Commentary on Reuter et al. - Data triangulation for substance abuse research

By: Martijn van Hasselt

Abstract:

Reuter *et al*. [1] highlight the limitations of using a general population survey (GPS) to determine the prevalence of frequent heroin use. Data from such surveys are likely to suffer from item non-response and under-reporting. In addition, frequent heroin use is relatively rare in the overall population, and the subpopulation of heavy users may, therefore, be poorly covered by the sampling frame.

Keywords: data triangulation | population survey | sampling bias | substance use | survey non-response | under-reporting

Article:

> *Estimating the prevalence of illegal drug use is notoriously difficult when using a general population survey. Data triangulation, or combining data from multiple sources, is a promising way forward but comes with its own set of challenges.*

Reuter *et al*. [1] highlight the limitations of using a general population survey (GPS) to determine the prevalence of frequent heroin use. Data from such surveys are likely to suffer from item non-response and under-reporting. In addition, frequent heroin use is relatively rare in the overall population, and the subpopulation of heavy users may, therefore, be poorly covered by the sampling frame. These problems are well-understood by survey scientists and do not just apply to heroin. Rather, the potential for survey bias arises whenever participants are asked about behaviors that are illegal, or when they worry about the "social desirability" of their responses [2-5]. In a striking illustration of survey bias, Reuter *et al*. [1] calculated two sets of estimates of the number of daily and near-daily heroin users in the United States between 2010 and 2016; one is based on data from the National Survey on Drug Use and Health (NSDUH), whereas another is obtained from a combination of multiple alternative data sources. The NSDUH estimates turn out to be an order of magnitude too low and unrealistically volatile from year to year.

It is timely to call renewed attention to the difficulty of quantifying substance use in the general population, and to the limitations of using survey samples for that purpose. Emerging evidence suggests that the COVID-19 pandemic may have led to significant increases in substance use and abuse [6-8]. Understanding these shifting patterns is critical for devising effective policy responses, and the demand for appropriate data and statistical methods will remain high for years to come.

One option to address problems with a GPS is to try and improve the quality of the survey itself, or simply switch to a different survey altogether; however, to quote Reuter *et al*. [1], "(…), thinking in terms of 'replace data system X with data system Y' perhaps misses the most fundamental point." The authors argue that when it comes to tracking problematic and harmful drug use, the key to advancing research and informing policy is to triangulate or combine data from multiple sources. In an age of ever-increasing availability of data, both from probability-based samples and non-survey sources such as administrative records and convenience samples, triangulation becomes a viable strategy that is particularly valuable when individual data sources are known to contain a weak signal at best.

Although triangulation is arguably appealing, its implementation is anything but straightforward. Specific methods for combining data from multiple sources are highly dependent on the features of those sources. For example, summary statistics from different samples may be combined using meta-analysis, which has a long history in statistics [9]. A more recent twist on meta-analysis is integrative data analysis (IDA), which uses pooled individual-level data from multiple sources [10-12]. In 2017, the National Academies of Sciences, Engineering and Medicine convened an expert panel that focused on data triangulation as a means to improve federal statistics for policy and research. This panel documented the heavy reliance of government agencies on survey samples, which suffer from increasing costs of administration and sustained declines in response rates [13]. The panel also discussed several state-of-the-art approaches for data triangulation as a promising way forward [14]. These approaches include record linkage, multiple frame methods, and imputation. Record linkage is based on matching multiple observations, either deterministically or probabilistically, from different data sets to the same sampling unit (e.g. an individual). Multiple frame methods provide a way to combine two or more samples with different but possibly overlapping target populations [15, 16]. Finally, imputation-based methods use statistical models to extrapolate or "fill in the gaps" that arise when combining data sets (e.g. [17]).

The potential benefits of data triangulation are significant. This approach can improve coverage of the target population, save resources by leveraging existing information, and help provide information about specific subpopulations or spatial and temporal patterns [16, 18]. At the same time, using data from multiple sources creates new challenges. Differences in respondents, in modes of data collection, and in measurements of key constructs may prevent meaningful triangulation. Multiple frame and imputation approaches are based on statistical models, the application of which often requires significant expertise. Responsible use of these statistical models also demands transparency about the underlying assumptions, as well as model checking and sensitivity analysis [14].

If data triangulation can rise to the challenge, it can significantly add to our understanding of substance use in hard-to-track populations. There is, unfortunately, no one-size-fits-all approach, and particular methods will have to be judged on their merits on a case-by-case basis. Nonetheless, such methods are bound to become critical tools for applied policy research in the future, and continued advocacy for their use remains essential.

References

1. Reuter P., Caulkins J. P., Midgette G. Heroin use cannot be measured adequately with a general population survey. *Addiction* 2021; https://doi.org/10.1111/add.15458

2. Kroutil L. A., Vorburger M., Aldworth J., Colliver J. D. Estimated drug use based on direct questioning and open-ended questions: responses in the 2006 National Survey on drug use and health. *Int J Methods Psychiatr Res* 2010; **19**: 74– 87; https://doi.org/10.1002/mpr302

3. McNeely J., Gourevitch M. N., Paone D., Shah S., Wright S., Heller D. Estimating the prevalence of illicit opioid use in new York City using multiple data sources. *BMC Public Health* 2012; **12**: 443 http://www.biomedcentral.com/1471-2458/12/443

4. Pepper J. V. How do response problems affect survey measurement of trends in drug use? In: C. F. Manski, J. V. Pepper, C. V. Petrie, editors. *Informing America's Policy on Illegal Drugs: What We Don't Know Keeps Hurting Us*. Washington, DC: The National Academies Press; 2001, pp. 321– 348.

5. Rehm J., Kilian C., Rovira P., Shield K. D., Manthey J. The elusiveness of representativeness in general population surveys for alcohol. *Drug Alcohol Rev* 2021; **40**: 161– 165; https://doi.org/10.1111/dar.13148

6. Slavova S., Rock P., Bush H. M., Quesinberry D., Walsh S. L. Signal of increased opioid overdose during COVID-19 from emergency medical services data. *Drug Alcohol Depend* 2020; **214**: 108176; https://doi.org/10.1016/j.drugalcdep.2020.108176

7. Niles J. K., Gudin J., Radcliff J., Kaufman H. W. The opioid epidemic within the COVID-19 pandemic: drug testing in 2020. *Popul Health Manag* 2021; **24**: S-43– S-51; https://doi.org/10.1089/pop.2020.0230

8. Pollard M. S., Tucker J. S., Green H. D. Changes in adult alcohol use and consequences during the COVID-19 pandemic in the US. *JAMA Netw Open* 2020; **3**: e2022942; https://doi.org/10.1001/jamanetworkopen.2020.22942

9. O'Rourke K. An historical perspective on meta-analysis: dealing quantitatively with varying study results. *J R Soc Med* 2007; **100**: 579– 582.

10. Curran P. J., Hussong A. M. Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychol Methods* 2009; **14**: 81– 100.

11. Horwood L. J., Fergusson D. M., Coffey C., Patton G. C., Tait R., Smart D., *et al*. Cannabis and depression: an integrative data analysis of four Australasian cohorts. *Drug Alcohol Depend* 2012; **126**: 369– 378; https://doi.org/10.1016/j.drugalcdep.2012.06.002

12. Silins E., Fergusson D. M., Patton G. C., Horwood L. J., Olsson C. A., Hutchinson D. M., *et al*. Adolescent substance use and educational attainment: an integrative data analysis comparing cannabis and alcohol from three Australasian cohorts. *Drug Alcohol Depend* 2015; **156**: 90– 96; https://doi.org/10.1016/j.drugalcdep.2015.08.034

13. National Academies of sciences, Engineering, and Medicine *Innovations in Federal Statistics: combining data sources while protecting privacy*. Washington, DC: The National Academies Press; 2017 https://doi.org/10.17226/24652

14. National Academies of sciences, Engineering, and Medicine *Federal Statistics, multiple data sources, and privacy protection: next steps*. Washington, DC: The National Academies Press; 2017 https://doi.org/10.17226/24893

15. Lohr S. L. Alternative survey sample designs: sampling with multiple overlapping frames. *Surv Methodol* 2011; **37**: 197– 213.

16. Lohr S. L., Raghunathan T. E. Combining survey data with other data sources. *Stat Sci* 2017; **32**: 293– 312; https://doi.org/10.1214/16-STS584

17. Jones H. E., Harris R. J., Downing B. C., Pierce M., Millar T., Ades A. E., *et al*. Estimating the prevalence of problem drug use from drug-related mortality data. *Addiction* 2020; **115**: 2393– 2404; https://doi.org/10.1111/add.15111

18. Citro C. F. From multiple modes for surveys to multiple data sources for estimates. *Surv Methodol* 2014; **40**: 137– 161.