Residential Anonymous Linking Fields (RALFs): a novel information infrastructure to study the interaction between the environment and individuals’ health

Sarah E. Rodgers1, Ronan A. Lyons1, Rohan Dsilva1, Kerina H. Jones1, Caroline J. Brooks1, David V. Ford1, Gareth John2, Jean-Philippe Verplancke1

1Centre for Health Information Research and Evaluation, School of Medicine, Grove Building, Swansea University, Singleton Park, Swansea SA2 8PP, UK
2Health Solutions Wales, 12th Floor, Brunel House, 2 Fitzalan Place, Cardiff CF24 0HA, UK
Address correspondence to Sarah E. Rodgers, E-mail: s.e.rodgers@swansea.ac.uk

ABSTRACT

Background The necessity of aggregating health data over areas can impede our understanding of health determinants.

Methods We demonstrate the possibility of creating anonymous links between individual residences and the local environment using digital map data and a data linkage system.

Results Digital map data were used successfully to anonymously link 1.3 million addresses to the local environment. The data linkage system allows detailed environment data surrounding each residence to be linked both to each resident therein and to their medical records.

Conclusions Local environment data specific to each house can be effectively and anonymously linked to the population registered with the National Health Service. Our integrated approach potentially enables flexible fine-scale, large-area observational studies of communities and health.

Keywords epidemiology, environment, public health

Background

Ideally spatial epidemiological studies would consist of data on individuals, including accurate residential location, characteristics, travel, personal exposures together with a longitudinal health record. However, health statistics are not generally handled at the individual level because of data confidentiality issues demanded by legislation. If disease data are available for individuals then it is usual to use a geographical masking process to protect the confidentiality of individuals, but at the same time accurate geographical analyses must be ensured if the results are to be meaningful. Health records are aggregated within a geographical area that has a sufficiently large population to ensure prevention of disclosure. This is known as an ecological approach. Critics of studying aggregated level population data argue that the results may wrongly attribute average population characteristics to individuals within the population; this is known as ecological fallacy.

Methods such as Bayesian mapping have been developed to compensate for some characteristics of area mapping. Although aggregated data protects the privacy of individuals, it limits the types of research that can be conducted. For example, the exposure risk from lead associated with road traffic pollution requires population areas to be defined linearly along the road, which are different to the pre-defined areas often associated with administrative or political units, such as electoral wards or lower super output areas (LSOA). In addition, there is much interest in the built environment and the extent to which it supports walking. Ideally, individual walking routes would be defined local to each residence, with the environment along the route characterized and...
analysed along with individuals’ health data. It is currently difficult to have such well-defined environments and corresponding health data without breaching confidentiality legislation, as access to high spatial resolution health data are extremely limited.

Data suppression is an alternative method to data aggregation used to produce non-identifiable data. We aggregate data across bands (e.g. in 10 year age bands), and counts with less than an arbitrary number of cases, often set at five, are suppressed; replaced by a symbol so the exact number is not known. There have been some concerns that even with some level of data suppression and aggregation, individuals may be identified, albeit with some considerable effort. This is particularly true for non-spatial data. For identification of individuals from spatial data, authors generally provide no practical solutions for avoidance of this risk. Seemingly in opposition to the concerns of many dealing with confidentiality, there is a great deal of research attempting to increase the spatial accuracy of disease cases using geocoding. The majority of this work has been completed in the USA. Many concede there is a margin of error for the most accurately geocoded data. It is agreed that advances in the availability of health, social and spatial environmental data, and the related techniques used to manage and analyse these data, provide researchers with important new methods to study health issues that are worth developing further. It is likely that methods that both increase geocoding accuracy and preserve individuals’ anonymity would be well received.

The removal of personal identifiers also removes precise spatial links and therefore locational identifiers. In effect, data are again aggregated within an area but removal of high-resolution spatial identifiers means data now comply with confidentiality legislation. Ordinarily this approach would exclude the possibility of linking all data for a particular individual. Generally, the resulting level of data aggregation from the de-identification process prevents researchers from performing microlevel spatial analyses necessary for important research problems. Once personal identifiers are removed, it also prevents researchers from linking together health incidents for the same person. However, to overcome part of the problem, it is possible to create an anonymous linking system for individuals’ health data. Even though spatial identifiers are not retained, the ability to combine all health events for one person together under a unique, anonymous identifier is possible. One can argue that such an anonymous linking system would in fact be more powerful than simply spatially co-locating diseases. This is because health events would be anonymously linked to the same person and the individual’s health could be attributed to the local community and even lifestyle choices.

The Health Information Research Unit (HIRU) has developed methods to overcome data confidentiality issues for records pertaining to individuals in the SAIL (Secure Anonymous Information Linkage) databank whilst retaining the capability to link data at the individual anonymized record level. We have overcome the challenge of storing health and social data for over 2 million people. These data are drawn from a range of routine data sources such as general practitioner (GP) records, hospital admissions, outpatient appointments, accident and emergency attendances, school achievement, screening programmes, cancer registration as well as pathology results. The investment that has been made in developing the infrastructure to obtain, store and secure these data means that optimum use should be made of the facility and as wide a range of research as possible supported.

Briefly, SAIL data have all identifiers removed and are therefore anonymous. However, data linkages can increase risk of disclosure, particularly if secondary data users combine the linked data with other spatially explicit information about individuals. This could affect not only the research participants, but also the reputation of HIRU and other health research groups whose credibility depends on participants’ confidence and promises of confidentiality. There are rigorous procedures used to ensure that data are not made available to view when there is the possibility of identification. Checks for rare records are made using HIRU’s NEMO (Numerical Evaluation of Multiple Outputs) algorithm and additional suppression is applied in such cases to minimize disclosure risk.

The method we describe here retains individuals’ anonymity while environmental metrics are created with sufficient precision to allow interesting and powerful analyses. Our novel technique not only makes it possible to link residences to individual’s medical data, but it does this in accordance with information governance by adhering strictly to data protection legislation and confidentiality guidelines.

At HIRU an Enterprise GIS system using ArcServer GIS is enabled on the Blue-C Super Computer, containing digital map data for Wales. This system has allowed us to construct a spatial health database. In addition to the Anonymous Linking Field (ALF), data in the SAIL databank retain a LSOA code; an area system developed by the UK’s Office of National Statistics (ONS). LSOAs have consistent boundaries and a minimum population of 1000, with an average of 1500 people. They are the smallest geographical unit available for the Census data. They are also presently the highest spatial resolution at which health data are made available to view by authorized researchers. The ONS allow research institutions to use these data because at this level of
aggregation data are unlikely to allow identification of individuals. Researchers are provided with restricted access to anonymous individually linked HIRU data after we have completed minimal data suppression ensuring researchers cannot identify individuals. The anonymous linking method uses a process of data splitting, a trusted third party, double encryption to anonymize and link individuals’ health records, and replaces identities with ALFs.26 This method makes it possible to follow a series of health service encounters and health outcomes through the health-care system for large numbers of anonymized individuals, now combining them with high-resolution spatial environment data.

Although an individual’s health and social data can presently be linked together using unique ALFs, the relation to other individuals within the same household and the household characteristics remain unknown. The novel method we have developed produces Residential Anonymous Linking Fields (RALFs) for every household and embeds individual ALFs within this framework. Such a system has many potential uses for environmental health research. Using the RALF system we are able to determine if there are chronic disease patterns within co-habiting groups because we have information about who shares a residence. We are likely, but not necessarily, looking at a family unit when the ages of the resident’s fall within certain age ranges. Therefore, we can investigate the influence of adult health conditions on the children in the household. The list of possible applications using this method is infinite but one final example, and perhaps the most important benefit of this method, will be following the long term effects of exposures measured in cross-sectional surveys. Often causality cannot be proven from a cross-sectional survey and the anonymous linkage of individual medical records to their environment will allow us to follow changes in environmental exposure along with changes in health for a number of years. It will also allow us to assess historical health conditions.

Of the three basic epidemiological characteristics, time, place and person, place has been the least used because of difficulties in defining spatial locations in a meaningful manner without disclosing identities.6 A particular individual’s relation to features in the environment is generally unknown and is most usually estimated from small area averages. The method presented here is able to recombine derived spatial information from precisely the point of residence with the anonymized individual data whilst always retaining anonymity. The next section will describe details of RALFs; the spatial extension to the anonymous linking method.

**Objective and method**

We demonstrate the possibility of creating anonymous links between individual residences and their local environment using digital map data, ensuring that both residents and the residence retain their anonymity. We achieve this using a data linkage system, which links health data to individuals and residences using RALFs, providing for a two-stage process linking individuals to spatial information derived from the local environment at a high spatial resolution (Fig. 1). The method we present here will enable the environment to be quantified precisely without directly
linking locational information with health data and therefore maintaining confidentiality. There are several steps to this process; however, the main processes are covered under the headings: address matching and environment metrics.

**Address matching**
The NHS Administrative Register (NHSAR) is a key database for this method, together with standardized address data. In the UK, the Royal Mail Postcode Address File (PAF) data has 28 million addresses. Additionally, each person registered with a General Practitioner has their address recorded in the NHSAR dataset. The original address matching process is performed by Health Solution Wales (HSW) using Royal Mail PAF data. The NHSAR addresses are matched to the PAF using software, such as QAS Pro address software from Experian and Matchcode, and PAF address keys are then derived and assigned to those addresses. A RALF is created from this process; a RALF, which in effect is an encrypted equivalent of the PAF address key. A dynamic table of ALFs and their relation to RALFs is also produced and held by HSW and HIRU. These data are double encrypted by HIRU meaning that data stored on the secure SAIL database remain unidentifiable even by HSW. The process is stated briefly here.

(i) Creation of a list of all addresses—combining and comparing address lists—NHSAR, PAF, Address Layer 2 (AL2).

(ii) Assigning unique number to each address and encrypting number to produce RALFs.

(iii) Development of a table which links ALFs to RALFs with entry and exit dates.

**Environment metrics**
AL2 data contain a point for each residential address and all points of interest, for example, shops, public houses, recreation grounds and bus shelters. This information, together with the Topography and Integrated Transportation Network layers, can be used to create environment metrics suitable to be used in analyses exploring the built environment’s effect on health. The process of capturing access to services is shown in Fig. 1 and is the first step in the total process (Fig. 2, a). The spatial error of Ordnance Survey Topography layer for urban areas is less than ±1 m.27 The points contained within the AL2 data are generally positioned towards the centre of buildings surveyed for the Topography layer. The list of AL2 addresses and GIS-derived metrics are provided to HSW (Fig. 2, b) who then match addresses with NHSAR list, attaching RALFs (Fig. 2, c). The environment metrics undergo data masking and necessary categorization (e.g. NEMO) before sending to HSW to ensure the absence of unique values. Metrics received by HSW will not be identified and labelled as metric_1, for example, therefore, although HSW can view these data they will not be meaningful. The anonymous RALF and corresponding environment metric only are returned to HIRU (Fig. 2, d). Finally health and environment data may be combined after matching ALFs with RALFs (Fig. 2, e).

Environment data are independent and completely separate from health data to allow environment data

![Fig. 2 Residential Anonymous Linking Field (RALF) matching process.](https://academic.oup.com/pubhealth/article-abstract/31/4/582/1494003)
collection using the original precise location of a residence and later linked anonymously. This innovative technique, using links between RALFs and the AL2 layer, allows an area of exposure to be created specific to a residence instead of relying on pre-defined regions. Examples of possible environment metrics for each residence include distance to the nearest doctor's surgery, dwelling density within 500 m, travel distance to a leisure centre using public transport, road junction density within 1000 metres. Metrics for the newly defined areas resulting from this process are not unique and therefore cannot be traced back to a particular residence. Finally, within a secure environment, these environment data are linked with health data, as shown in Fig. 2 (step e). This process is also shown diagrammatically in Fig. 1.

Results and discussion

Main finding

We have successfully devised a secure method anonymously linking individuals and their health data both to others in the same household and to their local environment. Two benefits as a direct result of the method using RALFs have been described in this paper. These are, firstly, linkage of individuals to each other when they share the same residence, and secondly, linkage of residences (and implicitly individuals) to the local environment at a high spatial resolution.

The matching and RALF creation process was tested in October 2008 and it is possible to assign a valid current address key to 96.5% of individuals. Historical data matching is also of a high quality; it is possible to assign an accurate address key to 95.7% of individuals, based on their address 10 years ago. The matching process created 1 321 709 RALFs, or distinct residences. At a census date, there were 2 947 193 ALFs associated with these RALFs. The use of RALFs has made possible a number of previously unachievable analyses. For example, there is now the potential to ascertain if co-habiting people have similar health issues. People in the same household can be linked together to provide valuable information analysing infectious diseases. Environment metrics specific to individuals and each project can be designed using GIS data.

What is already known on this topic

Anonymously linking health and environmental data at a high spatial resolution carries potential risks of identification through sparse data. We have developed a method to ensure we avoid this risk. Recently published literature is concerned with the inadvertent or deliberate re-identification of original addresses from low resolution map publications, even after obscuring algorithms have been applied. Our method does not contain spatial identifiers; therefore, we cannot publish the specific location of all individuals relative to one another in map form. Therefore, it is an advantage of our method that it does not publish or provide the researcher with geocoded disease cases. The possibility of increasing the spatial accuracy of residential geocodes, but at the same time maintaining confidentiality is possible when returning results in a non-spatial database format. All non-spatial statistics are possible, for example, regression or time series analyses.

What this study adds

The method we describe in this paper allows researchers to have access to data to make progress in understanding a variety of social and environmental processes operating at high resolutions. For example, analysis of housing and health data analyses will benefit from the specific type of linkages provided by the RALF system. Housing conditions, including the date of household insulation improvements from a government sponsored intervention, will be linked to RALFs as an environment metric. Using anonymous data stored in the SAIL databank we will be able to longitudinally monitor health conditions (e.g. respiratory tract infections, cardiovascular conditions and falls) of co-habiting groups in residences with and without insulation improvements. If health data were not collected before the intervention then it would not ordinarily be possible to complete a full assessment of health changes. However, retrospective health assessment for individuals in residences is possible with the RALF method linking to routinely collected health data.

Importantly, this approach is much less expensive than setting up large-scale longitudinal studies.

This method also allows the definition of the built environment surrounding each residence with a view to characterizing the walking potential of the neighbourhood. Environment metrics, such as dwelling density or road connectivity, may then be joined with individuals’ medical information to analyse the health effects of living in a ‘walkable’ environment. Additionally, confounders to environmental pollution, such as smoking status, can be captured using datasets received from GP practices. Additional information can also be anonymously linked from social and medical surveys.

Concentrating on associations observed at the individual level may not capture the processes operating at the collective level of the community. The method described here goes some way towards not only retaining the accuracy of individuals but also putting individuals in the context of a
family. This construct is more realistic because we have included the acknowledgement that people do not operate in isolation but are in fact part of a complex society. Geographical perspectives seek to develop theories explaining how individuals interact with their wider environment. Atomistic (individual) and family (co-habiting) data are combined with ecological data for areas. An extension to this method would be the identification of people who interact, on a daily or weekly basis, in the community. We will also investigate opportunities to link children to their school through the National Pupil Database. An environmental variable may be weighted for pollutants in proximity to school, or altered to include the walking potential of the route from home to school.

Limitations
The many benefits of this method have been discussed and this novel system has succeeded in all aspects as intended. However, it should be noted that because results are returned in database format and have no remaining spatial identifiers; it is not possible to complete spatial analyses. The lack of important confounder variables at the level of the individual may be a limitation. If important data are not available then these could be collected in a survey, which would then be anonymously linked to the routine data. A sample collected from survey data would require additional funding to complete, however, this is likely to be worthwhile in some studies.

Conclusions
The combination of data linkage methods, large-scale computing power with Enterprise GIS system and high-resolution Ordnance Survey data have allowed first environmental metrics to be calculated from map data and allowed secondly environmental data specific to each individual’s residence to be anonymously linked to routinely collected health, social and education data. In this way, we can explore the effect the local environment has on their physical and mental health in a way that has not been possible previously on a large scale with routinely collected data. Comparisons are planned between standard ecological analyses and those using data obtained from the RALF method to quantify differences. The use of routine data will also allow for savings in both time and money. With the aid of qualitative data, it may be possible to extend the environment to include the community with which an individual interacts to enhance further the focus on an individual.

Authors’ contributions
G.J. carried out the matching process, S.E.R. wrote the first draft of the manuscript. K.H.J., C.J.B., D.V.F., R.A.L. reviewed and critiqued the manuscript and made substantial intellectual contributions to subsequent drafts. G.J., S.E.R., K.H.J., D.V.F. and R.A.L. conceived the idea for this method. R.D. and J.-P.V. participated in its design and coordination.

Funding
HIRU is supported by a grant from the Wales Office of Research & Development, National Assembly for Wales.

References
22. Ford DV, Jones KH, Verplancke J-P et al. The SAIL Databank: building a national architecture for e-health research and evaluation (Submitted to BMC HSR for review, Feb 2009).