successfully pooled data from multiple years to study subpopulations of Asian Americans.[2]

COST

It is important to recognize that addressing power through increasing sample size or oversampling requires resources. For existing surveys this may require expenditures beyond existing budgets and for new investigations this requires budgeting appropriately. Choosing to collect sexual orientation or gender identity data in my experience is never a cost-saving decision and almost always involves tough trade-offs.

Money is not the only resource that investigators need to consider. An additional cost is space on data collection instruments. For existing surveys, the addition of a question may require the removal of another question, resulting in an opportunity cost.

AN OPTIMISTIC FUTURE

Despite the methodological challenges I have outlined, I want to mention how optimistic I am about the future. We can now show empirically that these data can be collected. Investigators and society just need to prioritize the collection of this information as they do with race and ethnicity. I am also optimistic because the number of researchers who want to collect this information is growing faster than I can respond to their requests for assistance. And I am mostly optimistic because I believe that emerging technologies, particularly methods that harness large online social networks and panels, may soon make the collection of information about rare and stigmatized populations ever more possible and routine.[7] AJPH

Randall L. Sell, ScD, MA, MS

REFERENCES

1. Sell RL; LGBTData.com. 2017. Available at: www.lgbtdata.com. Accessed May 10, 2017.

2. Paulose-Ram R, Burt V, Broitman L, Ahluwalia N. Overview of Asian American data collection, release, and analysis: National Health and Nutrition Examination Survey 2011–2018. *Am J Public Health*. 2017;107(6):916–921.

3. Jans M, Viana J, Grant D, Cochran SD, Lee AC, Ponce NA. Trends in sexual orientation missing data over a decade of the California Health Interview Survey. *Am J Public Health*. 2015;105(5):e43–e50.

4. Dahlhamer JM, Galinsky AM, Joestl SS, Ward BW. Sexual orientation in the 2013 National Health Interview Survey: a quality assessment. *Vital Health Stat 2*. 2014;2(169):1–24.

5. Herman JL; GenIUSS Group, eds. *Best Practices for Asking Questions to Identify Transgender and Other Gender Minority Respondents on Population-Based Surveys*. Los Angeles, CA: Williams Institute; 2014.

6. Sell RL, Kates J, Brodie M. Use of a telephone screener to identify a probability sample of gays, lesbians, and bisexuals. *J Homosex*. 2007;53(4):163–171.

7. Sell R, Goldberg S, Conron K. The utility of an online convenience panel for reaching rare and dispersed populations. *PLoS One*. 2015;10(12):e0144011.

Oversampling in Health Surveys: Why, When, and How?

Professional survey and polling firms often "oversample"[1] certain groups to better estimate attributes of that group and then use sampling weights in analyses to avoid unintended biases associated with oversampling.

WHY OVERSAMPLE?

How is this fact relevant to learning about the LGBT (lesbian, gay, bisexual, or transgender) population? Say that we wanted to do a survey of adults in America. That is our population of interest. Further say that we wanted to know the rates of hypertension among those who identify as "straight" and those who identify as "LGBT."

We do not have the time or money to assess all straight and LGBT people in the population, so we take a sample from the population (just as we would in any poll); for purposes of illustration and example, say we had the time and money to collect information about hypertension for 100 people. But if you simply took a random sample of 100 people, you might expect something like 96 people in that sample to identify as straight and about four to identify as LGBT.[2] If you were trying to describe the health characteristics of straight people, you would probably be fairly confident of your estimate of hypertension rates based on 96 people. You would probably feel much less comfortable characterizing the hypertension rates of LGBT people on the basis of answers from only four people.

So what to do? You could decide that that you will not report information about LGBT individuals because only four people identified as such, or you could decide that obtaining information about LGBT individuals is important and sample from the population in a different way to ensure that you surveyed more people identifying as LGBT. This intentional sampling process, designed to incorporate more (typically low-prevalence) members of a certain community into your sample, is called oversampling.

HOW TO OVERSAMPLE?

To learn more about this (relatively) small group of the population, one would intentionally include more of its members in the sample. Say that it is known from other surveillance data that there is a higher prevalence of LGBT individuals in certain cities, zip code areas, or metropolitan statistical areas, so we might

ABOUT THE AUTHOR

Roger Vaughan is an AJPH *associate editor and is with the Department of Biostatistics, Columbia University, New York, NY.*

Correspondence should be sent to Roger Vaughan, DrPH, MS, Mailman School of Public Health, Columbia University, 722 W 168th St, New York, NY 10032 (e-mail: roger.vaughan@columbia.edu). Reprints can be ordered at http://www.ajph.org by clicking the "Reprints" link.

This editorial was accepted May 5, 2017.

doi: 10.2105/AJPH.2017.303895

TABLE 1—Hypothetical Population and Sampling Percentages, and Creation and Application of Weights

Variable	Straight	LGBT
Population, %	96	4
Sample, %	80	20
Weight	1.2 (96/80)	0.2 (4/20)
Weight × sample n	96 (1.2 × 80)	4 (0.2 × 20)

Note. LGBT = lesbian, gay, bisexual, or transgender.

decide to oversample in those areas first until we selected, for example, 17 people who identified as LGBT. We would then choose the remaining 83 people randomly from the population (assuming that population proportions would result in about 80 people who say that they are straight and about three who say that they are LGBT[2]) to keep our sample size at 100. We are now much more confident about characterizing the hypertension rates of LGBT individuals on the basis of our sample of 20 people as opposed to four.

What we would not do is say that the prevalence of LGBT individuals in the population is 20% (20/100), because we purposefully sampled 20 such individuals to better describe their hypertension rates. When doing prevalence analyses, we would statistically "downweigh" those 20 observations to equal four, so the prevalence would not change (i.e., the true prevalence would still be four per 100, or 4%). But now we have used oversampling to learn something about a perhaps hard-to-reach or low-prevalence group.

Table 1 illustrates this process numerically; the first data row provides the estimated population prevalence for the two groups, and the second row shows the percentage of each group in our sample after oversampling (note that the "amount" of oversampling would be determined by the research team). The "weights" are calculated by taking the ratio of the population prevalence to the sample percentage, and one can see that when those weights are "applied" to the data, the rates return to the correct population proportions. Clearly, this example is simplified; the process of oversampling and calculation and application of weights is complex and a discipline unto itself, but the principle is the same.

WHEN TO OVERSAMPLE?

There are readily available sampling and statistical tools that can help one learn more about lower-prevalence populations without inducing bias in calculating prevalence rates. Therefore, the decision of whether to oversample in an LGBT health survey depends on the answer to a simple question: "Is learning about the health of LGBT individuals important or not?" AJPH

Roger Vaughan, DrPH, MS

REFERENCES

1. Mercer A. Oversampling is used to study small groups, not bias poll results. October 25, 2016. Available at: http://www.pewresearch.org/fact-tank/2016/10/25/oversampling-is-used-to-study-small-groups-not-bias-poll-results. Accessed May 16, 2017.

2. Gates GJ. In US, more adults identifying as LGBT. January 11, 2017. Available at: http://www.gallup.com/poll/201731/lgbt-identification-rises.aspx. Accessed May 16, 2017.

Recording Sexual Orientation in the UK: Pooling Data for Statistical Power

We know that sexual minority health disparities exist, but in the United Kingdom, the research demonstrating disparities in sexual minority health has been dominated by small convenience samples that do not represent clearly defined populations. Recently, UK population health surveys began to include a question on sexual orientation identity that makes available high-quality data. However, very few studies collect sexual orientation within their demographic data.[1] There need to be more, as it is this important, high-quality evidence that can be used to make a political impact and determine policy change.

Studies that collect data on sexual orientation and on health outcomes or behaviors and therefore allow prevalence of to be captured are the United Kingdom national longitudinal cohort study called "Understanding Society" (bit.ly/259UCLb) and several population cross-sectional studies. Data sets can be accessed through the UK Data Service (bit.ly/1Nz5cl3). Participant recruitment by the surveys is through random or stratified random sampling of their target population, which establishes generalizability of findings.

IDENTITY, ATTRACTION, BEHAVIOR

Sexual orientation was recorded in all of these included health surveys, using the standardized wording to capture sexual orientation identity that has been developed by the UK Office of National Statistics.[2] The sexual orientation identity question asks, "Which of the following options best describes how you think of yourself?" Participants can respond "heterosexual or straight," "gay or lesbian," "bisexual," or "other," or they can refuse to respond. This question does not measure sexual attraction or sexual behavior. These are different concepts well described in other literature.[3] A test of the impact of including the sexual orientation identity question in the Integrated Household Survey (2009–2010), which had a sample

ABOUT THE AUTHOR

Joanna Semlyen is with the Norwich Medical School, University of East Anglia, Norwich, United Kingdom.

Correspondence should be sent to Joanna Semlyen, Norwich Medical School, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ United Kingdom (e-mail j.semlyen@uea.ac.uk). Reprints can be ordered at http://www.ajph.org by clicking the "Reprints" link.

This editorial was accepted May 10, 2017.

doi: 10.2105/AJPH.2017.303910