RESEARCH · QUANTITATIVE AND MIXED METHODS

Secondary Data Research: Using Existing Datasets Well

How to use existing datasets well in secondary data research, covering types of secondary data, advantages and limitations, how to evaluate a dataset's fit and quality, analysis practices that avoid common pitfalls, and ethical questions.

ModQu research article · 1,924 words · 20 references (APA 7th edition)

Not every research question needs new data. Governments, survey organisations, researchers, businesses and public bodies collect vast amounts of information, much of which can be reused. **Secondary data research** analyses data that were collected by someone else, usually for a different purpose. Hakim (1982) defined secondary analysis as any further analysis of an existing dataset that presents interpretations, conclusions or knowledge additional to, or different from, those presented in the first report. Done well, it can answer questions faster and more cheaply than primary research, and sometimes it can answer questions that no new study could. Done badly, it can produce confident conclusions from data that were never fit for the purpose. This article explains how to use existing datasets well.

Primary and secondary data

Primary data are collected by the researcher for the study at hand. **Secondary data** already exist. Hox and Boeije (2005) noted that the distinction depends on the relationship between the researcher and the data, not on the data themselves: a survey is primary data for the team that designed it and secondary data for anyone who later reanalyses it.

Glass (1976) distinguished primary analysis, the original analysis of data; secondary analysis, the reanalysis of data to answer the original question with better methods or to answer new questions; and meta-analysis, the statistical combination of results from many studies. This article focuses on secondary analysis of datasets. Desk research using published sources and reports is covered separately.

A long tradition

Secondary analysis is one of the oldest forms of social research. Durkheim (1951), in a study first published in 1897, used official statistics on suicide from several European countries to argue that suicide rates varied with social conditions such as religion, family life and economic change. He collected no new data; his contribution lay in the questions he asked of records that already existed. The study also illustrates the main risk of the approach. Later critics pointed out that suicide statistics depend on how coroners and officials classify deaths, which can vary with religion and local custom, the very factors Durkheim was studying. A pattern in secondary data may reflect how the data were produced rather than the reality they were meant to record. That lesson applies as much to modern datasets as to nineteenth-century records, and it is why understanding how data were generated is the first task of any secondary analysis.

Types of secondary data

Secondary data come from many sources. The diagram below shows the main types.

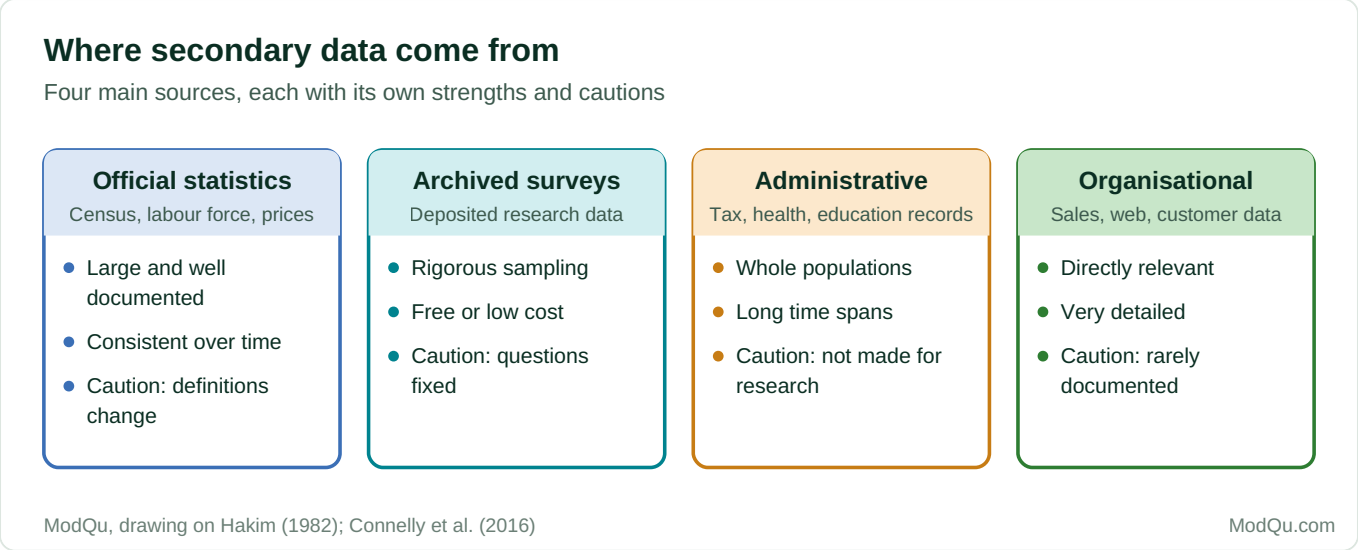


Figure 1. Main types of secondary data. Diagram by ModQu, drawing on Hakim (1982) and Connelly et al. (2016).

Official statistics are produced by national statistical agencies and government departments: population censuses, labour force surveys, trade statistics and price indices. They are usually large, carefully documented and collected consistently over time.

Archived research data are datasets deposited by researchers in data archives, such as national social science data services. Kiecolt and Nathan (1985) described how large national surveys, deposited for reuse, made it possible for researchers without large budgets to study national trends.

Administrative data are records created in running a service or organisation: tax records, health records, school enrolments, benefit claims. Connelly et al. (2016) argued that administrative data are among the most valuable sources for social research because they often cover whole populations over long periods, but noted that they were not designed for research and require careful handling.

Organisational and commercial data include sales records, customer databases, web analytics and loyalty card data. These can answer business questions directly, but may be proprietary and are rarely documented for outside users.

Qualitative data can also be reused. Heaton (2008) described secondary analysis of interview transcripts, field notes and diaries, which raises particular questions about whether data can be understood without the context in which they were collected.

Why use secondary data

Secondary data offer several advantages (Cheng & Phillips, 2014; Smith, 2008):

- **Cost and time.** The data already exist, so collection costs are avoided.
- **Scale.** National surveys and administrative records often have samples far larger than any single project could achieve.
- **Quality.** Major surveys are designed and run by specialists with rigorous sampling and fieldwork procedures.
- **Time span.** Repeated surveys and long-running records allow trends to be studied over decades.
- **Access to the past.** Historical questions can only be answered with data collected at the time.
- **Reduced burden.** Reusing data avoids asking more people to take part in research.

Elder (1974) demonstrated the value of reuse in a classic study. He used data collected decades earlier in a longitudinal study of Californian children to examine how the economic hardship of the Great Depression shaped their adult lives, a question the original researchers had not set out to answer. Because the original study had followed the same children over many years, Elder could compare those whose families lost much of their income with those who did not, and trace the differences into adulthood. No new study could have recreated those records.

Limitations

The main limitation is that the data were collected for someone else's purpose. Vartanian (2011) described several consequences:

- **Fit.** The variables may not measure exactly what the researcher wants, the population may differ from the one of interest, or key variables may be missing.
- **Control.** The researcher cannot change how questions were asked, who was sampled or when data were collected.
- **Documentation.** Codebooks may be incomplete, definitions may change over time, and important details of collection may be unrecorded.
- **Familiarity.** Learning a complex dataset takes time, and errors in understanding it are easy to make.
- **Age.** Data may be several years old by the time they are released.

Boslaugh (2007) emphasised that these limitations do not make secondary data inferior; they make careful evaluation essential before any analysis begins.

Evaluating a dataset

Before using a dataset, the researcher should be able to answer a set of basic questions about it. The table below lists the main ones.

Is this dataset fit for my question?

Six questions to answer before analysis

Question	What to check
Purpose	Who collected the data, why, and could that purpose shape them?
Population	Who was included or left out; sampling method; response rate
Measures	Exact definitions and wording; changes between years
Timing	When data were collected; how old they are now
Quality	Missing data, known errors, earlier checks by other users
Access	Licence, conditions of use, limits on publication

Based on Smith (2008); Vartanian (2011); Groves et al. (2009) ModQu.com

Figure 2. Questions to ask before using a secondary dataset. Based on Smith (2008), Vartanian (2011) and Groves et al. (2009); table by ModQu.

Purpose. Why were the data collected, and by whom? Data collected to support a campaign or to meet a performance target may be shaped by that purpose.

Population and sample. Who was included, and who was left out? How were participants selected? What was the response rate? Groves et al. (2009) showed how coverage, sampling and non-response errors can bias survey estimates, and these errors remain in the data when they are reused.

Measures. How exactly was each variable defined and measured? What question wording was used? Did definitions change between years? A change in how unemployment is defined can create an apparent trend that is entirely artificial.

Timing. When were data collected, and does that matter for the question?

Quality. How much is missing, and why? Are there known errors? Has the dataset been used and checked by other researchers?

Access and conditions. Who may use the data, under what licence, and with what restrictions on publication?

Smith (2008) recommended reading the original documentation and reports in full, and replicating published figures from the dataset before starting new analysis, to confirm that the data are being read correctly.

Administrative data in particular

Administrative records require extra care. Connelly et al. (2016) noted that they record what was needed for administration, not what researchers would like to know. A health record notes diagnoses that were made and treatments that were billed, not every illness people had. Changes in policy or practice, such as new coding rules or eligibility

criteria, can change the records without any change in the underlying reality. Benchimol et al. (2015) developed the RECORD reporting guideline for studies using routinely collected health data, which asks researchers to report how data were linked, which codes were used to define conditions and how the validity of those codes was assessed.

Analysis practices

Secondary data pose a particular risk for analysis. Because the data already exist, researchers can explore them freely, trying many variables, subgroups and models until something interesting appears. Gelman and Loken (2014) described this as the **garden of forking paths**: even without deliberate fishing, the many choices available in analysing an existing dataset make it easy to find patterns that arise by chance. Weston et al. (2019) recommended practices for increasing transparency when analysing pre-existing data, including:

- writing down hypotheses and an analysis plan before examining the key variables;
- stating what the researcher already knew about the data when the plan was written;
- distinguishing confirmatory analyses from exploratory ones in the report;
- holding back part of the data to test findings from the exploratory part;
- sharing code so that others can reproduce the analysis.

Other good practices include using survey weights where the data require them, handling missing data with appropriate methods rather than simply dropping incomplete cases, and checking whether results change under reasonable alternative definitions.

Data sharing and archiving

Secondary research depends on primary researchers making their data available. Wilkinson et al. (2016) set out the **FAIR principles**: data should be findable, accessible, interoperable and reusable. Corti et al. (2019) provided practical guidance on preparing data for sharing, including documentation, consent for reuse and anonymisation. Researchers who reuse data should cite the dataset itself, not only the original publication, so that data creators receive credit.

Ethics and privacy

Reusing data raises ethical questions. Participants consented to the original study, and may not have anticipated reuse for different purposes. Most data archives handle this by requiring that data be anonymised and that users agree to conditions of use. But anonymisation is harder than it seems. Sweeney (2002) showed that a large share of the population of the United States could be uniquely identified from just their postcode, date of birth and sex, and proposed **k-anonymity**, a standard requiring that each record be indistinguishable from at least a set number of others. Narayanan and Shmatikov (2008) showed that people in a supposedly anonymous dataset of film ratings could be re-identified by matching their ratings with public reviews. Researchers using secondary data should never attempt to identify individuals, should combine datasets only when permitted, and should take particular care with small groups and detailed geographical information.

An applied example

A hypothetical example: a company considering opening stores in a new region wants to estimate demand. Rather than commission a survey, it uses census data on population and income, a national household expenditure survey showing spending on its product category, and its own sales data from similar regions. Before analysis, the team checks that the expenditure survey's product categories match its own, that census income bands are consistent across years, and that the comparison regions are similar in key respects. The combined analysis gives an estimate of demand with a range of uncertainty, at a fraction of the cost of primary research. The team notes the main limitation, that the expenditure survey is three years old, and plans a small primary survey to check the most uncertain assumption.

Conclusion

Secondary data research can answer questions quickly, cheaply and at scale, and sometimes it is the only way to study the past or whole populations. Its value depends on understanding where the data came from, what they measure and what they leave out. Researchers who evaluate datasets carefully, analyse them transparently, and respect the privacy of the people behind the records can turn existing data into new knowledge.

References

Benchimol, E. I., Smeeth, L., Guttmann, A., Harron, K., Moher, D., Petersen, I., Sørensen, H. T., von Elm, E., Langan, S. M., & RECORD Working Committee. (2015). The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLOS Medicine, 12*(10), Article e1001885. https://doi.org/10.1371/journal.pmed.1001885

Boslaugh, S. (2007). *Secondary data sources for public health: A practical guide*. Cambridge University Press. https://doi.org/10.1017/CBO9780511618802

Cheng, H. G., & Phillips, M. R. (2014). Secondary analysis of existing data: Opportunities and implementation. *Shanghai Archives of Psychiatry, 26*(6), 371–375. https://doi.org/10.11919/j.issn.1002-0829.214171

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research, 59*, 1–12. https://doi.org/10.1016/j.ssresearch.2016.04.015

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2019). *Managing and sharing research data: A guide to good practice* (2nd ed.). SAGE.

Durkheim, E. (1951). *Suicide: A study in sociology* (J. A. Spaulding & G. Simpson, Trans.). Free Press. (Original work published 1897)

Elder, G. H., Jr. (1974). *Children of the Great Depression: Social change in life experience*. University of Chicago Press.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460–465. https://doi.org/10.1511/2014.111.460

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3–8. https://doi.org/10.3102/0013189X005010003

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Wiley.

Hakim, C. (1982). *Secondary analysis in social research: A guide to data sources and methods with examples*. George Allen & Unwin.

Heaton, J. (2008). Secondary analysis of qualitative data: An overview. *Historical Social Research, 33*(3), 33–45. https://doi.org/10.12759/hsr.33.2008.3.33-45

Hox, J. J., & Boeije, H. R. (2005). Data collection, primary versus secondary. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. 1, pp. 593–599). Elsevier. https://doi.org/10.1016/B0-12-369398-5/00041-4

Kiecolt, K. J., & Nathan, L. E. (1985). *Secondary analysis of survey data*. SAGE. https://doi.org/10.4135/9781412985796

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy* (pp. 111–125). IEEE. https://doi.org/10.1109/SP.2008.33

Smith, E. (2008). *Using secondary data in educational and social research*. Open University Press.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(5), 557–570. https://doi.org/10.1142/S0218488502001648

Vartanian, T. P. (2011). *Secondary data analysis*. Oxford University Press.

Weston, S. J., Ritchie, S. J., Rohrer, J. M., & Przybylski, A. K. (2019). Recommendations for increasing the transparency of analysis of preexisting data sets. *Advances in Methods and Practices in Psychological Science, 2*(3), 214–227. https://doi.org/10.1177/2515245919848684

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*, Article 160018. https://doi.org/10.1038/sdata.2016.18