




## Editorial

## Ecological and individualistic fallacies in health disparities research

Ya-Chen Tina Shih , PhD,<sup>1\*</sup> Cathy Bradley , PhD,<sup>2</sup> K. Robin Yabroff , PhD<sup>3</sup><sup>1</sup>Section of Cancer Economics and Policy, Department of Health Services Research, University of Texas MD Anderson Cancer Center, Houston, TX, USA<sup>2</sup>University of Colorado Comprehensive Cancer Center and Department of Health Systems, Management & Policy, Colorado School of Public Health, Aurora, CO, USA<sup>3</sup>Department of Surveillance and Health Equity Science, American Cancer Society, Atlanta, GA, USA**\*Correspondence to:** Ya-Chen Tina Shih, Department of Health Services Research, University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd, Unit 1444, Houston, TX 77030, USA (e-mail: yashih@mdanderson.org).

Ecological fallacy, defined by Robinson in 1950 as incorrect inferences about individuals based on characteristics and associations observed among groups (1), is a well-recognized concept in epidemiology and statistics. Research has shown when aggregate values for variables of interest (eg, median area-level income) were used as proxies for individual-level variables (eg, household income), covariates estimated from regression models may be biased (2,3) and the sign of regression coefficients could change (4). Policy makers and health-care providers who rely on these estimations could inadvertently draw the wrong conclusions or target the wrong group for interventions.

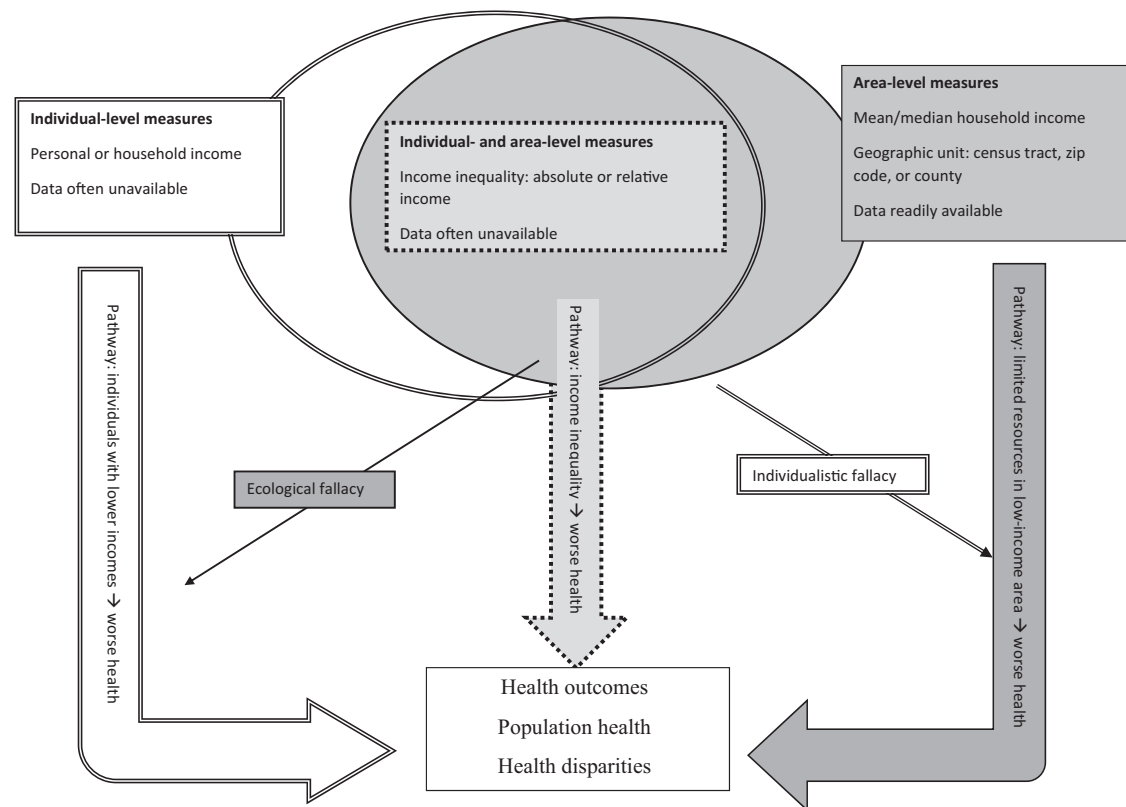
Health disparities research, including cancer disparities research, using observational data from registries, medical records, or administrative claims often lacks information on individual-level socioeconomic status (SES) variables, such as income, educational attainment, and employment status. Many studies use aggregate statistics at selected geographic units (eg, county, zip code, or census tract) as a substitute for individual-level SES. This approach, known as the census-based approach (5–7), is common practice in disparities research, where these variables are treated as “proxies” for individual SES and interpreted as if SES had been measured among individuals. Although some studies acknowledge ecological fallacy as a limitation, the use of the census-based approach is widely accepted by researchers as well as peer reviewers and is frequently viewed as an inevitable compromising analytical strategy driven by the lack of individual-level data on SES variables.

Davis et al. (8) made an important contribution to the literature of cancer disparities research by identifying appropriate data for neighborhood and individual income to showcase the issue of ecological fallacy when linking area-level factors to individual outcomes. The authors documented poor agreement between neighborhood and individual income measures, especially in rural communities (8). In addition, associations between neighborhood income and survival among patients with colorectal cancer were much smaller than associations of individual income and survival (8), suggesting that misclassification bias from using neighborhood income as a proxy for individual income may contribute to an underestimation of the income effect. Findings from this study serve as a cautionary tale for

cancer disparities research, especially for researchers exploring the association between income inequality, social determinants of health (SDoH), and rurality and health outcomes.

An extension of Davis et al. (8) is to expand the literature on income inequality and health (9) to cancer-related outcomes. The income variable, quantified as neighborhood or individual income, can be used to construct income inequality measures, such as Gini index (10), decile ratios (11), or Robin Hood index (12). This information can then be used to test hypotheses to explain the pathway from income inequality and health, such as absolute income, relative income, deprivation, or relative position hypothesis (13). Each hypothesis offers a different mechanism of potential policy actions to improve population health and reduce disparities. For example, empirical evidence supporting the absolute income hypothesis suggests reducing health disparities through improving incomes for all individuals, whereas the deprivation hypothesis recommends improving income and reducing income inequality specifically for individuals who are poor. The ability to distinguish and test these hypotheses hinges on having individual-level data because one cannot rely on neighborhood income alone to determine the relative economic standing of individuals in their communities. Furthermore, studies relying on aggregate incomes have shown conflicting findings, with stronger association between income inequality and health reported mostly in studies in the United States but not in many other countries. This observation, nicknamed “American exceptionalism” (14), has led Kawachi and Kennedy (15) to ponder the way income inequality manifested in the United States appears to differ from other countries. That is, income at an aggregate level appears to represent a socioeconomic construct at the ecologic level and should be analyzed and interpreted as such.

Many socioeconomic constructs exist mainly at the ecologic level, especially SDoH, which are the economic and environmental conditions under which people are born, live, work, play, and age that affect health, well-being, quality of life, morbidity, and mortality (16,17). Structural factors at the national, state, and local levels include aspects of the physical environment (eg, availability of safe and affordable housing, public transportation); laws, regulations, and policies (eg, housing assistance policies,



**Figure 1.** Conceptual framework on the relationship between income-related measures and health. This figure uses income-related measures to illustrate 3 pathways that explain the relationship between income and health. Each pathway is associated with a different measure of income, including individual income (**double line border**), area-level income (**single line border**), and income inequality (**dashed line border**), which combines income at the individual and area levels. This figure also comments on data availability and visualizes individual vs ecological fallacy.

paid sick leave mandates); and aspects of the social environment (eg, structural racism) can affect individual access to health care from prevention, screening, and diagnosis to treatment, survivorship, and end-of-life care. Many aspects of SDoH underlie the long-standing inequities in cancer outcomes by geography, SES, and race and ethnicity in the United States (16). For example, discriminatory housing practices, such as historic redlining, began limiting mortgage lending during the 1930s in neighborhoods largely inhabited by Black populations that were mapped in red. By limiting home ownership, historic redlining also limited intergenerational wealth transfer and resulted in disinvestment in these neighborhoods and lower area-level income. Contemporary residence in historically redlined neighborhoods is associated with later-stage cancer diagnoses, worse treatment, and higher cancer mortality rates (18-20). In this context, one must be wary of individualistic fallacy to avoid generalizing neighborhood-level relationships from only individual-level characteristics (21).

Davis et al. (8) reported that neighborhood income was particularly inaccurate for patients who lived in rural communities. Several reasons for this observation exist; each reinforces the need to exercise caution when interpreting area-level variables, especially in rural areas. First, people who live in rural areas often have addresses that correspond to a post office box that could be at considerable distance from where they live. Therefore, neighborhood income may reflect attributes of a different community than the community in which the patient resides. This problem is exacerbated by large rural communities described as Frontier and Remote (22). Frontier and Remote areas are characterized by low population size and high geographic remoteness, which is determined by the time it takes to travel by car to the edges of a

nearby urban area. Second, because of low population size, the area for which information is released for individuals is often much larger for rural residents than would be released for individuals who live in urban areas. For example, in a densely populated area, it may be possible for researchers to obtain neighborhood income at a small geographic unit such as a census block, whereas only county-level income may be available for an individual living in a rural area. Such data restrictions are intended to protect the patient's identity. Often, area-level data used to link to census information are missing from public use datasets altogether for patients who live in sparsely populated areas. Third, the oncology care team for rural patients may be located in a different community that is not necessarily adjacent or closest to where the patient lives. Rural residents who have the means to travel to a National Cancer Institute–designated center, for example, may receive more state-of-the-art treatment than their urban counterparts (23). Thus, the ability to travel to a National Cancer Institute center cannot be accurately predicted by median neighborhood income.

What can the research community do to mitigate biases resulting from ecological fallacy? An obvious solution to avoiding the issue of ecological fallacy is to improve the data quality by developing creative algorithms to make individual-level variables accessible without compromising patient confidentiality. Examples include use of categorical variables (eg, income category) to mask individual values while preserving the variations and relative difference across individuals. This will require concerted effort from government agencies that release public use databases for research and govern the terms of user agreement as well as efforts from data vendors of proprietary databases.

Another solution is through innovations in research methodology. One such example is the spatially-adjusted Bayesian additive regression tree model that borrows individual-level variables from another database through geographic commonality between the 2 databases (4). This model-based approach, however, is computationally intensive and requires a level of geographic granularity (eg, zip code) that often is not available in public use data. A third solution is through novel privacy preservation record linkage methods, such as tokenization of patient records through hashing technique (24). Although the performance of these data linkage methods is satisfactory (25), it requires considerable resources to create “tokens” because it requires licensing both the data linkage software and databases to be tokenized. All of these approaches, however, may reduce the timeliness underlying data for research.

It is critical for health disparities researchers to generate high-quality empirical evidence to identify populations living in vulnerable conditions to accurately inform policy actions. Figure 1 uses income-related measures to exemplify 3 plausible pathways that explain the relationship between income and health, highlighting the importance of understanding the implications of adopting measures at various levels. Without the availability of perfect data that provide both neighborhood and individual SES variables, accessibility of robust and user-friendly statistical methods, or affordable software with validated data linkage techniques, researchers using observational data to conduct disparities research must be mindful of ecological fallacy. Even when individual-level SES variables are available, it is important to mitigate individualistic fallacy by including variables at both the individual and area levels and applying multilevel statistical models to properly analyze cross-level processes (26). Davis et al. (8) illustrates the importance for disparities research to conceptualize measures at the appropriate level and exercise caution in interpreting study findings.

## Data availability

Not applicable.

## Author contributions

Ya-Chen Tina Shih, PhD (Conceptualization; Resources; Supervision; Writing – original draft; Writing – review & editing) Cathy J. Bradley, PhD (Conceptualization; Writing – original draft; Writing – review & editing) K. Robin Yabroff, PhD (Conceptualization; Writing – original draft; Writing – review & editing).

## Funding

Shih acknowledges funding from the National Cancer Institute (R01CA207216, R01CA225646), Bradley acknowledges the National Cancer Institute grant (P30CA46934), and Yabroff acknowledges funding from the National Cancer Institute (R01CA269488).

## Conflicts of interest

Y-CTS received consulting fees, travel, and accommodations for serving on a grants review panel for Pfizer Inc and an advisory board for AstraZeneca in 2019. KRY serves on the Flatiron Health Equity Advisory Board. YKR, a JNCI Deputy Editor, and author of this editorial, was not involved in the editorial review of the

manuscript or decision to publish the editorial. CJB and Y-CTS, JNCI Associate Editors and authors of this editorial, were not involved in the editorial review of the manuscript or decision to publish the editorial.

## Acknowledgements

The funder had no role in the writing of this editorial or the decision to submit it for publication.

## References

1. Robinson WS. Ecological correlations and the behavior of individuals. *Am Sociol Rev.* 1950;15(3):351-357.
2. Schwartz S. The fallacy of the ecological fallacy: the potential misuse of a concept and the consequences. *Am J Public Health.* 1994;84(5):819-824.
3. Willis A, Krewski D, Jerrett M, et al. Selection of ecologic covariates in the American Cancer Society study. *J Toxicol Environ Health Part A.* 2003;66(16-19):1563-1589.
4. Zhang S, Shih YCT, Muller P. A spatially-adjusted Bayesian additive regression tree model to merge two datasets. *Bayesian Anal.* 2007;3(2):611-634.
5. Devesa SS, Diamond EL. Socioeconomic and racial differences in lung cancer incidence. *Am J Epidemiol.* 1983;118(6):818-831.
6. Geronimus AT, Bound J. Use of census-based aggregate variables to proxy for socioeconomic group: evidence from national samples. *Am J Epidemiol.* 1998;148(5):475-486.
7. Gornick ME, Eggers PW, Reilly TW, et al. Effects of race and income on mortality and use of services among Medicare beneficiaries. *New Engl J Med.* 1996;335(11):791-799.
8. Davis LE, Mahar AL, Strumpf EC. Agreement between individual and neighborhood income measures in patients with colorectal cancer in Canada. *J Natl Cancer Inst.* 2023;115(5):514-522.
9. Lynch J, Smith GD, Harper S, et al. Is income inequality a determinant of population health? Part 1. A systematic review. *Milbank Q.* 2004;82(1):5-99.
10. Sen AK. *On Economic Inequality.* Oxford: Clarendon Press; 1973.
11. Wilkinson RG. Income distribution and life expectancy. *BMJ.* 1992;304(6820):165-168.
12. Kennedy BP, Kawachi I, Prothrow-Stith D. Income distribution and mortality: cross sectional ecological study of the Robin Hood index in the United States. *BMJ.* 1996;312(7037):1004-1007.
13. Wagstaff A, van Doorslaer E. Income inequality and health: what does the literature tell us? *Annu Rev Public Health.* 2000;21:543-567.
14. Muntaner C. Teaching social inequalities in health: barriers and opportunities. *Scand J Public Health.* 1999;27(3):161-165.
15. Kawachi I, Kennedy BP. *Health of Nations. Why Inequality is Harmful to Your Health.* New York: The New Press; 2006.
16. Alcaraz KI, Wiedt TL, Daniels EC, et al. Understanding and addressing social determinants to advance cancer health equity in the United States: a blueprint for practice, research, and policy. *CA Cancer J Clin.* 2020;70(1):31-46.
17. WHO. *Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health: Commission on Social Determinants of Health Final Report.* WHO Commission on Social Determinants of Health, World Health Organization; 2008.
18. Zhou Y, Bermanian A, Beyer KM. Housing discrimination, residential racial segregation, and colorectal cancer survival in southeastern Wisconsin. *Cancer Epidemiol Biomarkers Prev.* 2017;26(4):561-568.

19. Bikomeye JC, Zhou Y, McGinley EL, et al. Historical redlining and breast cancer treatment and survival among older women in the US. *J Natl Cancer Inst.* 2023. doi:10.1093/jnci/djad034.
20. Fan Q, Nogueira L, Yabroff KR, et al. Housing and cancer care and outcomes: a systematic review. *J Natl Cancer Inst.* 2022;114(12):1601-1618.
21. Alker HA Jr. A typology of ecological fallacies. In: Dogan M, Rokkan S, eds. *Quantitative Ecological Analysis*. Cambridge, MA: Massachusetts Institute of Technology; 1969:69-86.
22. U.S. Department of Agriculture. Frontier and remote area codes. <https://www.ers.usda.gov/data-products/frontier-and-remote-area-codes/>. Accessed August 20, 2019.
23. Bradley CJ, Eguchi M, Perrailon MC. Factors associated with use of high-cost agents for the treatment of metastatic non-small cell lung cancer. *J Natl Cancer Inst.* 2020;112(8):802-809.
24. Vatsalan D, Christen P, Verykios VS. A taxonomy of privacy-preserving record linkage techniques. *Inf Syst.* 2013;38(6):946-969.
25. Mirel LB, Resnick DM, Aram J, et al. A methodological assessment of privacy preserving record linkage using survey and administrative data. *Stat J IAOS.* 2022;38(2):413-421.
26. Subramanian SV, Jones K, Kaddour A, et al. Revisiting Robinson: the perils of individualistic and ecologic fallacy. *Int J Epidemiol.* 2009;38(2):342-360. Author reply 370-373.