Statistical Processing of JSTAR Datasets for the Protection of Personally Identifiable Information

The use of the JSTAR datasets is permitted only for research purposes. However, even if the use of the datasets is limited for research purposes, special care must be taken to ensure the confidentiality of personally identifiable information of individuals surveyed. Therefore, the Research Institute of Economy, Trade and Industry (hereinafter "RIETI") will take appropriate measures as set forth below to mask the following four types of personally identifiable information:

1. Detailed and municipality-level geographic information
2. Personal data on individuals and family attributes
3. Economic variables such as income, expenditures, and assets
4. Data provided in medical and nursing care insurance claims

1. Restrictions on Access to Geographic Information

The JSTAR datasets contain both detailed geographic information (specific addresses of individuals surveyed including residential districts) and municipality-level geographic information. Such information, which could lead to the identification of the individuals, naturally requires the highest level of confidentiality. Accordingly, access to such information is prohibited unless the use of such information is deemed, in the judgment of RIETI, to be essential to the particular research being conducted, as stipulated in the Application Criteria for Use of the JSTAR Datasets.

2. Top Coding and Restrictions on Access to Personal and Family Attributes

The JSTAR datasets include information not only about respondents' attributes but also those of their families. Specifically, in cases where an individual has unique personal attributes or family composition, a cross-reference of such data with other information may lead to the identification of the individual. In order to eliminate this possibility, information about respondents' attributes and those of their families is top coded and processed in a way so as to prevent the individuals from being identified. Access to data on respondents' attributes or those of their families is allowed only when the use of such data is deemed, in the judgment of RIETI, to be essential to the particular research being conducted. For details, please refer to Table 5.2.

3. Top Coding of Economic Variables Such as Income, Expenditures, and Assets

This section of the JSTAR survey was conducted on elderly people aged 50 or older regarding their economic conditions. As a general tendency, elderly people are more diverse than young people in economic attributes, most conspicuously, in income and assets. Given this tendency, it may be possible to track down the identity of a specific individual or a household by cross-referencing such economic attributes with other information when the individual or household has unusually high income and/or large assets. In order to eliminate this possibility, economic data related to income, expenditures, and assets are, in principle, top coded. Access to the data subject to top coding is permitted only when the use of such data is deemed, in the judgment of RIETI, to be essential to the particular research being conducted. Refer to Table 5.3 for the list of economic data subject to top coding.

4. Restrictions on Access to Data Provided in Medical/Nursing Care Insurance Claims and Data Aggregation

As part of the JSTAR survey, data provided in medical and nursing care insurance claims are collected subject to the consent of individuals surveyed and with the help of the governments of the five municipalities in which the survey is conducted. Such data are included in the JSTAR datasets in such a way that such data can be cross-referenced with other information and data collected in the survey.

Data provided in insurance claims includes not only the amount of insurance benefits received but also a code number by which a specific medical institution or nursing care service provider can be easily identified. Thus, datasets containing such code numbers shall not be provided to any researcher regardless the level of confidentiality protection management.

However, using a limited scope of processed medical and/or nursing care insurance data may be permitted, provided that the use of such data is deemed, in the judgment of RIETI, to be justifiable given the nature of the particular research being conducted. Specifically, researchers approved to use such data under the confidentiality protection management level of UH (ultra high) or VH (very high) will be provided with data that have been processed in such a manner that the code numbers of medical and/or nursing care service providers are replaced by serial numbers and that attribute data (type of institution, location, and the number of beds, etc.) are substituted by certain code numbers or are otherwise obscured. Furthermore, the scope of processed attribute data

made available is limited to particular medical institutions and/or nursing care service providers whose attribute data are deemed essential to the particular research being conducted. Meanwhile, researchers approved for use under the confidentiality level of H (high) will be allowed access only to data in which all the code numbers have been eliminated without substitution by serial numbers. Furthermore, information concerning medical and nursing care service expenditures and benefits will be provided only in the form of aggregate data in broad categories. For instance, the amount of expenditures spent or the number of insurance benefit points used would be aggregated by type of services (inpatients and outpatients in the case of medical services; in-home nursing care support, day services at care centers, and residential care services at care centers in the case of nursing care services) for a certain period of time.

5. Items Subject to Confidential Treatment and Confidentiality Protection Measures

5.1 Geographic Information

Survey Item	Confidentiality Protection Measure
Detailed geographic information (specific addresses)	This information is accessible only when approved for use under the confidentiality protection management level of UH (ultra high).
Municipality-level geographic information	This information is accessible only when approved for use under the confidentiality protection management level of UH (ultra high) or VH (very high).

5.2 Personal and Family Attributes

Survey Item	Confidentiality Protection Measure
Respondent's birth month and year	Only the year of birth is provided.
Spouse's birth month and year	Only the year of birth is provided.
Respondent's academic background	Detailed description is omitted in the case of an atypical academic background.
Spouse's academic background	Detailed description is omitted in the case of an atypical academic background.
Month and year of marriage	Only the year of marriage is provided.
Month and year of spouse's death	Only the year of spouse's death is provided.
Month and year of divorce	Only the year of divorce is provided.
Number of children	The data are top coded to the 97th percentile.
Month and year of children's birth	Only the year of birth is provided.
Children's academic background	Detailed description is omitted in the case of an atypical academic background.
Father's age	The data are top coded at the 97th percentile.
Mother's age	The data are top coded at the 97th percentile.
Father-in-law's age	The data are top coded at the 97th percentile.
Mother-in-law's age	The data are top coded at the 97th percentile.
Type of current job	Detailed description is omitted in the case of an atypical job.
Description of respondent's current job	Detailed description is omitted.
Description of the business of respondent's current employer	Detailed description is omitted.
Number of persons employed by respondent's current employer	The data are top coded at the 97th percentile.
Respondent's job type at age 54	Detailed description is omitted in the case of an atypical job.
Description of the respondent's job when the respondent was aged 54	Detailed description is omitted.
Description of the business of respondent's employer when the respondent was aged 54	Detailed description is omitted.

Number of persons employed by respondent's employer when the respondent was aged 54	The data are top coded at the 97th percentile.
Spouse's type of current job	Detailed description is omitted in the case of an atypical job.
Description of spouse's current job	Detailed description is omitted.
Description of the business of spouse's employer	Detailed description is omitted.
Number of persons employed by spouse's employer	The data are top coded at the 97th percentile.

5.3 Economic Variables

Survey Item	Confidentiality Protection Measure
Respondent's income after taxes and social insurance deductions	The data are top coded at the 97th percentile.
Total amount of taxes and social insurance premiums paid by respondent	The data are top coded at the 97th percentile.
Amount of taxes paid by respondent	The data are top coded at the 97th percentile.
Amount of social insurance premiums paid by respondent	The data are top coded at the 97th percentile.
Spouse's income after taxes and social insurance deductions	The data are top coded at the 97th percentile.
Total amount of taxes and social insurance premiums paid by spouse	The data are top coded at the 97th percentile.
Amount of taxes paid by spouse	The data are top coded at the 97th percentile.
Amount of social insurance premiums paid by spouse	The data are top coded at the 97th percentile.
Amount of deposits held by respondent	The data are top coded at the 97th percentile.
Amount of bonds held by respondent	The data are top coded at the 97th percentile.
Value of stocks held by respondent	The data are top coded at the 97th percentile.
Amount of deposits held by spouse	The data are top coded at the 97th percentile.
Amount of bonds held by spouse	The data are top coded at the 97th percentile.
Value of stocks held by spouse	The data are top coded at the 97th percentile.
Respondent's current hourly wage	The data are top coded at the 97th percentile.
Respondent's daily wage	The data are top coded at the 97th percentile.
Respondent's monthly wage	The data are top coded at the 97th percentile.
Amount of bonus(es) received in the past one year	The data are top coded at the 97th percentile.
Amount of respondent's business income	The data are top coded at the 97th percentile.

Respondent's annual income at the age of 54	The data are top coded at the 97th percentile.
Annual amount of public pension benefits received by respondent	The data are top coded at the 97th percentile.
Expected annual amount of public pension benefits to be received by respondent	The data are top coded at the 97th percentile.
Annual amount of public pension benefits received by spouse	The data are top coded at the 97th percentile.
Expected annual amount of public pension benefits to be received by spouse	The data are top coded at the 97th percentile.
Annual amount of private pension benefits received and/or expected annual amount of private pension benefits to be received by respondent	The data are top coded at the 97th percentile.
Amount of money spent on food per month	The data are top coded at the 97th percentile.
Amount of money spent on eating out per month	The data are top coded at the 97th percentile.
Monthly cost of living	The data are top coded at the 97th percentile.
Amount of money spent on the purchase of durable goods	The data are top coded at the 97th percentile.
Purchase price of car(s) owned	The data are top coded at the 97th percentile.
Number of years since respondent's home was built	The data are top coded at the 97th percentile.
Total floor space of respondent's home	The data are top coded at the 97th percentile.
Estimated current value of respondent's home	The data are top coded at the 97th percentile.
Amount of residential mortgage loans outstanding	The data are top coded at the 97th percentile.
Estimated current value of property (excluding home) owned by respondent	The data are top coded at the 97th percentile.
Total amount of debt outstanding	The data are top coded at the 97th percentile.
Target of savings balance one year later	The data are top coded at the 97th percentile.
Target of savings balance five years later	The data are top coded at the 97th percentile.
Ultimate target of savings balance	The data are top coded at the 97th percentile.
Amount of gift(s) before death or inheritance received by respondent	The data are top coded at the 97th percentile.
Amount of gift(s) to be received before death or inheritance	The data are top coded at the 97th percentile.

expected to be received by respondent	
Amount of gift(s) before death the respondent plans to give or inheritance the respondent plans to leave	The data are top coded at the 97th percentile.
Amount of public health insurance premiums paid by respondent	The data are top coded at the 97th percentile.
Amount of private health insurance premiums paid by respondent	The data are top coded at the 97th percentile.
Amount of nursing care insurance premiums paid by respondent	The data are top coded at the 97th percentile.

5.3 Data Provided in Medical/Nursing Care Insurance Claims

Survey Item	Confidentiality Protection Measure
Code numbers of medical institutions	All code numbers are replaced by serial numbers. Data containing such serial numbers are made available only to researchers approved for use under the confidentiality protection management level of UH (ultra high) or VH (very high).
Code numbers of nursing care service providers	All the code numbers are replaced by serial numbers. Data containing such serial numbers are made available only for use under the confidentiality protection management level of UH (ultra high) or VH (very high).
Data about the attributes of medical institutions and nursing care service providers	Attributes are substituted by certain code numbers or otherwise obscured, made available only for use under the confidentiality protection management level of UH (ultra high) or VH (very high), and only regarding particular medical institutions and/or nursing care service providers whose attributes are deemed essential to the particular research being conducted.
Data on expenditures for medical and nursing care services and insurance benefits	These data are made available only in the form of aggregate data in broad categories and for a certain period of time when provided for use under the confidentiality protection management level of H (high).

6. Random Selection of Data for Ensuring Confidentiality

Researchers approved for use under the confidentiality protection level of H (high) are provided with randomly selected data constituting 90% of the total sample.