"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

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.

© 2013 Elsevier Ireland Ltd. All rights reserved.

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.

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,