PNAS nexus

PNAS Nexus, 2024, **3**, 1–11
https://doi.org/10.1093/pnasnexus/pgae109
Advance access publication 11 March 2024
Research Report

Interviewer biases in medical survey data: The example of blood pressure measurements

Pascal Geldsetzer[a,b,c,†,*], Andrew Young Chang[b,d,e,†], Erik Meijer[f], Nikkil Sudharsanan[g,h], Vivek Charu[i,j], Peter Kramlinger[k,‡] and Richard Haarburger[l,‡]

[a]Division of Primary Care and Population Health, Department of Medicine, Stanford University, 3180 Porter Drive, Palo Alto, CA 94304, USA
[b]Department of Epidemiology and Population Health, Stanford University, 300 Pasteur Dr., Palo Alto, CA 94305, USA
[c]Chan Zuckerberg Biohub – San Francisco, 499 Illinois Street, San Francisco, CA 94158, USA
[d]Division of Cardiology, Department of Medicine, University of California San Francisco, 1001 Potrero Ave, San Francisco, CA 94110, USA
[e]Center for Innovation in Global Health, Stanford University, 3180 Porter Drive, Palo Alto, CA 94304, USA
[f]Center for Economic and Social Research, University of Southern California, 635 Downey Way, Los Angeles, CA 90089-3332, USA
[g]Professorship of Behavioral Science for Disease Prevention and Health Care, Technical University of Munich, Georg-Brauchle-Ring 60, 80992 Munich, Germany
[h]Heidelberg Institute of Global Health, Heidelberg University, Im Neuenheimer Feld 130.3, 69120 Heidelberg, Germany
[i]Quantitative Sciences Unit, Department of Medicine, Stanford University, 1070 Arastradero Road, Palo Alto, CA 94394, USA
[j]Department of Pathology, Stanford University, 300 Pasteur Dr., Palo Alto, CA 94305, USA
[k]Department of Statistics, University of California Davis, One Shields Avenue, Davis, CA 95616, USA
[l]Research Training Group: Globalization and Development, Faculty of Business and Economics, Georg-August-University Göttingen, Platz d. Göttinger Sieben 3, 37073 Göttingen, Germany
*To whom correspondence should be addressed: Email:pgeldsetzer@stanford.edu
†P.G. and A.Y.C. are joint first authors.
‡P.K. and R.H. are joint senior authors.

Edited By: Adelia Bovell-Benjamin

Abstract

Health agencies rely upon survey-based physical measures to estimate the prevalence of key global health indicators such as hypertension. Such measures are usually collected by nonhealthcare worker personnel and are potentially subject to measurement error due to variations in interviewer technique and setting, termed “interviewer effects.” In the context of physical measurements, particularly in low- and middle-income countries, interviewer-induced biases have not yet been examined. Using blood pressure as a case study, we aimed to determine the relative contribution of interviewer effects on the total variance of blood pressure measurements in three large nationally representative health surveys from the Global South. We utilized 169,681 observations between 2008 and 2019 from three health surveys (Indonesia Family Life Survey, National Income Dynamics Study of South Africa, and Longitudinal Aging Study in India). In a linear mixed model, we modeled systolic blood pressure as a continuous dependent variable and interviewer effects as random effects alongside individual factors as covariates. To quantify the interviewer effect-induced uncertainty in hypertension prevalence, we utilized a bootstrap approach comparing subsamples of observed blood pressure measurements to their adjusted counterparts. Our analysis revealed that the proportion of variation contributed by interviewers to blood pressure measurements was statistically significant but small: $\sim 0.24 - -2.2\%$ depending on the cohort. Thus, hypertension prevalence estimates were not substantially impacted at national scales. However, individual extreme interviewers could account for measurement divergences as high as 12%. Thus, highly biased interviewers could have important impacts on hypertension estimates at the subdistrict level.

Keywords: blood pressure, hypertension, measurement error, interviewer effects, health survey

Significance Statement

Physical measurements such as blood pressure are important indicators of countries’ health system performance. These measures are usually obtained in household surveys by study-specific interviewers, who are not clinical healthcare workers. Thus, there is a concern that they may contribute substantial measurement error. We used three large nationally representative health surveys from India, Indonesia, and South Africa to model the magnitude of the interviewer effect on blood pressure measurements, and then projected their impact on estimations of country-level hypertension prevalence. At smaller geographic units, “extreme” interviewers could substantially bias hypertension estimates. Overall, however, the magnitude of the interviewer effects was small and, thus, unlikely to substantially bias hypertension prevalence estimates at the national level.

OXFORD UNIVERSITY PRESS

Competing Interest: The authors declare no competing interest.
Received: September 6, 2023. **Accepted:** February 27, 2024

Introduction

Global health indicators such as blood pressure, weight, and height are critical for monitoring both national and international health system performance. Such markers are largely collected through household surveys, which are often seen as the gold standard methodology due to their population-representative nature (1–5).

Interviewer collected physical measures such as heart rate or body mass index (BMI) may appear to hold greater "objectivity" than self-reported indicators or subjective social indicators. Self-reported data are frequently prone to not only random measurement error but also systematic measurement error due to interviewee attitudes such as recall bias and social desirability bias (6). Nevertheless, physical measures are still subject to a substantial degree of random measurement error due to administrator technique and environmental context during acquisition (7–10). This phenomenon may possibly be magnified in the case where medical measurements are taken by nonclinician interviewers who may not routinely perform such measures outside of the research setting.

Nevertheless, many household surveys make the implicit assumption that, after their training, interviewers all perform to the same standard as one another (11). Subsequent analyses therefore assume that the interviewers are not a source of measurement error and that uncertainty estimates are purely based on the sampling strategy.

At the national level, these "interviewer effects" may average out from the large number of interviewers contributing both positive and negative measurement error. At finer geographic divisions, however, the relatively smaller number of interviewers may lead to greater variation or even potential bias in the measurement of a target indicator. This is particularly important because estimates from small areas are increasingly being used in public health decision-making and for mapping disease prevalence at subnational levels, sometimes in resolutions as fine as 5 × 5 km (12–15).

Prior analyses have queried the intraobserver and interobserver reproducibility of specific physical measures, but such investigations have tended to focus on the reliability of these markers for clinical situations (7, 9, 16). Furthermore, most such studies have utilized healthcare workers like nurses and medical trainees as the measurement-takers given their applicability to the medical setting, and have examined high-income country populations (17, 18). Large-scale empirical analyses of nonclinician interviewers' reliability for physical measures for public health purposes, especially in low- and middle-income countries (LMICs), remain sparse. The amount of random measurement error found in such global health indicators varies, with some exhibiting relatively low degrees (e.g. controlled laboratory-based tests), while others with increased operator inputs suffer from potentially greater degrees of interviewer-introduced measurement error. For example, anthropometry for newborns, adult waist circumference, and blood pressure measurements require interviewers to make subjective decisions about how and where to place the instruments and in what settings to do so (8, 19).

Here, we assess the magnitude of interviewer-induced measurement error in large-scale global health surveys using the case study of high blood pressure. High blood pressure is an ideal case study because it is already a disease of considerable importance in LMICs (20, 21). Blood pressure is readily and frequently measured noninvasively, and nonclinician study personnel can be taught how to collect blood pressure assessments (11). This is particularly important as community health workers and other nonnurse/nonphysician healthcare workers are increasingly being called upon to care for noncommunicable diseases in primary care in poor countries, and they are also frequently called upon for survey data collection as well (22–24).

As such, in the present analysis (assuming that interviewers are randomly allocated to households within primary sampling units), we examine the magnitude of uncertainty attributable to interviewer effects on blood pressure measurements and hypertension (systolic blood pressure ≥ 140 mmHg) in three large longitudinal health surveys from the Global South.

Results

Sample characteristics

Table 1 shows descriptive statistics for the datasets used in this study after preprocessing. Data from 169,681 total encounters were utilized, with 26,554 from the Indonesia Family Life Survey (IFLS), 55,469 from the Longitudinal Aging Study in India (LASI), and 87,658 from the National Income Dynamics Study (NIDS) of South Africa, respectively.

Variation shares in hypertension prevalence

To interpret the effect sizes of the interviewer-level effects, we compare their shares in total variation to the shares of other level effects and the residual from the same estimations. Table 2 presents the variance components of the fitted linear mixed models (LMMs) for the IFLS, NIDS, and LASI datasets.

The bootstrap likelihood ratio test (LRT) tests give $P < 0.0001$ for all three datasets. This strongly suggests the presence of interviewer effects in all three datasets, although they are numerically small.

Uncertainty in sample hypertension prevalence

Figure 1 displays the nonparametric bootstrap densities for hypertension prevalence, based on the original data (blue, dashed), and the corrected measurements (red, dotted). Panels (a), (b) and (c) indicate results for the IFLS, NIDS and LASI data, respectively. The vertical line represents the observed prevalence by data source.

Effect study

In order to illustrate the interviewer-introduced uncertainty in hypertension prevalences, we perform an effect study. Using the set of observed systolic blood pressure measurements and the measurements corrected for the estimated interviewer effects, we can compare observed interviewer-specific prevalences of hypertension to the respective corrected interviewer-specific prevalences. Alternatively, we can also illustrate differences in prevalences for geographic areas, such as subdistricts.

Interviewer-specific prevalences: observed and corrected

Figure 2 illustrates a subsample of the interviewer-specific observed and adjusted prevalences of hypertension for the IFLS dataset. The subsample is created based on the distribution of differences in observed and adjusted prevalences. For example, to focus on the most extreme cases, we depict the prevalences for all interviewers for whom the difference between observed and adjusted prevalence lies above the 70th percentile of these differences. In other words, we show the 30% of cases subject to

Table 1. Descriptives of IFLS, LASI, and NIDS data.

	Individual data sources			Overall
	IFLS (*n* = 26,554)	**LASI (*n* = 55,469)**	**NIDS (*n* = 87,658)**	**(*n* = 169,681)**
Average systolic blood pressure measurement				
Mean (SD)	129 (19.9)	129 (19.1)	121 (21.3)	125 (20.8)
Median (Min, Max)	126 (68.0, 241)	127 (60.0, 234)	117 (44.0, 240)	122 (44.0, 241)
Sex				
Male	16,569 (62.4%)	25,694 (46.3%)	36,446 (41.6%)	78,709 (46.4%)
Female	9,985 (37.6%)	29,775 (53.7%)	51,212 (58.4%)	90,972 (53.6%)
Age				
Mean (SD)	41.7 (13.1)	59.5 (10.4)	36.1 (17.1)	44.6 (18.0)
Median (Min, Max)	42.0 (15.0, 101)	58.0 (45.0, 108)	32.0 (14.0, 108)	46.0 (14.0, 108)
Education 1[a]				
Less than primary	0 (0%)	6317 (11.4%)	8224 (9.4%)	14541 (8.6%)
Primary or secondary	22,629 (85.2%)	42,613 (76.8%)	67,901 (77.5%)	133,143 (78.5%)
Tertiary	3,925 (14.8%)	6,539 (11.8%)	11,533 (13.2%)	21,997 (13.0%)
BMI				
Mean (SD)	23.3 (4.28)	22.8 (4.73)	26.1 (6.72)	24.6 (6.00)
Median (Min, Max)	22.7 (10.7, 57.1)	22.3 (10.5, 55.6)	24.6 (10.4, 60.0)	23.4 (10.4, 60.0)
Ever diagnosed with hypertension				
Not diagnosed	23,406 (88.1%)	39,600 (71.4%)	75,814 (86.5%)	138,820 (81.8%)
Diagnosed	3,148 (11.9%)	15,869 (28.6%)	11,844 (13.5%)	30,861 (18.2%)
Log income				
Mean (SD)	12.5 (4.04)	11.2 (1.66)	7.95 (0.942)	9.71 (2.72)
Median (Min, Max)	13.7 (0, 19.5)	11.4 (0, 18.8)	7.82 (4.25, 13.0)	9.13 (0, 19.5)
Smoking				
Nonsmoker	14,850 (55.9%)	47,686 (86.0%)	75,814 (86.5%)	138,350 (81.5%)
Smoker	11,704 (44.1%)	7,783 (14.0%)	11,844 (13.5%)	31,331 (18.5%)
Urban/Rural				
Urban	15,300 (57.6%)	19,263 (34.7%)	43,477 (49.6%)	78,040 (46.0%)
Rural	11,254 (42.4%)	36,206 (65.3%)	44,181 (50.4%)	91,641 (54.0%)

[a] Education levels were harmonized for the sake of simplifying descriptive statistics.

Table 2. Variance components of the fitted LMMs by dataset for IFLS, NIDS, and LASI.

Data source	Effect	Variance	Percentage (%)
IFLS	Interviewer	1.47	0.53
	Household	37.6	13.6
	Province	2.96	1.07
	Municipality	2.29	0.83
	Subdistrict	2.78	1.01
	Residuals	229	82.9
	Total	276.1	≈100
NIDS	Interviewer	7.19	2.2
	Household	39.6	12.1
	Cluster	3.74	1.15
	Province	3.06	0.94
	Residuals	273	83.7
	Total	328.17	≈100
LASI	Interviewer	0.785	0.24
	Household	21.3	6.55
	State	7.99	2.46
	District	7.75	2.39
	Village/ward	4.96	1.53
	Residuals	282	86.8
	Total	324.785	≈100

the most drastic adjustment effects. The top 50, 30, 10, and 1% cases are presented.

The analogous findings for NIDS and LASI are provided in Figs. S1 and S2.

Subdistrict specific prevalences: observed and corrected

Analogously to the interviewer-specific prevalences, we can also depict changes in prevalences for geographical units, as illustrated in Figure 3. The higher the granularity in geographical division, the larger the influence of single interviewers. We thus depict adjustment-induced changes in prevalences on the most granular level available for each respective data source. In case of LASI and IFLS, the most granular geographical level are subdistricts. In case of NIDS, less granular level data are available, so that we are limited to the cluster level.

Discussion

In the present analysis, we found that interviewer effects in blood pressure measurements were statistically significant, although numerically trivial, in three large longitudinal health surveys from Indonesia, India, and South Africa. This was achieved by calculating the proportion of total variance attributable to various sources, one of which was the interviewer. Nevertheless, both the absolute and relative contribution of the interviewer to blood pressure measurement variation was not particularly high, especially when compared to geographic/community-level effects. In the IFLS cohort, interviewer-level effects comprised 0.5% of the variance, while in NIDS, 2.2%, and in LASI, 0.2%. In fact, household effects (13.6, 12.1, 6.6%, respectively) dominated the variance of all three datasets, with residential effects (i.e. province, state, subdistrict, municipality) larger than interviewer effects except in NIDS.

On the population level, however, the combined interviewer effect could potentially impact the uncertainty in hypertension prevalence. As such, we generated nonparametric bootstraps of prevalence estimates unadjusted and adjusted for the interviewer effect, which show very small but consistently lower point estimates of hypertension prevalence in all three datasets on the order of a fraction of a percent. This may have minor implications

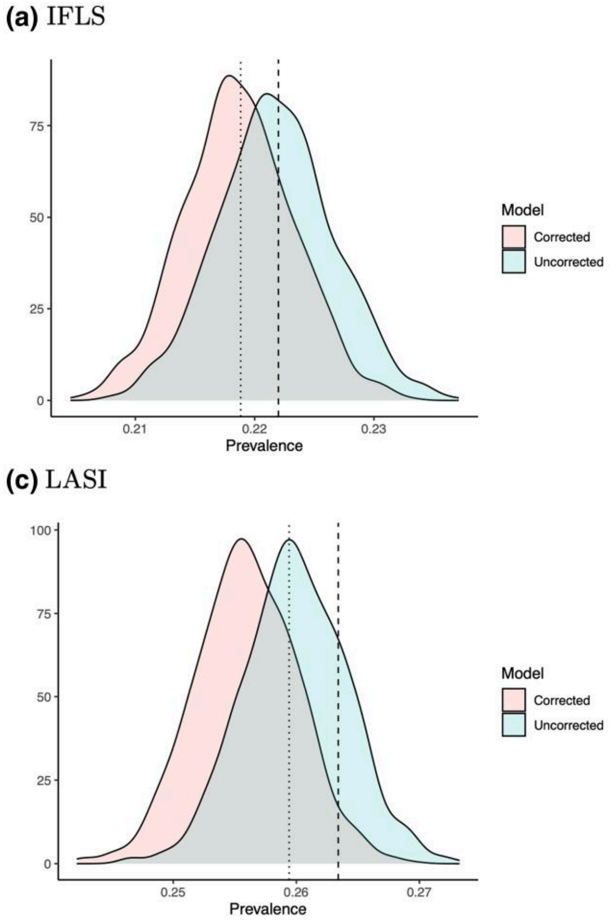

Fig. 1. Bootstrap densities for hypertension prevalence, based on the original data (uncorrected, dashed), and the corrected measurements (corrected, dotted). The vertical line represents the observed prevalence.

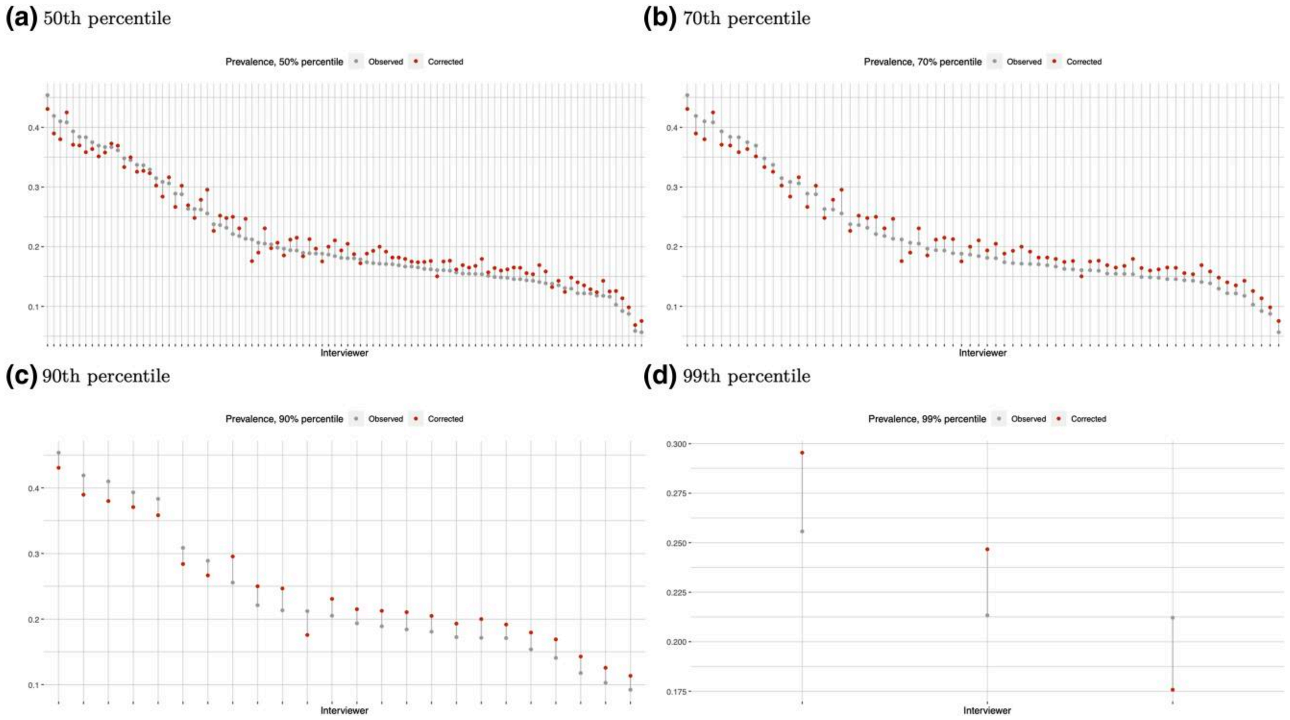

Fig. 2. IFLS: Observed and adjusted interviewer-specific prevalences of hypertension, 50, 30, 10, 1% of cases subject to largest adjustment effects.

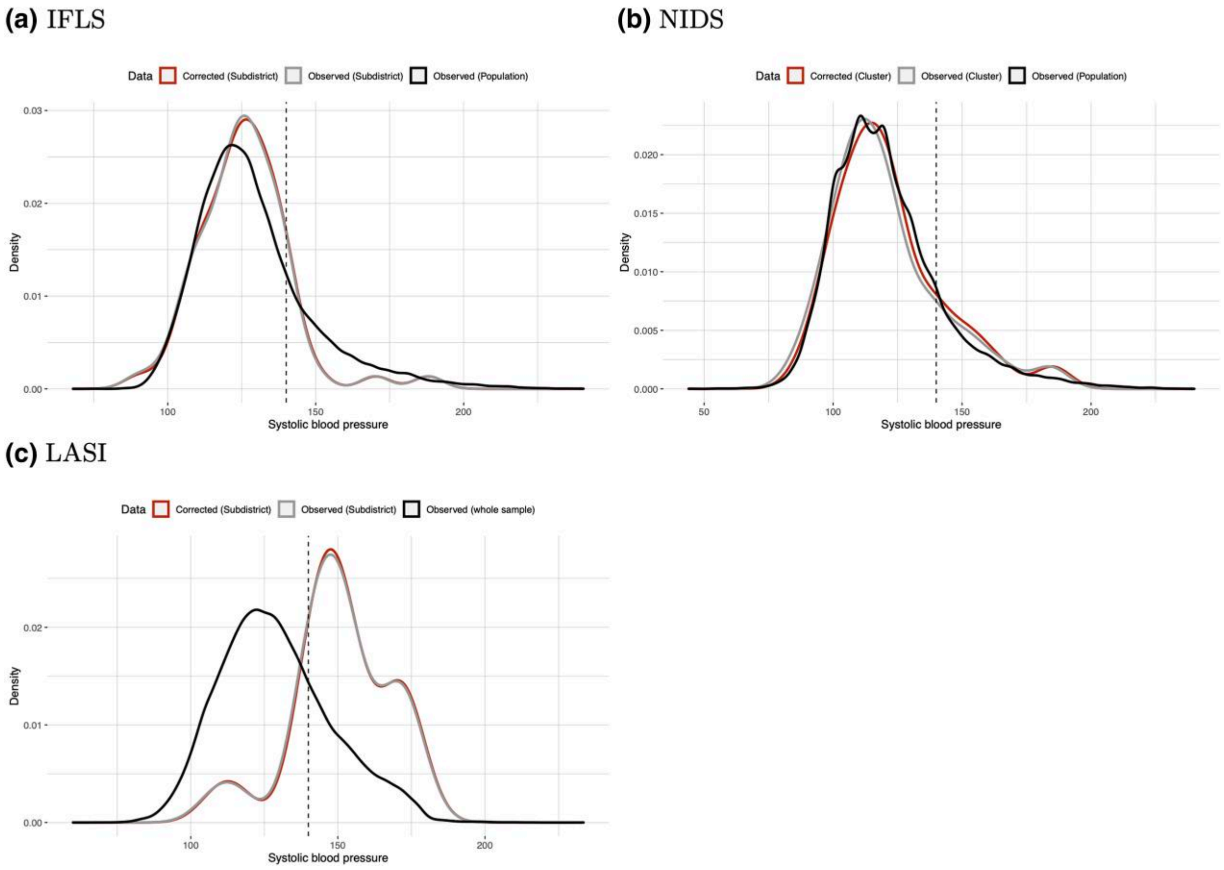

Fig. 3. Systolic blood pressure densities, observed and adjusted for estimated interviewer effects, for selected subdistricts subject to large adjustment-induced changes by data source. Population densities are added as comparison.

for public policy targeting hypertension suggesting slight current overestimation of true hypertension prevalence in these settings.

Nevertheless, the magnitude of the discrepancies is not exceedingly high at these larger scales—where we found the interviewer effect to carry the greatest possibility of influencing hypertension estimation was at smaller geographic divisions. Taking the most "extreme" individual interviewers responsible for the greatest adjustment effects in each dataset and comparing their observed and adjusted hypertension prevalences revealed divergences as high as 12% in NIDS. We therefore assessed their impacts by comparing the observed and interviewer-effect adjusted subdistrict specific hypertension prevalences subject to the greatest adjustment effects. These revealed up to 5–7 percentage points (p.p.) prevalence differences between observed and corrected values at subdistrict levels for the top 1% of cases subject to adjustment effects. The substantial degree of bias that these may introduce at the local level compared to the population (or whole sample) level are well visualized in the resultant cluster-specific blood pressure density plots. For example, in LASI, the modal systolic blood pressure signed difference between subdistrict and total population was nearly 25 mmHg.

Our study represents the largest empirical estimation of interviewer effects on blood pressure. We also believe it to be the first of its kind involving low- and middle-income country populations. Thus, it contributes to the growing body of work examining and quantifying interviewer-based sources of measurement error for survey-based global public health indicators. The results are reassuring that the present strategy of utilizing nonclinician study interviewers is likely not generating a critical degree of variation in blood pressure measurement for populations, and we propose one possible method by which analysts may adjust for these small interviewer effects.

Because our investigation is, to our knowledge, the first to assess interviewer effects for blood pressure in household surveys from low- and middle-income countries, we are only able to compare our findings to those from much smaller samples in two surveys from the UK. Cernat and Sakshaug found that there are interviewer effects on measurement error from both nurses and trained nonclinician interviewers in these two UK-based surveys (8, 19). For nonclinician interviewers, they noted that in measures such as height, weight, blood pressure, and pulse, interviewer effects similarly comprised only a small fraction of the variance—for blood pressure, <1%. Much like our findings, these studies also identified that area-level effects contributed a greater source of variation than the interviewer effect for many physical measures.

Nevertheless, our work further models the public health implications of the interviewer effect by estimating the impact of these forces on hypertension prevalence estimates at multiple geographic levels. In doing so, our analyses also identified that extremely biased interviewers could lead to markedly biased hypertension estimates, and that if there is disproportionate allocation of these "extreme" interviewers to a locale at the level of a subdistrict or smaller, that there may be substantially biased hypertension prevalence estimates in these geographic units.

Strengths of our study include the size of the analytic cohort (total 169,681 observations), as well as the use of three different nationally representative datasets from Africa, South Asia, and Southeast Asia. There is substantial heterogeneity in the resultant populations, not just by the distribution of gender, age, and urban/rural breakdown, but also the underlying true prevalence of hypertension. Blood pressure measurements from years 2008 through 2019 were included, further capturing time-related variation. The most important limitation of our analysis is that, analytically, our modeling strategy relies upon the assumption that all interviewers were quasirandomly allocated to participants within the primary sampling units. Moreover, we restrict our analysis to measurements of systolic blood pressure measurements only rather than adding measurements of diastolic blood pressure to infer on changes in the prevalence of hypertension for reasons of parsimony. A definition of hypertension that depends on two measurements would further complicate the already complex analysis. However, diastolic measurements depict a way to further investigate interviewer effects in future studies. Diastolic hypertension (both independently and in conjunction with systolic hypertension) may be a risk factor for adverse cardiovascular outcomes (25, 26). In addition, the LASI cohort was substantially older than the IFLS and NIDS cohorts. Furthermore, the full dataset does not constitute a random sample of all household surveys in low- and middle-income countries. Lastly, all three survey cohorts involved interviewers who were highly trained using established, high-quality protocols and closely monitored by study administration. As biased interviewers have higher impact on measurement error in small geographic units, our results may underestimate the magnitude of interviewer effects for less-rigorously trained/observed interviewers in LMIC settings.

We conclude by noting that interviewer effects appear to be present, but small at best in household surveys of blood pressure in lower middle- and middle-income countries. Future work could involve targeted empirical analyses of the influence of "extreme interviewers" on quantifying the local burden of disease, as well as replication of our methods in other cohorts from different continents and from low-income countries. Additionally, we recognize that blood pressure is but one physical measure from a large pool of monitored global health indicators. As prior research in other settings has suggested that interviewer effects vary with the type of measurement performed, independent analyses of these other markers such as weight and BMI should be pursued to provide a more comprehensive understanding of the phenomenon.

Materials and methods

Data sources

We demonstrate the implications of interviewer measurement biases using three common longitudinal health surveys. Besides waves 4 and 5, as well the east extension of the IFLS, we use all five waves of the NIDS, and the first wave of the LASI in our analysis. All three datasets were collected with the purpose to document socioeconomic and health outcomes over time. Moreover, they were designed to provide sufficient sample size and adequate sampling schemes to be nationally representative. Thus, they are generally considered suitable to estimate prevalences of diseases for whose documentation adequate examinations were conducted as part of the survey, such as hypertension. Various weights are available to make the datasets nationally representative. Since our primary objective is not to estimate nationally representative prevalences, but to identify interviewer effects as rigorously as possible, we refrain from using these weights in order to keep the model specifications simple. This applies to all three data sources. The preprocessing of the datasets consists in merging data from the various waves and reducing the data to the variables of interest. Moreover, we make minor changes to harmonize the datasets, such as simplifying individual variables to fewer values, as in the case of education. The further preprocessing consists in restricting the data to complete observations.

Sampling strategy

National Income Dynamics Study

The NIDS data were collected in five waves between February 2008 and December 2017 (27–31). Since the NIDS data are of longitudinal nature, the households interviewed in the first wave were recontacted for the following waves. However, new individuals entered the sample between waves 2 and 5 by joining the households comprising the original NIDS sample from wave 1. Additionally, in wave 5, the sample was topped up to account for undersampled socioeconomic groups and attrition. A two-stage stratified cluster sample design was applied in the data generation process of the first wave.

The underlying 2003 master data used to generate NIDS were provided by Statistics South Africa, comprised 3,000 primary sampling units (PSUs), and were stratified with respect to 53 district councils. The NIDS data depict a subset of 400 PSUs which were randomly drawn within the strata, while conserving proportionality. Within each PSU, eight nonoverlapping samples of dwelling units had been drawn for the creation of the master data, which are referred to as clusters in the NIDS documentation. The majority of clusters were assigned various household surveys before the creation of NIDS. Two clusters in each PSU however had never been involved in surveys and became the base for NIDS. For further details, see Leibbrandt et al. (32). NIDS wave 1 comprises completed surveys of 7,296 households from the aforementioned subsampled 400 PSUs. In order to establish national representativeness, different sets of weights were constructed as described in Wittenberg (33). Since our analysis does not aim for national representativeness, but focuses on interviewer effects only, we do not apply the weights provided within the NIDS data and thus do not further discuss the computation of the weights here.

After cleaning and preprocessing the NIDS data as outlined above, 87,658 observations remain, which we use throughout our analysis.

Indonesia Family Life Survey

The IFLS data used in the scope of this analysis comprise waves 4, 5, and the east extension (34–36). As is the case with NIDS, due to the IFLS data being a longitudinal survey, the households interviewed during the first wave were recontacted for all following waves. Thus, the sampling scheme of the first wave determined the sample composition of all following waves. IFLS1 stratified on provinces and urban vs. rural locations within which simple random sampling was applied. Out of a total of 27 Indonesian provinces, only 13 are included in the sample, which however represented 83% of the population in 1993 (35). Within the selected 13 provinces, 321 enumeration areas (EAs) were randomly chosen, with proportions being selected to cause oversampling of urban EAs and smaller provinces to ensure the comparability of rural and urban EAs. While within each urban EA 20 households were selected, 30 were selected within each rural EA, resulting in a total of 7,224 completed household interviews in IFLS1. For a more detailed description of the sampling scheme, please refer to Strauss et al. (35). IFLS East includes most of the provinces not covered by the main IFLS. Within each selected province, 14 villages or urban villages were randomly drawn. These were then subdivided into units/areas with about 100–150 households, from which one was drawn at random. Within each of these, again 20 households were drawn if urban and 30 if rural. See Sikoki et al. (34) for more details. After initial data cleaning and processing, 26,554 individual-level observations from IFLS 4, 5, and East remain, which we use in the scope of this analysis.

Longitudinal Aging Study in India

We use the first wave of LASI data which were collected between 2017 and 2019 (37). The sampling scheme applied throughout the LASI data collection followed the 2011 census and implemented a multistage, stratified cluster sample design. While in the case of urban areas three sampling stages were conducted, four stages were conducted in the case of rural areas. The first stage consisted in the selection of PSUs within states. In the second stage, villages were selected in the rural PSUs and wards within the urban PSUs. Stage three included the selection of households in rural areas and the selection of Census Enumeration Blocks (CEBs) in wards. The final and fourth stage applied in urban areas comprised the selection of households. The LASI data used in the scope of this analysis comprise 55,469 observations postpreprocessing and cleaning.

Interviewer training, characteristics, and monitoring

National Income Dynamics Study

Interviewer training was held at the same time as the pretest was conducted, and specifics on the training of blood pressure measurements are not documented. The NIDS documentation does not mention specially trained health professionals taking the health measurements as is common in similar surveys. Thus, health measurements have been taken by the interviewer conducting the rest of the household surveys.

With wave 5, a set of interviewer demographics and experience variables were added to the available data.

The use of paradata was implemented to oversee interviewers and thereby reduce interviewer effects. Precisely, paradata are used to monitor questionnaire duration, refusal rates, magnitude of anthropometric measurement differences between current waves and previous waves, flag extreme BMI measures, and run other similar checks. The checks were taken periodically from about 6 weeks into fieldwork or when there were enough data to estimate meaningful averages. When interviewers' performance measures were conspicuous, they were investigated, retrained, moved to different teams for closer supervision or removed. In some cases, the respective households were reinterviewed. The Southern Africa Labour and Development Research Unit (SALDRU) carried out a range of pattern searches and consistency checks on the data during fieldwork to identify interviewer effects and potential general cases of miscapture.

The NIDS sample used in our analysis comprises a total of 513 distinct interviewers taking blood pressure measurements.

Indonesia Family Life Survey

Supervisory training was held for all senior personnel. In the case of IFLS5, this training of trainers included reviewing all parts of the survey: household, community facility, health, computer-assisted personal interview system (CAPI) tracking, and the management information systems used in the scope of the data collection. Household interviewer training was conducted in two phases. Training sessions were divided into two parts, classroom training and field practice. Household interviewers received 19 days of classroom training and 4 days of field practice. The collection of health data was conducted by regular interviewers, i.e. no health professionals were involved in the data collection on site during the interviews. Training for health-related measurements was part of the regular interviewer training. In the case of IFLS4 and IFLS East, the CAPI system had not been implemented yet and blood pressure measurements were conducted by nurses, i.e. professional health workers, and nonprofessional interviewers, respectively.

The combined IFLS data contains a total of 409 distinct interviewers taking blood pressure measurements.

Longitudinal Aging Study in India

A series of manuals were designed to standardize different aspects of surveys conducted in the scope of the LASI data collection. These manuals were instrumental in the training of interviewers. One of the manuals specifically focuses on the physical measures section of LASI and thus includes instructions for the measurement of blood pressure. The training duration of interviewers and health investigators was 35 days, of which 5 took place in the field. Even though the interviewers were employed via subcontractors, they were trained by trainers, who themselves were trained by the International Institute for Population Sciences (IIPS). After training was completed, investigators were individually assessed to assure that their work met the requirements previously defined by the manuals.

The LASI sample used in our analysis comprises a total of 504 distinct interviewers taking blood pressure measurements.

Definition of hypertension, blood pressure measurement

Multiple systolic blood pressure measurements were taken in the scope of all surveys included in this study. In the case of the IFLS and LASI data, three measurements were taken per individual, in the case of the NIDS data only two. In order to mitigate the white coat effect and to average out idiosyncratic fluctuations in measurements, we average the second and third measurement, while disregarding the first in the case of IFLS and LASI. In the case of NIDS, we only consider the second measurement, disregarding the first. Following this procedure, we obtain a single systolic blood pressure value for each interviewee. We consider interviewees to be hypertensive if their resulting single systolic blood pressure measurement is equal to or greater than 140 mmHg.

Measurements were conducted using an Omron HEM 7121 BP monitor in the case of LASI and an Omron HEM 7203 in the case of IFLS. Information on the exact device used for blood pressure measurement throughout NIDS data collection is not part of the publicly available documentation.

Definition of covariates

We add covariates to the model, which we consider potential determinants of blood pressure. To keep the results comparable, we use mostly the same set of covariates across all datasets. Besides using interviewees' sex, age, BMI, and smoking status, we proxy interviewees' socioeconomic background with income and education. The variables we choose in the respective datasets to compose our income proxy refer to monthly salaries and wages or monthly profits from entrepreneurship for NIDS and IFLS, and the logarithm of total household income for LASI. While the resulting income variables are hardly comparable across datasets, we assume comparability within datasets. To align the information on the education of individuals, we recode education into three categories, namely less than primary schooling, primary and/or secondary schooling, and tertiary education, except in LASI, where we added a fourth category for no schooling. In order to proxy for the possible use of blood pressure lowering medication, we include a variable which depicts whether an interviewee has ever been diagnosed with hypertension before.

Statistical analysis

We model a linear relationship of systolic blood pressure and available covariates. Formally, systolic blood pressure for individual i in household j at location k measured by interviewer l is denoted as Y_{ijkl}, so that

$$Y_{ijkl} = \beta_0 + \sum_{d=1}^{p} x_{ijkld}\beta_d + u_j + v_k + w_l + \varepsilon_{ijkl}, \tag{1}$$

where $\varepsilon_{ijkl} \sim N(0, \sigma_\varepsilon^2)$ is a Gaussian error term. Furthermore, $x_{ijkl1}, \ldots, x_{ijklp}$ are the available covariates, $\beta_0 \in \mathbb{R}$ is a common intercept and u_j, v_k, w_l are the respective level effects of household, location, and interviewer, for $j = 1, \ldots, J$, $k = 1, \ldots, K$, $l = 1, \ldots, L$. These level effects, as well as the parameter vector $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^t$ are unknown and have to be estimated given a sample of independent measurements.

Since the main objective lies in investigating systolic blood pressure, this model includes a selection of socioeconomic control covariates. The separately modeled level effects include a household effect, the interviewer effect, and the maximum number of geographical-level effects supported by the respective dataset. In the following, we will motivate the use of these level effects individually. We suspect that the interviewer effect significantly influences systolic blood pressure measurements, and is at the core of our analysis, as described above. Of note, due to the inability to trace interviewers across waves of the datasets, we treat all observations individually and ignore the time dimension.

We motivate the use of geographical-level effects based on the assumption that geographical cultural clusters, geographical differences in the availability of food, geographical differences in health care access, and similar factors might affect systolic blood pressure spatially.

It is common practice to assign interviewers to households and not to interviewees directly. An interviewer then interviews all eligible individuals belonging to an assigned household. Variation in systolic blood pressure on the household level therefore potentially confounds the estimation of the interviewer effect. Thus, we include household effects to absorb household level variation.

We are interested in investigating $Y_{ijkl} - w_l$, which is the systolic blood pressure adjusted for the true measurement error induced by interviewers, which we are estimating with our approach. Accordingly, we consider $Y_{ijkl} - \widehat{w}_l$, where $\widehat{w}_l$ is a suitable estimator for w_l. In regression problems with multiple dimensions such as the present case outlined in Eq. 1, the question arises as to which effects are best modeled as random vs. modeled as fixed. In general, with a large number of coefficients to be estimated, the potential loss in degrees of freedom associated with modeling fixed effects is considered an argument in favor of random effects. In the large surveys considered in this article, several 100 interviewers were involved in taking measurements. Estimating a fixed effect for each interviewer is thus prohibitively expensive in terms of degrees of freedom. We therefore proceed in line with common practice and assume that the household effect u_j and interviewer effect w_l are stochastic (38–40). In case of the location effect v_k, the optimal choice is less clear. The potential loss of degrees of freedom is lower due to the lower number of coefficients to be estimated, especially at the highest level of geography. However, in order to maintain maximum comparability of the level effects, we consider it sensible to model all of them as random.

These random level effects are assumed to be independently drawn from underlying normal distributions (39). As part of the estimation procedure, we obtain estimates for the respective second moments of these distributions, which then can be used for simulation exercises or the calculation of reliability ratios. With the assumption of random effects, Eq. 1 constitutes a LMM, that is:

$$\begin{aligned} u_j &\sim N(0, \sigma_u^2), \quad j = 1, \ldots, J; \\ v_k &\sim N(0, \sigma_v^2), \quad k = 1, \ldots, K; \\ w_l &\sim N(0, \sigma_w^2), \quad l = 1, \ldots, L. \end{aligned}$$

Omitted variable bias

An individual's blood pressure depends on various factors, only some of which can be fully captured in large-scale surveys. Genetic preconditions for example are practically impossible to capture sufficiently in survey settings. Thus, we are agnostic about facing omitted variable bias in explaining systolic blood pressure independent of the particular survey dataset considered. However, depending on the survey, some essential predictors of blood pressure are missing, which in principle could be recorded in a survey setting.

Recalling that our main interest lies in investigating interviewer effects, we are mostly concerned about falsely attributing variation in systolic blood pressure measurements to interviewers. Confounding is most likely to occur if an interviewer's specific subset of individuals substantially differs from the overall population, along a dimension relevant for variation in systolic blood pressure.

The risk of confounded interviewer intercept estimates caused by small samples is mitigated by using the best linear unbiased predictor (BLUP) for random effects (41, 42). This estimator is a weighted average of the pooled sample and the sample from the level-specific subgroup, i.e. all measurements taken by one specific interviewer. The former exhibits a bias and small variance, whereas the latter is unbiased but has a large variance. It is constructed so that the more observations there are in the level-specific subgroup, the more weight is attributed to it. Conversely, if the level-specific subgroup sample is very small, the BLUP relies more heavily on the pooled sample. The estimation procedure therefore amounts to a variance-bias tradeoff in which the BLUP is optimal in terms of the mean squared error (MSE). Consequently, the potential small sample bias that leads to confounded interviewer intercept estimates is small, and we therefore consider its impact negligible.

It is in the nature of large-scale medical health surveys that observations may be subject to an intricate dependency structure. Clusters of dependent observations are imposed by the sampling design, e.g. by randomizing households to include in the study and not individuals, or by collecting repeated measurements on the same individuals. In general, not appropriately accounting for the resulting dependencies may skew statistical results.

To address such potential dependencies of interviewer effects, our model incorporates household and location effects besides the interviewer effects, aiming to capture the correlation structure inherent in the sampling design as effectively as possible. Furthermore, we argue that within-individual variation is at least partially accounted for by including covariates which were measured in every survey wave. We argue that the inclusion of these effects strongly mitigates the risk of incorrectly attributing variation in the measurement to the interviewers.

Testing for the presence of interviewer effects

We are interested in investigating the presence and significance of interviewer effects. This relates to the formal test of the hypothesis H_0: $\sigma_w^2 = 0$ vs. H_1: $\sigma_w^2 > 0$. This test is performed by evaluating the likelihood ratio statistic

$$\mathrm{LRT} = 2\left(\ell_{H_1} - \ell_{H_0}\right),$$

where ℓ_{H_0} is the log-likelihood of the model under the null and ℓ_{H_1} for the alternative. In our concrete case, ℓ_{H_1} nests ℓ_{H_0} and additionally includes interviewer random effects. As fundamental problem, the null lies at the boundary of the parameter space. The asymptotic distribution of the LRT has the inconvenient distribution of a point-mass on zero with weight 0.5 and χ_1^2-distribution elsewhere. The finite sample distributions however may severely differ from the asymptotic distribution (43, 44). For multiple random effects as in the present model, a parametric bootstrap can approximate the finite sample distribution well enough (45, 46). In particular,

$$\mathrm{LRT} \overset{\mathrm{d}}{\approx} aU\chi_1^2,$$

where $\overset{\mathrm{d}}{\approx}$ denotes approximate equality in distribution, $U \sim \mathrm{Bern}(1-p)$. Both a and p are unknown and have to be estimated by bootstrap replications. Eventually, P-values for the LRT under the null can be provided.

Adjusting for interviewer effects

Once we have established the presence and significance of interviewer effects, we adjust blood pressure measurements for these interviewer effects. Since we obtain not only an estimate of the second moment of the interviewer effect distribution but also intercepts for all individual interviewers, we can individually adjust systolic blood pressure measurements. A simple adjustment then takes the form

$$\widehat{\mathrm{Y}}_{ijkl}^{\mathrm{adj}} = \mathrm{Y}_{ijkl} - \widehat{w}_l, \tag{2}$$

where $\widehat{w}_l$ are the interviewer intercept effects (the BLUPs).

Assessing uncertainty in sample hypertension prevalence

In order to quantify the uncertainty in hypertension prevalence induced by interviewer measurement error we use a nonparametric bootstrap approach. Precisely, for this approach we repeatedly take subsamples of observed systolic blood pressure measurements and their corrected counterparts and compare resulting prevalences of hypertension. We depict the two generated sets of prevalences as densities, which allows for a straightforward comparison.

Bootstrap

We employ a nonparametric cluster bootstrap approach to infer about the uncertainty of hypertension prevalence given the corrected observations. We refer to this approach as nonparametric, since we do not use estimated parameters from the estimated model to generate new data, but only use the predicted interviewer effects to create adjusted measurements postestimation. Thus, we compare the density of hypertension prevalences based on corrected observations to the density of prevalences based on uncorrected observations. In order to account for the clustered structure of our data, we fix the coarsest geographic level (e.g. provinces) in the data and within these levels we draw from the second coarsest geographical level (e.g. municipalities).

The location level effects depict multiple levels of granularity and thus can also be represented as distinct effects. Let $p = 1, \ldots, P$ indicate the coarsest geographical level effect (e.g. province), and $m = 1, \ldots, M(p)$ represent the second coarsest geographical level effect (e.g. municipality).

Formally, let y_{ipm}, $m = 1, \ldots, M(p)$ be the ith individual measurements in province p and $\mathrm{y}_{ipm}^{\mathrm{adj}}$ the adjusted measurements respectively. Then, R bootstrap replications are generated via Algorithm 1.

Algorithm 1 *R* bootstrap replication

```
1: for $r = 1, \ldots, R$:
2: for $p = 1, \ldots, P$:
3: Draw $M(p)$ municipalities with replacement
4: Obtain composite sample
   $B(p) \subset \{y_{ipm} |$ for individual $i$ in municipality $m\}^{M(p)}$
5: Pool random samples to obtain $B = \cup B(p)$
6: Calculate $p_r(B) = |B|^{-1} \sum_{y \in B} \mathbb{1}(y > 140)^{p}$, and $p_r^{adj}$ analogously
```

The bootstrap prevalences $(p_r)_{r=1,\ldots,R}$ and $(p_r^{\mathrm{adj}})_{r=1,\ldots,R}$ allow for inferring about the difference in prevalences induced by the adjustment for interviewer effects.

Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

P.G. is a Chan Zuckerberg Biohub investigator. E.M. acknowledges funding from the National Institute on Aging (2R01AG030153). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. R.H. acknowledges funding from the German Research Foundation as part of Research Training Group 1723 "Globalization and Development."

Author Contributions

P.G. developed the research question, provided major contributions to the drafting of the manuscript, and provided major contributions to the methodological approach. A.Y.C. provided major contributions to the drafting of the manuscript and major contributions to the literature review. E.M. prepared the LASI data using Stata, provided minor contributions to the methodological approach, and applied the developed methodology to the LASI data using R. N.S. developed the research question and provided major contributions to the methodological approach. V.C. provided valuable comments about the methodology. P.K. and R.H. both provided major contributions to the drafting of the manuscript, major contributions to the methodological approach and conducted the coding in R.

Preprints

A preprint of this article is published at https://www.medrxiv.org/content/10.1101/2023.04.11.23288399v1.

Data Availability

The data used in this analysis cannot be shared. We used the Harmonized LASI data, version A.2 (47), augmented with three restricted variables (the interviewer identifier, the district identifier, and the identifier of the secondary sampling unit). The Harmonized LASI data are publicly available to registered users on https://g2aging.org (registration is free). The restricted variables are not publicly available. The complete data used in this analysis are available for IFLS and NIDS and can be accessed via the respective provider websites free of charge. The authors do not have permission to share these survey data.

We provide all analysis code in a publicly accessible repository https://github.com/rhaarb/IBMSD.

References

1 Boerma JT, Ghys PD, Walker N. 2003. Estimates of HIV-1 prevalence from national population-based surveys as a new gold standard. *Lancet*. 362(9399):1929–1931.

2 Clark A, Sanderson C. 2009. Timing of children's vaccinations in 45 low-income and middle-income countries: an analysis of survey data. *Lancet*. 373(9674):1543–1549.

3 Corsi DJ, Neuman M, Finlay JE, Subramanian SV. 2012. Demographic and health surveys: a profile. *Int J Epidemiol*. 41(6):1602–1613.

4 Mbondji PE, *et al*. 2014. Health information systems in Africa: descriptive analysis of data sources, information products and health statistics. *J R Soc Med*. 107(Suppl. 1):34–45.

5 Boerma JT, Sommerfelt AE. 1993. Demographic and health surveys (DHS: contributions and limitations. *World Health Stat Q*. 46(4):222–226.

6 Althubaiti A. 2016. Information bias in health research: definition, pitfalls, and adjustment methods. *J Multidiscip Healthc*. 9:211–217.

7 Ali S, Rouse A. 2002. Practice audits: reliability of sphygmomanometers and blood pressure recording bias. *J Hum Hypertens*. 16(5):359–361.

8 Cernat A, Sakshaug JW. 2020. Nurse effects on measurement error in household biosocial surveys. *BMC Med Res Methodol*. 20(1):1–9.

9 Svensson JC, Theorell T. 1982. Cardiovascular effects of anxiety induced by interviewing young hypertensive male subjects. *J Psychosom Res*. 26(3):359–370.

10 Ulijaszek SJ, Kerr DA. 1999. Anthropometric measurement error and the assessment of nutritional status. *Br J Nutr*. 82(3):165–177.

11 Jaszczak A, Lundeen K, Smith S. 2009. Using nonmedically trained interviewers to collect biomeasures in a national in-home survey. *Field Methods*. 21(1):26–48.

12 Dwyer-Lindgren L, *et al*. 2019. Mapping HIV prevalence in Sub-Saharan Africa between 2000 and 2017. *Nature*. 570(7760):189–193.

13 Graetz N, *et al*. 2018. Mapping local variation in educational attainment across Africa. *Nature*. 555(7694):48–53.

14 Osgood-Zimmerman A, *et al*. 2018. Mapping child growth failure in Africa between 2000 and 2015. *Nature*. 555(7694):41–47.

15 Reiner Jr RC, *et al*. 2018. Variation in childhood diarrheal morbidity and mortality in Africa, 2000–2015. *N Engl J Med*. 379(12):1128–1138.

16 Schulze MB, Kroke A, Bergmann MM, Boeing H. 2000. Differences of blood pressure estimates between consecutive measurements on one occasion: implications for inter-study comparability of epidemiologic studies. *Eur J Epidemiol*. 16:891–898.

17 Bogan B, Kritzer S, Deane D. 1993. Nursing student compliance to standards for blood pressure measurement. *J Nurs Educ*. 32(2):90–92.

18 Dickson BK, Hajjar I. 2007. Blood pressure measurement education and evaluation program improves measurement accuracy in community-based nurses: a pilot study. *J Am Acad Nurse Pract*. 19(2):93–102.

19 Cernat A, Sakshaug JW. 2021. Interviewer effects in biosocial survey measurements. *Field Methods*. 33(3):236–252.

20 Yusuf S, *et al*. 2020. Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (pure): a prospective cohort study. *Lancet*. 395(10226):795–808.

21 Zhou B, *et al*. 2017. Worldwide trends in blood pressure from 1975 to 2015: a pooled analysis of 1479 population-based measurement studies with 19.1 million participants. *Lancet*. 389(10064):37–55.

22 Jeet G, Thakur JS, Prinja S, Singh M. 2017. Community health workers for non-communicable diseases prevention and control in developing countries: evidence and implications. *PLoS One*. 12(7):e0180640.

23 Otieno CF, Kaseje D, Ochieng' BM, Githae MN. 2012. Reliability of community health worker collected data for planning and policy in a peri-urban area of Kisumu, Kenya. *J Community Health*. 37:48–53.

24 Singh P, Sachs JD. 2013. 1 million community health workers in Sub-Saharan Africa by 2015. *Lancet*. 382(9889):363–365.

25 Flint AC, *et al*. 2019. Effect of systolic and diastolic blood pressure on cardiovascular outcomes. *N Engl J Med*. 381(3):243–251.

26 Strandberg TE, Salomaa VV, Vanhanen HT, Pitkälä K, Miettinen TA. 2002. Isolated diastolic hypertension, pulse pressure, and mean arterial pressure as predictors of mortality during a follow-up of up to 32 years. *J Hypertens (Los Angel)*. 20(3):399–404.

27 Southern Africa Labour and Development Research Unit. National Income Dynamics Study (NIDS) Wave 1, 2008 [dataset]. Version 7.0.0. Pretoria: SA Presidency [funding agency]. Cape Town: Southern Africa Labour and Development Research Unit [implementer], 2018. Cape Town: DataFirst [distributor], 2018. https://doi.org/10.25828/e7w9-m033

28 Southern Africa Labour and Development Research Unit. National Income Dynamics Study Wave 2, 2010-2011 [dataset]. Version 4.0.0. Pretoria: SA Presidency [funding agency]. Cape Town: Southern Africa Labour and Development Research Unit [implementer], 2018. Cape Town: DataFirst [distributor], 2018. https://doi.org/10.25828/j1h1-5m16

29 Southern Africa Labour and Development Research Unit. National Income Dynamics Study Wave 3, 2012 [dataset]. Version 3.0.0. Pretoria: SA Presidency [funding agency]. Cape Town: Southern Africa Labour and Development Research Unit [implementer], 2018. Cape Town: DataFirst [distributor], 2018. https://doi.org/10.25828/7pgq-q106

30 Southern Africa Labour and Development Research Unit. National Income Dynamics Study 2014-2015, Wave 4 [dataset]. Version 2.0.0. Pretoria: Department of Planning, Monitoring, and Evaluation [funding agency]. Cape Town: Southern Africa Labour and Development Research Unit [implementer], 2018. Cape Town: DataFirst [distributor], 2018. https://doi.org/10.25828/f4ws-8a78

31 Southern Africa Labour and Development Research Unit. National Income Dynamics Study 2017, Wave 5 [dataset]. Version 1.0.0 Pretoria: Department of Planning, Monitoring, and Evaluation [funding agency]. Cape Town: Southern Africa Labour and Development Research Unit [implementer], 2018. Cape Town: DataFirst [distributor], 2018. https://doi.org/10.25828/fw3h-v708

32 Leibbrandt M, Woolard I, de Villiers L. 2009. Methodology: report on NIDS Wave 1. Technical Paper 1.

33 Wittenberg M. 2009. Weights: report on NIDS Wave 1. NIDS Technical Paper 2.

34 Sikoki BS, Witoelar F, Strauss J, Meijer E, Suriastini NW. 2013. Indonesia family life survey east 2012: user's guide and field report. Technical Report, SurveyMETER.

35 Strauss J, Witoelar F, Sikoki B. 2016. *The fifth wave of the Indonesia family life survey: overview and field report*. Vol. 1. Santa Monica (CA): Rand.

36 Strauss J, Witoelar F, Sikoki B, Wattie AM. 2009. The fourth wave of the Indonesia family life survey: overview and field report. RAND Labor and Population Working Paper WR-675/1-NIA/NICHD. Santa Monica, CA.

37 International Institute for Population Sciences (IIPS), MoHFW, Harvard T. H. Chan School of Public Health (HSPH) and the University of Southern California (USC). 2020. Longitudinal Ageing Study in India (LASI) wave 1, 2017–18, India Report.

38 Fielding A. 2004. The role of the Hausman test and whether higher level effects should be treated as random or fixed. *Multilevel Model Newsl*. 16:3–9.

39 Hodges JS. 2013. *Richly parameterized linear models*. New York (NY): CRC Press.

40 Hsiao C. 2014. *Analysis of panel data*. 3rd ed. New York (NY): Cambridge University Press.

41 Henderson CR. 1975. Best linear unbiased estimation and prediction under a selection model. *Biometrics*. 31(2):423–447.

42 Rao JNK, Molina I. 2015. *Small area estimation*. Hoboken (NJ): John Wiley & Sons.

43 Crainiceanu CM, Ruppert D. 2004. Likelihood ratio tests in linear mixed models with one variance component. *J R Stat Soc B*. 66: 165–185.

44 Crainiceanu CM, Ruppert D. 2005. Exact likelihood ratio tests for penalised splines. *Biometrika*. 92:91–103.

45 Crainiceanu CM. 2008. Likelihood ratio testing for zero variance components in linear mixed models. New York (NY): Springer. p. 3–17.

46 Greven S, Crainiceanu CM, Küchenhoff H, Peters A. 2008. Restricted likelihood ratio testing for zero variance components in linear mixed models. *J Comput Graph Stat*. 17(4):870–891.

47 Chien S, *et al*. 2021. Harmonized LASI documentation, version A.2 (2017–2019). RAND Working Paper Series WR-1018.