Research Article

Use of secondary data analyses in research: Pros and Cons

LL Pederson*, E Vingilis, CM Wickens, J Koval and RE Mann

Adjunct Professor, Department of Family Medicine, Western University, Canada

Introduction

What are secondary data? Secondary data refer to data that are collected by someone other than the user or are used for an additional purpose than the original one. A wide range of sources can be used as secondary data: censuses, information collected by government departments, organizational records and data that were originally collected for other research purposes [1-3]. Yee and Niemeier [4] discuss the benefits of longitudinal data as compared to repeated cross-sectional information.

Use of repeated cross sectional or longitudinal secondary data to explore social and health issues can result in the ability to provide comparative information about important environmental issues. For example, social or health related information could be examined before, during and after the current COVID 19 pandemic to gain some understanding of the course and impact of the outbreak and to inform resource allocation. Using secondary analyses of survey data collected by the China CDC, Gao, et al. [5] was able to provide timely information to demonstrate geographical differences and duration of Coronavirus in health care workers in China.

Secondary data can answer two types of questions: descriptive and analytical. 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. Numerous secondary data bases exist and many are available online (e.g., The European Bioinformatics Institute database [6] provides a searchable database of biologic sources that can be linked to survey data). The Centre for Addiction and Mental Health (CAMH) conducts surveys in adults in Ontario, Canada (CAMH Monitor) that are repeated cross-sectional studies. The Monitor has been used in both descriptively and analytically and has provided important information on a multitude of health behaviors and policies.

Examples

An analysis of CAMH Monitor data from 1996-2006 provided important descriptive information about quitting smoking among individuals who were categorized as regular or occasional smokers. We found that the prevalence of having quit smoking for at least one year increased over time. In addition, females were more likely to show this increase than males, and older individuals more likely than younger ones [7]. These results provide us with the backdrop for examining additional questions in future research about why people quit, what programs might help people quit, and whether those who do quit are using new products that have become available such as e-cigarettes, waterpipes, smokeless tobacco and bidis. In addition, future research could be undertaken to explore whether methods of quitting have changed over time. Either survey questions could be developed to examine these issues or qualitative interviews could be used to supplement the information from the survey.

CAMH Monitor data have also been used descriptively to analyze effects of new legislation or policies by examining trends before and after the introduction of the legislation or policy, such as the potential impact of legislation on motor vehicle collisions in Ontario among smokers and nonsmokers. Legislation was enacted in Ontario in 2006 to prohibit smoking in vehicles when children and adolescents were present. We found that before the law was enacted the rate of reported collisions was higher among smokers than non smokers. Following the enactment of the legislation the rate among smokers decreased and there was no statistical difference between smokers and nonsmokers [8]. What is not known is whether drivers are in fact smoking while they are driving, their awareness of the legislation and whether their driving—smoking patterns changed because of the legislation. Another study examining cross-sectional CAMH data over time to assess legislative effects, found that texting and driving declined after introduction of more severe penalties [9].

058

Other examples of the use of CAMH Monitor data to evaluate policy interventions include Wickens, et al. [9] who assessed the impact of legislation to increase penalties for distracted driving on rates of texting and driving, and Mann, et al. [10] who evaluated the impact of legislation introducing administrative sanctions for impaired driving in on rates of driving after drinking in the province. These secondary analyses can also be supplemented with qualitative interviews to provide some explanation and background for the original findings.

Other types of secondary databases are longitudinal where large samples of individuals are followed over a number of years. For example, Wiesenthal and Vingilis [11] analyzed the Canadian National Population Health Survey (NPHS) descriptively and analytically to examine trends over time and relationships among variables. Specifically, they examined trajectories of distress in participants after they reported being injured from a motor vehicle collision. The NPHS, a Statistics Canada survey, is a repeated measures longitudinal survey to monitor the health and wellbeing of 20,000 Canadians. Participants were interviewed biennially from 1994/95 to 2002/03 (5 waves of interviews over a 9-year span). Because of the longitudinal nature of the secondary database, hierarchical linear modelling was used to identify within person trends; men experienced greater overall distress over time than women and a greater increase in distress over time. Moreover, the level of pre-injury distress predicted post-injury distress. This study revealed more complex and nuanced relationships among variables in their prediction of post-motor vehicle injury psychological distress. This secondary database provided numerous benefits. First, motor vehicle injuries are rare events; however, a sample of 20,000 individuals interviewed over 9 years provided enough cases of motor vehicle injury to examine the effects of injuries on distress. Additionally, evidence was mixed on whether pre-morbid distress predicted post-injury distress as all previous studies only had retrospective data on pre-injury distress levels. The use of a longitudinal secondary database provided information on distress levels before the injury occurred. The large sample size of injured individuals in this secondary database allowed for examination of mediators and moderators of the effects.

Finally, secondary data can be administrative data, that is, official records, such as hospital or police records. For example, the impact of new stunt driving legislation using stunt driving charges and collision casualty statistics, identified a decrease in charges and collision casualties among young males after the 2007 street racing legislation was introduced [12,13]. In addition, different types of secondary data can complement each other. Secondary data of hospital and police records can identify cases where individuals were apprehended or injured severely enough to go to hospital while self-report data identifies cases that might be missed by more official secondary data tools.

Discussion

Of course, there are some important factors that need to be considered in the use of secondary data.

Pros: First, there is much information available that has been collected in the past. This information can be used to make important contributions to knowledge, provide recommendations for policy, and provide the backdrop for future research.

Second, because the information is already available, subsequent research can be conducted in a timely manner, without the longer timelines for submitting proposals for funding and collecting original data. This is particularly salient because often events happen, such as the introduction of policies or historical events such as the current COVID 19 pandemic, before there is any opportunity for researchers to prepare to collect the relevant information needed to evaluate their impact. Third, often large sample sizes are available with secondary datasets, which is particularly important when investigating rare events. Moreover, certain types of secondary data have added benefits. For example, longitudinal secondary datasets have increased statistical power and can estimate a greater range of conditional probabilities compared to repeated cross-sectional secondary datasets [4].

The use of secondary data also gives researchers who have conducted the original surveys additional information that they can use to justify continuation of their original research. For example, there is strong epidemiological evidence connecting cannabis use to collision risk [13-16] that has spurred and informed experimental simulation studies examining precisely how cannabis affects driving [18,19].

Cons: As noted, secondary data may not provide all of the information of interest. Questions may not be worded as precisely as we would like to answer specific questions of interest. Analyses become more complicated if the question wording or methods of administration vary. In these cases, it is particularly difficult to decide how information from a range of years can be considered together. It is also critical to understand how the information was originally collected. Response rates to surveys have decreased over time, calling into question how representative the responses might be, which must be considered in the interpretation of secondary analyses. However, many well designed surveys include sampling weights to counter the biases that may occur from non-representative sampling. Longitudinal secondary datasets can suffer from attrition, although this is sometimes addressed by replacing lost respondents [4].

Online surveys are limited to those with access to the technology; targeted sub-groups who may not be the groups of interest when doing secondary analysis; and are correlational precluding cause and effect conclusions. Finally, ethics approval may be required if the information is being used for a purpose not originally proposed

Conclusion

It is important to make note of the limitations when presenting the information from secondary data and what the potential impact on the interpretation of the results can be. Nevertheless, secondary analysis can make important contributions to knowledge as well as provide directions for

future research and programs. Tripathy (2013) [20] notes that while secondary data analysis can make important contributions to knowledge, it is important to follow specific guidelines in the use of such information, one of the most important being anonymization of the information.

Acknowledgment

References

1. Boslaugh S (2007) Secondary data sources for public health: A practical guide. Cambridge University Press. **Link:** https://bit.ly/3iJblls

2. Smith E (2008) Using secondary data in educational and social research. New York, NY: McGraw-Hill Education. **Link:** https://bit.ly/3fbbf3U

3. Johnston MP (2014) Secondary Data Analysis: A Method of which the time has come. Qualitative and Quantitative Methods in Libraries 3: 619-626. **Link:** https://bit.ly/2Z5YhPo

4. Yee JL, Niemeier D (1996) Advantages and disadvantages: Longitudinal vs. repeated cross-section surveys. **Link:** https://bit.ly/31UcK2J

5. Gao W, Sanna M, Tsai MK, Wen CP (2020) Geo-temporal distribution of 1,688 Chinese healthcare workers infected with COVID-19 in severe conditions—A secondary data analysis. Plos one 15: e0233255. **Link:** https://bit.ly/3gBsH1V

6. The European Bioinformatics Institute. **Link:** https://bit.ly/2ZRK3AQ

7. Pederson LL, Koval J, Ialomiteanu AR, Chaiton M, Mann RE (2020) What proportion of ever smokers quit? Analysis of information from CAMH from 1996-2016. J Addict Med Ther Sci 6: 21-25. **Link:** https://bit.ly/31V3hIz

8. Pederson LL, Koval J, Vingilis E, Seeley J, Ialomiteanu AR, et al. (2019) The relationship between motor vehicle collisions and cigarette smoking in Ontario: Analysis of CAMH survey data from 2002 to 2016. Prev Med Rep. 13: 327-331. **Link:** https://bit.ly/2Z72jHi

9. Wickens CM, Ialomiteanu AR, Cook S, Hamilton H, Haya M, et al. (2020) Assessing the impact of the 2015 introduction of increased penalties and enhanced public awareness and enforcement activities on texting while driving among adults in Ontario, Canada. Traffic Inj Prev 21: 241-246. **Link:** https://bit.ly/2ZMugmV

10. Mann RE, Smart RG, Stoduto G, Adlaf EM, Vingilis E, et al. (2000) Changing drinking-driving behaviour: The effects of Ontario's administrative driver's licence suspension law. Can Med Assoc J 162: 1141-1142. **Link:** https://bit.ly/2ZU0gWu

11. Wiesenthal, N., Vingilis, E. (2013) The impact of motor vehicle injury on distress: Moderators and trajectories over time. Transportation Research Part F: Traffic Psychology and Behaviour 21, 1-13. **Link:** https://doi.org/10.1016/j.trf.2013.08.004

12. Meirambayeva A, Vingilis E, McLeod A, Elzohairy I, Xiao Y, et al. (2014) Road safety impact of Ontario street racing and stunt driving law. Accid Anal Prev 71: 72-81. **Link:** https://bit.ly/2BESBmQ

13. Meirambayeva A, Vingilis E, Zou G, Elzohairy Y, McLeod AI, et al. (2014) Evaluation of deterrent impact of Ontario's street racing and stunt driving law on extreme speeding convictions. Traffic Inj Prev 15: 786-793. **Link:** https://bit.ly/3gDEBsk

14. Asbridge M, Hayden JA, Cartwright JL (2012) Acute cannabis consumption and motor vehicle collision risk: Systematic review of observations studies and meta-analysis. BMJ 344: e536. **Link:** https://bit.ly/2ZQnZGG

15. Li MC, Brady JE, DiMaggio CJ, Lusardi AR, Tzong KY, et al. (2012) Marijuana use and motor vehicle crashes. Epidemiol Rev 341: 65-72. **Link:** https://bit.ly/2Z6wwX5

16. Rogeberg O (2019) A meta-analysis of the crash risk of cannabis-positive drivers in culpability studies-Avoiding interpretational bias. Accid Anal Prev 123: 69-78. **Link:** https://bit.ly/2ACfq9Y

17. Rogeberg O, Elvik R (2016) The effects of cannabis intoxication on motor vehicle collision revisited and revised. Addiction 111: 1348-1359. **Link:** https://bit.ly/3gGMRaH

18. Brands B, Mann RE, Wickens CM, Sproule B, Stoduto G, et al. (2019) Acute and residual effects of smoked cannabis: Impact on driving speed and lateral control, heart rate, and self-reported drug effects. Drug Alcohol Depend 205: 107641. **Link:** https://bit.ly/3iE5vC1

19. Downey LA, King R, Papafotiou K, Swann P, Ogden E, et al. (2013) The effects of cannabis and alcohol on simulated driving: Influences of dose and experience. Accid Anal Prev 50: 879-886. **Link:** https://bit.ly/38z0Ixc

20. Tripathy JP (2013) Secondary data analysis: Ethical issues and challenges. Iran J Public Health 42: 1478-1479. **Link:** https://bit.ly/3fa9Fzf