The Secure Medical Research Workspace: An IT Infrastructure to Enable Secure Research on Clinical Data

Michael Shoffner, B.A.[1], Phillips Owen, B.S.[2], Javed Mostafa, Ph.D.[3], Brent Lamm, B.S.[4], Xiaoshu Wang, Ph.D.[2], Charles P. Schmitt, Ph.D.[2], and Stanley C. Ahalt, Ph.D.[5]

Abstract
Clinical data have tremendous value for translational research, but only if security and privacy concerns can be addressed satisfactorily. A collaboration of clinical and informatics teams, including RENCI, NC TraCS, UNC's School of Information and Library Science, Information Technology Service's Research Computing and other partners at the University of North Carolina at Chapel Hill have developed a system called the Secure Medical Research Workspace (SMRW) that enables researchers to use clinical data securely for research. SMRW significantly minimizes the risk presented when using identified clinical data, thereby protecting patients, researchers, and institutions associated with the data. The SMRW is built on a novel combination of virtualization and data leakage protection and can be combined with other protection methodologies and scaled to production levels. Clin Trans Sci 2013; Volume 6: 222–225

Keywords: protected health information (PHI), security, cybersecurity, HIPAA, HITECH, electronic health records (EHR), patient data, virtualization, data leakage protection (DLP), Institutional Research Board (IRB), computer security, systems analysis, medical informatics

Introduction

To support clinical and translational research, the University of North Carolina at Chapel Hill (UNC) has established the Carolina Data Warehouse for Health (CDW-H), a clinical data warehouse holding data continuously pulled from the Electronic Health Record system for the UNC HealthCare System.[1] The CDW-H contains Protected Health Information (PHI) from patients, including lab results, physician notes, diagnoses, and other sensitive medical information. It is federated with several other data systems at UNC that contain valuable and sensitive patient data, such as the Lineberger Cancer Registry being built and maintained by the UNC Lineberger Comprehensive Cancer Center. NC TraCS, the holder of UNC's National Institutes of Health (NIH) Clinical and Translational Science Award (CTSA), maintains the data warehouse and provides researchers with access to and help with using it. The CDW-H houses almost two billion rows of data, and on a typical day, more than one million data elements are added to it. The data that reside there represent an incredible opportunity for research. To date, over 700 studies have been based on the data provided by the CDW-H.

NC TraCS is tasked with both enabling and promoting research on data containing PHI and with ensuring such data are used only in accordance with University and federal policies. A critical concern in this dual role is ensuring that data provisioned to researchers remain secure, yet easy to use. Prior solutions, such as provisioning data to researchers through encrypted email or via a secured Network Attached Storage disk space, were deemed insecure as there was no control over what researchers did with the data after these were provisioned; e.g., nothing prevented researchers from emailing data to external systems to do work off-hours or from accidentally leaving data in nonsecure locations. To address this concern, NC TraCS, the Renaissance Computing Institute (RENCI), UNC School of Information and Library Science, and UNC Information Technology Services (ITS) prototyped and deployed a production framework called Secure Medical Research Workspace (SMRW). In this paper, we highlight the benefits and weaknesses of this system.

Methods

Three main factors drove the design for the SMRW. First, an analysis of Institutional Review Board (IRB)-mandated security requirements at UNC classified studies into Levels I, II, and III, with Level III reflecting the highest security needs. This analysis indicated that 27% of the 4,210 studies tabulated would require a Level III solution. Second, recognition from interviews with data security vendors and from reports on data breaches (see *Table 1*) that data disclosures by internal personnel are a dominant risk vector, especially within medicine, but that externally originating hacks are also important. These results suggest that it is as important to keep institutional personnel from unintentionally allowing data to leave the medical research enterprise as it is to keep hackers from pulling data out of the enterprise. Third, from discussions with researchers and research analysts within UNC's CTSA, it was decided that a high priority was to find a solution that allows researchers to work in a "natural environment," that is, an environment that they are accustomed to.

With the above drivers as context, the team performed a lightweight formal requirements analysis to arrive at the functional specifications for the system. We interviewed a set of researchers and institutional stakeholders and identified a set of eight use case actors (roles) in three classes (user, user assistant, and administrator) performing a total of 12 use cases. *Figure 1* shows a Unified Modeling Language diagram[2] of the use cases performed by the Researcher actor/role. From these use cases, the team derived a list of 15 functional "business requirements" for the system. Once the SMRW system architecture was finalized these business requirements were used to generate a set of evaluation criteria for vendor products.[3] When the system prototype was ready, it was tested with NC TraCS research analysts and selected target users for the purpose of improving usability where possible.

SMRW was under development for roughly 2 years. The initial phase consisted of working with users and stakeholders and assessment of preexisting components, with the development team writing prototype code for aspects of the solution that were unavailable off the shelf. The second phase involved selecting

[1]Renaissance Computing Institute and School of Information and Library Science; [2]Renaissance Computing Institute; [3]School of Information and Library Science; [4]NC TraCS Institute; [5]Biomedical Informatics, NC TraCS Institute and Department of Computer Science, University of North Carolina, Chapel Hill, North Carolina, USA.

Correspondence: Michael Shoffner (shoffner@renci.org)

DOI: 10.1111/cts.12060

Percentage of breaches	DISC	HACK	INSD	PHYS	PORT	STAT
EDU	30.42%	32.17%	1.75%	6.35%	21.44%	7.88%
MED	14.01%	5.10%	14.97%	14.33%	42.36%	9.24%

Table 1. Results pulled from the Privacy Rights Clearinghouse, a grant-supported nonprofit corporation. The data include publically documented breaches from 2005 to 2012 in the United States classified the type of entity and the type of breach, and the number of records breached. The types of breaches include: DISC (unintended disclosure), HACK (hacking or malware), CARD (debit/credit card fraud), INSD (breach made by someone with legitimate access), PHYS (lost, discarded, or stolen nonelectronic data), PORT (lost, discarded, or stolen physical device), STAT (lost, discarded, stolen nonportable electronic device), and UNKN (unknown).

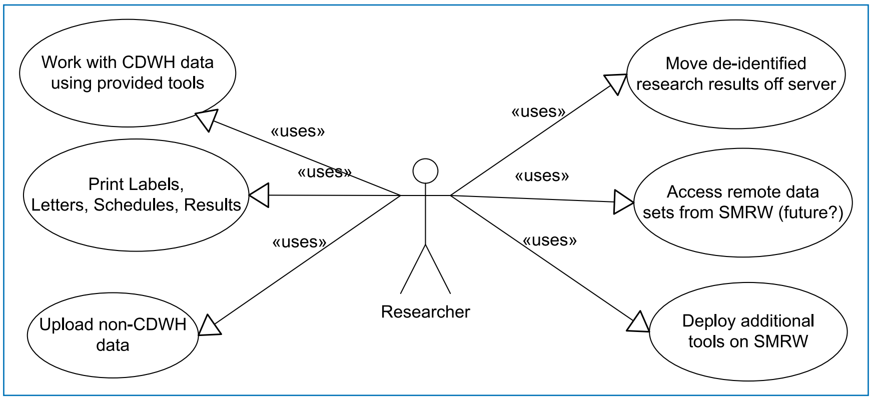
Work with CDWH data using provided tools
Print Labels, Letters, Schedules, Results
Upload non-CDWH data
«uses»
«uses»
«uses»
Researcher
«uses»
«uses»
«uses»
Move de-identified research results off server
Access remote data sets from SMRW (future?)
Deploy additional tools on SMRW

Figure 1. Researcher use cases for SMRW.

vendors, integrating technologies, and ensuring that all feature requirements could be met by selected vendors. The SMRW is now in production use at UNC.

Results

Figure 2 shows the conceptual high-level design of the SMRW. In this design, a researcher requests data from the CDW-H by contacting an NC TraCS research analyst, who issues a private SMRW "workspace" to the researcher. The researcher logs into the workspace from her local computer over a virtual private network connection. The workspace is a virtual remote desktop computer with analysis tools installed and a mounted drive that contains data requested from CDW-H. The researcher can bring any additional necessary data from her local computer into the workspace; however, if she attempts to take data out of the workspace, the SMRW system may either allow the transfer, stop the transfer, stop the transfer if the data meet certain conditions (e.g., contain social security numbers), or challenge the transfer based upon policy settings applied to that research project. Furthermore, any attempted data movement is also logged for future audit. *Figure 3* shows the prototypical platform workflow from the point of view of the researcher.

SMRW's administrative interface provides mechanisms to change data transfer policies and make use of audit capabilities. Several policies and technologies are integrated into the base system to ensure the data are kept secured and that researchers can easily work with their data. In the following, we review critical aspects under the assumption that the SMRW is operated with administrative, networking, and physical security typical of a modern data center (e.g., locked doors, firewalls). Note that this assumption is an integral "ingredient" in the overall security of the system.

Desktop virtualization

Researchers are required to use a centrally administered virtual machine (VM), which constitutes the workspace, to access sensitive data provisioned to them from the data warehouse. The workspaces contain preinstalled user-oriented software, for example R, SPLUS, and SAS, and preinstalled administrative software including antimalware software. Directory and policy services are used to further enforce university and IRB policies, e.g., what kinds of software can be run and whether administrative access is allowed on the workspace. NC TraCS provisions sensitive data from the data warehouse to disk space that is mounted as a file system folder on the researcher's workspace. Users are also allowed to mount local resources, such as local hard drives and printers, to the workspace when they connect to it. This ensures that users can freely copy data from their machines to their workspace (this ability can be turned off to protect against database linkage attacks if they are considered a potential problem). To reduce the possibility of introducing malware into the workspace, researchers are, by default, not permitted to install software (although this can be allowed by policy). Workspaces also update antimalware software and are configured to update the operating system automatically.

Data leakage protection

Each workspace is protected by Data Leakage Protection (DLP) software.[4] DLP comprises a suite of technology components that can control what data are allowed to leave a specific computer ("endpoint protection"), what data can be transferred over a network ("network protection"), and what data are stored within a set of computer resources ("discovery"). A key capability for the SMRW is the ability to adjust the policy for removing data on a project-by-project basis, with the following configuration choices possible:

(1) Disallow any removal of data.
(2) Allow removal of any data.
(3) Allow removal of data only if it passes a scan of the data (e.g., for PHI data elements).
(4) Allow removal of data only if the researcher acknowledges permission to do so.
(5) Allow removal of data only after receiving permission from an administrator.

We believe choice no. 4 represents a balance between providing researchers with easy access to their data while still ensuring no data egress without an acknowledgment of responsibility. As such,

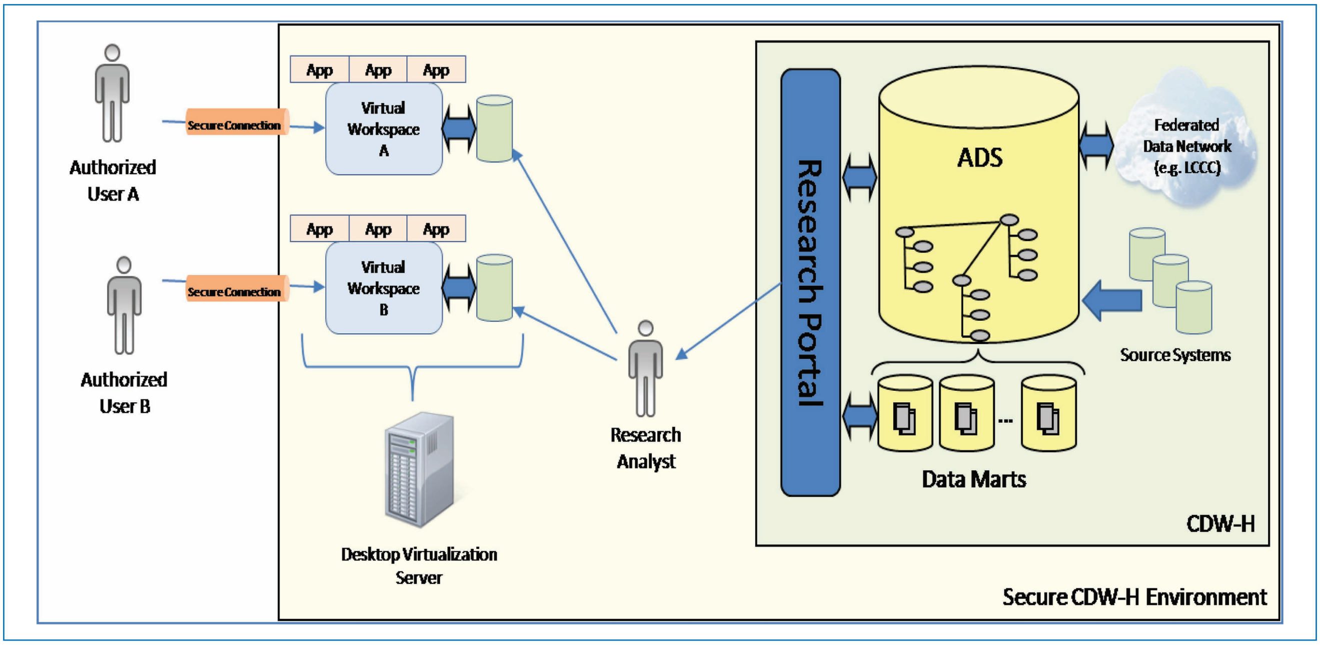

Figure 2. Conceptual view of the SMRW environment.

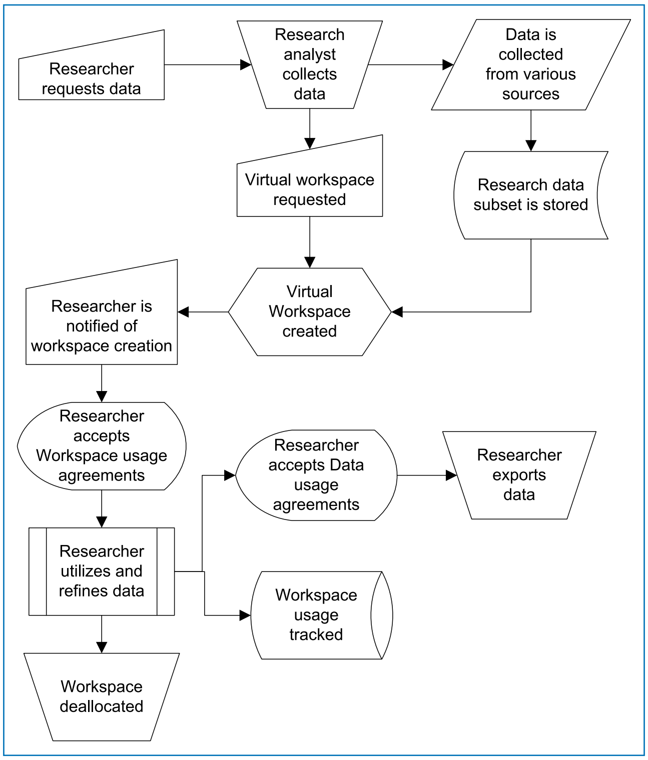
Researcher requests data
Research analyst collects data
Data is collected from various sources
Virtual workspace requested
Research data subset is stored
Researcher is notified of workspace creation
Virtual Workspace created
Researcher accepts Workspace usage agreements
Researcher accepts Data usage agreements
Researcher exports data
Researcher utilizes and refines data
Workspace usage tracked
Workspace deallocated

Figure 3. Workflow showing SMRW platform's typical (researcher) usage scenario.

in reviewing data leakage solutions, the ability to interrupt an attempt to remove data, present the researcher with a data use agreement, and require the researcher to actively acknowledge data removal was a key feature. The data leakage endpoint protection technology component allows the researcher to instantly give an acknowledgment.

For the SMRW, DLP client endpoint protection is present on each workspace provisioned to researchers and a central DLP server is used to administer all clients. Several additional factors were considered in assessing data leakage technology, including: (1) ability to control data being removed from the computer via different mechanisms (e.g., file copy, Instant Messenger, Web forms, email); (2) ability to define approaches used to scan the data (e.g., regular expression matching, file types, data locations, file signature); (3) ability to log all attempts to remove data, including logging of the data removed, as well as availability of additional forensics capabilities; (4) easy to use management software; and (5) correctness under a variety of test cases (e.g., whether copying worked correctly when the user cancelled the copy after being challenged with the data use agreement, a scenario that more than one vendor solution did not pass).

Technical implementation

The SMRW architecture is realized by selecting and deploying vendor products, configuring networking infrastructure, and integrating against the institution's IT services as relevant.[5] In UNC's case, virtualization capabilities are provided by VMware in a clustered configuration. All researcher workspace VMs are behind the UNC firewall to protect against external network-based attacks such as port scanning; the SMRW network is partitioned into three Virtual Local Area Networks, with one containing researcher workspaces and two containing infrastructure services such as a file server and the DLP server node. A proxy server and Remote Desktop Gateway mediate interaction between the workspaces and the Internet. UNC's institution-wide Active Directory infrastructure provides authentication as well as authorization policies enforced at the researchers' workspace instances through Windows General Policy Objects. Each workspace also uses Symantec antivirus as provided by UNC and uses UNC's servers for Windows operating system updates. Websense's DLP product is used for the system's DLP capabilities, including policy management and auditing of data access patterns on a per user basis.

Discussion

Evaluation

The two most crucial requirements of the SMRW environment are that it prevents data leakage (in accordance with policy settings) and that it logs attempts to move data. The SMRW team therefore tested the system for interdiction of data movement across monitored channels. These tests showed 100% efficacy when the system was configured to interdict all outbound file transfers. Preliminary testing also showed excellent recognition of test spreadsheets and files that contained mocked up PHI data, although we did not create a test suite with comprehensive coverage of all possible types of PHI data. In addition, we tested the logging capabilities of the system and found them to be satisfactory. These tests were sufficient for our purposes although additional testing could be undertaken for increased system hardening.

Benefits

The SMRW and other somewhat related approaches[6,7] have several benefits over alternative approaches of which we are aware. First, distributing sensitive data using mechanisms that do not provide accountability or policy-based control (e.g., distributing data using "flash drives") is a possible alternative, although it is clearly the riskiest. An alternative is to use encryption "at rest" and "in flight," which requires encryption, and encryption verification, to be present at all endpoints. While this solution provides a greater degree of security, it still allows sensitive data to be exposed if an encrypted device is commandeered while "open," that is, when it is being used to analyze data in an unencrypted form. The major DLP-based alternative to our system is to use the DLP endpoint on the end user devices themselves (i.e., on the laptop or desktop of the researcher). Our analysis determined that this was, from an institutional perspective, difficult and presented additional administration burdens and licensing costs. Furthermore, this impedes researchers because it "locks down" their computers. It also puts the control point at the wrong place in the system—sensitive data still migrate out to nodes when the preference of the institution is to keep it in the data center where it is safer and easier to manage.

A vendor Virtual Desktop Infrastructure (VDI) could also be used to supply the researcher VMs used in the system. VDI provides each user with a consistent desktop experience across different underlying VMs used to support user sessions. The VDI option did not satisfy a cost/benefit analysis for UNC's deployment, but may be an option in other settings.

Disadvantages

While there are clearly benefits to utilizing a solution like the SMRW, there are also drawbacks that should be considered. First, the cost of the infrastructure must be weighed. SMRW setup costs include: (1) hardware for virtualization, (2) personnel time to set up virtualization, (3) integration with institutional identity management systems (e.g., Windows Active Directory), (4) setup of the networking and firewall environment, (5) licensing costs for virtualization and (6) DLP products, respectively. Ongoing personnel time is also required for systems administration, monitoring DLP usage, and setting up and supporting researchers on the system. Second, this architecture places additional constraints on researchers in comparison with simply handing them the (possibly encrypted) data they require.

Conclusion

While SMRW has been designed to be as frictionless as possible for researchers to use, lower convenience is an inescapable trade-off. Two issues arise from this point. For a system such as SMRW to be successful from an institutional perspective, it must be widely used, and thus needs to be viewed as the only sanctioned method of sensitive data access. Consequently, it should be very easy to use and present minimal barriers to even relatively unsophisticated computer users. We plan to issue further reports on user experiences, as well as data from broad use statistics, in a subsequent paper.

Acknowledgments

We would like to thank UNC School of Medicine, UNC Hospitals, Stephen Galla, Ken Langley, Casey Averill, Erik Scott, Fabian Monrose, and Hye-Chung Kum. This work was supported by funds from RENCI, a research institute jointly sponsored by UNC, Duke University, and North Carolina State University, and by the UNC ITS Division. The project was also supported by the NC TraCS Institute, the recipient of UNC's NIH CTSA via support by the National Center for Research Resources and the National Center for Advancing Translational Sciences, NIH, through Grant Award Number UL1TR000083. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Mostafa J, Moore C. The North Carolina Translational and Clinical Sciences Institute: activities of the Biomedical Informatics Core. *Clin Transl Sci.* 2010; 3(3): 71–72.

2. Booch, G. Software Architecture and the UML. UML World Keynote Speech, 1999. http://researcher.watson.ibm.com/researcher/view_pubs.php?person=us-gbooch&t=1. Accessed January 2, 2013.

3. Owen P, Shoffner M, Wang X, Schmitt C, Lamm B, Mostafa J. Technical Report TR-11–01, Secure Medical Research Workspace, February 2011.

4. Gartner: Magic Quadrant for Content-Aware Data Loss Prevention. http://www.gartner.com/id=1379314. Accessed January 2, 2013.

5. Shoffner, M, Mostafa, J. Secure Medical Research Workspace. Poster presented at: CTSA Informatics KFC 2012 Annual Meeting; November 7–8, 2012; Chicago.

6. MyResearch@UCSF. http://it.ucsf.edu/projects/myresearch-0. Accessed January 2, 2013.

7. Bradford W, Hurdle J, LaSalle B, Facelli J. Development of a Protected Computer Environment for Translational Research Data and Analytics. Poster presented at: CTSA Informatics KFC 2012 Annual Meeting; November 7–8, 2012; Chicago: University of Utah.