**Name of Journal:** *World Journal of Methodology*

**Manuscript NO:** 86114

**Manuscript Type:** MINIREVIEWS

**Using national census data to facilitate healthcare research**

Colwill M *et al*. Census data in healthcare research

Michael Colwill, Andrew Poullis

**Michael Colwill, Andrew Poullis,** Department of Gastroenterology, St George’s Hospital London, London SW17 0QT, United Kingdom

**Author contributions:** Colwill M and Poullis A were involved in conception, literature review, writing and review.

**Corresponding author: Michael Colwill, BSc, MBBS, MRCP, Doctor,** Department of Gastroenterology, St George’s Hospital London, Blackshaw Road, London SW17 0QT, United Kingdom. michael.colwill@nhs.net

**Received:** May 31, 2023

**Revised:** September 9, 2023

**Accepted:** September 26, 2023

**Published online:**

**Abstract**

National censuses are conducted at varying intervals across both the developed and developing world and collect detailed data on a wide range of societal, economic and health questions. This immense volume of data has many potential uses in the field of healthcare research and can be utilised either in isolation or in conjunction with other information sources such as hospital records. At a governmental level census data can be used for healthcare service planning by providing accurate population density information but also, through the use of more detailed data collection, by helping to identify high-risk populations that may require increased resource allocation. It can also be a key tool in addressing and improving healthcare inequality and deprivation by both identifying those populations with poorer healthcare outcomes and through helping researchers to better understand the causes of this inequality. Similarly, it has utility when studying the complex causes of disease and assessing the success of strategies designed to tackle these aetiologies. However, the maximum benefit from these various uses can only be realised if the data collection and analysis processes utilised are robust and this requires that census bureaus regularly review and modify their methods in a transparent and thorough way.

**Key Words:** Census data; Methodology; Epidemiology

Colwill M, Poullis A. Using national census data to facilitate healthcare research. *World J Methodol* 2023; In press

**Core Tip:** National census data is collected widely across the world. Recently, more detailed data on a wide range of societal, economic and health questions has begun to be collected and this vast volume of data has enormous potential in healthcare research. Examples of potential utility are in assisting with healthcare service planning, analysing healthcare workforces, identifying healthcare inequality and it’s causes and understanding the causes of disease. However, census data’s utility is dependent upon robust and scientific data collection and analysis and this requires regular methodological review and improvement by national census bureaus.

**INTRODUCTION**

National censuses are performed in the majority of countries in the developed world and a growing number of developing countries (see Figure 1). The breadth of the data collected has increased and diversified significantly in recent decades with many countries now collecting data on socioeconomic status and health conditions as well as basic population demographics[1]. Whilst the idea of using census data for healthcare research is not new[2], this vast collection of data remains underutilised. This article will discuss areas in which this data has utility and some of the pitfalls associated with it.

***Healthcare service planning***

The basic demographic data that is provided by a national census is crucial for all elements of government planning including healthcare provision. At a very basic level, providing accurate data on population density can be used to decide the location and provision of healthcare facilities[3]. An example of this was demonstrated by a Tanzanian study focusing on maternal outcomes in obstetric care based upon proximity to healthcare centres and found that a greater distance to healthcare facilities was associated with worse maternal outcomes[4]. This data has since been used to justify the construction of new healthcare centres in appropriate under-served regions in order to address this disparity.

Some countries go further than just basic population data, such as in the United States where the American Community Survey (ACS) is performed along with the decennial census. The ACS tracks social determinants of health such as income, housing and national origins as well as insurance coverage, fertility and disability. This allows the department of health and human services (HHS) to more precisely target resources to match the anticipated needs of each region and is a key part of HHS’ long-term strategy.

Census data have also been used to analyse the healthcare services themselves as well as the populations they serve. A study from Japan in 2018 used several decades of census data and cross-referenced it with physician surveys and municipality borders to investigate concerns of disparity between the number of physicians in urban and rural practice[5]. The study identified an uneven distribution, which had been worsening over time, with a lack of physicians in rural settings and prompted government departments to start to develop strategies to mitigate this. Workforce analysis was performed in the United Kingdom using census data to review the make-up of the healthcare workforce and identified a heavy dependence upon foreign-trained workers indicating that domestic training programmes needed reforming[6].

Similarly, work by Gupta *et al*[7] used national census data from three different countries to analyse and compare inter-country healthcare worker provision. Their study gave a detailed snapshot of differences between these countries and the various challenges they faced which provided a mandate for the international community, including non-governmental organisations and charities, to direct and focus their resources demonstrating the utility of this data in a transnational, as well as national, setting. It should however also be noted that they found significant variability in the quality of data provided and this imposed limitations on the conclusions they were able to make and this will be discussed later in this article.

***Addressing healthcare inequality***

Studies have repeatedly shown that healthcare inequality, both at national and international levels, can impact upon mortality and morbidity[8,9] and contribute to deprivation. A recent example was data collected during the coronavirus pandemic which identified differences in outcome with those from deprived health systems having significantly worse survival[10] and thrust the issue of health inequality into the spotlight. Addressing this inequality is a complex political, societal and public health conundrum but census data can be a key element to identify inequalities and guide reform.

In order to allow politicians to appropriately allocate the resources required to address healthcare inequality, the location and nature of the inequality needs to be clearly identified and this is where census data has a role. In the United Kingdom, the 2019 NHS long term plan made addressing healthcare inequality a priority specifically targeting the most deprived 10% of the United Kingdom population. This plan, along with the Core20Plus5 initiative, combined the use of national census data, general practitioner records and hospital records to identify, at a local level, those groups who suffer from the highest levels of healthcare inequality[11]. Some examples of ‘at-risk groups’ were those from an ethnic minority or those with a disability.

Similar use of census data to identify populations with higher risk of deprivation have been used in the United States[12] with some interesting epidemiological findings. One example was the so called ‘Hispanic paradox’ where historically this population was believed to have better healthcare outcomes despite their high deprivation scores and risk profiles. However, recent and more detailed census data analysis has found that this may not be the case[13] demonstrating the importance of high quality data and statistical analysis.

Census data is also important when monitoring the progress, or lack of, with regards to tackling healthcare inequality. A large study in the United States entitled ‘The Public Health Disparities Geocoding Project’ used a five step data analysis process to determine, through census and health surveillance data, a picture of health inequalities over time. It identified both areas where improvements had been made but also areas where the problem persisted or had worsened and has been used to inform public policy[14].

Other work has focused on using census data to identify the causes of healthcare inequality. A study in the United States analysed this data and identified a significant association between the presence of greater numbers of liquor stores and the risk of health-related social problems in low income neighbourhoods[15]. Whilst there is clearly not a single cause for poorer health outcomes, this interesting analysis sheds light on possible environmental factors that will be an important part of reducing healthcare inequality.

***Understanding the causes of disease***

As previously mentioned, census data also have a further role in addressing healthcare inequality by assisting researchers to understand the causes of disease. Canadian researchers, through combining primary care records and census data, demonstrated a link between socioeconomic status and obesity[16] whilst a study in Spain used a similar methodology to demonstrate a link between deprivation and common cancers in order to better target screening programmes[17]. These studies show the utility of census data in assessing health disparities and environmental factors associated with chronic disease.

There are also examples of more detailed and complex use of census data for similar purposes. Moceri *et al*[18] used, in a case-control study, census data and birth certificates to reconstruct the early-life socioeconomic environment of elderly Alzheimer’s patients and, through examining variables such as paternal occupation, parental age and birth order amongst others, found higher odds-ratios for developing Alzheimer’s for certain characteristics. They also then combined this with genetic analysis of these patients to study the interaction between apolipoprotein ε4 allele and these socioeconomic risk factors.

As well as identifying risk factors for disease, census data have also been used to demonstrate effective interventions in improving public health. Patterson *et al*[19] used census data in England and Wales to demonstrate that active commuting, such as cycling or walking, was associated with lower cardiovascular risk. This is, in theory, a relatively easily achieved public health initiative and there are an increasing number of programmes attempting to increase this method of commuting with the aim of improving public health and reducing the risk of a wide variety of diseases.

***Pitfalls***

Whilst census data have multiple potential uses there are caveats that need to be addressed and recognised. Firstly, it’s utility is dependent upon having robust and accurate data and there have been instances where poor or incorrect collection has had profound social impact. In the United States, the 1840 census incorrectly identified higher levels of insanity amongst the ‘coloured’ population, an argument then used by slave-owners to suggest that African-American populations were not able to live as free people[20]. There has also been historically inaccurate data about native American populations and rates of disease leading to worsening healthcare inequality fuelled by the misappropriation of federal funding. A more recent example showed persisting inequalities when it comes to accurate population and health data collection in the Maori population in New Zealand[21] meaning that they receive less resource allocation from government funding. Moreover, there have been documented examples where census data has been deliberately falsified in order to obtain greater funding and support for specific regions. This was discussed and reviewed in detail by Adele with regards to the national census in Nigeria and identified chronic and deliberate falsification of data to obtain benefits from the government[22]. Given the implications for strategic planning and resource allocation, an inaccurate census can have profound impacts for communities and citizens and this is also true when the data is used for healthcare research. These examples underpin the need for a robust scientific process enabling accurate data collection and interpretation in order to have maximum benefit for those who need it the most.

Secondly, there has also been debate around the ethical considerations of using census data for healthcare research. Whilst the data is often anonymised, the nature of census data involves categorisation of respondents and there can be discontent with even simply the names of these categories. A recent example is the controversy surrounding the inclusion of a question asking for respondent’s gender identification in England and Wales census for the first time[23]. Similarly, there have been concerns about racial categorisation in association with health labels and stigma[24] although interestingly a public panel consultation in 2018 found that members of the public were in support of census data collection and it’s use in healthcare research[25].

Thirdly, there is also a concern that there is inherent bias[26] within the census process itself. This can take the form of non-respondent bias, as described by the United States census bureau[27], which has been shown to skew data significantly such as during the ‘poll tax’ era in the United Kingdom when non-respondent rates increased. There is also a concern about accurate female representation with this group being historically under-represented[26]. These pitfalls demonstrate the need for regular review of the methodological and analytical practices employed by census bureaus and appropriate improvements if indicated.

**CONCLUSION**

Internationally, census data collection is becoming more widespread, detailed and robust particularly in the developing world. This data, provided it is accurate to avoid defects, is an immensely rich resource which has utility in research to help with healthcare planning, reducing healthcare inequality and understanding more about the causes of disease. Provided that accurate data is collected in-line with good scientific practice and remains widely and freely available to researchers, it has the ability to be an invaluable resource in healthcare research.

**REFERENCES**

1 **The Leadership Conference Education Fund**. The Census and Health Care. [cited 22 May 2023]. Available from: http://civilrightsdocs.info/pdf/census/2020/Census-Health-Care-Factsheet.pdf

2 Cherubic Rescues by H. E. Meeker, M.D.: Cocaine Prescription and Safety-pin Extraction. *Anesthesiology* 2018; **129**: 535 [PMID: 30106781 DOI: 10.1097/ALN.0000000000002393]

3 **Office for National Statistics**. Census 2021 - Census stories. [cited 22 May 2023]. Available from: https://census.gov.uk/census-stories/

4 **Hanson C**, Cox J, Mbaruku G, Manzi F, Gabrysch S, Schellenberg D, Tanner M, Ronsmans C, Schellenberg J. Maternal mortality and distance to facility-based obstetric care in rural southern Tanzania: a secondary analysis of cross-sectional census data in 226 000 households. *Lancet Glob Health* 2015; **3**: e387-e395 [PMID: 26004775 DOI: 10.1016/S2214-109X(15)00048-0]

5 **Matsumoto M**, Kimura K, Inoue K, Kashima S, Koike S, Tazuma S. Aging of hospital physicians in rural Japan: A longitudinal study based on national census data. *PLoS One* 2018; **13**: e0198317 [PMID: 29856807 DOI: 10.1371/journal.pone.0198317]

6 **Yar M**, Dix D, Bajekal M. Socio-demographic characteristics of the healthcare workforce in England and Wales-- results from the 2001 Census. *Health Stat Q* 2006: 44-56 [PMID: 17165469]

7 **Gupta N**, Zurn P, Diallo K, Dal Poz MR. Uses of population census data for monitoring geographical imbalance in the health workforce: snapshots from three developing countries. *Int J Equity Health* 2003; **2**: 11 [PMID: 14697099 DOI: 10.1186/1475-9276-2-11]

8 **National Confidential Enquiry into Patient Outcome and Death**. How data captured by NCEPOD supports the identification of healthcare inequalities a review - 2022. [cited 22 May 2023]. Available from: https://www.ncepod.org.uk/pdf/current/Healthcare%20Inequalities.pdf

9 **The Health Foundation**. Quantifying health inequalities in England. [cited 23 May 2023]. Available from: https://www.health.org.uk/news-and-comment/charts-and-infographics/quantifying-health-inequalities#:~:text=This%20analysis%20uses%20a%20novel%20approach%20to%20explore,illness%20on%20people%20and%20their%20health%20care%20needs.

10 **Mishra V**, Seyedzenouzi G, Almohtadi A, Chowdhury T, Khashkhusha A, Axiaq A, Wong WYE, Harky A. Health Inequalities During COVID-19 and Their Effects on Morbidity and Mortality. *J Healthc Leadersh* 2021; **13**: 19-26 [PMID: 33500676 DOI: 10.2147/JHL.S270175]

11 **NHS England**. Core20PLUS5 (adults) - an approach to reducing healthcare inequalities. [cited 23 May 2023]. Available from: https://www.england.nhs.uk/about/equality/equality-hub/national-healthcare-inequalities-improvement-programme/core20plus5/

12 **Giachello AL**, Bell R, Aday LA, Andersen RM. Uses of the 1980 census for Hispanic health services research. *Am J Public Health* 1983; **73**: 266-274 [PMID: 6824113 DOI: 10.2105/ajph.73.3.266]

13 **Montanez-Valverde R**, McCauley J, Isasi R, Zuchner S, Carrasquillo O; SouthEast Enrollment Center Investigators and the All of Us Research Program Demonstration Projects Subcommittee. Revisiting the Latino Epidemiologic Paradox: an Analysis of Data from the All of Us Research Program. *J Gen Intern Med* 2022; **37**: 4013-4014 [PMID: 35505219 DOI: 10.1007/s11606-022-07625-y]

14 **Krieger N**, Chen JT, Waterman PD, Rehkopf DH, Subramanian SV. Painting a truer picture of US socioeconomic and racial/ethnic health inequalities: the Public Health Disparities Geocoding Project. *Am J Public Health* 2005; **95**: 312-323 [PMID: 15671470 DOI: 10.2105/AJPH.2003.032482]

15 **LaVeist TA**, Wallace JM Jr. Health risk and inequitable distribution of liquor stores in African American neighborhood. *Soc Sci Med* 2000; **51**: 613-617 [PMID: 10868674 DOI: 10.1016/s0277-9536(00)00004-6]

16 **Biro S**, Williamson T, Leggett JA, Barber D, Morkem R, Moore K, Belanger P, Mosley B, Janssen I. Utility of linking primary care electronic medical records with Canadian census data to study the determinants of chronic disease: an example based on socioeconomic status and obesity. *BMC Med Inform Decis Mak* 2016; **16**: 32 [PMID: 26969124 DOI: 10.1186/s12911-016-0272-9]

17 **Garcia-Gil M**, Elorza JM, Banque M, Comas-Cufí M, Blanch J, Ramos R, Méndez-Boo L, Hermosilla E, Bolibar B, Prieto-Alhambra D. Linking of primary care records to census data to study the association between socioeconomic status and cancer incidence in Southern Europe: a nation-wide ecological study. *PLoS One* 2014; **9**: e109706 [PMID: 25329578 DOI: 10.1371/journal.pone.0109706]

18 **Moceri VM**, Kukull WA, Emanual I, van Belle G, Starr JR, Schellenberg GD, McCormick WC, Bowen JD, Teri L, Larson EB. Using census data and birth certificates to reconstruct the early-life socioeconomic environment and the relation to the development of Alzheimer's disease. *Epidemiology* 2001; **12**: 383-389 [PMID: 11416775 DOI: 10.1097/00001648-200107000-00007]

19 **Patterson R**, Panter J, Vamos EP, Cummins S, Millett C, Laverty AA. Associations between commute mode and cardiovascular disease, cancer, and all-cause mortality, and cancer incidence, using linked Census data over 25 years in England and Wales: a cohort study. *Lancet Planet Health* 2020; **4**: e186-e194 [PMID: 32442494 DOI: 10.1016/S2542-5196(20)30079-6]

20 **Krieger N**. The US Census and the People's Health: Public Health Engagement From Enslavement and "Indians Not Taxed" to Census Tracts and Health Equity (1790-2018). *Am J Public Health* 2019; **109**: 1092-1100 [PMID: 31219723 DOI: 10.2105/AJPH.2019.305017]

21 **Harris R**, Paine SJ, Atkinson J, Robson B, King PT, Randle J, Mizdrak A, McLeod M. We still don't count: the under-counting and under-representation of Māori in health and disability sector data. *N Z Med J* 2022; **135**: 54-78 [PMID: 36521086]

22 **Bamgbose AJ**. Falsification of population census data in a heterogeneous Nigerian state: The fourth republic example. *Afr J Polit Sci Int Relat* 2009; **3**: 311-319

23 **Cooley L**. LGBT activism and the census: A battle half-won? [cited 23 May 2023]. Available from: https://blogs.lse.ac.uk/gender/2019/02/11/lgbt-activism-and-the-census-a-battle-half-won/

24 Race and the Census: The “Negro” Controversy. [cited 30 May 2023]. Available from: https://mixedracestudies.org/wp/?p=29861

25 **Douglas A**, Ward HJT, Bhopal R, Kirkpatrick T, Sayed-Rafiq A, Gruer L; SHELS researchers. Is the linkage of census and health data justified? Views from a public panel of the Scottish Health and Ethnicity Linkage study. *J Public Health (Oxf)* 2018; **40**: 435-440 [PMID: 28541459 DOI: 10.1093/pubmed/fdx060]

26 **UK Statistics Authority**. Ethical considerations in the use of geospatial data for research and statistics. [cited 30 May 2023]. Available from: https://uksa.statisticsauthority.gov.uk/publication/ethical-considerations-in-the-use-of-geospatial-data-for-research-and-statistics/

27 **US Census Bureau**. An Overview of Addressing Nonresponse Bias in the American Community Survey During the COVID-19 Pandemic Using Administrative Data. 2021. [cited 30 May 2023]. Available from: <https://www.census.gov/newsroom/blogs/random-samplings/2021/11/nonresponse-acs-covid-administrative-data.html>

**Footnotes**

**Conflict-of-interest statement:** All the authors report no relevant conflicts of interest for this article.

**Open-Access:** This article is an open-access article that was selected by an in-house editor and fully peer-reviewed by external reviewers. It is distributed in accordance with the Creative Commons Attribution NonCommercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: https://creativecommons.org/Licenses/by-nc/4.0/

**Provenance and peer review:** Invited article; Externally peer reviewed.

**Peer-review model:** Single blind

**Peer-review started:** May 31, 2023

**First decision:** August 24, 2023

**Article in press:**

**Specialty type:** Medical laboratory technology

**Country/Territory of origin:** United Kingdom

**Peer-review report’s scientific quality classification**

Grade A (Excellent): 0

Grade B (Very good): B, B

Grade C (Good): 0

Grade D (Fair): 0

Grade E (Poor): 0

**P-Reviewer:** Wang KJ, China; Hariyanto IT, Indonesia **S-Editor:** Wang JJ **L-Editor:** A **P-Editor:**

**Figure Legends**



**Figure 1 World map showing which countries and territories have produced national census data since 2015.** The figure is created from Powerpoint insert map tool.