%)	402,674 (93.4%)	1	0.98	0.98
LGBQT+ status	Missing	206,625 (8.6%)	157,010 (7.9%)	49,615 (11.5%)	1.09	1.46	1.34
	Person Is LGBTQ+	74,218 (3.1%)	57,129 (2.9%)	17,089 (4%)	1.07	1.38	1.29
	Person Is Not LGBTQ+	2,134,417 (88.4%)	1,769,838 (89.2%)	364,579 (84.5%)	0.99	0.95	0.96
Year arrive UK	Arrived After 2000	116,670 (4.8%)	86,141 (4.3%)	30,529 (7.1%)	1.12	1.65	1.48
	Arrived Up To 2000	59,320 (2.5%)	48,234 (2.4%)	11,086 (2.6%)	1.04	1.08	1.04
	Born In UK	2,239,270 (92.7%)	1,849,602 (93.2%)	389,668 (90.4%)	0.99	0.97	0.98
Marital status	Married/Civil Partnership	1,097,445 (45.4%)	931,301 (46.9%)	166,144 (38.5%)	0.97	0.82	0.85
	Never Married/Civil Partnership	863,730 (35.8%)	684,295 (34.5%)	179,435 (41.6%)	1.04	1.21	1.16
	Separated/Divorced/ Widowed	454,085 (18.8%)	368,381 (18.6%)	85,704 (19.9%)	1.01	1.07	1.06
Religion	Christian	1,142,091 (47.3%)	954,695 (48.1%)	187,396 (43.5%)	0.98	0.9	0.92
	Missing	141,575 (5.9%)	109,713 (5.5%)	31,862 (7.4%)	1.07	1.35	1.25
	No Religion	1,056,267 (43.7%)	860,762 (43.4%)	195,505 (45.3%)	1.01	1.04	1.04
	Non-Christian Religion	75,327 (3.1%)	58,807 (3%)	16,520 (3.8%)	1.03	1.27	1.23
Highest qualification	Higher Level Qualifications	772,751 (32%)	631,675 (31.8%)	141,076 (32.7%)	1.01	1.03	1.02
	Intermediate	605,319 (25.1%)	492,638 (24.8%)	112,681 (26.1%)	1.01	1.05	1.04
	Lower Level Qualifications	562,396 (23.3%)	471,807 (23.8%)	90,589 (21%)	0.98	0.88	0.9
	No Qualifications	474,794 (19.7%)	387,857 (19.5%)	86,937 (20.2%)	1.01	1.04	1.03
Activity last week	Economically Inactive	1,023,950 (42.4%)	847,061 (42.7%)	176,889 (41%)	0.99	0.96	0.97
	Missing	487 (0%)	487 (0%)	0 (0%)			
	Unemployed	71,537 (3%)	57,221 (2.9%)	14,316 (3.3%)	1.03	1.14	1.1
	Working	1,319,286 (54.6%)	1,079,208 (54.4%)	240,078 (55.7%)	1	1.02	1.02
Is carer	Carer	295,987 (12.3%)	250,775 (12.6%)	45,212 (10.5%)	0.98	0.83	0.85
	Not A Carer	2,119,273 (87.7%)	1,733,202 (87.4%)	386,071 (89.5%)	1	1.02	1.02
Disability	Disability	769,761 (31.9%)	644,355 (32.5%)	125,406 (29.1%)	0.98	0.9	0.91
	No Disability	1,645,499 (68.1%)	1,339,622 (67.5%)	305,877 (70.9%)	1.01	1.05	1.04
Accommodation ownership	Home Owner (Owns Outright)	914,828 (37.9%)	773,431 (39%)	141,397 (32.8%)	0.97	0.84	0.87
	Home Owner (With Mortgage)	782,256 (32.4%)	654,871 (33%)	127,385 (29.5%)	0.98	0.89	0.91
	Rental	718,176 (29.7%)	555,675 (28%)	162,501 (37.7%)	1.06	1.35	1.27
Household deprivation	Household Deprived In One Or More Dimension	1,278,322 (52.9%)	1,053,950 (53.1%)	224,372 (52%)	1	0.98	0.98
	Household Not Deprived In Any Dimension	1,136,938 (47.1%)	930,027 (46.9%)	206,911 (48%)	1	1.02	1.02

Table 4: Unadjusted and adjusted Odds Ratios (OR) and 95% confidence intervals (CI) regression analysis results examining the sociodemographic and household characteristics that influence whether individuals in the Census 2021 could be linked within SAIL

Characteristics	Univariate		Adjusted			
	OR	p-value	AOR	p-value	95% CI	
Male (ref: Female)	1.27	<0.001	1.28	<0.001	1.27	1.29
Age 16-30 (ref: 46-60)	1.49	<0.001	1.19	<0.001	1.18	1.21
Age 31-45 (ref: 46-60)	1.26	<0.001	1.11	<0.001	1.1	1.12
Age 61-74 (ref: 46-60)	1.04	<0.001	1.13	<0.001	1.12	1.14
Age 75+ (ref: 46-60)	1.16	<0.001	1.27	<0.001	1.25	1.29
Ethnicity: Asian, Asian British Or Asian Welsh (ref: White)	1.49	<0.001	1.27	<0.001	1.24	1.3
Ethnicity: Black, Black British, Black Welsh, Caribbean Or African (ref: White)	1.71	<0.001	1.21	<0.001	1.17	1.26
Ethnicity: Mixed Or Multiple Ethnic Groups (ref: White)	1.37	<0.001	1.16	<0.001	1.12	1.19
Ethnicity: Other Ethnic Group (ref: White)	1.50	<0.001	1.08	<0.001	1.04	1.12
LGBTQ+: Missing (ref: non-LGBTQ)	1.53	<0.001	1.41	<0.001	1.39	1.43
LGBTQ+: Person Is LGBTQ+ (ref: non-LGBTQ)	1.45	<0.001	1.27	<0.001	1.25	1.29
Year arrived UK: Arrived After 2000 (ref: Born In UK)	1.68	<0.001	1.3	<0.001	1.28	1.32
Year arrived UK: Arrived Up To 2000 (ref: Born In UK)	1.09	<0.001	1.07	<0.001	1.05	1.1
Marital status: Never Married/Civil Partnership (ref: Married/Civil Partnership)	1.47	<0.001	1.24	<0.001	1.22	1.25
Marital status: Separated/Divorced/Widowed (ref: Married/Civil Partnership)	1.30	<0.001	1.28	<0.001	1.27	1.29
Religion: Christian (ref: No Religion)	0.86	<0.001	0.97	<0.001	0.96	0.98
Religion: Missing (ref: No Religion)	1.28	<0.001	1.14	<0.001	1.13	1.16
Religion: Non-Christian Religion (ref: No Religion)	1.24	<0.001	0.94	<0.001	0.92	0.96
Highest Qualification: Higher level qualifications (ref: Intermediate)	0.98	<0.001	1.04	<0.001	1.03	1.05
Highest Qualification: Lower level qualifications (ref: Intermediate)	0.84	<0.001	0.86	<0.001	0.86	0.87
Highest Qualification: No qualifications (ref: Intermediate)	0.98	<0.001	0.98	<0.001	0.97	0.99
Activity last week: Economically Inactive (ref: Working)	0.94	<0.001	0.95	<0.001	0.94	0.96
Activity last week: Unemployed (ref: Working)	1.12	<0.001	0.94	<0.001	0.93	0.96
Is a carer: Carer (ref: non-carer)	0.81	<0.001	0.92	<0.001	0.91	0.93
Has long-term condition or disability? Yes (ref: No)	0.85	<0.001	0.87	<0.001	0.86	0.88
Family status: Home Owner (With Mortgage) (ref: owns outright)	1.06	<0.001	1.02	<0.001	1.01	1.03
Family status: Rental (ref: owns outright)	1.60	<0.001	1.47	<0.001	1.45	1.48
Deprivation: Household Deprived In One Or More Dimension (ref: No)	0.96	<0.001	0.95	<0.001	0.94	0.96

Discussion

This study was designed to explore the completeness and representativeness of Census 2021 data linkage within the SAIL Databank for research on the population of Wales, UK and, to understand which subgroups of the population might be disproportionately affected in studies using Census population linkage.

In total, 3,615,981 individuals aged 0–105+ years were included in this cross-sectional cohort comparison study. Overall, when examining the two population data sources, the WDSD and the Census 2021, the resident populations of Wales were 3,090,976 and 2,965,196 respectively. After linkage, 2,440,191 individuals were present in both populations which equates to a coverage of 78.9% and 82.3% for the WDSD and Census populations respectively. A further 650,785 (21.1%) individuals could only be found in the WDSD population and 525,005 (17.7%) individuals could only be found in the Census 2021 population.

Our results show that being male, an older individual (aged 75+), of Asian ethnicity, a more recent migrant (arriving to UK after 2000), member of the LGBTQ+ community or not disclosing LGBTQ+ status, being separated, divorced or widowed, or living in rental accommodation were the sociodemographic and household characteristics impacted most for not having Census linkable data in SAIL for further research. Conversely, having a long-term condition or disability, a lower level qualification, being a carer, following a non-Christian religion, being economically inactive or unemployed, and living with any dimension of deprivation were the sociodemographic and household characteristics impacted the least and were more likely to have Census linkable data in SAIL.

There exist multiple reasons postulated for varied linkage rates between WDS and Census as applied in this work, whilst outside the remit of this work, we discuss key themes here. Initially, the difference in the primary reason for existence for WDS and Census as well as varying processes and policies have implications. For example, Census is not a true complete enumeration of the full population and includes adjustment by imputation, additionally, Census returns are completed by either the main or proxy respondents which may introduce inconsistencies. WDS exists to support healthcare and administration in Wales and reasons exist why some individuals may not be registered, especially at specific dates which coincide with events such as Census 2021. Variation in interaction with healthcare services across certain demographic groups with different behaviours or systematic issues may affect linkage rates. For example, evidence suggests younger males engage less with primary care services

potentially resulting in less accurate and timely data flowing into administrative records (17), and some ethnic groups are more likely to have lower linkage rates due to various factors including systematic issues related to data capture systems and linkage algorithms designed predominantly around Western style name conventions (first name, middle name, last name) (18,19). Response rates specifically for Census, and data collection processes and quality of personal identifiable information for both sources will impact upon linkage success rates of downstream data linkage processes. The downstream processing of personal identifiable information data to create the Anonymised Linkage Fields for two separate data sources also contains intricacies which may affect linkage, for instance, variation in use of names and middle names, and general data collection and recording errors will influence linkage outcomes.

This work aims to inform the research community for those involved in using linked population-level data using population data sources such as Census and WDS within SAIL or other Trusted Research Environments. The work does not aim to fully analyse the drivers behind variation in data linkage across populations, indeed, due to necessary SAIL governance processes, investigations to understand specific sources of error are impossible. Neither does the work aim to suggest that one population data source is better than the other; we support a combination of the datasets, which provides greater potential for impactful research.

Among the strengths of this study was the ability to link multi-sourced population-wide demographic data for the population of Wales through the SAIL Databank. Having access to the Census 2021 allowed for the inclusion and analysis of individuals by protected characteristics. Additionally, to our knowledge, this is the first study examining the completeness and representativeness of Census 2021 with a total population register of Wales outside of the Office of National Statistics sources [21–23].

Although our study provides important findings, several limitations need to be considered. Due to a number of Census questions having age restrictions applied, the regression analysis was restricted to those aged 16+ years of age. We also restricted our final regression analysis to individuals living in private households and therefore excluded individuals living in communal establishments; future work could explore such populations, although there are known data quality and collection limitations related to communal establishments [24].

A further limitation of this study is that the comparisons between cohorts D and E was limited to age and sex only due to being created from two mutually exclusive data sources, which lacked other common variables. Whilst it might be assumed that these two similarly sized cohorts are in fact the same population, it is not possible to infer that from the data.

Conclusion

This study provides important research implications for consideration. Results show that certain personal characteristics and sub-groups of the population of Wales are disproportionately represented when combining population estimates and utilising Census data in data linkage population-wide studies in SAIL.This is an important finding for researchers to understand when carrying out future linked research on using census data and electronic health records on the Welsh population.

Acknowledgements

This study makes use of anonymised data held in the Secure Anonymised Information Linkage (SAIL) Databank. We would like to acknowledge all the data providers and people who make anonymised data available for research. The authors would like to extend their gratitude and acknowledgement to the NHS, the SAIL Consumer panel as well as the IGRP who approved this project (SAIL project 1650).

Contributors

RDJ, JL, and RAL conceived and designed the study. RDJ and JL had full access to all data used in this study. Due to data permission restrictions, not all authors were able to access the underlying data used in the study. RDJ checked and verified the data used in the analysis, and conducted the analysis in consultation with JL and RAL. JL wrote the original draft. RDJ, JL, LG, RF, ST, ME, and RL reviewed, edited, and approved the final manuscript. All authors were responsible for submitting the article for publication.

Funding

This work is supported by Administrative Data Research (ADR) Wales (Grant ref: ES/W012227/1), part of the ADR UK investment, uniting research expertise from Swansea University Medical School and WISERD (Wales Institute of Social and Economic Research and Data) at Cardiff University with analysts from Welsh Government. ADR UK is funded by the Economic and Social Research Council (ESRC), part of UK Research and Innovation.

Competing interest

The authors declare no conflict of interest.

Patient consent for publication

Not required.

Ethical approval

The use of deidentified data in SAIL complies with National Research Ethics Service (NRES) guidance. Applications to use data held within the SAIL Databank, an ISO: 27001 and UK Statistics Authority (UKSA) Digital Economy Act (DEA) accredited Trusted Research Environment, must first be approved by the independent Information Governance Review Panel (IGRP). This panel contains individuals with expertise in data governance and protection, including the Chair of the Wales NRES Committee, Caldicott Guardians and members of the public. The IGRP approved SAIL project 1650 on 19th September 2023.

Data availability

This study makes use of anonymised, individual-level data held in the SAIL Databank, a Trusted Research Environment, at Swansea University, Swansea, UK. Due to the nature and level of the data, data are not publicly available. All proposals to use SAIL data are subject to review by the independent IGRP. The IGRP gives careful consideration to each project proposal to ensure proper and appropriate use of SAIL data. If a project is approved, access to the requested data is gained through a privacy-protecting safe haven and remote access system referred to as the SAIL Gateway. SAIL has established an application process to be followed by anyone who would like to access data via SAIL at: https://www.saildatabank.com/application-process/.

References

1. Rudolph JE, Zhong Y, Duggal P, Mehta SH, Lau B. Defining representativeness of study samples in medical and population health research. BMJ Med [Internet]. 2023 May;2(1):e000399. Available from: https://doi.org/10.1136/bmjmed-2022-000399.

2. Moorthie S, Hayat S, Zhang Y, Parkin K, Philips V, Bale A, et al. Rapid systematic review to identify key barriers to access, linkage, and use of local authority administrative data for population health research, practice, and policy in the United Kingdom. BMC Public Health [Internet]. 2022 Dec 28;22(1):1263. Available from: https://doi.org/10.1186/s12889-022-13187-9.

3. van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, et al. A systematic review of barriers to data sharing in public health. BMC Public Health [Internet]. 2014 Dec 5;14(1):1144. Available from: https://doi.org/10.1186/1471-2458-14-1144.

4. ADRUK. Trusted research environments - ADR UK [Internet]. [cited 2025 Apr 11]. Available from: https://www.adruk.org/data-access/trusted-research-environments/#tab-c4813.

5. SERP. Home - SeRP - for Trusted Research Environment solutions [Internet]. [cited 2025 Apr 11]. Available from: https://serp.ac.uk/.

6. Harron K. Data linkage in medical research. BMJ Med [Internet]. 2022 Mar;1(1):e000087. Available from: https://doi.org/10.1136/bmjmed-2021-000087.

7. Brook EL, Rosman DL, Holman CDJ. Public good through data linkage: measuring research outputs from the Western Australian Data Linkage System. Aust N Z J Public Health [Internet]. 2008 Feb;32(1):19–23. Available from: https://doi.org/10.1111/j.1753-6405.2008.00160.x.

8. Office for National Statistics. Why we have a census [Internet]. Office for National Statistics. 2001 [cited 2025 Apr 11]. Available from: https://www.ons.gov.uk/census/2011census/whywehaveacensus.

9. SAIL Databank. SAIL Databank [Internet]. 2021 [cited 2025 Apr 11]. Available from: https://saildatabank.com/.

10. Lyons J, Akbari A, Agrawal U, Harper G, Azcoaga-Lorenzo A, Bailey R, et al. Protocol for the development of the Wales Multimorbidity e-Cohort (WMC): data sources and methods to construct a population-based research platform to investigate multimorbidity. BMJ Open [Internet]. 2021 Jan 1 [cited 2025 Jul 2];11(1):e047101. Available from: https://bmjopen.bmj.com/content/11/1/e047101.

11. Lyons RA, Jones KH, John G, Brooks CJ, Verplancke J-P, Ford D V, et al. The SAIL databank: linking multiple health and social care datasets. BMC Med Inform Decis Mak [Internet]. 2009 Dec 16;9(1):3. Available from: https://doi.org/10.1186/1472-6947-9-3.

12. Office for National Statistics. Coverage estimation for Census 2021 in England and Wales - Office for National Statistics [Internet]. 2022 [cited 2025 Jun 30]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/coverageestimationforcensus2021inenglandandwales?.

13. Norman P, Simpson L, Sabater A. "Estimating with confidence" and hindsight: New UK small-area population estimates for 1991. Popul Space Place [Internet]. 2008 Sep 1 [cited 2025 Jun 24];14(5):449–72. Available from: https://doi.org/10.1002/psp.492

14. Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML, et al. Challenges in administrative data linkage for research. Big Data Soc [Internet]. 2017 Dec 1 [cited 2025 Aug 4];4(2). Available from: https://doi.org/10.1177/2053951717745678.

15. Ford D V, Jones KH, Verplancke J-P, Lyons RA, John G, Brown G, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. BMC Health Serv Res [Internet]. 2009 Dec 4 [cited 2016 Sep 5];9(1):157. Available from: https://doi.org/10.1186/1472-6963-9-157.

16. Rodgers SE, Lyons RA, Dsilva R, Jones KH, Brooks CJ, Ford D V, et al. Residential Anonymous Linking Fields (RALFs): a novel information infrastructure to study the interaction between the environment and individuals' health. J Public Health (Bangkok) [Internet]. 2009 Dec 1;31(4):582–8. Available from: https://doi.org/10.1093/pubmed/fdp041.

17. Johnson RD, Griffiths LJ, Hollinghurst JP, Akbari A, Lee A, Thompson DA, et al. Deriving household composition using population-scale electronic health record data—A reproducible methodology. Ramagopalan S V., editor. PLoS One [Internet]. 2021 Mar 29 [cited 2021 Apr 1];16(3):e0248195. Available from: https://doi.org/10.1371/journal.pone.0248195.

18. Welsh Government. Welsh Index of Multiple Deprivation, [Online] [Internet]. Gov.Wales. 2020 [cited 2025 Apr

11]. Available from: https://www.gov.wales/welsh-index-multiple-deprivation.

19. United Kingdom. Equality Act 2010: Chapter 15 [Internet]. London: The Stationery Office; 2010. Available from: https://www.legislation.gov.uk/ukpga/2010/15/contents.

20. Office for National Statistics. Census [Internet]. Office for National Statistics. [cited 2025 Apr 11]. Available from: https://www.ons.gov.uk/census.

21. Pereira E, Robertson Z, Barnes A, Jones F, Mylles S, Pearce C. Comparing the 2021 Census to administrative data to better understand the population estimation challenge. Int J Popul Data Sci [Internet].2023 Sep 14 [cited 2025 Apr 11];8(2). Available from: https://doi.org/10.23889/ijpds.v8i2.2260.

22. Plachta J, Collyer S. Linking of the 2021 Census to massive linked administrative data to understand coverage and quality. Int J Popul Data Sci [Internet]. 2023 Sep 14 [cited 2025 Apr 11];8(2). Available from: https://doi.org/10.23889/ijpds.v8i2.2201.

23. Wilk M, Harper G, Firman N, Dibben C, Fry R, Dezateux C. Validation of a dynamic method of measuring households and populations from primary care Electronic Health Records: Cross-sectional comparison with Office for National Statistics Census 2021 estimates. Int J Popul Data Sci [Internet]. 2023 Sep 14 [cited 2025 Apr 11];8(2). Available from: https://doi.org/10.23889/ijpds.v8i2.2189.

24. Office for National Statistics. Communal establishment (CE) estimation and adjustment: Census 2021 - Office for National Statistics [Internet]. [cited 2025 Jun 30]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/communalestablishmentceestimationandadjustmentcensus2021.