Proceedings of Statistics Canada Symposium 2016
Growth in Statistical Information: Challenges and Benefits

A Systematic Review: Evaluating Extant Data Sources for Potential Linkage

Erin Tanenbaum, Michael Sinclair, Jennifer Hasche, Christina Park[1]

Abstract

The National Children's Study Vanguard Study was a pilot epidemiological cohort study of children and their parents. Measures were to be taken from pre-pregnancy until adulthood. The use of extant data was planned to supplement direct data collection from the respondents. Our paper outlines a strategy for cataloging and evaluating extant data sources for use with large scale longitudinal. Through our review we selected five evaluation factors to guide a researcher through available data sources including 1) relevance, 2) timeliness, 3) spatiality, 4) accessibility, and 5) accuracy.

Key Words: Extant Data; Linkage; Administrative Data; Data Evaluation.

1. Introduction

1.1 Background

The National Children's Study was designed to study environmental influences on child health and development. The planned Main Study would have surveyed 100,000 children and their parents from before birth to age 21 (NIH, 2011). Recruitment ended in July 2013 and the methods tested may benefit other longitudinal surveys. To reduce respondent burden, the NCS planned to supplement primary data collection by providing a list of data sources which could be linked to the NCS data by researchers. Use of extant data sources is not unusual, the science behind linking files to cross-sectional and longitudinal surveys continues to grow. In this paper we examine the need for data source evaluation tools. Such tools are needed to take into consideration single point in time linkages as well as linking longitudinal surveys to other longitudinal data sources.
NCS supplemental data sources covered a vast array of topics including the environment, health, education, crime, and socioeconomic factors. Although hundreds of data sources may be linked to NCS data, most future analytic research efforts would link only a few supplemental files or sources. Identifying appropriate data files may sound daunting and thus in this paper we summarize findings from a literature scan of data evaluation methods. We begin with some background on the use of extant data, and then move to describe a hierarchical structure for describing extant data including primary identifiers and evaluation criteria. Lastly, we discuss the challenges with longitudinal evaluations.

1.2 Types of Extant Data

Within the NCS, extant data sources include data about individuals, groups of people, or neighborhoods if the data was not originally collected through the NCS. As such, extant data includes various data sources including registries, administrative, "big data", and others. Although researchers may typically think of government agencies and statistical programs (Iezzoni, 1997) as primary generators of such data, an increasing number of non-statistical organizations are also producing extant data. Given that, the term extant data has come to encompass a much broader array of available

[1] Erin Tanenbaum, NORC at the University of Chicago, NICHD HHSN275201000123U, 4350 East-West Highway, 8th Floor, Bethesda, MD USA 20814 (Tanenbaum-Erin@norc.org); Michael Sinclair, Mathematica Policy Research, NICHD HHSN275201000123U, 1100 1st Street NE, 10th Floor, Washington, DC USA 20002-4221 (MSinclair@mathematica-mpr.com); Jennifer Hasche, NORC at the University of Chicago, NICHD HHSN275201000123U, 55 East Monroe Street, 20th Floor, Chicago IL USA 60603 (Hasche-Jennifer@norc.org); Christina Park, National Institute of Child Health and Human Development (NICHD), 6100 Executive Blvd., MSC 7510, Bethesda MD USA 20892-7510 (parkchris@mail.nih.gov).

data to include marketing, inventory control, exposure and damage assessment, aerial photographs, navigation databases, administrative data, and the like. Limiting the type of data source would be premature for a longitudinal and multi-faceted study like the NCS.

1.3 Rationale for Using and Evaluating Extant Data in the NCS

When using extant data, the benefits often outweigh the cons. Yet, it is still important to carefully weigh the pros and cons of supplemental data use. The NCS planned to rely heavily on extant data files for area and respondent level data. From a collection standpoint, supplemental sources are often less expensive to acquire than primary data (Iezzoni, 1997), readily available, can encompass large populations, and can reduce the burden to survey respondents. In addition, some external sources may offer the best source of information on a topic, making it irrational to collect through the primary study (e.g. census for population estimates). Extant data can also be used by statisticians to improve the coverage of sampling frames, correct biases in imputation, allow direct editing of survey items, and test accuracy or consistency of survey responses (Chappell, 2005, Bradburn, 1993, Czajka, 2003).

For researchers, extant data can fill gaps when studies cannot predict or collect all data needs. Additionally, the use of administrative health data continues to grow as a way to potentially reveal and improve personal health (Daver, 2013, Ng, 2010, Zhan, 2003). For example, a Medicare patient may know they had “bypass surgery,” but will likely be unable to report the procedures performed and associated costs whereas Medicare claims could bridge this gap revealing what a survey alone cannot. Thus, in today’s data rich environment, the question of whether a researcher should use extant data has changed to which extant data should be used and what limitations exist with said source.

1.4 Benefits of an Evaluation

While the usefulness of extant data is clear, challenges have been noted. For example, some may inappropriately link extant data, or disregard data quality concerns (Iezzoni, 1997, Jabine, 1985). In addition its use is susceptible to inadvertent misuse. For example, when using electronic health records (EHR) physicians may code with different requirements, procedures, degrees of thoroughness, or accuracy (Iezzoni, 1997) making it difficult to tell if differences in administrative records reflect true differences. Case in point, birth certificates are often limited in the number of categories for type of insurance and the categories changes by state (Martin, 2013). Additionally, extant data is often lacking documentation. Ideally, thorough and consistent documentation across years should be available to the researcher, but this is rarely the case, which makes data evaluations extremely difficult to perform (Reidy, 1998).

Also, privacy concerns continue to grow; it seems that more (not less) data fall under access restrictions hindering potential benefits (Lane, 2010). Thus, a mechanism which has been heralded as a faster and cheaper data source (The World Bank) is not without its costs – including the potential for deriving incorrect inferences (Davern, 2013) or assuming data access is available when it is not. Still, linking restricted-access sources can be extremely beneficial. For example, linking by personally identifiable information (PII), such as name, date of birth, address, and Social Security number can reduce respondent burden and increase a researcher’s knowledge, especially when linking information that the respondent does not know (e.g. diagnostic codes). Being able to evaluate extant data is thus paramount given the challenges and potential benefits related to use of extant data sources. Still, one must answer the question: does the analytic improvement outweigh the costs of linking extant data? In response, an extant data library and evaluation methodology was crafted for the NCS.

2. Assessing Extant Data

2.1 A Multifactorial Approach

We performed a literature scan of data evaluation methods, starting with well-established standards from the medical, statistical, and social survey literature. After reviewing over 80 publications, we enumerated evaluation concepts by source and compiled the concepts into categories. After an initial scan, we realized reviewing the concepts individually would not have taken into consideration nuances of the literature.

Some authors chose a structured evaluation related to type of information about the data. For example, multiple authors identified quality aspects related with 1) data sources (keeper and delivery of the data source), 2) its metadata (clarity of data definitions, etc.), and 3) the data itself (facts in the data source) (ABS, 2006). Karr, Sanil, and Banks (2006) took the concept further by ordering the three according to an increasing level of detail needed since they hypothesize that a report on data quality would be more detailed than those on metadata and source. The authors intended for this ordering to allow a user to stop searching for information if an aspect was considered irrelevant or to omit a data system from consideration if it failed at a higher level.

Based on our review, we recommend a multifactorial approach with two factors: a point-in-time hierarchy and a concept hierarchy. Such a hierarchy can reduce the level of effort if initial results about the data source reveal the it is not fit for its intended use. For example, researchers would not want to link a national vaccine rate from a survey if 97% of respondents did not reply to vaccine questions. In such an example, the researcher could stop evaluating the source and look for other files for similar information. We call our system an Appropriate Evaluation Protocol, as the researcher's specific needs are at the core of the evaluation.

2.2 Appropriate Evaluation Protocol

To anchor our extant data library, we began by creating a framework. An extant data library was created to aid researchers in identifying potential data sources which may be used as primary filters. This library may be thought of as the first tool in our proposed appropriate-use evaluation protocol. Yet, reviewing hundreds of data sources is both impractical and unreasonable for most research efforts. As such, filters were created to allow researchers to quickly filter through available sources. NCS extant data library filters included the name of the database, the data provider, topic, sub-topics, and key data elements. Additional information was also collected on each source including the provider type, contact information (website, etc.) and a list of similar databases allow a researcher to quickly review available sources before digging into the next layer of our proposed evaluation protocol: the evaluation criteria.

From the literature we reviewed evaluation criteria and intended to incorporate the criteria into our extant data library. Yet, we found many of the ideas were not actionable. For example, "accuracy" was mentioned often and yet a data user cannot measure accuracy without further instructions. As such we crafted five backbone factors which relate to 27 concepts (the questions behind the factors), and 50 actionable elements (the measure behind the concept). The elements may in turn be summarized to create a quick way to evaluate a source on a given factor or concept.

It's important to note that a broad evaluation may not be mandatory prior to using a source. Sometimes prior experience provides enough information to select an extant data file. Even so, an evaluation system provides a structured way to quickly study quality in its entirety or by factor for a specific application or potential new data source.

Our recommended library evaluation factors are similar to those in the literature (Kasprzyk, 2001) with one exception. We divided timeliness into temporal and geographical attributes. Separation is important within the NCS since environmental data sources are key and yet chemical half-lives vary greatly (Lioy, 2009) and recommended distances between pollution sources and residences also varies. By separating time and space, a source could meet geographic needs and yet be out of date or collected at an irregular interval.

Our five factors help answer the following questions:

1. Accessibility: how difficult is it to acquire and use the data?
2. Accuracy: are measures documented for being correct, precise, and/or consistently across geographies?
3. Relevance: does the data serve the analytic objectives? Does the source provide confounders or outcomes which are difficult to collect as part of the NCS survey collection system?
4. Spatial: at what level of geographic specificity is the data available?
5. Timeliness: when was data collected, and how quickly is it distributed?

After identifying 50 actionable elements, we assigned them to sequential tiers: tier one may be used by all studies as a filtering tool. Tier two contains elements only available with data access or in relation to a study. After tier one elements are collected, a researcher may score the elements based on their own needs and decide which data sources to explore further. Tables 2.2-1 and 2.2-2 include a complete list of proposed factors, components, and answer types.

Although these evaluation criteria are resourceful, simply evaluating a database by itself is not sufficient. Most of the literature stressed data evaluation depends on the intended use. For example, a study on urban populations may overlook the fact that rural coverage is limited in a data source. Since data needs will vary depending on the research question at hand, a final scoring mechanism is not part of our list of elements. While some sources proposed a scorecard, we believe that each element's information should be collected and the researcher should then determine its importance instead, not an extant data librarian. In addition, not all elements may be feasible to collect and as such filtering data sources may be necessary with limited information. In these cases, we believe the researchers will weigh the pros and cons of the missing information.

Table 2.2-1
Tier 1 Criteria across Studies

Factors	Concepts	Elements	Answers
Accessibility	Ease of data access?	Data publically available?	Yes/no.
		Available via a website?	Yes/no.
		Accessing data instructions	URL or short description.
		Restrictions on use or dissemination?	Yes/no.
		Time required to receive file from date of request.	# of days or months.
	Well-documented?	Files available including: 1) questionnaire, 2) methodology, 3) quality report, 4) crosswalk between data product vintages, and 5) codebook?	Yes/no for each; URL if available.
		Documentation appropriate (including language, symbols, units, etc.)?	Checklist for each documentation type.
	Cost-effective?	Cost to purchase data	In dollars.
	Available in standard software packages?	How are the data packaged? (SAS, ASCii, Microsoft Access, etc.).	Yes/no for each.
Accuracy	Linkage possible?	Presence of unique identifier to link.	Yes/no. Plus instructions.
	Unit nonresponse?	Percent units with all data missing.	Rate
	Precision?	Standard error, sampling error available?	Yes/no or NA (census)
	Edits appropriate?	Imputations, weighting, logic check edits, corrections published after release, etc.	Yes/no for each.
Relevance	Data use appropriate?	Data collection method appropriate?	List data methods and short explanation about its appropriateness.
	Reputation	Is the data regarded as true and credible?	Score, # of citations using the dataset. Negative points if citation finds faults with the source.
	Linkage appropriate?	Planned linkage variable appropriate?	Rating (1 to 10) against study's needs.
		How comparable are linkage unit definitions?	Written explanation of the differences.
		Linkage validated? Linked previously (other studies)?	Yes/no.
	Consistent representation?	Data definitions and methodologies stable over time?	Stability score (1 to 10) with overview of changes.
Spatial	Geographic level available?	Listed as a) public access, b) restricted access, c) not available	Answers for each geography including national, state, regional, county, track, block, household, individual, etc.
		Is data complete?	Yes/no for each level.

Factors	Concepts	Elements	Answers
	GPS data	Coordinates available or easy to calculate?	Yes/no and level
Timeliness	Time period compatibility	Frequency of data collection?	Frequency (monthly, etc.)
		What period of time was data collected?	Month/year started to planned
		Time to release from data collection.	Time period
		Sufficient temporal documentation?	Yes/no.

Table 2.2-2
Tier 2 Criteria within a Study

Factors	Concepts	Elements	Answers
Accessibility	Analysis Program Codes Provided?	Supplemental programs to import data and/or analyze data.	Yes/no by data type.
	Cost to Process Data?	Estimated costs to format data.	In dollars
Accuracy	Completeness?	The extent of missing-ness.	% missing by data element.
		Fields imputed?	% imputed, and Yes/No documentation for methods
	Precision consistent?	Examine changes in response rates, attrition, etc. over time.	Rating (1 to 10).
	Coverage?	Coverage of target population?	% coverage or a rating (1 to 10).
	Measurement error?	Statements about biases	Open-ended; rating (1 to 10).
Relevance	Consistency of Logical Structure	Tests for logical relationships.	Yes/no sequencing, skip patterns, consistency check, etc.
	Data use appropriate?	Keywords related to needed content?	Text list from codebook or source website
		Data Elements utility/usefulness?	List elements from codebooks. Rating (1 to 10) by element.
		Logic structure appropriate?	Yes/no if tests are appropriate.
		Data source appropriate for study?	Categorical (government, private, other); rating (1 to 10).
	Target Population Compatible?	Review the population included in the data in contrast to the NCS population.	Score 1 to 10. Include concepts such as group quarters, definition of a household, etc.
	Utility, Value-Added, Used by Others	Based on intended purpose. Data serves to supplement/ replace direct collection?	Research specific.
		Data used by others?	# citations.
Timeliness	Time period compatibility	Frequency meets study's needs?	Rating 1 to 10 for intended use.
		Data period meet study's needs?	Rating 1 to 10 for intended use.

Longitudinal surveys such as the NCS provided unique data linkage challenges as data sources, collection methods, and level of documentation changes over time. Unfortunately, it is often difficult to identify and locate data documentation years later and yet analyses are often funded or conceptualized years after data collection is complete. For example, imagine an analysis is planned two years after data collection is complete. At that point in time, extant data sources are evaluated and linked by a research. Yet, the team who created the extant data may change, the company supplying the data may close or merge, and documentation may go missing. Ideally, full documentation should be available to the researcher when the data is used for analysis, but this is rarely the case, and makes extant data assessments extremely difficult to perform after the fact (Reidy, George, & Lee, 1998).

3. Conclusion

We propose researchers use an appropriate evaluation protocol to help researchers to quickly identify extant data that are appropriate to answer their individual research questions. This paper provides a summary of the evaluation criteria many have used to study the appropriate use of specific extant data sources to augment program data under the assumption that such data will improve the analytical utility and quality of the study results and may have the potential to reduce data collection costs and burden. These criteria provide the means to make decisions among a competing set of extant data sources and we provide a suggested infrastructure for collecting and storing this information at multiple levels of review. We hope these methods will assist other researchers that face a similar challenge to design their own data linkage program plan.

Since an analysis is only as good as the data on which it is based, and data is only as good as the process with which it is collected, we the authors look forward to a future where data transparency meets the needs of all researchers. Until then, we will continue to research data evaluation criteria.

References

Australian Bureau of Statistics (ABS). (2006). Information Paper: Evaluation of Administrative Data Sources for Use in Quarterly Estimation of Interstate Migration, 2006 to 2011. (Cat. no. 3127.0.55.001). Canberra, Australia.

Bradburn, N.M. (1993). A Census that Mirrors America [electronic resource]: Interim Report / Panel to Evaluate Alternative Census Methods, Committee on National Statistics, Commission on Behavioral and Social Sciences and Education, National Research Council. Washington, DC: National Academy Press.

Chappell, G., Obenski, S., and Farber, J. (2005). Research to Improve Census Imputation Methods: Item Results and Conclusions. Presentation at the Joint Statistical Meetings of the American Statistical Association, Survey Research Methods Section. Minneapolis, MN, August 10.

Czajka, J.L., Jacobson, J.E., & Cody, S. (2003). Survey estimates of wealth: a comparative analysis and a review of the Survey of Income and Program Participation. Social Security Bulletin, 65(1), 63.

Davern, M., Roemer, M., and Thomas W. (2013). Investing in a Data Quality Research Program for Administrative Data Linked to Survey Data for Policy Research. Unpublished book.

Iezzoni, L. (1997). Assessing quality using administrative data. Annals of Internal Medicine, 127(8 Pt 2), 666-674.

Jabine, T.B., & Scheuren, F.J. (1985). Goals for statistical uses of administrative records: the next 10 years (with discussion). Journal of Business and Economic Statistics, 3, 380-404.

Karr, A.F., Sanil, A.P., & Banks, D.L. (2006). Data quality: a statistical perspective. Statistical Methodology, 3(2), 137. doi:10.1016/j.stamet.2005.08.005

Kasprzyk, Daniel (2001). Talk at the National Statistics Office (NSO) at the Federal Committee on Statistical Methodology (FCSM) conference. http://www.fcsm.gov/01papers/Kasprzyk.pdf Accessed July 1, 2013.

Lane, J., & Schur, C. (2010). Balancing access to health data and privacy: a review of the issues and approaches for the future. Health Services Research, 45(5 Pt 2), 1456-1467. doi:10.1111/j.1475-6773.2010.01141.x

Lioy, P., Isukapalli, S., Trasande, L., Thorpe, L., Dellarco, M., Weisel, C., & ... Landrigan, P. (2009). Using national and local extant data to characterize environmental exposures in the National Children's Study: Queens County, New York. Environmental Health Perspectives, 117(10), 1494-1504. doi:10.1289/ehp.0900623

Martin, J. A., Wilson, E. C., Osterman, M. J., Saadi, E. W., Sutton, S. R., & Hamilton, B. E. (2013). Assessing the quality of medical and health data from the 2003 birth certificate revision: results from two states. National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System, 62(2), 1-19.

National Institutes of Health (NIH), The National Children's Study: It's All About Our Children. NIH MedlinePlus: the magazine [Internet]. 2011 Summer;6(2):4-5. Available from: https://www.nlm.nih.gov/medlineplus/magazine/issues/summer11/articles/summer11pg4-5.html

Ng, C. (2010). Population and Administrative Datasets for Research and Evaluation. Presentation for Fraser Health.

Reidy, M., George, R., & Lee, B.J. (1998). Developing an integrated administrative database. Exploring Research Methods in Social Policy Research. Asldershot, UK: Ashgate Publishing Company.

The World Bank, What Happens When Big Data Meets Official Statistics? - Live Webcast. Speakers include Robert Groves, Provost, Georgetown University; formerly Director, U.S. Census Bureau. Accessed from: http://live.worldbank.org/what-happens-when-big-data-meets-official-statistics-live-webcast

Zhan, C., & Miller, M. (2003). Administrative data-based patient safety research: a critical review. Quality & Safety in Health Care, 12(Suppl 2), ii58-ii63.