*International Journal of Epidemiology*, 2020, i4–i14
doi: 10.1093/ije/dyz051
Supplement Article

Supplement Article

# Availability, access, analysis and dissemination of small-area data

**Susan Hodgson ,[1]* Daniela Fecht,[1,2] John Gulliver,[1] Hima Iyathooray Daby,[1,2] Frédéric B Piel ,[1,2] Fuyuen Yip,[3] Heather Strosnider,[3] Anna Hansell [1,2] and Paul Elliott[1,2]**

[1]MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, London, UK, [2]UK Small Area Health Statistics Unit, MRC-PHE Centre for Environment and Health, Imperial College London, London, UK and [3]Environmental Health Tracking Section, National Center for Environmental Health, Centers for Disease Control and Prevention, Atlanta, USA

*Corresponding author. Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, Imperial College London, St Mary's Campus, Norfolk Place, London W2 1PG, UK. E-mail: susan.hodgson@imperial.ac.uk

Accepted 11 March 2019

## Abstract

In this era of 'big data', there is growing recognition of the value of environmental, health, social and demographic data for research. Open government data initiatives are growing in number and in terms of content. Remote sensing data are finding widespread use in environmental research, including in low- and middle-income settings. While our ability to study environment and health associations across countries and continents grows, data protection rules and greater patient control over the use of their data present new challenges to using health data in research. Innovative tools that circumvent the need for the physical sharing of data by supporting non-disclosive sharing of information, or that permit spatial analysis without researchers needing access to underlying patient data can be used to support analyses while protecting data confidentiality. User-friendly visualizations, allowing small-area data to be seen and understood by non-expert audiences, are revolutionizing public and researcher interactions with data. The UK Small Area Health Statistics Unit's Environment and Health Atlas for England and Wales, and the US National Environmental Public Health Tracking Network offer good examples. Open data facilitates user-generated outputs, and 'mash-ups', and user-generated inputs from social media, mobile devices and wearable tech are new data streams that will find utility in future studies, and bring novel dimensions with respect to ethical use of small-area data.

**Key words:** Small-area studies, open data, remote sensing, environment and health

**Key Messages**

- Availability of spatially resolved data is increasing, with new data being put to use in small-area environment and health studies.
- Access to data is supported by open data initiatives, however the tightening of information governance and greater autonomy over use of personal data will impact health research.
- Analysis tools that support and audit the ethical and legal use of health data are likely to find increasing utility in small-area and multi-cohort studies.
- Dissemination of data to a wide audience can support public understanding of environment and health research.
- User-generated data, from social media, smart phones and wearable tech, will support future environment and health studies.

## Introduction

This paper reflects on some of the issues associated with the availability, access, analysis and dissemination of small-area data. Small-area data in this context refers to data describing characteristics—e.g. health, environment, demographic, economic—of a defined geographic area. The definition of 'small area' will vary, study by study, depending on access and availability of data (more below), but also the scale appropriate to the analyses to be undertaken (which is contingent on rarity of outcome and/or the size of the population under study).[1] Ideally the small areas will describe relatively homogeneous populations in terms of exposures and/or key variables of interest, as it is this characteristic of a small-area study that can reduce components of ecological bias.[2] The small-area approach, then, benefits from the efficiencies of being able to utilize data collected and made available at a range of postal, census and/or administrative geographies, while minimizing the bias inherent in the ecological approach. Here, we offer our view on upcoming opportunities and challenges in this field, drawing on our 30 years of experience working at, or with, the UK Small Area Health Statistics Unit (SAHSU).

## Availability of small-area data

Spatially resolved datasets are becoming increasingly available for research, although it is important that the quality of the data, and their limitations, are well understood in order to be able to make meaningful inferences from their use.[3,4] In the UK, existing public data are becoming more readily available via platforms such as the UK Open Data portal (data.gov.uk), which currently provides free access to >40 000 datasets from more than 1000 sources, under the principle that all information created by the central government, local authorities and public sector bodies should be made available for re-use. Similar web portals hosting open government data are available, for example, in the USA (www.data.gov), France (www.data.gouv.fr) and Singapore (www.data.gov.sg).[5]

Such data are being accessed and downloaded with increasing frequency for a wide range of purposes. Figure 1 shows monthly data downloads for the top 10 publishers from the Data.gov.uk portal. There is also evidence emerging to suggest an increasing use of open government data specifically for scientific research, most prominently in high-income countries. Amongst 1229 studies using open government data investigated in a review by Yan and Weber,[6] 25.5% and 11.6% used open access data from the UK and USA, respectively; however, 7% and 6.3% of studies used data from India and Kenya, suggesting open data are also a valuable scientific resource in some low- and middle-income countries.

### Health data

As seen in Figure 1, open data from the Office for National Statistics and NHS Digital are frequently downloaded from the Data.gov.uk portal. NHS digital publish more than a thousand datasets on health and social care provided by the National Health Service (NHS). These datasets describe aspects of primary care, secondary care, emergency care, community services, maternity services, mental health, social care, clinical audits and disease registers, prescribing, population health, the NHS workforce and estates, clinical indicators, healthcare resources, data standards and data quality.[7] The Office for National Statistics (ONS) provides annual statistics on births and deaths in England and Wales, with near complete ascertainment.[8,9] The ONS also holds national data on child health, disability, drug use, alcohol and smoking, and life expectancy. Public Health England are custodians of the National Cancer Registration and Analysis Service, which provides data on 30 years of cancer registrations.[10]

In addition to these established datasets that are well used in research, new health datasets are being made

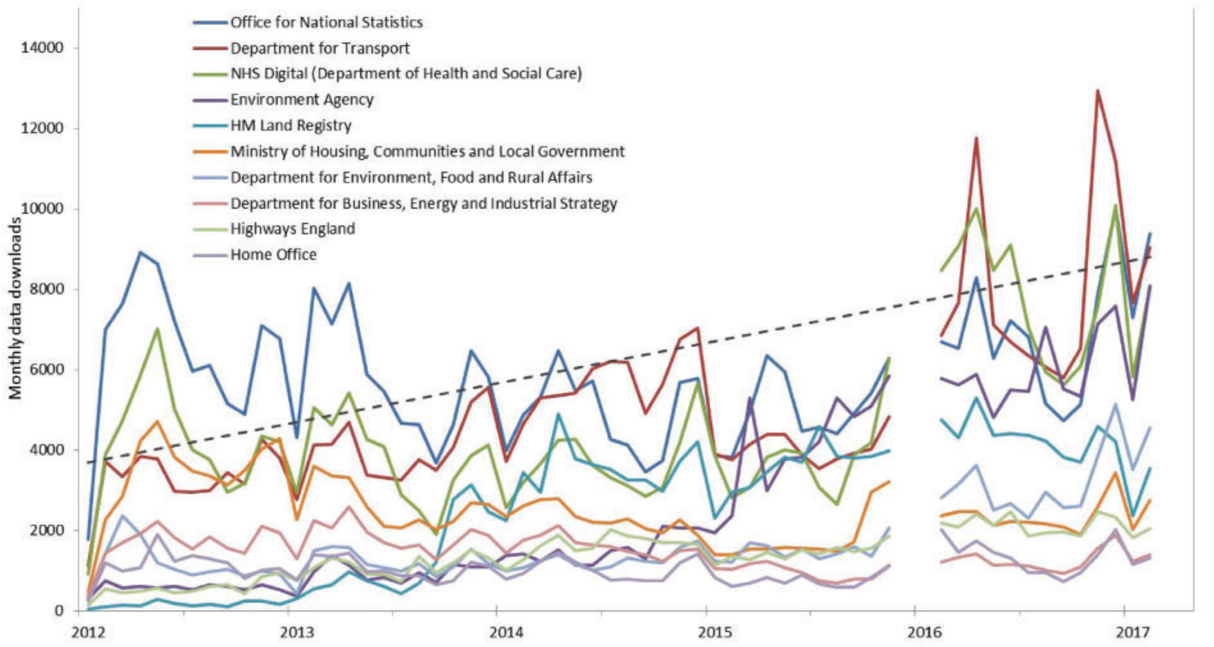

**Figure 1.** Usage statistics (downloads), by publisher for each month (December 2012–March 2018) from Data.gov.uk for top 10 data publishers, with linear trend line for total monthly downloads. Data generated April 5, 2018, from Google Analytics. Contains public sector information licensed under the Open Government Licence v3.0.

available for research. For example, general-practice-level prescribing data, detailing all medicines, dressings and appliances prescribed and dispensed each month in England since August 2010 (https://openprescribing.net/), have found utility in studies assessing health service inequalities[11] and in work linking environmental exposures to volume and cost of prescribing.[12] Hospital outpatient data for England was accredited as a national statistic in 2008. This valuable resource, detailing 119 million outpatient appointments in the year February 2017 to January 2018[13] is now also being used in research alongside the more established inpatient data.[14–16] The quality of these data, in terms of ascertainment, has however been found lacking (e.g. self-harm episodes were under-ascertained when compared with local bespoke data collection methods[16]) and the diagnosis fields were reported to be too infrequently populated to be useful (although it was possible to use date and specialty of appointment to study hospital-service use following heart failure[15]).

## Environmental data

With respect to environmental data describing the distribution of levels of pollutants, the increased availability of highly spatially resolved data, for large geographic areas, has supported the development of harmonized exposure assessment on a national/international scale, for certain pollutants. This up-scaling and harmonization supports cross-country comparisons and increases the statistical power to look at subtle effects and/or rare outcomes. Recent air pollution exposure assessments on national and international scales have been based on land use regression (LUR) modelling approaches that combine two types of data: (i) a chemical transport model or information from satellite data [e.g. aerosol optical depth (AOD)] to describe regional/background concentrations of particulate pollution, typically with a granularity of 1–10 km, and (ii) localized spatial predictors summarized in circular buffers of varying radii from tens to hundreds of metres (e.g. lengths of roads, traffic intensity, area of housing, industry, green space etc.) or distance to source. A spatially distributed network of air pollution monitors is required to develop and validate the LUR model. Using this approach, de Hoogh *et al.*[17] (2016) developed air pollution surfaces at a resolution of 100 meters for the whole of western Europe using data from the European network of air pollution monitoring sites (Airbase), data from a chemical transport model ($NO_2$) and AOD [particulate matter with an aerodynamic diameter <2.5 micrometers ($PM_{2.5}$)] for background sources, pan-European land cover (CORINE), and national models of road traffic. This LUR model explained 60% of the variability in $PM_{2.5}$ concentrations (48% related to AOD; 12% related to local predictors); without information on AOD, the local spatial predictors only explained 38% of the variability in $PM_{2.5}$ concentrations. Such air pollution surfaces produced on a 100 m grid provide sufficient granularity to characterize the variability in exposures within and between small areas. Similar examples include $NO_2$ for Australia,[18] the USA,[19] and a global model of $NO_2$ albeit with a reduced spatial resolution in some countries.[20]

Whereas such models have been used for estimation of outdoor pollutant concentrations at specific locations, e.g. residential address, for defined periods, such as annual[21,22] or trimester-specific estimates for birth outcomes,[23,24] they are not suitable for estimating time-varying exposures at a much higher temporal resolution, e.g. with respect to daily commuting patterns, nor for indoor exposures. Methods to address these areas require further development, as discussed further below.

Similar advances due to better accessibility to environmental data have been observed in other fields as well. Noise is largely a local pollutant, propagating over tens to hundreds of metres in most cases. Accurate noise modelling requires detailed geographical data to characterize source emissions (e.g. individual road links with time-varying traffic flows/speeds) and noise propagation (e.g. geometry and height of individual buildings, and highly resolved land cover data differentiating between different types of man-made and natural surfaces) and is computationally demanding. These considerations have presented a substantial challenge for modelling noise over large geographical areas, especially for national- or international-scale epidemiological studies. Although proprietary software (e.g. CadnaA, SoundPlan) has long been available, open-source software has only recently been developed,[25,26] with some simplifications both to data requirements (e.g. using average building height and fewer categories of land cover) and emissions/propagation models to serve the needs of large-scale studies. Noise levels are normally calculated at one or more façade locations of individual dwellings,[26–28] except, for example, in EU strategic noise mapping in agglomerations (>100 000 residents) where surfaces are produced. Morley *et al.*[25] describe the implementation of a version of the CNOSSOS-EU (Common NOise aSSessment MethOdS) noise modelling framework, which was used to undertake harmonized noise exposure assessment for four large cohort studies (UK Biobank and EPIC Oxford in the UK, Lifelines in the Netherlands and HUNT in Norway) participating in the EU FP7 funded Biobank Standardisation and Harmonisation for Research Excellence in the European Union (BioSHaRE-EU) project. This work found road traffic noise exposure to be associated with blood biochemistry[29] and heart rate,[30] but not with incident cardiovascular disease,[31] blood pressure[30] or asthma prevalence.[32]

The examples above highlight the benefits of increased data availability and processing power to generate improved resolution exposure models with greater geographical coverage. However, the resultant exposure metrics still represent outdoor exposure concentrations that may not reflect personal exposure. The benefits of studying a larger population, i.e. the ability to study rare diseases, undertake sub-group analyses and assess interactions, need to be balanced by the loss of power/interpretability introduced when using such proxy measures of personal exposure. As discussed later, geo-location data from mobile devices and sensor data from wearable tech might permit modelled outdoor exposures to be calibrated to individuals and/or sub-groups of the population to better reflect personal exposures.

Use of satellite-derived data is of particular importance for low- and middle-income countries (LMICs) where routinely collected environmental data might not be available, as well as for global studies such as those conducted by the Global Burden of Disease project (GBD, www.healthdata.org/gbd ).[33] Remotely sensed data are now extensively used for epidemiological purposes to evaluate global risks to human health via climate change or environmental pollution.[33] They are also crucial in identifying suitable habitats for vector-borne diseases and thereby to monitor, control and prevent infectious diseases such as malaria and neglected tropical diseases.[34–36] The Normalized Difference Vegetation Index (NDVI), for example, is a widely applied indicator from remotely sensed data to evaluate the extent of live green vegetation. This has been used to track agricultural patterns to estimate the burden of child malnutrition in African countries[37] and, more often in developed countries, to estimate exposure to salutogenic urban green space (e.g.).[38] Satellite-derived environmental data have also contributed to a better understanding of distributions and geographical patterns of infections and chronic diseases that are largely driven by environmental factors such as temperature, soil characteristics and land use. Examples include soil-transmitted helminth infections in South America, where remotely sensed data have allowed spatial distributions of infection prevalence to be predicted, to support the targeting of populations for treatment,[39] and Zika virus, where remotely sensed data contributed to vector distribution modelling, suggesting that 2.17 billion people inhabit areas that are environmentally suitable for Zika transmission.[40]

## Socio-economic data

Deprivation is strongly associated with risk of disease and with disease risk factors, with inequalities in socio-economic status accounting for half of the inequalities in some diseases.[41] As such, it is important to gather area-level information on deprivation so that adjustment for this key confounding variable can be made. Within the UK, several area-level measures of deprivation, derived from census data, are available. The most commonly used include (i) the Carstairs Index, which comprises four indicators of material disadvantage—lack of car ownership, low occupational social class, overcrowded households and male unemployment[42]; (ii) Townsend Deprivation Score, which also includes four indicators of deprivation—unemployment, non-car ownership, non-home ownership and overcrowding,[43] and (iii) the Index of Multiple Deprivation (IMD), which combines information from seven domains of deprivation—income, employment, education/skills/training, health and disability, crime, housing and services, and living environment.[44] The choice of deprivation measure will depend on the geographic scale and/or time frame of analysis (not all measures are able to be calculated for all geographic areas, nor from each census), but also on measures used in previous studies (if direct comparison is important), and on the composition of the index [e.g. if it is important to adjust (or not) for health (which is a component of IMD)]. Whichever measure is used, it should be remembered that these are all ecological measures of deprivation, i.e. describe area-based, not individual circumstance.

### Population data

Prerequisite for the accurate identification of populations at risk from geographically varying diseases is detailed, high spatial resolution information on human population distributions. In developed countries, various methods have been put forward to obtain accurate information on population at the small-area scale.[45] But this is a particular challenge in LMICs where extensive mapping resources and detailed population counts are often lacking. Recent developments such as the WorldPop project (http://www.worldpop.org.uk/) have tried to fill this gap using geospatial data from satellites on land cover and light-at-night, together with locally available census information, to estimate small-area population distributions globally.[46]

The health, environment and socio-economic data holdings at SAHSU, to support small-area health analyses within the UK, are summarized in Table 1.

## Access to small-area data

Against this backdrop of increased availability of health and environmental data, there has been a tightening of information governance regulation, particularly in the UK and EU. For example, in January 2014, the UK government formalized the right of patients in England to request that their confidential information is not used beyond their own care and treatment. Patients who 'opt-out' will no longer have their personal confidential information appear in data disseminations from data providers; their information can only be made available in anonymized form (so that the individuals are not identified in the data). Data from patients who opt-out may not therefore be represented in epidemiological or health service research studies where personal confidential data are required for data linkages. Patients choosing to opt out are demographically and geographically clustered, and increasing in number over time. For instance, opt-out rates were reported to be significantly higher in older versus younger patients, in female versus male patients, and differed by ethnicity and area-level deprivation.[47] As age, sex, ethnicity and socio-economic status are important confounders in health studies, differential loss of numerator and/or denominator data due to opt-outs is likely to bias observed associations between risk factors and health outcomes.

Beyond the UK, Australia is currently debating the pros and cons of opting in/out of the electronic 'My Health Record', that advocates claim will benefit patients via the sharing of key health information, and lead to efficiencies in healthcare, while others are concerned about data security issues.[48] In 1995 Denmark established a system whereby citizens could opt-out of having their details shared for research projects, however this option was revoked in 2014 when an estimated 16% of the population had opted out.[49] Information on how data are used, who has access to sensitive information, and the safeguards in place need to be carefully communicated if the public are to be convinced of the case for sharing health data for research. The growing demand for patients to control their data in the era of big data is understandable, but there are important consequences for public health research that relies on the availability of comprehensive datasets and patient identifiable data to improve health and social care.

The EU General Data Protection Regulation (GDPR), enforced May 25, 2018, aims to harmonize data privacy laws across Europe (replacing the Data Protection Act 1998 in the UK), with important implications for health research. Health data fall within the 'special category' of personal data, and require, in addition to a 'lawful basis' for processing, that further conditions are met before use. Of relevance here, data processing for preventative or occupational medicine, the management of health or social care systems and services, and public health are included as further conditions for processing such special category data. The harmonization of data protection regulation might improve and facilitate data sharing across the EU, supporting multi-country studies.[50]

Although data sharing between countries within the EU might be supported by the GDPR, the ethical sharing of

**Table 1.** Health, environment and socio-economic data holdings of the Small Area Health Statistics Unit (SAHSU) database (more details at: https://www.sahsu.org/content/sahsu-database)

| Dataset | Geographic extent | Provider |
|---|---|---|
| Health data | | |
| Hospital Episodes Statistics (HES) Admitted Patient Care | England | NHS Digital |
| HES Accident and Emergency | England | NHS Digital |
| HES Critical Care | England | NHS Digital |
| Cancer registrations | England | Office for National Statistics |
| Cancer registrations | Wales | The Welsh Cancer Intelligence and Surveillance Unit |
| Deaths registrations | England and Wales | Office for National Statistics |
| Birth and still births registrations | England and Wales | Office for National Statistics |
| Local Congenital Anomaly Registers | Regions of the United Kingdom and the Republic of Ireland | British Isles Network of Congenital Anomaly Registers |
| Scottish births, mortality and congenital anomalies | Scotland | NHS Scotland (Information Services Division) |
| Environment data | | |
| Land Cover Map | Great Britain | Centre for Ecology and Hydrology |
| CORINE Land Cover | EU | European Environment Agency |
| Agricultural Census | England | Edinburgh |
| Air temperature | England and Wales | SAHSU/British Atmospheric Data Centre |
| Sunshine duration | England and Wales | SAHSU/British Atmospheric Data Centre |
| Light emissions | EU | SAHSU/NOAA National Geophysical Data Centre |
| $NO_2$ and $PM_{10}$ | Great Britain | RGI |
| $NO_2$ and $PM_{10}$ background concentrations | EU | APMOSPHERE EU project |
| $NO_2$ and $PM_{10}$ | EU | ESCAPE EU project |
| Historic black smoke and $SO_2$ | Great Britain | CHESS Wellcome project |
| $NO_2$ | Great Britain | CHESS Wellcome project |
| Heavy metals (lead and cadmium) in soil | England and Wales | SAHSU/Countryside Survey |
| Road Traffic Noise | Great Britain | BioSHaRE EU project |
| Road Traffic Noise | London | TRAFFIC NERC project |
| Socio-economic data | | |
| Carstairs Index | Great Britain | Office for National Statistics (Census data) |
| Townsend Index | England | Office for National Statistics (Census data) |
| Index of Deprivation | England, Wales, Scotland, Northern Ireland | Ministry of Housing, Communities & Local Government |
| Urban–rural classifications | Great Britain | Office for National Statistics |

CHESS, Chronic Health Effects on Smoke and Sulphur; ESCAPE, European Study of Cohorts for Air Pollution Effects; HES, Hospital Episodes Statistics; NERC, Natural Environment Research Council; NOAA, national oceanic and atmospheric administration; RVI, Ruimte voor Geo-Informatie.

data between different jurisdictions presents a challenge to the scaling-up of research. 'Compute to data' methods provide one approach to avoid the need for physical sharing of data. As an example, DataSHIELD (www.datashield.ac.uk/) has been developed to support the non-disclosive sharing of information. This approach can facilitate research in circumstances where data governance might prevent the release of data and/or the combination of multiple datasets for unified analysis, or where data providers are happy to share information but do not wish to cede control of the governance of those data and/or the intellectual property they represent by physically sharing.[51] Such an approach was used in the BioSHaRE-EU project to permit the combined individual-level analysis of harmonized data from participants from several European cohorts, some of which were held by cohort custodians and which were queried remotely. For example, this approach was used to assess the associations between ambient air pollution and traffic noise and adult asthma prevalence (using data from 646 731 participants from HUNT3, Lifelines and UK Biobank[32]) and cardiovascular risk factors (in 144 082 participants from HUNT3 and Lifelines[29]) and air quality on wheeze/shortness of breath (in 377 954/173 560 participants from Lifelines and UK Biobank[52]).

## Analysis of small-area data

The need for user-friendly tools, capable of processing a large amount of information and supporting the linkage of datasets, their analysis and visualization is also increasing. One such tool is the SAHSU 'Rapid Inquiry Facility' (RIF), developed in the late 1990s[53] and refined for use in the EU.[54] The RIF software was then adapted and enhanced for use in the US Centers for Disease Control and Prevention (CDC) National Environmental Public Health Tracking Network, as one of several tools used by the Tracking Program.[55] The RIF facilitates environmental health analysis, by linking health, environmental, socioeconomic, population and geographic data. The RIF supports disease mapping studies (standardized disease rates and relative risks across a user-specified area to explore the spatial distribution of disease) and risk analysis (to investigate whether a putative exposure source is associated with adverse health outcomes in the exposed population). The latest version of the RIF (RIF 4.0) is an open source, freely accessible web platform on a spatially enabled database, PostGIS.[56] In addition to the integration of advanced methods in statistics, exposure assessment and data visualization, the RIF also generates an audit trail, to facilitate adherence to data protection and information governance requirements mentioned above.

The RIF has been used to assess, for example, kidney disease mortality following a historic industrial contamination incident in the UK,[57] mortality from cardiovascular and cerebrovascular disease and drinking water hardness in Spain,[58] the association between deprivation and circulatory system disease mortality in Hungary[59] to investigate cancer rates in residents living over contaminated groundwater plumes near an Air Force Base in Utah, USA,[60] and to explore the geographic variation of cancer incidence at the neighbourhood level in Ontario, Canada.[61]

Tools, such as the RIF, can facilitate small-area analysis, but cannot replace the need for epidemiological, geographical and statistical input in the planning, analysis and interpretation of small-area analyses. This expertise is necessary to ensure that: (i) appropriate health, exposure, covariate and population data are selected; (ii) associations are assessed at a meaningful geographic (and temporal) scale; (iii) an appropriate statistical approach is applied; and (iv) the resulting output is interpreted, with full appreciation of any data quality issues and an understanding of the limits of the small-area approach. Considerations regarding data choice, study area, time period and analytic method are further discussed by Piel *et al*,[1] and a review of the main methodologic issues associated with the small-area approach is presented in Beale *et al*.[2]

## Dissemination of small-area data

Development of user-friendly interfaces and visualizations has allowed small-area data to be disseminated to a wide audience including researchers, health professionals, policy makers and the public. The Environment and Health Atlas for England and Wales and US CDC National Environmental Public Health Tracking Network website, are two examples where health and environment data are presented, along with supporting information, to facilitate public understanding.

The SAHSU's Environment and Health Atlas for England and Wales (www.envhealthatlas.co.uk) provides interactive maps for a range of health conditions and environmental agents at the small-area scale (census ward level, average population 6000) in England and Wales. The maps were developed as a resource for the public, researchers and those working in public health and policy, to support the understanding of the geographic distribution of environmental agents and health conditions in England and Wales. To facilitate interpretation, the Atlas presents age- and deprivation-adjusted disease risks for males and females separately, with statistical smoothing to adjust for chance fluctuations in disease risk that can occur when using small numbers of cases or small populations (Figure 2). The print version of the Atlas includes additional interpretative text, and a detailed explanation of the statistical methods used.[62] To ensure that the Atlas was useful for the target audience, SAHSU worked closely with the independent charity, Sense about Science (senseaboutscience.org/), who brought together a range of specialists including epidemiologists, health geographers, statisticians, medical doctors, journalists, science communicators, representatives from government organizations, local government as well as interested members of the public. They critically assessed the Atlas material and highlighted issues concerning the display of maps and clarity of content, leading to improvements in presentation and interpretability.

The mission of the US CDC National Environmental Public Health Tracking Program is to provide information from a nationwide network of integrated health and environmental hazard and exposure data to drive actions to

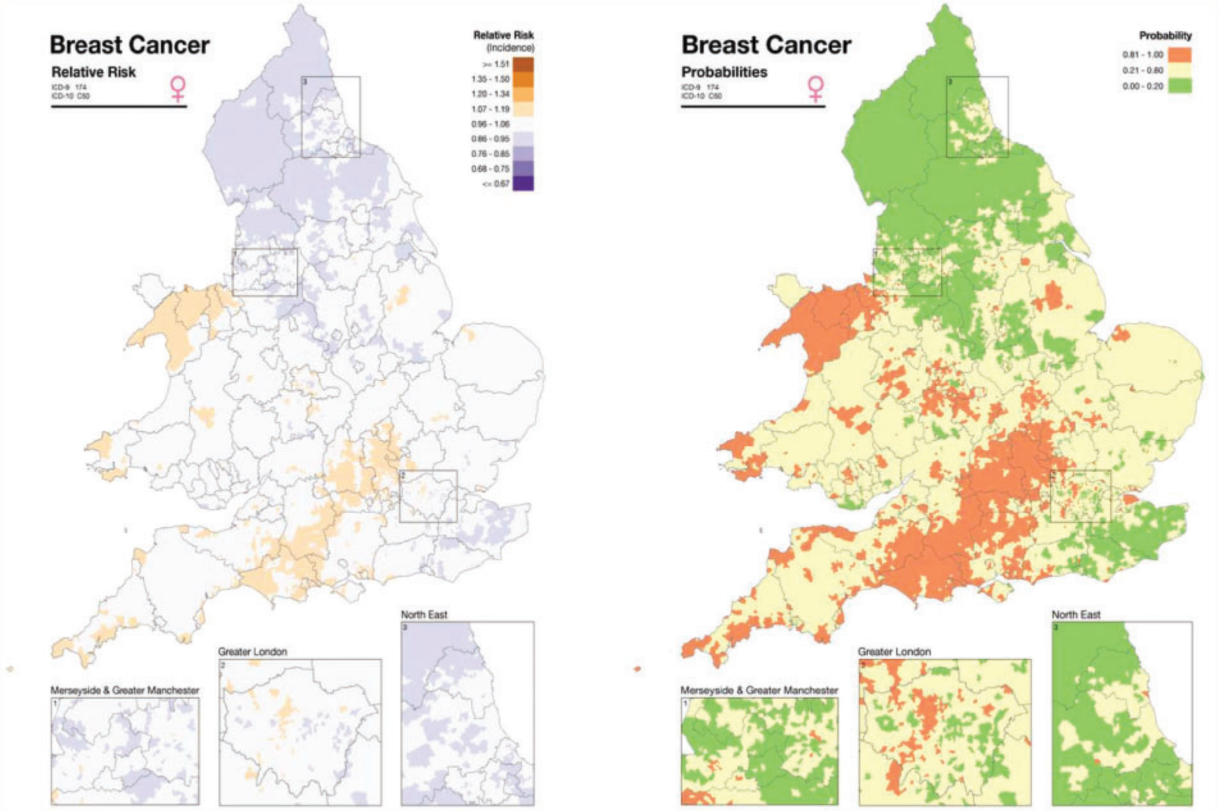

**Figure 2.** Disease map from the Small Area Health Statistics Unit (SAHSU) Environment and Health Atlas for England and Wales (www.envhealthatlas.co.uk). Left-hand side: smoothed relative risk of female breast cancer incidence in England and Wales, adjusted for age and deprivation, 1985–2009; right-hand side: posterior probabilities for female breast cancer incidence in England and Wales, adjusted for age and deprivation, 1985–2009. Contains National Statistics and Ordnance Survey data © Crown copyright and database right 2013. Cancer incidence for Wales was supplied by welsh cancer intelligence and surveillance unit (WCISU).

improve the health of communities.[63,64] In collaboration with partners, the Tracking Program identifies priority environmental health issues, determines key surveillance questions, and evaluates the utility of existing data for answering the question and informing the issue. Selected data are integrated into the National Environmental Public Health Tracking Network and used to (i) describe temporal and spatial trends in disease and potential environmental exposures, (ii) identify populations most affected, (iii) generate hypotheses about associations between health and environmental exposures, and (iv) develop, guide, and assess the environmental public health policies and interventions aimed at reducing health outcomes associated with environmental factors. Gaps in data are addressed by developing standards for new data collections, models, or new methodologies for using existing data, or by expanding the utility of non-traditional public health data. The Tracking Network (https://ephtracking.cdc.gov) permits viewers to explore interactive maps, tables and charts, view information by location (county), and visit state and local tracking websites. The Tracking Program is currently working to improve the spatial resolution (to geographic units smaller than county) of the publicly disseminated data to better address local-level issues. Efforts must balance the need for small-area health data with the need to protect confidentially and produce stable, reliable disease rates. Additionally, the Tracking Program is enhancing the Tracking Network to facilitate the delivery of real-time data to mitigate acute exposures to elevated levels of environmental hazards.

User-generated output, e.g. mash-ups and web applications that combine data from multiple sources and provide additional information and/or functionality, are also supporting the dissemination of small-area data. For instance, the Environmental Research Group of King's College London have developed the London Air app (www.londonair.org.uk), which displays up to date air pollution levels based on measurements taken within the previous hour from monitoring stations that comprise the London Air Quality Network, combined with a detailed model, to show a prediction of air quality at a 20 m resolution across the whole of Greater London.

## Future opportunities and challenges

User-generated inputs, including data from web searches and social media (e.g. for influenza surveillance[65,66]), accelerometer data and geolocations from mobile devices (e.g. for assessing active transport[67]) and physiological and environmental sensor data from wearable tech (e.g. for assessing air pollution exposure[68]) are opening up new data streams for future studies. Most of these datasets are currently underused for epidemiological studies, but combining, for example, time-activity data from travel surveys, Global Positioning Systems (GPS) devices and/or accelerometer data with high-resolution modelled exposure surfaces will permit improved characterization of exposure to spatially and temporally varying risk factors. For example, travel survey data were used in conjunction with estimated micro-environmental concentrations of $PM_{2.5}$, black carbon, and $NO_2$ to characterize air pollution exposures for 89 000 individuals in Hong Kong, with 'dynamic' exposure estimates differing significantly from the 'static' exposure assigned to residential address.[69] Smaller studies have also trialled the use of smart phone based GPS/physical activity data, along with highly spatially and temporally resolved air pollution mapping, e.g. to better understand activities contributing to air pollution exposure in Barcelona.[70] The availability of, for example, accelerometer-measured physical activity in >100 000 participants of the UK Biobank study[71] and >27 000 children in the International Children's Accelerometry Database[72] indicates that these data can be collected at scale.

New data brings new challenges, including demographic bias due to access/availability/trust in digital tools and apps,[73] along with considerations in establishing terms for ethical data use, that will need to be carefully considered.

## Funding

The work of the UK Small Area Health Statistics Unit is funded by Public Health England as part of the MRC-PHE Centre for Environment and Health, funded also by the UK Medical Research Council.

The findings ad conclusions in this article are those of the authors and do not necessarily reflect the views of Public Health England or the Department of Health, nor represent the official position of the Centers for Disease Control and Prevention.

**Conflict of interest:** None declared.

## References

1. Piel F, Fecht D, Hodgson S *et al*. Small-area methods for investigation of environment and health. *Int J Epidemiol* 2020;**49**. doi: 10.1093/ije/dyaa006.
2. Beale L, Abellan JJ, Hodgson S, Jarup L. Methodologic issues and approaches to spatial epidemiology. *Environ Health Perspect* 2008;**116**:1105–10.
3. Sadiq S, Indulska M. Open data: quality over quantity. *Int J Inf Manage* 2017;**37**:150–54.
4. Abiteboul S, Dong L, Etzioni O *et al*. The elephant in the room: getting value from Big Data. In *Workshop on Web and Databases (WebDB)*, ACM Press, 2015.
5. Hendler J, Holm J, Musialek C, Thomas G. US government linked open data: semantic. data. gov. *IEEE Intell Syst* 2012;**27**: 25–31.
6. Yan A, Weber N. Mining open government data used in scientific research. In *International Conference on Information*. Cham, Switzerland: Springer International Publishing, 2018; 303–313.
7. NHS Digital, *Data, Insights and Statistics* September 2018. https://digital.nhs.uk/data-and-information/data-insights-and-statistics. (29 October 2018, date last accessed).
8. ONS. *ONS Statistical Bulletin: Deaths Registered in England and Wales: 2017*. 2018a. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsregistrationsummarytables/2017 (18 July 2018, date last accessed).
9. ONS. *ONS Statistical Bulletin: Births in England and Wales: 2017*. 2018b. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/birthsummarytablesenglandandwales/2017 (18 July 2018, date last accessed).
10. The Cancer Alliance Data, Evidence and Analysis Service (CADEAS*)* 2018. *Cancer Statistics: Availability and Location*. September 2018. http://www.ncin.org.uk/publications/ (29 October 2018, date last accessed).
11. Rowlingson B, Lawson E, Taylor B, Diggle PJ. Mapping English GP prescribing data: a tool for monitoring health-service inequalities. *BMJ Open* 2013;**3**:e001363.
12. Gidlow CJ, Smith G, Martinez D *et al*. Research note: natural environments and prescribing in England. *Landsc Urban Plan* 2016;**151**:103–108.
13. NHS Digital. *Provisional Monthly Hospital Episode Statistics for Admitted Patient Care, Outpatients and Accident and Emergency Data - April 2017 to January 2018: NHS Digital*. 2018 [updated March 13, 2018]. https://digital.nhs.uk/catalogue/PUB30242 (04 March 2018, date last accessed).
14. Bernal JAL, Lu CY, Gasparrini A, Cummins S, Wharam JF, Soumerai SB. Association between the 2012 Health and Social Care Act and specialist visits and hospitalisations in England: a controlled interrupted time series analysis. *PLoS Med* 2017;**14**: e1002427.
15. Bottle A, Goudie R, Bell D, Aylin P, Cowie MR. Use of hospital services by age and comorbidity after an index heart failure admission in England: an observational study. *BMJ Open* 2016; **6**:e010669.
16. Clements C, Turnbull P, Hawton K *et al*. Rates of self-harm presenting to general hospitals: a comparison of data from the Multicentre Study of Self-Harm in England and Hospital Episode Statistics. *BMJ Open* 2016;**6**:e009749.
17. de Hoogh K, Gulliver J, Donkelaar AV *et al*. Development of West-European PM2.5 and NO2 land use regression models incorporating satellite-derived and chemical transport modelling data. *Environ Res* 2016;**151**:1–10.
18. Knibbs LD, Hewson MG, Bechle MJ, Marshall JD, Barnett AG. A national satellite-based land-use regression model for air

pollution exposure assessment in Australia. *Environ Res* 2014; **135**:204–11.

19. Novotny EV, Bechle MJ, Millet DB, Marshall JD. National satellite-based land-use regression: NO2 in the United States. *Environ Sci Technol* 2011;**45**:4407–14.
20. Larkin A, Geddes JA, Martin RV *et al*. Global land use regression model for nitrogen dioxide air pollution. *Environ Sci Technol* 2017;**51**:6957–64.
21. Eeftens M, Beelen R, de Hoogh K *et al*. Development of land use regression models for PM2. 5, PM2. 5 absorbance, PM10 and PMcoarse in 20 European study areas; results of the ESCAPE project. *Environ Sci Technol* 2012;**46**:11195–205.
22. Beelen R, Hoek G, Vienneau D *et al*. Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe–the ESCAPE project. *Atmos Environ* 2013;**72**:10–23.
23. Gulliver J, Elliott P, Henderson J *et al*. Local-and regional-scale air pollution modelling (PM 10) and exposure assessment for pregnancy trimesters, infancy, and childhood to age 15 years: Avon Longitudinal Study of Parents And Children (ALSPAC). *Environ Int* 2018;**113**:10–19.
24. Smith RB, Fecht D, Gulliver J *et al*. Impact of London's road traffic air and noise pollution on birth weight: retrospective population based cohort study. *BMJ* 2017;**359**:j5299.
25. Morley DW, de Hoogh K, Fecht D *et al*. International scale implementation of the CNOSSOS-EU road traffic noise prediction model for epidemiological studies. *Environ Pollut* 2015;**206**: 332–41.
26. Gulliver J, Morley D, Vienneau D *et al*. Development of an open-source road traffic noise model for exposure assessment. *Environ Model Softw* 2015;**74**:183–93.
27. Gan WQ, McLean K, Brauer M, Chiarello SA, Davies HW. Modeling population exposure to community noise and air pollution in a large metropolitan area. *Environ Res* 2012;**116**:11–16.
28. Brink M. Parameters of well-being and subjective health and their relationship with residential traffic noise exposure—a representative evaluation in Switzerland. *Environ Int* 2011;**37**: 723–33.
29. Cai Y, Hansell AL, Blangiardo M *et al*. Long-term exposure to road traffic noise, ambient air pollution, and cardiovascular risk factors in the HUNT and lifelines cohorts. *Eur Heart J* 2017;**38**: 2290–96.
30. Zijlema W, Cai Y, Doiron D *et al*. Road traffic noise, blood pressure and heart rate: pooled analyses of harmonized data from 88, 336 participants. *Environ Res* 2016;**151**:804–13.
31. Cai Y, Hodgson S, Blangiardo M *et al*. Road traffic noise, air pollution and incident cardiovascular disease: a joint analysis of the HUNT, EPIC-Oxford and UK Biobank cohorts. *Environ Int* 2018;**114**:191–201.
32. Cai Y, Zijlema WL, Doiron D *et al*. Ambient air pollution, traffic noise and adult asthma prevalence: a BioSHaRE approach. *Eur Res J* 2017;**49**:1502127.
33. Cohen AJ, Brauer M, Burnett R *et al*. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *Lancet* 2017;**389**:1907–18.
34. Clennon JA, Kamanga A, Musapa M, Shiff C, Glass GE. Identifying malaria vector breeding habitats with remote sensing data and terrain-based landscape indices in Zambia. *Int J Health Geogr* 2010;**9**:58.
35. Hamm NA, Soares Magalhaes RJ, Clements AC. Earth observation, spatial data quality, and neglected tropical diseases. *PLoS Negl Trop Dis* 2015;**9**:e0004164.
36. Subnational Estimates Working Group of the HIV Modelling Consortium. Evaluation of geospatial methods to generate subnational HIV prevalence estimates for local level planning. *AIDS (Lond Engl)* 2016;**30**:1467–74.
37. Johnson K, Brown ME. Environmental risk factors and child nutritional status and survival in a context of climate variability and change. *Appl Geogr* 2014;**54**:209–21.
38. Dadvand P, Nieuwenhuijsen MJ, Esnaola M *et al*. Green spaces and cognitive development in primary schoolchildren. *Proc Natl Acad Sci USA* 2015;**112**:7937–42.
39. Brooker S, Clements AC, Bundy DA. Global epidemiology, ecology and control of soil-transmitted helminth infections. *Adv Parasitol* 2006;**62**:221–61.
40. Messina JP, Kraemer MUG, Brady OJ *et al*. Mapping global environmental suitability for Zika virus. *eLife* 2016;**5**:e15272.
41. Di Cesare M, Khang Y, Ho Asaria P *et al*. Inequalities in non-communicable diseases and effective responses. *Lancet* 2013; **381**:585–97.
42. Carstairs V, Morris R. Deprivation, mortality and resource allocation. *Community Med* 1989;**11**:364–72.
43. Townsend P, Phillimore P, Beattie A. *Health and Deprivation: Inequality and the North*. London: Routledge, 1988.
44. Smith T, Noble M, Noble S, Wright G, McLennan D, Plunkett E. *The English Indices of Deprivation* 2015 *Technical Report*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/464485/English_Indices_of_Deprivation_2015_-_Technical-Report.pdf (29 October 2018, date last accessed).
45. Fecht D, Cockings S, Hodgson S, Piel F, Martin D, Waller L. Advances in mapping population and demographic characteristics at small area levels. *Int J Epidemiol* 2020;**49**(**Suppl 1**): i15–i25.
46. Tatem AJ. WorldPop, open data for spatial demography. *Sci Data* 2017;**4**:170004.
47. Piel FB, Parkes BL, Daby H, Hansell AL, Elliott P. The challenge of opt-outs from NHS data: a small-area perspective. *J Public Health* 2018;**40**:e594–600.
48. McCall C. Opt-out digital health records cause debate in Australia. *Lancet* 2018;**392**:372.
49. Nordfalk F, Hoeyer K. The rise and fall of an opt-out system. *Scand J Public Health* 2018; doi:10.1177/1403494817745189.
50. Rumbold JMM, Pierscionek B. The effect of the General Data Protection Regulation on medical research. *J Med Internet Res* 2017;**19**:e47. doi:10.2196/jmir.7108.
51. Gaye A, Marcon Y, Isaeva J *et al*. DataSHIELD: taking the analysis to the data, not the data to the analysis. *Int J Epidemiol* 2014;**43**:1929–44.
52. Doiron D, de Hoogh K, Probst-Hensch N *et al*. Residential air pollution and associations with wheeze and shortness of breath in adults: a combined analysis of cross-sectional data from two large European cohorts. *Environ Health Perspect* 2017;**125**:097025.
53. Aylin P, Maheswaran R, Wakefield J *et al*. A national facility for small area disease mapping and rapid initial assessment of

apparent disease clusters around a point source: the UK Small Area Health Statistics Unit. *J Public Health Med* 1999;**21**:289–98.

54. Jarup L. Health and environment information systems for exposure and disease mapping, and risk assessment. *Environ Health Perspect* 2004;**112**:995–7.
55. Beale L, Hodgson S, Abellan JJ, Lefevre S, Jarup L. Evaluation of spatial relationships between health and the environment: the rapid inquiry facility. *Environ Health Perspect* 2010;**118**:1306–12.
56. Piel F, Parkes B, Hambly P *et al*. Software application profile: the rapid inquiry facility 4.0: an open access tool for Environmental Public Health Tracking. *Int J Epidemiol* 2020;**49**(**Suppl 1**): i38–i48.
57. Hodgson S, Nieuwenhuijsen MJ, Hansell A *et al*. Excess risk of kidney disease in a population living near industrial plants. *Occup Environ Med* 2004;**61**:717–19.
58. Ferrandiz J, Abellan JJ, Gomez-Rubio V *et al*. Spatial analysis of the relationship between mortality from cardiovascular and cerebrovascular disease and drinking water hardness. *Environ Health Perspect* 2004;**112**:1037–44.
59. Juhasz A, Nagy C, Paldy A, Beale L. Development of a Deprivation Index and its relation to premature mortality due to diseases of the circulatory system in Hungary, 1998-2004. *Soc Sci Med* 2010;**70**:1342–49.
60. Ball W, LeFevre S, Jarup L, Beale L. Comparison of different methods for spatial analysis of cancer data in Utah. *Environ Health Perspect* 2008;**116**:1120–24.
61. Holowaty EJ, Norwood TA, Wanigaratne S, Abellan JJ, Beale L. Feasibility and utility of mapping disease risk at the neighbourhood level within a Canadian public health unit: an ecological study. *Int J Health Geogr* 2010;**9**:21.
62. Hansell AL, Ghosh RE, Fecht D, Fortunato L. *The Environment and Health Atlas for England and Wales*. Oxford: Oxford University Press, 2014.
63. McGeehin MA, Qualters JR, Niskar AS. National environmental public health tracking program: bridging the information gap. *Environ Health Perspect* 2004;**112**:1409–13.
64. Qualters JR, Strosnider HM, Bell R. Data to action: using environmental public health tracking to inform decision making. *J Public Health Manag Pract* 2015;**21**(**Suppl 2**):S12–22.
65. Corley CD, Cook DJ, Mikler AR, Singh KP. Text and structural data mining of influenza mentions in web and social media. *IJERPH* 2010;**7**:596–615.
66. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS. Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. *PLoS Comput Biol* 2015;**11**:e1004513.
67. Griffin GP, Jiao J. Where does bicycling for health happen? Analysing volunteered geographic information through place and plexus. *J Transp Health* 2015;**2**:238–47.
68. Dons E, Laeremans M, Orjuela JP *et al*. Wearable sensors for personal monitoring and estimation of inhaled traffic-related air pollution: evaluation of methods. *Environ Sci Technol* 2017;**51**: 1859–67.
69. Tang R, Tian L, Thach T-Q *et al*. Integrating travel behavior with land use regression to estimate dynamic air pollution exposure in Hong Kong. *Environ Int* 2018;**113**:100–108.
70. De Nazelle A, Seto E, Donaire-Gonzalez D *et al*. Improving estimates of air pollution exposure through ubiquitous sensing technologies. *Environ Pollut* 2013;**176**:92–99.
71. Doherty A, Jackson D, Hammerla N *et al*. Large scale population assessment of physical activity using wrist worn accelerometers: the UK Biobank Study. *PLoS One* 2017;**12**: e0169649.
72. Cooper AR, Goodman A, Page AS *et al*. Objectively measured physical activity and sedentary time in youth: the International children's accelerometry database (ICAD). *Int J Behav Nutr Phys Act* 2015;**12**:113.
73. Cesare N, Grant C, Hawkins JB, Brownstein JS, Nsoesie EO. Demographics in social media data for public health research: does it matter? arXiv Preprint arXiv:171011048. 2017.