SSOAR

Open Access Repository
www.ssoar.info

Evidence from big data in obesity research: international case studies

Wilkins, Emma; Aravani, Ariadni; Downing, Amy; Drewnowski, Adam; Griffiths, Claire; Zwolinsky, Stephen; Birkin, Mark; Alvanides, Seraphim; Morris, Michelle A.

Postprint / Postprint
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Wilkins, E., Aravani, A., Downing, A., Drewnowski, A., Griffiths, C., Zwolinsky, S., ... Morris, M. A. (2020). Evidence from big data in obesity research: international case studies. *International Journal of Obesity*, *44*, 1028-1040. https://doi.org/10.1038/s41366-020-0532-8

Nutzungsbedingungen:
Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:
This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.
By using this particular document, you accept the above-stated conditions of use.

gesis Leibniz-Institut für Sozialwissenschaften

Mitglied der Leibniz-Gemeinschaft

Diese Version ist zitierbar unter / This version is citable under:
https://nbn-resolving.org/urn:nbn:de:0168-ssoar-70720-3

This is a post-peer-review, pre-copyedit version of an article published in
International Journal of Obesity, Vol. 44, 2020, pp. 1028-1040
The final authenticated version is available online at: https://doi.org/10.1038/s41366-020-0532-8
Page numbers have been adjusted to the publishers version, whereby this postprint is fully quotable.

Evidence from big data in obesity research: international case studies

Emma Wilkins[1] · Ariadni Aravani[1] · Amy Downing[1] · Adam Drewnowski[2] · Claire Griffiths[3] · Stephen Zwolinsky[3] · Mark Birkin[4] · Seraphim Alvanides[5,6] · Michelle A. Morris[1]

Received: 23 May 2019 / Revised: 20 December 2019 / Accepted: 7 January 2020 / Published online: 27 January 2020

Abstract

Background/objective Obesity is thought to be the product of over 100 different factors, interacting as a complex system over multiple levels. Understanding the drivers of obesity requires considerable data, which are challenging, costly and time-consuming to collect through traditional means. Use of 'big data' presents a potential solution to this challenge. Big data is defined by Delphi consensus as: *always digital*, has a large sample size, and a large volume or variety or velocity of variables that require additional computing power (Vogel et al. Int J Obes. 2019). 'Additional computing power' introduces the concept of big data analytics. The aim of this paper is to showcase international research case studies presented during a seminar series held by the Economic and Social Research Council (ESRC) Strategic Network for Obesity in the UK. These are intended to provide an in-depth view of how big data can be used in obesity research, and the specific benefits, limitations and challenges encountered.

Methods and results Three case studies are presented. The first investigated the influence of the built environment on physical activity. It used spatial data on green spaces and exercise facilities alongside individual-level data on physical activity and swipe card entry to leisure centres, collected as part of a local authority exercise class initiative. The second used a variety of linked electronic health datasets to investigate associations between obesity surgery and the risk of developing cancer. The third used data on tax parcel values alongside data from the Seattle Obesity Study to investigate socio-demographic determinants of obesity in Seattle.

Conclusions The case studies demonstrated how big data could be used to augment traditional data to capture a broader range of variables in the obesity system. They also showed that big data can present improvements over traditional data in relation to size, coverage, temporality, and objectivity of measures. However, the case studies also encountered challenges or limitations; particularly in relation to hidden/unforeseen biases and lack of contextual information. Overall, despite challenges, big data presents a relatively untapped resource that shows promise in helping to understand drivers of obesity.

Introduction

Obesity is a complex health, social and economic challenge. It is widely recognised as a product of numerous multi-level factors, including individual, social, economic, environmental and political influences [1–3]. This complexity is represented in the Foresight Obesity System Map [4], which lists 108 contributing factors, depicted as nodes in a system diagram. It is also reflected in the multi-disciplinary nature of obesity research, which covers disciplines as diverse as medicine, public health, geography and computer science. Whole systems approaches, which intervene across these multiple levels and domains, have been touted as a way to tackle the growing problem of obesity [5]. Understanding the drivers of obesity and responses to interventions within such a complex system

✉ Michelle A. Morris
m.morris@leeds.ac.uk

[1] Leeds Institute for Data Analytics and School of Medicine, University of Leeds, Leeds, UK
[2] Center for Public Health Nutrition, University of Washington, Seattle, WA, USA
[3] School of Sport, Leeds Beckett University, Leeds, UK
[4] Leeds Institute for Data Analytics and School of Geography, University of Leeds, Leeds, UK
[5] Engineering and Environment, Northumbria University, Newcastle, UK
[6] GESIS—Leibniz Institute for the Social Sciences, Cologne, Germany

requires considerable data. Use of 'big data' and associated analytics, presents a potential solution to this challenge.

Various definitions of 'big data' have been adopted [6–8]. In this paper, we adopt a definition of 'big data' established by a recent Delphi survey of international obesity and big data experts [9], which agreed that, in contrast to traditional data, big data:

> *"is always digital, has a large sample size, and a large volume or variety or velocity of variables that require additional computing power. It can include quantitative, qualitative, observational or interventional data from a wide range of sources (e.g. government, commercial, cohorts) that have been collected for research or other purposes, and may include one or several datasets. Specialist skills in computer programming, database management and data science analytics are usually required to analyse big data."*

According to the Delphi survey, 'big data' can include not only 'novel' types of data such as social media, loyalty cards and sensors, but also routinely collected data, such as health records, government and census data.

The Economic and Social Research Council (ESRC) Strategic Network for Obesity ('the Network') was established to consider the use of big data in obesity research [10]. Several outputs from the Network, which form part of this paper series, have demonstrated that research applications using big data, and associated analytics, within obesity research are rich and diverse. Timmins et al. [11] report a wide range of studies already using big data in obesity research. They reveal that big data could provide many benefits such as increased scope and objectivity, access to unreached populations, and the potential to evaluate real-world interventions. Big data and big data analytics have also been used to produce innovative data visualisation tools, with demonstrable policy impact [12]. Looking to the future, a mapping exercise [13] demonstrated that big data sources can provide information spanning almost 80% of the nodes in the Foresight Obesity System Map. The remainder of the nodes could be covered by more traditional data sources, demonstrating how synergy of big and traditional data can support whole systems approaches to obesity.

Big data also has limitations, such as concerns around data validity and representativeness [11], which need to be balanced alongside benefits. Challenges exist around ethics, data governance, data handling and processing capabilities [6, 9, 14]. Consistent reporting of data sources, such as through the use of the recently developed BEE-COAST framework [13], better enables critique of these strengths and limitations.

Applications of big data in obesity research include use of retail sales data to evaluate the impact of obesity policy [15], use of geotagged social media data to explore patterns in obesity-related behaviours [16, 17], and the use of smartphone data to assess physical activity patterns over space and time [18, 19]. These examples draw on data from diverse sectors, highlighting again the multi-disciplinary nature of obesity research.

Uses of big data include both hypothesis generation ('exploratory analyses') and hypothesis testing. Recognising the distinction between these two forms of enquiry is important to avoid hypothesising after the results are known [20]. This may be particularly problematic in the case of big data research, as large sample sizes, coupled with repeated exploratory analyses, will lead to increased chance of detecting statistically significant associations that are of limited clinical and practical importance.

The aim of this paper is to showcase international research case studies presented during seminars held by the Network in the UK [10]. These are intended to complement existing high-level reviews of big data in obesity research [11, 13] by providing an in-depth view of how big data can be used in this field, and the specific benefits, limitations and challenges encountered.

Methods and results

Three case studies presented at the Network Seminar Series [10] are reported. Each employed several sources of data, including 'big' and 'traditional' data to measure obesity-related exposures and/or outcomes. These data are reported using a standardised BEE-COAST framework [13] that cross references to the Foresight Obesity System Map nodes [4] highlighting the breadth of data coverage (Table 1).

Table 2 summarises all Network seminar presentations. Further information and seminar recordings can be found at https://www.cdrc.ac.uk/research/obesity/.

Case Study 1: Uptake of physical activity in Leeds, UK

Griffiths and Zwolinsky, Seminar: May 2016, London School Hygiene Tropical Medicine

Background Physical activity can help prevent and man-age a number of chronic health conditions, including obesity [21, 22]. The World Health Organisation [23], and other bodies internationally [24, 25] have called upon authorities to increase opportunities for physical activity as a means to tackle obesity. Repurposing existing 'big' spatial data on the physical activity environment provides novel opportunities to support policy.

Data Leeds Let's Get Active Programme, Points of Interest (Table 1).

Table 1 BEE-COAST framework reporting the data used in our case studies.

Case Study 1: Uptake of physical activity in Leeds, UK		
Data: Leeds Let's Get Active (LLGA)		
Background	Key features	This data set consists of two elements. Firstly, a local population level survey of sixty thousand adult residents to determine current levels of physical activity and sedentary behaviour. Secondly, attendance data at local authority leisure centres from individuals completing the physical activity survey.
	History	LLGA was originally commissioned by Sport England as part of their 'get healthy, get into sport' funding stream. In addition, it received match funding from Leeds Health and Wellbeing Board. Open to all adults in the area, LLGA is a community-based physical activity intervention that encourages inactive residents to be more active. Participants signing-up get free unlimited access to around 150 h of timetabled sessions each week at Leeds based local authority leisure centres.
	Purpose	To assess physical activity levels and attendance rates at LLGA.
Elements	Content	Demographics: • Registration data • Member number • Gender • Year of birth • Post code of residence Physical activity data: The following variables were collected using the short form International Physical Activity Questionnaire (sIPAQ) [51], an instrument which was designed for population surveillance of physical activity among adults: • Metabolic equivalent (METs)—energy expenditure minutes of physical activity per week • Vigorous intensity physical activity (minutes per week) • Moderate intensity physical activity (minutes per week) • Light intensity physical activity (minutes per week) • Walking (minutes per week) • Sedentary time (minutes per weekday) Attendance data: • Venue • Type of session attended (e.g. swimming or gym class) • Timestamp
	Ownership	Leeds City Council
	Aggregation	Data are at the level of an individual.
	Sharing	These data may be accessed through Leeds City Council's Sport and Active Lifestyles department. Data are controlled by this group and access is determined on a project by project basis.
	Temporality	The sIPAQ was collected on a cross sectional basis at the point of registration. LLGA attendance data are collected every time a member accesses an LLGA session through a local authority swipe card system. Data used in Case Study 1 were collected between September 2013 and July 2015.
Exemplars	Indicative use cases	There are many examples of sIPAQ data being used to determine physical activity and sedentary behaviour levels within epidemiological research, both nationally and internationally. For example, sIPAQ data have been used to understand clustering of health behaviours [52] and correlates of sedentary behaviour [53]. To the best of our knowledge, these data have not been used in conjunction with efforts to investigate associations with objective measures of physical activity opportunities. A small number of studies have used swipe card data to measure leisure centre attendance [54, 55]. However, to our knowledge, none have used swipe card data to investigate associations between leisure centre proximity and attendance.
	Foresight nodes	3.1 Physical activity 3.4 Level of recreational activity 4.3 Access to opportunities for physical exercise
Data: Points of Interest		
Background	Key features	Points-of-Interest is a dataset detailing over 4 million geographic features (both natural and built) across Great Britain.
	History	The dataset is created and maintained by PointX Ltd on behalf of Ordnance Survey, the national mapping agency of Great Britain. PointX is an independent company jointly owned by Ordnance Survey and Landmark Information Group. Points-of-Interest data have been available since 2000, and are updated quarterly (see below).
	Purpose	Points-of-Interest data were developed for the purpose of mapping features of public interest in Great Britain. They have various uses including both administrative (e.g. service provision and emergency planning) and commercial (e.g. driver routing and location based services).
Elements	Content	The scope of features covered is broad, including commercial services, education and healthcare establishments, transportation infrastructure, attractions, and public infrastructure. Of particular relevance to the obesity system, the dataset contains information on food outlets (various classifications), public transportation nodes (e.g. bus stops), formal green spaces (e.g. commons and parks), and sport and recreational facilities. For each feature, the following data are available: • Unique reference number • Feature name • Feature classification (600 classifications available) • Feature address • Feature location (British National Grid coordinates) • Positional accuracy of feature location • Unique property reference number (allows linkage to Ordnance Survey Address Base suite of products) • Topographic ID and version Identifier (allows linkage to Ordnance Survey MasterMap Topography Layer product). • ITN easting and northing (allows linkage to OS MasterMap ITN layer) • Telephone number and/or web address
	Ownership	Ordnance Survey
	Aggregation	Data are available at the level of individual features.
	Sharing	Points-of-Interest data can be accessed for free online via the EDINA Digimap website using an educational institution login. However, use of the data via this means is restricted to 'educational use' and/or limited 'administrative use', as defined by Ordnance Survey's end user agreement. Data can be shared with others who have entered into the end user agreement/a data handlers' agreement with Ordnance Survey. Less restrictive access to the data can be purchased at a cost.

Table 1 (continued)

Data: Points of Interest		
	Temporality	A new version of Points of Interest is released every quarter. EDINA Digimap hold previous versions of Points of Interest back to March 2015. With each new release, Ordnance Survey publish details on the changes that have been made as compared with the previous release. Note, feature classification codes have also changed over time (last update at time of writing: January 2013). The data used in the present case study were from 2011.
Exemplars	Indicative use cases	Points of Interest has been used to characterise access to local amenities relating to diet and physical activity such as food outlets [56], and sport and recreational facilities [57].
	Foresight nodes	4.2 Opportunity for team based activity 4.3 Access to opportunities for physical exercise 4.6 Reliance on labour saving devices and services 4.9 Opportunity for un-motorised transport 4.11 Dominance of motorised transport 4.13 Walkability of living environment 7.4 Food exposure, 7.5 Food abundance, 7.7 Convenience of food offerings, 7.8 Food variety
Case Study 2: Obesity and Colorectal Cancer in England, UK		
Data: Hospital Episode Statistics (HES)		
Background	Key features	HES is a dataset that contains details of hospital inpatient admissions, outpatients appointments and accident and emergency (A&E) attendances, covering all National Health Service (NHS) trusts in England (including acute hospitals, primary care trusts and mental health trusts)
	History	HES is collated by NHS Digital. NHS Digital is the national provider of information, data and IT systems for commissioners, analysts and clinicians in health and social care in England.
	Purpose	HES data are collected during a patient's time at hospital. It is an administrative dataset which allows hospitals to be paid for the care they deliver. HES data can also be used for non-clinical purposes (secondary use), including research.
Elements	Content	The data collected include patient information, clinical details and administrative details, including: <u>Patient information:</u> • Age • Sex • Ethnicity • Residence location <u>Clinical details:</u> • Diagnoses • Operative procedures • Consultant information • Specialty information <u>Administrative details:</u> • NHS trust • General practitioner • Admission dates • Discharge dates • Method of admission • Referrer
	Ownership	NHS digital
	Aggregation	Data are collected on a patient level.
	Sharing	Through application to NHS Digital.
	Temporality	HES extracts are taken from the Secondary Uses Service database (a secure data repository of healthcare data in England) on a monthly basis, at pre-arranged dates during the year.
Exemplars	Indicative use cases	HES provides data for the purpose of healthcare analysis to the NHS, government and others including: • National bodies and regulators, such as the Department of Health, NHS England, Public Health England, NHS Improvement and the Care Quality Commission. • Local Clinical Commissioning Groups. • Provider organisations. • Government departments. • Researchers and commercial healthcare bodies. • National Institute for Clinical Excellence (NICE). • Patients, service users and carers. • The media. HES statistics are known to be used for: • National policy making. • Benchmarking performance against other hospital providers or Clinical Commissioning Groups. • Academic research, such as investigating trends over time in obesity surgery rates [58] and obesity-related hospital admissions among children and adolescents [59]. • Analysing service usage and planning change. • Providing advice to ministers and answering a wide range of parliamentary questions. • National and local press articles. • International comparison.
	Foresight nodes	2.1 Self-esteem, 2.4 Stress, 7.9 Alcohol consumption.
Data: National Cancer Registration and Analysis Service (NCRAS)		
Background	Key features	The NCRAS dataset is a collection of data items relating to cancer diagnosis and treatment.
	History	NCRAS is a service that manages the collection of cancer registration data. Every year, NCRAS collects information on over 300,000 cases of cancer.
	Purpose	To build a complete picture of the incidence and prevalence of cancer in England, as well as understanding how cancer patients are diagnosed, treated and their outcomes.

Table 1 (continued)

Data: National Cancer Registration and Analysis Service (NCRAS)		
Elements	Content	The data collected include: • Patient details such as name, address, age, sex, ethnicity. • Details on type of cancer. • How advanced a cancer case is. • Details of treatment a patient underwent.
	Ownership	Public Health England
	Aggregation	Data are collected on a tumour level
	Sharing	Data can be obtained via the Public Health England Office for Data Release or from the NHS Digital Data Access Request Service. In the present case study, NCRAS data were obtained from the NHS Digital Data Access Request Service who performed linkage with the HES and Office for National Statistics mortality data.
	Temporality	NCRAS data are refreshed yearly.
Exemplars	Indicative use cases	Examples of how NCRAS data are used to support cancer epidemiology, public health and research are: • Monitoring how many people are diagnosed with cancer • Improving cancer care through feedback of data to the clinical community • Supporting cancer research via investigation of possible causes of cancer and effectiveness of treatments Research applications include building projections of future cancer incidence and mortality rates [60] and examine mortality rates after cancer treatment [61].
	Foresight nodes	Not applicable
Data: Office For National Statistics Mortality Data		
Background	Key features	The Office for National Statistics is a non-ministerial government department which produces national statistics on a variety of topics, such as health, economy and crime, for the UK government. One such area is data on mortalities across England and Wales.
	History	The Office for National Statistics was formed in 1996, and since that time has collected data on the UK population via population census, surveys, and analysis of data generated by businesses and organisations such as the National Health Service and the register of births, marriages and deaths.
	Purpose	The purpose of the data collected by the Office for National Statistics is to inform policymaking, enable tracking of population changes over time, and international comparisons.
Elements	Content	The Office for National Statistics mortality data contains information taken from the death certificate for all deaths registered in England and Wales. This includes: • Cause of death • Date of death • Place of death
	Ownership	Office for National Statistics
	Aggregation	Data available at the level of individuals
	Sharing	Data can be obtained directly from the Office for National Statistics. However, in the present case study, the data were obtained pre-linked with HES and NCRAS data from the NHS Digital Data Access Request Service.
	Temporality	Data are released monthly in provisional format (without quality assurance and subject to change due to e.g. delays in death registration) and yearly in finalised format.
Exemplars	Indicative use cases	Office for National Statistics mortality data have been used in a wide variety of research contexts e.g. to explore changes in mortality rates over time [62], or between populations [63].
	Foresight nodes	Not applicable
Case study 3: Sociodemographic determinants of obesity in Seattle, USA		
Data: Seattle Obesity Studies I, II and III (SOS I, II, III)		
Background	Key features	A population-based health survey of adults living in King County, Washington State, USA.
	History	The original SOS study (SOS I) is a telephone-based survey of 2,001 adults living in King County. Stratified random sampling was used to select telephone numbers for residential properties in King County. Zip codes with a high percentage of low-income or ethnic minority residents were over-sampled. A pre-notification letter was sent to selected households, which was followed up with a telephone call inviting one household member to complete a 20 min telephone survey (where there were multiple household members ≥18 years, one was selected at random). Surveys took place between 2008 and 2009. Survey questions closely replicated those used in the national Behavioural Risk Factor Surveillance System (BRFSS). Participants were also invited to complete a Food Frequency Questionnaire, which was mailed to participants by post. The SOS II and III was a prospective cohort study comprising two waves (wave 1: SOS II; wave 2: SOS III; 1 year apart) conducted between 2011 and 2013. A new sample of 516 King County adults was recruited through similar methods as for SOS I, with the exception that sampling was stratified based on tax parcel property values rather than zip code income and ethnicity. Survey questions were similar to those in SOS I. However, surveys were administered through in-person interviews with objective measures of height and weight collected. Participants also wore a GPS tracker to record travel patterns in time and space, and an accelerometer to measure physical activity (SOS III only). Further information is available at Aggarwal, Monsivais et al. [64] and Drewnowski, Aggarwal et al. [65].
	Purpose	The Seattle Obesity Study (SOS) was designed to addresses the socioeconomic and environmental determinants of health inequities, with a focus on obesity.
Elements	Content	Collectively the data contain information on: • Age • Gender • Ethnicity • Number of children <12 years • Number of children between 12 and 18 years

Table 1 (continued)

Case study 3: Sociodemographic determinants of obesity in Seattle, USA		
Data: Seattle Obesity Studies I, II and III (SOS I, II, III)		
		• Household size • Education • Employment • Annual household income • Home ownership • Height • Weight • Self-rated health • CVD • Diabetes • Smoking status • Travel diary • Self-reported physical activity outside work • Diet (food frequency questionnaire of intakes and portion sizes of ~150 foods and beverages). • Dietary and food shopping behaviours (SOS II and III only) • GPS (SOS II and III only) • Accelerometry (SOS III only)
	Ownership	University of Washington, Seattle, Washington, USA
	Aggregation	Data are available at the level of individual participants.
	Sharing	Available on a case-by-case basis upon request to the University of Washington.
	Temporality	SOS I data were collected between 2008 and 2009. SOS II and III data were collected between 2011 and 2013, with wave 2 (SOS III) conducted 1 year after wave 1 (SOS I).
Exemplars	Indicative use cases	The SOS data have been used in a variety of studies investigating built and social inequalities, health and obesity [39].
	Foresight nodes	1.1 Education, 1.5 Sociocultural valuation of food, 1.16 Smoking cessation, 3.1 Physical activity, 3.4 Level of recreational activity, 3.5 Level of domestic activity, 3.6 Level of occupational activity, 3.7 Level of transport activity, 4.11 Dominance of motorised transport, 4.12 Dominance of sedentary employment, 4.13 Walkability of living environment, 5.15 Predisposition to physical activity, 7.1 force of dietary habits, 7.3 tendency to graze, 7.8 Food variety, 7.9 Alcohol consumption, 7.12 Fibre content of food and drink, 7.13 Portion size, 7.14 Demand for convenience, 7.16 Nutritional Quality of food and drink.
Data: King County Tax Parcel Values		
Background	Key features	This dataset includes the estimated market value of tax parcels within King County (Washington, USA) as assessed by the King County tax assessor.
	History	The state of Washington imposes a property tax based upon assessed tax parcel values. Tax parcels are plots of land, often containing one or more residential units. Tax parcel values are determined in King County every 6 years by the county tax assessor. Valuations are based on the combined value of both land and any improvements attached to the land (such as drive-ways, buildings etc.). Valuations aim to estimate the market value of the land and improvements, taking into account comparable bare land sales, building square footage, year built, and other property characteristics. The assessed value per parcel is the sum of a parcel's land and improvement values.
	Purpose	The data are primarily generated for the purpose of determining property taxes.
Elements	Content	The data includes information on the estimated market value of each tax parcel, and the number of residential units per parcel.
	Ownership	King County
	Aggregation	Data are available at the level of tax parcels.
	Sharing	The data are publicly available.
	Temporality	Land parcel values are re-assessed every 6 years.
Exemplars	Indicative use cases	This is the first known use of King County tax assessor property values within health research.
	Foresight nodes	Property values appear to indirectly capture the social and economic context and composition of a neighbourhood, which might capture diverse nodes of the Foresight map, including: 1.1 Education, 1.2 Acculturation, 1.5 Sociocultural valuation of food, 1.7 Social acceptability of fatness, 1.15 Social rejection of smoking, 6.11 Desire to minimise cost, 6.13 Market price of food offerings, 7.14 Demand for convenience.

HES hospital episode statistics, *LLGA* Leeds Let's Get Active, *NCRAS* National Cancer Registration and Analysis Service, *NHS* National Health Service, *sIPAQ* short form International Physical Activity Questionnaire; *SOS* Seattle Obesity Study; *GPS* Global Positioning System.

Methods Links with Leeds City Council facilitated secondary analysis of data emerging from the Leeds Let's Get Active (LLGA) programme; a council initiative to increase physical activity through exercise classes delivered at leisure centres. Exploratory, cross-sectional analyses investigated (i) the association between the number of neighbourhood physical activity opportunities and separate outcomes of sedentary behaviour and physical activity, controlling for neighbourhood deprivation, and (ii) whether residential proximity to participating leisure centres was related to attendance. Physical activity opportunities were derived from Points-of-Interest data; a large dataset detailing the locations of a wide range of features across the whole of Great Britain.

Participant postcodes were analysed in a Geographic Information System together with data on the locations of physical activity opportunities from Points-of-Interest data and the locations of participating leisure centres. Physical activity opportunities separately included (i) green spaces and (ii) built facilities such as gyms, climbing facilities, and swimming pools. Neighbourhoods were defined using

Table 2 Summary of Network Seminar Series Presentations.

Presenter(s)	Presentation title
Seminar 1: Policy, Impact and Data (Leeds, 15 November 2015)	
Prof Pinki Sahota (Association for Obesity; Leeds Beckett University)	Using big data to tackle obesity: perspective from the Association for Obesity.
Mr Michael Chang (Town and Country Planning Association)	Town and Country Planning Association: reuniting health with planning and links to the Network.
Prof Mark Birkin (Consumer Data Research Centre; University of Leeds)	The Consumer Data Research Centre
Prof Jamie Pearce (Administrative Data Research Centre; University of Edinburgh)	Network for Obesity
Seminar 2: Data, Methods and Models (Cambridge, 16 March 2016)	
Dr James Woodcock (University of Cambridge) and Dr Robin Lovelace (University of Leeds)	Modelling and visualising large and complex datasets to guide active travel policies: a case study from the Propensity to Cycle Tool
Dr Darren Greenwood (University of Leeds)	Interpreting results from analysis with big data: examples from epidemiology
Prof Adam Drewnowski (University of Washington)	Big Data and the obesity epidemic
Seminar 3: Novel Results and Visualisation (London, 18 May 2016)	
Dr Pablo Monsivais (Washington State University)	Data visualisation: Why bother?
Dr Claire Griffiths (Leeds Beckett University)	Physical activity, sedentary behaviour and the environment—the importance of using local level data to inform local level decisions
Prof Jaap Seidell (Vrije Universiteit Amsterdam)	Integrated approaches for childhood obesity: the Dutch experience
Seminar 4: Application and Policy (Edinburgh, 13th September 2016)	
Prof Nanette Mutrie MBE (University of Edinburgh)	How big(gish) data were used in informing Scottish physical activity policy
Mrs Lorraine Tulloch (Obesity Action Scotland)	The killer stat, the elevator pitch: Using data to influence obesity prevention
Prof Paul Gately (Leeds Beckett University and MoreLife)	Using data to drive improvements in weight management
Seminar 5: Opportunities and Challenges (Leeds, 25 April 2017)	
Prof Alex Singleton and Dr Mark Green (University of Liverpool)	Consumers in context: Consumer Data Research Centre indicators and applications for health
Prof Mark Gilthorpe (University of Leeds)	Methodological challenges in the analysis of large and complex (big) data: a causal inference perspective.
Dr Amy Downing and Mrs Ariadni Aravani (University of Leeds)	Obesity surgery and risk of colorectal and other obesity-related cancers in England: the challenges of using routinely collected data

'Lower Super Output Area' (LSOA) boundaries (a UK administrative geography containing ~1500 people) and 2 km circular buffers.

Results LLGA data contained 29,796 self-reports of physical activity and sedentary behaviours, together with leisure centre attendance data from swipe cards. Analyses revealed no associations between any measure of physical activity opportunities and physical activity or sedentary behaviours, with the exception of counts of green spaces within LSOAs. Those with the highest counts of green spaces within LSOAs were more likely to meet physical activity guidelines.

Fewer than 50% of participants who registered for the programme attended a session. Of those that did, over one third did not attend the centre closest to them. Having a leisure centre within the residential Middle Super Output Area (an administrative geography containing ~8000 people) or a 2 km circular buffer only accounted for a small proportion of the variability in attendance rates. On further investigation, circular buffers of at least 4 km around leisure centres were required to capture over 50% of attendees.

Conclusion There is some indication that neighbourhood greenspace is related to physical activity. However, in agreement with other literature [26, 27], this study shows different definitions of environment can produce different results. Future work must use measures that are relevant, consistent and transferable. Mere proximity to opportunities from home may not be a good indicator of actual exposure/opportunities. People frequently visit leisure centres that are not closest to home.

Case Study 2: Obesity and Colorectal Cancer in England, UK

Aravani and Downing, Seminar: April 2017, Leeds Beckett University

Background Obesity is linked to an increased risk of several malignancies, including colorectal cancer [28, 29]. Counterintuitively, some research suggests surgery to reduce obesity ('obesity surgery') may increase the risk of colorectal cancer [30–33]. However, this association remains unclear, with the majority of studies having short follow-up time or lacking statistical power. This study tested the a-priori hypothesis that obesity surgery is associated with the risk of colorectal cancer and also explored associations with other obesity-related cancers (breast, kidney or endometrial) across the English National Health Service (NHS).

Data Hospital Episode Statistics (HES), National Cancer Registration and Analysis Service (NCRAS), Office for National Statistics (ONS) mortality data (Table 1).

Methods This was a national population-based retrospective observational study. Individuals who underwent obesity surgery (the 'OS group') or had a hospital episode with a diagnosis of obesity but no obesity surgery (the 'no-OS

group'), between April 1997 and September 2013, were identified using HES data. HES data were obtained pre-linked with NCRAS and ONS mortality data. This allowed the identification of individuals in the OS and no-OS groups who were subsequently diagnosed with colorectal cancer, or other obesity-related cancers. It also allowed the identification of the time 'at risk'—the time from obesity diagnosis/surgery to development of a cancer, death or last follow-up (30 September 2013). Standardised incidence ratios (SIR) with 95% confidence intervals (CI) were calculated to define the risk of developing cancer in the OS and no-OS groups relative to the background English population, accounting for age and calendar year.

Results A total of 1,002,607 obese patients were identified, of whom 4% (*n* =39 747) underwent obesity surgery. The OS group and no-OS groups had a median follow-up period of 3 years (range 1–16 years) and 2.5 years (range 1–16 years), respectively. In the no-OS cohort, 3237 developed colorectal cancer with an SIR of 1.12 (95%CI 1.08–1.16) relative to the background population. In the OS cohort, 43 developed colorectal cancer with an SIR of 1.26 (95%CI 0.92–1.71). There was a significantly increased risk of colorectal cancer among the oldest (≥50 years) in the OS group (SIR: 1.47, 95% CI: 1.02–2.06). High SIRs for renal and endometrial cancers were found in both the OS and non-OS groups [34]. By contrast, OS was associated with reduced breast cancer risk [34].

Conclusion Although the association between obesity surgery and subsequent colorectal cancer risk was limited by the small OS group size and short follow-up time, this study showed an elevated colorectal cancer risk continues after obesity surgery in individuals older than 50 years. The high SIRs for renal and endometrial cancers require further investigation.

Case Study 3: Sociodemographic determinants of obesity in Seattle, USA

Drewnowski, Seminar March 2016, University of Cambridge

Background Socioeconomic status (SES), both at the individual and neighbourhood level is thought to contribute to obesity. However, studies of obesity and its determinants often do not contain important socioeconomic variables or include only self-reported measures, which are simplistic and subject to bias. Neighbourhood measures of SES are often only available for administrative geographies, which are subject to bias from the modifiable areal unit problem (MAUP) [35] and may not be suited to capturing neighbourhood effects on obesity [36]. A series of exploratory studies were conducted to examine whether residential property values—the second largest source of wealth in the US [37]—could be used as a proxy measure of individual and neighbourhood level SES, and to simulate obesity prevalence at a micro-scale.

Data Seattle Obesity Study (SOS) I, II and III, King County Tax Parcel Values (Table 1).

Methods Data from the Seattle Obesity Study (SOS) I, II and III were used to assess associations between socio-economic variables and health-related outcomes, including diet and obesity. Participants' residential addresses were geocoded to tax parcel centroids; plots of land owned by a single landowner and typically containing a single residential property or a block of properties e.g. flats. In a series of studies, SOS participants were ascribed individual and neighbourhood measures of SES based on King County Tax Parcel Values. Individual SES was operationalised as the average property value in the tax parcel of residence. Neighbourhood SES was operationalised as the average property value in the residential neighbourhood (various definitions including residential census tracts and home-centric buffers spanning multiple tax parcels). Multivariable linear regressions examined associations between these measures and obesity-related outcomes, including behaviours, diet quality (e.g. measures of soda and salad consumption) and obesity, controlling for age, gender, and race/ethnicity. This was contrasted with the performance of more traditional measures of SES, including education and income, at predicting obesity-related outcomes.

Results Obesity-related outcomes were related both with property value measures of SES and more traditional SES measures. However, effect sizes for property value measures were typically equal to or greater than effect sizes for traditional measures. For example, among women, the prevalence ratio for obesity was 3.4 times greater among those having average residential tax parcel values in the lowest quartile compared with the highest (95% CI: 2.2–5.3) [38]. Contrastingly, education explained less variation in obesity rates (high-school vs college, prevalence ratio: 1.7, 95% CI: 1.2–1.7). Average residential property values within residential census tracts were also associated with soda and salad consumption [39], whereas income and education were not.

Conclusion Residential property values present a convenient and readily-available measure of both individual and neighbourhood SES. They appear to better capture the multi-faceted nature of SES compared with single, self-reported measures such as education or income. They also have potential to be applied to spatial micro-simulation models (a technique for estimating the

characteristics of a population [40]) to model obesity rates at the micro-scale.

Discussion

These case studies demonstrate how big data and traditional data both have an important role in understanding the aetiology of obesity, alongside responses to obesity interventions. An earlier mapping exercise [13] demonstrated that combining big data with traditional data could provide information spanning over 82% of the 108 nodes in the Foresight Obesity System Map. The data used in our three case studies spanned 34 nodes (31%), of which 59% were covered by big data sources. These case studies demonstrate that big data can successfully be used to augment traditional data to cover a wider scope of the obesity system, or to provide increased size, coverage, temporality, or objectivity of measures. The remaining discussion provides an in-depth review of the specific benefits, limitations and challenges encountered within these case studies.

Benefits

Large size and coverage

A key benefit, evident in all three case studies, was the potential size and coverage of the data. For example, by combining HES, NCRAS and ONS mortality data, Case Study 2 was able to assess cancer rates among over 1 million obese people. Moreover, the data were representative of the entire UK population with a recorded hospital admission since 1997, including populations that are often unreached. Furthermore, as there was no option to opt in or out, recruitment and attrition biases, which hamper traditional cohort studies, were minimised.

While the data used in both Case Studies 1 and 3 were confined to relatively small geographic regions (Leeds, UK and King County, USA, respectively), both had the potential to be extended nationally, or even internationally. For example, the Points-of-Interest data used in Case Study 1 is available across the whole of Great Britain. Property values from county tax assessors are publicly available at the level of tax parcels for all US states, with alternate sources of property values (such as commercial property sales data) being available internationally [41].

Better temporality

Traditional epidemiologic obesity studies are largely cross-sectional or take repeated measures of exposures and/or outcomes at discrete time points [42]. The data used in these case studies provided improved temporality over traditional data in several respects. For example, the Points-of-Interest data used in Case Study 1 are updated quarterly, allowing fine-grained assessment of built environment dynamics, and close temporal linkage to obesity outcomes data. Financial and time constraints would make it unfeasible to collect environmental data at this frequency and scale through primary means. Historical Points-of-Interest data also allows older cohort studies to be linked with built environment variables. Data used in Case Study 2 currently span several decades and are updated continually, allowing tracking of health outcomes (hospital admissions, cancer incidences etc.) for an ever-growing cohort of people. The property values data used in Case Study 3, while only updated every 6 years, still has more frequent updates than decennial census data, which is typically used to measure SES [43].

Objective measures

The data used in all three case studies also provided the benefit of objective measures. Case Study 1 used spatial data from the UK's national mapping agency to objectively measure neighbourhood physical activity opportunities. This is in contrast with other studies, which have asked participants about perceptions of their local environment [44]. Perception measures do not correlate well with objective measures, and both may be important to comprehensively capture built environment influences on obesity [45]. Case Study 2 used highly robust data from the NHS, Public Health England and ONS, which importantly included objective data on obesity diagnoses, surgery, cancer incidences and deaths. Finally, Case Study 3 demonstrated how property values could provide an objective proxy for individual SES, which performs better than self-reported education or income at predicting obesity-related outcomes.

Augmentation of other data

In Case Studies 1 and 3, big data were used to augment traditional data, illustrating the potential for big and traditional data to work in harmony. Both utilised location information (residential addresses) to link traditional data with built and socioeconomic environmental data. These represent important areas of the Foresight Obesity System Map frequently missing from traditional datasets. Case Study 3 also demonstrated that property values may provide improved measures of individual SES, even where alternate measures are included in traditional datasets. Moreover, measures of neighbourhood SES can be computed at a range of geographical scales, and unconstrained by administrative boundaries, minimising bias due to the MAUP [35]. These datasets also offer the potential for linkage with other big datasets such as electronic medical

records. Indeed, an ongoing study ('Moving2Health') is seeking to link longitudinal electronic medical records with historical property values data [46] in an entirely new approach to studying built environment influences on health and disease.

Limitations and challenges

As well as the many benefits described above, limitations and challenges were also encountered. These can be divided into two categories: hidden/unforeseen biases and lack of contextual information.

Hidden/unforeseen biases

Bias within data is a concern for most research. Traditional studies seek to eliminate or reduce bias through design, with the well-established 'gold standard' being the randomised controlled trial. In epidemiological research, observational and case-control studies seek to minimise biases through methodological sampling techniques and rigorous data cleaning and handling procedures. 'However, the process of collection, manipulation and extraction of value from big data—the big data analytics—is often opaque and may not follow expected research norms, making it challenging to identify and account for potential sources of bias'.

As an example, while the data used in Case Study 2 was a national sample, differences in demographics between the general population and those (i) having a hospital episode and (ii) being eligible for obesity surgery, may lead to selection biases. In particular, people undergoing obesity surgery were required by the NHS to meet certain criteria (BMI ≥40 kg m^{-2} or 35–40 kg m^{-2} alongside at least one other obesity-related condition and inability to sustain weight loss through standard techniques). These factors may be associated with cancer risk independently of obesity treatment, confounding any observed associations. Indeed, in a negative control analysis, Case Study 2 found a higher incidence of lung cancer among those with obesity, and particularly those undergoing obesity treatment, compared with the background population [34]. This was unexpected given lung cancer is not an obesity-related cancer and suggests residual confounding in the data; potentially due to the increased smoking rates among those with obesity.

Another example of bias relates to systematic differences in the handling of data. Tax parcel values, as used in Case Study 3, are determined by independent counties according to state-level regulations. There may therefore be variability in valuation methods both at the county and state levels, leading to systematic biases in property valuations nationally. While not an issue in Case Study 3, as the study area was confined to one county, appropriate methods, such as multi-level modelling, would need to be considered in research spanning multiple counties or states. Comparability of house prices across large geographical areas also requires careful analysis in view of the known tendency towards spatial autocorrelation [47].

Sources of bias can be hard to predict. A recent validation study showed that Points-of-Interest data, as used in Case Study 1, has variable completeness across different types of facilities (in this case, types of food outlets) [48]. This was thought to be due to differences in turnover/closure rates across outlet types, and the way Points-of-Interest data is sourced—with information on different outlet types being sourced from different data providers. Variability in data quality across outlet types led, in turn, to geographically varying errors due to differences in food outlet composition across environment types (e.g. deprived areas having more fast food outlets). It is unclear whether such bias would exist for listings of physical activity opportunities, as used in Case Study 1, but in any event, this example highlights how sources of bias may be difficult to anticipate.

Lack of contextual information

Lack of contextual information about the data was an additional challenge encountered across the case studies. This can lead to poorly performing predictive models and bias in causal models if confounders cannot be controlled for. Case Study 2 met a number of challenges in this respect. Firstly, the data did not include an earliest date of obesity diagnosis. This induces a time-related bias, with those undergoing surgery potentially having lived for longer with obesity than those not undergoing surgery.

Secondly, the HES data only classified procedures by type and not purpose, and it was not always clear whether procedure codes related to obesity surgery or to some other procedure (notably, some procedure codes could have encompassed both surgeries to treat obesity and surgeries to treat cancer). Procedural codes also changed over time. For example, prior to 2004 there were no codes for sleeve gastrectomy or gastric banding. It was unclear what coding was used to capture these surgeries prior to 2004 leading to further challenges in identifying obesity surgeries within the HES data.

A further 'missing information' challenge encountered in Case Study 2 was the absence of data on important covariates; notably BMI and other variables that may lead to increased cancer risk, and which may vary between the OS and no-OS groups. As mentioned above, using negative control analyses, the researchers detected potential residual confounding with the data. This highlights that even if sources of bias are identified, it may not be possible to control for them.

Challenges relating to missing contextual information were also evident, albeit to a lesser extent, in Case Studies 1 and 3. In Case Study 1, proprietary classifications were used to extract physical activity opportunities from Points-of-Interest data, but it is unclear how these classifications were applied by the data provider, and how suitable they were for capturing physical activity opportunities relevant to obesity. For example, the classifications 'swimming pools' and 'tennis facilities' were likely to include both public and private (e.g. members-only) facilities. The data also did not include factors such as facility quality, cost or opening hours—all of which may influence facility utilisation. Similarly, while the property values data used in Case Study 3 appears to provide a good predictor of individual and neighbourhood socioeconomic context, it does not include information on other assets owned by people, and therefore may not perform well in areas where property represents only a small proportion of total assets.

Future directions and conclusion

The case studies presented in this paper highlight a variety of ways in which big data and associated analytics, have been used, alongside traditional data, in whole systems obesity research. They have provided detailed examples of how big data can present improvements over traditional data in relation to size, coverage, temporality, and objectivity of measures. Case study 3 also demonstrated that big data and big data analytics could be used to simulate data that is missing/unavailable from other datasets. For example, spatial micro-simulation could be used to estimate neighbourhood obesity rates through combination of individual and area based characteristics [40]. However, these case studies also highlight that bigger data does not necessarily mean fewer challenges or limitations. Hidden/unforeseen biases and missing contextual information caused problems. Researchers should be mindful of these limitations, and look to mitigate them wherever possible, for example through using negative control analyses to test for biases, and linkage with additional datasets to provide additional contextual information.

The data used in the presented case studies, while meeting the definition of 'big data' as agreed by consensus of experts in a recent Delphi study [9], may be regarded by some as being relatively simple, and perhaps not showcasing big data to its full potential. However, we feel the case studies presented here reflect the present state of big data and obesity research, which undoubtedly still has room for advancement in harnessing the full breadth and variety of big data. Other studies that are advancing the field of big data and obesity research in terms of the complexity of data and/or associated analyses have, for example, used loyalty card data to explore associations between objectively measured food purchases and individual characteristics [49], or linked loyalty card food purchase data across the whole of London with medical prescription data to predict hypertension, high cholesterol, and diabetes at a fine spatial resolution [50]. Spatial microsimulation using census data has also been used to build a synthetic population for the UK, which has been linked via demographic characteristics to a nationally representative dietary survey (The National Diet and Nutrition Survey, allowing modelling of small-area variations in body mass index, calorie intake and physical activity level [40]. Nevertheless, there is still considerable scope for future innovation, such as through combining a greater number of diverse datasets to better capture the myriad of obesity drivers [13] and harnessing the temporal dimension of quickly-evolving datasets to track or predict changes over time.

Overall, in spite of challenges, big data and associated analytics, present a relatively untapped resource that shows promise in helping to understand obesity. We feel it is best utilised as a complement to traditional data, for example through data linkage or by providing a platform to test new methods to establish best practices in future research.

Acknowledgements The ESRC Strategic Network for Obesity was funded via ESRC grant number ES/N00941X/1. The authors would like to thank all of the network investigators (https://www.cdrc.ac.uk/research/obesity/investigators/) and members (https://www.cdrc.ac.uk/research/obesity/network-members/) for their participation in network meetings and discussion which contributed to the development of this paper.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Davison KK, Birch LL. Childhood overweight: a contextual model and recommendations for future research. Obes Rev. 2001;2:159–71.
2. Egger G, Swinburn B. An "ecological" approach to the obesity pandemic. BMJ. 1997;315:477–80.
3. Harrison K, Bost KK, McBride BA, Donovan SM, Grigsby-Toussaint DS, Kim J, et al. Toward a developmental conceptualization of contributors to overweight and obesity in childhood: the six-Cs model. Child Dev Perspect. 2011;5:50–8.
4. Butland B, Jebb S, Kopelman P, McPherson K, Thomas S, Mardell J et al. Foresight. Tackling obesities: future choices—project report. Government Office for Science; 2007.
5. Rutter HR, Bes-Rastrollo M, de Henauw S, Lahti-Koski M, Lehtinen-Jacks S, Mullerova D, et al. Balancing upstream and downstream measures to tackle the obesity epidemic: a position statement from the European association for the study of obesity. Obes Facts. 2017;10:61–3.

6. Mittelstadt BD, Floridi L. The ethics of big data: current and foreseeable issues in biomedical contexts. Sci Eng Ethics. 2016;22:303–41.
7. Kaisler S, Armour F, Espinosa JA, Money W. Big data: issues and challenges moving forward. In: Proceedings of the 46th Hawaii International Conference on System Sciences. Association for Computing Machinery Digital Library; 2013. p. 995–1004.
8. Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. J Big Data. 2014;1: https://doi.org/10.1186/2196-1115-1-2.
9. Vogel C, Zwolinsky S, Griffiths C, Hobbs M, Henderson E, Wilkins E. A Delphi study to build consensus on the definition and use of big data in obesity research. Int J Obes. 2019. https://doi.org/10.1038/s41366-018-0313-9.
10. Morris M, Birkin M. The ESRC strategic network for obesity: tackling obesity with big data. Int J Obes. 2018;42:1948–50.
11. Timmins K, Green M, Radley D, Morris M, Pearce J. How has big data contributed to obesity research? A review of the literature. Int J Obes. 2018;42:1951–62.
12. Monsivais P, Francis O, Lovelace R, Chang M, Strachan E, Burgoine T. Data visualisation to support obesity policy: case studies of data tools for planning and transport policy in the UK. Int J Obes. 2018;42:1977–86.
13. Morris M, Wilkins E, Timmins K, Bryant M, Birkin M, Griffiths C. Can big data solve a big problem? Reporting the obesity data landscape in line with the Foresight obesity system map. Int J Obes. 2018;42:1963–76.
14. Vayena E, Salathé M, Madoff LC, Brownstein JS. Ethical challenges of big data in public health. PLOS Comput Biol. 2015;11: e1003904.
15. Silver LD, Ng SW, Ryan-Ibarra S, Taillie LS, Induni M, Miles DR, et al. Changes in prices, sales, consumer spending, and beverage consumption one year after a tax on sugar-sweetened beverages in Berkeley, California, US: a before-and-after study. PLoS Med. 2017;14:e1002283.
16. Gore RJ, Diallo S, Padilla J. You are what you tweet: connecting the geographic variation in america's obesity rate to Twitter content. PLoS ONE. 2015;10:e0133505.
17. Nguyen QC, Li D, Meng H-W, Kath S, Nsoesie E, Li F, et al. Building a national neighborhood dataset from geotagged Twitter data for indicators of happiness, diet, and physical activity. JMIR Public Health Surveill. 2016;2:e158.
18. Hirsch JA, James P, Robinson JR, Eastman KM, Conley KD, Evenson KR, et al. Using MapMyFitness to place physical activity into neighborhood context. Front Public Health. 2014;2:1–9.
19. Althoff T, Hicks JL, King AC, Delp SL, Leskovec J. Large-scale physical activity data reveal worldwide activity inequality. Nature. 2017;547:336–9.
20. Kerr NL. HARKing: hypothesizing after the results are known. Pers Soc Psychol Rev. 1998;2:196–217.
21. Lee IM, Shiroma EJ, Lobelo F, Puska P, Blair SN, Katzmarzyk PT, et al. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. Lancet. 2012;380:219–29.
22. Bennett JE, Li G, Foreman K, Best N, Kontis V, Pearson C, et al. The future of life expectancy and life expectancy inequalities in England and Wales: Bayesian spatiotemporal forecasting. Lancet. 2015;386:163–70.
23. World Health Organisation. Report of the Commission on ending childhood obesity. Geneva, Switzerland: World Health Organisation; 2016.
24. Centers for Disease Control and Prevention. Recommended community strategies and measurements to prevent obesity in the United States. Atlanta, GA, U.S.: Centers for Disease Control and Prevention; 2009.
25. Local Government Association. Building the foundations: tackling obesity through planning and development. London, UK: Local Government Association; 2016.
26. Burgoine T, Alvanides S, Lake AA. Creating 'obesogenic realities'; Do our methodological choices make a difference when measuring the food environment? Int J Health Geogr. 2013;12. https://doi.org/10.1186/1476-072X-12-33.
27. Wilkins E, Morris M, Radley D, Griffiths C. Methods of measuring associations between the Retail Food Environment and weight status: Importance of classifications and metrics. SSM Popul Health. 2019. https://doi.org/10.1016/j.ssmph.2019.100404.
28. Bardou M, Barkun AN, Martel M. Obesity and colorectal cancer. Gut. 2013;62:933–47.
29. Siegel R, Desantis C, Jemal A. Colorectal cancer statistics, 2014. CA Cancer J Clin. 2014;64:104–17.
30. Derogar M, Hull MA, Kant P, Östlund M, Lu Y, Lagergren J. Increased risk of colorectal cancer after obesity surgery. Ann Surg. 2013;258:983–8.
31. Kant P, Hull MA. Excess body weight and obesity—the link with gastrointestinal and hepatobiliary cancer. Nat Rev Gastroenterol Hepatol. 2011;8:224–38.
32. Östlund MP, Lu Y, Lagergren J. Risk of obesity-related cancer after obesity surgery in a population-based cohort study. Ann Surg. 2010;252:972–6.
33. Sainsbury A, Goodlad RA, Perry SL, Pollard SG, Robins GG, Hull MA. Increased colorectal epithelial cell proliferation and crypt fission associated with obesity and roux-en-Y gastric bypass. Cancer Epidemiol Biomark Prev. 2008;17:1401–10.
34. Aravani A, Downing A, Thomas JD, Lagergren J, Morris EJA, Hull MA. Obesity surgery and risk of colorectal and other obesity-related cancers: an English population-based cohort study. Cancer Epidemiol. 2018;53:99–104.
35. Openshaw S. The modifiable areal unit problem. In: Concepts and techniques in modern geography. Norwich: Geo Books; 1984. p. 1–41.
36. Kwan M-P. The uncertain geographic context problem. Ann Assoc Am Geogr. 2012;102:958–68.
37. Di Zhu X, Yang Y, Liu X. The importance of housing to the accumulation of household net wealth. Harvard, USA: Joint Center for Housing Studies, Harvard University; 2003.
38. Rehm CD, Moudon AV, Hurvitz PM, Drewnowski A. Residential property values are associated with obesity among women in King County, WA, USA. Soc Sci Med. 2012;75:491–5.
39. Drewnowski A, Buszkiewicz J, Aggarwal A. Soda, salad, and socioeconomic status: findings from the Seattle Obesity Study (SOS). SSM Popul Health. 2019;7:e100339.
40. Birkin M, Morris MA, Birkin TM, Lovelace R. Using census data in microsimulation modelling. In: Stillwell J, Duke-Williams O, editors. The Routledge handbook of census resources, methods and applications. 1st ed. Routledge: IJO publication; 2018.
41. Jiao J, Drewnowski A, Moudon AV, Aggarwal A, Oppert J-M, Charreire H, et al. The impact of area residential property values on self-rated health: a cross-sectional comparative study of Seattle and Paris. Prev Med Rep. 2016;4:68–74.
42. Nguyen DM, El-Serag HB. The epidemiology of obesity. Gastroenterol Clinics. 2010;39:1–7.
43. Pickett KE, Pearl M. Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review. J Epidemiol Commun Health. 2001;55:111–22.
44. Timperio A, Salmon J, Telford A, Crawford D. Perceptions of local neighbourhood environments and their relationship to childhood overweight and obesity. Int J Obes. 2005;29:170–5.
45. Roda C, Charreire H, Feuillet T, Mackenbach JD, Compernolle S, Glonti K, et al. Mismatch between perceived and objectively measured environmental obesogenic features in European neighbourhoods. Obes Rev. 2016;17 S1:31–41.

46. Drewnowski A, Arterburn D, Zane J, Aggarwal A, Gupta S, Hurvitz PM, et al. The Moving to Health (M2H) approach to natural experiment research: a paradigm shift for studies on built environment and health. SSM Popul Health. 2019;7:100345.
47. Bourassa SC, Cantoni E, Hoesli M. Predicting house prices with spatial dependence a comparison of alternative methods. J Real Estate Res. 2010;32:139–60.
48. Wilkins EL, Radley D, Morris MA, Griffiths C. Examining the validity and utility of two secondary sources of food environment data against street audits in England. Nutr J. 2017;16:1–13.
49. Nevalainen J, Erkkola M, Saarijarvi H, Nappila T, Fogelholm M. Large-scale loyalty card data in health research. Digit Health. 2018;4:2055207618816898.
50. Aiello L, Schifanello R, Quercia D, Del Prete L. Large-scale and high-resolution analysis of food purchases and health outcomes. EPJ Data Sci. 2019;8:14.
51. Craig CL, Marshall AL, Sjostrom M, Bauman AE, Booth ML, Ainsworth BE, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc. 2003;35:1381–95.
52. Zwolinsky S, McKenna J, Pringle A, Widdop P, Griffiths C, Mellis M, et al. Physical activity and sedentary behavior clustering: segmentation to optimize active lifestyles. J Phys Act Health. 2016;13:921–8.
53. Bauman A, Ainsworth BE, Sallis JF, Hagströmer M, Craig CL, Bull FC, et al. The descriptive epidemiology of sitting: a 20-country comparison using the International Physical Activity Questionnaire (IPAQ). Am J Prev Med. 2011;41:228–35.
54. Guerin PB, Diiriye RO, Corrigan C, Guerin B. Physical activity programs for refugee somali women: working out in a new country. Women & Health. 2003;38:83–99.
55. Pope L, Harvey J. The efficacy of incentives to motivate continued fitness-center attendance in college first-year students: a randomized controlled trial. J Am Coll Health. 2014;62:81–90.
56. Cetateanu A, Jones A. Understanding the relationship between food environments, deprivation and childhood overweight and obesity: evidence from a cross sectional England-wide study. Health Place. 2014;27:68–76.
57. Harrison F, Burgoine T, Corder K, van Sluijs EM, Jones A. How well do modelled routes to school record the environments children are exposed to? A cross-sectional comparison of GIS-modelled and GPS-measured routes to school. Int J Health Geogr. 2014;13:5.
58. Ells LJ, Macknight N, Wilkinson JR. Obesity surgery in England: an examination of the health episode statistics 1996–2005. Obes Surg. 2007;17:400–5.
59. Nielsen JDJ, Laverty AA, Millett C, Mainous AG, Majeed A, Saxena S. Rising obesity-related hospital admissions among children and young people in England: National time trends study. PLoS ONE. 2013;8:e65764.
60. Smittenaar C, Petersen K, Stewart K, Moitt N. Cancer incidence and mortality projections in the UK until 2035. Br J Cancer. 2016;115:1147–55.
61. Wallington M, Saxon EB, Bomb M, Smittenaar R, Wickenden M, McPhail S, et al. 30-day mortality after systemic anticancer treatment for breast and lung cancer in England: a population-based, observational study. The Lancet Oncol. 2016;17:1203–16.
62. Smolina K, Wright FL, Rayner M, Goldacre MJ. Determinants of the decline in mortality from acute myocardial infarction in England between 2002 and 2010: Linked national database study. BMJ. 2012;344:d8059.
63. Hanratty B, Lowson E, Grande G, Payne S, Addington-Hall J, Valtorta N, et al. Transitions at the end of life for older adults–patient, carer and professional perspectives: A mixed-methods study. Health Serv Deliv Res. 2014. https://doi.org/10.3310/hsdr02170.
64. Aggarwal A, Monsivais P, Cook AJ, Drewnowski A. Does diet cost mediate the relation between socioeconomic position and diet quality? Eur J Clin Nutr. 2011;65:1059–66.
65. Drewnowski A, Aggarwal A, Tang W, Moudon AV. Residential property values predict prevalent obesity but do not predict 1-year weight change. Obesity. 2015;23:671–6.