Comment

Pitfalls of Using Data Portals as Sources for Psychological Research
The Example of Cross-national Homicide Data

Manuel Eisner
Institute of Criminology, University of Cambridge

Pasco Fearon
Research Department of Clinical, Educational and Health Psychology, University College London

Abstract (150 words)
Increasingly, data portals allow researchers to download data-sets generated by others for futher analyses. Open-access data have many benefits. However, researchers should be aware of the need to carefully assess the data before they are used. We illustrate the challenges using the example of the cross-national homicide data generate by the World Health Organisation (WHO) as part of their Global Health Estimates. For many countries these data are statistically modelled on the basis of complex algorithms, and WHO warns against their uncritical use. However, in some publications researchers have downloaded these data for regression models, using the same predictors that had been used by WHO for generating the estimates, meaning that the same information is used on both sides of the equation. In this comment we alert data providers, researchers, and reviewers to the need for caution when using secondary data.

Statement of Relevance (150 words)
The foundations of good science are valid and reliable data. The past two decades have seen an unprecedented increase in easily available social science data. If used with caution, they can help to advance knowledge, but they can lead to bad science if used uncritically. We illustrate the problem using the example of cross-national homicide data. Since 2004, the World Health Organization (WHO) releases estimates of the number of homicides for most countries globally, which can be accessed from various data portals. For many countries, these data are based on complex statistical models rather than actual counts, and the WHO has warned against their uncritical use. Yet recent publications have relied on these data without paying sufficient attention to how they were generated. More efforts by providers of data portals in documenting limitations, better training of researchers, and reviewers familiar with the data sources can help to limit the problem.

The past decades have seen an extraordinary growth of large open access data portals available for cross-national comparative social science research and a corresponding increase in the use of such data. In particular, international organisations like the World Bank, the United Nations Office on Drugs and Crime (UNODC), the United Nations Children's Fund (UNICEF), and the World Health Organization (WHO) provide users with comprehensive indicator systems on, e.g., the economy, the environment, education, health, child well-being, gender inequality, and conflict and violence. If adequately used, data retrieved from such data portals can help researchers to address important research questions. However, they can also lead to bad science if researchers fail to pay close attention to how data were generated. We illustrate this risk using a real-life example, namely data on cross-national homicide rates.

Homicide Rates

Homicide rates, the number of recorded intentional killings standardised by the population, are the most widely used indicator in large-N cross-national comparative research on interpersonal violence and its predictors. Cross-national studies have examined, for example, whether levels and trends in national homicide rates can be predicted by social inequality, trust in fellow citizens, poverty, family instability, ethnic cleavages, climate change, or urbanisation (Nivette, 2011; Trent & Pridemore, 2012). Researchers working in this field typically use either of two types of data: The first are national vital registration statistics that record the number of people killed a cording to the death certificates. They are collected by public health authorities and compiled internationally by WHO (World Health Organization, 2008). The second are criminal justice statistics that report the number of homicides recorded by the police. They are compiled internationally by UNODC (United Nations Office of Drugs and Crime, n.d.).

A Sudden Growth in Data Coverage

Both data sources have their limitations, but data derived from national vital registration statistics have widely come to be considered the gold standard in comparative homicide research (Andersson & Kazemian, 2018). However, until around 2010 limited sample sizes due to the lack of national cause of death data were a major impediment to research. Generally, earlier research included fewer than 70, mostly high-income countries, leading to concerns about whether findings are generalisable.

Around 2010 the situation seemingly improved radically. Between 2009 and 2016, at least 10 articles published in peer reviewed journals reported findings of cross-sectional or panel regression models that include 160 or more countries, meaning that that virtually all countries were considered (Kanis et al, 2017). How was this sudden growth in data coverage achieved and can the results reported in these studies be considered valid and reliable?

A Cautionary Note -- The Kanis et al (2017) Study

In 2017 a group of researchers including one of the authors of this comment examined the origin of the data used in these studies, and cautioned researchers about their appropriate use (Kanis et al, 2017). They showed that studies with large Ns had relied on data either directly retrieved from the WHO Global Health Estimates data portal (World Health Organization, 2008) or on the UNODC Global Study on Homicide data portal (United Nations Office on Drugs and Crime, 2011). The latter was partly based on data from the WHO Global Health Estimates.

Since 2004, the Global Health Estimates represent an effort, by WHO, to generate cause-specific estimates of the severity of health problems for societies across the globe, and to track progress towards global health targets (Mathers, 2020). In particular, they use complex statistical modelling techniques to quantify the mortality and the loss of health in countries where no or limited data are currently available. For homicide rates, the WHO researchers test

different prediction models with a large pool of covariates, iteratively optimising their models until they arrive at the final model. For the 2012 estimates, for example, six predictor variables were included in the final models: the gender inequality index, alcohol consumptions patterns, the percentage of people residing in urban areas, the male proportion of the population aged 15 to 30, the infant mortality rate, and the religious fractionalization (i.e. the degree of religious heterogeneity in a society) measure. The prediction model is adapted for each year for which estimates are produced.

As part of their Global Health Estimates program, WHO has released estimate-based homicide rates every four years since 2004. Kanis at al. (2017) showed that for a large number of countries, these data are not based on reports by national public health authorities to WHO, but on models computed by scientists at WHO. For the 2012 estimates, for example, national vital registration data are used for 54 countries (31%). For 17 countries criminal justice data are used (10%), for 30 countries national criminal justice data were adjusted by WHO (17%). However, for 72 countries (42%) the data were modelled by WHO researchers in the complete or partial absence of empirical data supplied by national authorities. This includes almost all countries on the African continent as well as a substantial number of countries in Asia and the Middle East (for a list see Kanis at al. 2017).

Kanis at al. (2017) note that the uncritical usage of regression-based estimates as though they were derived from national data collection efforts can lead to serious problems. First, the predictors used for modelling purposes include variables that are conceptually and empirically related to constructs used by researchers in their regression models. The same information is hence used on both sides of the equation. Second, as WHO adapts estimation paramaters in each estimation round, studies that estimate panel models wrongly attribute change in the estimation procedure to true change in violence levels. Third, missingness of homicide data is strongly concentrated geographically, with especially large gaps across the African continent. The estimated data therefore imply that covariates of homicide found in some parts of the world generate unbiased predictions in other regions, an assumption that has been shown does not always hold (Nivette, 2012).

Why did Researchers Fail to Notice?

WHO describes the estimation procedure and warns against the uncritical use of estimated data (World Health Organisation, 2014). Yet at least 10 peer-reviewed publications published between 2009 and 2017 present findings based on an uncritical use of the WHO estimates (Kanis et al., 2017: 319). Several others have been published after the publication of Kanis et al (2017).

Various factors may contribute to this problem: First, the ease of downloading data from large data portals may have lead researchers to pay not enough attention to documentation on the generation of the data. Second, several data portals including the World Development Indicators (World Bank, n.d.), the 2011 UNODC Homicide Statistics (UNODC, 2011), and university-led initiatives such as Clio Infra disseminate homicide data generated through WHO homicide estimation procedures. Their meta-data generally recommend that users consult the original source of the data. However, researchers may too frequently assume that downloading data from a respectable data portal is sufficient. Third, reviewers may not be aware of the underlying issues and fail to alert researchers to the challenges.

Conclusions

The past two decades have seen an unprecendented increase in the availability of open-access data that can easily be downloaded from the internet. Using secondary data-sets for social science research requires careful scrutiny of how the data were generated at the level of the primary data source, and whether they are adequate for the research purposes (Atkinson and

Brandolini, 2001). Failure to do so can result in poor science based on inadequate data. Several measures can help to reduce such problems: Data portals may consider additional efforts to document their data and alert readers to possible limitations; researchers should always track secondary data downloaded from portals to the primary data source and carefully document any challenges; And, finally, reviewers of manuscripts that use secondary data should ideally be familiar with the data so that they can assess whether the data have been used adequately.

References

Andersson, C., & Kazemian, L. (2018). Reliability and validity of cross-national homicide data: A comparison of UN and WHO data. *International Journal of Comparative and Applied Criminal Justice*, *42*(4), 287-302.

Atkinson, A. B., & Brandolini, A. (2001). Promise and pitfalls in the use of" secondary" data-sets: Income inequality in OECD countries as a case study. *Journal of Economic Literature*, *39*(3), 771-799.

Clio Infra (n.d.). Clio Infra – Reconstructing Global Inequality. https://clio-infra.eu.

Kanis, S., Messner, S. F., Eisner, M. P., & Heitmeyer, W. (2017). A cautionary note about the use of estimated homicide data for cross-national research. *Homicide Studies*, *21*(4), 312-324.

Mathers, C. D. (2020). History of global burden of disease assessment at the World Health Organization. *Archives of Public Health*, *78*(1), 1-13.

Nivette, A. E. (2011). Cross-national predictors of crime: A meta-analysis. *Homicide Studies*, *15*(2), 103-131.

Nivette, A. E. (2012). Spatial patterns of homicide and political legitimacy in Europe. *International Journal of Comparative and Applied Criminal Justice*, *36*(3), 155-171.

Trent, C. L. S., & Pridemore, W. A. (2012). *A review of the cross-national empirical literature on social structure and homicide.* In M. C. A. Liem & W. A. Pridemore (Eds.), *Handbook of European homicide research: Patterns, explanations, and country studies* (p. 111–135). Springer Science + Business Media.

United Nations Office on Drugs and Crime. (2011). 2011 *Global study on homicide.* Retrieved from http://www.unodc.org/unodc/en/data-and-analysis/statistics/crime/global-study-on-homicide-2011.html

United Nations Office on Drugs and Crime (n.d.) *DataUNODC homicide rates*, https://dataunodc.un.org

World Bank (n.d.) *World development indicators.* Washington, D.C. :The World Bank.

World Health Organization (2008). *Global health estimates - Global burden of disease 2008.* Retrieved from http://www.who.int/healthinfo/global_burden_disease/estimates_country/en/

World Health Organization, Violence and Injury Prevention (2014). *Global status report on violence prevention*. WHO, Geneva 2014.