



“Virtual Data Access” for Statistical and Research Purposes

2014 FCSM Statistical Policy Seminar
December 15-16, Washington D.C.



Overview of the Substance Abuse and Mental Health Data Archive (SAMHDA)

- SAMHDA website was first published on December 3rd, 1997
 - Data Portal was implemented in January 2013
- SAMHSA's main goal for creating a virtual data access system
 - Comply with M-13-13 & Open Data
 - Maximizing the potential of expensive datasets, and committing to CBHSQ's designation as a statistical unit (2006)
 - Provide safe and secure access to restricted data for approved researchers
- SAMHSA's 'buy-in' for this virtual data access system was achieved by:
 - Universal agreement within the agency to make data more available due to its analytic and research potential
 - Weighed pros/cons of having an RDC, VDE, etc. and this was the least costly, and optimized the ease of access with high data security

Development of SAMHDA Data Portal

BRIEF DESCRIPTION

- The SAMHDA Data Portal works by:
 - Remote access through a virtual data enclave
 - Tri-factor authentication (unique name and password, Mtoken, IP address)
- The Data Portal is maintained through:
 - Contract that is currently with University of Michigan
 - SAMHSA staff
- The Data Portal employs:
 - No statistical disclosure, researchers have access to full restricted data files

Development of the SAMHDA Data Portal

RESOURCES/CHALLENGES

- The Data Portal requires certain resources:
 - A yearly budget of ~200,000
 - FTEs, IT to ensure security, troubleshoot, monitor accounts and users, upload files, and conduct disclosure reviews
- Challenges face when developing the Data Portal:
 - No precedence to follow or model after
 - Contractor inexperience, and developing everything from scratch
 - Internal Staffing
- These Challenges were overcome by:
 - Every decision was treated as extremely important
 - Extensive thought and deliberation went into every aspect of the project, and there is very little that is not customized

Development of the SAMHDA Data Portal

TESTING

- Testing for Data Portal before going live involved:
 - On-site inspections of the Contractor
 - Hiring an ethical hacker
 - First user was an NIH employee
- Glitches/Potential issues found when testing the Data Portal:
 - A few vulnerabilities
 - Always the human element
- These were corrected by:
 - All glitches were corrected
 - We have an ATO and systems are currently being checked/ tested for potential issues
 - Contracts, inspections of user sites

Development of the SAMHDA Data Portal

IT & RISKS

- IT Challenges faced when developing the Data Portal:
 - HHS did not support VMWare, so testing wasn't possible for some staff
 - To connect to the Data Portal, a static IP address is needed so there was a need to work closely with IT
- These challenges were mitigated through:
 - Working cooperatively with IT to get resources needed
 - Pilot sites were created to help identify IT risks

Conducting Business for the SAMHDA Data Portal *COSTS*

- Charges and costs associated with the Data Portal:
 - There is no charge for the use of the Data Portal
 - There is no attempt at cost recovery

Conducting Business for the SAMHDA Data Portal *SYSTEM USERS*

- Most of the researchers using the Data Portal come from the academic community
- Reactions of users:
 - Researchers were extremely happy with system and availability of data
 - The application process can be long, we are moving towards an electronic version of applying vs. PDFs
 - Researchers submit multiple files for disclosure review and turnaround time can be longer than they would like

Conducting Business for the SAMHDA Data Portal *USAGE*

- Currently over 40 research groups using the Data Portal, more applications coming in for our 3rd call for applications
- Questions regarding data quality are rarely an issue, if at all
- Users are in constant contact with Data Portal administrators through an email help address, and 800 number
- Call for applications (month long), applications are submitted and reviewed by a team of people, approved based on criteria
 - There have been about 20 applications each call
 - Each application is reviewed by 4-5 people
 - The number of applications seems to increase with each call for applications

Conducting Business for the SAMHDA Data Portal *IMPACT*

- Always great to make data available to researchers, however costs to keep the Data Portal are steadily increasing
- The Data Portal has allowed us to make restricted data collected under CIPSEA available to researchers for statistical purposes
- Tremendously positive feedback
- Ongoing concerns with the Data Portal are:
 - Increasing the systems capabilities (i.e. moving from 32 bit to 64 bit environment, increasing server size, etc.)
 - Site inspections occurring within one year of project access
 - Human behavior

Conducting Business for SAMHDA Data Portal *MOVING FORWARD*

- So far the original vision aligns fairly reasonably with the actual Implementation
 - Staff resources don't allow for applications on a rolling basis
 - Researchers don't submit final versions for disclosure review
- Virtual access is growing and it seems to be a vision that is adopted by more agencies as a way to disseminate data for researchers
- Moving forward: we are constantly working on improving the ease of access and ensuring the data is protected. VDE is a way to cut costs and reduce the work load on employees that conduct data runs. With VDE's growing, there are also implications of more secondary review, data security, and revisions of DUA's.

Questions?

www.datafiles.samhsa.gov

Brooklyn Lupari

Brooklyn.Lupari@samhsa.hhs.gov

Jonaki Bose

Jonaki.Bose@samhsa.hhs.gov