

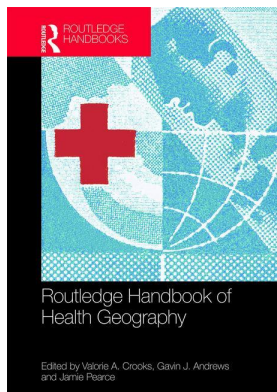
This article was downloaded by: 10.3.97.143

On: 01 Apr 2023

Access details: *subscription number*

Publisher: *Routledge*

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: 5 Howick Place, London SW1P 1WG, UK



## **Routledge Handbook of Health Geography**

Valorie A. Crooks, Gavin J. Andrews, Jamie Pearce

### **Health geography and the future of data**

Publication details

<https://www.routledgehandbooks.com/doi/10.4324/9781315104584-45>

Daniel Lewis

**Published online on: 11 Jun 2018**

**How to cite :-** Daniel Lewis. 11 Jun 2018, *Health geography and the future of data from:* Routledge Handbook of Health Geography Routledge

Accessed on: 01 Apr 2023

<https://www.routledgehandbooks.com/doi/10.4324/9781315104584-45>

**PLEASE SCROLL DOWN FOR DOCUMENT**

Full terms and conditions of use: <https://www.routledgehandbooks.com/legal-notices/terms>

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

# HEALTH GEOGRAPHY AND THE FUTURE OF DATA

*Daniel Lewis*

Health geography is concerned with the collection, analysis and presentation of data that measures or represents the health of individuals or populations in light of geographical factors. A guiding principle of health geography is that *places matter* for health (Kearns and Moon, 2002), the validity of which has been tested in according to a diversity of research practices (Macintyre, Ellaway and Cummins, 2002). In this chapter, the ongoing role of quantitative data in health geography is explored in light of developments sometimes termed the *data revolution* (Kitchin, 2014).

The chapter will first define what quantitative data are, describe how to distinguish between different types of quantitative data, and suggest some questions relevant to working with any dataset. The changing nature of data is then discussed, focusing on examples in health geography that illuminate *big data*, *open* and *volunteered* data and then data *infrastructures* and the value of *indexical* data. This should all be read as an introduction to thinking critically about data.

Health geographers are concerned with social justice (Kearns and Moon, 2002). For quantitative health geographers, evidence for social justice relies on quantitative data about people's personal and environmental circumstances. A solid appreciation of data is fundamental to making meaningful contributions to health geography and social justice.

## **Defining quantitative data: what are data?**

Quantitative health geographers typically approach *data* as the building blocks of analysis. Data are collected, analyzed, and interpreted in the hope of producing meaningful knowledge about the world; this process is known as *enrichment*.

The word "data" comes from the Latin word *dare*, meaning *to give*, the idea being that facts about the world are given. However, Checkland and Holwell (2005) suggest that the real world is made up of an overwhelming mass of facts, or *data*, and so some practical selection must take place. They term data that has been selected *capta*, reflecting that most data are in fact *taken* for a specific purpose, from the Latin *capere*. Almost all data used in quantitative health geography are actually *capta*, reflecting both what it is feasible to measure and the decision of what to include in, and what to leave out of, our models of reality. For practical reasons, we tend to refer to this all as "data."

Quantitative data are numeric records and are usually seen as distinct from qualitative data, which are non-numeric and can take a range of forms: texts, images, audio and video recordings, material objects and so forth. The actual values that quantitative data take fall into two main groups: numerical and categorical,

largely determined by what is actually being quantified. Most data are either *extensive* or *representative* (Kitchin, 2014); extensive data deals with quantities such as height, weight or distance, which relate to the physical reality of something being studied, while representative data encode non-physical or *latent* (Bollen, 2002) characteristics such as occupational, social class or quality of life. Table 45.1 further breaks down numerical and categorical data into sub-types of quantitative data.

In addition to the data types in Table 45.1, quantitative data are often typed based on how they are produced. Conventionally, *primary data* are collected for a specific study or purpose; a researcher either personally collects the data or sets the rules by which data are collected. *Secondary data* are primary data that have been made available to other researchers; these researchers reuse the data to answer their own questions. Sometimes a distinction is made between secondary data and *tertiary data* (Kitchin, 2014); census data and hospital-admissions data are generally released as tertiary data in which counts, rates, or summaries (e.g., means) are derived for aggregate units (such as census output zones, hospitals, clinics) to preserve confidentiality. Research with census secondary data or individualized hospital admissions may require visiting secure data labs.

When considering what quantitative data are, it can be tempting to see all data as just a mass of numbers; however, data require structure. Data about different phenomena are called *variables* (if the values of the data they hold vary; if not, they are called *constants*), which are gathered into datasets for specific research purposes. It is important to know what kind of data (numerical, categorical), each variable in a dataset holds and whether it is primary, secondary or tertiary data; however, we should also be aware of other aspects of the data. Table 45.2 gives an overview of some dimensions to take note of (identified by Kitchin, 2014) and the kinds of questions we could ask to get a better understanding of the *raw materials* of our research.

The answers to some of these questions will often be found in the *metadata*; that is, the data about the data. This will likely cover technical and spatial/temporal aspects and, depending on the type of data, may include references to ethical review. Other aspects, particularly the political/economic and philosophical dimensions of data, may not be easily identified, instead requiring independent critical thought.

There is much to consider, and much taken for granted, on the topic of data. Nonetheless, it is extremely important that researchers have a clear grasp of what data are, particularly in light of recent and ongoing changes to how data are produced.

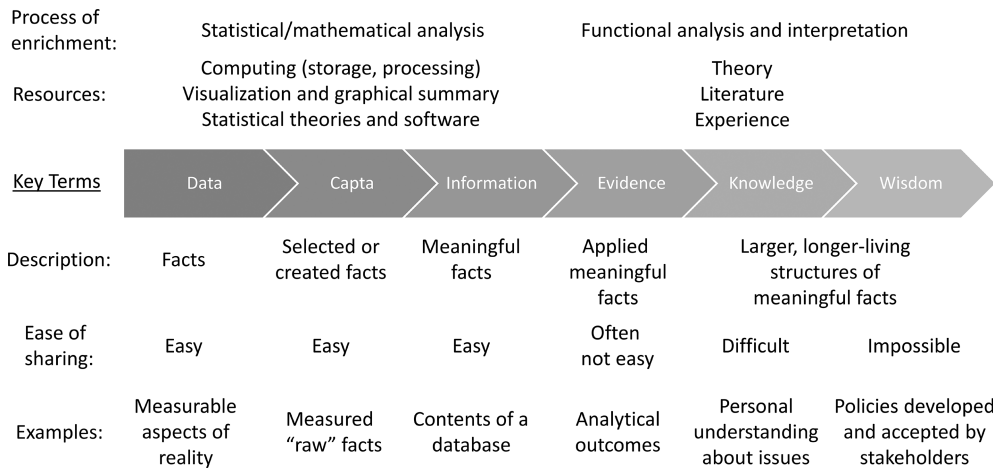


Figure 45.1 Enriching data: a hierarchy of information

Table 45.1 Types of quantitative data, their definitions and examples

<i>Data type – sub-type</i>	<i>Definitions</i>	<i>Examples in health geography</i>
<b>Numerical</b>	Continuous (decimal numbers) or discrete (whole numbers). The size of the value is meaningful, implying more or less of something.	The Townsend index is a measure of material deprivation based on census variables. Areas with high Townsend scores are more deprived relative to areas with low Townsend scores.
Interval	In addition to above, the distance between observations (the interval) is fixed and interpretable. Averages can be computed, but not ratios. Zero values (origins) are arbitrary.	Temperatures (Celsius/Fahrenheit) are interval data. A degree is a fixed unit, making differences between values as well as the size of a value meaningful. Temperature is associated with malaria transmission, affecting its geographic distribution and seasonality.
Ratio	As with interval data, but with true zero values, which indicate an absence of something, so fractions can be computed (e.g., twice as much).	In addition to measures such as weight or distance, counts are also ratio data. The physician-to-population ratio is a simple measure of health-service access. The UK has 2.81 physicians per 1,000 people, more than the US, which has 2.55, and almost twice that of China at 1.49 (WHO, 2017).
<b>Categorical</b>	Data take on a fixed value indicating belonging to a group or category. Numerical data can be grouped according to certain schemes (e.g., quantiles – ordered, evenly sized groups). Categories are often representative, based on non-physical properties.	The Townsend index calculated for small areas can be divided into deciles (quantiles with 10 groups). Average life expectancy can be computed for each decile of deprivation and has been shown to decrease as deprivation increases in what is known as a social gradient.
Binary	Binary data express on-off, true-false, present-absent responses as data, or otherwise dichotomize (split into two categories) data.	In medical geography, whether or not someone has a specific disease at a given point in time is usually represented as a binary variable. A value of 1 indicates they are a case (i.e., they have the disease), while 0 indicates they are not a case.
Nominal	Nominal data is divided into named, discrete categories. There is no express order to the categories, and any given category is not intrinsically better or worse than any other.	Black, white, and Latino are examples of nominal categories expressing race or ethnicity. The welfare geography of who gets what, where, when and why (Smith, 1977) may consider disparities in access to green spaces by racial or ethnic group as evidence of environmental injustice.
Ordinal	Ordinal data are categorical data that can be sorted into an order. The position of a category relative to another is meaningful.	Disease severity is a classic example of ordinal data. In research on health-service utilization, a person's willingness to travel for surgery or care can be patterned according to the severity or stage of a disease. This can have implications for the provision of health care.

Table 45.2 Potential questions when dealing with data

<i>Dimension</i>	<i>Relevant questions</i>
Technical	What instruments were used to collect these data? ◦ How accurate are those instruments? ◦ How precise are those instruments? ◦ What are the uncertainties and/or biases in these data? ◦ Has the validity or reliability of these data been assessed? ◦ Should we be concerned about error? ◦ Were these data subject to any cleaning or post-processing? ◦ Should these data be cleaned or truncated?
Spatial/temporal	Are these data about particular places or times? ◦ Are they subject to particular aggregations: spatial zones or time periods? ◦ Are the data longitudinal? ◦ Are they about particular cohorts of people? ◦ Does the way these data are collected change over time or space? ◦ Have instruments or questions changed over time? ◦ How consistent are these data spatially and/or temporally? ◦ Can these data be compared across space and/or time?
Ethical	How were these data produced? ◦ Was any consent/permission obtained necessary and appropriate? ◦ Are there rules governing how these data are stored and/or shared? ◦ Are there any prohibitions, restrictions or guidance on how these data are to be used? ◦ Is it possible that generating or publishing these data will cause harm or distress? ◦ Are there any privacy or confidentiality issues that need to be addressed? ◦ Are these data anonymous or identifiable? ◦ Have disclosure controls been applied?
Political/ economic	What is the wider context within which these data were generated? ◦ Whom do these data represent? ◦ Are these data routinely collected? ◦ Is there anything controversial or unusual about these data, the way they were defined or the way they were generated? ◦ Whom are these data for? ◦ Are these data freely and openly available? ◦ Who benefits from these data? ◦ Who paid for the collection, creation and/or dissemination of these data, and why?
Philosophical	Why were these data created/collected? ◦ Do these data measure what they are supposed to, or claim to, measure? ◦ Can these data be taken at face value? ◦ How were these data originally conceived of? ◦ Are there omissions in what these data capture? ◦ Have the backgrounds/beliefs of the researchers producing these data influenced what is collected, and how?

### Health geography and the data revolution

The way in which data are collected, stored, shared and analyzed is changing. Kitchin (2014) suggests that there are three main factors underpinning these changes: *big* data, *open* data, and data *infrastructures*.

*Big data* refers to the masses of data continually pulled from the world around us using sensors and information-communication technologies, documenting our actions, interactions and transactions. This data is given the prefix *big* because it differs from conventional *small* data, being high-volume (many observations/measurements), high-velocity (real-time or near real-time collection, continually updating) and high-variety (datasets come in many different formats – e.g., text files, PDFs, web data formats, SMS messages). *Volume*, *velocity* and *variety*, known as the *three Vs* (Sagiroglu and Sinanc, 2013) is one common way to think of big data, but many authors offer expanded or competing approaches (boyd and Crawford, 2012; Kitchin, 2014).

*Open data* are freely available to use, reuse and redistribute, provided you attribute the original source and make any changes or improvements to the data similarly open (Molloy, 2011). The intent is that the data be provided in a convenient form for use and that any monetary cost of data not exceed the reproduction cost.

*Data infrastructure* refers to the means to join everything together to promote the successful storage, sharing, linking, use and reuse of data. Government open data has been made possible by the digital data infrastructures that underpin online portals such as data.gov in the United States ([www.data.gov/](http://www.data.gov/)) and data.gov.uk in the United Kingdom (<https://data.gov.uk/>).

In the next three sections, examples of big data, open data and data infrastructures are given that have a health-geographic focus.

### **Big data: Google Flu Trends and electronic health records**

Unlike familiar *small* data, which is captured for a specific research purpose, big data is often termed *exhaust* data (Kitchin, 2014). “Exhaust” implies that the data were not originally collected for the purpose for which they are being used but are a by-product of some other process, such as finding information on the internet or managing the treatment of patients.

The most well-known big-data experiment in health, Google Flu Trends, aimed to predict the rate of *influenza-like illness* (possible cases of influenza) by sifting through the tens of thousands of searches made on Google every second (Ginsberg et al., 2008). Using secondary data, researchers found the common search queries associated with physician visits for influenza-like illness, which they could then use to estimate present influenza activity. Conventional approaches to estimating influenza activity, such as those used by the US Centers for Disease Control and Prevention (CDC), rely on a number of sophisticated surveillance approaches, including using hospital admissions records, mortality reports, and clinical laboratory testing (CDC, 2016). However, estimates based on these data usually lag behind the actual rate of influenza activity; the CDC publishes estimated flu data with a one-to-two-week latency. Google Flu Trends was intended to *nowcast* influenza activity using big data, effectively predicting influenza outbreaks in near real-time (around a one-day latency) based on search activity. Google Flu Trends worked reasonably well, until suddenly in 2013 it did not, dramatically overestimating the flu activity that flu season. Lazer et al. (2014) discuss the reasons for Google Flu Trends’ well-publicized failure of prediction, citing something they term *big data hubris* – the implicit assumption that big data could supplant conventional data rather than complement it.

The application of big-data approaches in health care has been described as inevitable (Murdoch and Detsky, 2013), partly due to the challenges and opportunities presented by electronic health records (EHRs). Nowadays, health-care providers can store all the data generated by hospital visits digitally as a diverse set of records including patient demographics; diagnosis, treatment and operations data; prescriptions; medical history; test results and scans; hospital administrative data, and so on. Data in EHRs are primarily aimed at improving the coordination of patient care by medical professionals. However, EHRs also offer researchers the chance to improve the health care of individuals and populations. EHRs allow for the creation of large observational datasets that integrate multiple data sources, as well as the potential for experimental datasets where patients have naturally received different treatments *as if at random* or for longitudinal datasets that track patients over time. Analytical uses for EHRs are still emerging, and the EHR is not yet a comprehensive resource in many countries. In the United Kingdom, for instance, Clarke et al. (2017) report limited or fragmented progress with EHRs, while Jha et al. (2009) have reported slow adoption among US hospitals.

### **Open data: health dashboards and volunteered health data**

Typically, *small* data were expensive to produce, and, as a result, highly valuable. Often access to such data was tightly controlled. Pressure for transparency and accountability in government, as well as movements in the technology sector supporting open-source software and open collaboration, provided the conditions for a shift to open data. Open data provide opportunities for people to make informed decisions in their own lives as well as the tools for participating in and critiquing policy-making. Moreover, open data offers the possibility that communities of individuals could produce their own data and volunteer it for the reuse of others.

The diffusion of health data through *dashboards* is a visible product of the data revolution; dashboards are visual displays (statistical measures or graphical representations) that take often dynamic data and report

key indicators or important summaries *at a glance*. Batty (2015) notes that monitoring human systems has long been a feature of modern medicine – for instance, the electrocardiogram (EKG) machine tracing the beating of a patient’s heart is a kind of dashboard. Monitoring human organizations, such as cities, is more recent, requiring not only advances in computing, but also in many cases the advent of open data. Health and health-care dashboards offer the opportunity to transcend the EKG, curating multiscalar data relevant to individuals and their environments into effective summaries. Applications in personalized medicine (aka precision medicine) might involve integrating individual clinical data with smoking-behavior data from sensors in electronic cigarettes and open geospatial and residential neighborhood demographic data. Such an overview may support patients and doctors in choosing between different courses of action for preventative health care or, suitably anonymized, might help local policy-makers contextualize population health decision-making.

Volunteered data, or *crowdsourced* data, are an important part of the open-data picture. In general, crowdsourcing means opening up a task for public contribution and collaboration (Ranard et al., 2014); by the same measure, crowdsourced data is data produced (knowingly or unknowingly) by members of the public. The idea that people choose to volunteer data is particularly prominent in geography (Elwood, Goodchild and Sui, 2012), where the term *volunteered geographic information* (VGI) covers the recent creation of web technologies that allow people to associate data with places, features or locations on Earth using a computer. Volunteered data has become increasingly important in humanitarian crises and disaster response, as in the case of the 2010 Haitian earthquake; Zook et al. (2010) describe how users of OpenStreetMap (a free and open-source web map) from around the world traced aerial photos to provide digital maps to aid workers on the ground. Ranard et al. (2014) also review a number of crowdsourcing initiatives that are rooted in conventional scientific inquiry.

### **Data infrastructure: indexical data for individual, longitudinal and geographical linkage**

“Data infrastructures” refers to the hardware and software systems that make data available, shareable and reusable. They are present in the background of the examples discussed above, facilitating the use of big and open data. However, data infrastructures are also important for data linkage, a specific type of data – *indexical* data – underpinning this process. Linking previously independent data can enhance their usefulness for research. For instance, researchers may wish to link survey data on self-reported minor psychiatric morbidity and demographics with clinical admissions data on actual psychiatric episodes. Indexical data is crucial to achieving this aim. Often, a key outcome of data linkage is the integration of important *attributional* data; attributional data are measures such as those discussed in Table 45.1. Kitchin (2014) suggests a range of possible indexical data, developed in Table 45.3.

Indexical data offers the potential to transform conventional small data through data linkage and substantially improve the usefulness of those data. Indexical data can be used to track an individual over time, as well as to integrate data from diverse sources or devices and locate individuals in their geographical context. However, linkage can be extremely difficult – not only practically, but also legally and ethically. Indexical data may be subject to laws that govern whether, and under what conditions, data linkage can take place.

### **Consolidation: a critical appreciation of data**

Krieger et al. (2015) argue that when people are killed by law enforcement it is a public-health as well as a criminal-justice issue and, further, that public responses such as protests or violence can have public-health consequences. Although killings of law-enforcement officials are well documented, similar data do not exist



Table 45.3 Types of indexical data

<i>Indexical data</i>	<i>Description</i>	<i>Health-geography use case</i>
<i>Personal</i> Name, date of birth, date of death, sex, ethnicity, place of birth/death.	Personal data is the basis for distinguishing between individuals.	Health-inequalities research makes extensive use of these data, often not indexically, but attributionally to define population subgroups of analytical interest.
<i>Administrative</i> Passport, social security, national identity card, national insurance, immigration documents, tax records, school/college/university enrollments, public health records.	Governments hold numerous unique identifiers covering both citizens and non-citizens to manage state services and systems.	Increasingly used by health researchers interested in integrating health data with other data. For instance, linking medical records to more general survey data using administrative identifiers.
<i>Geographical</i> Address, postal/zip code, district, polling area/constituency etc.	Addresses may allow linkage of individuals across datasets, with support from other identifying data (name, age, sex).	Commonly used by health geographers to link individuals or households to aggregate data about their spatial context or to measure proximity to resources (e.g., hospitals, green spaces).
<i>Commercial</i> Accounts and memberships, loyalty schemes/cards, financial services (bank accounts, credit cards etc.), social media/networking accounts.	Companies can build up substantial personal data based on transactions both online and offline. Accounts and memberships act as possible points of linkage for these data.	Limited current research in health geography, but considerable interest. Topics include food purchase through supermarket loyalty cards, ideally linked to attribute-rich survey data.
<i>Electronic</i> Email address, cellular phone, device ID (e.g., MAC (media access control) address, IP (internet protocol) address).	Electronic devices that connect ubiquitously to the internet broadcast unique identity numbers, which can be used to link activity on these devices to their owners.	Research in health geography very limited, but scope for applications in mobile health, and with the tech industry, such as the Apple health app, which tracks users' health.
<i>Biomedical</i> Fingerprint, DNA sequence, retina, voice, biomarkers.	Fingerprints and retinal scans are common access controls that uniquely identify individuals. Law enforcement and immigration commonly hold these data. DNA and other identifiable biological data are increasing a part of biomedical research data and could be used for linkage.	Health geography is currently more interested in the attributional value of these kind of data for population health than as a form of linkage.

on those people who were killed by the actions of law enforcement. Krieger et al. (2015) explain that these data must be gathered from multiple sources and combined. As attempts to obtain these data directly from law enforcement have hitherto proven unsuccessful, Krieger et al. (2015) suggest that deaths resulting from the actions of law enforcement should be made *notifiable* – a distinct cause of death, a *health condition*, in public-health mortality reporting.



What we think of as data are changing and, as a result, so are the conditions for health–geographic research. In the future, data will come from a diverse range of sources, in a diverse range of formats and at high speed. There will also be a lot more data, enabling health geographers to explore new questions and requiring new approaches to storing, analyzing and interpreting these *raw materials*. It is important, therefore, that we be clear on what these raw materials are: observations taken for a reason, conventionally for the direct purpose of learning something about the world, but increasingly as a by-product, an *exhaust*. Big data are being captured by machines, sensors, and proprietary algorithms whose primary intent is not research, and increasingly private interests are compiling and sharing data without disclosing their motivations for doing so.

Being critical about data and thinking carefully about what a given data source captures is as important as ever. Data are not strictly *neutral*; they reflect the context in which they were produced. Health geographers need a secure grasp of the changing nature of data if they are to continue to produce work that is timely and relevant.

## References

- Batty, M. (2015). A perspective on city dashboards. *Regional Studies, Regional Science*, 2(1), pp. 29–32.
- Bollen, K. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, 53, pp. 605–634.
- boyd, d. and Crawford, K. (2012). Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication and Society*, 15(5), pp. 662–679.
- Centers for Disease Control and Prevention (CDC). (2016). *Overview of influenza surveillance in the United States*. [online] Available at: [www.cdc.gov/flu/weekly/overview.htm](http://www.cdc.gov/flu/weekly/overview.htm) [Accessed 6 Aug. 2017].
- Checkland, P. and Holwell, S. (2005). Data, capta, information and knowledge. In: M. Hinton, (ed.), *Introducing information management: the business approach*. London: Routledge, pp. 47–55.
- Clarke, A., Watt, I., Sheard, L., Wright, J. and Adamson, J. (2017). Implementing electronic records in NHS secondary care organization in England: policy and progress since 1998. *British Medical Bulletin*, 121(1), pp. 95–106.
- Elwood, S., Goodchild, M. and Sui, D. (2012). Researching volunteered geographic information: spatial data, geographic research, and new social practice. *Annals of the Association of American Geographers*, 102(3), pp. 571–590.
- Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M. and Brilliant, L. (2008). Detecting influenza epidemics using search engine query data. *Nature*, 457, pp. 1012–1014.
- Jha, A., DesRoches, C., Campbell, E., Donelan, K., Rao, S., Ferris, T., Shields, A., Rosenbaum, S. and Blumenthal, D. (2009). Use of electronic health records in U.S. hospitals. *The New England Journal of Medicine*, 360, pp. 1628–1638.
- Kearns, R. and Moon, G. (2002). From medical to health geography: novelty, place and theory after a decade of change. *Progress in Human Geography*, 26(5), pp. 605–625.
- Kitchin, R. (2014). *The data revolution: big data, open data, data infrastructures and their consequences*. London: Sage.
- Krieger, N., Chen, J., Waterman, P., Kiang, M. and Feldman, J. (2015). Police killings and police deaths are public health data and can be counted. *PLoS Medicine*, 12(12), e1001915.
- Lazer, D., Kennedy, R., King, G. and Vespignani, A. (2014). The parable of Google Flu: traps in big data. *Science*, 343(6176), pp. 1203–1205.
- Macintyre, S., Ellaway, A. and Cummins, S. (2002). Place effects on health: how can we conceptualise, operationalise and measure them? *Social Science & Medicine*, 55(1), pp. 125–139.
- Molloy, J. (2011). The Open Knowledge Foundation: open data means better science. *PLoS Biology*, 9(12), e1001195.
- Murdoch, T. and Detsky, A. (2013). The inevitable application of big data to health care. *Journal of the American Medical Association*, 309(13), pp. 1351–1352.
- Ranard, B., Ha, Y., Meisel, Z., Asch, D., Hill, S., Becker, L., Seymour, A. and Merchant, R. (2014). Crowdsourcing – harnessing the masses to advance health and medicine, a systematic review. *Journal of General Internal Medicine*, 29(1), pp. 187–203.
- Sagiroglu, S. and Sinanc, D. (2013). *Big data: a review*. International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, pp. 42–47.
- Smith, D. (1977). *Human geography: a welfare approach*. London: Edward Arnold.
- World Health Organization. (2017). *Global Health Observatory (GHO) data*. [online]. Available at: [www.who.int/gho/health\\_workforce/physicians\\_density/en/](http://www.who.int/gho/health_workforce/physicians_density/en/) [Accessed 1 Sept. 2017].
- Zook, M., Graham, M., Shelton, T. and Gorman, S. (2010). Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake. *World Medical and Health Policy*, 2(2), pp. 7–33.