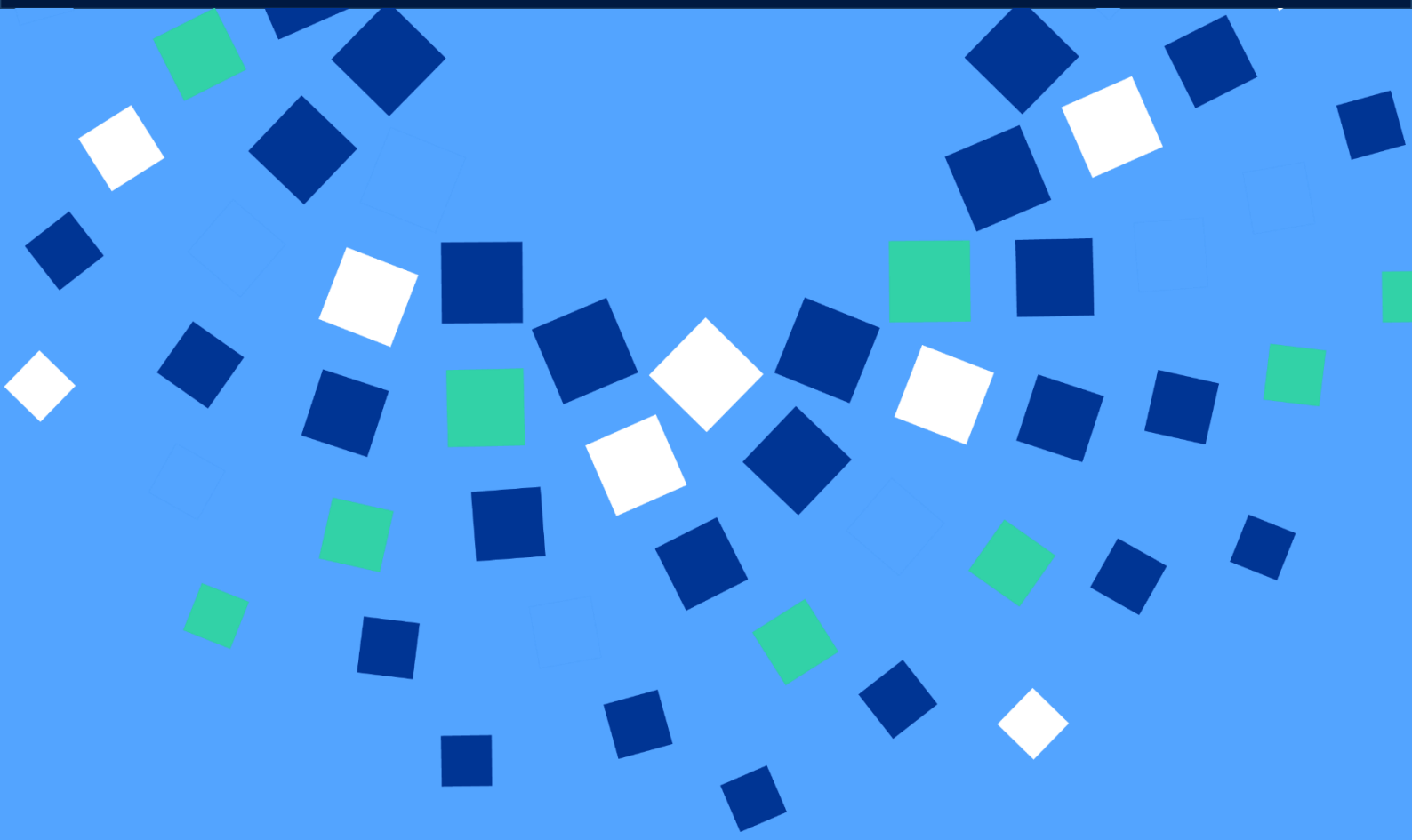
YDG CYMRU | ADR WALES

Data Explained

Welsh Demographic Service Dataset (WDSD)

Authors: Victoria Best, Sarah Aldridge, Roberta Chiovoloni, Ashley Akbari
Date: October 2026

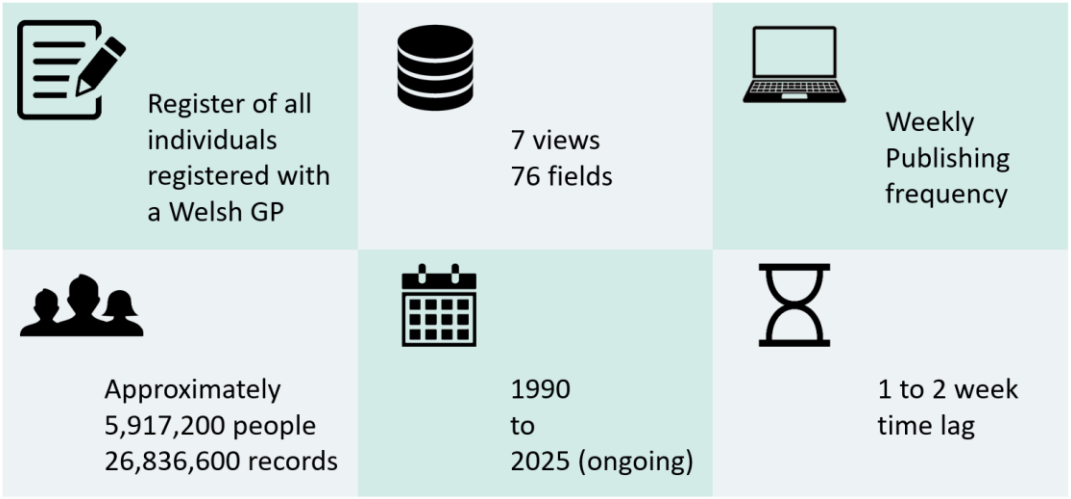


Summary

This Data Explained summarises experiences and learning from working with the Welsh Demographic Service Dataset (WDSD) data. This publication is intended to help guide future researchers using this data and to provide feedback on future data development and documentation.

The administrative data discussed in this Data Explained was made securely available through an approved SAIL project via the independent Information Governance Review Panel (IGRP) – project 1934. The data used in this research project comes from Digital Health and Care Wales (DHCW) and was accessed through the Secure Anonymised Information Linkage (SAIL) Databank. The data were not originally collected for research, and gaps and inconsistencies in their recording are expected; some are detailed below.

Overview



Introduction

The Welsh Demographic Service Dataset (WDSD) is a patient-level data source containing demographic information about Welsh residents registered at Welsh General Practices (GPs). The WDSD is sourced from WDS (Welsh Demographic Service) data, which includes information such as address, date of birth and GP registration history (Tingay et al., 2015).

The WDSD contains information about all patients registered with a Welsh General Practice (GP). As such, inclusion in the WDSD within SAIL is not dependent on the GP contributing data to SAIL. WDSD coverage began in 1990, although fewer active records existed before this date, peaking around July 1948, when the NHS was founded (Office for National Statistics, 2016).

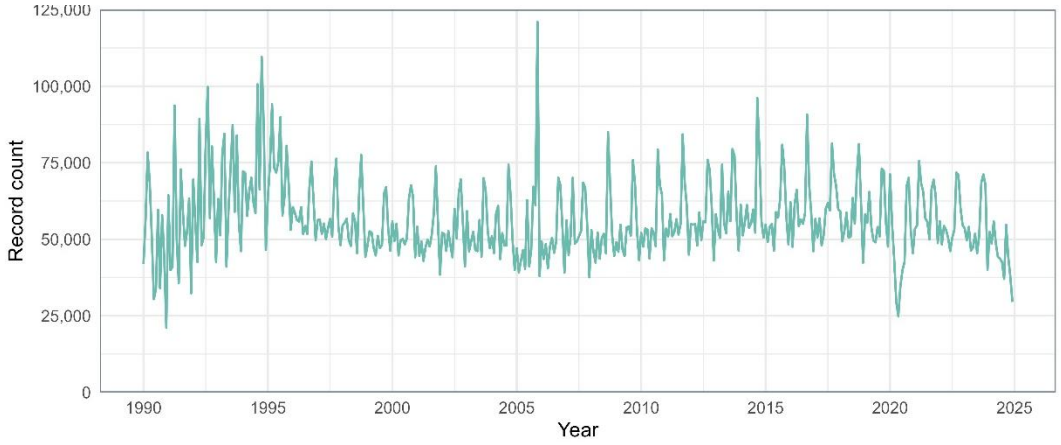


Figure 1 - Count of all WDSD active_from records per month from 1990 to 2024, sourced from the raw WDSD table per_residence_gpreg.

Seasonal variation is observable within the WDSD, with new active records increasing in September and October each year. Grouping by age shows this increase is largely associated with those aged 15-29, suggesting it may relate to relocation at the start of the academic year.

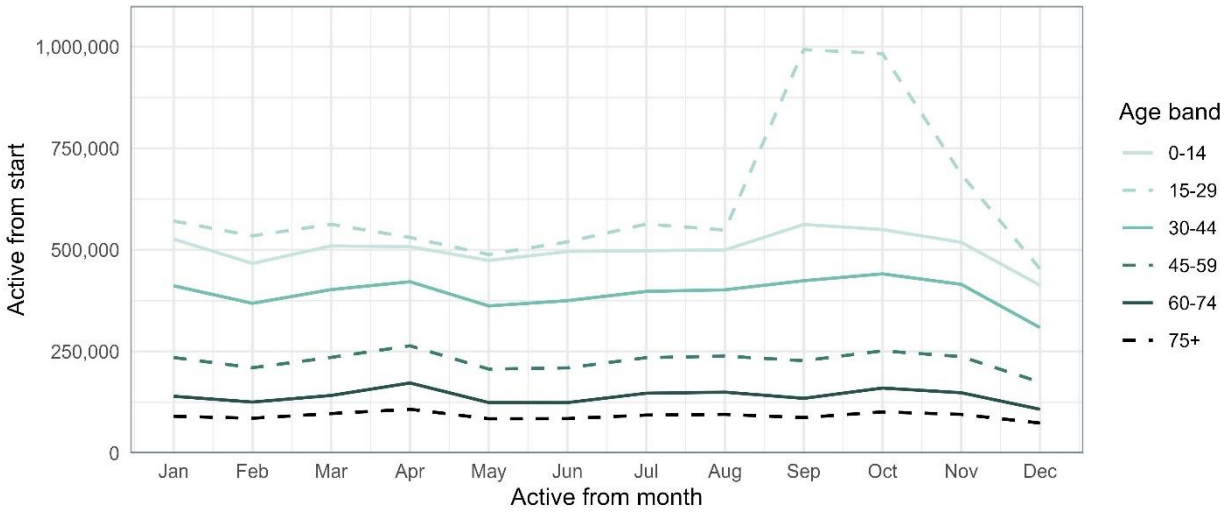


Figure 2 - Count of records by active_from month presented by age band, sourced from the original WDSD table per_residence_gpreg.

How is the data collected?

The WDSD data is originally captured and recorded by administrative staff at individual Welsh GPs via data entry to the local GP system, such as EMIS or Vision. As such, the WDSD contains information for all Welsh residents registered with a Welsh GP.

WDSD information was historically captured through the National Health Application and Infrastructure Services (NHAIS) system, often referred to as the Exeter system. This records key information for the WDSD, including a patient's name, address, date of birth, sex and dates of GP registration (Faculty of Public Health, n.d.). As of April 2024, the primary data source shifted from NHAIS to the Personal Demographics Service (PDS): the national electronic database of NHS patient details covering England, Wales, and the Isle of Man (NHS England, n.d.). Patient details are updated in

PDS via GP systems or directly via FHIR application programming interfaces (APIs), and WDS draws its updates from PDS.

WDSD data is used as the basis of a population spine by Digital Health and Care Wales (DHCW) to link wider pseudonymised data via an encrypted patient ID. Within SAIL, an Anonymised Linkage Field (ALF) is assigned to each record in the WDSD data where there is sufficient confidence in the person's identity(R. A. Lyons et al., 2009). This field can then be used to link information for the same person across data sources. The WDSD also contains address details and dates, allowing a Residential Anonymised Linkage Field (RALF) to be assigned from the Unique Property Reference Number (UPRN), which can be used to identify people living in the same residence at the same time (Heaven & Lowe, 2014; Johnson et al., 2021).

Format and linkage

There are 7 WDSD tables, totalling 76 fields.

The **per_residence_gpreg** table is the original WDSD table with multiple rows per person. The SAIL team use this to produce single clean views of the WDSD data, adding linked information such as deprivation and a pseudonymised residential ID (*RALF_PE*).

The SAIL team creates four main single clean tables**:**
single_clean_ar_pers,
single_clean_geo_wales,
single_clean_geo_char_lsoa
and **single_clean_geo_ralf_lsoa**.

The single clean tables are often easier sources to work with than the original **per_residence_gpreg** table, as the cleaning process reduces duplication and overlapping records, closes small gaps of up to 30 days (since a short gap doesn't likely represent a move out of Wales and therefore shouldn't be a gap in the follow-up period), joins consecutive records with the same information to make processing easier, and extends follow-up time to birth when records start within 6 weeks of birth (Thayer et al., 2020).

The tables **single_clean_geo_char_lsoa** and **single_clean_geo_ralf_lsoa** are provisioned only to projects with permission to access Lower-layer Super Output Area (LSOA) and RALF data, respectively, within their IGRP. LSOAs are geographic areas based on the latest census data comprising between 400 and 1,200 households with a population generally between 1,000 and 3,000 (Office for National Statistics, n.d.). The Office for National Statistics makes minor adjustments to LSOAs and boundaries following each decennial census, so to reflect this, additional versions of the **single_clean_geo_char_lsoa** and **single_clean_geo_ralf_lsoa** tables are produced each time the LSOA codes and boundaries are updated.

WDSD data tables have no primary key, so unique rows are often identified through a combination of fields. The diagram below indicates the linkage structure and some commonly accessed fields within the WDSD data. Row ID indicates the fields that, together, can identify a unique row in that table. *ALF_PE* is the common linkage field between the WDSD tables, but for tables containing address or GP information, which may change over time, using the date fields is important when linking to ensure the correct records are chosen.

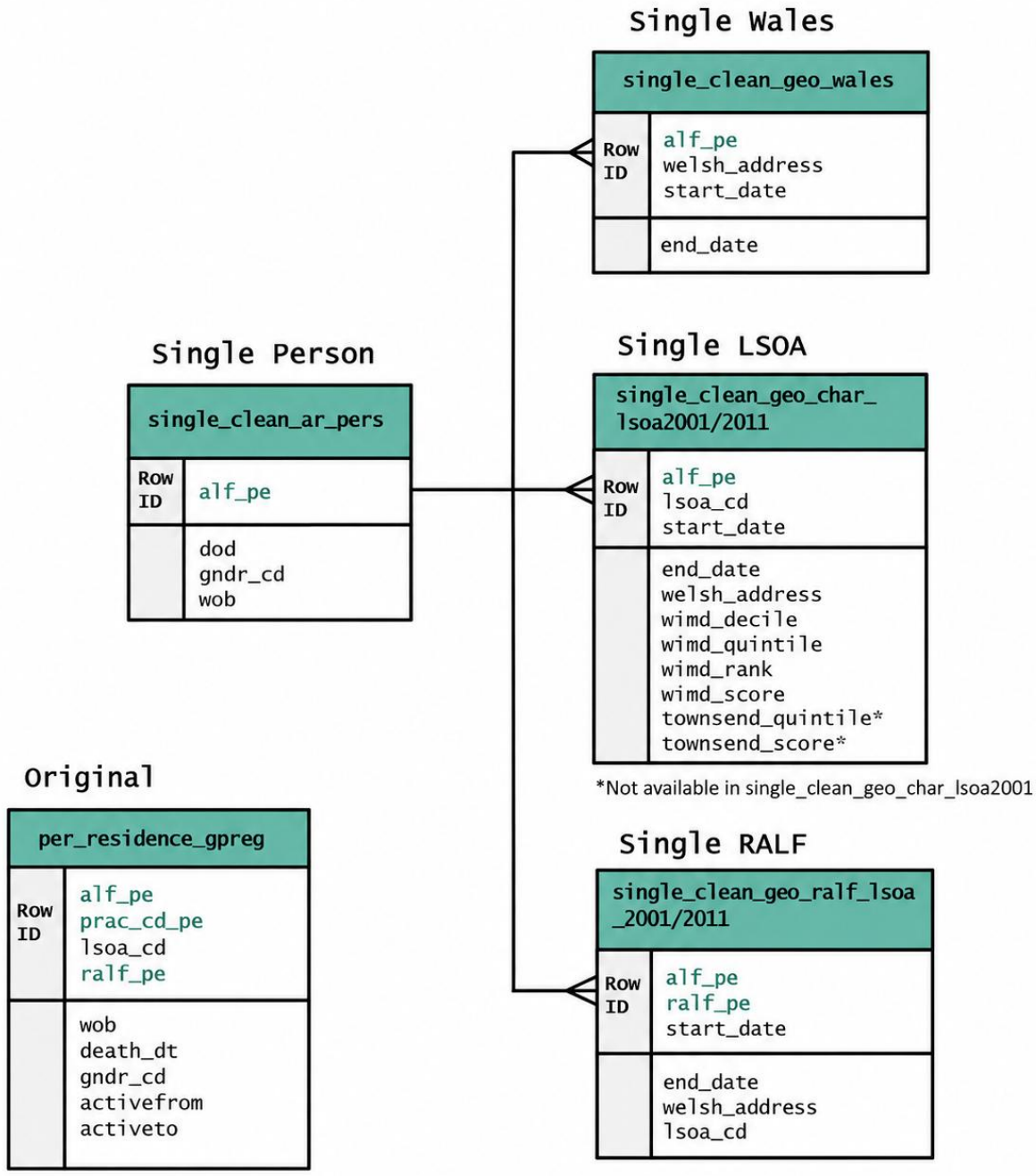


Figure 3 - Visualisation of the data linkage and key fields in the WDSD tables.

WDSD Tables

Original WDSD table

per_residence_gpreg
This is the original WDSD long data table, with multiple rows per person, where each row represents a unique combination of circumstances such as GP registration and address. The fields in the original WDSD table are week of birth (WOB), death date (DEATH_DT), gender code (GNDR_CD), and GP practice code (PRAC_CD_PE); if the project has sufficient permission, LSOA codes are also present.

Two additional identifier fields are available: the person anonymised linkage field (*ALF_PE*), which is available in all projects, and the residential anonymised linkage field (*RALF_PE*), which requires specific permission.

When using this table, cleansing prior to use is advised to remove duplication and close small gaps in continuous records that can result from administrative errors or differences between recording systems. It may therefore be more efficient to use the single clean tables where possible, which have already been data cleansed.

Single Clean Person Table

single_clean_ar_pers

This table is useful for identifying basic demographics such as date of birth, date of death, and gender. It contains a single row per person, including both living and deceased. This table does not include residence start or end dates and includes people who have migrated into and out of Wales; therefore, it should not be used in isolation to identify point-in-time Welsh population estimates or to develop a cohort spine that depends on inclusion dates.

Single Clean Residence Tables

The remaining single clean tables contain residence information. Cleaning for all single clean residence tables includes closing gaps of 30 days or less and extending the address start to the week of birth if the record starts within 6 weeks. LOSA and RALF residence tables are only provisioned to projects with specific permission in the project IGRP approval.

single_clean_geo_wales

This table helps identify Welsh residents at a point in time or over the study period. It can therefore be used to develop a cohort spine or to produce Welsh population estimates. It contains information about Welsh residence, with each row representing a single continuous period of Welsh or non-Welsh residence per person (*ALF_PE*). The *WELSH_ADDRESS* field provides a binary indicator of whether or not the person was a Welsh resident between the start and end dates.

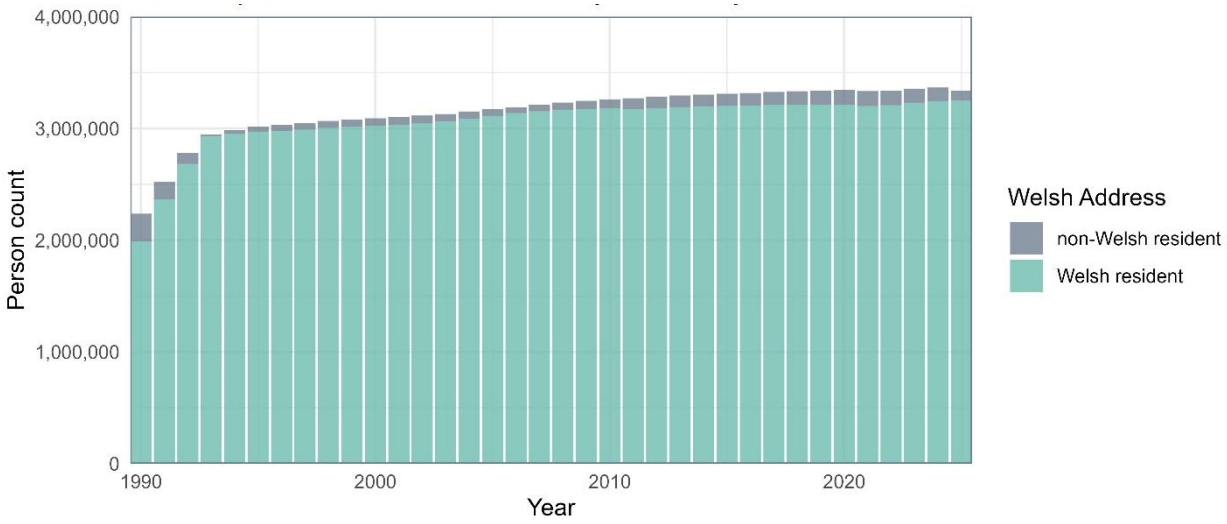


Figure 4 - Count of Welsh resident records as at the 1st January each year within the single_clean_geo_wales view.

single_clean_geo_char_lsoa2001 and _lsoa2011

Each row in these tables represents a single continuous period of residence within a single LSOA per person (*ALF_PE*), linked to the various LSOA versions (depending on the table used).

Each record within the **single_clean_geo_wales** table with an *ALF_PE* should have at least one corresponding record within the **single_clean_geo_char_lsoa** table. These tables provide an indicator of Welsh residence, as well as the Welsh Index of Multiple Deprivation (WIMD) and the Townsend deprivation indices. LSOA tables are useful for identifying the geographic location of residence, which can add covariate information such as deprivation, rurality, and Health Board. Further data are published at the LSOA level, such as crime rates and population density, which can be linked via the LSOA. LSOA data can also be used to generate visual maps if required.

single_clean_geo_ralf_lsoa2001 and _lsoa2011

These tables contain pseudonymised address information, with each row representing a single continuous period of residence at a single address (*RALF_PE*) per person (*ALF_PE*), linked to the various LSOA versions (depending on the table used). These tables also provide WIMD and Townsend deprivation indices.

The *RALF_PE* is an encrypted residential identifier (Johnson et al., 2021). This enables researchers within SAIL to identify people living at the same residence while maintaining anonymity (Rodgers et al., 2009). Not all individuals have sufficient address information to be assigned a *RALF_PE; as a result, only a subset of people within the single_clean_geo_char_lsoa tables are includ*ed. The figure below shows the percentage of people living in a Welsh LSOA each year who have a corresponding record in the RALF table.

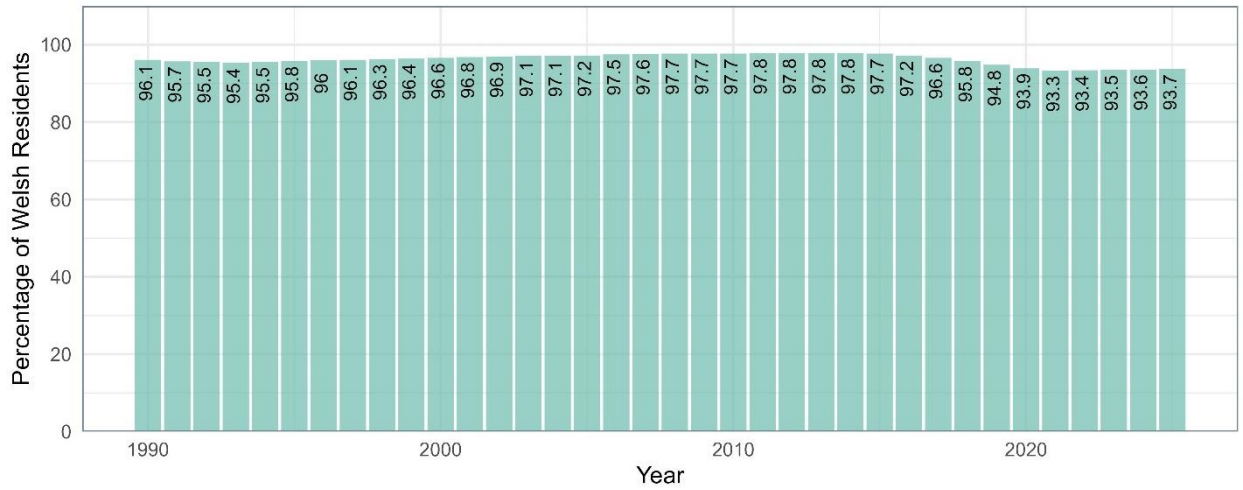


Figure 5 - Percentage of Welsh residents in the single_clean_geo_char_lsoa2011 table who also have a record within the single_clean_geo_ralf_lsoa2011 view as of the 1st January for each year.

What can the data be used for?

WDSD data is often used as the basis to create a cohort of Welsh residents when conducting research within SAIL. Electronic health data often includes information on non-Welsh residents, so linkage to the WDSD enables researchers to ensure they include only data relating to the population of interest.

The WDSD enables the creation of a longitudinal population spine due to the nature of the data held; this is an advantage over census data, which provides only a cross-sectional snapshot of a point in time (J. Lyons et al., 2025).

The inclusion of LSOA allows many covariates to be linked that would otherwise not be available, strengthening the statistical analysis. The inclusion of RALF enables the health conditions of cohabiting groups to be analysed, and the RALF can also be linked to wider SAIL data to identify individuals living in certain residential groups, such as care homes.

Data limitations encountered

The WDSD is sourced from GP registration data and is therefore reliant on GP records being accurate and up to date. There are a number of reasons why addresses may not be updated in GP records, including adult children choosing to remain registered at their parents' address, or shared custody arrangements for children in which only one address is recorded. Short-term stays in Care Homes may also not result in a change of address on a GP system (Office for National Statistics, 2016).

A further issue with the accuracy of GP registration data is the time lag between a change of address and the patient informing their GP. This makes it likely for some addresses to never be recorded, while others may be updated long after the move has taken place. Further, the time lag varies depending on patient demographics and health conditions. Typically, mothers with young children, patients with ongoing health conditions, and the very elderly update their GP address more quickly, whereas healthy young adults, the highly mobile, and males are generally slower to update their GP address (Office for National Statistics, 2016).

Lyons et al. (2025) compared the WDSD with 2021 Census data and identified that non-linkage between the two data sources was more likely for males, those aged 75+, those of Asian ethnicity, migrants to the UK since 2000, LGBTQ+ or not disclosing LGBTQ+ status, being separated, divorced or widowed, or living in rental accommodation. This is likely to result in bias in the WDSD having accurate and up to date information.

References

Faculty of Public Health. (n.d.). *Health Information: Use of information technology in processing and analysis*. Health Knowledge. Retrieved 2 December 2025, from https://www.healthknowledge.org.uk/e-learning/health-information/applications-health-information-specialists/use-information-technology-processing-analysis

Heaven, M., & Lowe, S. (2014). *Data Linking Demonstration Project—Flying Start*. Welsh Government. https://www.gov.wales/sites/default/files/statistics-and-research/2019-01/data-linking-demonstration-project-flying-start.pdf

Johnson, R. D., Griffiths, L. J., Hollinghurst, J. P., Akbari, A., Lee, A., Thompson, D. A., Lyons, R. A., & Fry, R. (2021). Deriving household composition using population-scale electronic health record data—A reproducible methodology. *PLOS ONE*, *16*(3), e0248195. https://doi.org/10.1371/journal.pone.0248195

Lyons, J., Johnson, R., Edwards, M., Turner, S., Fry, R., Griffiths, L. J., & Lyons, R. (2025). Administrative data linkage to Census 2021 in Wales, UK: A cross-sectional study examining completeness and representativeness for population analytics. *International Journal of Population Data Science*, *10*(1). https://doi.org/10.23889/ijpds.v10i1.2994

Lyons, R. A., Jones, K. H., John, G., Brooks, C. J., Verplancke, J.-P., Ford, D. V., Brown, G., & Leake, K. (2009). The SAIL databank: Linking multiple health and social care datasets. *BMC Medical Informatics and Decision Making*, *9*(1), 3. https://doi.org/10.1186/1472-6947-9-3

Office for National Statistics. (n.d.). *Statistical geographies—Office for National Statistics*. Office for National Statistics. Retrieved 16 December 2025, from https://www.ons.gov.uk/methodology/geography/ukgeographies/statisticalgeographies

Office for National Statistics. (2016, December 20). *Patient Register: Quality assurance of administrative data used in population statistics, Dec 2016—Office for National Statistics*. Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/patientregisterqualityassuranceofadministrativedatausedinpopulationstatisticsdec2016

Rodgers, S. E., Lyons, R. A., Dsilva, R., Jones, K. H., Brooks, C. J., Ford, D. V., John, G., & Verplancke, J.-P. (2009). Residential Anonymous Linking Fields (RALFs): A novel information infrastructure to study the interaction between the environment and individuals' health. *Journal of Public Health*, *31*(4), 582–588. https://doi.org/10.1093/pubmed/fdp041

Thayer, D., Rees, A., Kennedy, J., Collins, H., Harris, D., Halcox, J., Ruschetti, L., Noyce, R., & Brooks, C. (2020). Measuring follow-up time in routinely-collected health datasets: Challenges and solutions. *PLOS ONE*, *15*(2), e0228545. https://doi.org/10.1371/journal.pone.0228545

Tingay, K., Heaven, M., & Lowe, S. (2015, December 17). *Review into the capability of the 'Secure Anonymised Information Linkage' (SAIL) Databank to provide data for the Social Services National Outcomes Framework for people who need care and support and carers who need support—Final report*. https://www.gov.wales/sites/default/files/statistics-and-research/2019-07/151217-review-capability-secure-anonymised-information-linkage-database-en_0.pdf

Glossary

Abbreviation	Description
ADR UK	Administrative Data Research UK
ALF	Anonymous Linkage Field
API	Application Programming Interfaces
DHCW	Digital Health and Care Wales
FHIR	Fast Healthcare Interoperability Resources
GP	General Practice
IGRP	Information Governance Review
LGBTQ+	Lesbian, Gay, Bisexual, Transgender, and Queer
LSOA	Lower-layer Super Output Area
NHAIS	National Health Application and Infrastructure Services
NHS	National Health Service
ONS	Office for National Statistics
PAF	Postcode Address File
PDS	Personal Demographics Service
RALF	Residential Anonymised Linkage Field
SAIL	Secure Anonymised Information Linkage (SAIL) Databank
UPRN	Unique Property Reference Number
WIMD	Welsh Index of Multiple Deprivation

Disclaimer

This work was produced using administrative data accessed through the SAIL Databank. The use of the data in this work does not imply the endorsement of the SAIL Databank or data owners in relation to the interpretation or analysis.

This work uses research data which may not exactly reproduce National Statistics aggregates.

Acknowledgements

This study uses anonymised data held in the Secure Anonymised Information Linkage (SAIL) Databank, which is part of the national population data research infrastructure for Wales. We would like to acknowledge all the data providers who make anonymised data available for research. This study was carried out as part of the ADR Wales programme of work. ADR Wales unites 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.

This Data Insight has been produced by the ADR Wales Ymateb research team. It provides a snapshot of informative research currently underway at ADR Wales but is not intended to provide a complete picture of work undertaken within this field or the ADR Wales programme of work. The information presented in this Data Insight has been reviewed by ADR Wales colleagues with expertise within this thematic area and accepted to be accurate at the point of publication. Views expressed in this Data Insight are those of the researchers and not necessarily those of ADR Wales partner organisations.

The work is supported by ADR UK (Administrative Data Research UK). ADR UK is a partnership transforming the way researchers access the UK's wealth of public sector data, to enable better informed policy decisions that improve people's lives. ADR UK is an Economic and Social Research Council (ESRC) investment (part of UK Research and Innovation). [Grant number: ES/W012227/1].

Publication contact

Please contact the SAIL Databank via the Helpdesk.