UNIVERSITY OF LEEDS

This is a repository copy of *Evidence from big data in obesity research: International case studies*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/154851/

Version: Accepted Version

Article:

Wilkins, E, Aravani, A orcid.org/0000-0002-3976-7888, Downing, A orcid.org/0000-0002-0335-7801 et al. (6 more authors) (2020) Evidence from big data in obesity research: International case studies. International Journal of Obesity, 44 (5). pp. 1028-1040. ISSN 0307-0565

https://doi.org/10.1038/s41366-020-0532-8

© The Author(s), under exclusive licence to Springer Nature Limited 2020. This is an author produced version of a paper published in International Journal of Obesity. Uploaded in accordance with the publisher's self-archiving policy.

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Title: Evidence from big data in obesity research: International case studies

Running title: Evidence from big data in obesity research

Authors: 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]

[1] Leeds Institute for Data Analytics & School of Medicine, University of Leeds, United Kingdom

[2] Center for Public Health Nutrition, University of Washington, Seattle, USA

[3] School of Sport, Leeds Beckett University, United Kingdom

[4] Leeds Institute for Data Analytics & School of Geography, University of Leeds, United Kingdom

[5] Engineering & Environment, Northumbria University, United Kingdom

[6] GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany

Corresponding author: Michelle A Morris

Email: m.morris@leeds.ac.uk

Tel: +44 113 343 0883

Conflict of interest statement

Authors report no conflict of interest.

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*[1]*. '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 sociodemographic 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[2-4]. This complexity is represented in the Foresight Obesity System Map[5], 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[6]. Understanding the drivers of obesity and responses to interventions within such a complex system 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[7-9]. In this paper, we adopt a definition of 'big data' established by a recent Delphi survey of international obesity and big data experts[1], 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 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, Green *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[1, 7, 14]. Consistent reporting of data sources, such as through 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 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[5] 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 manage 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)

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 (GIS) 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 ‘Lower

Super Output Area' (LSOA) boundaries (a UK administrative geography containing ~1,500 people) and 2km 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 ~8 000 people) or a 2km circular buffer only accounted for a small proportion of the variability in attendance rates. On further investigation, circular buffers of at least 4km 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 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 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 identification of the time 'at risk' - the time from obesity diagnosis/surgery to development of a cancer, death or last follow-up (30^{th} 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, 3 237 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 (Table1).

Methods: Data from the Seattle Obesity Study (SOS) I, II and III were used to assess associations between socioeconomic 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 to 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 to single, self-reported measures such as education or income. They also have potential to be applied to spatial microsimulation 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 socioeconomic status, 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 ≥40kg·m^{-2} or 35-40kg·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 to 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 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 microsimulation 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 [49], 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 [50], 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 [51]. 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 [52]. 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 [53] 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 Economic and Social Research Council grant number ES/N00941X/1.

The authors would like to thank all of the network investigators[a] and members[b] for their participation in network meetings and discussion which contributed to the development of this paper.

[a] - www.cdrc.ac.uk/research/obesity/investigators/

[b] - www.cdrc.ac.uk/research/obesity/network-members/

Conflict of interest statement

Authors report no conflict of interest.

References

1. 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. *International Journal of Obesity* 2019**:** e-pub ahead of print 17 Jan 2019; doi:10.1038/s41366-018-0313-9.

2. Davison KK, Birch LL. Childhood overweight: A contextual model and recommendations for future research. *Obesity Reviews* 2001; **2**(3)**:** 159-71.

3. Egger G, Swinburn B. An “ecological” approach to the obesity pandemic. *BMJ* 1997; **315**(7106)**:** 477-480.

4. 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 Development Perspectives* 2011; **5**(1)**:** 50-58.

5. Butland B, Jebb S, Kopelman P, McPherson K, Thomas S, Mardell J *et al.* Foresight. Tackling obesities: future choices. Project report. 2007.

6. 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. *Obesity Facts* 2017; **10**(1)**:** 61-63.

7. Mittelstadt BD, Floridi L. The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics* 2016; **22**(2)**:** 303-41.

8. Kaisler S, Armour F, Espinosa JA, Money W. Big data: Issues and challenges moving forward. *2013 46th Hawaii International Conference on System Sciences* 2013**:** 995-1004.

9. Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. *Journal of Big Data* 2014; **1**(2)**:** e-pub 24 June 2014; https://doi.org/10.1186/2196-1115-1-2.

10. Morris M, Birkin M. The ESRC Strategic Network for Obesity: Tackling obesity with big data. *International Journal of Obesity* 2018; **42**(12)**:** 1948-1950.

11. Timmins K, Green M, Radley D, Morris M, Pearce J. How has big data contributed to obesity research? A review of the literature. *International Journal of Obesity* 2018; **42:** 1951–1962.

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. *International Journal of Obesity* 2018; **42**(12)**:** 1977-1986.

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. *International Journal of Obesity* 2018; **42**(12)**:** 1963-1976.

14. Vayena E, Salathé M, Madoff LC, Brownstein JS. Ethical challenges of big data in public health. *PLOS Computational Biology* 2015; **11**(2)**:** 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 Medicine* 2017; **14**(4)**:** 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**(9)**:** 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 and Surveillance* 2016; **2**(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. *Frontiers in Public Health* 2014; **2**(19)**:** 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**(7663)**:** 336–339.

20. Kerr NL. HARKing: hypothesizing after the results are known. *Pers Soc Psychol Rev* 1998; **2**(3)**:** 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**(9838)**:** 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**(9989)**:** 163-70.

23. World Health Organisation. Report of the Commission on ending childhood obesity. 2016.

24. Centers for Disease Control and Prevention. Recommended community strategies and measurements to prevent obesity in the United States. 2009.

25. Local Government Association. Building the foundations: Tackling obesity through planning and development. 2016.

26. Burgoine T, Alvanides S, Lake AA. Creating 'obesogenic realities'; Do our methodological choices make a difference when measuring the food environment? *International journal of health geographics* 2013; **12**(33)**:** e-pub 3 July 2013; doi: 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 - population health* 2019**:** e-pub ahead of print 4 May 2019; doi: https://doi.org/10.1016/j.ssmph.2019.100404.

28. Bardou M, Barkun AN, Martel M. Obesity and colorectal cancer. *Gut* 2013; **62**(6)**:** 933-947.

29. Siegel R, Desantis C, Jemal A. Colorectal cancer statistics, 2014. *CA Cancer Journal for Clinicians* 2014; **64**(2)**:** 104-117.

30. Derogar M, Hull MA, Kant P, Östlund M, Lu Y, Lagergren J. Increased risk of colorectal cancer after obesity surgery. *Annals of Surgery* 2013; **258**(6)**:** 983-988.

31. Kant P, Hull MA. Excess body weight and obesity—the link with gastrointestinal and hepatobiliary cancer. *Nature Reviews Gastroenterology and Hepatology* 2011; **8**(4)**:** 224-238.

32. Östlund MP, Lu Y, Lagergren J. Risk of obesity-related cancer after obesity surgery in a population-based cohort study. *Annals of Surgery* 2010; **252**(6)**:** 972-976.

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 Epidemiology Biomarkers & Prevention* 2008; **17**(6)**:** 1401-1410.

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 Epidemiology* 2018; **53:** 99-104.

35. Openshaw S. The modifiable areal unit problem. In: *Concepts and Techniques in Modern Geography*. Geo Books: Norwich, 1984, pp 1-41.

36. Kwan M-P. The uncertain geographic context problem. *Annals of the Association of American Geographers* 2012; **102**(5)**:** 958-968.

37. Di Zhu X, Yang Y, Liu X. The importance of housing to the accumulation of household net wealth. In: 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. *Social Science & Medicine* 2012; **75**(3)**:** 491-495.

39. Drewnowski A, Buszkiewicz J, Aggarwal A. Soda, salad, and socioeconomic status: Findings from the Seattle Obesity Study (SOS). *SSM-Population 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 (eds). *The Routledge Handbook of Census Resources, Methods and Applications*, 1 edn. Routledge, 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. *Preventive Medicine Reports* 2016; **4:** 68-74.

42. Nguyen DM, El-Serag HB. The epidemiology of obesity. *Gastroenterology Clinics* 2010; **39**(1)**:** 1-7.

43. Pickett KE, Pearl M. Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review. *Journal of Epidemiology & Community Health* 2001; **55**(2)**:** 111-122.

44. Timperio A, Salmon J, Telford A, Crawford D. Perceptions of local neighbourhood environments and their relationship to childhood overweight and obesity. *International Journal of Obesity* 2005; **29**(2)**:** 170-175.

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. *Obesity Reviews* 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-Population Health* 2019; **7:** 100345.

47. Bourassa SC, Cantoni E, Hoesli M. Predicting house prices with spatial dependence a comparison of alternative methods. *The Journal of Real Estate Research* 2010; **32**(2)**:** 139-160.

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**(82)**:** 1-13.

49. 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 (Lond)* 2019.

50. Nevalainen J, Erkkola M, Saarijarvi H, Nappila T, Fogelholm M. Large-scale loyalty card data in health research. *Digit Health* 2018; **4:** 2055207618816898.

51. Aiello L, Schifanello R, Quercia D, Del Prete L. Large-scale and high-resolution analysis of food purchases and health outcomes. *EPJ Data Science* 2019; **8**(14).

52. Birkin M, Morris M, Birkin T, Lovelce R. Using census data in microsimulation modelling. In: Stillwell J, Duke-Williams O (eds). *The Routledge Handbook of Census Resources, Methods and Applications*, 2018.

53. Morris MA, Wilkins E, Timmins KA, 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 (Lond)* 2018; **42**(12)**:** 1963-1976.