Research Use of Restricted Data: The HRS Experience

Michael A. Nolte  
Senior Research Associate

Janet J. Keller  
Research Associate

The Health and Retirement Study  
The University of Michigan
Introduction

- The **Health and Retirement Study** (HRS) contacts a sample of more than 22,000 Americans over the age of 50 every two years (1992-2003, 2004 survey in field)
- Supported by the National Institute on Aging (U01 AGO 9740)
- Covers: physical and mental health, insurance coverage, financial status, family support systems, labor market status, and retirement planning.
- Level of detail increases confidentiality risks.
HRS Data Products

- **Public Data**: available without cost to researchers and analysts (user registration required)
- **Restricted Data**: available only under specific conditions
  - Affiliation with an institution with an NIH-certified Human Subjects Review Process
  - Current Receipt of United States Federal Research Funds
HRS Restricted Data Products

- SSA Administrative Data (Earnings, Benefits, WSEI, SSI)
- Detailed Geographic Identifiers: State, County, ZipCode, Tract/BNA
- Respondent Geographic Markers (Birthplace, Residence, Mobility)
- Detailed Industry and Occupation codes
- Pension Information
- Medicare Claims and Summary Information (from CMS)
- Cause of Death (from National Death Index)
# Health and Retirement Study: Data Collection Path

## Sample Development

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HRS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AHEAD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coda/WB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Early Boomers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Interview Data

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Core</td>
<td>HRS</td>
<td>AHEAD</td>
<td>HRS</td>
<td>AHEAD</td>
<td>HRS</td>
<td>All Cohorts</td>
<td></td>
<td>All Cohorts</td>
<td></td>
<td>All Cohorts</td>
<td></td>
<td>All Cohorts</td>
<td></td>
</tr>
<tr>
<td>Exit</td>
<td>HRS</td>
<td>AHEAD</td>
<td>HRS</td>
<td>AHEAD</td>
<td>HRS</td>
<td>All Cohorts</td>
<td>All Cohorts</td>
<td></td>
<td>All Cohorts</td>
<td></td>
<td>All Cohorts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modules</td>
<td>HRS</td>
<td>AHEAD</td>
<td>HRS</td>
<td>AHEAD</td>
<td>HRS</td>
<td>All Cohorts</td>
<td>All Cohorts</td>
<td></td>
<td>All Cohorts</td>
<td></td>
<td>All Cohorts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imputations</td>
<td>HRS</td>
<td>AHEAD</td>
<td>HRS</td>
<td>AHEAD</td>
<td>HRS</td>
<td>All Cohorts</td>
<td>All Cohorts</td>
<td></td>
<td>All Cohorts</td>
<td></td>
<td>All Cohorts</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Supplements

### Aging, Demographics and Memory Study

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mail Surveys</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- ADAMS
- 1999 Mailout
- CAMS
- CAMS
- HUMS
- Diabetes

## Linkages

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Employer Pension Plans</td>
<td>HIPPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Security Permissions</td>
<td>HRS</td>
<td>AHEAD</td>
<td>HRS</td>
<td>AHEAD</td>
<td>HRS</td>
<td>All Cohorts</td>
<td></td>
<td>All Cohorts</td>
<td></td>
<td>All Cohorts</td>
<td></td>
<td>All Cohorts</td>
<td></td>
</tr>
<tr>
<td>NDI Match</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Notes:

1. For more information, click on the arrow or box of interest.
2. Initial interview year for cohort:
3. Use the [Question Concordance](#) as a tool for cross-referencing Core and Exit variables by content across time.
4. Color Codes:
   - Available: Released data product
   - In process: Data being collected or processed
   - Planned: Future data collection effort
   - Core: Imputation products available
   - Exit: Imputation products available
Why Restrict Access to Data: A Compromise Solution

- Allow researchers to access confidential data
- Preserve respondent privacy and prevent re-identification
Restricted Data Access Methods

- Restricted Data Agreement
  ([http://hrsonline.isr.umich.edu/rda/index.html](http://hrsonline.isr.umich.edu/rda/index.html))
- MiCDA Data Enclave
  ([http://micda.psc.isr.umich.edu/enclave/](http://micda.psc.isr.umich.edu/enclave/))
- Virtual Enclave (under development)
HRS Restricted Data Application Process

1. Determine eligibility, review RDA procedures, select restricted files, submit application (research plan and data protection plan)

2. Data Confidentiality Committee Working Group (DCC-WG) receives, logs and reviews application; contacts researcher if problems are found.

3a. Submit application to local IRB for review

3b. Obtain Contracting Authority approval

3c. Submit signed originals of IRB and Contracting Authority approval documents to HRS

4. DCC-WG receives, logs and reviews IRB/CA forms; then forwards the application package to the Data Confidentiality Committee (DCC)

5. Principal Investigator signs contract

6. Restricted data are sent to researcher

7. Post application activities: RDA modification
   RDA renewal
   Yearly Recertification
   Audits

FINISH
RDA Pros and Cons

Users obtain access to HRS restricted data products, but...

User burden
- Eligibility requirements
- Data protection plan
- Yearly inspection, annual report and bi-annual renewal
- Pre-publication review (under certain circumstances)

HRS burden
- Administrative overhead
- Data protection plans: users expect support from HRS.
...The Need to Protect Restricted Data

- Users may not be aware of security issues
- HRS staff must keep up with Internet developments
- Complexity grows in direct proportion to Internet security threats

Options (see Checklist for details)
  - Standalone workstation or private network (good)
  - Internet-connected workstation/network (bad)
Internet connections are a problem...

**Standalone workstation requires...**
- Physical protection (locked office)
- System OS configured using NIST guidelines
- File encryption

**Internet workstation requires...**
- Physical protection for workstation (also server, network components)
- Firewall supplemented by intrusion detection system
- Anti-virus and anti-spyware software installed and updated
- System OS configured using NIST guidelines (+ update procedures)
- File encryption
Alternative: The HRS/MiCDA Data Enclave

- Solution for users who cannot implement their own data protection plan
- Restricted data users conduct all analysis with the confines of a secure facility maintained and staffed by HRS.
- Private network: server, workstations, printers not connected to outside world
- Workstation access to restricted data limited to authenticated users
- Physical access to facility is controlled
The Enclave Solution: Advantages

- Allows access for researchers who do not qualify for a Restricted Data Agreement (graduate students, junior faculty, commercial users)
- Provides a venue for special projects. (cross-category merges of files from various sources)
- Disclosure review procedures limit risk of potential confidentiality breaches
- Serves users with special needs
The Enclave Solution: Disadvantages

- Inconvenient for non-local users
- Expensive to implement and operate
- Staffing issues for HRS
- Users are subject to restrictions on research activities (pre-export disclosure review)
Option 3?

We want to…

- Minimize the security burden for researchers
- Give HRS better control over security model definition and implementation
- Help non-local users who do not qualify for RDA

Let’s see what it would take to set up a “Virtual Enclave”
The Virtual Enclave Concept: Overview

- All analysis takes place on server
- Server is hardened, located behind firewall, monitored by Intrusion Detection System
- Thin client connection to the server
  - VPN tunnel using SSH-2 (authentication via digital certificate)
  - Remote Desktop Protocol connection (high encryption) to server running Windows Terminal Services
- Procedures for supervision of remote workstation specified in MOU negotiated between HRS and remote site
Virtual Enclave: Technical Background

- Custom Linux client with built-in SSH and RDP support
- Virtual Private Network (VPN) connection from client to server using SSH (authentication via Digital Certificate)
- Desktop Protocol (RDP) connection from client through VPN to server

Diagram:
- Management Console (Intrusion Detection)
- Secure Server
- Remote Workstation
- Private Network
- Public Network
- Firewall (port 22 open)
- RDP Link
- RDP thru VPN
Server Issues...

- Server configured for high security environment (non-routable network behind firewall, etc.)
- Programs and data files accessible by user are limited
- User programs requiring inbound or outbound Internet connections are removed or modified
- Access to client resources (disk storage, printers, serial ports) removed
- No user access to server network resources (e.g., printers)
Example...

```
ssh2.exe -c 3des-cbc -L 5000:localhost:3389 user1@xxx.xxx.xxx.xxx
user1's password:
Authentication successful.
Last login: Sun Jul  4 15:27:33 2004 from 198.108.24.2
Microsoft Windows XP [Version 5.1.2600]
(C) Copyright 1985-2001 Microsoft Corp.
C:\Documents and Settings\user1>
```
Example...

user1 now has full access to server resources
Virtual Enclave: Advantages

- Researchers obtain access to all HRS restricted data products.
- Researchers avoid travel and lodging costs.
- Researchers and/or their sponsors are responsible only for the remote site.
- Once the server side environment is implemented, multiple sites and VPN links can be added at little cost to HRS.
- User fees help defray costs charged to HRS.
Virtual Enclave: Potential Problems

- Remote site requires a sponsoring organization (monitoring staff, equipment, workspace)
- Server side maintenance costs: (software, server, communications equipment, staff, security, VPN link management)
- Users need funding to pay HRS cost recovery fees
- Staff costs at HRS end: Confidentiality review of user analysis results, custom linux client maintenance, server maintenance, long-distance interaction between HRS staff and remote users…and more!
Conclusion

- RDA most convenient, but strings are attached
- Enclave works for local users; does not meet external user needs
- Virtual enclave still under development; implementation uncertain
- Restricted data = non-trivial overhead costs
If you want to learn more...

- HRS restricted data Web site: http://hrsonline.isr.umich.edu/rda
- Developing a data protection plan: http://hrsonline.isr.umich.edu/rda/rdapkg_prot.htm
- The Restricted Data Environment: Issues Relating to Network-Connected Clients: http://hrsonline.isr.umich.edu/rda/RstDataEnv.htm
- Michigan Center on the Demography of Aging Data Enclave (MiCDA) Web site: http://micda.psc.isr.umich.edu/enclave/
- Maintaining restricted data confidentiality http://hrsonline.isr.umich.edu/rda/rdapkg_disclim.htm
References