International Journal of Population Data Science

Journal Website: www.ijpds.org

Including household effects in Big Data research: the experience of building a longitudinal residence algorithm using linked administrative data in Wales

Tingay, KS[1*], Roberts, M[1], and Musselwhite, CBA[1]

Submission History

Submitted:	27/12/2017
Accepted:	10/10/2018
Published:	20/11/2018

[1]Swansea University

Abstract

The effect of the wider social-environment on physical and emotional health has long been an area of study. Extrapolating the impact of the individual's immediate environment, such as living with a smoker or caring for a chronically-ill child, would potentially reduce confounding effects in health-related research. Surveys, including the UK Census, are beginning to collect data on household composition. However, these surveys are expensive, time consuming, and, as such, are only completed by a subsection of the population. Large-scale, linked databanks, such as the SAIL Databank at Swansea University, which hold routinely collected secondary use clinical and administrative datasets, are broader in scope, both in terms of the nature of the data held, and the population. The SAIL databank includes demographic data and a geographic indicator that makes it possible to identify groups of people that share accommodation, and in some cases the familial relationships among them. This paper describes a method for creating households, including considerations for how that information can be securely shared for research purposes. This approach has broad implications in Wales and beyond, opening up possibilities for more detailed population-level research that includes consideration of residential social interactions.

Background

Our immediate physical, social and emotional environment impacts on our health and wellbeing. For example passive smoking has been linked to cancer (e.g. [1]); carrying the burden of responsibility for ill or disabled family members increased the risk of depression and anxiety in Greek caregivers [2]; frequent house moves or changes to the household composition have been connected to increased depression, emotional distress, and marijuana use in adolescents [3, 4], and in lower educational attainment in younger children [5]. Fowler, Henry & Marcal [6], found that unstable household composition can have long-standing impacts on mental health, lifestyle and antisocial behaviour in adolescence and early adulthood. Moreover, adverse life events experienced by other members of the family, such as racist abuse, illness or financial deprivation, have also been shown to increase an individual's socio-emotional difficulties [7, 8].

In other respects, though, close social contact can be beneficial. Many studies have not only found an increase in survival time following diagnosis of cancer in married people compared with non-married patients (see, for example, [9, 10]), but also that married people tend to be diagnosed at an earlier stage than non-married patients [11]. Moreover, cohabitation with a significant other was a positive mitigating factor in survival of diseases, such as ovarian cancer [12].

It is clear from the above research that the immediate household environment can be seen as an important determinant of health and wellbeing. Surveys such as the UK Millennium Cohort Study [13], Add Health [14], and Understanding Society [15], as well as government-led surveys and censuses such as the Scottish [16] and Welsh [17] Health Surveys, have included questions about household composition and stability, but the ability to model this information using routinely collected health and administrative data sets is some way behind. Administrative datasets may be incomplete, particularly in transient populations, such as asylum seekers [18], or where data is context-specific [19]. Population-wide censuses, such as that conducted by the Office of National Statistics (ONS) in the United Kingdom (UK), are sufficiently financially costly and resource-intensive to only be carried out every ten years.

The ONS has done considerable work to construct households [20] using data from the dienally-collected census. Considerable interest in using administrative instead of census data resulted in a methodology for measuring household size and composition [21]. However, these estimates are currently based on 1% of addresses from existing 2011 Census data (the Population Coverage Survey (PCS)) [45] and only relate to snapshots in time [21]. For research purposes, a longitudinal view of household changes is desired. The ONS method also omits what are known as "complex addresses", such as blocks of flats and other communal residence types. Again, these are households of interest for at least some research purposes, especially for research involving issues such as socioeconomic

*Corresponding Author:
Email Address: k.s.tingay@swansea.ac.uk (KS Tingay)

https://doi.org/10.23889/ijpds.v3i1.452

factors, or spread of infection.

There are plans announced to use linked data to construct households, especially in relation to validating the PCS estimates, but these are for England first before being tested in other UK nations. Given existing Welsh linked data, there is scope for Welsh household modelling to be conducted alongside ONS's work, and to provide a dataset with a focus on research instead of population estimates.

While standardised terminologies such as Read and the Systematized Nomenclature of Medicine (SNOMED-CT) are widely used, many Electronic Health Records (EHRs) also use local codes that need to be mapped to a standardised terminology system in order to be linked with other EHRs of the same type [23]. Individuals must be anonymised in order to securely store their information, but this method must be robust enough to allow for datasets to link information on the same individual with few errors that might otherwise lead to spurious research results [24].

The ability to link records from different datasets, each containing a variety of different information, opens up huge possibilities for researchers, as well as challenges to keep data secure and non-identifiable. While there is increasing support for the benefits of data linkage for research purposes [25], aggregating individuals into groups risks increasing their identifiability. Given the well-known cost, relatively small sample size, and methodological issues of surveys [26], using routinely collected health and administrative datasets to produce household-level data appears to be an as yet untapped resource [27]. This article outlines the protocol for research modelling households using anonymised routine data.

Methods

Design and conceptual framework

The Secure Anonymous Information Linkage (SAIL) databank, developed by the Swansea University Medical School, has collated routine clinical and administrative datasets from around Wales since 2006 [28]. At present, SAIL holds data on inpatient hospital admissions, outpatient hospital visits, birth and immunisation records, cancer screening, emergency department attendances, mortality records, congenital anomalies, and data from over 70% of GP practices in Wales, among other, more specific, datasets [29]. Although SAIL does not hold data from all GP practices in Wales, all Welsh residents and associated socio-demographic data are available through the Wales Demographic Service (WDS), which is the population spine for data linkage [30]. Individual-level data are anonymised using a split-file process, where identifiable information such as name, address, and date of birth, is separated from study data, such as GP events, and sent to a Trusted Third Party for anonymisation. An Anonymised Linkage Field (ALF) is sent to SAIL for probabilistic linkage with health datasets. This ALF is unique to each individual, allowing for information to be securely linked across different datasets.

Each ALF is associated with a Residential Anonymised Record Linkage field (RALF, a de-identified address code based on the Unique Property Reference Number (UPRN) provided by the UK's Ordnance Survey (OS). Figure 1 shows the types of data held in SAIL and how these could be used to link individuals to a household, as well as to answer health and wellbeing related research questions.

UK geographical coding

The OS is the UK's national mapping agency, focusing on geographical surveying. The UPRN is a unique, linkable identifier for every British spatial address, which remains consistent across the life cycle of that address [31]. UPRNs are assigned by the governmental local authority responsible for that area, and form part of the National Address Gazetteer infrastructure, the UK address database of over 40 million addresses. UPRNs are allocated to new properties upon planning application approval, to sub-divided properties using a parent/child relationship of the subdivisions to the original ("parent") UPRN, and to merged properties (i.e. where two properties are knocked together to make a single property). Upon demolition or merging, a UPRN is considered "historic" [32].

These UPRNs are de-identified as RALF codes, which are placed in relation to a Lower Super Output Area (LSOA) code, a computer-generated geographic area created by the ONS. LSOA's are a sub-layer of the ONS population estimate hierarchical areas, known as Super Output Areas (SOAs), which have been in use UK-wide since the 2001 Census [32]. Each SOA level contains similar geographical and social populations and fit within existing government administrative boundaries.

The smallest SOA, Output Areas (OAs), consist of a minimum of 100 residents in 40 households, although up to 125 households was recommended where possible. In rural areas, such as parts of Wales, this minimum size is particularly important given the sparse population densities, although this does lead to some geographically large OAs compared to those in more urban areas.

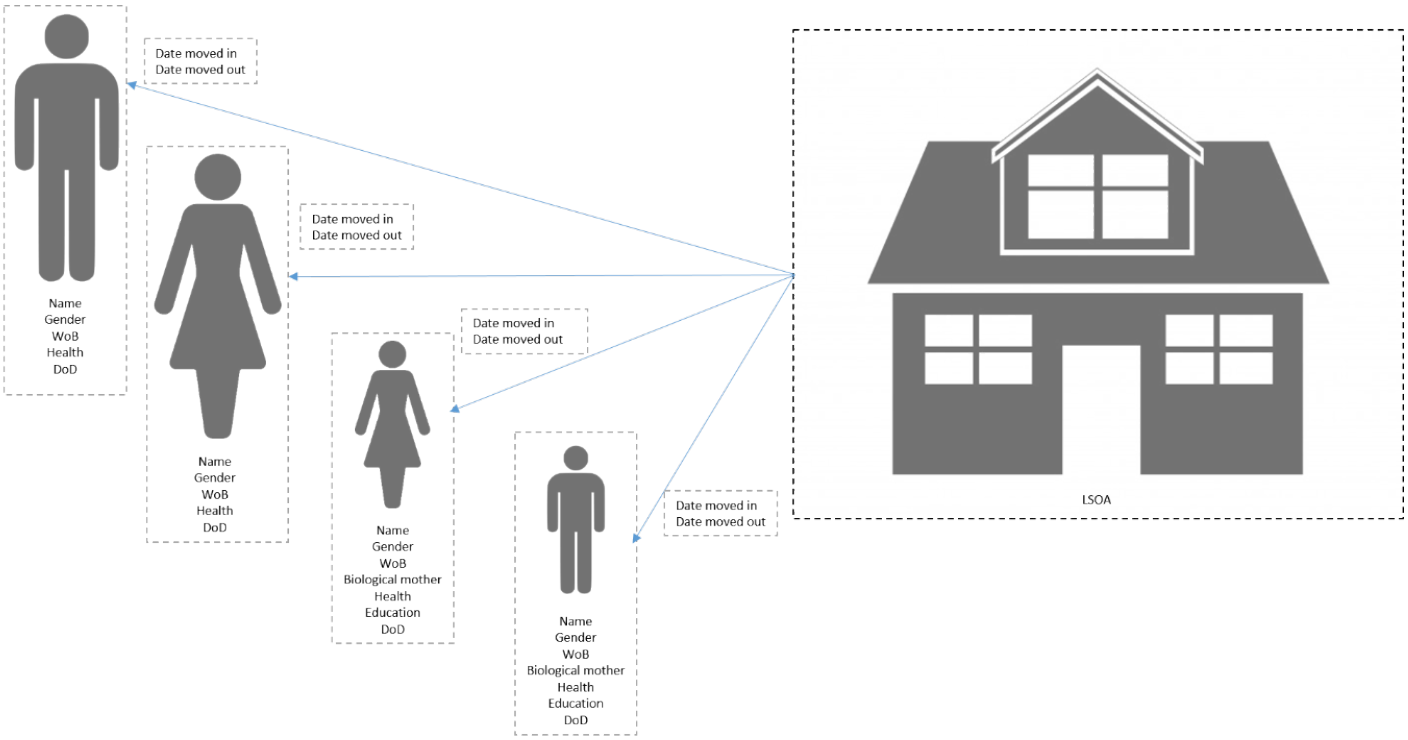
LSOAs consist of between 4 and 6 OAs and cover an average of 1,500 residents in 650 households [33], typically with a minimum of 1,000 and a maximum of 3,000 residents. Approximately 4 to 5 LSOAs make up a Middle Super Output Areas (MSOAs), which hold an average of 7,200 residents (minimum of 5,000) in 2,000 households [34]. The minimum sizes for the ONS area hierarchies aim to ensure that confidentiality is maintained while allowing for population estimates to be calculated [32].

In the 2016 ONS outputs, Wales contains 1,909 LSOAs in 22 areas, consisting of an average of 1,631 people (minimum 900, maximum 4,512) [34].

RALF linkage and maintenance

RALFs are linked to an individual's ALF based on the person's place of residence as recorded on their GP registration [35]. When an individual registers with a GP, their address is recorded for contact and identification purposes. While the address may be confirmed when the individual receives a prescription, or is referred for specialist care, the onus is on the patient to notify any changes in address while registered at that practice. In the UK, GP registration is on a location basis, with GPs only accepting patients who live within their practice boundary. If a patient moves out of the practice boundary, and notifies their GP of that fact, they risk being deregistered with that practice.

Figure 1: Data variables which can be used to identify people living at the same address



Despite the risk of inaccurate address data, this linkage between the RALFs and LSOAs allows geographical outputs to be mapped, while retaining individual-level anonymity. To date, RALFs have been used to analyse a variety of environmental factors, such as access to alcohol outlets [36], vaccination uptake [37], fuel poverty [38], and housing regeneration [39].

Development method

The ONS Census (2011) defines a household as "one person living alone or in a group of people (not necessarily related) living at the same address with common housekeeping - that is, sharing either a living room or sitting room or at least one meal a day" (pg. 2) [40]. Purely using routine administrative data, it is not possible to use the 2011 Census definition of a household, as existing routine datasets do not record whether household members share common housekeeping. Instead, a modified definition is proposed: "one person living alone or a group of people (not necessarily related) living at the same address as defined by the same UPRN". As the RALF for each resident contains the individual's moving in and out date, the possibility exists to group people living at a RALF at the same time. Using privacy-preserving methods, this project will use RALFs and moving dates as a base to create Household Anonymised Linkage Fields (HALFs) of individuals living at the same residence during the same time period. A set of rules will be created which define the creation and completion of households.

These HALFs will be validated using existing survey data that contain questions on household composition. As the time periods in which individuals living at a residence are likely to overlap, it is likely that further HALF iterations will be required, e.g. HALF 1.1, 1.2... 1.n. These iterations can be used as a measure of the stability of the household. Such versioning allows for both longitudinal modelling, in which changes can be measured as they occur, and cross-sectional modelling, where the state of the household is measured at a specified point in time, either as a date or as a life stage.

Households may further be divided into biological (or partly biological) and non-biological. For children born in Wales, the ALF of the biological mother is recorded. These Maternal ALFs make it possible to infer mothers living with their biological children. While it is not possible, presently, to identify a biological father, an adult male living in the same HALF as children from or prior to the birth of the child/ren could be used to infer, if not a biological relationship, then a significant relationship to both adult male and child/ren. Stable households which move to different RALFs, either as full households or, for example, two adults moving repeatedly together, can be implied as a "family". Part of this project will address the amount of time a household unit remains stable in order for members to impact on each other.

By keeping the ALF, RALF, Maternal ALF, and HALF, individual effects within a household can be measured, through a "target" individual or person type that is the focus of study, e.g. children, or individuals with particular diagnoses. At the same time, different households can be studied, such as single-parent families, non-familial households, students living in university halls accommodation, or residents of a care facility. The Maternal ALF field allows extended families to be studied, and using the RALF and associated data allows for the geographical distance between families to be included, which could be an indicator of family support. Work has already been completed on a family identifier that allows both resident and non-resident family members to be identified via the Maternal

ALF [41].

Challenges in identifying families and households

One of the challenges in identifying family groups from routine clinical data, whether biological or sociological, is that electronic record systems are geared at individuals, rather than at the wider familial context. While there is growing recognition of the importance of collecting family history (see, for example, [42]), it is also recognised that current EHR designs put the onus for collecting this information on the patient [43].

Data quality is a further challenge. The dates recorded for an individual's residence in a RALF are based on their registration at a GP practice. However, not all of the GP practice registrations are up to date. Welsh Government figures report that 3,197,633 individuals were registered with a Welsh GP in 2016 [44]. Mid-year estimates for that same year give a population total of only 3,113,150 individuals [34], suggesting that some individuals may not have de-registered from their old practice, or even that some are registered with multiple practices. GP registration relies on the patient to provide accurate information. Pharmacies and specialist health services require a current address for the purposes of posting appointment or test letters, or to confirm identity. Furthermore, parents may register their children before themselves. As a result, a child may be seen to be living at a RALF before any adults [45]. These data quality challenges will be addressed further in a separate article.

Validation

In this project, validation takes the algorithmic definition of "..the degree to which a model is an accurate representation of the real world from the perspective of its intended uses" [46, pg 6562]. In this instance, the best "real world" representation is the 2011 UK Census. The Census data are currently not available at the individual level in a way that would make it linkable to the routine data in SAIL, but aggregates of household sizes by LSOA are freely available from the ONS.

The HALF identifier will be measured against linked survey responses for different population groups, including children, and older adults. These surveys have been selected for their target populations, inclusion of household environment questions, and longitudinal nature. The latter will allow for the comparison of the identifier at different life stages. Where possible, survey respondents will be linked to their RALFs and HALFs at the time of the survey. This point prevalence will allow for direct comparisons between the survey and routine data. To allow for variations in population sizes as a whole, and any data issues beyond our control, the model will be considered valid if the results are within the 95% confidence level [47].

Variations in household composition between the data sources may show strengths and limitations of both methods of data collection. Where direct linkage is not possible, for example for aggregated results such as in the ONS Census, aggregated point prevalence comparisons will be drawn using the date of the data collection. As with the direct comparisons, there is likely to be some degree of variability between the observed and predicted data, requiring the use of 95% confidence levels.

Further detail on validation methods will be published separately.

Ethical approval and considerations

The more information collected on an individual, the greater the chance of inadvertent re-identification, even by trained researchers. Grouping individuals for whom so much is known only increases this risk. While SAIL currently employs policies to minimise accidental disclosure [48], these have not yet been tested on aggregated groups in this way. In order to apply privacy-preserving research controls, one must first understand the nature of the problem. Analyses will be performed to identify epidemiological areas of small numbers that can then be used to create a set of rules by which household-level data can be released. It may be, for example, that Lower Super Output Area (LSOA) level information would prove too disclosive for some research questions, and that MSOA or Health Board geographic divisions would be more appropriate. Equally, individual responses to living with a household member with a rare condition may lead to too-small numbers for analysis, and aggregating to household-level responses may allow for more rigorous methodology. This is likely to need to be addressed on a project-by-project basis.

Discussion

The option to include household-level details would add a level of complexity currently missing from Big Data studies. This would allow for a better understanding of deprivation and other factors, which could potentially impact on health and wellbeing. Studying the genome is an established research field, and advances in GIS methods and data capture is allowing the study of green and blue spaces on health and wellbeing (see, for example, [49, 50, 51]). By comparison, household effects, while known to be influential on population health, are lacking in Big Data research. That this is probably due to the current lack of available methodologies outside cohort studies, making this project, and others like it, all the more relevant to health and wellbeing research.

As well as for research, the model is also relevant to policy makers, who could make decisions based upon a more sophisticated portfolio of evidence than is available with existing methods. While there are known methodological issues around the existing data, not to mention unknown unknowns which will become apparent as the work progresses, the creation of such a tool promises to open up a range of research possibilities to researchers. While the identifier is being developed in Wales, the RALF and HALF methodologies could be applied throughout the UK and indeed to other nations with similar census type measures.

It is important to note the distinction between a research dataset and population statistics. While the UK's ONS work on household estimates will be of tremendous value to policy makers, there are questions about its utility as a research tool. The lack of longitudinal reporting, inability to link to other data sources, such as health and education, and omission of communal residences limit its usefulness in health and

social science research. In contrast, the model we are building includes household history, including following household members across different addresses, and changes to existing households over time. We have already conducted work into communal residences and are able to classify RALFs which are student halls of residence, residential care homes, and blocks of flats [52]. More work is required to refine the algorithms, however, particularly in relation to residential care homes and identifying households within blocks of flats.

The rules established by this project are likely to be relevant to any models developed by the ONS and other government household projects. Issues such as how households are defined using routine data, the strength of any mother-child linkages, multi-generational households, longitudinal and/or transient households, non-familial households, and communal households are common across residency datasets. These methods, while using a country-specific dataset, are likely to be of interest to international researchers and population data scientists interested in replicating building households from administrative data in their local jurisdiction. Such research opens up the possibility for more detailed population-level research through the analysis of residential social interactions.

Declarations

This study was approved by SAIL's independent Information Governance Review Panel (IGRP, project 0495).

Authors' contributions

KT conceptualized the study and its design and drafted the manuscript. MSR and CBAM contributed to the conception and design of the study. All authors reviewed the manuscript and approved it for publication.

Acknowledgements

This project was supported by Farr@CIPHER and the Administrative Data Research Centre Wales, and makes use of the SAIL Databank.

Availability of data and materials

The data supporting this study are available to access within the SAIL Databank and ADRC-W.

Statement on conflicts of interest

The authors declare that there is no conflict of interest.

References

1. Kim, C.H., Lee, Y.C., McNallan, S.R., Cote, M.L., Lim, W-Y., Chang, S-C., Kim, J.H., Ugolini, D., Chen, Y., Liloglou, T., Andrew, A.S., Onega, T., Duell, E.J., Field, J.K., Lazarus, P., Le Marchand, L., Neri, M., Vineis, P., Kiyohara, C., Hong, Y-C., Morgenstern, H., Matsuo, K., Tajima, K., Christiani, D.C., McLaughlin, J.R., Bencko, V., Holcatova, I., Boffetta, P., Brennan, P., Fabianova, E., Foretova, L., Janout, V., Lissowska, J., Mates, D., Rudnai, P., Szeszenia-Dabrowska, N., Mukeria, A., Zaridze, D., Seow, A., Schwartz, A.G., Yang, P., Zhang, Z-F. Exposure to secondhand tobacco smoke and lung cancer by histological type: A pooled analysis of the International Lung Cancer Consortium (ILCCO). International Journal of Cancer. 2014; 135(8): 1918-1930. https://doi.org/10.1002/ijc.28835

2. Govina, O., Kotronoulas, G., Mystakidou, K., Katsaragakis, S., Vlachou, E., & Patriraki, E. Effects of patient and personal demographic, clinical and psychosocial characteristics on the burden of family members caring for patients with advanced cancer in Greece. European Journal of Oncology Nursing. 2015; 19: 81–88. https://doi.org/10.1016/j.ejon.2014.06.009

3. Cavanagh, S. E. Family Structure History and Adolescent Adjustment. Journal of Family Issues. 2008; 29(7): 944–980. https://doi.org/10.1177/0192513X07311232

4. Cavanagh, S. Family structure and adolescent adjustment. Journal of Family Issues. 2015; 29(7): 944-980. https://doi.org/10.1177/0192513X07311232

5. Hutchings, H. A., Evans, A., Barnes, P., Demmler, J., Heaven, M., Hyatt, M. A., James-Ellison, M., Lyons, R., Maddocks, A., Paranjothy, S., Rodgers, S., Dunstan, F. Do children who move home and school frequently have poorer educational outcomes in their early years at school? An anonymised cohort study. PLOS One. 2013; 8(8), e70601. https://doi.org/10.1371/journal.pone.0070601

6. Fowler, P. J., Henry, D. B., & Marcal, K. E. (2015). Family and housing instability: Longitudinal impact on adolescent emotional and behavioral well-being. Social Science Research. 2015; 53: 364–374. https://doi.org/10.1016/j.ssresearch.2015.06.012

7. Becares, L., Nazroo, J., & Kelly, Y. A longitudinal examination of maternal, family, and area-level experiences of racism on children's socioemotional development: patterns and possible explanations. Social Science & Medicine. 2015; 142: 128–135. https://doi.org/10.1016/j.socscimed.2015.08.025

8. Pittman, L. & Boswell, M. Low-income multigenerational households: variation in family functioning by mothers' age and race/ethnicity. Journal of Family Issues. 2008; 29(7): 851-881. https://doi.org/10.1177/0192513X07312107

9. Aizer, A. A., Chen, M.-H., McCarthy, E. P., Mendu, M. L., Koo, S., Wilhite, T. J., Graham, P., Choueiri, T., Hoffman, K., Martin, N., Hu, J., Nguyen, P. L. Marital Status and Survival in Patients With Cancer. Journal of Clinical Oncology. 2013; 31(31): 3869–3876. https://doi.org/10.1200/JCO.2013.49.6489

10. Osborne, C., Ostir, G. V, Du, X., Peek, M. K., & Goodwin, J. S. The influence of marital status on the stage at diagnosis, treatment, and survival of older women

with breast cancer. Breast Cancer Research and Treatment. 2005; 93: 41–47. https://doi.org/10.1007/s10549-005-3702-4

11. Li, Q., Gan, L., Liang, L., Li, X., & Cai, S. The influence of marital status on stage at diagnosis and survival of ppatient with colorectal cancer. Oncotarget. 2015; 6(9): 7339–7348. https://doi.org/10.18632/oncotarget.3129

12. Ibfelt, E. H., Dalton, S. O., Høgdall, C., Fagö-Olsen, C. F., Steding-Jessen, M., Osler, M., Johansen, C., Frederiksen, K., Kjær, S. K. (2015). Do stage of disease, comorbidity or access to treatment explain socioeconomic differences in survival after ovarian cancer? – A cohort study among Danish women diagnosed 2005–2010. Cancer Epidemiology. 2015; 39: 353–359. https://doi.org/10.1016/j.canep.2015.03.011

13. Smith, K. & Joshi, H. The Millennium Cohort Study. Population Trends. 2002; Spring(107): 30-34

14. Klein, J.D. The National Longitudinal Study of Adolescent Health: Preliminary Results: Great Expectations. JAMA. 1997; 278(10): 864-865 https://doi.org/10.1001/jama.1997.03550100090045

15. Buck, N. & McFall, S.L. Understanding Society: design overview. Longitudinal and Life Course Studies. 2012; 3: 5-17 https://doi.org/10.14301/llcs.v3i1.159

16. Scottish Government. Scottish Health Survey. Retrieved from https://www.gov.scot/Topics/Statistics/Browse/Health/scottish-health-survey

17. NHS Wales Public Health Wales Observatory. Welsh Health Survey. Retrieved from http://www.publichealthwalesobservatory.wales.nhs.uk/welsh-health-survey

18. Aspinall, P. J. The extent of collection of information on migrant and asylum seeker status in routine health and social care data sources in England. International Journal of Migration, Health and Social Care. 2007; 3(4): 3–13. https://doi.org/10.1108/17479894200700020

19. Avidan, A., & Weissman, C. Record completeness and data concordance in an anesthesia information management system using context-sensitive mandatory data-entry fields. International Journal of Medical Informatics. 2012; 81(3): 173-181. https://doi.org/10.1016/j.ijmedinf.2011.12.009

20. Harper, G., Mayhew, L. Using administrative data to count and classify households with local applications. Applied Spatial Analysis. 2016; 9: 433-462 https://doi.org/10.1007/s12061-015-9162-2

21. Office for National Statistics. Annual assessment of the ONS's progress on the Administrative Data Census: July 2018. Retrieved from https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusannualassessments/annualassessmentofonssprogressontheadministrativedatacensusjuly2018

22. Office for National Statistics. Research Outputs: An update on developing household statistics for an Administrative Data Census. Retrieved from https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusresearchoutputs/householdsandfamilies/researchoutputsanupdateondevelopinghouseholdstatisticsforanadministrativedatacensus

23. Abhyankar, S., Demner-Fushman, D., & McDonald, C. J. Standardizing clinical laboratory data for secondary use. Journal of Biomedical Informatics. 2012; 45(4): 642–650. https://doi.org/10.1016/j.jbi.2012.04.012

24. Anderson, J. G. Security of the distributed electronic patient record: a case-based approach to identifying policy issues. International Journal of Medical Informatics. 2000; 60(2): 111–118. https://doi.org/10.1016/S1386-5056(00)00110-6

25. Jones, K.H., Laurie, G., Stevens, L., Dobbs, C., Ford, D.V., Lea, N. The other side of the coin: Harm due to the non-use of health-related data. International Journal of Medical Informatics. 2017; 97: 43-51 https://doi.org/10.1016/j.ijmedinf.2016.09.010

26. Lilford, R. J., Richardson, A., Stevens, A., Fitzpatrick, R., Edwards, S., Rock, F., & Hutton, J. L. Issues in methodological research: perspectives from researchers and commissioners. Health Technology Assessment. 2001; 5(8). https://doi.org/10.3310/hta5080

27. Lyons, R. A., Ford, D. V., Moore, L., & Rodgers, S. E. Use of data linkage to measure the population health effect of non-health-care interventions. The Lancet. 2014; 383(9927): 1517–1519. https://doi.org/10.1016/S0140-6736(13)61750-X

28. Lyons, R. A., Hutchings, H., Rodgers, S. E., Hyatt, M. A., Demmler J., Gabbe, B.J., Brooks, C. J., Brophy, S., Jones, K.H., Ford, D.V., Paranjothy, S., Fone, D., Dunstan, F., Evans, A., Kelly, M., Watkins, W., Maddocks, A., Barnes, P., James-Ellison, M., John, G. & Lowe, S. Development and use of a privacy-protecting total population record linkage system to support observational, interventional, and policy relevant research. The Lancet. 2012; 380(Supplement 3): S6. https://doi.org/10.1016/S0140-6736(13)60362-1

29. Ford, D., Jones, K., Verplancke, J.-P., Lyons, R., John, G., Brown, G., Brooks, C., Thomspon, S., Bodger, O., Couch, T. & Leake, K. The SAIL Databank: building a national architecture for e-health research and evaluation. BMC Health Services Research. 2009; 9, e157. https://doi.org/10.1186/1472-6963-9-157

30. NHS Wales Informatics Service. Welsh Demographic Services. Retrieved from http://www.wales.nhs.uk/nwis/page/52552

31. GeoPlace. The UPRN lifecycle from planning to demolition. Retrieved from https://www.geoplace.co.uk/documents/10181/41984/2015%20the%20UPRN%20lifecycle%20V3%20%28CMS%20ID%20-%201429701616057%29

32. Office for National Statistics. Census geography: an overview of the various geographies used in the production of statistics collected via the UK Census. Retrieved from https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography

33. Office for National Statistics. Census Geography. Retrieved from https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography

34. Office for National Statistics. Mid-2016 Population Estimates for Lower Layer Super Output Areas in England and Wales by Single Year of Age and Sex – Supporting Information. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/lowersuperoutputareamidyearpopulationestimates

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

36. Fone, D., Dunstan, F., White, J., Webster, C., Rodgers, S., Lee, S., Shiode, N., Orford, S., Weightman, A., Brennan, I., Sivarajasingam, V., Morgan, J., Fry, R. & Lyons, R. Change in alcohol outlet density and alcohol-related harm to population health (CHALICE). BMC Public Health. 2012; 12(1): 428. https://doi.org/10.1186/1471-2458-12-428.

37. Hutchings, H. A., Evans, A., Barnes, P., Healy, M. A., James-Ellison, M., Lyons, R. A., Maddocks, A., Paranjothy, S., Rodgers, S.E. & Dunstan, F. (2015). Does frequent residential mobility in early years affect the uptake and timeliness of routine immunisations? An anonymised cohort study. Vaccine. 2015; 34(15): 1773–1777. https://doi.org/10.1016/j.vaccine.2016.02.049

38. Heaven, M., & Lowe, S. Data Linking Demonstration Project - Examining Fuel Poverty using Home Energy Efficiency Data (HEED) and Routinely Collected Health Data. Welsh Government. 2013; Number 77/2013. Retrieved from http://gov.wales/docs/caecd/research/131206-data-linking-demonstration-project-examining-fuel-poverty-heed-routinely-collected-health-data-summary-en.pdf

39. Rodgers, S. E., Heaven, M., Lacey, A., Poortinga, W., Dunstan, F. D., Jones, K. H., Palmer, S., Phillips, C., Smith, R., John, A., Davies, G. & Lyons, R. A.. Cohort profile: The housing regeneration and health study. International Journal of Epidemiology. 2014; 43(1): 52–60. https://doi.org/10.1093/ije/dys200

40. Office for National Statistics. 2011 Census. Household Questionnaire. Retrieved from https://calls.ac.uk/wp-content/uploads/2013/05/EngWalescensus2011.pdf

41. French, R., Tingay, K.S. Identifying families in Welsh administrative data. Administrative Data Research Network conference, Edinburgh, UK. 2017

42. Murray, M.F., Giovanni, M.A., Klinger, E., George, E., Marinacci, L., Getty, G., Brawarsky, P., Rocha, B., Orav, E.J., Bates, D.W., Haas, J.S.. Comparing electronic health record portals to obtain patient-entered family health history in Primary Care. Journal of General Internal Medicine. 2013; 28(12): 1558-1564. https://doi.org/10.1007%2Fs11606-013-2442-0

43. Nathan, P.A., Johnson, W., Clamp, S., Wyatt, J.C.. Time to rethink the capture and use of family history in primary care. British Journal of General Practice. 2016; 66(653): 627-628. https://doi.org/10.3399/bjgp16X688273

44. Statistics for Wales. GPs in Wales, 2006-2016. Retrieved from http://gov.wales/docs/statistics/2017/170329-general-medical-practitioners-2006-2016-en.pdf

45. Tingay, K.S., Hyatt, M., Demmler, J.C., Brooks, C.J., Lyons, R. Generation of family units in the SAIL Databank. In Scottish Health Informatics Programme. 2013.

46. Sornette, D., Davis, A.B., Ide, K., Vixie, K.R., Pisarenko, V., Kamm, J.R.. Algorithm for model validation: Theory and applications. PNAS. 2007; 104(16): 6562-6567. https://doi.org/10.1073pnas.0611677104

47. Marcus, A.H., Elias, R.W. Some Useful Statistical Methods for Model Validation. Environmental Health Perspectives. 1998; 106(Suppl 6): 1541-1550 https://doi.org/10.1289/ehp.98106s61541

48. Jones, K. H., Ford, D. V, Jones, C., Dsilva, R., Thompson, S., Brooks, C. J., Heaven, M., Thayer, D., McNerney, C. & Lyons, R. A.. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: A privacy-protecting remote access system for health-related research and evaluation. Journal of Biomedical Informatics. 2014; 50: 196–204. https://doi.org/10.1016/j.jbi.2014.01.003

49. Rodgers, S., Wheeler, B., White, J., White, M., Lovell, R., Fry, R., Berridge, D., Stratton, G., Nieuwenhiujsen, M., Lyons, R. Green-blue space exposure changes and impact on individual-level well-being and mental health: a population-wide record-linked natural experiment. 2018. Retrieved from https://www.journalslibrary.nihr.ac.uk/programmes/phr/160707/#/

50. Song, J., Fry, R., Mizen, A., Akbari, A., Wheeler, B., White, J., White, M., Lovell, R. Association between blue and green space availability with mental health and wellbeing. 2018; 3(3): 330. https://doi.org/10.23889/ijpds.v3i4.921

51. De Vries, S., Ten Have, M., van Dorsselaer, S., van Wezep, M., Hermans, T., de Graaf, R. Local availability of green and blue space and prevalence of common mental disorders in the Netherlands. BJPsych Open. 2016; 2(6): 366-372. https://doi.org/10.1192/bjpo.bp.115.002469

52. Tingay, K.S., Roberts, M.S., Musselwhite, C.B.A. Identifying anonymous residence types using administrative data. Informatics for Health conference, Manchester, UK. 2017. https://doi.org/10.13140/RG.2.2.35870.95045