Exeter, D. J., Rodgers, S., & Sabel, C. E. (2013). "Whose data is it anyway?" The implications of putting small area-level health and social data online. Health policy (Amsterdam, Netherlands), 114. 10.1016/j.healthpol.2013.07.012

Health Policy xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Health Policy

journal homepage: www.elsevier.com/locate/healthpol

"Whose data is it anyway?" The implications of putting small area-level health and social data online

Daniel John Exeter[a,*], Sarah Rodgers[b,1], Clive Eric Sabel[c,d,2]

[a] School of Population Health, The University of Auckland, Private Bag 92019, Wellesley Street, Auckland 1142, New Zealand
[b] Centre for Health Information Research and Evaluation, Swansea University, Singleton Park SA2 9PP, United Kingdom
[c] Department of Geography, College of Life & Environmental Sciences, University of Exeter, Amory Building, Rennes Drive, Exeter EX4 4RJ, United Kingdom
[d] European Centre for Environment and Human Health, University of Exeter Medical School, Knowledge Spa, Royal Cornwall Hospital, Truro TR1 3HD, United Kingdom

ARTICLE INFO

Article history:
Received 31 May 2012
Received in revised form 23 May 2013
Accepted 15 July 2013

Keywords:
Web 2.0
Privacy
Medical record linkage
Access
Data collection
Confidentiality

ABSTRACT

Data from electronic patient management systems, routine national health databases, and social administrative systems have increased significantly over the past decade. These data are increasingly used to create maps and analyses communicating the geography of health and illness. The results of these analyses can be easily disseminated on the web often without due consideration for the identification, access, ethics, or governance, of these potentially sensitive data. Lack of consideration is currently proving a deterrent to many organisations that might otherwise provide data to central repositories for invaluable social science and medical research. We believe that exploitation of such data is needed to further our understanding of the determinants of health and inequalities. Therefore, we propose a geographical privacy-access continuum framework, which could guide data custodians in the efficient dissemination of data while retaining the confidentiality of the patients/individuals concerned. We conclude that a balance of restriction and access is needed allowing linkage of multiple datasets without disclosure, enabling researchers to gather the necessary evidence supporting policy changes or complex environmental and behavioural health interventions.

1. Introduction

Developments in the secure management of large routine health and demographic datasets and efforts to democratise data availability [1] over the past decade have led to their increased use by policy analysts, academics, and NGOs. Many studies continue to use such data for cross-sectional analyses [2–15], however there is a growing recognition among the academic community that the strength of routine data is the ability to create 'cohorts' by linking records from multiple health and social datasets to better examine an individual's interaction with the health system and its association with particular outcomes [16–20]. Given the significant improvements in geographic referencing (the process of converting street addresses or postcodes/zip codes to map coordinates) of health events, it is not surprising that a large proportion of database-derived cohorts are interested in the geography of health. One example is the Secure Anonymised Information Linkage (SAIL) Databank [21], which uses probabilistic linkage to construct a cohort comprising the health trajectories of over 2 million Welsh residents.

* Corresponding author. Tel.: +64 9 923 4400; fax: +64 9 373 7503.
E-mail address: d.exeter@auckland.ac.nz (D.J. Exeter).
[1] Tel.: +44 01792 602308; fax: +44 1792 513430.
[2] Tel.: +44 01392 722297; fax: +44 01392 723342.

0168-8510/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.healthpol.2013.07.012

The SAIL databank is being used to examine variations in health service costs and the association between health and the built environment [22,23].

Geographical Information Systems (GIS) provides substantial support for the management and availability of (spatial) data. GIS have undergone considerable changes over the past decade with commercial GIS packages progressing from standalone software packages to the development of GIS applications for desktop, server, web and mobile GIS, not to mention the inclusion of Cloud Computing [24]. Similar developments have been observed in the development of Open Source GIS. As Evans and Sabel [25] have demonstrated, extensive spatial analytical functionality can now be incorporated to webGIS. For example, MySQL and PostgreSQL, (coupled with PostGIS) are two popular open source database management systems (DBMS) widely used for GIS applications These DBMS may be integrated with the MapServer (http://www.mapserver.org) and GeoServer (http://www.geoserver.org) packages to provide open source WebGIS, with limited functionality. Major multinational corporations interested in the management of (spatially enabled) data, such as Google Inc. are now leveraging these developments via inter-linked databases and (mapping) products to provide tools to users over the web to be able to query and explore data.

The plethora of health and social data and tools to analyse them now becoming available on the web, combined with both a Web 2.0 savvy generation and an increasing workforce of non-geographically trained 'experts' in WebGIS has led to a further development in the visualisation of these data over the web. Use of 'mash-ups' of spatially enabled data from a variety of sources, raises a concern that one can use the additive power of datasets to infer results more revealing than the individual datasets allow. At present, the transmission of health data over the Internet varies immensely by geographical region, geographic scale, in the method of delivery and extent of user interaction. For example, users interested in the global variations in life expectancy might extract tables from the United Nations for analyses not online. Indeed, data downloaded from the United Nations, World Bank and World Health Organisation were used in the production of the WorldMapper online atlas (http://www.worldmapper.org). Alternatively, users interested in regional health variations may be attracted to the NHS atlas of healthcare variation, available at http://www.sepho.org.uk/extras/maps/NHSatlas2011/atlas.html. Here, users choose a topic of interest and the InstantAtlas software presents a regional map of England, linked to a histogram that outlines the region's performance (Fig. 1). At the other extreme, users visiting the US National Cancer Institute's website (http://ratecalc.cancer.gov/) select a specific type of cancer and the strata to produce a map at their chosen geographical scale. The user can export these maps as an image and also drill down to extract further information regarding cancers at the county level. Glover and Jenkins [26] used a similar but Flash-based webGIS to enable community mapping in Australia, that allowed community members to upload and map their own (health) datasets to share, entrusting the administration and maintenance of their 'projects' to a third party.

Glover and Jenkins' webGIS is an example of the dual-use dilemma that confronts users of health and social data on the Internet. In the life sciences, the dual-use dilemma refers to instances where the same scientific work can have a beneficial or hazardous use – the dilemma being the inability to prevent the misuse without foregoing the beneficial uses [27]. While DNA synthesis, for example, may have numerous potential benefits, there is potential for this technology to be used for bioterrorism. We contend that there is also a dual-use dilemma with respect to the proliferation of health and social data: On the one hand, for the benefit of society and specifically advancements in medical understanding, publically funded data should be disseminated widely. On the other hand, some of these data are potentially sensitive and should be carefully managed.

In this paper, we discuss some of the opportunities and concerns associated with making available potentially sensitive data and outline a proposed spatial-privacy framework to guide researchers. First we outline concerns over 'Big Data' before describing the benefits that may be achieved through the use of high resolution spatial data. In doing so, we consider why health and social data should be released and to whom. We conclude by proposing a framework for the efficient use of health and spatial data whilst preventing misuse, in response to the concerns and issues that we raise throughout the paper.

2. What are the concerns?

In the digital era, there is growing concern that potentially identifiable information is increasingly available without an individual's consent. Real concerns centre around so called 'mash-ups' of data – combination of multiple data sources independent of each other, but which together could potentially reveal more as a whole than the sum of the individual parts. With smart-phone technology increasingly widely used, so called 'Big-Data' is available at our finger-tips. There is now the potential to electronically track in space and time a user either covertly [28] or overtly, for example when users manually enable geo-tagging in Twitter.

Civilian access to more accurate geospatial digital data from Global Positioning System (GPS), coupled with digital imagery was pivotal in the development of Google's "Street View" product. Although undoubtedly a commercial product, one could reasonably argue that Google are providing 'Street View' in good faith, allowing users to familiarise themselves with a destination they are locating. Despite Google's capture of geo-coded photos from public spaces however, privacy advocates have objected to the 'Street View' service as some images reveal individuals in compromising circumstances, such as patients leaving abortion clinics, individuals climbing residential security gates, and other lewd behaviour. Thus, care must be taken when sensitive information accompanies location data [29].

Such concerns do vary from country to country, however. Socially conservative countries such as the USA appear to be at one end of the (protectionist) spectrum,

Please cite this article in press as: Exeter DJ, et al. "Whose data is it anyway?" The implications of putting small area-level health and social data online. Health Policy (2013), http://dx.doi.org/10.1016/j.healthpol.2013.07.012

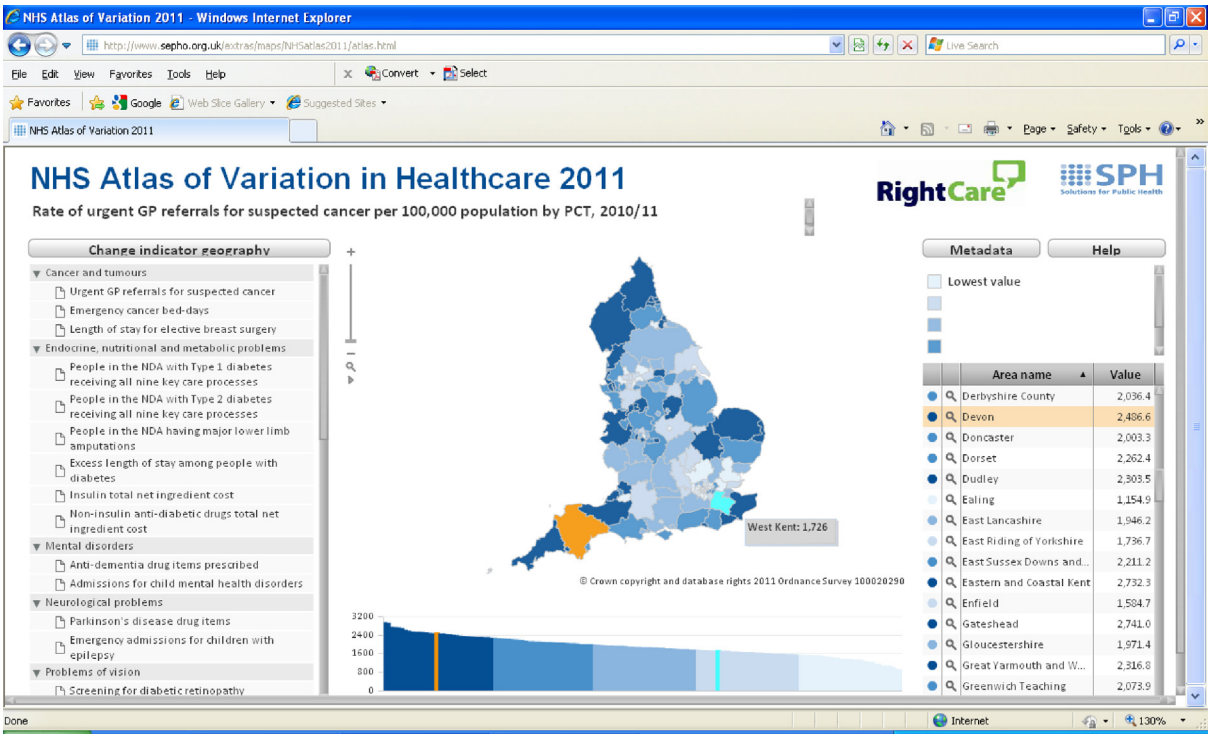

Fig. 1. Screen shot of the NHS Atlas of Healthcare Variation, with a region selected demonstrating brushing and linking functionality.

with perhaps Scandinavian countries at the other. In Scandinavia, it is possible, for example, to legally obtain detailed tax records for individuals if a social security number is known [30]. There is an underlying philosophy in Scandinavian countries (e.g. [31]) that information gathered with public funds should be made available for the benefit of society, subject to appropriate safeguards. Contrast that with recent debate in anglophile countries where the trend is moving to ever tighter protectionism, but note in the UK the debate opening out as British PM David Cameron articulates his 'Big Society' concept – a concept that places wider society at least on a par with narrower individual self-interest [1].

Consider the nature of the data contained within electronic medical records (EMRs), which has evolved extensively over the past decade [32–35], to provide accurate socio-demographic and clinical information, across the health system with relative ease. From a patient's perspective, confidentiality is paramount in the patient–provider relationship, and a threat to this relationship may restrict their willingness to reveal information that is crucial for a correct diagnosis [36]. A recent study in New Zealand interviewed 203 patients concerning their perceived attitudes to EMRs and asked about their willingness to share information with four different types of EMR user: health professional; health administrator; researchers; and 'other'. As expected, respondents were significantly more likely to share non-identifiable information and that the type of EMR users influenced the responses. Over 70% of respondents would consider release of information to health professionals, this reduced substantially to about 40% for release of information to the research community. When data were identifiable however, patients were more conservative, with approximately 50% and less than 20% of patients likely to allow health professionals and researchers access to their records, respectively [37].

3. Why should data be released?

The data social scientists use for their research, whether obtained from small surveys or from large national datasets, are expensive to collect. Given that these data are often funded from the public purse, there is a strong argument that these datasets should be made available for analysis by other researchers for the benefit of society. Especially when we consider that the amount of data obtained can far exceed the research questions for which results can be published. The UK, Canada, Australia and New Zealand provide data repositories for the archival of health and social data. Indeed, the UK Data Archive (http://www.data-archive.ac.uk) was established in 1967 and continues to provide a valuable service designed to encourage scientific enquiry and debate. A report by the UKDA [38] outlines the many benefits of having a resource that include: maximising transparency of research; enabling new collaborations; minimising the duplication of data collection; and responding to the increasing demands by some journals that data be deposited for use by the wider scientific community. Given the continued reduction of research funds available for population health research, there is a strong argument for supporting such a resource.

Please cite this article in press as: Exeter DJ, et al. "Whose data is it anyway?" The implications of putting small area-level health and social data online. Health Policy (2013), http://dx.doi.org/10.1016/j.healthpol.2013.07.012

There has been significant financial investment in a number of longitudinal studies, such as the Longitudinal Study of England and Wales (e.g. [39]), the Scottish Longitudinal Study [40] and the Millennium Birth Cohort [41], yet access to these remains relatively difficult. This raises the question of whether the investment to creating expensive studies such as these is worthwhile, if access to the data is so restricted that they are effectively unusable. Conversely, allowing researchers to more easily access these data would surely help understand the impact that determinants have on various health outcomes.

4. Why individual level data are needed

Environment, social and individual factors all have a role to play in an individual's health and wellbeing [42,43]. Linking social and health data to a particular place is important because where we live can and does influence our health. Health outcomes are related to an individuals' physical and social environment, including factors such as water, soil and air content, exposure to hazardous materials, tobacco smoke, occupation, marital status, social support, characteristics of the home, in addition to the composition of the local built environment [44,45].

Spatial epidemiologists employ a powerful suite of analytical methods in search of explanations of how the environment affects people's health and disease risk [46,47]. To conduct this research in practice, disease cases for individuals are often required; these are often located in space by researchers using the residential address of the person. The availability of an address, or proxy coordinates, enable the most powerful spatial analyses, potentially providing new insights into disease aetiology [46,48,49]. The distance over which some environmental factors are influential are understood to be small, in the order of less than 400 m [50]. This implies that any (partial) suppression of geographic location would render the data inadequate for purpose and the results less useful than might otherwise be possible. Naturally, results obtained from these analyses should never be published at the individual level, but rather should either be suitably aggregated, either spatially or statistically, or appropriately masked [51].

Disease cases aggregated into small areas may allow spatial patterns closer to those in reality to emerge, however the potential for individuals to be identified is also increased [52,53], so area units are often disproportionately larger in rural areas partly to prevent identification of individuals in sparsely populated areas [54]. The aggregation of health data into different sized spatial units, and indeed different zone configurations at the same spatial resolution, (the so-called Modifiable Areal Unit Problem) [55,56] also introduces errors in which the observed relationship between disease outcomes and exposures may vary immensely. Aggregated data mapped "on the fly" may contain errors, including large variance, and should be interpreted with caution. Prevalence calculations rationalise the number of disease cases by the number of people at risk of having the disease per area, therefore generally restricting users to pre-defined census areas with population estimates. In another branch of health geography, health service researchers are interested in the supply of health facilities as determined by a maximum distance from the facility to homes. Many calculations of population demand per facility use (road) network-based distance measurements and represent individuals using (census) area centroids because of their compliance with patient privacy law [57].

Research has often relied on cross-sectional studies using data captured at a single point in time, however longitudinal and intervention studies with a time series of environmental exposures and health events are required to infer causal relationships [58,59]. A sequence of health events associated with changing residential exposure is often the highest spatio-temporal resolution data available for individuals, [60] although availability of such data is limited. Such analyses often rely upon databases supported by national population and residential registers, with the Scandinavian countries notably leading the world in this regard. An example of the type of longitudinal residential exposure analysis that these registers allow is the work by Sabel et al. exploring potential environmental or genetic explanations of ALS in Finland, which revealed place of birth effects with important implications for disease causation, which otherwise would remain hidden [61,62]. Advances in spatial analytical techniques have followed technological advances and now specialist software designed to analyse these complex data are available, including SatScan for spatio-temporal cluster detection [63], SpaceStat (http://www.biomedware.com), for exploratory spatial data analysis; spatial econometric analyses, and GeoDa [64] for descriptive spatial data analysis, such as spatial autocorrelation statistics, in addition to basic spatial regression functionality and a number of geovisualisation techniques. In addition the "R" statistical software (http://www.r-project.org/) has an extensive library of open source packages that enable the analysis and visualisation of spatial data. Therefore, demand for access to longitudinal individual level health data with residential based locations by researchers is high, for example, to explore spatio-temporal patterns aiming to improve our understanding of disease causality.

5. Who should have access to these data: legal aspects of having access to individual level data

Geoprivacy is concerned with the privacy and confidentiality of geographic data and is particularly important in health research. Health data collected with point georeferences are increasingly unavailable to researchers because laws (in the UK and the US, for example) protect sensitive information held for individuals [65,66]. Restrictions preserving privacy are often placed by the data collection organisation at the time of data collection, either before release to researchers or online into the public domain. The information that is most often the first to be subjected to suppression is geographic location; knowledge of an individual's address will almost certainly disclose their identity. Often, address information for an individual will be aggregated into a local small area unit, for which population estimates are available. Nevertheless, research has shown that, if desired, reverse-geocoding can in some cases be used to identify unique addresses [67],

Please cite this article in press as: Exeter DJ, et al. "Whose data is it anyway?" The implications of putting small area-level health and social data online. Health Policy (2013), http://dx.doi.org/10.1016/j.healthpol.2013.07.012

alternatively the use of non-geographic attributes used in combination may be used for identification of individuals, especially if the level of aggregation is too small [68]. A balance is needed – it may be possible to release data at a higher spatial resolution if fewer proxy identifiers such as age, gender or occupation, accompany the data. Different levels of data privacy may be released to trusted researchers in comparison to that released online into the public domain.

The organisation providing data (data custodian) may place restrictions on data usage by a researcher to only answer the specific research question(s) using the methods outlined in the data request and/or ethics application. In general, the more sensitive the data, the greater will be the number of conditions of use imposed by ethical review and the data custodian. For example, diseases such as tick-transmitted human *babesiosis* may have individual level address data released on the understanding the data are password protected and will only be used by specified researchers for the submitted purpose [69]. However, the release of individual-level data for more sensitive health conditions is unlikely to be released to researchers. This is likely the case for particular groups of population; disease cases for children and other vulnerable groups will be released more cautiously, if at all. These restrictions are likely to create omissions in the research completed for those groups for whom research could be most beneficial.

While some data custodians may be reluctant to share their data initially, there is certainly a need to establish a trusted custodian-user relationship. By developing protocols that ensure the data are managed (by both parties) in an ethical manner, systems can be created to facilitate a trustworthy working relationship. This then removes the "onus" from an individual researcher to a system built with information governance in mind. In these cases, submitting the work to an information governance review committee prior to publication would be reassuring to both the data provider and researcher. Although some epidemiologists have access to coordinates for individual disease cases and could be viewed as privileged, there are obvious disclosure pitfalls of this system compared to anonymised health databanks, incorporating expert information governance checks [21,70]. Researchers may publish results believing them to be anonymised and adhering to disclosure rules, to find later there is sufficient material available either on their maps, or the internet, to identify individuals [67,71].

Researchers from industry and academia in the UK already have access to anonymised clinical and demographic data governed by the NHS, but other routine datasets exist that are under-utilised. The CEO of the Economic and Social Research Council (ESRC), Professor Paul Boyle ([72], p.19) strongly believes that "*[w]e need more active engagement with the public – a 'social contract' based on an informed understanding of research benefits... We have to explain how data are reliable, valuable, and can be properly managed. [and the] failure to make better use of routinely collected public data can be argued to be a criminal waste of public resources*" and we strongly support this perspective. The extent to which health data should be available to researchers or the general public is currently the subject of wider societal debate. The European Union is currently considering revisions to its data protection guidelines, and researchers in Europe are concerned that the revisions would restrict access to individual-level data currently used in medical research. Similarly, David Cameron, the UK Prime Minister, proposed a number of changes for NHS-maintained data access. Of particular relevance was his vision to ensure that all patient data would be included in clinical research, although patients could opt-out if they chose [73].

6. A framework for providing access to spatial health and social data

Here, we suggest that access to health and social data could be determined by the degree to which individuals may be identified and the potential benefit of analysis of such data, and propose a privacy-access continuum framework (Fig. 2) drawing on our collective experiences working with geographic health data in different countries. Within this framework, the level of anonymity would specify the potential levels of access for the different types of end-users. Consider electronic medical records for example, and assume that individuals have a unique patient identification code (PIC). This may be the National Health Service number in the UK, Social Security number in the US, or the National Health Indicator in New Zealand. Typically, the PIC and key demographic characteristics of an individual are associated with each unique contact a patient has with the health system: visiting the GP, medicines dispensed, blood tests at a community laboratory and procedures in hospital. These routine datasets would have a range of end-users, including clinicians, researchers, policy analysts and the public. With regard to levels of access, one would reasonably expect the clinician to have the most access, in order to offer the best treatment regime to a patient. In this instance, all of the patient's data is 'live' – including their PIC, demographics, and past medical history. Depending on the research question of interest, researchers and policy analysts would require data from one or more routine databases. Under these circumstances, the data released may be for individuals following the removal of 'live' PICs. Requests from researchers and analysts would require careful consideration of the level of geographical identifiers released and providers could use the disclosure rules endorsed by their national statistics department for guidance. Often, the questions asked by the public are answered using aggregate data, such as prevalence and incidence rates, stratified by age, gender, and/or geographical areas. However, one would recommend considering disclosure rules again, ensuring small numbers do not increase the risk of identification.

A further example of the privacy-access continuum (Fig. 2) relates to the use of residential address data for geographical analyses. Using residential data provides opportunities to understand the aetiology of diseases, as individuals can be identified in relation to specific exposures. Small areas may be used for spatial analyses, so individuals are "lost in the crowd", but when used with data for other key determinants (e.g. age, occupation) these data once again have the ability to identify individuals. The use of a full UK postcode (Unit postcode), containing an average

Please cite this article in press as: Exeter DJ, et al. "Whose data is it anyway?" The implications of putting small area-level health and social data online. Health Policy (2013), http://dx.doi.org/10.1016/j.healthpol.2013.07.012

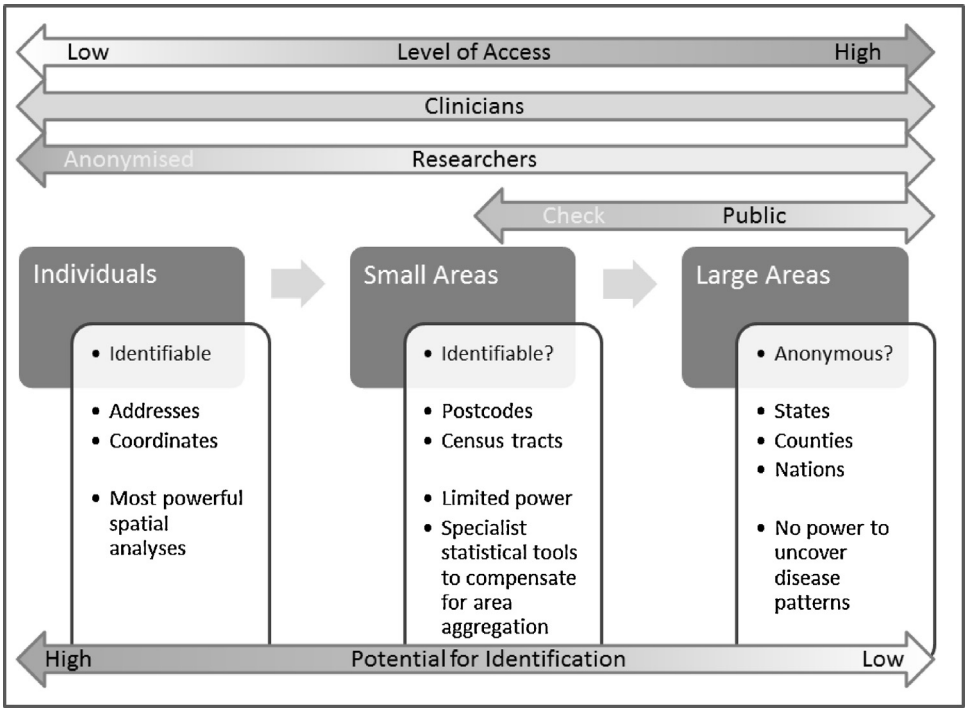

Fig. 2. The privacy-access continuum framework.

of 16 homes, will nearly always make individual records identifiable. This is particularly a problem with rare diseases because individuals will be likely to be unique within small areas. Large area aggregation is less likely to enable identification of individuals, but the ability to perform analyses that usefully identify environment or social causes of disease reduces significantly. Ideally, researchers will use anonymised systems to reduce disclosure risk. Thus, any raw health or social data entering the public realm would be checked to ensure they comply with information governance. However, this should not be at the expense of research that may benefit society by reducing the burden of ill-health, particularly to disadvantaged groups.

When considering the privacy-access continuum for large health datasets, at one extreme are anonymised databanks. Databanks require a large initial investment but continue to make large strides forward assessing public health using anonymous individual-level data. Although the data can be made available to researchers, mechanisms put in place to maintain privacy standards are time consuming to implement. Indeed, the anonymisation procedure is likely to render certain types of high resolution spatial analyses intractable. Therefore, a balance must be struck between access and privacy.

There is an increasing trend for some custodians of national surveys to provide users the opportunity to access 'microdata' [74–76]. In the UK, for example, the academic community have access to a 1% sample of anonymised records from the 1991 and 2001 censuses. Data security initiatives, such as ensuring an individual is 'present' in only one of the many different tables available, minimises the risk of identification. Furthermore, geographic analyses can only be conducted at the Local Government Area scale. While some NZ microdata can be accessed on CD-ROM, (subject to approval) [74] the more common approach is to allow users to conduct analyses at a secure location (i.e. at the Statistics Office, on a dedicated machine). Users are therefore are able to exploit the benefits of individual data, and their spatial attributes, BUT only the results from the analyses are removed from the secure data laboratories. Some organisations now have systems in place to allow researchers remote access to project-specific data [70].

At the other extreme are the population registers maintained in Scandinavian countries. By maintaining a near real-time longitudinal population register, with the ability to link to other registers such as births and deaths, housing, buildings and tax registers [30], these data provide numerous exciting research possibilities subject to appropriate ethical approval procedures. It is interesting to speculate about the future of national population registers elsewhere with debate raging over the decline of decennial population censuses in the United Kingdom.

7. Linkage – and putting a system of checks in place:

Secure microdata facilities and anonymised databanks employ data linkage mechanisms and these serve not only to maintain privacy but also link disparate silos of health data that would not ordinarily be associated. The combination of different data types, for example, social care and health data, is made possible using anonymous linkages that maintain privacy of individuals. Incoming data (such as quarterly updates of hospitalisation) may be joined to data already in the system for a particular individual even though the identity of that person is unknown. Many data providers would be less likely to provide such enriched

social and health data for research unless a trusted system was in place to assure that privacy was maintained and that the researcher has a good reputation in handling these complex datasets. Some organisations maintaining databanks physically house the data in one location [21,70] whereas others maintain data with the providers (e.g. Office of National Statistics, Police, Social Development) and have a mechanism for combining data tailored to a particular research question [77,78]. It is beyond the scope of this paper to design or detail the technical mechanisms to implement this framework, however, there are many papers which do so (see for example, Ford et al. [70]).

One of the many benefits of data linkage systems is their ability to conduct more complete longitudinal studies for a reduced cost [23]. Aggregated data could be approved through information governance procedures and subsequently placed online, however, raw data and complete analysis including results should be kept on a secure server, thereby providing a data release prevention mechanism to reassure data providers. Under this model, the researcher should only be allowed to download/receive the aggregated results, either in the form of data for large administrative units, tables, figures and summaries of statistical procedures. Performing the quality assurance processes may be time consuming and therefore delay the publication of data online.

8. Conclusions

We have reflected on the numerous opportunities that the increasing availability of health and social data have created, particularly with respect to the publication of spatial health and social data on the Internet. We argue that there is a dual use dilemma in making these data available. On the one hand, publically funded data should be disseminated widely to benefit society. On the other, there is a risk that the release of such information can be used for detrimental purposes. To this end, we present a privacy-access framework that proposes an opportunity to maximise the possibilities to better understand the geography of health while mitigating adverse effects of releasing individual-level data. There is a plethora of research demonstrating the value of spatial data in a population health context [3,5,10,20,22,25,26,39,41,46,54,55,57,59,61–63,69,70], yet the use of large, anonymised datasets is in its relative infancy. While each data provider has policies (e.g. [70,77]) concerning the release of their datasets, we argue that there is a need to develop guidelines with a view to standardising data management policies. Therefore, our paper welcomes a debate on both sides of the argument regarding access to such datasets. Further conversations are required to negotiate mechanisms that maximise societal benefit from access to data while simultaneously preserving the confidentiality of individuals.

References

[1] McHale JV. Using anonymized NHS data without consent: a step too far? British Journal of Nursing 2012;21(1):54–5 [Mark Allen Publishing].
[2] Adamson JA, Ebrahim S, Hunt K. The psychosocial versus material hypothesis to explain observed inequality in disability among older adults: data from the West of Scotland Twenty-07 Study. Journal of Epidemiology and Community Health 2006;60((11) November):974–80 [Research Support, Non-U.S. Gov't].
[3] Cox M, Boyle PJ, Davey P, Morris A. Does health-selective migration following diagnosis strengthen the relationship between Type 2 diabetes and deprivation? Social Science & Medicine 2007;65((1) July):32–42 [Research Support, Non-U.S. Gov't].
[4] Doran T, Drever F, Whitehead M. Is there a north-south divide in social class inequalities in health in Great Britain? Cross sectional study using data from the 2001 census. BMJ 2004;328((7447) May):1043–5.
[5] Exeter DJ, Boyle PJ. Does young adult suicide cluster geographically in Scotland? Journal of Epidemiology and Community Health 2007;61((8) August):731–6 [Research Support, Non-U.S. Gov't].
[6] Exeter DJ, Boyle PJ, Norman P. Deprivation (im)mobility and cause-specific premature mortality in Scotland. Social Science & Medicine 2011;72((3) February):389–97 [Research Support, Non-U.S. Gov't].
[7] Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R. Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: The Public Health Disparities Geocoding Project (US). Journal of Epidemiology and Community Health 2003;57((3) March):186–99 [Research Support, U.S. Gov't. P. H. S.].
[8] Mackenbach JP, Cavelaars AE, Kunst AE, Groenhof F. Socioeconomic inequalities in cardiovascular disease mortality; an international study. European Heart Journal 2000;21((14) July):1141–51 [Comparative Study Research Support, Non-U.S. Gov't].
[9] Manuel DG, Leung M, Nguyen K, Tanuseputro P, Johansen H. Burden of cardiovascular disease in Canada. Canadian Journal of Cardiology 2003;19((9) August):997–1004 [Research Support, Non-U.S. Gov't].
[10] Norman P, Boyle P, Exeter D, Feng Z, Popham F. Rising premature mortality in the U.K.'s persistently deprived areas: only a Scottish phenomenon? Social Science & Medicine 2011;73((11) December):1575–84 [Research Support, Non-U.S. Gov't].
[11] Pearce J, Witten K, Bartie P. Neighbourhoods and health: a GIS approach to measuring community resource accessibility. Journal of Epidemiology and Community Health 2006;60((5) May):389–95 [Research Support, Non-U.S. Gov't].
[12] Prus SG, Gee E. Gender differences in the influence of economic, lifestyle, and psychosocial factors on later-life health. Canadian Journal of Public Health Revue canadienne de sante publique 2003;94((4) July–August):306–9 [Comparative Study Research Support, Non-U.S. Gov't].
[13] Weich S, Twigg L, Holt G, Lewis G, Jones K. Contextual risk factors for the common mental disorders in Britain: a multilevel investigation of the effects of place. Journal of Epidemiology and Community Health 2003;57((8) August):616–21 [Research Support, Non-U.S. Gov't].
[14] Yngwe MA, Diderichsen F, Whitehead M, Holland P, Burstrom B. The role of income differences in explaining social inequalities in self rated health in Sweden and Britain. Journal of Epidemiology and Community Health 2001;55((8) August):556–61 [Research Support, Non-U.S. Gov't].
[15] Boyle P, Exeter D, Feng Z, Flowerdew R. Suicide gap among young adults in Scotland: population study. BMJ 2005;330((7484) January):175–6 [Research Support, Non-U.S. Gov't].
[16] Blakely T, Woodward A, Salmond C. Anonymous linkage of New Zealand mortality and Census data. Australian and New Zealand Journal of Public Health 2000;24((1) February):92–5 [Research Support, Non-U.S. Gov't].
[17] Brewer N, Wright CS, Travier N, Cunningham CW, Hornell J, Pearce N, et al. A New Zealand linkage study examining the associations between A1C concentration and mortality. Diabetes Care 2008;31((6) June):1144–9 [Research Support, Non-U.S. Gov't].
[18] Mehta S, Wells S, Riddell T, Kerr A, Pylypchuk R, Marshall R, et al. Under-utilisation of preventive medication in patients with cardiovascular disease is greatest in younger age groups (PREDICT-CVD 15). Journal of Primary Health Care 2011;3((2) June):93–101 [Research Support, Non-U.S. Gov't].
[19] Reitsma JB, Kardaun JW, Gevers E, de Bruin A, van der Wal J, Bonsel GJ. [Possibilities for anonymous follow-up studies of patients in Dutch national medical registrations using the Municipal Population Register: a pilot study]. Nederlands tijdschrift voor geneeskunde 2003;147((46) November):2286–90 [Research Support, Non-U.S. Gov't].
[20] Rodgers SE, Demmler JC, Dsilva R, Lyons RA. Protecting health data privacy while using residence-based environment and demographic data. Health & Place 2011;18:209–17.
[21] Lyons RA, Johnes KH, John G, Brooks CJ, Verplancke JP, Ford DV, et al. The SAIL databank: linking multiple health and social care

Please cite this article in press as: Exeter DJ, et al. "Whose data is it anyway?" The implications of putting small area-level health and social data online. Health Policy (2013), http://dx.doi.org/10.1016/j.healthpol.2013.07.012

datasets. BMC Medical Informatics and Decision Making 2009;9((3) January):24.
[22] Rodgers SE, Lyons RA, Dsilva R, Jones KH, Brooks CJ, Ford DV, et al. 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) December):582–8 [Research Support, Non-U.S. Gov't].
[23] Lyons RA, Fone D, Rodgers SE, Paranjothy S, Hyatt M. Wales Electronic Cohort for Children. clinicaltrialsgov: NCT01136681. [Observational Study]. 2010 (NCT01136681).
[24] ESRI. http://support.esri.com/en/content/productlifecycles. Redlands, USA: Environmental Science Research Institute, (Inc.); 2012 [cited 01.09.12].
[25] Evans B, Sabel CE. Open-source web-based geographical information system for health exposure assessment. International Journal of Health Geographics 2012;11((1) January):2.
[26] Map your own data: supporting community action.Glover J, Jenkins P, editors. 14th international medical geography symposium (IMGS). 2011.
[27] UK Parliamentary Office of Science and Technology. The Dual-Use Dilemma. London: Parliamentary Office of Science and Technology; 2009. p. 4.
[28] BBC Technology News. iPhone tracks users' movements. BBC. Available from: http://www.bbc.co.uk/news/technology-13145562; 2011 [cited 01.31.13].
[29] nypost.com. Google Street Views cool or creepy? New York Post Online; 2007 [07.06.07].
[30] Sabel CE, Dorling D, Hiscock R. Sources of income, wealth and the length of life: an individual level study of mortality. Critical Public Health 2007;17(4):293–310.
[31] Ministry of Justice (Sweden). Personal data protection: information about the personal data act. Ministry of Justice (Sweden), ed., Stockholm. http://www.government.se/content/1/c6/07/43/65/0ea2c0eb.pdf; 2006.
[32] Brandeis GH, Hogan M, Murphy M, Murray S. Electronic health record implementation in community nursing homes. Journal of the American Medical Directors Association 2007;8((1) January):31–4 [Evaluation Studies].
[33] Bria 2nd WF, Shabot MM. The electronic medical record, safety, and critical care. Critical Care Clinics 2005;21((1) January):55–79, viii [Review].
[34] Rosenbeck KH, Randorff Rasmussen A, Elberg PB, Andersen SK. Balancing centralised and decentralised EHR approaches to manage standardisation. Studies in Health Technology and Informatics 2010;160(Pt. 1):151–5.
[35] Yoo S, Kim B, Park H, Choi J, Chun J. Realization of real-time clinical data integration using advanced database technology. In: AMIA annual symposium proceedings/AMIA symposium AMIA symposium. 2003. p. 738–42.
[36] NZHIS. Health information privacy and confidentiality. Wellington: New Zealand Health Information Service; 1995.
[37] Whiddett R, Hunter I, Engelbrecht J, Handy J. Patients' attitudes towards sharing their health information. International journal of medical informatics 2006;75((7) July):530–41 [Research Support, Non-U.S. Gov't].
[38] UK Data Archive. Managing and Sharing Data: a best practice guide for researchers. Available from: http://www.data-archive.ac.uk/media/2894/managingsharing.pdf; 2011 [17.01.12].
[39] Hattersley L, Creeser R. Longitudinal Study 1971–1991: history, organisation and quality of data. London: OPCS; 1995.
[40] Boyle PJ, Feijten P, Feng Z, Hattersley L, Huang Z, Nolan J, et al. Cohort profile: the Scottish Longitudinal Study (SLS). International Journal of Epidemiology 2009;38((2) April):385–92.
[41] Smith K, Joshi H. The Millenium birth cohort study. Population Trends 2002;107:30–4.
[42] McLeroy KR, Bibeau D, Steckler A, Glanz K. An ecological perspective on health promotion programs. Health Education Quarterly 1988;15((4) Winter):351–77.
[43] Dahlgren G, Whitehead M. Policies and strategies to promote social equity in health. Stockholm: Institute of Future Studies; 1991.
[44] Pickle L, Waller L, Lawson A. Current practices in cancer spatial data analysis: a call for guidance. International Journal of Health Geographics 2005;4(1):3.
[45] Marmot M. Social determinants of health: from observation to policy. Medical Journal of Australia 2000;172((8) April):379–82.
[46] Goovaerts P. Combining area-based and individual-level data in the geostatistical mapping of late-stage cancer incidence. Spatial and Spatio-temporal Epidemiology 2009;1:61–71.
[47] Norman GJ, Nutter SK, Ryan S, Sallis JF, Calfas KJ, Partrick K. Community design and access to recreational facilities as correlates of adolescent physical activity and body-mass index. Journal of Physical Activity and Health 2006;3(Suppl.):S118–28.
[48] Sabel CE, Gatrell AC, Löytönen M, Maasilta P, Jokelainen M. Modelling exposure opportunities: estimating relative risk for motor neurone disease in Finland. Social Science & Medicine 2000;50(7/8):1121–37.
[49] Sabel CE, Boyle P, Raab G, Löytönen M, Maasilta P. Modelling individual space–time exposure opportunities: a novel approach to unravelling the genetic or environment disease causation debate. Spatial and Spatio-temporal Epidemiology 2009;1(1):85–94.
[50] Langford M, Fry R, Higgs G. Measuring transit system accessibility using a modified two-step floating catchment technique. International Journal of Geographical Information Science 2012;26(2):193–214.
[51] Armstrong MP, Rushton G, Zimmerman DL. Geographically masking health data to preserve confidentiality. Statistics in Medicine 1999;18((5) March):497–525 [Research Support, Non-U.S. Gov't Review].
[52] Kawakami N, Winkleby M, Skog L, Szulkin R, Sundquist K. Differences in neighborhood accessibility to health-related resources: a nationwide comparison between deprived and affluent neighborhoods in Sweden. Health & Place 2010;17:132–9.
[53] Malin B, Benitez K, Masys D. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. Journal of the American Medical Informatics Association 2011;18((1) January):3–10 [Article].
[54] Pearce J, Blakely T, Witten K, Bartie P. Neighborhood deprivation and access to fast-food retailing. A national study. American Journal of Preventive Medicine 2007;32(5):375–82.
[55] Fotheringham AS, Wong DWS. The modifiable areal unit problem in multivariate statistical analysis. Environment and Planning A 1991;23(7):1025–44.
[56] Sabel CE, Kihal W, Bard D, Weber C. Creation of synthetic homogeneous neighbourhoods using zone design algorithms to explore relationships between asthma and deprivation in Strasbourg, France. Social Science & Medicine 2012;91:110–21.
[57] Jones SG, Ashby AJ, Momin SR, Naidoo A. Spatial implications associated with using euclidean distance measurements and geographic centroid imputation in health care research. Health Services Research 2010;45((1) February):316–27 [Article].
[58] Trost SG, Owen N, Bauman AE, Sallis JF, Brown W. Correlates of adults' participation in physical activity: review and update. Medicine & Science in Sports & Exercise 2002;34((12) December):1996–2001.
[59] Goldman N. Social inequalities in health—disentangling the underlying mechanisms. In: Weinstein M, Hermalin AI, Stoto MA, editors. Population health and aging: strengthening the dialogue between epidemiology and demography. New York: Academy of Science; 2001. p. 118–39.
[60] Meliker JR, Slotnick MJ, Avruskin GA, Kaufmann A, Fedewa SA, Goovaerts P, et al. Individual lifetime exposure to inorganic arsenic using a space-time information system. International Archives of Occupational and Environmental Health 2007;80((3) January):184–97.
[61] Sabel CE, Boyle P, Raab G, Löytönen M, Maasilta P. Modelling individual space–time exposure opportunities: a novel approach to unravelling the genetic or environment disease causation debate. Spatial and Spatio-temporal Epidemiology 2009;1:85–94.
[62] Sabel CE, Gatrell AC, Löytönen M, Maasilta P, Jokelainen M. Modelling exposure opportunities: estimating relative risk for motor neurone disease in Finland. Social Science & Medicine 2000;50:1121–37.
[63] Kulldorff M, Information Management Services Inc. SaTScanTM v8.0: Software for the spatial and space-time scan statistics. http://www.satscan.org; 2010.
[64] Anselin L, Syabri I, Kho Y. GeoDa: an introduction to spatial data analysis. Geographical Analysis 2006;38(1):5–22.
[65] Health Insurance Portability and Accountability Act of 1996, Report 104-736 (1996).
[66] Office of Public Sector Information. The Data Protection Act. Available from: http://www.opsi.gov.uk/Acts/acts1998/19980029.htm; 1998 [07.07.09].
[67] Brownstein JS, Cassa CA, Kohane IS, Mandl KD. Reverse geocoding: concerns about patient confidentiality in the display of geospatial health data. In: AMIA Annu. Symp. Proc. 2005. p. 905.
[68] El Emam K, Brown A, AbdelMalik P. Evaluating predictors of geographic area population size cut-offs to manage re-identification risk. Journal of the American Medical Informatics Association 2009;16:256–66.

Please cite this article in press as: Exeter DJ, et al. "Whose data is it anyway?" The implications of putting small area-level health and social data online. Health Policy (2013), http://dx.doi.org/10.1016/j.healthpol.2013.07.012

[69] Rodgers SE, Mather TN. Human Babesia microti incidence and *Ixodes scapularis* distribution, Rhode Island, 1998–2004. Emerging Infectious Diseases 2007;13((4) April):633–5.
[70] Ford D, Jones K, Verplancke J-P, Lyons R, John G, Brown G, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. BMC Health Services Research 2009;9(1):157.
[71] Cassa CA, Wieland SC, Mandl KD. Re-identification of home addresses from spatial locations anonymized by Gaussian skew. International Journal of Health Geographics 2008;7(45).
[72] Boyle PJ. Should we be making better use of public data in health research? The Foundation of Science and Technology; London. Accessed at: http://www.foundation.org.uk/events/pdf/20110608_boyle.pdf; 2011 [08.06.11].
[73] Vogel G. Clinical research, U.K. to open health records as E.U. considers restrictions. Science 2011;334:1483–4, 6062 [News].
[74] Statistics New Zealand. Confidentialised unit record files. Wellington: Statisics New Zealand. Available from: http://www.stats.govt.nz/methods_and_services/~/link.aspx?_id=78BBB257AC734729B8147724176FE9B7&_z=z; 2010 [cited 01.20.12].
[75] Marsh C. The sample of anonymised records. ESRC Data Archive Bulletin 1991;48:3–9.
[76] US Census Bureau. Public Use Microdata Sample (PUMS) files. US Census Bureau. Available from: http://www.census.gov/main/www/pums.html; 2010 [cited 20.01.12].
[77] Holman CD. An end to suppressing public health information. Medical Journal of Australia 2008;188((8) April):435–6.
[78] Holman CD, Bass AJ, Rosman DL, Smith MB, Semmens JB, Glasson EJ, et al. A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. Australian Health Review 2008;32((4) November):766–77.

Please cite this article in press as: Exeter DJ, et al. "Whose data is it anyway?" The implications of putting small area-level health and social data online. Health Policy (2013), http://dx.doi.org/10.1016/j.healthpol.2013.07.012