Shanghai Archives of Psychiatry, 2014, Vol. 26, No. 6

•Research methods in psychiatry•

Secondary analysis of existing data: opportunities and implementation

Hui G. CHENG[1]*, Michael R. PHILLIPS[1,2]

Summary: The secondary analysis of existing data has become an increasingly popular method of enhancing the overall efficiency of the health research enterprise. But this effort depends on governments, funding agencies, and researchers making the data collected in primary research studies and in health-related registry systems available to qualified researchers who were not involved in the original research or in the creation and maintenance of the registry systems. The benefits of doing this are clear but the barriers are many, so the effort of increasing access to such material has been slow, particularly in low- and middle-income countries. This article introduces the rationale and concept of the secondary analysis of existing data, describes several sources of publicly available datasets, provides general guidelines for conducting secondary analyses of existing data, and discusses the advantages and disadvantages of analyzing existing data.

Key words: statistical data interpretation; secondary analysis; existing data; data collection; National Institute of Health

[*Shanghai Arch Psychiatry*. 2014; **26**(6): 371-375. doi: http://dx.doi.org/10.11919/j.issn.1002-0829.214171]

1. Background

A typical mental health research project begins with the development of a comprehensive research proposal and is (hopefully) followed by the successful acquisition of funding; the researcher then collects data, analyzes the results, and writes-up one or more research reports. Another less common, but no less important, research method is the analysis of existing data. The analysis of existing data is a cost-efficient way to make full use of data that are already collected to address potentially important new research questions or to provide a more nuanced assessment of the primary results from the original study. In this article we discuss the distinction between primary and secondary data, provide information about existing mental health-related data that are publically available for further analysis, list the steps of conducting analyzes of existing data, and discuss the pros and cons of analyzing existing data.

2. Data sources

2.1 'Primary data', 'secondary data', or 'existing data'?

There is frequently confusion about the use of the terms 'primary data', 'primary data analysis', 'secondary data', and 'secondary data analysis'. This confusion arises because it is never completely clear whether data employed in an analysis should be considered 'primary data' or 'secondary data'. Based on the usage of the National Institute of Health (NIH) in the United States, 'primary data analysis' is limited to the analysis of data by members of the research team that collected the data, which are conducted to answer the original

[1] Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[2] Departments of Psychiatry and Global Health, Emory University, Georgia, United States
*correspondence: chengyaojin@yahoo.com

A full-text Chinese translation of this article will be available at www.shanghaiarchivesofpsychiatry.org on January 25, 2015.

hypotheses proposed in the study. All other analyses of data collected for specific research studies or analyses of data collected for other purposes (including registry data) are considered 'secondary analyses of existing data', whether or not the persons conducting the analyses participated in the collection of the data. This replacement of the traditional term 'secondary data analysis' with the term 'secondary analysis of existing data' is a much clearer categorization because it avoids the confusion of trying to decide whether the data employed in an analysis is 'primary data' or 'secondary data'.

Of course, there are cases where the distinction is less clear. One example would be the analysis of data by a researcher who has no connection with the data collection team to address a research question that overlaps with the hypotheses considered in the original study. Another example would be when a member of the original research team subsequently revisits the original hypothesis in an analysis that uses different statistical methods. These situations commonly occur in the analyses of large-scale population surveys where the research questions are generally broad (e.g., sociodemographic correlates of depression) and when the participating researchers share the cleaned data with the broader research community. In both of these situations, based on a strict application of the NIH usage, the analyses would be considered 'secondary analysis of existing data' NOT 'primary data analysis' and NOT 'secondary data analysis'. In fact, we recommend avoiding the ambiguous term 'secondary data analysis' entirely.

2.2 Sources of existing data

Existing data can be private or public. To maximize the output of data collection efforts, researchers often assess many more variables than those strictly needed to answer their original hypotheses. Often times, these data are not fully used or explored by the original research team due to restrictions in time, resources, or interest. Unfortunately, the vast majority of these completed datasets are not made available, and in many countries (including China), there isn't even a registry or other means of determining what data have been previously collected about a specific research topic (so there are many unnecessarily duplicated studies). However, if the research team is willing to share their data with other researchers who have the interest, skills, and resources to conduct additional analyses, this can greatly increase the productivity of the research team that conducted the original study. This type of exchange usually involves an agreement between the data collection team and the data analysis team to clarify details about data sharing protocols and how the data should be used.

There are several publically available health-related electronic databases that can be used to address a variety of research topics. A few examples follow.

(a) The World Health Organization (WHO) Global Health Observatory Data Repository (http://apps.who.int/gho/data/?theme=main) provides statistics on an array of health-related topics for countries around the world. However, these statistics are generally at the country-level so regional or population subgroup-specific data are not usually available. Another similar source is data available on the website of the Institute of Health Metrics and Evaluation at the University of Washington in the United States (http://www.healthdata.org/). This website includes the Global Burden of Disease (GBD) estimates which quantify country-level health-related burden (i.e., cause-specific mortality and disability) from 1990 to 2010 and data visualization tools which make it possible to compare the relative importance of different health conditions (including mental disorders) between countries and between different population groups within countries (http://www.healthdata.org/gbd/data-visualizations).

(b) Established in 1962, the Inter-university Consortium for Political and Social Research (ICPSR, http://www.icpsr.umich.edu/icpsrweb/landing.jsp) is a major data source for scholars in the social sciences. Located at the University of Michigan in the United States, ICPSR is a membership-based network that includes 65,000 datasets from over 8,000 discrete studies or surveys, including a number of large-scale population surveys conducted in the United States and other countries. The website provides online analysis tools to generate simple descriptive statistics including frequencies and cross-tabulations. In addition to ASCII and .txt format, the website also provides options for downloading data in formats that are compatible with popular statistical software packages such as SAS, Stata, SPSS, and R. The website also provides technical support in data analysis and in the identification of potential data sources. In order to download data, users need to register with the system.

(c) A variety of government agencies in the United States regularly collect data on different health-related topics and post them online for free download once data cleaning is completed. For example, the United States Census Bureau (http://www.census.gov/data.html) provides basic demographic data and the Centers for Disease Control and Prevention (http://www.cdc.gov) provides access to data on cause-specific disability, mortality, and an array of health conditions including injuries and violence, alcohol use, and tobacco smoking. The Substance Abuse and Mental Health Services Administration have a range of datasets posted on their website (http://www.samhsa.gov/data/) about various mental and substance use disorders. Users interested in more information about publicly available health-related data can refer to *Secondary data sources for public health: A practical guide* by Boslaugh.[1]

3. Conducting a secondary analysis of existing data

There are two general approaches for analyzing existing data: the 'research question-driven' approach and the 'data-driven' approach. In the research question approach, researchers have an a priori hypothesis or a question in mind and then look for suitable datasets to address the question. In the data-driven approach researchers glance through variables in a particular dataset and decide what kind of questions can be answered by the available data. In practice, the two approaches are often used jointly and iteratively. Researchers typically start with a general idea about the question or hypothesis and then look for available datasets which contain the variables needed to address the research questions of interest. If they do not find datasets that contain all variables needed, they usually modify the research question(s) or the analysis plan based on the best available data.

When conducting either research question-driven or data-driven approaches to the analysis of existing data, researchers need to follow the same basic steps.

(a) There needs to be an analytic plan that includes the specific variables to be considered and the types of analyses that will be conducted. (In the research question-driven approach this is determined before the researchers look at the actual data available in the dataset; in the data-driven approach this is determined after the researchers look through the dataset.)

(b) Researchers must have a comprehensive understanding of the strengths and weaknesses of the dataset. This involves obtaining detailed descriptions of the population under study, sampling scheme and strategy, time frame of data collection, assessment tools, response levels, and quality control measures. To the extent possible, researchers need to obtain and study in detail all survey instruments, codebooks, guidebooks and any other documentation provided for users of the databases. These documents should provide sufficient information to assess the internal and external validity of the data and allow researchers to determine whether or not there are enough cases in the dataset to generate meaningful estimates about the topic(s) of interest.

(c) Before conducting the analysis, researchers need to generate operational definitions of the exposure variable(s), outcome variable(s), covariates, and confounding variables that will be considered in the analysis.

(d) The first step in the analysis is to run frequency tables and cross-tabulations of all variables that will be included in the main analysis. This provides information about the use of the coding pattern for each variable and about the profile of missing data for each variable. Due attention should be paid to skip patterns, which can result in large numbers of missing values for certain variables. In comprehensive surveys that take a long time to complete, skipping a group of questions that are not relevant for a particular respondent (i.e., 'skips') is a common method used to reduce interviewee burden and to avoid interviewee burn-out. For example, in a survey about alcohol-related problems, the survey module typically starts with questions about whether the interviewee has ever drunk alcohol. If the answer is negative, all questions about drinking behaviors and related problems are skipped because it is safe to assume that this interviewee does not have any such problems. Prior to conducting the full analysis, these types of missing values (which indicate that a particular condition is not relevant for the respondent) need to be distinguished from missing values for which the data is, in fact, missing (which indicate that the status of the individual related to the variable is unknown). Researchers should be aware of these skips in order to make a strategic judgment about the coding of these variables.

(e) Finally, the researcher should recode the original variables in order to properly handle missing values and, if necessary, to transform the distribution of the variables so that they meet the assumptions of the statistical model to be used in the intended analysis. The recoded variables should be stored in a new dataset and all syntax for the recoding of variables (and for the analysis itself) should be documented. The original dataset should NEVER be altered in any way.

(f) When using data from longitudinal surveys or when using data stored in different datasets, it is critical to check the accuracy of the identifier variable(s) to ensure that the data from different time periods or from different datasets is matched correctly when merging the datasets.

(g) For longitudinal studies, the assessment methods and the coding methods for key variables can change over time. Thus, close examination of the survey questionnaires and codebooks are essential to ensure that each variable in the combined dataset has a uniform interpretation throughout the study. This may require the creation of separate uniform variables that are constructed in different ways at different points in time throughout the study, such as the crosswalks to convert diagnostic categories between DSM-III, DSM-IV, and DSM-5.

(h) Many population-based surveys, particularly those focused on assessing the prevalence of relatively uncommon conditions such as schizophrenia, employ multi-stage sampling strategies to enrich the sample. In this case, the data set usually includes design variables for each case (including sampling weight, strata, and primary sampling unit)

that are needed to adjust the analysis of interest (such as the prevalence of a condition, odds ratios, mean differences, etc.). Researchers who conduct secondary analysis of existing data should consider the design variables used in the original study and apply these variables appropriately in their own analyses in order to generate less biased estimates.[2,3]

4. Pros and cons of the secondary analysis of existing data

4.1 Advantages

The most obvious advantage of the secondary analysis of existing data is the low cost. There is sometimes a fee required to obtain access to such datasets, but this is almost always a tiny proportion of what it would cost to conduct an original study. Also, the data posted online are usually cleaned by professional staff members who often provide detailed documentation about the data collection and data cleaning process. Moreover, teams conducting large-scale population-based surveys that are made available to others usually employ statisticians to generate ready-to-use survey weights and design variables – something that most users of the data are unable to do – so this helps users make necessary adjustments to their estimates. This is a great boon to graduate students and others who have lots of good ideas but no money to conduct the studies that could test their ideas.

Researchers who would rather spend their time testing hypotheses and thinking about different research approaches rather than collecting primary data can find a large amount of data online. The increasing availability of such data online encourages the creative use and cross-linking of information from different data sources. For example, experts in hierarchical models can combine data from individual surveys with aggregate data from different administrative levels of a community (e.g., village, township, county, province, etc.) to examine the factors associated with health-related outcomes at each level. The availability of such databases also provides statisticians with real-life data to test new statistical models. Such analyses could identify potential new interventions to existing problems that can subsequently be tested in prospective studies.

4.2 Disadvantages

Inherent to the nature of the secondary analysis of existing data, the available data are not collected to address the particular research question or to test the particular hypothesis. It is not uncommon that some important third variables were not available for the analysis. Similarly, the data may not be collected for all population subgroups of interest or for all geographic regions of interest. Another problem is that to protect the confidentiality of respondents, publicly available datasets usually delete identifying variables about respondents, variables that may be important in the intended analysis such as zip codes, the names of the primary sampling units, and the race, ethnicity, and specific age of respondents. This can create residual confounding when the omitted variables are crucial covariates to control for in the secondary analysis.

Another major limitation of the analysis of existing data is that the researchers who are analyzing the data are not usually the same individuals as those involved in the data collection process. Therefore, they are probably unaware of study-specific nuances or glitches in the data collection process that may be important to the interpretation of specific variables in the dataset. Sometimes, the amount of documentation is daunting (particularly for complex, large-scale surveys conducted by government agencies), so users may miss important details unless they are prominently presented in the documents. Succinct documentation of important information about the validity of the data (by the provider) and careful examination of all relevant documents (by the user) can mitigate this problem.

5. Government support for secondary analysis of existing data

This paper discusses several issues related to the secondary analysis of existing data. There are definitely limitations to such analyses, but the great advantage is that secondary analyses can dramatically increase the overall efficiency of the research effort and – a secondary advantage – give young researchers with good ideas but little access to research funds the opportunity to test their ideas. Recognizing the importance of making the most of high-quality research data and of rapidly translating research findings into actionable knowledge, starting in 2003 the United States National Institute of Health, the largest funding agency for biomedical research in the world, required all projects with annual direct costs of 500,000 US dollars or more to include data-sharing plans in their proposals. Moreover, NIH has released several program announcements specifically designed to promote secondary analysis of existing datasets. Other countries and some large health care providers also make registry data available to qualified researchers. These practices ensure that other researchers not involved in the studies or in the creation and maintenance of the registries will be able to use the data generated by these big projects or by the registries to test a wide range of hypotheses. Other governments (including the Chinese government), health-related non-government organizations, and other funders of biomedical research need to follow these examples. Failure to provide qualified researchers access to government-generated registry data or to government-supported research data results in a huge but unnecessary wastage of economic and intellectual resources that could be better employed to improve the health of the nation.

Conflict of interest

The authors declare no conflict of interest related to this article.

Funding

This work was supported by a grant from the China Medical Board (13-165) to HGC.

现有数据的分析：机遇与实施

程辉，费立鹏

概述：现有数据的二次分析已成为提升卫生研究机构整体效率的一种日益流行的方法。该工作取决于政府、资助机构以及研究者，取决于他们能不能让没有参与原始研究、没有参与创建和维护登记系统的其他合格研究人员获得原始研究数据或登记系统的数据。二次分析的好处是显而易见的，但面临的障碍很多。因此提高这些数据可获得性的工作进展缓慢，在低收入和中等收入国家尤为如此。本文介绍了现有数据二次分析的基本原理和概念，描述了若干个可公开获得的数据库，为现有数据的二次分析提供一般准则，并讨论了现有数据分析的优势和不足。

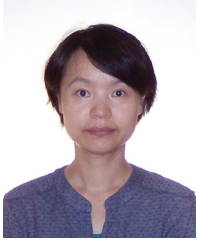
关键词：统计学数据解释，数据采集，美国国立卫生研究所

本文全文中文版从 2015 年 1 月 25 日起在 www.shanghaiarchivesofpsychiatry.org 可供免费阅览下载

References

1. Boslaugh S. *Secondary data sources for public health: A practical guide.* New York, NY: Cambridge; 2007
2. Lohr SL. *Sampling: Design and analysis (2nd Ed.).* Boston, MA: Brooks/Cole; 2010
3. Graubard BI, Korn EL. Modelling the sampling design in the analysis of health surveys. *Stat Methods Med Res.* 1996; **5**(3): 263-381

(received, 2014-11-11; accepted, 2014-12-04)



Dr. Hui Cheng is an epidemiologist by training. She is currently a post-doctoral research associate at Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine. She has published findings from studies on mental health related topics using public data. Her main interest is substance use and related problems, and public mental health.