DOI: 10.1111/1475-6773.14026

METHODS ARTICLE

HSR Health Services Research

Describing and assessing a new method of approximating categorical individual-level income using community-level income from the census (weighting by income probabilities)

Uriel Kim PhD[1,2,3,4] | Siran M. Koroukian PhD[2,3,5] | Kurt C. Stange MD, PhD[1] | James C. Spilsbury PhD[3] | Weichuan Dong PhD[3,5] | Johnie Rose MD, PhD[1,2,5]

[1]Center for Community Health Integration, Case Western Reserve University School of Medicine, Cleveland, Ohio, USA

[2]Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, Ohio, USA

[3]Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine, Cleveland, Ohio, USA

[4]Kellogg School of Management, Northwestern University, Evanston, IL, USA

[5]Population Cancer Analytics Shared Resource, Case Comprehensive Cancer Center, Cleveland, Ohio, USA

Correspondence
Uriel Kim, Center for Community Health Integration, Case Western Reserve University School of Medicine, Cleveland, OH 44106-7136, USA
Email: uriel.kim@case.edu

Funding information
American Cancer Society, Grant/Award Numbers: 132678-RSGI-19-213-01-CPHPS, RWIA-20-111-02 RWIA; National Cancer Institute, Grant/Award Number: P30 CA043703; National Center for Advancing Translational Sciences, Grant/Award Numbers: 1TL1TR002549, UL1TR000439; National Institute of General Medical Sciences, Grant/Award Number: 5T32GM007250; Pharmaceutical Research and Manufacturers of America Foundation, Grant/Award Number: PDHO18; National Institutes of Health, Grant/Award Numbers: UH3-DE025487, R15 NR017792; Ohio Medicaid Technical Assistance and Policy Program; The Centers for Disease Control and Prevention, Grant/Award Numbers: U48 DP006404-03S7, U48 DP005030-05S1

Abstract

Objective: To assess a new approach (weighting by “income probabilities [IP]”) that uses US Census data from the patients' communities to approximate individual-level income, an important but often missing variable in health services research.

Data Sources: Community (census tract level) income data came from the 2017 5-year American Community Survey (ACS). The patient data included those diagnosed with cancer in 2017 in Ohio ($n = 65{,}759$). The reference population was the 2017 5-year ACS Public Use Microdata Sample ($n = 564{,}357$ generalizing to 11,288,350 Ohioans).

Study Design/Methods: We applied the traditional approach of income approximation using median census tract income along with two IP based approaches to estimate the proportions in the patient data with incomes of 0%–149%, 150%–299%, 300%–499%, and 500%+ of the federal poverty level (FPL) (“class-relevant income grouping”) or 0%–138%, 139%–249%, 250%–399%, and 400%+ FPL (“policy-relevant income grouping”). These estimated income distributions were then compared with the known income distributions of the reference population.

Data Collection/Extraction Methods: The patient data came from Ohio's cancer registry. The other data were publicly available.

This manuscript represents independent research supported in part by National Institute of General Medical Sciences (5T32GM007250); National Center for Advancing Translational Science (1TL1TR002549); Pharmaceutical Research and Manufacturers of America Foundation (PDHO18); National Cancer Institute, Case Comprehensive Cancer Center (P30 CA043703); The Centers for Disease Control and Prevention, U48 DP005030-05S1 and U48 DP006404-03S7; Ohio Medicaid Technical Assistance and Policy Program (MEDTAPP); National Institutes of Health (N and UH3-DE025487); The American Society (132678-RSGI-19-213-01-CPHPS and RWIA-20-111-02); and contracts from Cleveland Clinic Foundation, including a subcontract from Celgene Corporation.

Principal Findings: Both IP based approaches consistently outperformed the traditional approach overall and in subgroup analyses, as measured by the weighted average absolute percentage point differences between the proportions of each of the income categories of the reference population and the estimated proportions generated by the income approximation approaches ("average percent difference," or APD). The smallest APD for an IP based method, 0.5%, was seen in non-Hispanic White females in the class-relevant income grouping (compared with 16.5% for the conventional method), while the largest APD, 7.1%, was seen in non-Hispanic Black females in the policy-relevant income grouping (compared with 18.0% for the conventional method).

Conclusions: Weighting by IP substantially outperformed the conventional approach of estimating the distribution of incomes in patient data.

KEYWORDS

residence characteristics/statistics and numerical data, social class, income/statistics and numerical data, health status disparities, censuses, data collection/methods

What is known on this topic

- Individual-level income is often a key variable of interest in biomedical research studies since the impact of socioeconomic status on health has been repeatedly demonstrated.
- When individual-level income is not directly available, it is often proxied from community-level data using approaches that can lead to significant misclassifications.

What this study adds

- This study finds that using income probabilities (IP) can more accurately estimate individual-level income compared with the traditional approach of proxying individual-level income using median census tract income.
- Weighting by IP performs particularly well at the extremes of the income distribution, which is often of key importance in health disparities studies.

1 | INTRODUCTION

Individual-level measures of income are often a key variable of interest in biomedical research studies since the impact of socioeconomic status on health has been repeatedly demonstrated. However, clinical data rarely capture this information directly, given the impracticality of soliciting income data from patients in health care settings. In datasets where patients' communities of residence are known, researchers have approximated individual-level income using community-level income from the US Census.

The most common and conventional implementation of this approach assigns individuals the median income of their communities of residence.[1–4] The infidelity between community-level (aggregate) and individual-level income has been documented in a broad range of settings[5–9] and can lead to large numbers of income misclassifications.[3,10,11] Nonetheless, because individual-level income proxied using median community-level income can often explain at least some of the variability observed in study outcomes,[3] and the lack of reasonable alternatives, such an approach is generally accepted.[12]

In studies where income is a key exposure or a stratifying variable of interest, more accurate approaches to income approximation are desirable. Thus, this study describes and assesses a method of approximating individual-level income from community-level income data, coined here as weighting by "income probabilities [IP]," that can potentially reduce income misclassification compared with conventional approaches. When weighting by IP, the probability of community residents having a certain, investigator-defined categorical income value is calculated using community-level count income data from the American Community Survey (ACS) of the US Census. These probabilities are then used as observation weights for individual patients, leading to synthetic datasets that are stratified into the different investigator-defined income categories.

This IP based approach holds theoretical promise for two reasons. First, it is a "probabilistic" method of individual-level income approximation compared with conventional "deterministic" approaches. Conventional approaches are deterministic in the sense that all patients from the same community are proxied a single income value. A deterministic approach can lead to many misclassifications by underestimating the numbers of patients on both the low and high end of the income spectrum, while overestimating the number of patients in the middle of the spectrum. Conversely, in weighting by IP, patients are assigned a probability of having a particular categorical income given their community of residence rather than a specific income value. When these probabilities are used as observation weights for the patients, and when the patients are aggregated, the distribution of the approximated incomes of the patients is expected to faithfully recapitulate the income distribution of the patients' communities of residence (assuming the patients are representatively sampled from their communities). The second related advantage of weighting by IP is that no patients are explicitly excluded from a particular income category. This preserves power compared with conventional approaches.

To evaluate whether the theoretical promise of weighting by IP translates practically, we evaluated two proposed IP based approaches against the conventional approach of assigning individual-level income categories to patients using median community-level income data. The three approaches were assessed by applying them to patient data from the Ohio Cancer Incidence Surveillance System (or OCISS, Ohio's state cancer registry) and comparing the resulting distribution of the income categorizations against a reference population derived from the ACS's Public Use Microdata Sample for Ohio.

2 | METHODS

2.1 | The calculation of IP and the general approach of weighting by IP

An investigator interested in identifying patients with incomes less than 200% of the federal overty level (FPL) might proxy individual-level income for the patients using the median community-level income data from their communities of residence in the conventional, prevailing approach. Commonly, census tract-level community income data from the US Census are used. Thus, the presumed patients with individual-level incomes <200% FPL would only include those living in census tracts where the median census tract income is <200% FPL.

Like the conventional approach of income estimation, weighting by IP uses community-level (census tract) income data from the US Census to approximate individual-level income. In its most basic form, IP are calculated by simply taking the count of individuals within a certain income bracket in each census tract, divided by the total population of that census tract. To extend the above example, an investigator interested in identifying patients with incomes <200% FPL using weighting by IP would calculate two IP for each census tract—one for <200% FPL and one for 200%+ FPL. Once these IP for the census tracts are created, they are applied back to the individual patients based on the patient's census tract of residence. The IP are then used as observation weights to generate income-specific statistics for the patients.

Consider an example of a dataset of 10 patients for which an investigator wishes to estimate their incomes. From the available patient address information, it is known that half of these patients live in a "wealthy" census tract ("Community B"), and half live in a "poorer" census tract ("Community A"). From the US Census, it is ascertained that a total of 50 people live in the wealthy census tract, of which 20 have incomes <200% FPL. Thus, the probability of having an individual-level income of <200% FPL is 0.40 in this census tract (and the probability of having an individual-level income of 200%+ FPL is 0.60). In the poorer census tract, the US Census indicates that there are 50 residents, of which 40 have incomes <200% FPL; thus, the probability of having an individual-level income <200% FPL is 0.80 for this census tract, while the probability of having an individual-level income of 200%+ FPL is 0.20. These probabilities are the IP. The IP are then applied back to the 10 patients using their census tracts of residence (for example, each of the five patients from the wealthy census tract is assigned a "<200% FPL income probability" of "0.40" and a "200%+ FPL income probability" of "0.60," while each of the five patients from the poorer census tract is assigned a "<200% FPL income probability" of "0.80" and a "200%+ FPL income probability" of "0.20"). To calculate the total number of patients estimated to have incomes <200% FPL, the "<200% FPL income probability" for each patient is used as the observation weight. Thus, $(5 \text{ patients} \times 0.40) + (5 \text{ patients} \times 0.80)$, or 6 patients would be estimated to have incomes of <200% FPL. By extension, $(5 \text{ patients} \times 0.60) + (5 \text{ patients} \times 0.20)$, or 4 patients would be estimated to have incomes of 200%+ FPL.

Note that in weighting by IP, no patients are explicitly excluded from a specific income category of interest as in the conventional approach. Rather, a patient's contribution to an income category subpopulation is modulated by the act of weighting by the IP. In the previous example, patients hailing from the poorer census tract were assigned an income probability of 0.80, which was closer to the maximum value of 1. Since these patients were highly likely to have an individual-level income of interest (<200% FPL) based on their census tract of residence, they contributed maximally to the estimated subpopulation comprised of individuals with incomes <200% FPL. Conversely, patients hailing from the wealthier community were assigned an income probability of 0.40, which was closer to the minimum value of 0. Thus, they contributed comparatively less to the estimated subpopulation comprised of individuals with incomes <200% FPL. Figure 1 provides a conceptual summary of conventional versus IP based approaches.

The weighting by IP approach can be extended to classify patients into multiple investigator-specified income categories. For example, an investigator interested in classifying patients into incomes of 0%–100%, 101%–200%, and 201%+ FPL would first calculate IP for each of the three income categories ("$\text{IP}^{0\%–100\%}$," "$\text{IP}^{101\%–200\%}$," and "$\text{IP}^{201\%+}$") for all of the census tracts in the study area. Then, applying $\text{IP}^{0\%–100\%}$, $\text{IP}^{101\%–200\%}$, and $\text{IP}^{201\%+}$ to the patient data creates three synthetic datasets containing those with an estimated individual-level income of 0%–100%, 101%–200%, and 201%+ FPL, respectively. Note that the IP for a given census tract always sum to 1 (e.g., $\text{IP}^{0\%–100\%} + \text{IP}^{101\%–200\%} + \text{IP}^{201\%+} = 1$, for a given census tract), a key feature of this approach. This feature ensures that the total counts across the income-specific

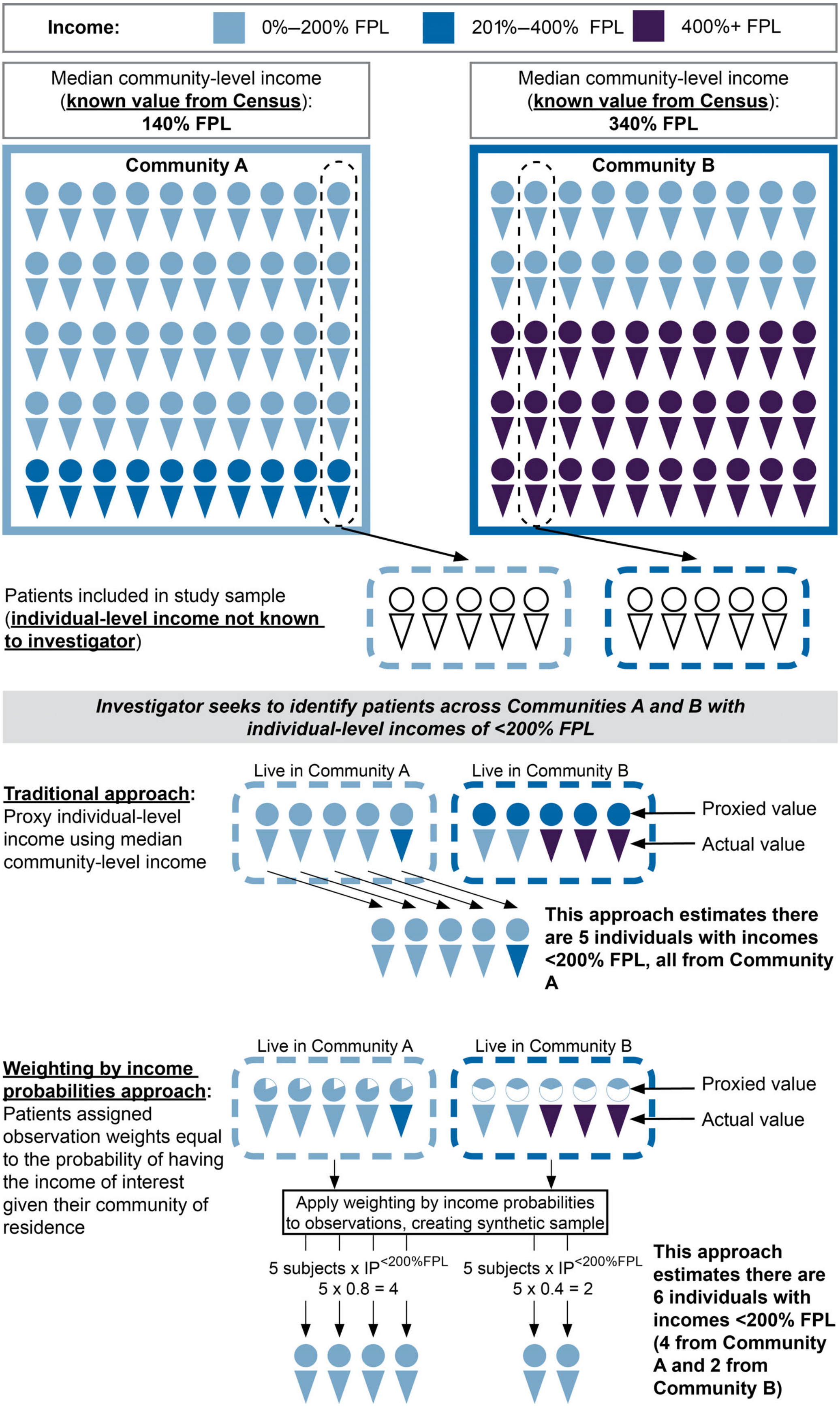

Income:
0%–200% FPL
201%–400% FPL
400%+ FPL
Median community-level income (known value from Census): 140% FPL
Median community-level income (known value from Census): 340% FPL
Community A
Community B
Patients included in study sample (individual-level income not known to investigator)
Investigator seeks to identify patients across Communities A and B with individual-level incomes of <200% FPL
Traditional approach: Proxy individual-level income using median community-level income
Live in Community A
Live in Community B
Proxied value
Actual value
This approach estimates there are 5 individuals with incomes <200% FPL, all from Community A
Weighting by income probabilities approach: Patients assigned observation weights equal to the probability of having the income of interest given their community of residence
Live in Community A
Live in Community B
Proxied value
Actual value
Apply weighting by income probabilities to observations, creating synthetic sample
5 subjects x IP<200%FPL
5 x 0.8 = 4
5 subjects x IP<200%FPL
5 x 0.4 = 2
This approach estimates there are 6 individuals with incomes <200% FPL (4 from Community A and 2 from Community B)

FIGURE 1 Conceptual summary of conventional versus the weighting by income probabilities approach to individual-level income approximation. FPL, federal poverty level; IP, income probabilities [Color figure can be viewed at wileyonlinelibrary.com]

subpopulations generated by weighting by income probability equal the original, unweighted count of patients.

2.2 | Community-level income data from the US Census

The ACS program of the US Census is an authoritative source of community-level income estimates. The ACS asks more detailed questions on social, economic, housing, and demographic characteristics than the decennial census; additionally, sampling for the ACS occurs on a rolling basis (with 3.5 million Americans surveyed annually).[13] Thus, among the publicly available data from the US Census, the ACS provides the most detailed and timely statistics of community-level income. Since the ACS samples a smaller population annually than the decennial census, multiple years of data are aggregated to improve the reliability of estimates.[13] For example, "2017 ACS 5-year" estimates combine the data collected from 2013 to 2017.

The publicly available data from the ACS generally come pre-tabulated (in aggregate). Income in these pre-tabulated tables is either reported in relative (adjusting for family size and reported as income as a ratio of the FPL) or absolute terms (reported in inflation-adjusted dollars). Furthermore, within communities, income can be reported at the units of "households" (all persons living in a housing unit), "families" (all persons living in a housing unit headed by someone who is part of a family), or "individuals."[14] Finally, the ACS reports income statistics in counts, proportions, means, and medians. Not all combinations of these forms of income data are available from the ACS, and the availability also varies by the geographic unit.[15]

We preferred relative measures of income (i.e., FPL) since they are often more policy-relevant than absolute measures. FPL is calculated at the family unit (a family's total income in dollars is divided by the number of family members, which is then compared with established poverty thresholds), but can be attributed to individuals (all individuals from a family get assigned the same FPL value). We also preferred FPL data reported individuals (vs. families), since these data were more likely to be representative of the entire community (for example, the median family FPL in a community with many large, wealthy families would belie the number of wealthy individuals in a community). Most FPL data in the census are reported as counts (e.g., the number of individuals in a specific FPL category for each community).

When proxying individual-level income using community-level income, a granular geographic unit is preferable since aggregated data of smaller communities are more likely to be representative of the individual residents.[3] Commonly used geographic units are counties, zip codes, census tracts, and block groups (these units generally descend in size).[14] The evidence suggests that proxied values using census tract- or block group-level data are generally more reliable than those proxied from zip code-level data.[11,16–19] We opted to use census tract-level income data over block group-level data; while block groups represent a more granular geographic unit, the census tract-level estimates generally have smaller margins of error (higher statistical reliability). Additionally, the publicly available block group-level data often report categorical income variables more coarsely than analogous census tract-level data because of this issue of statistical reliability and the additional concern of respondent confidentiality.

2.3 | Individual-level income approximation methods evaluated

We sought to classify patients into two sets of income categories using both traditional and IP based approaches. The first set of income categories we specified was called the "class-relevant income grouping" and consisted of the following income categories: 0%–149%, 150%–299%, 300%–499%, and 500%+ FPL. These income categories generally correspond to the constructs of lower, lower-middle, middle, upper-middle/upper-class, though exact definitions vary widely.[20] The second set of income categories we specified was called the "policy-relevant income grouping" and consisted of the following income categories: 0%–138%, 139%–249%, 250%–399%, and 400%+ FPL. These income categories correspond to different income-based eligibility thresholds associated with the Affordable Care Act.[21]

"Approach 1" represented the conventional income approximation method. It used "2017 ACS 5-year Table B17024," which reports counts of individuals in pre-defined income categories (in FPL), stratified by 10 age groups, for census tracts. The pre-defined age categories in the table were <6, 6–11, 12–17, 18–24, 25–34, 35–44, 45–54, 55–64, 65–74, and 75+ years. The pre-defined income categories in the table were <50%, 50%–74%, 75%–99%, 100%–124%, 125%–149%, 150%–174%, 175%–184%, 185%–199%, 200%–299%, 300%–399%, 400%–499%, and 500%+ FPL. Using the count data across all age groups, the FPL category containing the 50th percentile observation was identified. The identified FPL category was the proxied individual income of patients. Of note, the categories of the policy-relevant income grouping do not fully correspond to the native FPL categories of the ACS. Therefore, some modest manipulation of "2017 ACS 5-year Table B17024" was needed before patients were assigned to the income categories of the policy-relevant income grouping. Specifically, the counts for the 125%–149% FPL and 200%–299% FPL income categories in "2017 ACS 5-year Table B17024" were split between 125%–138%/139%–149% FPL and 200%–249%/250%–399% FPL, respectively. The counts for these new income categories were allocated using linear interpolation, a reasonable approach given population-level income distribution patterns (see S1 for further detail). These new FPL income categories allow for patients to be assigned to all the income categories of the policy-relevant income grouping.

Approaches 2 and 3 were weighting by IP approaches. Approach 2 is the simplest implementation of weighting by IP, using the same census tract-level ACS table as Approach 1 ("2017 ACS 5-year Table B17024"). Using these data, IP for each of the four income categories of the class-relevant income and policy-relevant income groupings were calculated for each census tract.

Approach 3 was a slightly more sophisticated implementation. In Approaches 1 and 2, the 10 age strata of "2017 ACS 5-year

Table B17024" were consolidated into a single age group (all ages). Conversely, in Approach 3, the age-stratified income data of "2017 ACS 5-year Table B17024" were used to calculate age-specific IP for each of the four income categories of the class-relevant income and policy-relevant income groupings. Thus, the IP assigned to the observations in the patient dataset were age-dependent, accounting for the age-dependency of income. We strategically consolidated the 10 age strata in "2017 ACS 5-year Table B17024" to six (<18, 18–34, 35–44, 45–54, 55–64, 65+), reducing the overall complexity of Approach 3 and improving the statistical precision of the age-specific IP (see S2 for further detail).

S3 summarizes the conventional (Approach 1) and IP based (Approaches 2 and 3) methods evaluated in this study. S4 includes reproductions of "ACS Table B17024," along with the derived values needed for Approaches 1–3.

people.[23] Furthermore, the PUMS contains a limited set of variables (but includes sex, race, ethnicity, age, and individual income in % FPL).

Our measure of relative performance was "average percent difference." To calculate this measure, we first noted the absolute percentage point difference between the expected distribution (from the ACS PUMS reference data) and estimated distribution (using Approaches 1–3) of each of the income categories of the class-relevant and policy-relevant income groupings. The absolute percentage point differences were then averaged, weighted by the proportions of individuals in each income category in the reference population. Average percent difference is summarized by the following formula (with c = 1, 2, 3, 4 corresponding to the categories of 0%–149%, 150%–299%, 300%–499%, and 500%+ FPL for the class-relevant income grouping and 0%–138%, 139%–249%, 250%–399%, and 400%+ FPL for the policy-relevant income grouping, respectively):

$$\text{Average Percent Difference of Approach 1, 2, or 3} = \sum_{c=1}^{c=4} \frac{\%\text{income category c from ACS PUMS}}{100\%} \times \left|\%\,\text{income category c from ACS PUMS} - \text{Estimated}\,\%\,\text{income category c from OCISS using specified approach}\right|$$

2.4 | Assessment of income estimation approaches using patient data from the OCISS and reference data from the ACS Public Use Microdata Sample

We assessed Approaches 1–3 to understand the relative performance of each approach across sex and race and ethnicity groups, the income distribution (with particular attention to the lowest income categories), and different income categorizations (i.e., class-relevant vs. policy-relevant income grouping).

The patient data for which we wanted to estimate income came from the OCISS, Ohio's population-based cancer registry. All incident cases of cancer diagnosed or treated in the state must be reported to the OCISS.[22] Thus, the OCISS theoretically represents a complete, unbiased dataset of cancer patients in Ohio. The OCISS follows North American Association of Central Cancer Registries (NAACCR) data standards and contains detailed demographic information (including patient address at diagnosis), in addition to disease and treatment information. Our patient data included all those in the OCISS diagnosed with malignant cancer in 2017.

After applying Approaches 1–3 to the patient data from the OCISS, we compared the distribution of the resulting income estimates at the aggregate (state-wide) level against those from a reference population derived from the ACS Public Use Microdata Sample (PUMS) for Ohio. The PUMS differs from other ACS data because it is at the individual (un-tabulated) level. To maintain confidentiality, individuals are geographically identifiable only to "Public Use Microdata Areas," relatively large proprietary geographic areas consisting of approximately 100,000

The average percent difference was calculated by sex (male, female) and race and ethnicity (non-Hispanic White, non-Hispanic Black; race and ethnicity crosswalks between the OCISS and PUMS are provided in S5). Differences in age structure between the OCISS patient data and the ACS PUMS reference dataset were accounted for by directly age-adjusting both to the 2000 US Standard Population. When controlling for sex, race, ethnicity, and age structure, and when including all cancer types in the OCISS, we hypothesized that the best income estimation approach would closely recapitulate the expected income distributions established by the reference population. Thus, the best performing income estimation approach would be noted by the smallest average percent difference. The assessment procedure is summarized in Figure 2.

3 | RESULTS

The patient data from the OCISS included 65,759 individuals, of which 50.2% were female, 86.2% were non-Hispanic White, and 10.1% were non-Hispanic Black. The ACS PUMS reference population revealed that 26.5%, 26.9%, 24.9%, and 21.9% of female Ohioans had incomes of 0–149%, 150%–299%, 300%–499%, and 500%+ FPL, respectively; the distribution for men was 22.3%, 26.6%, 26.9%, and 24.3%, respectively. The income distribution for sex and race and ethnicity strata are provided in Figure 3 for the class-relevant income grouping, and analogous data for the policy-relevant income grouping are provided in Figure 4.

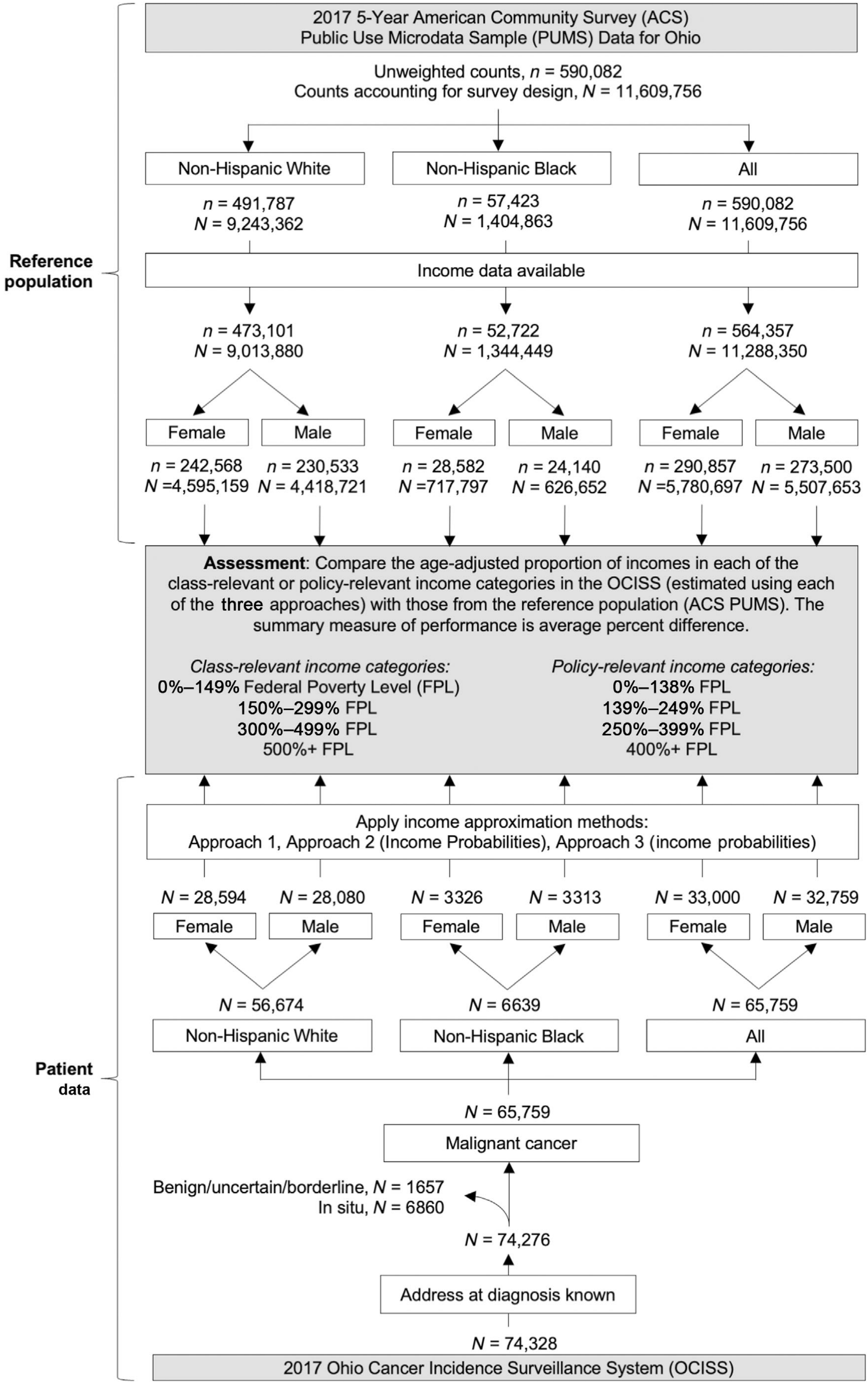

2017 5-Year American Community Survey (ACS)
Public Use Microdata Sample (PUMS) Data for Ohio
Unweighted counts, *n* = 590,082
Counts accounting for survey design, *N* = 11,609,756
Non-Hispanic White
Non-Hispanic Black
All
n = 491,787
N = 9,243,362
n = 57,423
N = 1,404,863
n = 590,082
N = 11,609,756
Reference population
Income data available
n = 473,101
N = 9,013,880
n = 52,722
N = 1,344,449
n = 564,357
N = 11,288,350
Female
Male
Female
Male
Female
Male
n = 242,568
N =4,595,159
n = 230,533
N = 4,418,721
n = 28,582
N =717,797
n = 24,140
N = 626,652
n = 290,857
N =5,780,697
n = 273,500
N = 5,507,653
Assessment: Compare the age-adjusted proportion of incomes in each of the class-relevant or policy-relevant income categories in the OCISS (estimated using each of the three approaches) with those from the reference population (ACS PUMS). The summary measure of performance is average percent difference.
Class-relevant income categories:
0%–149% Federal Poverty Level (FPL)
150%–299% FPL
300%–499% FPL
500%+ FPL
Policy-relevant income categories:
0%–138% FPL
139%–249% FPL
250%–399% FPL
400%+ FPL
Apply income approximation methods:
Approach 1, Approach 2 (Income Probabilities), Approach 3 (income probabilities)
N = 28,594
N = 28,080
N = 3326
N = 3313
N = 33,000
N = 32,759
Female
Male
Female
Male
Female
Male
N = 56,674
N = 6639
N = 65,759
Non-Hispanic White
Non-Hispanic Black
All
Patient data
N = 65,759
Malignant cancer
Benign/uncertain/borderline, *N* = 1657
In situ, *N* = 6860
N = 74,276
Address at diagnosis known
N = 74,328
2017 Ohio Cancer Incidence Surveillance System (OCISS)

FIGURE 2 Overview of assessment methodology for evaluating income estimation approaches

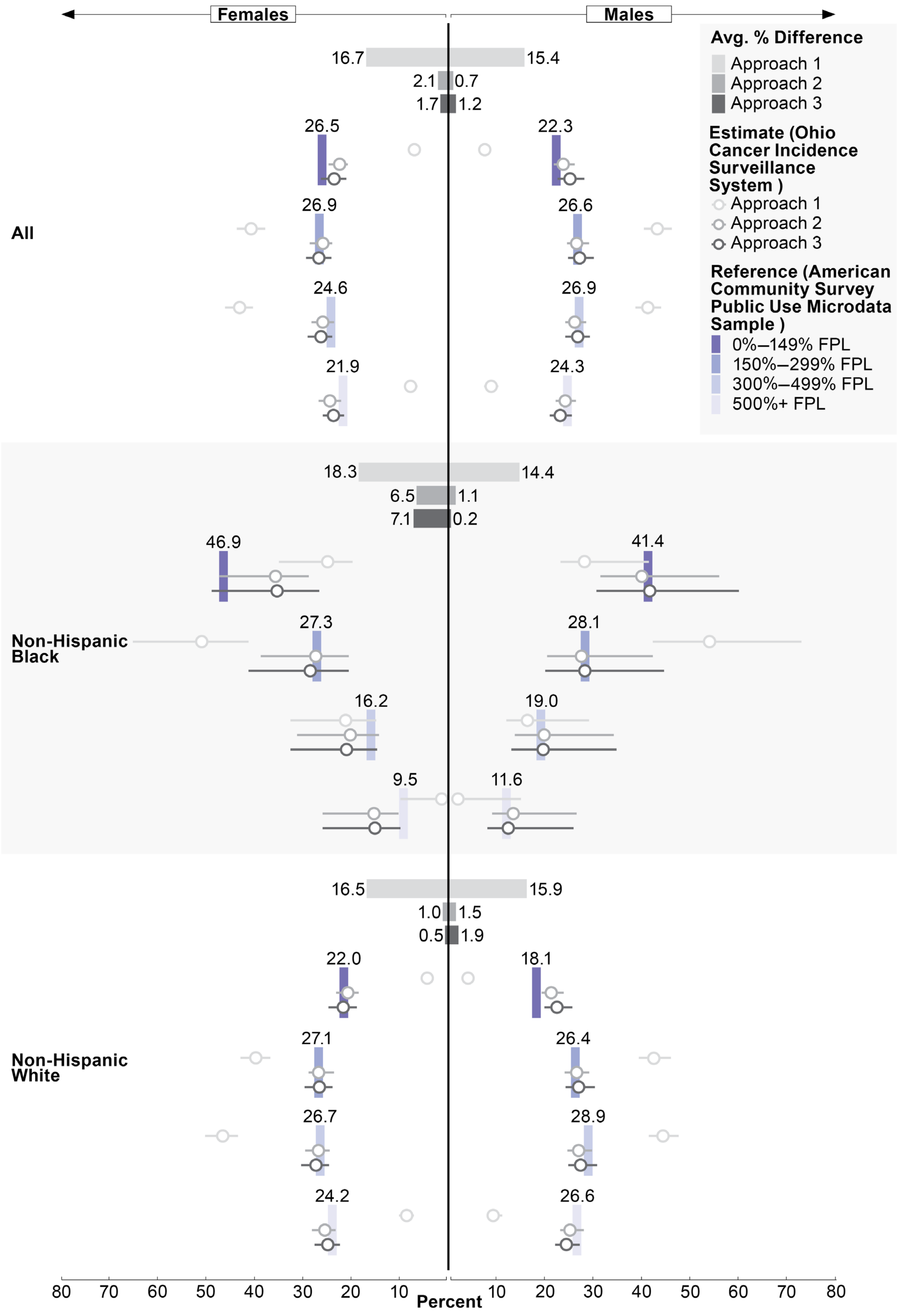


FIGURE 3 Assessment results for income estimation approaches for the class-relevant income grouping. FPL, federal poverty level [Color figure can be viewed at wileyonlinelibrary.com]

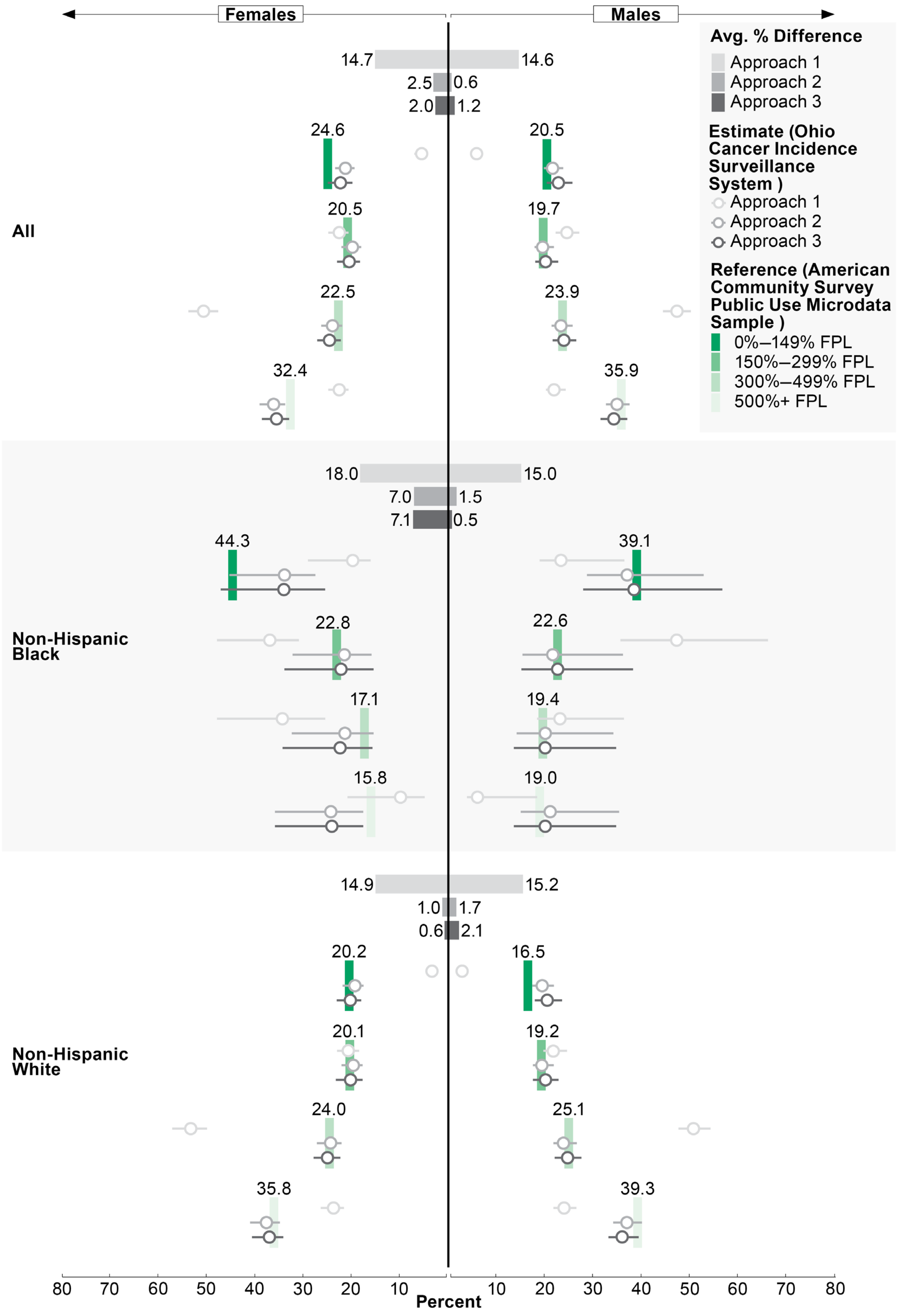
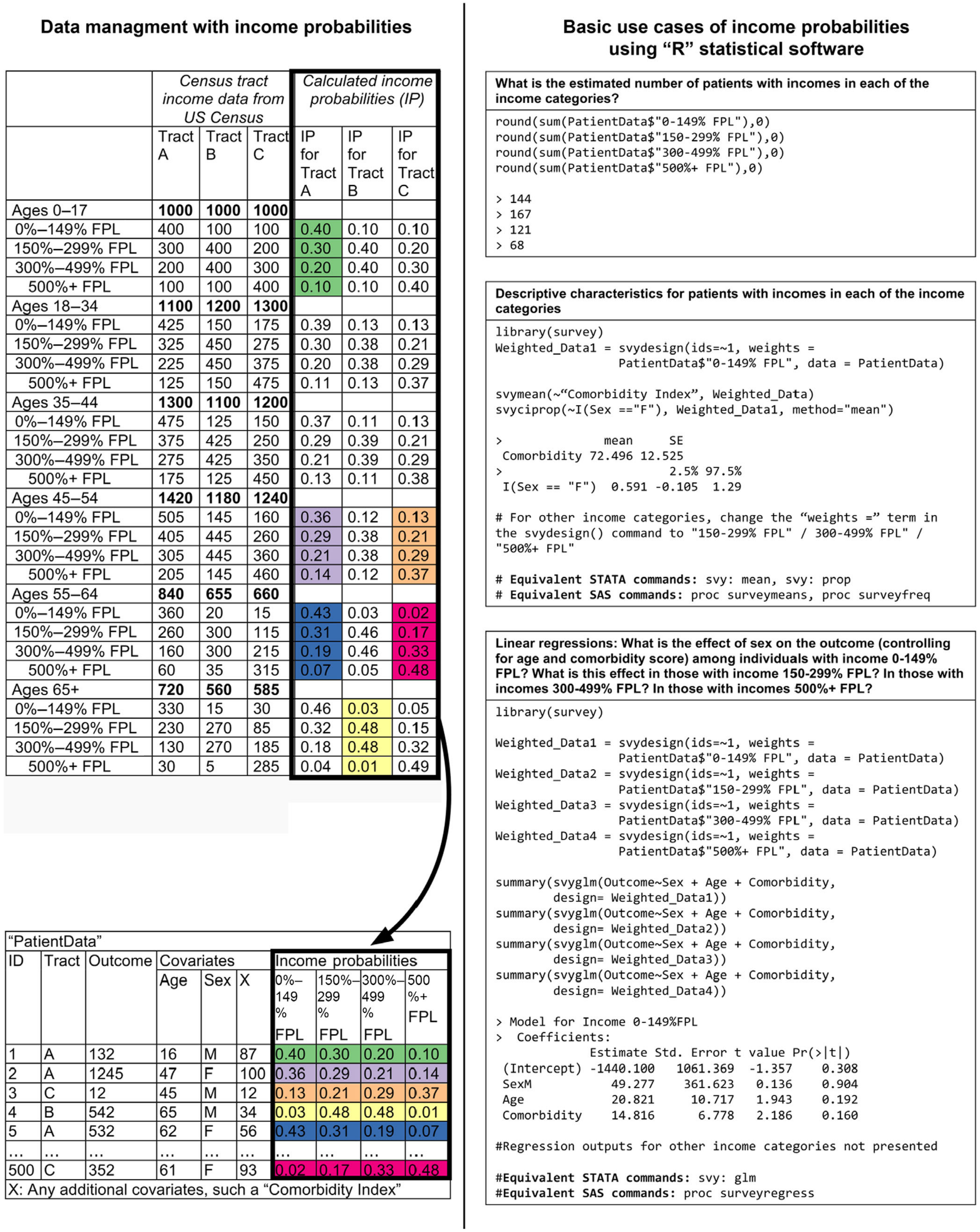


FIGURE 4 Assessment results for income estimation approaches for the policy-relevant income grouping. FPL, federal poverty level [Color figure can be viewed at wileyonlinelibrary.com]



Data managment with income probabilities

	Census tract income data from US Census			*Calculated income probabilities (IP)*		
	Tract A	Tract B	Tract C	IP for Tract A	IP for Tract B	IP for Tract C
Ages 0–17	**1000**	**1000**	**1000**			
0%–149% FPL	400	100	100	0.40	0.10	0.10
150%–299% FPL	300	400	200	0.30	0.40	0.20
300%–499% FPL	200	400	300	0.20	0.40	0.30
500%+ FPL	100	100	400	0.10	0.10	0.40
Ages 18–34	**1100**	**1200**	**1300**			
0%–149% FPL	425	150	175	0.39	0.13	0.13
150%–299% FPL	325	450	275	0.30	0.38	0.21
300%–499% FPL	225	450	375	0.20	0.38	0.29
500%+ FPL	125	150	475	0.11	0.13	0.37
Ages 35–44	**1300**	**1100**	**1200**			
0%–149% FPL	475	125	150	0.37	0.11	0.13
150%–299% FPL	375	425	250	0.29	0.39	0.21
300%–499% FPL	275	425	350	0.21	0.39	0.29
500%+ FPL	175	125	450	0.13	0.11	0.38
Ages 45–54	**1420**	**1180**	**1240**			
0%–149% FPL	505	145	160	0.36	0.12	0.13
150%–299% FPL	405	445	260	0.29	0.38	0.21
300%–499% FPL	305	445	360	0.21	0.38	0.29
500%+ FPL	205	145	460	0.14	0.12	0.37
Ages 55–64	**840**	**655**	**660**			
0%–149% FPL	360	20	15	0.43	0.03	0.02
150%–299% FPL	260	300	115	0.31	0.46	0.17
300%–499% FPL	160	300	215	0.19	0.46	0.33
500%+ FPL	60	35	315	0.07	0.05	0.48
Ages 65+	**720**	**560**	**585**			
0%–149% FPL	330	15	30	0.46	0.03	0.05
150%–299% FPL	230	270	85	0.32	0.48	0.15
300%–499% FPL	130	270	185	0.18	0.48	0.32
500%+ FPL	30	5	285	0.04	0.01	0.49

“PatientData”

ID	Tract	Outcome	Covariates			Income probabilities			
			Age	Sex	X	0%–149% FPL	150%–299% FPL	300%–499% FPL	500%+ FPL
1	A	132	16	M	87	0.40	0.30	0.20	0.10
2	A	1245	47	F	100	0.36	0.29	0.21	0.14
3	C	12	45	M	12	0.13	0.21	0.29	0.37
4	B	542	65	M	34	0.03	0.48	0.48	0.01
5	A	532	62	F	56	0.43	0.31	0.19	0.07
…	…	…	…	…	…	…	…	…	…
500	C	352	61	F	93	0.02	0.17	0.33	0.48

X: Any additional covariates, such a “Comorbidity Index”

Basic use cases of income probabilities using “R” statistical software

What is the estimated number of patients with incomes in each of the income categories?

```
round(sum(PatientData$"0-149% FPL"),0)
round(sum(PatientData$"150-299% FPL"),0)
round(sum(PatientData$"300-499% FPL"),0)
round(sum(PatientData$"500%+ FPL"),0)

> 144
> 167
> 121
> 68
```

Descriptive characteristics for patients with incomes in each of the income categories

```
library(survey)
Weighted_Data1 = svydesign(ids=~1, weights =
                 PatientData$"0-149% FPL", data = PatientData)

svymean(~“Comorbidity Index”, Weighted_Data)
svyciprop(~I(Sex =="F"), Weighted_Data1, method="mean")

>              mean     SE
 Comorbidity 72.496 12.525
>                       2.5% 97.5%
 I(Sex == "F")  0.591 -0.105  1.29

# For other income categories, change the “weights =” term in
the svydesign() command to "150-299% FPL" / 300-499% FPL" /
"500%+ FPL"

# Equivalent STATA commands: svy: mean, svy: prop
# Equivalent SAS commands: proc surveymeans, proc surveyfreq
```

Linear regressions: What is the effect of sex on the outcome (controlling for age and comorbidity score) among individuals with income 0-149% FPL? What is this effect in those with income 150-299% FPL? In those with incomes 300-499% FPL? In those with incomes 500%+ FPL?

```
library(survey)

Weighted_Data1 = svydesign(ids=~1, weights =
                 PatientData$"0-149% FPL", data = PatientData)
Weighted_Data2 = svydesign(ids=~1, weights =
                 PatientData$"150-299% FPL", data = PatientData)
Weighted_Data3 = svydesign(ids=~1, weights =
                 PatientData$"300-499% FPL", data = PatientData)
Weighted_Data4 = svydesign(ids=~1, weights =
                 PatientData$"500%+ FPL", data = PatientData)

summary(svyglm(Outcome~Sex + Age + Comorbidity,
        design= Weighted_Data1))
summary(svyglm(Outcome~Sex + Age + Comorbidity,
        design= Weighted_Data2))
summary(svyglm(Outcome~Sex + Age + Comorbidity,
        design= Weighted_Data3))
summary(svyglm(Outcome~Sex + Age + Comorbidity,
        design= Weighted_Data4))

> Model for Income 0-149%FPL
>  Coefficients:
             Estimate Std. Error t value Pr(>|t|)
 (Intercept) -1440.100   1061.369  -1.357    0.308
 SexM           49.277    361.623   0.136    0.904
 Age            20.821     10.717   1.943    0.192
 Comorbidity    14.816      6.778   2.186    0.160

#Regression outputs for other income categories not presented

#Equivalent STATA commands: svy: glm
#Equivalent SAS commands: proc surveyregress
```

FIGURE 5 Operationalizing weighting by income probabilities using common statistical software. FPL, federal poverty level [Color figure can be viewed at wileyonlinelibrary.com]

Using the conventional approach of income estimation (Approach 1), the average percent difference in the class-relevant income grouping was 16.7% for women and 15.4% for men. This compared with 2.1% and 0.7% for women and men using Approach 2 (weighting by IP), and 1.7% and 1.2% for women and men using Approach 3 (age-specific weighting by IP). Thus, the best performing IP based method represented a 15.0 percentage point (nearly 10-fold) improvement in income estimation in women, and a 14.7 percentage point (22-fold) improvement in income estimation in men. In non-Hispanic Black patients, the best performing IP based approach represented a 11.8 percentage point improvement in females and a 14.2 percentage point improvement in males. In non-Hispanic White patients, the best performing IP based approach represented a 16.0 percentage point improvement in females, and a 14.4 percentage point improvement in males. These findings are summarized in Figure 3.

In the class-relevant income grouping, the average percent difference was generally slightly higher for females than for males, although this relationship varied by race and ethnicity strata (e.g., Approaches 2 and 3 had higher average percent differences for non-Hispanic White males). Across all sex and race and ethnicity strata, Approach 1 consistently underestimated the numbers of patients in the lowest (0%–149% FPL) and the highest (500%+ FPL) income categories, while overestimating the numbers of patients in the middle-income categories (150%–299% and 300%–499% FPL) compared with Approaches 2 and 3. For example, against the reference of 26.5% of women having incomes 0%–149% FPL, Approach 1 estimated 7.1% (95% confidence interval [CI]: 6.1–8.5), while Approach 2 estimated 22.8% (95% CI: 20.9–25.1), and Approach 3 estimated 23.8% (95% CI: 21.3–26.6) of women to have that income.

Findings in the policy-relevant income grouping (Figure 4) were similar to those noted for class-relevant income grouping. With some exceptions, Approach 1 had a slightly smaller average percent difference in the policy-relevant income grouping compared with the class-relevant income grouping, while the converse was true for Approaches 2 and 3.

4 | DISCUSSION

Here we described weighting by IP as a method of approximating individual-level income using community-level income data from the census. In weighting by IP, the probability of community residents having a certain, investigator-defined categorical income value is calculated. Then, these probabilities were used as observation weights for individual patients, leading to synthetic datasets that were stratified into different investigator-defined income categories. The results from our assessment suggest that IP based approaches outperform the conventional approach of proxying individual-level income using median community-level income.

IP based methods likely outperform conventional methods because the accuracy of conventional methods is highly dependent on the underlying distribution of community-level income data. IP based methods use community-level count income data (rather than central tendency measures, such as median income), and thus account for the actual distribution of community-level income. Indeed, our assessment shows that IP based methods are especially superior to the conventional approach in the low and high end of the income spectrum. This is of relevance in studies of health disparities because they often focus on patients on the income extremes. Furthermore, weighting by IP can likely be applied to a broad range of patient datasets. Our assessment demonstrated that IP based approaches uniformly outperformed the conventional approach across different sex and race and ethnicity strata, suggesting that IP based approaches should perform well in other datasets with significantly different profiles of sex and race and ethnicity compared with the OCISS.

The average percentage difference for IP based approaches between the class-relevant income and policy-relevant income groupings were generally comparable. This suggests that the weighting by IP method is flexible to different investigator-defined income categorizations, within reason. The choice of investigator-defined income categorizations is limited somewhat by the prespecified income categorizations of the community-level income data that the US Census provides. For example, investigators interested in identifying patients beyond the top-coded income category of the ACS (500%+ FPL) will need additional statistical processing of the census data (such as mean-constrained integration over brackets[24]) before calculating IP.

The primary weakness of weighting by IP is that it is not suitable in analyses where patients need to be assigned a specific categorical income (as observations are weighted by the IP), or in the relatively rare instances where analysis of income as a continuous variable is desired. It should also be noted that weighting by IP is incrementally more difficult to implement compared with traditional approaches. However, in studies where individual-level income is a key variable of interest, the balance between methodological complexity and accurate income approximation will likely favor using this approach. Finally, income (and socioeconomic position, construed more broadly) can have meaningful impacts on health outcomes at multiple levels.[7,25,26] IP are used only to approximate individual-level income, and thus do not obviate the need for neighborhood- or area-level measures of socioeconomic position.

Our assessment could have been improved if we had a validation dataset that was relatively large, complete (i.e., contain an unbiased set of patients for a particular disease state), and contain accurate income information. The OCISS satisfies the first two criteria, but does not contain income information, necessitating a separate reference dataset for income (the ACS PUMS dataset). Indeed, biomedical datasets rarely contain accurate individual-level income information, which motivated both this study and the tradition of indirectly estimating patient income using US Census data. Thus, we could not identify datasets that we could reasonably acquire that would provide a more superior assessment of weighting by IP than the ones utilized in this study.

Community-level income data from the US Census has long been used to impute or approximate missing individual-level variables in patient data ("indirect estimation"). While the prevailing approach has been to simply use central tendency measures from the US Census to approximate these missing measures, more recent approaches have challenged this paradigm. A notable example is the Bayesian imputation of race/ethnicity using surname information and census block group of residence.[27,28] Weighting by IP complements these efforts to create more rich and complete patient-level datasets by leveraging census data. Weighting by IP is distinguished from other indirect estimation efforts with census data by its focus on income and its high accuracy payoff relative to its methodological simplicity.

We emphasize that a key strength of weighting by IP is that it is generalizable to different statistical techniques. Note that IP are assigned to patients as observation weights, so statistical tools and models (e.g., weighted regression) designed to analyze weighted (often survey) data can be directly utilized within income strata. Thus, once IP have been calculated and assigned to the patients, the analysis of the data is no more difficult than analyzing weighted data. Figure 5 illustrates how weighting by IP (in this case, age-specific IP as in Approach 3) can be operationalized using common statistical software to conduct key analyses, including regression. We include R code to calculate IP for all census tracts in any state in S6.

Future investigations of weighting by IP could be focused on mitigating its weaknesses and expanding its applications. Our assessment suggests that most variance in indirect estimation of individual-level income using census data is captured in the community of residence (census tract) because the age-specific weighting by IP (Approach 3) is at most associated with modest, incremental improvements in the average percent difference compared with generalized weighting by IP (Approach 2). Thus, including additional variables beyond age to calculate IP will likely lead to marginal improvements at best, while increasing the complexity of the approach. As a practical matter, the US Census does not publicly provide highly dimensional cross-tabulated income data for census tracts, so implementing such an approach would likely not be possible with the currently available data. Possible expansions of IP based approaches include applying it to variables beyond income, adapting IP to assign patients a specific categorical income, utilizing IP to impute other missing variables (as in multiple imputation), or incorporating the IP into models other than using them as observation weights.

In conclusion, weighting by IP represents a substantial methodological improvement in approximating individual-level income using community-level income data from the US Census compared with traditional methods. It is an approach that is robust across a variety of demographic populations, and thus can likely be applied to a broad range of patient data sets. Accurate approaches to individual-level income approximation such as weighting by IP can advance disparities studies and better inform policy interventions.

ACKNOWLEDGMENT

None.

ORCID

Uriel Kim https://orcid.org/0000-0001-9910-0682

REFERENCES

1. Cherkin DC, Grothaus L, Wagner EH. Is magnitude of co-payment effect related to income? Using census data for health services research. *Soc Sci Med*. 1992;34(1):33-41. doi:10.1016/0277-9536(92)90064-W
2. Chen FM, Breiman RF, Farley M, Plikaytis B, Deaver K, Cetron MS. Geocoding and linking data from population-based surveillance and the US Census to evaluate the impact of median household income on the epidemiology of invasive Streptococcus pneumoniae infections. *Am J Epidemiol*. 1998;148(12):1212-1218. doi:10.1093/oxfordjournals.aje.a009611
3. Bach PB, Guadagnoli E, Schrag D, Schussler N, Warren JL. Patient demographic and socioeconomic characteristics in the SEER-medicare database: applications and limitations. *Med Care*. 2002;40(8):IV19-IV25.
4. Chi GC, Hajat A, Bird CE, et al. Individual and neighborhood socioeconomic status and the association between air pollution and cardiovascular disease. *Environ Health Perspect*. 2016;124(12):1840-1847. doi:10.1289/EHP199
5. Pardo-Crespo MR, Narla NP, Williams AR, et al. Comparison of individual-level versus area-level socioeconomic measures in assessing health outcomes of children in Olmsted County, Minnesota. *J Epidemiol Community Health*. 2013;67(4):305 LP-310. doi:10.1136/jech-2012-201742
6. Narla NP, Pardo-Crespo MR, Beebe TJ, et al. Concordance between individual vs. area-level socioeconomic measures in an urban setting. *J Health Care Poor Underserved*. 2015;26(4):1157-1172. doi:10.1353/hpu.2015.0122
7. Buajitti E, Chiodo S, Rosella LC. Agreement between area- and individual-level income measures in a population-based cohort: implications for population health research. *SSM - Popul Heal*. 2020;10:100553. doi:10.1016/j.ssmph.2020.100553
8. Xie S, Hubbard RA, Himes BE. Neighborhood-level measures of socioeconomic status are more correlated with individual-level measures in urban areas compared with less urban areas. *Ann Epidemiol*. 2020;43:37-43.e4. doi:10.1016/j.annepidem.2020.01.012
9. Erdle SC, Birken CS, Parkin PC, Urquia ML, Maguire JL. Poor agreement between family-level and neighborhood-level income measures among urban families with children. *J Clin Epidemiol*. 2014;67(7):838-840. doi:10.1016/j.jclinepi.2014.02.006
10. Geronimus AT, Bound J, Neidert LJ. *On the validity of using census geocode characteristics to proxy individual socioeconomic characteristics*; 1995. National Bureau of Economic Research Technical Working Paper Series No. 189. doi:10.3386/t0189
11. Soobader M, LeClere FB, Hadden W, Maury B. Using aggregate geographic data to proxy individual socioeconomic status: does size matter? *Am J Public Health*. 2001;91(4):632-636. doi:10.2105/ajph.91.4.632
12. Marra CA, Lynd LD, Harvard SS, Grubisic M. Agreement between aggregate and individual-level measures of income and education: a comparison across three patient groups. *BMC Health Serv Res*. 2011;11(1):69. doi:10.1186/1472-6963-11-69
13. US Census Bureau. American Community Survey. Accessed November 25, 2017. https://www.census.gov/acs/www/data/data-tables-and-tools/
14. Subject definitions. Accessed January 9, 2020. https://www.census.gov/programs-surveys/cps/technical-documentation/subject-definitions.html
15. US Census Bureau. American Community Survey information guide. Accessed August 8, 2019. https://www.census.gov/content/dam/Census/programs-surveys/acs/about/ACS_Information_Guide.pdf
16. Thomas AJ, Eberly LE, Davey Smith G, Neaton JD, Group for the MRFIT (MRFIT) R. ZIP-code-based versus tract-based income measures as long-term risk-adjusted mortality predictors. *Am J Epidemiol*. 2006;164(6):586-590. doi:10.1093/aje/kwj234
17. Krieger N, Chen JT, Waterman PD, Soobader M-J, Subramanian SV, Carson R. Choosing area based socioeconomic

measures to monitor social inequalities in low birth weight and childhood lead poisoning: the Public Health Disparities Geocoding Project (US). *J Epidemiol Community Health*. 2003;57(3):186 LP-199. doi:10.1136/jech.57.3.186

18. Krieger N, Waterman PD, Chen JT, Soobader M-J, Subramanian SV. Monitoring socioeconomic inequalities in sexually transmitted infections, tuberculosis, and violence: geocoding and choice of area-based socioeconomic measures--the public health disparities geocoding project (US). *Public Health Rep*. 2003;118(3):240-260. doi:10.1093/phr/118.3.240
19. Krieger N, Chen JT, Waterman PD, Soobader M-J, Subramanian SV, Carson R. Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? The public health disparities geocoding project. *Am J Epidemiol*. 2002;156(5):471-482. doi:10.1093/aje/kwf068
20. Rose SJ. The growing size and incomes of the upper middle class; 2016.
21. Rosenbaum S. The patient protection and affordable care act: implications for public health policy and practice. *Public Health Rep*. 2011; 126(1):130-135.
22. Ohio Cancer Incidence Surveillance System (OCISS). Accessed April 25, 2020. https://odh.ohio.gov/wps/portal/gov/odh/know-our-programs/ohio-cancer-incidence-surveillance-system/welcome-to
23. PUMS documentation. Accessed May 9, 2020. https://www.census.gov/programs-surveys/acs/technical-documentation/pums.html
24. Jargowsky PA, Wheeler CA. Estimating income statistics from grouped data: mean-constrained integration over brackets. *Sociol Methodol*. 2018;48(1):337-374. doi:10.1177/0081175018782579
25. Krieger N, Williams DR, Moss NE. Measuring social class in US public health research: concepts, methodologies, and guidelines. *Annu Rev Public Health*. 1997;18:341-378. doi:10.1146/annurev.publhealth.18.1.341
26. Pichora E, Polsky JY, Catley C, Perumal N, Jin J, Allin S. Comparing individual and area-based income measures: impact on analysis of inequality in smoking, obesity, and diabetes rates in Canadians 2003–2013. *Can J Public Health*. 2018;109(3):410-418. doi:10.17269/s41997-018-0062-5
27. Haas A, Elliott MN, Dembosky JW, et al. Imputation of race/ethnicity to enable measurement of HEDIS performance by race/ethnicity. *Health Serv Res*. 2019;54(1):13-23. doi:10.1111/1475-6773.13099
28. Elliott MN, Morrison PA, Fremont A, McCaffrey DF, Pantoja P, Lurie N. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Serv Outcomes Res Methodol*. 2009;9(2):69-83.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Kim U, Koroukian SM, Stange KC, Spilsbury JC, Dong W, Rose J. Describing and assessing a new method of approximating categorical individual-level income using community-level income from the census (weighting by income probabilities). *Health Serv Res*. 2022;57(6):1348-1360. doi:10.1111/1475-6773.14026