Oxford Research Encyclopedia of Environmental Science

Big Data in Environment and Human Health

Lora Fleming, Anthony Kessel, Virginia Murray, Michael Depledge, Sabina Leonelli, Niccolò Tempini, Harriet Gordon-Brown, Gordon Nichols, Christophe Sarran, Paolo Vineis, Giovanni Leonardi, Brian Golding, and Andy Haines

Subject: Environment and Human Health, Quantitative Analysis and Tools

Online Publication Date: Jul 2017 DOI: 10.1093/acrefore/9780199389414.013.541

Summary and Keywords

Big data refers to large, complex, potentially linkable data from diverse sources, ranging from the genome and social media, to individual health information and the contributions of citizen science monitoring, to large-scale long-term oceanographic and climate modeling and its processing in innovative and integrated "data mashups." Over the past few decades, thanks to the rapid expansion of computer technology, there has been a growing appreciation for the potential of big data in environment and human health research.

The promise of big data mashups in environment and human health includes the ability to truly explore and understand the "wicked environment and health problems" of the 21st century, from tracking the global spread of the Zika and Ebola virus epidemics to modeling future climate change impacts and adaptation at the city or national level. Other opportunities include the possibility of identifying environment and health hot spots (i.e., locations where people and/or places are at particular risk), where innovative interventions can be designed and evaluated to prevent or adapt to climate and other environmental change over the long term with potential (co-) benefits for health; and of locating and filling gaps in existing knowledge of relevant linkages between environmental change and human health. There is the potential for the increasing control of personal data (both access to and generation of these data), benefits to health and the environment (e.g., from smart homes and cities), and opportunities to contribute via citizen science research and share information locally and globally.

At the same time, there are challenges inherent with big data and data mashups, particularly in the environment and human health arena. Environment and health represent very diverse scientific areas with different research cultures, ethos, languages, and expertise. Equally diverse are the types of data involved (including time and spatial scales, and different types of modeled data), often with no standardization of the data to allow easy linkage beyond time and space variables, as data types are mostly shaped by the needs of the communities where they originated and have been used. Furthermore, these "secondary data" (i.e., data re-used in research) are often not even originated for this purpose, a particularly relevant distinction in the context of routine health data re-use. And the ways in which the research communities in health and environmental sciences approach data analysis and synthesis, as well as statistical and mathematical modeling, are widely different.

There is a lack of trained personnel who can span these interdisciplinary divides or who have the necessary expertise in the techniques that make adequate bridging possible, such as software development, big data management and storage, and data analyses. Moreover, health data have unique challenges due to the need to maintain confidentiality and data privacy for the individuals or groups being studied, to evaluate the implications of shared information for the communities affected by research and big data, and to resolve the long-standing issues of intellectual property and data ownership occurring throughout the environment and health fields. As with other areas of big data, the new "digital data divide" is growing, where some researchers and research groups, or corporations and governments, have the access to data and computing resources while

others do not, even as citizen participation in research initiatives is increasing. Finally with the exception of some business-related activities, funding, especially with the aim of encouraging the sustainability and accessibility of big data resources (from personnel to hardware), is currently inadequate; there is widespread disagreement over what business models can support long-term maintenance of data infrastructures, and those that exist now are often unable to deal with the complexity and resource-intensive nature of maintaining and updating these tools.

Nevertheless, researchers, policy makers, funders, governments, the media, and members of the general public are increasingly recognizing the innovation and creativity potential of big data in environment and health and many other areas. This can be seen in how the relatively new and powerful movement of Open Data is being crystalized into science policy and funding guidelines. Some of the challenges and opportunities, as well as some salient examples, of the potential of big data and big data mashup applications to environment and human health research are discussed.

Keywords: digital divide, digital access divide, open data, data mashups, data access, data infrastructure, data reuse

Introduction and Definitions

There have been many definitions of *big data* since the first use of this term by NASA scientists in 1997 (Press, 2013). The early emphasis was on the “3 Vs”: *volume* (increasingly large size of the data), *variety* (the diversity of types of data, e.g., from free text to remote sensing), and *velocity* (the rapid generation and flow of data). Normandeau (2013) and others have added the growing importance of *veracity* (issues of data bias and cleanliness), *validity* (appropriate data for the intended use), and *volatility* (how long the data are valid, how long need to store) (Khoury & Ioannidis, 2014). Ward and Barker (2013) summarize their review of the big data definition literature review with: “big data is a term describing the storage and analysis of large and or complex data sets using a series of techniques,” including machine learning; while Mayer-Schönberger and Cukier (2013) suggest that essential to big data is: “The ability of society to harness information in novel ways to produce useful insights or goods and services of significant value” (p. 2) and the “. . . things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value” (p. 6).

While these definitions are relative to the specific point in time in which they are formulated, two aspects of Big Data are common to these and the many other definitions (see datascience@berkeley), and could thus be regarded as fundamental to the identification and understanding of this term. The first is the need for such data to be stored and analyzed with tools other than an Excel spreadsheet or other mainstream reporting and analysis applications. The other is the emphasis on the importance of narratives and interpretation in relation to big data. There is widespread agreement that simply being able to assemble and access large datasets is not interesting in itself, but rather a prelude to the opportunity to use such data as evidence for knowledge claims.

As an outgrowth of big data, the concept of *data mashups* (mashup being a term originating from jazz) is increasingly used. Data mashups are described as a dynamic, explorative, and ongoing exercise of processing, mixing, and analyzing different types of data together to produce a unified and unique output that can be potentially more useful than and accessed independently of the original individual datasets (Daniel & Matera, 2014). They thus highlight both the variety of formats and sources characterizing big data and the significance of data processing and analysis in bringing such data together harmoniously, so as to create something new.

Some of the important potential opportunities and challenges of using big data and data mashups in environment and human health are discussed and explored. Although many of these factors are generic to all uses of big data, there are unique aspects revealed within the field of environment and human health.

Possibilities and Opportunities

At the core of the field of environment and human health is the belief that the interactions between humans and ecosystems matter to the health of both humans and the environment in the short and long term. The field of environment and human health has been transformed in the last decade in large part due to the growing appreciation of the magnitude of anthropogenic change (climate and other environmental change), the impacts on the health of both local and planetary natural systems, and in particular, the potential for these impacts to effect human health and wellbeing both positively and negatively. After years of primarily anthropocentric views of public health in which the environment was primarily limited to the "built environment," this transformation has been expressed in the new visions of "Ecological Public Health" (Lang & Rayner, 2012) and more recently "Planetary Health" (Whitmee et al., 2015). This realization has developed in parallel with the environmental science community by the development and expansion of the concept of ecosystem services aligned with human health and wellbeing (Corvalan, Hales, & McMichael, 2005), and more recently the concept that humans are exceeding the "Planetary Carrying Capacity" of these natural systems at local and planetary levels (Rockström, Sachs, Öhman, & Schmidt-Traub, 2013).

With this increasing appreciation of the complex interactions between natural systems and human health and their expansive planetary temporal and spatial scales, the new and growing ability to potentially address the big questions of environment and human health is the major promise of big data and their mashups (Fleming et al., 2014). These data mashups of diverse environment and health data allow the exploration of and the ability to ask, new and different questions about these interactions, with a better understanding and appreciation of their complexity (McMichael & Haines, 1997). In addition, there is the opportunity (and really, the need) to take advantage of the new and diverse data obtained at different scales and on different aspects of society and the environment. Examples range from:

- Social media (see for instance, Patients Like Me, developing the "Open Research Exchange [ORE]" platform to put patients at the center of the clinical research process, and helps medical researchers pilot, deploy, share, and validate new ways to measure diseases) (Tempini, 2015; Wicks et al., 2014).
- Data collected through citizen science and self-tracking initiatives; analysis of the human genome and a more individual personalized medicine approach in preventing and treating disease (Del Savio et al., 2016).
- The rapid growth of planetary scale models to analyze past events and produce forecasts of how future climate and other environmental change may impact on human health and ecosystems. (Canali, 2016; Fleming et al., 2014; Panahiazar, Taslimitehrani, Jadhav, & Pathak, 2014; Velasco, Agheneza, Denecke, Kirchner, & Eckmanns, 2014)

What are some of the other potential uses of big data in environment and human health if these data were made more easily available to researchers, policy makers and the general public? Some of the potential uses include:

- Rapidly identify hot spots (where populations and/or ecosystems are particularly vulnerable to environmental changes) for targeted prevention, interventions, and research.
- Provide healthcare practitioners and public health planners with relevant information for improving services for locations and populations identified as being at risk.
- Establish surveillance and follow trends in the interactions between different ecosystems and populations over time.
- Provide early warning systems to prevent and anticipate environmental impacts on health and wellbeing.
- Initiate and evaluate natural experiments and even formal interventions in different places and among different ecosystems and populations of humans and other animals.
- Provide engagement, information, and communication through citizen science, scenario building, and shared decision making with the research community, policymakers, and civil society (Boyd & Crawford, 2012; Fleming et al., 2014; Granella, Belmonte Fernández, & Díaz, 2014; Haines, McMichael, & Epstein, 1993; Velasco et al., 2014).

In addition, such approaches can be important in hypothesis generation as well as analytical hypothesis testing.

As an example, the potential for the examination of environmental determinants in infectious diseases has been explored over a number of years, driven in part by a need to demonstrate potential human health impacts from climate change (see Box 1). The establishment, in 2005, of the European Centre for Disease Prevention and Control (ECDC) has established a means for collecting standardized (although heterogeneous) data on human infectious diseases across European Union Member States and non-EU countries. A European Environmental Epidemiology network (E3) has stimulated interactions between research groups in Europe, but funding has been limited. A review of environmental data resources for use with human disease datasets has established the diversity and potential of data sources (Nichols, Andersson, Lindgren, Devaux, & Semenza, 2014).

Box 1. Big Data and Infectious Disease Epidemiology

Gordon Nichols

Bacteria are diverse and have an ancestry that extends far back before the start of eukaryotic life. Carl Woese in the 1970s used sequencing of the 16S ribosomal DNA to clarify bacterial phylogenetics and separate archaebacteria from eubacteria (Woese & Fox, 1977; Woese, Kandler, & Wheelis, 1990). The enormous advances in the cost of DNA sequencing in recent years have allowed schemes to type isolates based on a multi-locus approach, where isolates are compared using seven or ten chosen genes (Dingle, Colles, Wareing, Ure, Fox, & Bolton, 2001), to be extended to approaches that use all genes for the 53 ribosomal proteins that are present in most bacteria (rMLST) (Jolley & Maiden, 2012) to schemes based on the whole genome or core genome (Sheppard, Jolley, & Maiden, 2012).

Such approaches take advantage of the highly conserved nature of the active sites of genes and the occurrence of areas that are under less selective pressure, and are therefore more variable. Whole genome sequencing is now commonly used across the world for following the indigenous spread and outbreaks of bacterial infections.

The source of bacterial pathogen, *Salmonella* isolates can commonly be determined through comparison with known isolates within England and Wales, or across Europe, by isolating the organism from a food, or by identifying common foods that the affected people with the same strain have eaten. Using a series of algorithms, the sequences from individual isolates are compiled into whole genome types and then single nucleotide polymorphism (snp) types that are compared to known isolates (Ashton et al., 2016). These can then be compared to other patient isolates, to animal, and to environmental isolates. A different approach, called source attribution, uses the sequence of isolates from animal strains to estimate the percentage of human infections (for example, the bacterial pathogen, *Campylobacter*) originating from different host species (Sheppard et al., 2010).

The international datasets from all this activity for all bacterial pathogens will build up hugely over the coming years and transform our understanding of infectious diseases, incorporating antibiotic resistance, pathogenicity factors, and the mechanisms of infectious disease emergence. Further extension of the utility and dynamism of this approach requires open access to large bacterial datasets of annotated genomes, as well as environmental exposure and human health data (Jolley & Maiden, 2010; Maiden & Harrison, 2016).

These advances are possible due to the rapid development of a variety of factors in the computer and information sciences, including:

- The ability to create, collect, and store diverse types of data.
- The increasing rapidity of the linking, processing, analysis, and synthesis of these different types of data in complex data mashups.

- The growth of different ways of understanding and communicating big data through visualization and other techniques (Boyd & Crawford, 2012; Brownstein, Freifeld, & Madoff, 2009; Daniel & Matera, 2014; Hey, Tansley, & Tolle, 2009; Mayer-Schönberger & Cukier, 2013).

Challenges

In addition to the many opportunities presented by big data and big data mashups in environment and human health and in many other fields, there are an increasing number of challenges to be considered and, where possible, addressed, to truly take advantage of these opportunities now and in the future (Boyd & Crawford, 2012; IEAG, 2014; Kitchin, 2014; Leonelli, 2014; Mayer-Schönberger & Cukier, 2013).

Taking a wide viewpoint, there are many different disciplines and communities potentially involved in big data and big data mashups in environment and human health, each with their own inherent cultures, traditions, and visions of how these data should be collected, stored, analyzed, and communicated. Furthermore, these are groups that have not necessarily worked together in the past, so even a shared language can be lacking. As a very small example, in the climate change community, *mitigation* refers solely to those activities that decrease emissions of greenhouse gases (and short lived climate pollutants), while in the public health and environmental pollution communities *mitigation* refers to attempts to lessen the impacts on human health and wellbeing from an exposure (e.g., containing a leaking toxic waste facility, evaluating potentially exposed people, etc.; Lang & Rayner, 2012). Nevertheless, this very diversity holds the promise of new and enriched uses of big data deriving from the exchange of data, personnel, resources, and approaches between the different disciplines and communities interested in the environment and human health.

At the *micro* level, there are a myriad of different types and scales of data being collected, as noted above, from social media to remote sensing (Jorm, 2015; Panahiazar et al., 2014; Velasco et al., 2014). Furthermore, much of the data re-used as *secondary data* for research analyses were not necessarily originally collected for this purpose. Although the linkage between environment and human health data is usually possible through space and time elements (e.g., geocode), there remains the challenge of different geographic and temporal scales, with human data often very densely collected for relatively few individuals at set intervals over a few years compared with very intense collection over a wide geographic area over many years for environmental data (e.g., visits to the GP vs. hourly rainfall data; Fleming et al., 2014).

There is also a difference in exposure data that are specific to the location (as seen with environmental data) compared to the exposure data specific to an individual (as with health data). In addition to the data quality issues already mentioned (Khoury & Ioannidis, 2014; Normandeau, 2013), there is a lack of standardization and coordination of research tools and practices concerning environment and health data (Kessel & Combs, 2016). New efforts by international groups such as the Open Geospatial Consortium (OGC) (see Health DWG for OGC human health data standardization, and ELIXIR for biomedical data) are cause for hope in the future.

These communities also have different traditions, capabilities, and resources for the processing, storage/archiving, curation, and management of these big data, which are fostered and stewarded by different institutions and funders, and which arch back to the diverse backgrounds, interests, and goals of the researchers and the communities in question. Original approaches, such as the use of social media for the engagement of wider publics are not immune, but they are exposed to the same issues as a result of the fluidity of the underlying social arrangements (Kallinikos &Tempini, 2014; Tempini, 2015).

Due to the several post-war International Council for Science (ICSU) projects encouraging data exchange (e.g., International Geological Year, International Polar Year, etc.), and the fact that much of the research in this area is carried out by academics funded by governments, the environmental sciences communities such as the oceanographic and atmospheric sciences have the tradition and experience of many decades of internationally shared data stored and curated with access to super computers and data management infrastructure (Hampton et al., 2013; Maeda &Arévalo Torres, 2012,). A similarly strong culture of data sharing and common infrastructure has arguably characterized the health genomics community, at least since the ratification of the Bermuda Rules in 2009 (Contreras, 2011; Ossorio, 2011; Wu et al., 2009).

Within the broader health community, big data are available for research, but often are increasingly difficult to access due to privacy concerns, ownership, and/or high costs. There are ongoing efforts to create joined up human health big data, particularly associated with environmental data. These include: the U.S. population-based datasets of the National Health and Nutrition Examination Survey (NHANES); large Pan European Union (EU)-sponsored studies (e.g., LIFEPATH and EXPOsOMICS; see Box 2); and the U.K. Biobank, sponsored by Wellcome and the U.K. Research Councils. However with respect to big data resources and the sharing of these data, the health community in general is highly fragmented. Health-related research is also becoming more privatized, and increasingly sponsored by competing companies. Furthermore, the health community does not have a strong tradition of, or even easy access to shared resources, nor to either health or environmental data. Thus, there are important ethical and equity issues, as well as major challenges of ownership and the perception of the high financial value of health data (Boyd & Crawford, 2012; Gliklich, Dreyer, & Leavy, 2014).

Box 2. EXPOsOMICS

Paolo Vineis

Although information on both environmental and genetic causes of disease is growing as a result of large-scale epidemiological research, environmental exposure data (including diet, lifestyle, environmental, and occupational factors) are often fragmentary (in time and depth), non-standardized, at crude resolution, and often do not include estimates at the level of the individual. As a result, important associations can go undetected.

This limitation has recently been framed within the context of the *exposome*, the environmental counterpart of the genome. The concept of the *exposome* refers to the totality of environmental exposures from conception onwards.

By comprehensively addressing the integration of the external and the internal exposomes at the individual level, EXPOsOMICS provides a holistic and consolidated approach to exposure science (Vineis et al., 2016). Building upon several research projects funded by the European Union, with rich sets of health data, exposure data, biomarker measurements, and publicly available data sources, this multidisciplinary project pools and integrates information from short-term, experimental human studies and long-term epidemiological cohorts/consortia (including adults, children, and newborns) to enable focused investigations to refine environmental exposure assessment based on the concept of life-course epidemiology.

The exposome is characterized by employing novel tools and drawing on experience gained in existing EU initiatives (personal exposure monitoring, databases coupled with geographic information systems (GIS), remote sensing), with a focus on air and water pollution; and measuring biomarkers of the internal exposome (xenobiotics and metabolites), using omic technologies (adductome, metabolome, transcriptome, epigenome, proteome) (Pecina-Slaus & Pecina, 2015). Overall data are integrated for almost 3,000 subjects with an extremely complex network of external and internal measurements. Novel statistical methods are developed to capture and understand this multi-level complexity.

Environment and human health communities also approach the synthesis, analysis, and interpretation of data differently. This is due in part to the types of data available, but also to the availability of resources and the scale of the analyses. For example, much epidemiologic data have been analyzed using regression approaches that assume a fairly simplistic inclusion of a limited number of exposure factors for a relatively small group of individuals focusing on often a single health outcome. This approach is not adequate to addressing the complexity of the impacts on human health of climate and other environmental change over long time periods and large diverse areas and populations. At the same time, climate prediction modeling involves complex models that can include a plethora of environmental factors using paleontological data to the present day and requiring super-computing to analyze.

These communities have much to learn from each other in terms of their respective approaches to analysis. Some reciprocal learning and helpful cross-contamination is now beginning to take place. For example, there are increasing applications of Bayesian statistics across the environment and human health sciences, and it is becoming more common to combine statistics and geographic information systems (GIS) to explore health and environment interactions (Bingenheimer & Raudenbush, 2004; Brownson et al., 2015; Granella et al., 2014; Kamel Boulos & Al-Shorbaji, 2014; Morrison, Marcot, & Mannan, 2012).

Another difference between these groups is the interpretation of the data mashups and analyses, and what constitutes, or does not constitute, causal inference. Different disciplines (and even different groups within the same discipline) can have diverging perceptions about the strength of correlation with unproven cause-effect relationships, associations or patterns found in the data, as well as the reliability of given data types. This means that different groups may justifiably interpret the strength and reliability of the causal claims extracted from the same datasets very differently. For example, genome-wide association study (GWAS) results would be taken by some clinicians as a fully reliable causal claim, while many biologists and epidemiologists would be skeptical and treat the same result as an unproven correlation to be tested experimentally and through other data sources (Bradford Hill, 1965; Leonelli, 2012).

Increasingly in human health, randomized intervention or clinical trial is seen as the *sine qua non* of proof or disproof of causality. On the other hand, most environmental sciences (and still much of health research) must rely on the repeated analysis of laboratory (with the challenge in understanding how the laboratory results relate to what is seen in the real world) and "natural experiments," and the creation of models based on collected data with follow-up on subsequent forecasts to prove causality. There are benefits and challenges from both approaches. In the case of the environmental sciences (including when applied to human health), logistical, ethical, and other limitations in the ability to apply experimental approaches to their research now or in the future.

Another challenge across the field of environment and health is the sustained inter/cross-disciplinary training and identification of skilled personnel, particularly those who can work (and communicate) across the range of disciplines and communities with these diverse types of big data (Eynon, 2013; Jorm, 2015; Yewell, 2015). Expertise in database management and linkage, software development, and/or statistics and modeling with understanding and interest in environment and human health are essential. Both disciplines tend to be weak in estimating the need for data management expertise, resulting in many projects producing rich datasets without a precise strategy for whether, how, when, and to whom to make them accessible. The issue tends to be further compounded in projects that link together both disciplines, as further complexities arise that are linked to the conditions of interdisciplinarity considered above (Edwards, Mayernik, Batcheller, Bowker, & Borgman, 2011).

Furthermore, personnel with the abilities and skills to engage and to seek out the questions by engaging stakeholders to ensure that they are relevant to policy and practice; and then to interpret and communicate the results with a range of stakeholders (from other researchers to individual human subjects and community groups to policy makers) are essential if the promise of big data mashups in environment and human health are to be realized (Boyd & Crawford, 2012). Again sharing of these personnel and their skills (including software development and communication techniques) hold the promise of bringing these diverse communities closer together in collaborative efforts to collect, analyze, and understand big data and the big questions in environment and health.

Associated with the issues of training and expertise (as well as infrastructure), is the need for long-term, sustainable funding of big data resources (Bastow & Leonelli, 2010). Given the short-term nature of funding cycles, many research projects in big data, environment, and human health are typically set up at most for up to five years, with no possibility to extend funding further so as to maintain and update the datasets and related infrastructures that have been produced. When the funding ends, access to data deteriorates and is sometimes lost entirely, leading to a loss of knowledge. This is a conservative culture favoring the repetition of prior projects, and a lack of thorough long term evaluation of impact (Boyd & Crawford, 2012; Hall, Fottrell, Wilkinson, & Byass, 2014).

As already alluded to, a major challenge for human health data, which is not present to the same extent and in the same forms for environmental data, lies in the requirement to protect data confidentiality and privacy at the individual and even group levels (Aicardi, Del Savio, Dove, Lucivero, Tempini, & Prainsack, 2016; Tempini, 2016; Tene & Polonetsky, 2012). Data confidentiality also involves issues of ownership, governance, and ethics (Boyd & Crawford, 2012; Gliklich et al., 2014; Jütting, 2016; Mayer-Schönberger & Cukier, 2013; Nuffield Council on Bioethics, 2015; Ossorio, 2011). These complex issues are emerging as a result of the exploding variety of data that certain projects make available (Aicardi et al., 2016) and the need for comprehensive information security strategies in the face of continuously changing risks and requirements (Burton et al., 2015; Ford et al., 2009; Tempini, 2016). Again, this can represent an organizational and cultural challenge and potential inequity between the environmental and human health communities. The environmental scientists are expected to freely give up their data to the human health researchers, but not vice versa, in the name of protecting the confidentiality of individual human subjects in the human health databases (Fleming et al., 2014; Maeda & Arévalo Torres, 2012). Increasingly, the ownership and governance of the data, particularly human health data with its requirements for long-term data protection and confidentiality, are adding to the expense and decreasing the ability of researchers and others (even those individuals and groups who participated in generating the data) to gain access to and use these data (Boyd & Crawford, 2012; IEAG, 2014; Roche, Jennions, & Binning, 2013).

The issue of big data access is now compounding the “digital divide” locally and internationally with a new “data access divide.” There are also issues of access to the appropriate technology, resources, and trained personnel. There is the expense and the myriad challenges involved in data management and dissemination. Furthermore the increasing perception already discussed of the value of (particularly) health data, the resources required to collect, store, and analyze big data, are increasingly being appropriated and developed by a few multi-national corporations and governments (Boyd & Crawford, 2012). Increasingly, there is very little opportunity for less powerful and internationally recognized players (e.g., researchers and citizen scientists in low-resource environments) to participate in shaping the relevant technologies and strategies (Bezuidenhout, Rappert, Kelly, & Leonelli, 2017; Boyd & Crawford, 2012).

This divide in who has and has not the capacity to become involved in the online sharing and re-use of big data results in an alarming lack of essential data relating to certain subgroups and geographical locations (Gibin, Mateos, Petersen, & Atkinson, 2009; IEAG, 2014). This, in turn, indicates that big data collections assembled in health and environmental sciences may be far from comprehensive. This strongly restricts the potential for big data use and their applications to our global environment and human health challenges. Moreover, as in the case of the digital divide, the opportunity to access, share, and use big data collections and infrastructures is not equitably distributed within societies or across nations. This further excludes already marginalized groups from potential future benefits and opportunities as well as decision making processes. Ironically, this is happening at a time when a wide diversity of research-relevant data are being collected, shared, and consulted by a widening portion of the population around the world. Examples range from cellphones that deliver weather data to farmers in developing nations to the successful crowdsourcing of monitoring environmental data by citizen scientists globally (IEAG, 2014).

Horizon Scanning

Looking forward into the future of big data in environment and human health, where are the new challenges and opportunities? There is much research needed into the process of the increasing number of big data mashups in environment and human health (see Box 3). For example, what are the factors that increase and incentivize the activities of integrating and sharing of knowledge and collaborations, and that lead to innovations in this area?

Box 3. Medical and Environmental Data Mash-up Infrastructure (MEDMI)

Niccolò Tempini

The MEDMI (Medical and Environmental Data-a Mash-up Infrastructure) Partnership brings together leading U.K. organizations and researchers in climate, weather, environment, and human health (University of Exeter, U.K. MET Office, Public Health England; and the London School of Hygiene and Tropical Medicine).

The aim is to investigate the links between environmental factors and public health. MEDMI hosts weather, environmental, and human health data, provided or sourced by different partner institutions, that can be accessed remotely (after information governance clearance) and linked through a library of scripting tools, developed to flexibly support a wide range of research needs. It also experiments with the opportunity to develop integrated, in-browser analysis tools, with a view of extending the secure exploration of the data to larger and not exclusively professional audiences. As part of the project, a number of demo and pilot projects were launched to explore aspects of tool development, or demonstrative research cases (Fleming et al., 2014).

Lacking strong antecedents, the project has pioneered this space and has accumulated a great amount of experience. It has also explored first-hand some of the challenges of interdisciplinary environment-health data integration and research, which have included the complexities of multi-partner coordination and project management; interdisciplinary, multi-stakeholder method and tool development; and data availability (including licensing) vs. infrastructure development interdependencies.

There is a need for:

- More qualitative research on how to accelerate the knowledge-to-diffusion cycle.
- A deeper understanding of the types of resources (including personnel) and training that will support these activities.
- A realistic estimate of the essential infrastructure, true costs, and other factors needed for sustained collection, storage, analysis, and sharing of big data around the major issues of environment and human health.

The mixture and use of different types of data to serve a variety of different research purposes (e.g., social media as a public health surveillance and even early warning tool, to predict forest fires or Ebola outbreaks before traditional public health institutions even know) holds great promise. However, significant resources need to be invested to truly evaluate the promise and limitations of these activities on different scales, locations, and communities (Hall et al., 2014). Some early examples provide some potential caveats about the uses of these data and the potential traps for researchers in big data analysis. These include: the substantial over-estimate by Google of potential influenza primary care visits; difficulties accessing corporate owned social media data; issues around sampling and the extent to which parts of relevant populations are excluded by high-tech data collection tools; and, due to the rapidly changing social media platforms and data (sometimes due to intentional corporate manipulation), issues around the reproducibility of findings and even accurate interpretation by researchers (Lazar, Kennedy, King, & Vespignani, 2014).

In particular, the formal objective evaluation of high profile exemplars (e.g., the internet of things, smart cities, soft GIS) taking into account both the predicted and the unpredicted outcomes, positive and negative, is essential (Hall et al., 2014; Kamel Boulos & Al-Shorbaji, 2014; Townsend, 2013). Unexpected positive and negative uses of big data mashups in environment and human health should be anticipated and discussed openly at regular intervals in the life of any project of this type (Boyd & Crawford, 2012; Jones, 2012; Leonelli, 2016; Michael, 2014). An obvious example consists of potentially dual-use research with national security implications, such as the "Missing Map Project," in which crowd-sourced volunteers digitize maps to aid in disaster response in developing nations, while also potentially producing useful tools for unintended users such as terrorists.

The promise of big data in low and middle income countries (LMICs) is balanced by the increasing international data access divide discussed above. Corporations and governments must be held accountable for providing access to environment and health data to all citizens, as well as arguably tools to be able to interact with those data and extract meaningful information from them (Royal Society, 2012). Recent threats in the United States, to remove access to or even destroy data funded by the Federal Government perceived to be politically sensitive (i.e., with regard to climate change or community health disparities/inequalities) should be challenged by researchers, communities, and policy makers. At the same time, national governments need to provide governance structures and legal framework within which health data sharing can happen without harm to citizens engaging in it.

The new concept of *open data* means that these data must be publicly available for anyone to use, and must be licensed in a way that allows for their reuse (Gurin, 2014). There is evidence for and the development of the FAIR Guiding Principles for scientific data management and stewardship. And there is the embrace of *Open Data* as a relatively new and powerful movement, which is being increasingly crystallized into science policy and funding guidelines, and which sends a strong signal about the importance of data access equity as a driving force for global innovation and creativity (EC, 2016; Science International, 2015; Wilkinson et al., 2016). In the future, could there be a "WIKI-style" (i.e., open access and moderated) Google Earth equivalent, where environment and health data of all types could be deposited, and from which these data could be readily accessed, mashed up, and analyzed?

Related to the previous discussion of data ownership, confidentiality, and privacy, there is a growing use of machine learning, algorithms, and application program interfaces (APIs). The approaches to harnessing big data could provide methods to use data in ways that protect ownership and confidentiality by design and with a lesser investment burden in terms of curation and sustainability (Jütting, 2016; Ribes & Bowker, 2009). For example, in the area of climate change in the context of city and/or nation level wide assessments, the uncertainty of individual health outcome responses to the environment (e.g., heat) grows with the scale of the aggregation in the health data. Nevertheless, we need to explore at what scale that data aggregation and uncertainty truly make a difference to the actual decision making in health policy and/or health management. It is possible that confidentiality of sensitive data may be achieved through the aggregation of these data and other techniques without significantly disrupting the usability of these data (Djennad et al., unpublished manuscript).

Similarly, there is an increasing potential to build in bespoke protections for confidential data (e.g., Enigma, developed by MIT, is a decentralized cloud platform with guaranteed privacy where private data are stored, shared, and analyzed without ever being fully revealed to any party). Examples of creative approaches specifically in the human health field to address confidentiality and privacy issues include the following:

- University College of London Data Safe Haven uses a “walled garden approach” allowing safe transfer and analysis of sensitive data.
- Datashield is a library that enables the remote and non-disclosive analysis of sensitive research data.
- Opal Project is a platform that unlocks the potential of private data for public good in a privacy-conscientious, scalable, socially, and economically sustainable manner.
- Patients Like Me, previously mentioned.
- My Clinical Outcomes is a platform that collects and reports relevant outcome measures from individual patients throughout their care according to local requirements.

The ability to link these clinical and other human health and well being data with the range of environmental and other data (while maintaining privacy and data confidentiality) would provide unique opportunities for data mashups in the future.

Box 4. Sendai Framework for Disaster Risk Reduction 2015–2030

Virginia Murray

As an example of the importance of environment and human health big data globally, of real import for the implementation of the Sendai Framework for Disaster Risk Reduction 2015–2030 (UN, 2015) is the need to use globally agreed minimum data sets, and other methods where appropriate, to support the assessment of global progress in achieving the agreed outcomes and the seven global targets. These targets are:

A. Substantially reduce global disaster mortality by 2030, aiming to lower the average per 100,000 global mortality rate in the decade 2020–2030 compared to the period 2005–2015.

B. Substantially reduce the number of affected people globally by 2030, aiming to lower the average global figure per 100,000 in the decade 2020–2030 compared to the period 2005–2015.

C. Reduce direct disaster economic loss in relation to global gross domestic product (GDP) by 2030.

D. Substantially reduce disaster damage to critical infrastructure and disruption of basic services, among them health and educational facilities, including by developing their resilience by 2030.

E. Substantially increase the number of countries with national and local disaster risk reduction strategies by 2020.

F. Substantially enhance international cooperation to developing countries through adequate and sustainable support to complement their national actions for implementation of the present Framework by 2030.

G. Substantially increase the availability of and access to multi-hazard early warning systems and disaster risk information and assessments to people by 2030.

Following the agreement at the United Nations Office for Disaster Risk Reduction (UNISDR) Open-Expert Working Group (OIEWG) at its Third Session in November 2016, the indicators for the Sendai Framework Goals have now been agreed and will be taken to the UN General Assembly in February 2017 (UNGA, 2016A). Targets A–D are outcome targets and will require disaster loss and damage data; E and G will require national self-assessment; and Target F will relate to overseas development system monitoring. These indicator outcomes (UNGA, 2016B) are described as follows:

i) *Compound Indicators*: Indicators to measure the achievement of the Global Targets, which can be constructed on the basis of a number of specific Global Indicators.

ii) *Global indicators*: Indicators ready to contribute to the global measurement of the target, for which a methodology exists, or has been proposed, and for which data is already available in a significant number of countries or can be generated through national self-assessment.

iii) *National Indicators*: Indicators for which a methodology exists or has been proposed, but for which data is not currently easily available in a significant number of countries. These indicators can be applied nationally in countries where the necessary data is available. When data becomes widely available in a larger number of countries, these indicators can potentially migrate to the Global Indicators category.

All the Sendai Framework indicators will be shared and used seamlessly by the Inter-agency and Expert Group on Sustainable Development Goal Indicators (IAEG-SDGs) to reduce the reporting burden on countries implementing both the Sendai Framework and the SDGs. A similar collaboration by UNISDR will hopefully lead to collaboration for the sharing of similar agreement for the Paris Climate Change Agreement.

The Sendai Framework for Disaster Risk Reduction and its Goals (including those for reduction of risk through early warning) and the United Nations (UN) Sustainable Development Goals (SDGs) are specifically linking the global sustainable development goals with individual health, well being, and environmental outcomes to monitor their delivery (Banning-Lover, 2016; IEAG, 2014; Lim, Fullman, Murray, & Mason-Jones, 2016; Sethi, 2016; see Box 4). Key big data goals in relation to the UN SDGs include:

- Developing a global consensus on principles and standards.
- Sharing technology and innovations for the common good.
- Creating and identifying new resources for capacity development.
- Providing leadership for coordination and mobilization as part of a UN-led “Global Partnership for Sustainable Development Data” (IEAG, 2014).

For example, the UN Global Pulse Initiative in their Pulse Lab Jakarta, in Indonesia, has developed a public-facing National Citizen Feedback Dashboard showing progress against Global Development Goals using environment and human health outcome measures.

In addition, lessons can be learned by the developed countries from LMICs and their innovative uses of cellular and other technologies to deliver a host of environment and health research and services. These range from cash transfers aligned with crop price reports, to climate change impacts, to innovative ways of gathering census data (Hall et al., 2014; Hussain et al., 2014). An example of such creative uses of environment and health data mashups taking place in LMICs is Flowminder, which uses mobile phone data to understand climate change and migration patterns in Bangladesh. As with other aspects of big data applications, Hall et al. (2014) note the continual need for scaling up small pilot projects, and throughout, for rigorous experimental and quasi-experimental studies to strengthen the evidence base.

Conclusions

The availability of big data and the tendency towards big data mashups are becoming more prevalent and widespread, and their multiple uses (and abuses) will continue to expand into the future. The growing impacts of climate and other environmental change on human health and well being and on the health of local and planetary natural systems necessitate the increased interdisciplinary-institutional-international use of these big data to fully understand and address these impacts.

The combination of growth in the volume and scale of data, alongside advances in data analytics, provide the potential for big data to be used to explore creatively, and suggest new understanding of and approaches for, the *wicked problems* of environment and human health. The data analysis and linkage tools, including machine learning, that are being developed should make it easier to link different types of data (i.e., environment and health) in the future.

At the same time, it is imperative to prioritize training in data management and analysis, as well as to incentivize widespread data sharing and equitable participation by researchers and the wider public. Better and more equitable access, as well as making available the capacity to work with these data; the ongoing sharing and evaluation of experience, resources, and data; careful consideration and extensive discussions of ethics, governance, data ownership, and privacy; and better communication and engagement with civil society, are all essential ingredients if we are to realize the full potential of big data and their mashups to address the current and future challenges of environment and human health.

Acknowledgments

This research was funded in part by the U.K. Medical Research Council (MRC) and the U.K. Natural Environment Council (NERC) for the MEDMI Project (MR/K019341/1); the U.K. National Institute of Health Research (NIHR) for the Health Protection Research Unit (HPRU) in Environmental Change and Health at the London School of Hygiene and Tropical Medicine with Public Health England (PHE), and in collaboration with the University of Exeter, University College London, and the Met Office; the European Research Council (FP7/2007–2013 Grant Agreement 335925) and the Australian Research Council (Discovery Project DP160102989); and the European FP7 Exposomics grant (grant agreement 308610).

References

Aicardi, C., Del Savio, L., Dove, E. S., Lucivero, F., Tempini, N., & Prainsack, B. (2016). **Emerging ethical issues regarding digital health data. On the World Medical Association Draft Declaration on Ethical Considerations Regarding Health Databases and Biobanks**. *Croatian Medical Journal*, *57*(2), 207–213.

Ashton, P. M., Nair, S., Peters, T. M., Bale, J. A., Powell, D. G., Painset, A., et al. (2016). **Identification of *Salmonella* for public health surveillance using whole genome sequencing**. *Peer J*, 4, e1752.

Banning-Lover, R. (2016, June 3). **The missing development trillions: How you would fund the SDGs**. *The Guardian*.

Bastow, R., & Leonelli, S. (2010). Sustainable digital infrastructure. *EMBO Reports*, *11*(10), 730–735.

Bezuidenhout, L., Rappert, B., Kelly, A., & Leonelli, S. (2017). **Beyond the digital divide: Towards a situated approach to open data**. *Science and Public Policy*.

Bingenheimer, J. B., & Raudenbush, S. W. (2004). Statistical and substantive inferences in public health: issues in the application of multilevel models. *Annual Review of Public Health*, *25*, 53–77.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication, & Society*, *15*, 5, 662–679.

Burton, P. R., Murtagh, M. J., Boyd, A., Williams, J. B., Dove, E. S., Wallace, S. E., et al. (2015). Data Safe Havens in health research and healthcare. *Bioinformatics*, *31*(20), 3241–3248.

Bradford Hill, A. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, *58*, 295–300.

Brownson, R. C., Samet, J. M., Chavez, G. F., Davies, M. M., Galea, S., Hiatt, R. A., et al. (2015). Charting a future for epidemiologic training. *Annals of Epidemiol*ogy, *25*(6), 458–465.

Brownstein, J. S., Freifeld, C. C., & Madoff, L. C. (2009). Digital disease detection: Harnessing the web for public health surveillance. *New England Journal of Medicine*, *360*, 2153–2157.

Canali, S. (2016). Big data, epistemology, and causality: Knowledge in and knowledge out in EXPOsOMICS *Big Data Society, 3*(2), 1–11.

Contreras, J. L. (2011). Bermuda's legacy: Policy, patents and the design of the genome commons. *Minnesota Journal of Law, Science, & Technology*, *12*, 61.

Corvalan, C., Hales, S., & McMichael, A. (2005). *Ecosystems and human well-being: Health synthesis: A report of the Millennium Ecosystem Assessment (MEA)*. Geneva, Switzerland: World Health Organisation (WHO).

Daniel, F., & Matera, M. (2014). *Mashups: Concepts, models and architectures (data-centric systems and applications*). New York: Springer.

Del Savio, L., Prainsack, B., & Buyx, A. M. (2016). Crowdsourcing the human gut. Is crowdsourcing also "citizen science"? *Journal of Science Communication*, *15*(3), A03 1–16.

Dingle, K. E., Colles, F. M., Wareing, D. R., Ure, R., Fox, A. J., Bolton, F. E., et al. (2001). Multilocus sequence typing system for Campylobacter jejuni. *Journal of Clinical Microbiology*, *39*(1), 14–23.

European Commission (EC). (2016). ***Open innovation, open science, open to the world***. Directorate-General for Research and Innovation.

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, *41*(5), 667–690.

Eynon, R. (2013). **The rise of big data: What does it mean for education, technology, and media research**? *Learning Media Technology*, *38*(3).

Fleming, L. E., Haines, A., Golding, B., Kessel, A., Cichowska, A., Sabel, C. E., et al. (2014). Data mashups: Potential contribution to decision support on climate change and health. *International Journal of Environmental Research and Public Health*, *11*, 1725–1746.

Ford, D. V., Jones, K. H., Verplancke, J.-P., Lyons, R. A., John, G., Brown, G., et al. (2009). The SAIL databank: Building a national architecture for e-health research and evaluation. *BMC Health Services Research*, *9*(1), 157.

Gibin, M., Mateos, P., Petersen, J., & Atkinson, P. (2009). Google maps mashups for local public health service planning. Planning support systems best practice and new methods. *The GeoJournal Library*, *95*, 227–242.

Gliklich, R. E., Dreyer, N. A., & Leavy, M. B. (2014). **Principles of registry ethics, data ownership, and privacy**. In R. E. Gliklich, N. A. Dreyer, & M. B. Leavy (Eds.), *Registries for Evaluating Patient Outcomes: A User's Guide* [Internet]. (3d ed.). Rockville, MD: U.S. Department of Health and Human Services.

Granella, C., Belmonte Fernández, O., & Díaz, L. (2014). Geospatial information infrastructures to address spatial needs in health: Collaboration, challenges, and opportunities. *Future Generation Computer Systems*, *31*, 213–222.

Gurin, J. (2014). **Big data and open data: What's what and why does it matter**? *The Guardian*, April 15.

Haines, A., McMichael, A. J., & Epstein, P. R.(1993). Global health watch: Monitoring impacts of environmental change. *Lancet*, *342*, 1464–1469.

Hall, C. S., Fottrell, E., Wilkinson, S., & Byass, P. (2014). **Assessing the impact of mHealth interventions in low- and middle-income countries: What has been shown to work**? *Global Health Action*, *7*.

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., et al. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, *11*(3), 156–162.

Hey, T., Tansley, S., & Tolle, K. (2009). *The fourth paradigm: Data-intensive scientific discovery*. Redmond, WA: Microsoft Research.

Hussain, S., Hun Bang, J., Han, M., Ahmed, M.I., Amin, M.B., Lee, S., et al. (2014). Behavior life style analysis for mobile sensory data in cloud computing through MapReduce. *Sensors (Basel)*, *14*(11), 22001–22020.

Independent Expert Advisory Group on a Data Revolution for Sustainable Development (IEAG). (2014). ***A world that counts: Mobilising the data revolution for sustainable development***. Report prepared at the request of the United Nations Secretary-General.

Jolley, K. A., Bliss, C. M., Bennett, J. S., Bratcher, H. B., Brehony, C., Colles, F. M., et al. (2012). Ribosomal multilocus sequence typing: Universal characterization of bacteria from domain to strain. *Microbiology*, *158*(4), 1005–1015.

Jolley, K. A., & Maiden, M. C. (2010). BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, *11*, 595.

Jones, M. (2012, March 28). **Big Data's unintended consequences**. *The Sydney Morning Herald*.

Jorm, L. (2015). **Routinely collected data as a strategic resource for research: Priorities for methods and workforce**. *Public Health Research & Practice*, *25*(4), e2541540.

Jütting, J. (2016, February 2). **Citizens, the private sector and SDG implementation: Scale success and scrap the rest**. *Huffington Post*.

Kallinikos, J., & Tempini, N. (2014). Patient data as medical facts: Social media practices as a foundation for medical knowledge creation. *Information Systems Research*, *25*(4), 817–833.

Kamel Boulos, M. N., & Al-Shorbaji, N. M. (2014). **On the Internet of things, smart cities and the WHO healthy cities**. *International Journal of Health Geographics*, *13*(1), 10.

Kessel, K. A., & Combs, S. E. (2016). **Review of developments in electronic, clinical data collection, and documentation systems over the last decade: Are we ready for big data in routine health care**? *Frontiers in Oncology*, *6*, 75.

Khoury, M. J., & Ioannidis, J. P. A. (2014). Big data meets public health: Human well-being could benefit from large-scale data if large-scale noise is minimized. *Science*, *346*(6213), 1054–1055.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. London: SAGE.

Lang, T., & Rayner, G.(2012). Ecological public health: The 21st century's big idea? *BMJ*, *345*, 17–20.

Lazar, D., Kennedy, R., King, G., & Vespignani A. (2014). **The parable of Google flu: Traps in big data analysis**. *Science*, *343*,1203–1205.

Leonelli, S. (2012). When humans are the exception: Cross-species databases at the interface of clinical and biological research. *Social Studies of Science*, *42*(2), 214–236.

Leonelli, S. (2014). What difference does quantity make? On the epistemology of Big Data in biology. *Big Data & Society*, *1*, 1–11.

Leonelli, S. (2016). **Locating ethics in data science: responsibility and accountability in global and distributed knowledge production**. *Philosophical Transactions of the Royal Society: Part A*, *374*(2083), 20160122.

Lim, S. S., Fullman, N., Murray, C. J., & Mason-Jones (2016). Measuring the health-related Sustainable Development Goals in 188 countries: A baseline analysis from the Global Burden of Disease Study 2015. *Lancet*, *388*, 1813–1850.

Maeda, E. E., & Arévalo Torres, J. (2012). Open environmental data in developing countries: Who benefits? *Ambio*, *41*(4), 410–412.

Maiden, M. C., & Harrison, O. B. (2016). Population and functional genomics of Neisseria revealed with gene-by-gene approaches. *Journal of Clinical Microbiology*, *54*(8), 1949–1955.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York: Eamon Dolan/Houghton Mifflin Harcourt.

McMichael, A. J., & Haines, A. (1997). Global climate change: The potential effects on health. *BMJ*, *315*(7111), 805–809.

Michael, C. (2014, October 6). **Missing maps: Nothing less than a human genome project for cities**. *The Guardian*.

Morrison, M. L., Marcot, B., & Mannan, W., (2012). *Wildlife-habitat relationships: concepts and applications*. New York: Island Press.

Nichols, G. L., Andersson, Y., Lindgren, E., Devaux, I., & Semenza, J. C. (2014). European monitoring systems and data for assessing environmental and climate impacts on human infectious diseases. *International Journal of Environmental Research & Public Health*, *11*(4), 3894–3936.

Normandeau, K. (2013). ***Beyond volume, variety and velocity is the issue of big data veracity. Inside big data***.

Nuffield Council on Bioethics. (2015). ***The collection, linking and use of data in biomedical research and health care: ethical issues***. The Nuffield Foundation

Ossorio, P. (2011). Bodies of data: Genomic data and bioscience data sharing. *Social Research*, *78*(3), 907–932.

Panahiazar, M., Taslimitehrani, V., Jadhav, A., & Pathak, J. (2014). Empowering personalized medicine with big data and semantic web technology: Promises, challenges, and use cases. *Proceedings of the IEEE International Conference on Big Data* (pp. 790–795). IEEE.

Pecina-Slaus, N., & Pecina, M. (2015). Only one health, and so many omics. *Cancer Cell International*, *15*, 64.

Press, G.(2014). **12 big data definitions: What’s yours**? *Forbes Magazine*.

Ribes, D., & Bowker, G. C. (2009). Between meaning and machine: Learning to represent the knowledge of communities. *Information and Organization*, *19*(4), 199–217.

Roche, D. G., Jennions, M. D., & Binning, S. A. (2013). Data deposition: Fees could damage public data archives. *Nature*, *502*(7470), 171.

Rockström, J., Sachs, J. D., Öhman, M. C., & Schmidt-Traub, G.(2013). ***Sustainable development and planetary boundaries***.

Royal Society. (2012). **Science as an open enterprise**.

Science International. (2015). ***Open data in a big data world***.

Sethi, T. (2016). **Crowding in funding for SDGs: Reflections on the missing development trillions**. *AidData*.

Sheppard, S. K., Colles, F., Richardson, J., Cody, A. J., Elson, R., Lawson, A., et al. (2010). Host association of Campylobacter genotypes transcends geographic variation. *Applied and Environmental Microbiology*, *76*(15), 5269–5277.

Sheppard, S. K., Jolley, K. A., & Maiden, M. C.(2012). **A gene-by-gene approach to bacterial population genomics: Whole genome MLST of *Campylobacter***. *Genes (Basel)*, *3*(2), 261–277.

Tempini, N. (2015). Governing *PatientsLikeMe*: Information production and research through an open, distributed and data-based social media network. *The Information Society*, *31*(2), 193–211.

Tempini, N. (2016). Science through the “golden security triangle”: Information security and data journeys in data-intensive biomedicine. In *Proceedings of the 37th International Conference on Information Systems (ICIS)*. Dublin, December 11–14.

Tene, O., & Polonetsky, J. (2012). **Privacy in the age of big data: A time for big decisions**. *Standford Law Review*.

Townsend, A. M. (2013). *Smart cities: Big data, civic hackers, and the quest for a new utopia*. New York: Norton.

United Nations (UN). (2015). ***Sendai framework for disaster risk reduction 2015–2030***.

UN General Assembly (UNGA). (2016a). **Open-ended intergovernmental expert working group on indicators and terminology relating to disaster risk reduction**. Geneva, September 29–30, 2015, February 10–11, 2016, and November 15–18, 2016.

UN General Assembly (UNGA). (2016b). ***Report of the open-ended intergovernmental expert working group on indicators and terminology relating to disaster risk reduction***. (Geneva, September 29–30, 2015, February 10–11, 2016, and November 15 & 18, 2016). Advance unedited version.

Velasco, E., Agheneza, T., Denecke, K., Kirchner, G., & Eckmanns, T. (2014). Social media and Internet-based data in global systems for public health surveillance: A systematic review. *Milbank Q*, *92*(1), 7–33.

Vineis, P., Chadeau-Hyam, M., Gmuender, H., Gulliver, J., Herceg, Z., Kleinjans, J., et al. (2016). **EXPOsOMICS Consortium. The exposome in practice: Design of the EXPOsOMICS project**. *International Journal Hygiene and Environmental Health*.

Ward, J. S., & Barker, A. (2013). **Undefined by data: A survey of big data definitions**. *School of Computer Science University of St Andrews, United Kingdom*.

Whitmee, S., Haines, A., Beyrer, C., Boltz, F., Capon, A. G., Ferreira de Souza Dias, B., et al. (2015). Safeguarding human health in the Anthropocene epoch: Report of The Rockefeller Foundation–Lancet Commission on planetary health. *The Lancet*, *386*, 10007.

Wicks, P., Vaughan, T., & Heywood, J. (2014, January 28–29). **Subjects no more: what happens when trial participants realize they hold the power?***BMJ*, *348*, g368.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). **The FAIR Guiding Principles for scientific data management and stewardship**. *Scientific Data*, *3*, 160018.

Woese, C. R., & Fox, G. E.(1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences USA*, *74*(11), 5088–5090.

Woese, C. R., Kandler, O., & Wheelis, M. L.(1990). Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences USA*, *87*(12), 4576–4579.

Wu, W., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., et al. (2009). **BioGPS: An extensible and customizable portal for querying and organizing gene annotation resources**. *Genome Biology*, *10*, R130.

Yewell, J. (2015). **Big data presents big challenges, big opportunities in environmental health**. *NIEHS Environmental Factor*.

Lora Fleming

University of Exeter Medical School, European Centre for Environment and Human Health

Anthony Kessel

Public Health England

Virginia Murray

Public Health England

Michael Depledge

University of Exeter Medical School, European Centre for Environment and Human Health

Sabina Leonelli

University of Exeter

Niccolò Tempini

University of Exeter

Harriet Gordon-Brown

European Centre for Environment and Human Health, University of Exeter Medical School

Gordon Nichols

Public Health England

Christophe Sarran

Met Office

Paolo Vineis

Imperial College London

Giovanni Leonardi

Public Health England

Brian Golding

Met Office

Andy Haines

London School of Hygiene and Tropical Medicine