RESEARCH USE OF RESTRICTED DATA: THE HRS EXPERIENCE

Michael A. Nolte, Senior Research Associate, The University of Michigan
Janet J. Keller, Research Associate, The University of Michigan

Key Words: researcher eligibility; distribution procedures; disclosure limitation; respondent confidentiality

Introduction

The Health and Retirement Study (HRS) is intended to provide data for researchers, policy analysts, and program planners who are making major policy decisions that affect retirement planning, health insurance, saving, and economic well-being of Americans over age 50, and to test theories and estimate the parameters of dynamic behavioral models. By collecting data in multiple domains typically studied by separate disciplines, the HRS seeks to facilitate interdisciplinary research.

The biennial interview covers a wide range of content areas including: detailed income and wealth; work, retirement, and work history; health care utilization, insurance coverage, and out-of-pocket spending; relations with other family members including monetary transfers and detailed data on time spent giving or receiving care; and self-reports of major health conditions.

HRS data offerings are supplemented by restricted files derived from U.S. Federal Government administrative data and elsewhere, pension estimation data access software, special-purpose modules targeting respondent sub-categories, and off-year mail surveys.

Since HRS is a longitudinal study, an enormous amount of information is gathered at multiple time-points for each respondent. This greatly increases the likelihood that an individual (or family, household, employer, or pension benefit provider) can be identified. In order to minimize the danger of breaches of respondent confidentiality, all direct identifiers are removed: names); addresses (lot/street, city, state, Zip Code); numeric identifiers (SSN, Medicare ID, Medicaid ID, telephone numbers); and exact dates (date of birth, date of death, date of interview, dates married/divorced, military service dates, disability periods). In addition, critical variables such as geographic location and occupation/industry are recoded to less specific levels than those provided by the respondent. Creating sanitized data sets suitable for *unrestricted*[1] distribution to the public imposes a serious cost on researchers who need access to detailed identifying information. To assist these researchers, special *restricted* data sets containing sensitive data are made available to qualified individuals under specific contractual conditions. These data sets fall into two classes: excised survey data and data linkages.

The first class of restricted data set is built from the kinds of sensitive information mentioned in the previous paragraph: geographic information; detailed occupation/industry codes; and date of interview. The second class of restricted data is administrative information obtained from third parties (with the permission of respondents). Data sets derived from such information include Social Security earnings and benefits files, pension information, National Death Index data, and Medicare claims information. The organizations that provide these data usually require the imposition of specific contractual conditions before they release the data. For example, the Social Security Administration provides earnings and benefits data to HRS under the terms of a Memorandum of Understanding that sets rules for data processing (rounding and top-coding) as well as conditions under which these data may be disseminated and used.

At present, the HRS makes both types of restricted data available to researchers through two methods: Restricted Data Agreements and the MiCDA Data Enclave. A third method, tentatively called the Virtual Enclave, is at the proof-of-concept stage, and is being evaluated by HRS staff members. All three methods are reviewed in this paper.

Restricted Data Agreements

The term "restricted data" refers to HRS data sets that cannot be distributed to the general public because of respondent confidentiality concerns, or because third-party user agreements prohibit redistribution. Researchers who wish to use HRS restricted data must meet a rigid set of qualifications before they can

[1] Freely available (after registration) through the HRS Web site (http://hrsonline.isr.umich.edu) on condition that the researcher not attempt to identify individual respondents

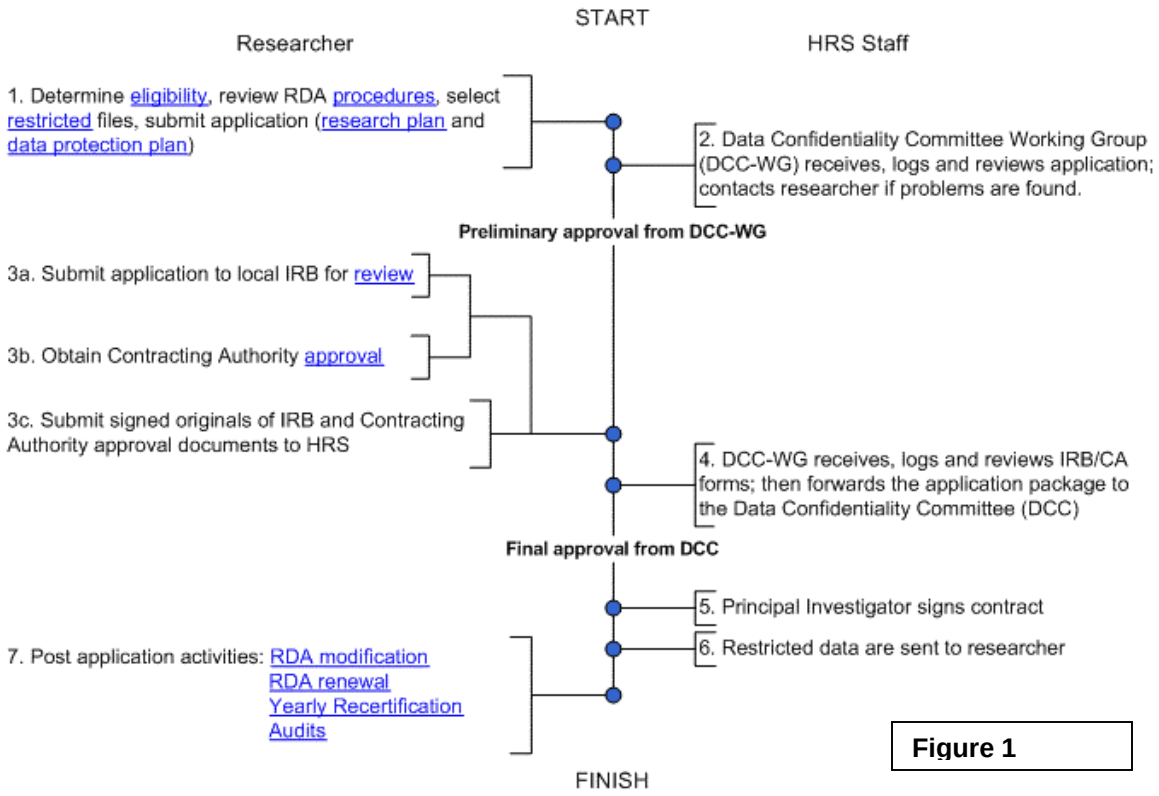


Figure 1

obtain access. Researchers are eligible to receive HRS restricted data sets only if they are affiliated with an institution with an NIH-certified Human Subjects Review Process, and are Principal Investigator of a project funded by the U.S. Federal Government. A visual representation of the restricted data application[2] process appears in Figure 1, above.

Restricted Data Agreements: Advantages and Disadvantages

Delivery: Restricted Data Agreements allow the HRS to distribute restricted data while retaining some measure of control over how researchers use these data. In return, the researcher benefits by obtaining access to all HRS restricted data products. The catch is that the researcher must: (1) meet eligibility requirements (U.S. Federal funding and NIH-approved IRB process); (2) develop and implement a data protection plan[3] to protect respondent confidentiality; and (3) agree to yearly inspections, annual reports, a renewal process every other year, and in some cases, pre-publication review of analysis results.

Application procedures: HRS typically grants quick approval to user requests as long as the research and data protection plans are straightforward. Unfortunately, non-standard research and data protection plans take significant time to review and approve. In addition, implementing a non-standard data protection plan may require extra expense and effort (e.g., additional equipment, software and/or workspace) on the part of the researcher.

Costs: There are no direct charges for HRS restricted data; however, the researcher and HRS share indirect costs. The HRS is responsible for application review, data preparation, and data delivery. The researcher is responsible for all local costs. HRS and the researcher share the costs of the application approval process.

Special Rules: Pre-publication review of analysis results is not required except for reporting geographic analysis below the census division level.[4] Certain merges (e.g. social security earnings records with geographic information) are prohibited. This

[2] See the Health and Retirement Study Web site http://hrsonline.isr.umich.edu/rda for more information on the requirements for obtaining restricted data.
[3]For data protection plan details, see: http://hrsonline.isr.umich.edu/rda/rdapkg_prot.htm

[4] For information on maintaining respondent confidentiality, see: http://hrsonline.isr.umich.edu/rda/rdapkg_disclim.htm

prohibition is enforced through the Restricted Data Agreement.[5]

Usage: Since the program began in 1996, 147 Restricted Data Agreements have been executed; 99 of these are active. In this time period, 462 researchers (294 active) at 47 research sites have used one or more of the data sets provided through this process.

The MiCDA Data Enclave

The process described above tends to discriminate against students and junior faculty members. To meet the needs of these potential users, in cooperation with the Michigan Center on the Demography of Aging (MiCDA) Data Enclave, the HRS established a facility where researchers who do not meet the standard qualifications can obtain access to restricted data. The MiCDA Data Enclave is designed to assist prospective users of restricted data files who do not meet the requirements imposed by restricted data contractual agreements, as well as researchers who have special data analysis needs that cannot be met under the terms of a standard restricted data agreement. Respondent confidentiality is maintained in a controlled, secure environment. In order to gain access to Enclave data resources, researchers agree that any data file provided to them will be used only for statistical reporting and analysis, and will not be published or released in identifiable form. In this context, the term "statistical summary information" means the result(s) of statistical analysis in any of the following forms: record listings; frequency tabulations; magnitude tabulations; means; variances; regression coefficients; correlation coefficients; graphical displays; and any other result of an analytic process. To help ensure a low probability of accidental individual identification, researchers promise not to remove any printouts, electronic files, documents, or media from the Data Enclave premises until they have been scanned for disclosure risk by Data Enclave staff. [6]

Data Enclave: Advantages and Potential Problems

Delivery: The Enclave provides restricted data to researchers who cannot meet standard licensing requirements. Researchers benefit by obtaining access to all HRS restricted data products without needing to worry about technical issues. The only difficulty is that the researcher must be physically present at the Enclave facility in Ann Arbor

Application procedures: Applications can be approved quickly as long as the research plan is straightforward. Special merges (e.g. social security earnings records and geographic information) are allowed, although such requests may require additional negotiations with third party data providers.

Costs: The HRS charges user fees to help defray costs; users must obtain funding in order to use Enclave facilities. The indirect costs of the application approval process are shared by HRS and the researcher. The HRS and the Michigan Center for the Demography of Aging are responsible for all direct Enclave costs (software, server, workstations, communications equipment, staff, workspace, security).

Special Rules: Enclave staff members must review all analysis output for confidentiality problems before the users are allowed to publish their results.

Usage: At present 24 researchers are using the Enclave on a regular basis (62 out of 65 business days during the period January 2, 2004 through March 31, 2004), and more can easily be added. In addition, sufficient Enclave capacity is available to provide HRS staff with a secure environment for processing restricted data.

A Third Option?

Although the MiCDA Data Enclave works well for University of Michigan users, the need for researchers to be physically present in Ann Arbor is clearly a problem for non-local users. In an attempt to solve this problem, we are investigating the possibility of implementing a *Virtual Enclave*. This is a method for remote access to restricted data that: gives non-local users a secure connection to Enclave resources; minimizes the administrative and technical burdens on researchers; and preserves HRS control over the data access process.

The security model that underlies this method assumes that restricted data will be stored on an HRS server, but that to users, these data will appear to be accessible from their desktop. The element that makes this solution attractive in terms of cost is our use of a virtual private network (VPN)[7] to encrypt all data

[5] The Social Security Administration will approve the use of geographic information with earnings and/or benefits information provided that (1) analysis results are submitted for approval; (2) all analysis is conducted under HRS supervision in the MiCDA Data Enclave

[6] For more information, including details on the disclosure review process visit the MiCDA Data Enclave Web site: http://micda.psc.isr.umich.edu/enclave

[7] Private network built atop a public network. Hosts within the private network use encryption to talk to

transmitted between the remote site and our secure server. In implementing our Virtual Enclave we tested two client-server setups: thick client and thin client. Thick client implementations map server resources to the local workstation. All processing occurs on the client with data being moved over the VPN from the server to the client whenever necessary. In contrast, thin client solutions move all processing to the server. Only user keystrokes and screen display information move across the network. A complete report on our testing process is available for download[8], but the clear answer to the thick vs. thin client question was that we preferred the latter. We found that run times for common types of analysis were several orders of magnitude longer when data were being moved over the network to the client. We also found that it was much easier to implement thin client solutions; as we shall see, the technical burden on the remote user is minimal.

Virtual Enclave: Technical Background

Server: We begin with a server running Windows Terminal Services[9] configured for high security environment. This machine is located behind a firewall on a non-routable network and is monitored by an intrusion detection system.[10] Programs and data files accessible by server users are severely limited. All user-accessible programs requiring inbound or outbound Internet connections are removed or modified. User access to server network resources (e.g., printers) is also removed. In addition, when clients are connected to the server, access to client-side resources (disk storage, printers, serial ports) is disallowed.

Client: On the client side, users may choose any operating system that supports the Secure Shell Protocol (SSH2)[11] and a version of the Remote Desktop Protocol that supports high encryption. Since all processing occurs on the secure server, no other client software products are required.

Network: We developed a client-server connection process that requires very little direct action by the remote user. The user begins by initiating a VPN tunnel using SSH2, and authenticating via digital certificate.[12] Once this connection is running, the user opens a Remote Desktop[13] connection to the server, and logs in. (Figure 2 provides an overview of this process.) Once the connection/authentication process is complete, the user's desktop on the remote server appears, and work begins. This combination of SSH and RDP was tested on remote workstations running Windows 2000, Windows XP, Macintosh OS-X, and Red Hat Linux v8. In each case, the connection was made with ease no matter what network topology was in use. An additional benefit is that client software is available through an Open Source License (linux) or from Microsoft (PCs and Macs).

Virtual Enclave: Advantages and Disadvantages

The Virtual Enclave solution has a number of advantages. The most important one is that researchers can obtain access to all HRS restricted data products by meeting the same conditions as local users of the MiCDA Enclave. The remote link replaces travel and lodging costs. Technical problems are minimized by the system; researchers and/or their sponsors are responsible only for an Internet-connected workstation at the remote site. For HRS there are a number of advantages. Once the server side environment is implemented, multiple sites and VPN links can be added at little cost to HRS. This allows more users to be served without losing control over our restricted data.

The Virtual Enclave concept does have some disadvantages. The remote site requires a sponsoring organization that is willing to provide user monitoring, equipment, and a workspace. Since it is likely that HRS will need to impose some sort of cost recovery system, users will also need to obtain funding to cover these fees. The HRS will need to provide for additional setup, staffing, and maintenance costs. Confidentiality review of user analysis results will still be required, complicated by the need for long-distance interaction between HRS staff and remote users. A real-world VPN implementation will require significant management resources, since support for installation and maintenance of VPN client software at remote sites

---

other hosts; the encryption excludes hosts from outside the private network even if they are on the public network. Source: http://www.visi.com/crypto/inet-crypto/glossary.html

[8] *The HRS Virtual Enclave Prototype: Testing Process and Preliminary Results* http://hrsonline.isr.umich.edu/docs/dmgt/VPNTesting.pdf

[9] When testing the SSH2 link, we used a Windows XP machine as a server.

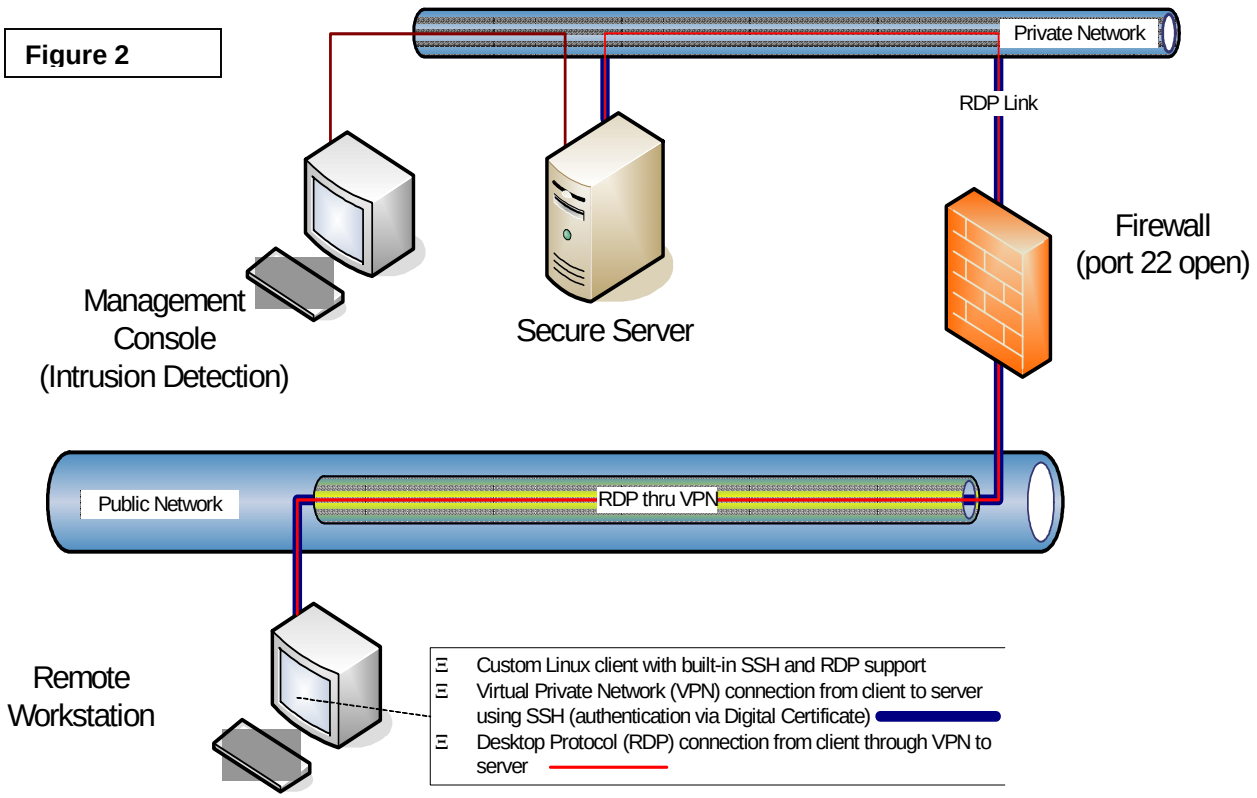
[10] Only port 25 needs to be open on the firewall; all others (including port 80) can be blocked.

[11] http://www.ssh.com

[12] While conducting our tests, we used password authentication; digital certificate authentication will be used in a production environment.

[13] For details on Remote Desktop see: http://www.microsoft.com/windows2000/techinfo/howitworks/terminal/rdpfandp.asp

Figure 2



will be necessary. Server and firewall resources must support multiple VPN connections without performance degradation. Since it is always necessary to consider the possibility that the firewall might be breached, an intrusion detection system is necessary to monitor traffic on the server subnet.[14]

Conclusion

The Health and Retirement Study has established special procedures that allow researchers to access a wide range of restricted data products. These procedures, as implemented, have allowed hundreds of researchers to access special data sets such as earnings and benefits data, detailed geographic information, medical information, and pension data. The HRS recognizes that neither Restricted Data Agreements nor the MiCDA Data Enclave can meet the needs of all researchers. As a result, development has begun on a Virtual Enclave system that will allow remote users to access restricted data. This concept shows promise in test situations, but implementation details must be carefully considered before a production system can be offered to the public. We concede that these solutions are complicated, inconvenient, and expensive; unfortunately there is no alternative if we are to maintain the confidentiality of our respondents.

[14] For a discussion of intrusion detection techniques, see S. Northcutt et al., *Inside Network Perimeter Security*

References

Bott, E. and Siechert, C. Microsoft Windows Security Inside Out for Windows XP and Windows 2000, Microsoft Press, ISBN 0-7356-1623-9

Lockhart, A. Network Security Hacks, O'Reilly, ISBN 0-596-00643-8

McNab, C. Network Security Assessment, O'Reilly, ISBN 0-596-00611-X

Northcutt, S., Zeltser, L., Winters, S., Fredrick, K.

Ritchey, R. Inside Network Perimeter Security, New Riders, ISBN 0-7357-1232-8

Souppaya, M., Johnson, P.M., Kent, K. Harris, A. Guidance for Securing Microsoft Windows XP Systems for IT Professionals: A NIST Security Configuration Checklist, Special Publication 800-68 (Draft), June 2004.