Peertechz

MEDICAL GROUP

JOURNAL OF
Addiction Medicine and Therapeutic Science
OPEN ACCESS JOURNAL

ISSN: 2455-3484 DOI: https://dx.doi.org/10.17352/jamts

Commentary

Analysis of secondary data: Considerations revisited

Linda L Pederson[1]*, John J Koval[2] and Evelyn Vingilis[3]

[1]Adjunct Professor, Department of Epidemiology and Biostatistics, Department of Family Medicine, The Western Centre for Family Medicine and Public Health, University of Western University, 1839 Aldersbrook Road, London, Ontario N6G 3S3, Canada

[2]Professor Emeritus, Department of Epidemiology and Biostatistics, The Western Centre for Family Medicine and Public Health, University of Western Ontario, London Ontario N6G 2M1, Canada

[3]Director, Population & Community Health Unit, Professor, Department of Family Medicine, Department of Epidemiology and Biostatistics, The Western Centre for Public Health and Family Medicine, University of Western Ontario London, ON, N6G 2M1, Canada

Received: 09 June, 2022
Accepted: 28 June, 2022
Published: 29 June, 2022

***Corresponding author:** Linda L Pederson, Adjunct Professor, Department of Epidemiology and Biostatistics, Department of Family Medicine, The Western Centre for Family Medicine and Public Health, University of Western University, 1839 Aldersbrook Road, London, Ontario N6G 3S3, Canada, Tel: 519-661-9369; E-mail: lindap@mindspring.com

Copyright License: © 2022 Pederson LL, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

https://www.peertechzpublications.com

Check for updates

In a recent publication, we discussed the benefits and cautions of using secondary data analyses in research on lifestyle and health behavior [1]. We provided some guidelines about the use of secondary data in terms of the contributions that can be made and at the same time considerations necessary in using data that are collected by someone else. The use of secondary data to explore social and health issues results in being able to provide information about important issues in a timely fashion. Secondary data can answer two types of questions: descriptive and analytical [2]. Hence, the information can be used to describe events or trends or it can be used to examine relationships among variables cross-sectionally or longitudinally.

What are secondary data? Secondary data refer to the use of information that is collected by someone other than the original user and/or for purposes other than those originally intended. Common sources of secondary data can include census information, information collected by government departments on health and other behaviors, organization records such as hospital, housing, and arrest data, and information that was originally collected for research purposes such as public health surveys [2]. Data that can be used include both large-scale surveys and data collected using qualitative methods. Information can be collected on a regular basis from different samples [repeated cross-sectional] or from the same individuals over time [longitudinal].

It is crucial that in looking at using secondary data, the original parameters of data collection, question-wording, and sampling are clear and any sources of biases are noted [3]. Understanding the sources of the information, how it was collected, and how it was analyzed are precursors to the future appropriate use of the information [3,4].

Initially, we noted several Benefits. First, a lot of information is available from available sources and that can be used to make important content contributions to knowledge, as well as providing the backdrop for future research. Second, since the information is already available, research can be conducted promptly, without the timelines needed for submitting funding proposals. While ethics approval may be required if the information is being used for a purpose not originally proposed or have identifiers or sensitive information such as found on clinical records [2], ethics approval for the original data collection would be available. Continued use of secondary data gives researchers who have conducted the original information that they can use to justify the continuation of their research. In other words, the information gathered and used in secondary analyses often raises questions and can provide the backdrop for additional research.

The original discussion was followed by the presentation of Cautions: As noted, the available data may not provide all of the information of interest. This situation can lead to additional research, but it may leave some questions and explanations open. Second, response rates to surveys have been decreasing over time. In a detailed analysis of surveys conducted by Health and Human Services in the US, the decline from 1995-2015 ranged from 10% to over 20% [5]. Not only were there declines in response rates, but there were declines in completion rates and increases in item non-response. These patterns leave open

Citation: Pederson LL, Koval JJ, Vingilis E (2022) Analysis of secondary data: Considerations revisited. J Addict Med Ther Sci 8(1): 010-013.
DOI: https://dx.doi.org/10.17352/2455-3484.000054

the issue of exactly how representative the responses are. In addition, questions may not be worded as precisely as we would like to address specific areas currently under consideration It is still possible to learn from existing information and to use that information to provide the basis for new research and policy development.

Additional considerations

The purpose of this article is to outline some additional benefits and cautions that have come to light since our original publication and expand on some of the positive and negative issues associated with the use of secondary data.

Comparisons with other use: First, the data have been used, at least to some extent – so information on the analysis and some findings are available for comparisons and guidelines for potential future analyses. The use of definitions from the original research can help to provide guidance for future analyses; for example, cessation of a tobacco product can be defined as no use during a specific time such as one year. It is recommended that additional analyses include attempts to understand and verify original analyses so that future analyses are consistent with earlier findings.

Analytic issues: Surveys that are designed to be representative of a population often require complex statistical methodology and statistical programs that incorporate sample weights, stratification, and clustering. Users of these types of secondary survey data should ensure that they use the required sampling weights and design-based data analyses associated with the survey when attempting to replicate the original findings and when planning additional analyses. To do otherwise could lead to inappropriate estimates and standard errors of these estimates. For example, Kim, et al. [6] examined the percentage of research papers that employed appropriate statistical methodology while analyzing information from a secondary dataset, the Korean National Health and Nutrition Examination Survey [KNHANES]. Over 80% of published studies using the KNHANES used statistical analyses that were appropriate for datasets generated by a simple random sampling method but not analyses requiring incorporation of the complex survey design used in the collection of the information. The consequences of the inappropriate analyses were that the means and standard errors of the ordinary statistical analyses and the design-based analyses differed; the standard errors from the design-based analyses tended to be larger [6].

The comparison with earlier analyses does not mean that the potential use of information has to be limited to the original purposes of data collection. For example, prevalence estimation of patterns of drug use and other risk behaviors were among the original purposes of surveys conducted by the Centre for Addiction and Mental Health [CAMH] [7,8]. Subsequently, the relationships between sociodemographic variables and patterns of tobacco use among elementary and high school students were examined and provided additional information from the existing survey [9].

Particularly in secondary data analysis, one should avoid data dredging, that is, performing many statistical procedures and reporting only on those which are statistically significant. One should construct hypotheses from theoretical arguments, or as suggested by other studies, and then perform analyses on the secondary data. Even in this case, when working at a level of 0.05, about 5% of the hypotheses tested will be falsely found to be statistically significant.

If it is the case that the secondary analyses do not meet the standards for the required complex design and weightings assigned to the original data, this needs to be acknowledged. It is important that additional analyses do not call into question either the findings from the original analyses as well as those from additional analyses, but that the objectives are specified. However, if the complex design information can be used in the analysis, the estimates of variance are more precise. For example, analyses of the relationship between vaping and cigarette smoking in high school students from the CAMH surveys have been examined to learn about possible risks and protective factors and incorporated the complex designs in the subgroups included [9]

Combining datasets: Researchers are often interested in combining datasets from different years in order to 1] increase the sample size to get more precise estimates or 2] compute averages or differences in estimates from different years. One method is to stack the datasets for different years on top of each other and perform a "pooled" statistical analysis, using appropriate weightings. Some datasets come in an appropriate format, e.g., the National Survey on Drug Use and Health [10], but any statistical computer package can be used to concatenate datasets from different years. Lee, et al. [11] used a stacked dataset to estimate averages over years and differences between years, whereas Pederson, et al. [12] looked at changes in outcomes over years using a regression model. An alternative approach, used by Thomas and Wannell [13], is to compute estimates and their variances for each year and then calculate means [and their variances] over two or more years. Pederson, et al. [9] used the same methodology to compute differences in means and percentages between two years. In the analysis of pooled data, it is possible to use propensity score matching to examine the change in relationships over different years. However, there are challenges with this methodology; see, for example Norris [14].

Data quality: Some databases may be well developed methodologically, [i.e., Statistics Canada [Statcan] with their questionnaire development, sampling, etc.] and have included checks on data quality, missing values and misinformation. As an example, the Better Outcomes Registry & Network [BORN], Ontario's birth registry on pregnancy, birth and the early childhood, was independently evaluated for accuracy of selected data elements entered into the database [15]. Data entry includes data collection methods at different points during pregnancy, birth and childhood, and multiple data collection of the same data elements during the course of care. These data form a unified maternal-newborn record based on robust linking and matching algorithms. Accuracy assessment

Citation: Pederson LL, Koval JJ, Vingilis E (2022) Analysis of secondary data: Considerations revisited. J Addict Med Ther Sci 8(1): 010-013.
DOI: https://dx.doi.org/10.17352/2455-3484.000054

of the database indicated agreement from 56.9 to 99.8%, with 76% of the data elements with greater than 90% agreement [5]. Hence, users of secondary databases should be aware of and assess what procedures were used to improve quality of original data collection and to eliminate inconsistent or incomplete information.

Things to watch for: When dealing with information from longitudinal or repeated cross-sectional surveys, there are some additional issues that need to be considered. For example, question wording can change. Question sequencing can change as well as skip patterns. As a result, attempting to examine trends in use of substance over time may need to take into consideration variations in question sequencing, wording and response categories [16,17]. For example, evidence indicates that survey respondents' have provided different frequency estimates for behaviors based on the range of the response categories that were offered in a closed-ended question [18]. Additionally, the social and political environments can change; for example, cigarette smoking was acceptable forty years ago and now it is not. Therefore, smokers may be more likely to misrepresent their behavior in light of social desirability concerns [18]. In these cases, it is particularly difficult to decide how information from a range of years can be considered together. As a result, it is critical to understand the environment and conditions that existed when information was originally collected.

Data collection methods may differ. Information has been collected on line, on the telephone or in person. For example, information on a health record can be provided by the individual patient at one point in time and collected by the provider at another. It is not clear what the impact of the use of different data collection methods might be. In last few years many omnibus surveys and polls have changed from random digit dialing telephone survey [Computer Assisted Telephone Interviewing-CATI] to probability internet panel [PIP] surveys because of time and cost and the predominance of robo calls and blocking of unknown telephone numbers. Surveys by Hemsworth, et al. [19] documented what can result from the use of different data collection procedures. They collected information from a CATI [CATI, n = 502] and a PIP survey [PANEL, n = 530] to examine differences regarding attitudes and behavior toward livestock use and welfare. There was little difference in demographics between the two surveys apart from highest level of education. However, there were differences in both attitudes and behavior toward the red meat industry after controlling for education.

Conclusion

There can be important contributions to knowledge as well as directions for future research and programs that emerge from secondary analyses. It is crucial to be aware of differences in methods and to acknowledge them and the possible impact they may have on responding. These factors do not preclude the use of secondary data, but need to be acknowledged when attempting to present findings from secondary data. It may be the case that the methodological and environmental differences can account for what appear to be changes or trends and these potential impacts should be noted. It is essential to be aware of the many environmental factors that can impact response rate and response content. It is also necessary to keep in mind that finding relationships and trends over time does not provide evidence of causality but only descriptive information about the relationship between variables. Moreover, it is important to use appropriate statistical methods with secondary datasets with complex sampling designs and weightings.

References

1. Pederson LL, Vingilis E, Wickens CM, Koval J, Mann RE. Use of secondary data analyses in research: Pros and cons. Journal of Addiction Medicine and Therapeutic Science. 2020.
2. Huston P, Naylor CD. Health services research: reporting on studies using secondary data sources. Canadian Medical Association Journal. 1996; 155: 1697-1702
3. ALCHEMER (formerly Survey Gizmo). Why you should consider secondary data for your next study (2021). https://www.alchemer.com/resources/blog/secondary-data-analysis/#:~:text=Secondary%20data%20analysis%20involves%20a,question%20of%20a%20previous%20study
4. Crossman A. Understanding Secondary Data and How to Use It in Research. ThoughtCo. 2019. https://www.thoughtco.com/secondary-analysis-3026573
5. US Department of Health and Human Services USDHHS (2022). https://aspe.hhs.gov/sites/default/files/private/pdf/255531/Decliningresponserates.pdf
6. Kim Y, Park S, Kim NS, Lee BK. Inappropriate survey design analysis of the Korean National Health and Nutrition Examination Survey may produce biased results. J Prev Med Public Health. 2013 Mar;46(2):96-104. doi: 10.3961/jpmph.2013.46.2.96. Epub 2013 Mar 28. PMID: 23573374; PMCID: PMC3615385.
7. Boak A, Hamilton HA, Adlaf EM, Henderson JL, Mann RE. Drug use among Ontario students, 1977-2017: Detailed findings from the Ontario Student Drug Use and Health Survey (OSDUHS). (CAMH Research Document Series No. 47) Toronto, ON: Centre for Addiction and Mental Health. 2017. OSDUHS MH & Well-Being Detailed Report (camh.ca)
8. Boak A, Elton-Marshall T, Mann RE, Hamilton HA. Drug use among Ontario students, 1977-2019: Detailed findings from the Ontario Student Drug Use and Health Survey (OSDUHS) Toronto, ON: Centre for Addiction and Mental Health. 2020. OSDUHS Drug Use Report (camh.ca)
9. Pederson LL, Vingilis E, Koval JE Cigarette Use among Elementary (Grades 7 and 8) and High School (Grades 9-12) Students: Are the correlates of sociodemographic variables related to use different in 2017 and 2019? Unpublished manuscript. 2022.
10. Center for Behavioral Health Statistics and Quality. National Survey on Drug Use and Health (NSDUH). 2020. https://www.datafiles.samhsa.gov/dataset/nsduh-2002-2019-ds0001-nsduh-2002-2019-ds0001.
11. Lee S, Davis WD, Nguyen HA, McNeel T S, Brick JM, Flores-Cervantes I. Examining Trends and Averages Using Combined Cross-Sectional Survey Data from Multiple Years, CHIS (California Health Interview Survey) Methodology Paper, 2007.
12. Pederson LL, Koval J, Ialomiteanu AR, Chaiton M, Mann RE. What proportion of ever smokers quit? Analysis of information from CAMH from 1996- 2016. Journal of Addiction Medicine and Therapeutic Science. 2020; 6(1): 021-025.
13. Thomas S, Wannell B. Combining cycles of the Canadian Community Health Survey. Health Rep. 2009 Mar;20(1):53-8. PMID: 19388369.
14. Norris P. Using matching techniques with pooled cross-sectional data. Scottish Centre for Crime and Justice Research, University of Edinburgh. 2008. https://www.restore.ac.uk/Longitudinal/surveynetwork/seminars/s4/Norris_matching_techniques_pooled_cross_sectional_data.pdf

Citation: Pederson LL, Koval JJ, Vingilis E (2022) Analysis of secondary data: Considerations revisited. J Addict Med Ther Sci 8(1): 010-013.
DOI: https://dx.doi.org/10.17352/2455-3484.000054

15. Dunn S, Lanes A, Sprague AE, Fell DB, Weiss D, Reszel J, Taljaard M, Darling EK, Graham ID, Grimshaw JM, Harrold J, Smith GN, Peterson W, Walker M. Data accuracy in the Ontario birth Registry: a chart re-abstraction study. BMC Health Serv Res. 2019 Dec 27;19(1):1001. doi: 10.1186/s12913-019-4825-3. PMID: 31881960; PMCID: PMC6935171.

16. Pew Research Center. U.S. Survey Methodology. 2022. https://www.pewresearch.org/our-methods/u-s-surveys/u-s-survey-methodology/

17. Pew Research Center (ND) Writing Survey Questions (2021). https://www.pewresearch.org/our-methods/u-s-surveys/writing-survey-questions/

18. Schwarz N, Hippler H-J Deutsch B, Strack F. Response scales: Effects of category range on reported behavior and comparative judgments. Public Opinion Quarterly. 1985; 49:388-395.

19. Hemsworth LM, Rice M, Hemsworth PH, Coleman GJ. Telephone Survey Versus Panel Survey Samples Assessing Knowledge, Attitudes and Behavior Regarding Animal Welfare in the Red Meat Industry in Australia. Front Psychol. 2021 Apr 8;12:581928. doi: 10.3389/fpsyg.2021.581928. PMID: 33897517; PMCID: PMC8060561.

Discover a bigger Impact and Visibility of your article publication with Peertechz Publications

Highlights

- Signatory publisher of ORCID
- Signatory Publisher of DORA (San Francisco Declaration on Research Assessment)
- Articles archived in worlds' renowned service providers such as Portico, CNKI, AGRIS, TDNet, Base (Bielefeld University Library), CrossRef, Scilit, J-Gate etc.
- Journals indexed in ICMJE, SHERPA/ROMEO, Google Scholar etc.
- OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting)
- Dedicated Editorial Board for every journal
- Accurate and rapid peer-review process
- Increased citations of published articles through promotions
- Reduced timeline for article publication

Submit your articles and experience a new surge in publication services (https://www.peertechz.com/submission).

Peertechz journals wishes everlasting success in your every endeavours.

Citation: Pederson LL, Koval JJ, Vingilis E (2022) Analysis of secondary data: Considerations revisited. J Addict Med Ther Sci 8(1): 010-013.
DOI: https://dx.doi.org/10.17352/2455-3484.000054