Working Paper 2015-15

A study on the micro data linkages to promote the production and utilization of health and welfare statistics

Miae Oh, Hyunsoo Choi

KIHASA Korea Institute for Health and Social Affairs

A study on the micro data linkages to promote the production and utilization of health and welfare statistics

Miae Oh, Associate Research Fellow

Korea Institute for Health and Social Affairs
Building D, 370 Sicheong-daero, Sejong city 30147 KOREA
http://www.kihasa.re.kr
ISBN: 978-89-6827-313-1 93510

Contents

List of Tables

List of Figures

1 Introduction

1 Introduction

Government 3.0 is the new administrative paradigm by which public information is actively shared and barriers between government agencies are eliminated in order to promote effective communication and cooperation, which will enable personalized government services, create employment opportunities, and support a creative economy[1]. The impetus for the initiation of Government 3.0 is big data, which emerged along with other changes in the information technology (IT) environment. Big data is a broad term for data sets so large and complex that traditional data processing applications are inadequate for analyzing them. Big data can also refer to a set of techniques and technologies that are required in order to uncover and analyze the hidden values in such large and diverse datasets[2]. The fact that the current administration is attempting to create value by incorporating existing diverse data is closely associated with the emergence of big data.

The three elements that comprise big data are data volume, velocity (speed of data in and out), and data variety. Data va-

1) For additional information about Government 3.0, see the Government 3.0 website (http://www.gov30.go.kr/).
2) Wikipedia.

riety, in particular, is the most important component for creating a fourth element of big data: new value. The core of big data analysis is to create value by linking a variety of data and predicting the future.

The most crucial technology required for creating value from big data is data linkage. Data linkage, or data matching, is one of the main methods available for addressing the limited amount of information that can be collected from a single data set. Data matching can be defined as a process that combines multiple data sets in order to create a single complete data set that provides richer information. As such, data matching is also interchangeable with record linkage and data integration.

Collected and compiled databases commonly contain items such as household or personal identification number, gender, and age. When data from multiple sources are combined through these common items, sometimes additional data from these same entities or entities that share similar values for the common variable can be collected, increasing the value of the data set. Linking existing data is meaningful, because it increases the utility of the existing data and is typically faster and less expensive than additional surveying.

The need has arisen for a study that examines how best to link data in order to better utilize the National Statistics Microdata integrated Service led by Statistics Korea.

Currently, Statistics Korea is working on mid- to long-term projects due by 2020 regarding data integration methods. This topic goes hand in hand with the government's plan for secondary data creation and expansion, which is designed to promote the utilization of linked data as part of an effort to advance national statistics.

Unfortunately, integrating various data sets can have unwanted side effects, such as the disclosure of personal information. The United States' PRISM data collection program incident in 2013 clearly demonstrates such risk. Korea's own credit card information incident in 2014 also illustrates the magnitude of the effect that leakage of personal information can have in today's society. Therefore, in data linkage, personal information must be protected, and an optimal balance must be struck between protection of personal information and value creation through data linkage and data utilization.

This paper aims to define the concept of data linkage, classify the types of data linkage, and investigate domestic and international data utilization cases—as well as various matching methods—through a theoretical lens. Additionally, by linking and analyzing the Korea Welfare Panel Survey and the National Survey of Tax and Benefit based on the results of a statistical matching method simulation, we hope to make use of the foundational data required for designing policy.

Linking and utilizing diverse survey data from various government and public agencies are tasks required for Government 3.0. Furthermore, the efforts toward achieving Government 3.0, in which value is created by linking and analyzing stored data, are expected to increase our nation's global competitiveness.

2 Concept and types of data linkage

1. Concept definition
2. Classification of data by characteristics
3. Data linkage methodology

2 Concept and types of data linkage

1. Concept definition

Data linkage, or data matching, is the connection of micro-data from multiple sources to create a completely integrated data set that provides richer information. Data matching is also known as record resolution, entity resolution, object identification, field matching, data integration, and data fusion[3].

This study classifies data linkage into two types: exact matching and statistical matching.

Exact matching is the integration of data about the same entity (individual, family, event, place, etc.). Statistical matching is the integration of data about similar entities (individuals, families, events, places, etc.).

Because data about the same entity are connected through a unique identifier, the exact matching method can be used to link administrative data from various sources or to link administrative data with survey data. In contrast, because statistical matching connects data about similar entities, for which no unique identifier is available, it is used to link survey data or to

3) Christen, P.(2012). Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer.

link administrative data with survey data (in cases in which accurate personal information cannot be obtained from the survey data).

The data linkage process[4] used for both exact matching and statistical matching is described below.

- Develop clearly defined objectives
- Address legal, policy, privacy, and security issues
- Establish relationships with data providers and data users
- Gain a thorough understanding of data sources
- Decide how you will do the matching
- Define and build information technology (IT) data storage and processing requirements
- Obtain the source data
- Carry out the matching
- Validate the matching and provide quality measures
- Consider provision of access to microdata and confidentiality of published outputs
- Carry out the analysis and disseminate results

Data linked according to this process offers many benefits, which are directly related to the need for data linkage. First, data linkage provides additional information. It adds additional items (columns) to the existing data set, which may enable complex problems to be solved. Second, it enables longitudinal analysis through personal history. Time-series data can also be

4) Statistics New Zealand (2006) "Data Integration Manual"

compiled through this mechanism, by matching different types of data created at various points in time. Third, data linkage can be used to test the accuracy and reliability of survey data and administrative data, by comparing those data sets to each other. Fourth, data linkage can provide information about missing values in survey data. A variety of linkage techniques can do this. Finally, data linkage can reduce the cost associated with conducting surveys and the potential burden on survey respondents. More accurate information can be obtained by linking administrative data and survey data than by relying solely on survey respondents' responses. Thus, data linkage reduces the time and cost associated with ensuring data accuracy.

2. Classification of data by characteristics

The data to be linked can be classified as either administrative data or survey data. Administrative data is contained in databases compiled and managed by public agencies for the purpose of providing official public services. In contrast, survey data, or microdata, is collected from individual entities, through a variety of surveying methods.

Matched files are created by joining recipient files and donor files from which additional information is to be extracted. There are four possible combinations of recipient files and do-

nor files in this process; linking administrative data with other administrative data, linking administrative data with survey data, linking survey data with administrative data, and linking survey data with other survey data.

In all four cases, if unique identifiers are present in both data sets, exact matching is possible. If the unique identifier is missing from either of the data sets, the data sets can be joined via a statistical matching method, which connects similar-but not necessarily identical-entities. Therefore, based on the four combinations of data sets and two types of matching techniques, eight types of linkage are available.

〈Table 1〉 8 Types of Data matching

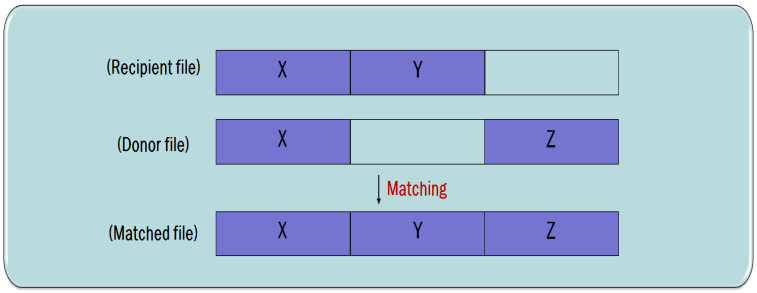
	Data characteristic	
	Recipient file	Donor file
Exact Matching	Administrative data	Administrative data
	Administrative data	Survey data
	Survey data	Administrative data
	Survey data	Survey data
Statistical Matching	Administrative data	Administrative data
	Administrative data	Survey data
	Survey data	Administrative data
	Survey data	Survey data

3. Data linkage methodology

Data matching joins disparate data files to create a single integrated data file. In the example depicted in Figure 1, there are two files, A and B. File A is composed of the variables X and Y, and file B is composed of the variables X and Z. Assuming that we are linking file B to file A, file A is the host file, or recipient file. A recipient file is defined as a "base file" in this paper. File B is used as a source of additional information to be extracted in the course of the data integration, and it is called a donor file. In this paper, it is termed a "linked file,"because it is the file to be linked. Variable X, which is found in both file A and file B, is a "common variable." Variable Y, which is found only in file A, and variable Z, which is found only in file B, are referred to as "unique variables."

In typical data matching, based on a common variable, is that variable Z from file B is added to file A. The common variable used to merge the two files is referred to as the "matching variable." The resulting file, integrated via the matching variable, is referred to as an "integrated file." In this paper, it is termed the "matched file."

[Figure 1] Data matching



Because exact matching joins data based on unique identifiers, it is typically free of uncertainty. Unique identifiers can be values of a single variable, such as social security numbers, or a combination of variables, such as name, date of birth, gender, and address. In cases in which the unique identifier is such a combination, data pertaining to the same entity may have been entered differently in the two data files due to changes in legal name, changes in address, input errors, disparity of when the data was created (age and educational attainment), and missing items.

In order to minimize such uncertainties and create reliable integrated data sets, in case of errors in common variables, probabilistic matching may be performed by assigning a weighted value for each identifier.

We collectively refer to the individual-level values a “record,” and the errors in matching records from two disparate data sets can be classified as type 1 error and type 2 error.

〈Table 2〉 Data Linkage Errors in Exacting Matching

Linkage \ Actual	True Match	True Non-Match
Link	True Positive	False Positive Link (type 1 error)
Non-link	False Negative Link (type 2 error)	True Negative

When the records from the two disparate data sets are matched via exact matching, they correspond to the same entity (an individual, a business, an event, etc.). When a matched record pair indeed originated from a single entity, it is a true positive. Likewise, a non-matched record pair that indeed originated from two different entities is a true negative. Errors that can occur during data linkage can be classified as type 1 error or type 2 error. A type 1 error occurs when a non-matched pair is falsely identified as a match (false positive). A type 2 error occurs when a matched pair is falsely identified as a non-match (false negative). The final process of data matching is performance measurement, and the type 1 error and type 2 error are computed during this process of assessing the quality of the matched data set.

The process of linking two disparate data begins with data standardization. This is a crucial preprocessing step, which affects the quality of the matched data set. This step, which includes data cleansing, is a process in which data are normalized into a consistent format and data expression and encoding are

streamlined. The next step is blocking, which reduces the number of record pairs for comparison when the data sets are particularly large. In blocking, the data set is divided into several blocks based on a variable or a combination of two or more variables. Records that have identical values for a combination of variables are grouped into the same block, and potential record pairs are compared via various methods. The next step is decision model determines whether each record pair is a match, a non-match, or a possible match. When a record pair is a match and the records indeed originated from a single entity, the linkage is accurate. However, if the records originated from two different entities, a type 1 error has occurred. Similarly, if a record pair is a non-match, but the records indeed originated from a single entity, a type 2 error has occurred. Possible matches are sent back to the blocking step to repeat this process. The quality of the linked data is indicated by the number of type 1 errors and type 2 errors. The decision model's performance is important to minimizing the number of errors.

When two or more data sets are to be matched and unique identifiers are not present, statistical matching is used. The basic assumptions required for statistical matching are as follows.

First, the data sets to be linked are obtained from the same population. According to Rässler (2002)[5]'s matching rules, the

5) Rässler R. (2002). Statistical Matching : a Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches, Springer.

averages of the common variables in the two data sets must not differ significantly. Second, given a common variable X(e.g., age), the two data sets' unique variables Y (e.g., income) and Z (e.g., debt) must be conditionally independent. This assumption is required in order to estimate the following joint probability distribution, $\hat{f}_{X,Y,Z}(x,y,z)$, when linking the base file and the donor file.

$$\begin{aligned}\hat{f}_{X,Y,Z}(x,y,z) &= f_{Z,X}(z,x) f_{Y|X}(y|x) \\ &= f_{Z|X}(z|x) f_X(x) f_{Y|X}(y|x) \\ &= f_{Z|X}(z|x) f_{Y,X}(y,x)\end{aligned}$$

The conditional matching distribution is expressed by the following equation.

$$\hat{f}_{Z,Y|X}(z,y|x) = f_{Z|X}(z|x) f_{Y|X}(y|x).$$

Classical statistical matching methods researched during the 1980s include unconstrained matching and constrained matching (Rodgers,1984)[6]. Unconstrained matching imposes no restrictions in linking two data sets. As such, the mean and standard deviation of Z^B vary in the linked files. Unconstrained matching has the advantage of finding a value closest to the

6) Rodgers, W.L. (1984) An evaluation of statistical matching. Journal of Business and Economic Statistics 2, 91-102.

value in a record of file A by allowing duplicate values in file B. However, it has the disadvantage of increasing the variance of estimates for variable Z. Despite such a disadvantage, unconstrained matching is widely used.

Constrained matching uses all of file B's records in the data linkage process, and the weighted value of the linked file has the following restrictions.

$$\sum_{j=1}^{m} w_{ij} = w_i (\text{for } i = 1, \cdots, n), \quad \sum_{i=1}^{n} w_{ij} = w_j (\text{for } j = 1, \cdots, m)$$

w_i represents file A(base file), w_j represents the weighted value of file B (donor file), and w_{ij} represents the adjusted weighted value of the matched file created by linking the two files.

The constrained matching method defines constraints based on the microdata's weighted value. Because the microdata to be linked are obtained from the same population, the implication is that the sum of the weighted values in each data set are equal (Paass 1985)[7]. The distance function (d_{ij}) and objective function used as constraints in the constrained matching method. In order to minimize the objective function, the weighted value w_{ij} is readjusted in the matched file (Barr and Tuner 1981)[8].

---

7) Paass, G. (1985) Statistical record linkage methodology: state of the art and future prospects. In Bulletin of the International Statistical Institute, Proceedings of the 45th Session, Vol. LI, Book 2. Voorburg, Netherlands: ISI.

8) Barr, R.S. and Turner, J.S. (1981). Microdata file merging through large-scale network technology. Mathematical Programming Study, Volume 15, pp. 1-22.

$$\min\left\{\sum_{i=1}^{n}\sum_{j=1}^{m}(d_{ij}\times w_{ij})\right\}$$

Constrained matching has the advantage that all of the records in file B are used and that the Z^B distribution in the matched file is the same as that in file B. However, the closest value in file B to a record in file A may not be linked, and computation time is increased in order to minimize the objective function associated with the constraints.

3 The cases of data linkage in the health and welfare sector

1. The case of exact matching
2. The case of statistical matching

3 The cases of data linkage in the health and welfare sector

Depending on the matching method (exact matching or statistical matching), data characteristics (administrative data or survey data), and data type (base file or donor file), eight types of data matching are available. In this section, we examine a case of exact matching currently utilized in the health and welfare sector, as well as a case in which the potential for utilizing statistical matching has been analyzed.

1. The case of exact matching

A. The adjustment of the eligibility rule of the Basic Pension (linking two administrative data sets)

The basic pension policy adopted in 2008 in order to address the blind spots in old-age income security was reformed to create the basic pension system in July 2014. The goal of the reform was to provide benefits (a maximum of KRW 200,000 per month per person) to 70% of all elderly over 65 years of age. In order to determine which individuals were eligible to receive pension benefits, the adjustment of the eligibility rule was set, in advance, as total recognizable elderly income being in the bottom

70% of all elderly (as of 2014, the 70% threshold was KRW 870,000 for a single elderly person and KRW 1,392,000 for an elderly couple). Acknowledged income was defined as the sum of appraised income and asset-conversion income.

In order to deduce the eligibility rule that satisfied the condition of acknowledged income being in the bottom 70%, while considering the basic pension policy's goal coverage and its policy characteristics, it was important to construct a database containing total acknowledged income for the elderly individual and his or her spouse and administrative data by asset item. Exact matching was used to join the two administrative data sets.

For the adjustment of the eligibility rule in 2014, a database of total elderly and spouses was constructed from the following two administrative data sets.

The base file contained the National Health Insurance's itemized income and asset administrative data collected from approximately 7.36 million Korean adults 65 years of age and older and their spouses, who were eligible for basic pension benefits in 2014. The donor file contained the social security information system's (Haengbok-eum) itemized income and asset administrative data collected from approximately 4.38 million elderly Koreans and their spouses. This donor file was linked to the base file in order to compile in one database all of the information required to calculate acknowledged income.

The two types of administrative data were extracted at an identical point in time and linked based on individuals' unique identifier in order to create a database (matched file) containing all necessary information about the adjustment of the eligibility rule.

Next, the acknowledged income of each elderly household (the elderly individual and his or her spouse) was analyzed to estimate the distribution of the acknowledged income of all elderly households. The criteria of acknowledged income and each household's eligibility for basic pension benefits was then determined based on the results of various policy simulation, such as adjusting the method of calculating the acknowledged income.

B. Korea Health Panel and the National Health Insurance Service's database (linking a survey data set and an administrative data set)

The Korea Health Panel began in 2008 as a consortium of the Korea Institute for Health and Social Affairs and the National Health Insurance Service. The Korea Health Panel enables a comprehensive review of the factors that influence individuals' healthcare service utilization behavior (socioeconomic factors, health awareness, health behavior, etc.), healthcare expenditures, and sources of funding for healthcare. In designing

the survey, unique identifiers were collected to link the National Health Insurance Service's administrative database. The content of the survey was updated to provide an accurate history of healthcare service utilization and the health insurance because the National Health Insurance Service's administrative data was linked to the survey data.

The Korean Health Panel is a database that tracks the utilization of healthcare services during the previous year, and it includes data pertaining to non-benefit item. On the other hand, the National Health Insurance Service's administrative database is focused on insurance benefits information for the patients who use healthcare services. As such, it provides accurate information regarding each patient's history of utilizing healthcare services but does not provide information pertaining to non-benefit item. For this reason, a data matching project between the Korea Health Panel and the National Health Insurance Service began in 2013 in an effort to address the inaccuracy of the Korea Health Panel's recollection-based survey data pertaining to insurance benefits. In this case, the Korea Health Panel data set was the base file. Because the National Health Insurance Service's database was used to reinforce the accuracy of the Panel's data, this merging is an example of exact matching, in which the Korea Health Panel's survey data (base file) were linked with the National Health Insurance Service's administrative data (donor file).

2. The case of statistical matching

The case of statistical matching for which we consider the potential of data matching is the merging of the Korea Welfare Panel Survey and the National Survey of Tax and Benefit. The goal of creating matched data by linking the two data sets is to compare taxpayer and welfare recipient attitudes toward welfare, in order to ultimately provide information useful in making effective policy decisions with regard to expanding welfare and raising taxes.

The Korea Welfare Panel Survey provides a diverse information pertaining to welfare benefits and attitudes toward welfare. The National Survey of Tax and Benefit reflects the Korea national tax service's year-end tax adjustment and thus contains accurate information regarding individual's tax payments.

Variables common to the two data sets include demographic characteristics, employment status, marital status, educational attainment, personal income, and household income based on equivalence scale. Of these, the matching variables used for merging are demographic characteristics, employment status, personal income, and household income based on equivalence scale. The Korea Welfare Panel Survey's unique variables are welfare status and attitude toward welfare. These variables are matched with the individual's tax payment information of the

National Survey of Tax and Benefit. The analytical results of linking the two data sets, as well as the method used, major variables, etc., are described in more detail in Chapter 4.

4 Simulation study of data linkage via statistical matching

1. Statistical matching method
2. Outline of the simulation study
3. Data linkage and analysis : Korea Welfare Panel Survey & National Survey of Tax and Benefit

4 Simulation study of data linkage via statistical matching

1. Statistical matching method

In this chapter, for a variety of distance functions, we study the performance of a random hot deck and of a nearest neighbor distance hot deck using R program's StatMatch package(D'Orazio, 2012). The common variables, those that are used for the matching, are $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{in})$, and the distance functions considered in the study and defined as $d(i,j)$ are as follows.

- Manhattan distance function

$$d(i,j) = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$$

- Mahalanobis distance function

$$d(i,j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)}$$

$\boldsymbol{S}$: the estimate of the covariance matrix

- Exact distance function

$$d(i,j) = \sum_{k=1}^{n} D_{(ij)k},$$

$$D_{(ij)k} = \begin{cases} 1, & x_{ik} = x_{jk} \\ 0, & x_{ik} \neq x_{jk} \end{cases},$$

In order to compare the statistical matching techniques, the data were taken from the 2012 Korea Health Panel. Excluding missing values and no responses, 7,646 observations were available. Of these, 3,500 were randomly extracted to comprise a base file, and the remaining 4,146 were compiled into a donor file. Because one data set was divided to create the base file and donor file, individuals did not exist in both files and thus exact matches did not exist, but the two files contained the same variables. In the experiment, the base file's values for variable Z were hidden, and the files were linked with the donor file via various statistical matching techniques. In order to test the reliability of the matched file, the base file's values for variable Z and the donor file's values for variable Z (the values used in the matched file) were compared.

The variables most closely related to variables Y and Z were selected as matching variables. Cramer's V was used as a measure of categorical variable analysis. In the base file, the unique variables used as variable Y were health-related variables, including private pensions, private life insurance, and chronic disease status. In the donor file, the unique variables used as variable Z were economic information, including economic activity participation status and annual personal income. The matching variables, used as variable X because they are most closely related to Y and Z, included gender, age, and educational attain-

ment, for which the value of Cramer's V was large.

Depending on the statistical matching technique used, the matched file's data values may vary. In order to examine how much the Z values differed between the base file and the donor file, matched files were constructed by use of six different techniques, and the resulting Z values were compared. The six techniques were the three distance functions combined with either the nearest neighbor distance hot deck or the random hot deck. These techniques were applied to 100 data sets, and the average prediction errors for income and chronic disease status were calculated. The prediction error is given by the following equation.

$$\sum_{i=1}^{N_{rec.A}} (rec.A_i - don.B_i)^2 / N_{rec.A}$$

〈Table 3〉 Comparison of statistical matching methods

Statistical matching method	distance function	Prediction Error (standard deviation)	
		log transformed Income	The presence or absence of chronic diseases
Nearest neighbor distance hot deck	Manhattan	1.625(0.065)	0.387(0.007)
	Mahalanobis	1.627(0.063)	0.386(0.006)
	Exact	1.627(0.072)	0.387(0.007)
Random hot deck	Manhattan	1.594(0.096)	0.385(0.008)
	Mahalanobis	1.591(0.106)	0.385(0.009)
	Exact	1.631(0.106)	0.393(0.007)

Of these, the random hot deck using the Manhattan distance function and the random hot deck using the Mahalanobis distance function performed the best. The same six techniques were applied to the Korea Welfare Panel Survey, which is to be used in the actual analysis, in order to compare performance. The results of the comparison were almost identical to those involving the Korea Health Panel. Given that, the random hot deck method using Mahalanobis distance function was used in this study in order to determine donor file groups that were closest to the base file's entities and from which one was then randomly selected.

2. Outline of the simulation study

In the simulation study of data linkage in the following section of this chapter, the Korea Welfare Panel Survey is a survey that has been conducted jointly by the Korea Institute for Health and Social Affairs and Seoul National University's Social Welfare Research Center since 2006. The panel is composed of 7,000 households; for sample extraction, 50% of the households were classified as low-income households earning less than 60% of median income. The Korea Welfare Panel Survey provides various data about respondents with regard to changes in socio-economic characteristics, welfare recipient status, desire for welfare benefits, and attitudes toward welfare. Its purpose is to

provide information enabling the analysis of the dynamic aspects and varying needs of people over the cycle of their life, with the aim of enhancing flexibility and reactivity in policy.

The National Survey of Tax and Benefit, established and conducted by the Korea Institute of Public Finance, was designed to provide data in support of the nation's policy analysis and research; since 2008, approximately 5,000 households comprise the panel. For the National Survey of Tax and Benefit's sample, 300 high-income (in the top 10% of income) and low-income (within the near-poverty group) households were extracted. The two data sets did not share the same exact distribution. However, they belonged to the same population, Population and Housing Census.

In this study, we construct the matched data combined with Korea Welfare Panel Survey and National Survey of Tax and Benefit by using statistical matching method. In this matched database, the distributions of common variables before and after the data integration are compared and analyzed, and groups are distinguish according to taxpayer and welfare status based on the representative variables from each survey. In addition, by comparing and analyzing attitudes toward welfare (e.g., support for the expansion of welfare, perceptions of the middle class's tax burden, opinions regarding the need for an increase in the tax rate), we aim to provide the empirical and advisable evidence for tax and welfare policy

3. Data linkage and analysis : Korea Welfare Panel Survey & National Survey of Tax and Benefit[9)]

A. Comparison of the distributions of common variables before the matching

Variables common to the Korea Welfare Panel survey(KoWePS) and the National Survey of Tax and Benefit(NaSTaB) include gender, educational attainment, age, economic activity participation status, personal income, and family income based on equivalence scale. With regard to gender, the proportion of males in the National Survey of Tax and Benefit was far greater than that in the Korea Welfare Panel Survey. With regard to educational attainment, the proportion of well-educated people in the National Survey of Tax and Benefit was greater than that in the Korea Welfare Panel Survey. With regard to age distribution, the proportion of individuals aged 60 years or older in the Korea Welfare Panel Survey was greater than that in the National Survey of Tax and Benefit. Given that the proportion of unemployed individuals in the Korea Welfare Panel Survey was greater than that in the National Survey of Tax and Benefit, it can be assumed that there were more elderly females in the Korea Welfare Panel Survey than in the National Survey of Tax

9) Choi, H. S., Oh, M. A. (2015). The comparative analysis of the welfare attitude in accordance with the dual status of tazpayer and welfare recipient by using data linkage, Journal of the Korean Data Analysus Society, 17(4), 1983-1994 (in Korean)

and Benefit. On the other hand, the distributions—including means and medians—of personal income and family income did not vary much between the Korea Welfare Panel Survey and the National Survey of Tax and Benefit. In order to compare the distribution of each variable in the Korea Welfare Panel Survey and the National Survey of Tax and Benefit, a Chi-squared test was performed on categorical variables, including gender, age, educational attainment, and employment status, and a Wilcoxon rank-sum test was performed on continuous variables, including personal income and household income based on equivalence scale. The results are shown in Table 4.

〈Table 4〉 Socio-demographic characteristics before matching

Common variable	Categories	KoWePS (N=4185)	NaSTaB (N=7550)	Test	
Sex	men	1793(42.8%)	4442(58.8%)	χ^2= 275.8	*
	women	2392(57.2%)	3108(41.2%)		
Age	20~39	1063(25.4%)	2010(26.6%)		
	40~49	728(17.4%)	1905(25.2%)		
	50~59	702(16.8%)	1439(19.1%)		
	60+	1692(40.4%)	2196(29.1%)		
Education	Uneducated	405(9.7%)	401(5.4%)	χ^2= 187.5	*
	Elementary school	891(21.3%)	956(12.7%)		
	Middle school	522(12.5%)	723(9.6%)		
	High school	1130(27.0%)	2420(32.2%)		
	College and over	1237(29.5%)	3009(40.1%)		
Employment Status	Regular wage workers	793(18.9%)	2938(39.1%)	χ^2= 324.4	*
	Temporary wage workers	481(11.5%)	569(7.6%)		
	Daily wage workers	258(6.2%)	412(5.5%)		
	Self-employed	582(13.9%)	1418(18.9%)		
	employer	53(1.3%)	235(3.1%)	χ^2= 867.8	*
	Unpaid family workers	238(5.7%)	244(3.2%)		
	Economically inactive population	1780(42.5%)	1693(22.6%)		
log transformed Personal Income (yearly)	Median	5.96	7.27	W= 12270510	*
	Mean	-1.09	1.79		
log transformed household Income (yearly, family number adjusted)	Median	7.55	7.63	W= 15382704	*
	Mean	7.44	7.45		

* : $p<0.05$

B. Studying the significance of matching variables using logistic regression

Because statistical matching requires the assumption of conditionally independent, the statistical significance of the matching variable in the matched file must be checked (Kim, Park, 2014). Because the Korea Welfare Panel Survey does not contain information about tax payments, we use the tax payment variable from the National Survey of Tax and Benefit.

However, only if the matching variable can sufficiently explain the tax payment variable, is the analysis of the matched file meaningful. As such, by performing a logistic regression analysis of using the National Survey of Tax and Benefit's data to explain the relationship between the matching variables and tax payment status, we can estimate the statistical significance of the matching variables.

Of the variables common to the Korea Welfare Panel Survey and the National Survey of Tax and Benefit, the matching variables used to match the two data sets include gender, age group, employment status, personal income, and household income based on equivalence scale, all of which are used as explanatory variables in the logistic regression. The dependent variable is tax payment status, from the National Survey of Tax and Benefit, and the results of the logistic regression are as follows.All matching variables are statistically significant. This can be interpreted as

concluding that the matching variables to be used in linking the Korea Welfare Panel Survey data and the National Survey of Tax and Benefit are closely related with tax payment status. Therefore, analyses based on the matched file are expected to yield meaningful results.

〈Table 5〉 Logistic regression analysis result of the National Survey of Tax and Benefit

Common variable		Estimate	Standard error	Wald χ^2	
intercept		-0.4828	0.1330		
Sex	male	0	0	5.8	*
	female	-0.1813	0.0751		
Age	20~39	0	0	26.1	*
	40~49	0.0690	0.0800		
	50~59	0.0885	0.0913		
	60+	-0.6246	0.1420		
Employment Status	Regular wage workers	0	0	537.8	*
	Temporary wage workers	-2.4344	0.1839		
	Daily wage workers	-17.048	91.318		
	Self-employed	-1.5062	0.0897		
	employer	-1.0384	0.1555		
	Unpaid family workers	-16.763	25.052		
	Economically inactive population	-3.6280	0.2781		
log transformed Personal Income (yearly)		0.00007	0.000025	7.9	*
log transformed Household Income (yearly, family number adjusted)		0.00016	0.000017	87.5	*

* : $p<0.05$

C. Matching results

Based on the results of the logistic regression analysis, the matching variables used to link the two data sets are gender, age group, employment status, personal income, and household income based on equivalence scale. Of these, gender, age group, and employment status are exactly matched by blocking each variable's category. Also, in cases in which personal income and household income based on equivalence scale bracket were similar, the random hot deck using the Mahalanobis distance function was used in order to link the Korea Welfare Panel Survey data and the National Survey of Tax and Benefit. The results of the matching showed that 4,185 respondents provided information pertaining to welfare attitudes in the Korea Welfare Panel Survey, the base file. Of the 7,550 respondents in the National Survey of Tax and Benefit, the donor file from which tax payment information will be obtained, 4,185(including duplicates) are linked with records in the Korea Welfare Panel Survey. After excluding errors, the total number of subjects in the matched file is 4,182.

Welfare status is not used as a common variable between the Korea Welfare Panel Survey and the National Survey of Tax and Benefit, in an aim to ensure reliability of the matched file created from National Survey of Tax and Benefit data. However, variables that exist in the National Survey of Tax and Benefit

were analytically compared with variables in the Korea Welfare Panel Survey.

Matching the National Survey of Tax and Benefit's national basic livelihood security benefits recipient status, basic pension benefits recipient status, and child care benefits recipient status with the corresponding variables contained in the Korea Welfare Panel Survey yielded correlations of 89%, 74%, 82%, indicating reliability.

D. The comparative analysis of the welfare attitude in accordance with the dual status of taxpayer and welfare recipient by using statistical matching

This study compares and analyzes attitudes toward welfare based on groups' taxpayer and welfare recipient status, in order to provide the fundamental data to inform effective tax and welfare policies. For the analysis, we created four population groups based on taxpayer and welfare recipient status, as shown in Figure 2. The matched data set, previously created via statistical matching, was used in the analysis, and information regarding individuals' actual taxpayer status was based on their final tax payment amounts, which take into account the year-end adjustment or the tax returns after various deductions have been applied from the National Survey of Tax and Benefit. Welfare recipient status is from the Korea Welfare Panel Survey in which respondents or household members indicate whether they receive

one or more forms of government welfare benefits or services. In addition, additional variables pertaining to individuals' attitude of welfare are from the Korea Welfare Panel Survey.

The average annual personal income for non-taxpayer receiving welfare is KRW 6.33 million. In contrast, the average annual personal income of taxpayers not receiving welfare is KRW 38.89 million, or six times greater, and the distributions of personal income within those groups are clearly different. The average annual personal income of taxpayers receiving welfare is KRW 33.99 million.

[Figure 2] 4 Groups based on tax-payer and welfare-recipient status

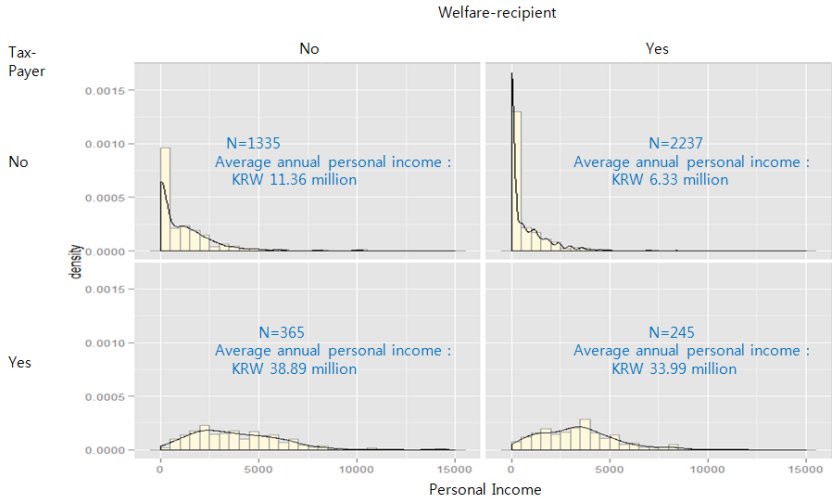


Table 6 reports the various attitudes toward welfare of the groups divided based on taxpayer and welfare-recipient status.

〈Table 6〉 Welfare attitude in accordance with taxpayer and welfare recipient

	Attitude	Group				x2
		Tax Yes & Welfare Yes	Tax Yes & Welfare No	Tax No & Welfare Yes	Tax No & Welfare No	
(a) Need to guarantee minimum living of the workable person	agreement	117 (47.8%)	161 (44.1%)	975 (43.6%)	514 (38.5%)	13.737 *
	neutral	41 (16.7%)	73 (20.0%)	444 (19.8%)	278 (20.8%)	
	disagreement	87 (35.5%)	131 (35.9%)	818 (36.6%)	543 (40.7%)	
(b) Universal welfare vs. Selective welfare	universal	117 (47.8%)	146 (40.0%)	616 (27.5%)	468 (35.1%)	91.555 *
	neutral	46 (18.8%)	76 (20.8%)	449 (20.1%)	254 (19.0%)	
	selective	82 (33.4%)	143 (39.2%)	1172 (52.4%)	613 (45.9%)	
(c) Universal childcare service	agreement	188 (76.7%)	235 (64.4%)	1421 (63.5%)	788 (59.0%)	49.082 *
	neutral	34 (13.9%)	53 (14.5%)	427 (19.1%)	229 (17.2%)	
	disagreement	23 (9.4%)	77 (21.1%)	389 (17.4%)	318 (23.8%)	
(d) Universal education service up to university	agreement	75 (30.6%)	71 (19.5%)	391 (17.5%)	242 (18.1%)	30.079 *
	neutral	36 (14.7%)	49 (13.4%)	369 (16.5%)	190 (14.2%)	
	disagreement	134 (54.7%)	245 (67.1%)	1477 (66.0%)	903 (67.7%)	
(e) Current tax burden of middle class	high	75 (31.5%)	137 (37.9%)	534 (24.8%)	391 (29.9%)	34.423 *
	proper	119 (50.0%)	167 (46.1%)	1248 (57.8%)	694 (53.2%)	
	low	44 (18.5%)	58 (16.0%)	376 (17.4%)	220 (16.9%)	
(f) Need to increase tax burden for welfare	agreement	137 (55.9%)	198 (54.3%)	1297 (58.0%)	716 (53.6%)	51.320 *
	neutral	56 (22.9%)	61 (16.7%)	469 (21.0%)	212 (15.9%)	
	disagreement	52 (21.2%)	106 (29.0%)	471 (21.0%)	407 (30.5%)	

* : $p<0.05$

First, opinions regarding the need for a guaranteed minimum living standard for individuals capable of working are analyzed. Of taxpayers receiving welfare , 47.8% are in support of the idea, a notably higher percentage than those of the other three groups and 12.3 percentage points greater than the 35.5% of taxpayers receiving welfare who are opposed to the idea. In contrast, 40.7% of non-taxpayer individuals not receiving welfare are opposed to the idea, more than the 38.5% of those who were in support of the idea.

Next, attitudes toward universal welfare vs. selective welfare, an important campaign issue which has arisen since the 2010 local election, is analyzed. The group of taxpayers receiving welfare has a relatively higher preference (47.8%) for universal welfare, compared to the other three groups. This most popular opinion was characterized by a preference for universal free child care and universal free school lunches regardless of income level or even an expansion of universal welfare programs, such as the basic pension, for which benefits have increased in recent years. The proportion of taxpayers who receive welfare and hold this opinion is 14.4 percentage points greater than the proportion that prefers selective welfare (33.4%). The group of taxpayers not receiving welfare has little preference for either universal welfare (40.0%) or selective welfare (39.2%). In contrast, the group of non-taxpayers receiving welfare has a clear, almost two-to-one, preference for selective welfare (52.4%) vs. uni-

versal welfare(27.5%).

Such a trend in attitudes is exhibited even more clearly with regard to universal free child care and universal free education up to college. Of taxpayers receiving welfare, 76.7% support universal free child care; this proportion is 12 to 17 percentage points greater than those in the other three groups, and only 9.4% of taxpayers receiving welfare oppose universal free child care. In contrast, however, a high proportion (54.7%) of taxpayers receiving welfare do not support universal free education up to college. Only 30.6%, of taxpayers receiving welfare expresses support for universal free education; however, this proportion is 11 to 13 percentage points greater than those of the other three groups.

Next, the perceptions of the middle class's tax burden are analyzed. In the group of non-taxpayers receiving welfare, only 24.8% express an opinion that the middle class burden is high. In contrast, 37.9% of taxpayers not receiving welfare express that the middle class burden is high. Of taxpayers receiving welfare, 31.5% think that the middle class tax burden is high; this proportion is 6.4 percentage points less than that of taxpayers not receiving welfare. Those who think the middle class tax burden is appropriate or low is 50.0% and 18.5%, respectively; these proportions are 3.9 percentage points and 2.5 percentage points, respectively, higher than those of the other three groups.

Finally, attitudes toward raising taxes to fund welfare ex-

pansion are analyzed. It is clear that a group's welfare status, rather than its taxpaying status, determines whether individuals in the group, on average, support such a tax increase.

Other than the group of non-taxpayers receiving welfare(58.0%), the group of taxpayers receiving welfare shows the strongest support (55.9%) for a tax increase to fund welfare expansion. Furthermore, only 21.2% of taxpaying welfare recipients oppose such a tax increase; this proportion is 7.9 percentage points lower than that of the group of taxpayers not receiving welfare.

In summary, the analysis of attitudes toward welfare expressed by the four groups based on taxpaying status and welfare-recipient status has found the following.

According to the analysis results, people with dual status of taxpayer above the tax break-even point and welfare recipient due to recent enlargement of welfare programs are relatively more positive about the expansion of the universal welfare than taxpayers who never receive welfare benefit.

Also, in the group with dual status, the recognition that the current tax burden of middle class is low is relatively higher and the negative thinking about the need to increase tax burden for the enlargement of the universal welfare programs is relatively lower than the group who pay tax but never receive welfare benefit. This finding may be interpreted as individuals who have such dual status have received welfare benefits such as free ed-

ucation, social service vouchers, and basic pension as a result of the recent expansion of welfare programs, and therefore have realized that their tax payments are utilized for expanding various welfare services to the middle class as well as the low-income class.

This is particularly meaningful in the sense that when people believe that their tax payments benefit their own lives, in addition to the lives of low-income individuals, and they trust that the government will spend their tax money responsibly, social agreement with regard to welfare expansion is more likely.

5 Conclusions

5 Conclusions

Statistical data collection in the health and welfare sector consists of 44 types of survey statistics (approval statistics). Survey statistics typically used to identify status account for 57% of statistics in the health and welfare sector[10]. Compared to survey statistics in other fields, a great deal of micro data is being produced in this sector. Individual micro data sets, which are used in research to inform policies, rarely contain all of the information researchers or policy makers desire. Even if researchers conduct surveys to collect the necessary information, some responses are missing, and the time and cost associated with surveying are constraints. In order to obtain reliable results, a large sample size and therefore a large budget is required.

In addition, with the rise of Government 3.0, which promotes active sharing of public data and elimination of barriers among agencies in order to facilitate communication and cooperation through which to provide personalized public services, create employment opportunities, and support a creative economy, the number of instances in which public data are disclosed is on the rise.

10) Source: The Korea Ministry of Health and Welfare Statistics website (http://stat.mw.go.kr). Last accessed July 2014.

Under the current circumstances, in which surveying is difficult and the availability of administrative data is increasing, data matching can increase the utility of existing survey data and administrative data, provide foundational data for evidence-based research, and contribute to the production of new statistics.

In an effort to characterize the role of national statistics, this study has aimed to examine data matching methodologies and classify data types. In addition, this study has aimed to increase the quality of the data resulting from data matching and, therefore, provide a strong foundation for various academic and policy applications.

Finally, we would like to suggest several points that should be considered in order to improve the reliability of micro data linkage methods, increase data utilization

When creating an integrated database, which of the files should be used as the base file, which of the files should be used as the donor file, and which matching method should be used are all important considerations. In addition, the quality of integrated data can vary depending on which common variables are selected for use as in the matching variable.

In particular, when statistical matching is used as the linkage technique, the basic assumptions must be reviewed theoretically, and the analytical results can differ depending on the researcher's chosen matching technique. If the sample weight differs between

the two data sets to be joined, how the integrated data's weighted value should be adjusted is another important issue.

An important issue in the realm of data matching and policy making is the protection of personal information. In fact, the protection of personal information can clash with the value of the data linkage. Especially when unique identifiers are used in exact matching, we have a responsibility to protect personal information.

In addition to taking a legal approach and utilizing masking techniques such as statistical matching, policies should be established to support the establishment of data linkage centers in order to share the responsibility of protecting personal information.

With the dawn of the big data era, we are demanding that the field of statistics take on an important role: to predict the future via analysis of big data. Previously, the role of statistics was focused on describing what is happening and why to provide information. Data matching suggests a new vision, that value can be created by maximizing predictive potential by expanding the diversity of big data and thereby facilitating new analyses. However, such prediction is not sufficient. The predictions must be capable of telling us what we need to do and how to do it. In order to formulate specific policies to solve society's problems and prepare for the future, we must seek plans for efficiently utilizing data linkage, which will maximize various effects through cooperation of researchers and policy makers.

References

Miae Oh(2014), Governance 3.0 and Bigdata: Examples in the health and welfare sector, Issue & Focus ,230, Korea Institute for Health and Social Affairs.

Kim, D. K., Park, M. G. (2014). Comparison of several micro matching methods : application to the economic activity census and time use survey, Journal of the Korean Data Analysis Society, 16(5), 2393-2408. (in Korean).

Arbeitsgemeinschaft Media-Analyse. (1996). Die Media-Analyse der Arbeitsgemeinschaft Media-Analyse eV.

Barr, R.S. and Turner, J.S. (1981). Microdata file merging through large-scale network technology. Mathematical Programming Study, Volume 15, 1-22.

Billard, L., & Diday, E. (2006). Descriptive statistics for interval-valued observations in the presence of rules. Computational Statistics, 21(2), 187-210.

Choi, H. S., Oh, M. A. (2015). The comparative analysis of the welfare attitude in accordance with the dual status of tazpayer and welfare recipient by using data linkage, Journal of the Korean Data Analysus Society, 17(4), 1983-1994 (in Korean)

Churches, T. and Christen, P. (2004) Some methods for blindfolded record linkage. BMC Medical Informatics and Decision Making, 4(9)

Dalenius, T. (1977). Towards a methodology for statistical disclosure control. Statistik Tidskrift, 15. 429-444.

Dalenius, T. (1986). Finding a needle in a haystack or identifying anonymous census record. Journal of Official Statistics, 2, 329-336.

Dalenius, T. and Reiss, S.P. (1978). Data-swapping: A technique for disclosure control. Preceedings of the Section on Survey Research Methods, American Statistical Association, Washington DC, USA, 191-194.

Dalenius, T. and Reiss, S.P. (1982). Data-swapping: A technique for disclosure control. Journal of Statistical Planning and Inference, 6, 73-85.

D'Orazio M.(2013), Statistical matching and imputation of survey data with StatMatch

Duncan, G. T., Elliot, M. and Slazar-Gonzalez, J. (2011). Statistical confidentiality principles and practice statistics for social and behavioral sciences, Springer, New York, NY, USA.

Dusserre, L., Quantin, C, and Bouzelat, H. (1995) A one way public key cryptosystem forthe linkage of nominal files in epidemiological studies. Medinfo, 8, 644-7

Dwork, C. (2006). Differential privacy. In 33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006), Springer, Venice, Italy, 1-12.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. Journal of the American Statistical Association, 64(328), 1183-1210.

Paass, G. (1985) Statistical record linkage methodology: state of the art and future prospects. In Bulletin of the International Statistical Institute, Proceedings of the 45th Session, Vol. LI, Book 2. Voorburg, Netherlands: ISI.

Radner, D.B., Allen, R., Gonzalez, M.E., Jabine, T.B. and Muller, H.J. (1980) Report on Exact and Statistical Matching Techniques. Statistical Policy Working Paper 5, Federal Committee on Statistical Methodology.

Rodgers, W.L. (1984) An evaluation of statistical matching. Journal of Business and Economic Statistics 2

Sanghoon Ahn(2000), Pro-welfare politics: A model for changes in European welfare states, Uppsala University.

Statistics New Zealand (2006), Data Integration Manual

Glossary

Data linkage(also referred to as data matching): Connecting micro data from multiple sources to create a completely integrated data set that provides richer information. This process is also used for record linkage, entity resolution, object identification, and field matching.

Record linkage (also referred to as exact matching): See exact matching.

Exact matching: A method of joining identical entities (individuals, families, events, places, etc.) found in different data sets.

Statistical matching: A method of joining similar entities (individuals, families, events, places, etc.) found in different data sets.

Administrative data: A database created and managed by a public agency for official purposes.

Survey data: Information about individual entities, collected via a survey method.

Unique identifier: A single variable or combination of multiple variables that can be used to identify an entity. Single variables often used as unique identifiers include social security numbers, business registration numbers, names, and titles. Name, dates of birth, gender, addresses, etc. can be combined to form unique identifiers.

Record: A set of information about an individual entity.

Base file (also referred to as a host file or recipient file): The file to be used as a base in the data linkage process.

Donor file: A file to be used to supplement data in the base file in the data linkage process.

Matched file: The single data file created by merging a base file and a donor file in the data linkage process.