American Journal of
Biomedical Science & Research

www.biomedgrid.com
ISSN: 2642-1747

Mini Review

Record Linkage and Matching in Survey Research

Kyu-Seong Kim*

Department of Statistics, University of Seoul, South Korea

***Corresponding author:** Kyu-Seong Kim, Professor of Department of Statistics, University of Seoul, South Korea.

To Cite This Article: *Kyu-Seong Kim. Record Linkage and Matching in Survey Research. Am J Biomed Sci & Res. 2023 18(4) AJBSR.MS.ID.002482,* DOI: 10.34297/AJBSR.2023.18.002482

Received: April 12, 2023; **Published:** April 14, 2023

Introduction

Sample surveys are usually conducted through a well pre-determined survey design, a questionnaire designs and a tightly controlled data collection process. And survey data are collected through a single survey in survey research. A high level of quality of survey data collected in this manner is an advantage, while some practical constraints imposed on the survey are limitations. In essence, the sample survey has survey time and survey cost constraints. And recently, the survey response rate and the coverage rate of the target population are on the decline and the response burden of respondents is on the rise.

These challenges potentially threaten the quality of the collected data and reduce the confidence of the conclusion derived from survey data [1]. In this situation, if there is a method to combine one survey data with other existing data to obtain more information, then this method will be a very attractive method for researchers. Linkage of records by the same entity (e.g. individuals, businesses etc.) at the micro-level across multiple data offers great opportunities for researchers [2]. In a particular survey data may be linked to other survey data, administrative data set or existing organic data in survey research. The key element when linking several data is identifier for connection. If there is a unique identifier or a unique set of identifiers for the record belong to several data, then several data will be linked. This method of linkage is called the record linkage. In the case of sample survey data, the unique identifier usually does not exist, so the method of record linkage is not appropriate. Instead, records can be linked using multiple, even if not unique, somewhat loose identifiers. At this time, the linked record can be seen as a connection by a similar identifier rather than a connection by the unique identifier. In other words, it is a method of linking records with the same statistical properties of several identifiers. This method of connection is called statistical matching. The primary goal of record linkage is to link the records that contain the same entity, but the primary goal of statistical matching is to link similar records [3].

Record Linkage

The idea of record linkage dates back to Dunn's paper titled 'Record Linkage' in 1946 [4]. Dunn named accumulating important events in life centered on the same person as record linkage. The method of record linkage can be broadly classified into deterministic and probabilistic. In the deterministic record linkage, records are uniquely linked according to pre-determined rules, so the only identifier must be available in almost all records.

When the identifier of the entity is not unique, probabilistic linkage using auxiliary identifiers is available [5]. This method has the framework of mathematical theory in Fellegi & Sunter's pioneering paper, where a match weight is calculated indicating the possibility that the two records are connected [6]. Later, their theory provides a theoretical foundation for record linkage applications in various fields. From the late 1990s, various advanced machine learning techniques began to be developed for record linkage. And some machine learning and neural network algorithms often showed higher accuracy [7]. Any kind of record linkage can lead to associated linkage error, so a potential record linkage error after a process of record linkage should be established. The quality of the record linkage will be evaluated as a linkage error and the magnitude of these errors [8,9]. Two important kinds of linkage error are false links where linked records have different individuals and missed links where records of the same individuals are not connected [8]. A linkage error rate may be calculated as the proportion of the sum of false and missed linked records among all records.

Statistical Matching

Basically, statistical matching presupposes a situation in which two types of data are given. The target data may be sample data,

administrative data, commercial data etc. Two types of data are given common variables that can identify entity of the record such as name, age, address etc. Statistical matching may be used when entity's unique identifiers are insufficient. The purpose of statistical matching is to create a distribution of characteristics of entities rather than exact record connection of entities. To integrate the two data through statistical matching, the definitions and concepts used in the two different data should be consistent. The details to be checked are as follows: harmonization of the definition of records, harmonization of reference periods, completion of populations, harmonization of survey variables, harmonization of classifications, adjustment for missing data etc. [10]. After matching the definitions and concepts used in two different data, the work of data matching should be performed.

Statistical matching is divided into macro matching and micro matching according to purpose of matching. Macro matching is a method used to estimate the joint probability function of linked characteristic variables or the important parameters of the estimated probability distribution. On the other hand, micro matching is a method to replace record-level values missed in one data with appropriately predicted value or with value from other data and then generate a combined micro data [10]. To evaluate statistical matching that links two or more different data to generate new micro-data, the following shall be considered: assumptions assigned to the joint probability model, estimators based on the joint probability function (macro matching), creation of an appropriate substitution for missing value (micro matching), the inference process in a linked data etc. [11].

Linkage errors are inevitable in linked data due to the lack of a unique identifier. Small linkage errors can have a substantial impact on statistical inference [3]. Statistical issues that arise newly due to data linkage include the potential bias and an alternative statistical framework in the basis of linked data. In addition to the usual survey bias, linked data are subject to potential biases such as consent bias, missed links bias, incorrect links bias etc. [12-14]. The key is how to deal with such potential biases that have been raised. Also, a new statistical framework for secondary analysis of linked data is required. Newly developed statistical theory using linked data is actively being presented [3,14].

Discussion

Data linkage can be viewed as an alternative method for conducting surveys, and there are challenges to be solved. Above all, public trust in record-linked data should be gained and maintained. Then the cost and time required for data linkage should be optimized, and the efficiency of linkage methods should be maximized [15]. Additionally, if the possibility of identity exposure of respondents increases due to data linkage, it is difficult to provide linked data to external researchers. Therefore, statistical disclosure control techniques should be applied to linked data to reduce the exposure risk [16]. The importance of linking data increases as a method of collecting data, then methods of improving the linkage quality will be developed and software that performs linking data will become widely available [14]. This direction is expected to continue for the time being.

References

1. Sakshaug JW, Couper MP, Ofstedal MB, Weir DR (2012) Linking survey and administrative records: mechanisms of consent. Social Methods Res 41: 535-569.
2. Schnell R (2013) Linking surveys and administrative data. Working paper series WP-GRLC-2013-03 in German Record Linkage Center.
3. Han Y, Lahiri P (2019) Statistical analysis with linked data. International Statistical Review 87: S139-S157.
4. Dunn HL (1946) Record linkage. American Journal of Public Health 36(12): 1412-1416.
5. Newcombe HB, Kennedy JM, Axford SJ, James AP (1959) Automatic linkage of vital records. Science 130: 954-959.
6. Fellegi IP, Sunter AB (1969) A theory of record linkage. Journal of the American Statistical Association 17: 17-35.
7. Wilson DR (2011) Beyond probabilistic record linkage: using neural networks and complex features to improve genealogical record linkage. Proceedings of international joint conference on neural networks, San Jose, California: 9-14.
8. Kvalsvig A, Gibb S, Teng A (2019) Linage error and linkage bias: a guide for IDI users.
9. Doidge J, Christen P and Harron K (2020) Quality assessment in data linkage. Working paper of the UK Government Analysis Function and Office for national Statistics.
10. D Orazio M, Di Zio M, Scanu M (2006) Statistical Matching. Wiley.
11. Rodgers WL (1984) An evaluation of statistical matching. Journal of Business and Economic Statistics 2: 91-102.
12. Harron K (2016) Introduction to data linkage. Administrative Data Research Network.
13. Moore JC, Smith PWF, Durrant GB (2018) Correaltes of record linkage and estimating risks of non-linkage biases in business data sets. Journal of Royall Statistical Society series A181: 1211-1230.
14. Chambers R, Da Silva AD (2020) Improved secondary analysis of linked data: a framework and an illustration. Journal of Royal Statistical Society Series A183: 37-59.
15. Harron K (2022). Data linkage in medical research. BMJMED 1(1): e2000087.
16. Skinner C (2009) Record linkage, correct match probabilities and disclosure risk assessment. in Insights on Data Integration Methodologies: 11-23.