

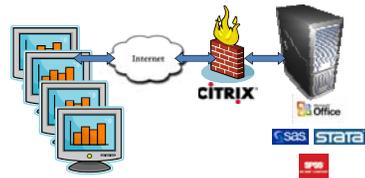
## Accessing CMS Claims Records: Data Enclave as a Virtual RDC

**Elizabeth C. Hair, PhD**  
 Senior Research Scientist  
 Public Health, NORC at the University of Chicago  
 Bethesda, MD  
[Hair.Elizabeth@norc.uchicago.edu](mailto:Hair.Elizabeth@norc.uchicago.edu)  
 301-634-9386



## What is the Data Enclave?

- ❑ The Data Enclave is a secure environment that allows for remote access to confidential microdata




ISPOR 2011 WS 2

## What is the Data Enclave? (cont.)

- ❑ The environment is firewalled from outside intrusion, and is only accessible to authorized users
- ❑ All information inflow and outflow is controlled and monitored by experienced confidentiality officers
- ❑ Researchers are provided access to numerous analytic tools in the secure environment to work with the data, including MySQL, SAS, Stata, SPSS, R, and more



ISPOR 2011 WS 3

## Comparing Traditional & Virtual RDC

- ❑ The Data Enclave resembles a physical RDC
  - Users conduct their work in a controlled environment
  - Information flows are strictly controlled and monitored
  - Researchers may access sensitive raw microdata securely
- ❑ The Data Enclave as a "Virtual RDC"
  - User activity is logged and may be audited at the keystroke level
  - Users need not make travel plans to obtain data access
  - Data providers can review & disclosure proof outputs remotely
  - Less costly to build infrastructure for a virtual environment
  - Sharing/collaboration environment allows users within the same group to interact with one another while working with the data



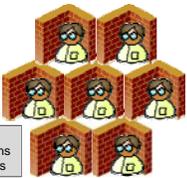
ISPOR 2011 WS 4

## The Data Enclave as a "Virtual RDC"

Geographically dispersed researchers travel to secure RDCs



Instead of creating costly brick and mortar RDCs, the Data Enclave can now roll out a virtual RDC to any university in the country



Thin client terminals are installed in secure locations at researchers' universities




ISPOR 2011 WS 5

## Challenges with Handling CMS Claims Records

- ❑ Volume of the data is enormous
  - 5% sample in 2008 (BEN, BASF, IP, OP, CARR, HHA, HOSPICE, PDE, DME, SNF) in uncompressed ASCII format is about 230GB
  - Processing large data files requires state-of-the-art cyber-infrastructure (processing power, RAM, storage space, etc)
- ❑ Administrative Records
  - Transforming administrative records into usable analytic files for users is a non-trivial exercise
  - Data mining process can be complex and time-consuming



ISPOR 2011 WS 6

## Customized Data Solution

- ❑ To address these challenges, NORC leverages an Advanced Computational Engine
  - The computational engine and sub-setting tool will be customized for CMS Claims Records
  - This customized version will include additional analytic functionality so researchers may subset, analyze, and merge data according to complex definitions
- ❑ The platform allows for fast, easy tabulation and data manipulation across multiple large datasets such as those necessary for effective CER



ISPOR 2011 WS

7

## Customized Data Solution (cont.)

- ❑ Advantages of the Advanced Computational Engine
  - CMS data subsets pertaining to complex inclusion/exclusion restrictions can be created on the fly.
  - Users can perform non-confidential tabulations of the entire dataset
  - This means that users can check the robustness of their subset against other possible subsets and alter their requests
- ❑ The Advanced Computational Engine will also allow users to merge data and construct analytic variables before downloading the data into the statistical program of their choice
  - Including MySQL, SAS, Stata, R, and more



ISPOR 2011 WS

8

## Use Case Example

- ❑ Many CMS data requests correspond to CER (Comparative Effectiveness Research) projects
- ❑ For instance, one request seeks examines the effectiveness of a particular treatment procedure on breast cancer patients.
  - Analysts first would need to identify those breast cancer patients that were treated by a particular procedure
  - Conditions and treatment procedures would be defined using the claims data files – and the Advanced Computational Engine would extract appropriate observations with their claims
  - The Advanced Computational Engine could compile beneficiary level data file using claims records and construct analytic variables
    - e.g. disease, cost, utilization, etc.



ISPOR 2011 WS

9

## Disclosure Review

- ❑ No data or analytic output can be removed from the Enclave without undergoing formal statistical disclosure control and approval of sponsor agency and NORC confidentiality officers
- ❑ Once users have finalized their results, they submit them to Enclave managers for review
- ❑ If the output is safe for release (i.e., no disclosive information about beneficiaries), results are distributed to researchers via a secure file transfer protocol site
- ❑ If there are problems with potential disclosure, the sponsor agency and Enclave staff work with users to resolve the remaining confidentiality issues.



ISPOR 2011 WS

10

## Phase 2: Research Data Center Pilot

- ❑ The Data Enclave as a “Virtual RDC”



ISPOR 2011 WS

11

## Process

- ❑ **Select pilot researchers.** In close coordination with CMS, we will select nine to twelve pilot test researchers, the goal of which is to simulate the actual experience of CER power users.
- ❑ **Sample data subset.** An initial proof of concept for this effort will be implemented based on the standard 5% LDS of all claim types subset currently available in the NORC Data Enclave.
- ❑ **Application process.** Pilot test researchers will submit applications via an online application process to NORC. To the extent possible, the application process will mirror current (or planned) processes.
- ❑ **Access Node.** NORC will issue all pilot test researchers customized thin-client machines from which they will remotely connect to the enclave.



ISPOR 2011 WS

12

## Process (cont.)

- ❑ **Data Access.** We will be able to compare and contrast the relative advantages and disadvantages of different subsetting methodologies.
- ❑ **Advanced Computational Engine.** Enclave test researchers will have access to an Advanced Computational Engine, providing high speed computational power for data aggregation, exploration and visualization.
- ❑ **Statistical and Other Online Discovery Tools.** All test researchers will have access to a suite of statistical applications (e.g., SAS, Stata, SPSS, Limdep, "R", Matlab, etc.) the MS Office Suite, a customized computational engine and collaboration tools.
- ❑ **Statistical Disclosure Control.** No output will be released to the public until reviewed and approved by NORC statisticians.

ISPOR 2011 WS



13

## Timeline

- ❑ Q3 2011: Researcher recruitment, Data Deposit and Processing
- ❑ Q4 2011: Review Applications, Distribute Access Nodes, Deliver Advanced Computational Engine and other tools, Provide Access
- ❑ Q1-Q2 2012: Support Researchers, Provide Statistical Disclosure Control

ISPOR 2011 WS



14