Data Management

June 2025

1) Introduction

The *Health, Aging, and Retirement in Thailand (HART)* project is a longitudinal household survey designed to capture multidimensional aspects of aging among Thais aged 45 and above. In addition to health and retirement outcomes, the survey collects information on social, economic, and family dynamics over time. As a panel study spanning multiple waves (2015–2025), the integrity, consistency, and usability of the data are contingent upon a well-structured data management system.

This document provides an overview of the HART data architecture and variable organization to facilitate accurate and efficient use of the dataset for longitudinal analysis. The contents are structured into five key components: (1) Data Structure, (2) Variable Code, (3) Personal Identification and Household Identification, and (4) Survey Status.

By elaborating on these components, this section aims to provide users with the necessary technical foundation to manage, merge, and analyze the HART dataset effectively. Special attention is paid to ensuring consistency across survey waves, facilitating robust longitudinal and multilevel analyses for researchers.

2) Data Structure

This section provides an overview of the organization and layout of the HART dataset. The data are structured in a panel format, comprising multiple waves of data collection. Each wave includes distinct files covering core interviews, household-level information, and individual-level data. Variables are standardized across waves where applicable, thereby supporting robust longitudinal analyses.

The HART dataset is structured across multiple levels of observation, allowing for comprehensive analysis of older adults within the context of their households and broader social environments. At the **household level**, data collected in Part Aa (Household Data) capture shared attributes relevant to all household information, such as household characteristics, residence area, number of members in household, and family structure. In contrast, most other parts of the dataset operate at the **individual respondent level**, such as Part Ab (Demographic Data), Part Ac (Social Activities), Part C (Health Status), and Part G (Life Satisfaction), which include personal characteristics, experiences, and outcomes specific to the focal respondent. This differentiation between household-level and individual-level data allows for nuanced examination of how household context interacts with individual outcomes over time.

Additionally, the data structure includes modules with **nested or repeated records** to represent multiple entities connected to a respondent. For instance, in Part Bb (Children), each respondent may report data on multiple children, meaning one respondent is associated with several child-level entries. Similarly, Parts Bc (Parent), Bd (Sibling), Be (Grandchildren), and Bf (Relatives and Friends) gather information on different types of relationships, and each sub-part can include multiple entries per respondent. In Part D (Work Status), a single respondent may report more than one current or past job, especially in sub-parts Dc (Past Work), reflecting job-level data. This multi-level data architecture is central to the panel design of HART, enabling linkage across individuals, families, and time while maintaining analytical flexibility for both cross-sectional and longitudinal studies.

3) Variable code

The HART variable coding system follows a consistent and logical naming convention. Each variable name typically encodes information regarding the wave of data collection, the specific section of the questionnaire, and the subject matter or content area. This systematic structure enables users to identify, trace, and align variables across different survey years with minimal ambiguity. Variable labels and value codes are comprehensively documented in the codebook to support accurate interpretation and statistical analysis.

The variable code structure in the HART dataset is designed to provide a systematic and interpretable framework for identifying, categorizing, and navigating survey items across parts, sub-parts, and topic domains. Each variable code consists of a combination of alphanumeric identifiers that encode three key dimensions:

(1) The part of the questionnaire (e.g., Part A for General Information, Part C for Health)
(2) The sub-part or section within the part (e.g., Household Data, Demographic Data, Cognition)
(3) The topic or item grouping within that sub-part.

These codes ensure that every survey item is uniquely labeled and traceable, allowing researchers to quickly determine the context and content of any given variable.

The coding scheme is also accompanied by hierarchical classifications visible in columns such as "Row1 is Part," "Row2 is Sub-part," and "Row3-4 is Topic," which group variables thematically. For example, a code like C104001 refers to a variable in Part C (Health), Sub-part Ca (Health Status), and Topic 04 (Body pain). The final three digits designate the specific item number. This structure enhances consistency across waves and supports both cross-sectional and longitudinal analyses by helping users locate corresponding variables across years. Researchers can use these codes in tandem with the HART codebook for efficient data management, merging, and interpretation.

Table 1 Example of the variable code structure in the HART dataset

Part	Sub-part	Topic	Item	Code
C = Part C (Health)	1 = Ca (Health Status)	01 = Health score	001-002	C101001-C101002
C = Part C (Health)	1 = Ca (Health Status)	02 = Physical impairment	001-026	C102001-C102026
C = Part C (Health)	1 = Ca (Health Status)	03 = Congenital disease	001-163	C103001-C103163
C = Part C (Health)	1 = Ca (Health Status)	04 = Body pain	001-044	C104001-C104044
C = Part C (Health)	1 = Ca (Health Status)	05 = Accident	001-025	C105001-C105025
C = Part C (Health)	1 = Ca (Health Status)	06 = Urinary problems	001-005	C106001-C106002
C = Part C (Health)	1 = Ca (Health Status)	07 = Eye health	001-057	C107001-C107057
C = Part C (Health)	1 = Ca (Health Status)	08 = Ear health	001-033	C108001-C108033
C = Part C (Health)	1 = Ca (Health Status)	09 = Dental health	001-013	C109001-C109013
C = Part C (Health)	1 = Ca (Health Status)	10 = Body information	001-012	C110001-C110012
C = Part C (Health)	1 = Ca (Health Status)	11 = Health care	001-052	C111001-C111052
C = Part C (Health)	1 = Ca (Health Status)	12 = Feeling	001-010	C112001-C112010
C = Part C (Health)	1 = Ca (Health Status)	13 = Activity daily life	001-007	C113001-C113007
C = Part C (Health)	1 = Ca (Health Status)	14 = Assistance	001-002	C114001-C114002
C = Part C (Health)	2 = Cb (Health insurance)	01 = Government welfare	001-003	C201001-C201003
C = Part C (Health)	2 = Cb (Health insurance)	02 = Insurance	001-013	C202001-C202013
C = Part C (Health)	2 = Cb (Health insurance)	03 = Annual check-up	001-004	C203001-C203004
C = Part C (Health)	2 = Cb (Health insurance)	04 = OPD	001-012	C204001-C204012
C = Part C (Health)	2 = Cb (Health insurance)	05 = IPD	001-013	C205001-C205013
C = Part C (Health)	2 = Cb (Health insurance)	06 = Other medical services	001-013	C206001-C206013
C = Part C (Health)	2 = Cb (Health insurance)	07 = Assistance	001-002	C207001-C207002
C = Part C (Health)	3 = Cc (Cognition)	01 = Self-assessment	001-003	C301001-C301003
C = Part C (Health)	3 = Cc (Cognition)	02 = Word recall	001-011	C302001-C302011
C = Part C (Health)	3 = Cc (Cognition)	03 = Count backwards	001-002	C303001-C303002
C = Part C (Health)	3 = Cc (Cognition)	04 = Subtraction	001-006	C304001-C304006
C = Part C (Health)	3 = Cc (Cognition)	05 = Date remembers	001-004	C305001-C305004
C = Part C (Health)	3 = Cc (Cognition)	06 = Assistance	001-002	C306001-C306002

Note: More information in Page “Data Codebook”

4) Personal Identification and Household Identification

Each respondent in the HART study is assigned a **unique personal ID (PID)** that remains constant across waves. This enables accurate longitudinal tracking of individuals over time. Similarly, a **household ID (HHID)** identifies the respondent's household and is used to connect individual data with household-level characteristics. These identifiers are anonymized and encrypted in the public dataset to protect participant confidentiality while preserving analytical utility.

To ensure relational integrity within and across data files, each dataset includes **primary keys**—unique identifiers for each record—and **foreign keys** that establish links between related datasets. For instance, individual-level data can be linked to corresponding household-level data through shared household identifiers, while responses across different waves can be merged using a stable personal ID. These key structures support a wide range of analytical strategies, including longitudinal tracking, multilevel modeling, and cross-sectional comparisons.

5) Survey Status

The survey status variable in the HART dataset captures the outcome of the data collection process for each respondent in each wave, providing information for tracking panel retention, sample refreshment, and overall data quality. This variable includes 7 codes, each indicating a specific type of interview status or household transition scenario.

To track longitudinal participation, the HART study classifies all respondents in each wave into distinct survey status categories:

(1) **R-completed**: Respondents who fully completed the full interview in each wave.
(2) **R-pass away and R***: The original respondent passed away, and a new eligible respondent (R*) from the same household continued participation and completed the interview.
(3) **R-pass away and No-R***: The respondent passed away and there was no replacement from the household, so the case was closed.
(4) **R-refuse/unable and R***: The original respondent refused or was unable to participate, and a new household member (R*) took their place to complete the interview.
(5) **R-refuse/unable and No-R***: The respondent refused or was unable to participate, and no one else in the household could replace them.
(6) **Un-contactable**: The research team attempted but failed to contact the respondent or household after multiple attempts; the case was closed as uncontactable.
(7) **New household**: Beginning in Wave 4 (2022) – refreshment the data –, a second cohort was introduced by adding new households into the sample frame to ensure population representativeness over time.

In Wave 5, due to protocol changes, the study discontinued the replacement respondent strategy. Thus, if an original respondent passed away or refused to participate, the case was closed without attempting to identify an R*, marking a shift in panel maintenance policy.

Citation

If you use or refer to any of the documents, please cite them properly using the suggested citation formats below. This ensures academic integrity and helps others locate the original source.

Health, Aging, and Retirement in Thailand (HART). (2025). *Data Management*. Bangkok: National Institute of Development Administration (NIDA). Retrieved from https://hart.nida.ac.th/survey-design/