EDM Forum
EDM Forum Community

eGEMs (Generating Evidence & Methods to improve patient outcomes)

Publish

8-8-2016

Framework for Deploying a Virtualized Computing Environment for Collaborative and Secure Data Analytics

Adrian Meyer
UNC Lineberger Comprehensive Cancer Center, adrian.meyer@unc.edu

Laura Green
UNC Lineberger Comprehensive Cancer Center, legreen@email.unc.edu

Ciearro Faulk
UNC Lineberger Comprehensive Cancer Center, ciearro.faulk@unc.edu

Stephen Galla
UNC Research Computing, Information Technology Services, stephen_galla@unc.edu

See next pages for additional authors

Recommended Citation

Meyer, Adrian; Green, Laura; Faulk, Ciearro; Galla, Stephen; and Meyer, Anne-Marie (2016) "Framework for Deploying a Virtualized Computing Environment for Collaborative and Secure Data Analytics," *eGEMs (Generating Evidence & Methods to improve patient outcomes)*: Vol. 4: Iss. 3, Article 4.
DOI: http://dx.doi.org/10.13063/2327-9214.1224
Available at: http://repository.edm-forum.org/egems/vol4/iss3/4

The Electronic Data Methods (EDM) Forum is supported by the Agency for Healthcare Research and Quality (AHRQ), Grant 1U18HS022789-01. eGEMs publications do not reflect the official views of AHRQ or the United States Department of Health and Human Services.

Framework for Deploying a Virtualized Computing Environment for Collaborative and Secure Data Analytics

Abstract

Introduction: Large amounts of health data generated by a wide range of health care applications across a variety of systems have the potential to offer valuable insight into populations and health care systems, but robust and secure computing and analytic systems are required to leverage this information.

Framework: We discuss our experiences deploying a Secure Data Analysis Platform (SeDAP), and provide a framework to plan, build and deploy a virtual desktop infrastructure (VDI) to enable innovation, collaboration and operate within academic funding structures. It outlines 6 core components: Security, Ease of Access, Performance, Cost, Tools, and Training.

Conclusion: A platform like SeDAP is not simply successful through technical excellence and performance. Its adoption is dependent on a collaborative environment where researchers and users plan and evaluate the requirements of all aspects.

Acknowledgements

Work on this study was supported by the Integrated Cancer Information and Surveillance System (ICISS), UNC Lineberger Comprehensive Cancer Center with funding provided by the University Cancer Research Fund (UCRF) via the State of North Carolina.

Keywords

Health Information Technology, System Architecture, Health Informatics, Virtual Computing, Virtual Desktop Infrastructure, VDI, Data Security

Disciplines

Computer and Systems Architecture | Health Information Technology | Health Services Research | Public Health

Authors

Adrian Meyer, *UNC Lineberger Comprehensive Cancer Center*; Laura Green, *UNC Lineberger Comprehensive Cancer Center*; Cieарro Faulk, *UNC Lineberger Comprehensive Cancer Center*; Stephen Galla, *UNC Research Computing, Information Technology Services*; Anne-Marie Meyer, *UNC Gillings School of Public Health, Epidemiology*.

eGEMs
Generating Evidence & Methods
to improve patient outcomes

Framework for Deploying a Virtualized Computing Environment for Collaborative and Secure Data Analytics

Adrian Meyer, MS;[i] Laura Green, MBA;[i] Ciearro Faulk, BS;[i] Stephen Galla, MBA;[ii] Anne-Marie Meyer, PhD[iii]

ABSTRACT

Introduction: Large amounts of health data generated by a wide range of health care applications across a variety of systems have the potential to offer valuable insight into populations and health care systems, but robust and secure computing and analytic systems are required to leverage this information.

Framework: We discuss our experiences deploying a Secure Data Analysis Platform (SeDAP), and provide a framework to plan, build and deploy a virtual desktop infrastructure (VDI) to enable innovation, collaboration and operate within academic funding structures. It outlines 6 core components: Security, Ease of Access, Performance, Cost, Tools, and Training.

Conclusion: A platform like SeDAP is not simply successful through technical excellence and performance. It's adoption is dependent on a collaborative environment where researchers and users plan and evaluate the requirements of all aspects.

[i]UNC Lineberger Comprehensive Cancer Center, [ii]UNC Research Computing, Information Technology Services, [iii]UNC Gillings School of Public Health, Epidemiology

Introduction

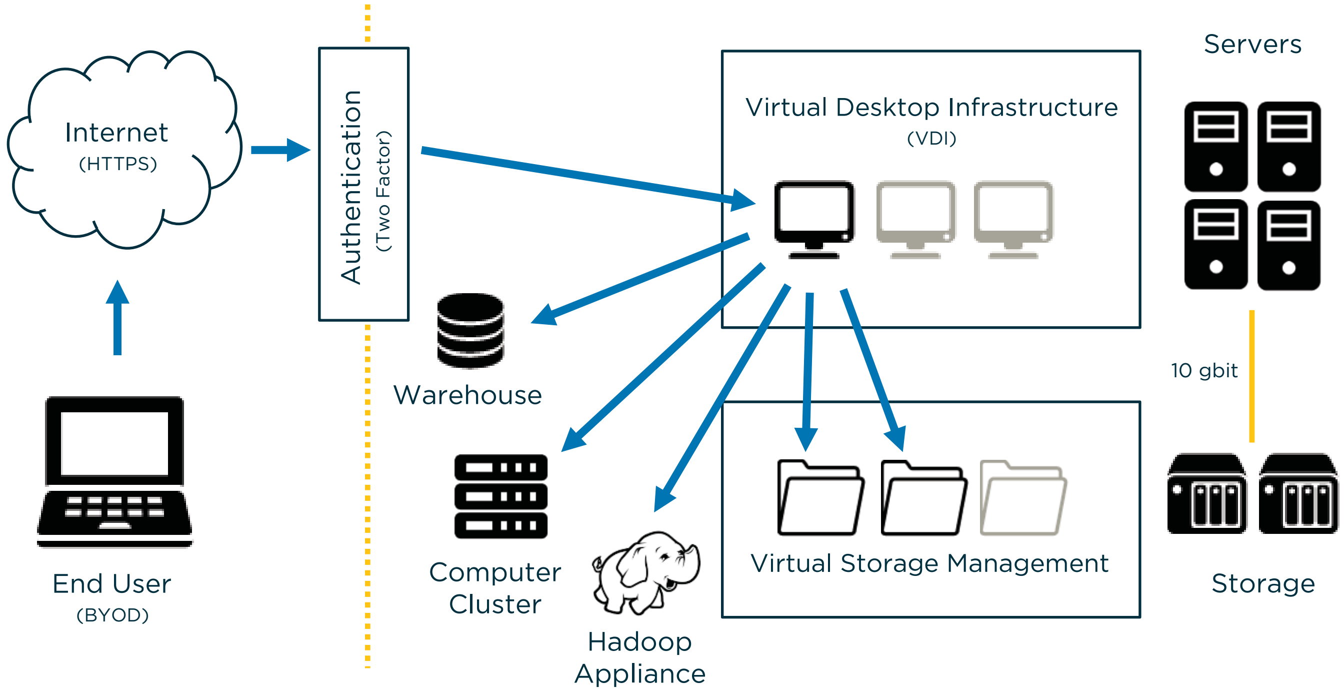
In this digital age, large amounts of valuable data are being generated by a wide range of health care applications across diverse systems. These data have the potential to offer tremendous insight into populations and health care systems. However, robust secure computing and analytic systems are required to leverage accumulated information. Multidisciplinary teams with diverse analytic expertise are required to analyze these data. These diverse users often have specific preferences for analytic tools that create endless permutations of potential system configurations.

This framework outlines our experience in deploying a Secure Data Analysis Platform (SeDAP), and the process of planning and designing a secure environment enabling access to the data from anywhere at any time. The framework addresses the challenges of complying with security requirements in a rapidly changing landscape of data privacy and regulations surrounding patient data.[1,2,3] While third-party cloud computing services offer great scalability and cost flexibility, building and operating a dedicated environment leverages existing organizational resources and provides complete control over requirements.

System Architecture

A virtualized computing environment provides the backbone to an adaptive and scalable solution. Virtualization is implemented on all possible components within the architecture. Traditional environments such as data warehouses, computing clusters, or other appliances are made available through the Virtual Desktop Infrastructure (VDI).

Figure 1. System Architecture



Objectives

The Secure Data Analysis Platform (SeDAP) is the result of a bottom-up, user-centric development, deployment, and maintenance approach taken by the Integrated Cancer Information and Surveillance System (ICISS) at the Lineberger Comprehensive Cancer Center for the University of North Carolina at Chapel Hill (UNC-CH) in conjunction with the UNC-CH Information Technology Services Research Consulting (ITS-RC) group. The SeDAP was built as a replacement for a dedicated Linux based server system and was designed with the following main objectives and user requirements:

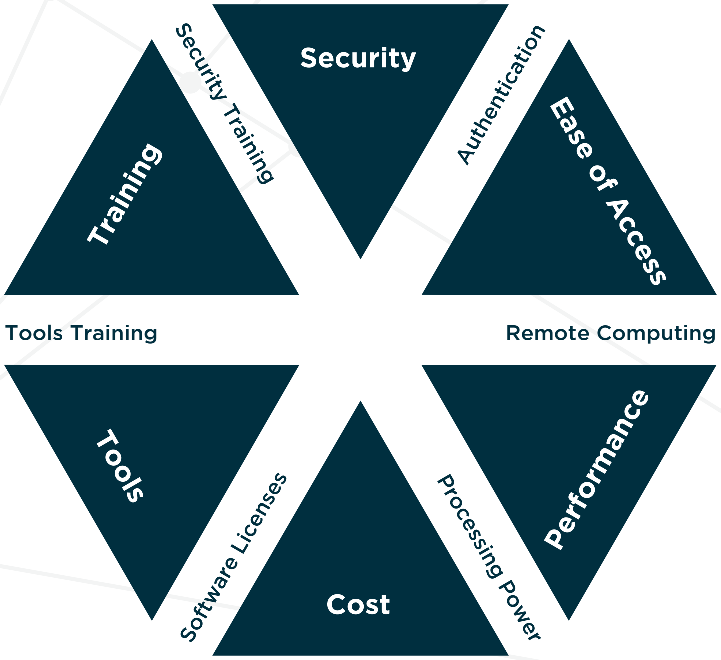
- Increased Usability: a modern graphic user interface (GUI)-based environment to get users up and running quicker;
- Increased Accessibility: secure and flexible remote access while supporting Bring Your Own Device (BYOD) where applicable; and
- Increased Security: closed and controlled environment with single common security management plan.

Framework Overview

A solid framework for deploying such an environment guides and accelerates the planning, design, and implementation stages while addressing six core components: (1) Security, (2) Ease of Access, (3) Performance, (4) Cost, (5) Tools, and (6) Training.

ICISS currently provides big data resources for clinical and population health research including comparative effectiveness research (CER) and outcomes research. The ICISS core data repository is composed of North Carolina Central Cancer Registry tumor data linked to public and private insurance claims data. Additional linkages to other patient registries, cohort studies, genomic, and clinical data provide additional layers of patient data.[4]

Figure 2. The Six Framework Components



The ICISS team includes clinicians, biostatisticians, epidemiologists, health services researchers, demographers, geographers, and computer scientists, trained at the PhD and master's degree levels. It represents the multidisciplinary "team science" approach needed to optimally leverage secondary data, as well as a large diversity in informatics technology expertise and user requirements.[5] When a legacy system expired, this multidisciplinary team environment combined with large data sets encompassing clinical PHI offered a unique opportunity to gather specific user requirements and test system optimization in order to develop the proposed framework.

Framework Components

Security

Security requirements for any system are directly connected to the data requirements and the data source. Most state and federal governments have enacted specific privacy protection laws to address personal and business information other common sources of research data such as health care organizations, government entities, or academic institutions that require implementation of specific security requirements. In the United States, these requirements might include, but are not limited to, the Federal Information Security Management Act (FISMA), Health Insurance Portability and Accountability Act (HIPAA), and Family Educational Rights and Privacy Act (FERPA).

Implemented security controls will directly affect operational procedures.[6] Therefore, careful planning is required to enumerate the types of data that will be analyzed over the lifetime of the system and to identify the most stringent data security requirements.

In SeDAP, data access is managed through authorization to folders by assigning one of four different access rules:

1. Role-based access, such as system administration or data management, is assigned to users based on task-specific responsibilities irrespective of project or data.
2. Institutional Review Board (IRB) and study-specific access is assigned to all staff specifically assigned to an active study and leverages the existing academic institutional IRB approval systems.
3. Data Use Agreement (DUA) access is assigned to the few individuals who require access to an entire "raw" data set, or who subset data for others and are involved in merging and linking data.
4. Team level access is used for exploratory work and collaboration within multidisciplinary settings.

A global security management plan outlining these access rules was developed and is compliant with all regulatory requirements associated with executed DUAs as well as institutional, state, and federal rules. This written plan helped to build trust with the data suppliers and provides the baseline for contractual agreements. Within our institution we were able to leverage global security controls such as common user IDs and network security, while others controls had to be developed from the ground up.

Ease of Access

Security and "Ease of Access" intersect through two-factor authentication. The infrastructure allows users to connect through dedicated remote software to virtual desktops such as Microsoft Windows or Linux. This approach allows a Bring Your Own Device (BYOD) model and accessibility from the wide variety of devices within the academic workplace. The environment allows users to disconnect from a virtual desktop from device A and reconnect from device B without disrupting jobs and analyses.

This model provides flexibility for users and projects while the configuration of SeDAP facilitates standardization of key scientific process such as project tracking, data structures, access controls, and documentation. SeDAP configuration has helped build a culture of transparency for an academic environment that requires replication of tasks and analyses across data samples. For research on large, disparate health care data sets, it has reduced duplication of effort and has increased efficiency as users are trained in the same way and have access to the same sets of tools and to a systematic platform for sharing programming code and other information.

A specific benefit of a VDI is that a computing job requiring extensive time will not be interrupted when the user disconnects. This is particularly important for complex queries or simulations using large data sets that may take days to run. The virtual machine inside the VDI continues to run and process any or parallel jobs that have been started. Users can check running jobs from another device. We further implemented a messaging system that can transmit messages from within the secure environment to a web portal and that notifies users via email of changes in job status. This allows them to know when jobs have completed or been interrupted without having to log into the computing environment. This messaging system has significantly helped with efficiency regarding large jobs that take days to run but may unexpectedly be stopped due to software failure, allotted disk space, or even model parameterization failure.

Performance

Before implementing the system, the performance requirements were assessed with a set of benchmarks that measured the existing computing resources available to the users. Table 1 lists the compared systems. The Disk I/O performance was measured at the end user level, meaning an average of the performance on top of the operating system. This number represents the average performance an analytic application has access to. Since our data size requirements exceed 10 terabytes, a Solid State Disk (SSD) setup for Laptops or Desktops was not measured.

Table 1. Benchmark Systems and Hard Disk Performance

	OPERATING SYSTEM	CPU	MEMORY (RAM)	DISK I/O
Laptop	Windows 7, 64bit	1 Socket, 4 Cores, 8 Threads	2 Slots, 8 GB	**7200 RPM Laptop Drive** Read: 54 MBs Write: 52 MBs
Desktop	Windows 7, 64bit	1 Socket, 2 Cores, 2 Threads	2 of 4 Slots, 4 GB	**10k RPM Desktop Drive** Read: 113.80 MBs Write: 110.30 MBs
Server	RHEL6, 64bit	2 Socket, 8 Cores, 16 Threads	72 GB	**Network Attached Storage (NAS, 1gbit)** Read: 67 MBs Write: 58 MBs
VDI	Windows 7, 64bit	1 Socket, 2 Cores, 2 Threads	1 Slot, 4 GB	**Fiber Attached Storage (10gbit)** Read: 392.73 MBs Write: 465.63 MBs

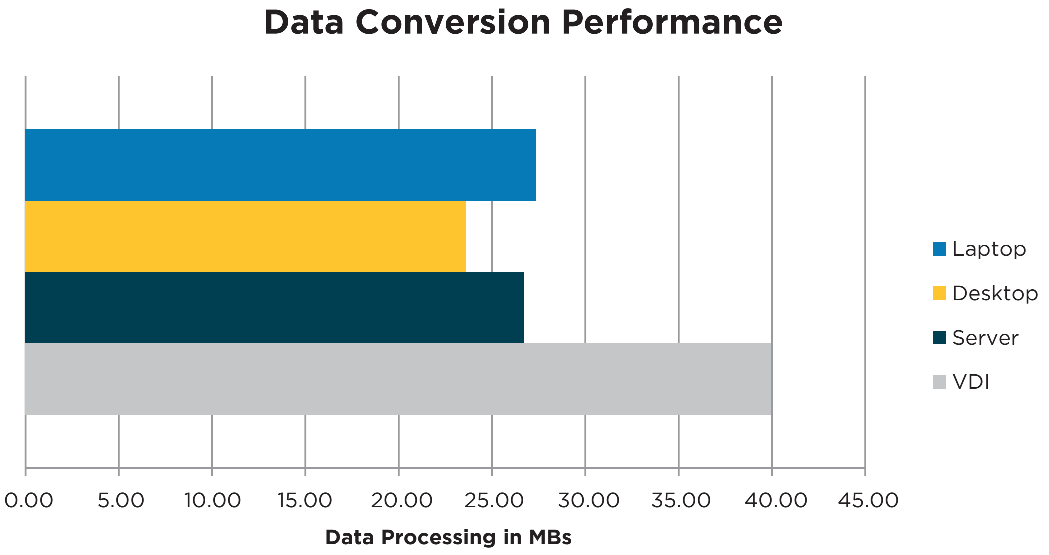
The outcomes of the data conversion benchmark (Figure 3) show that increasing I/O performance significantly affects the data processing performance. The ICISS daily workflow often requires scanning and creating more than 100 GB in a single analytical step. To reach high performance read and write rates of over 300 megabytes per second, the storage appliance is connected through a dedicated 10 gigabit optical line. Additionally, advanced data caching and buffering is implemented to absorb the high demand of data. The resulting VDI environment performs on data operations at a level of a high performance desktop environment with SSD.

During the first two years of operation, we experienced a shift in the performance bottleneck. With users no longer having to waiting for larger jobs to finish, the threshold of the data size per job increased. Inefficiencies in programming code and data management amplified this effect. We were able to address some of these with training and implementing best practices at the level of individual users.

Cost

Entertaining the concept of building a large platform can be daunting. Finding the right balance between performance and cost is primarily defined by processing power requirements. Costs add up quickly, and making a business case to replace a running legacy system with a newer and more expensive system can be hard to sell to stakeholders who may not understand the technical details or requirements of the team. We adopted a bottom-up approach to cost estimation and system budgeting by first calculating the costs of a traditional analytic workstation. Basing our estimate on the price of a high performance workstation with a cost of $4,500 and a three year warranty, we elected to use a yearly $1,500 per user as a rough baseline for hardware. The resulting aggregate of 50 users for four years resulted in $300,000 for infrastructure. By working closely with UNC ITS-RC, we were able to identify opportunities for the SeDAP to leverage existing campus infrastructure such as server room space, networks, power supply and backup, technicians, and operational staff.

Figure 3. Performance Results from Data Conversion CSV to Native SAS Files



Costs associated with storage comprise the biggest part of total system cost. These costs vary greatly depending on demands. Slower disks based on serial AT attachment (SATA) technology are considerably cheaper than high performing solid state drives (SSD). Using cloud storage solutions might also be an option as long as all the security considerations are met. We elected to operationalize both cheaper "slow" and a more expensive "fast" storage. Data that is in active analysis and temporary caching will use the "fast" environment while less performance-dependent information uses the "slow" space.

Costs associated with operationalizing the platform for users are another factor that is less easily estimated but cannot be overlooked. Account creating, access provisioning, troubleshooting, interfacing with representatives from ITS, and user training are critical activities that cost staff valuable time. This staff support time is also not easily recuperated or justified with extramural funding in an academic environment. Yet these activities and staff support are essential to transform the environment into a true solution for users, and the costs associated with delivery of these services must be incorporated. We learned that key aspects of personnel and staff cost were the identifying of and teaming up with existing organizational resources. In our case, we closely tracked hours spent by our service provider in these activities for each "seat" created in the first four months of system deployment (to account for the learning curve associated with new processes), and calculated the proportion of the employee's fully burdened salary that this time represented. Each research study utilizing the platform was then assessed a fee equivalent with these costs. The nature of this fee is dependent on funding type: for those funding vehicles permitting computing costs, a flat fee was assessed. For funding mechanisms that only fund personnel, a fraction of the service provider's salary was included in the budget of each proposal.

Tools

Licenses for a virtual platform and end user software can also be very expensive and contribute to overall and maintenance costs. We were able to leverage "educational institution" licenses, but to avoid overspending on these licenses user input was directly sought to limit the selection of tools within the SeDAP. A list of tools currently used in the legacy computing solution was expanded to include tools requested by users based on their existing skill set and planned professional development. Once the initial list was finalized, we were able to take advantage of institutional discounts and bundling of licenses to provide a wide variety of tools at reduced cost. Some of these tools include SAS, STATA, R, ArcGIS, QGis, Mathematica, Pearl, PHP, SQLite, LaTeX, and Oracle SQL Developer.

The tools are preinstalled in two different Microsoft Windows base images supported by the VDI. One image is intended for use of data containing identifiers and is focused on data management. Tools installed in this image are intended to support tasks including file decrypting and encrypting, packing and unpacking of archives, export-transform-load (ETL), database loading, and file conversion. The second image is focused on analytics. These tools support statistical analysis (including simulation), geospatial analysis, network analysis, and documentation tools. To guarantee rapid deployment an image is assigned to the new user. At the point of authentication, a copy of the image is prepared unless the user had an active image from a previous connection. All the tools are immediately available to the user upon initial log in. An alternate model would be to build an image for each research project. This configuration works well if physical separation of space is required, but it exponentially increases the administrative work to maintain updates.

Traditional computing cluster, data warehouses, or modern appliances can also easily be integrated into a VDI. Connections or tunnels can easily be established after security risk considerations. The SeDAP team currently manages tunnels to messaging services and a relational data warehouse.

Training

Adequate and continued education of the users increases both efficiency and security.

Tools Training: The configuration of commonly used tools such as SAS, STATA and R likely differs from the "out of the box" setup. The adjustments within the VDI can be specified to optimize performance based on user requirements. Having users continuously involved with defining best practices of specific features of these tools creates a valuable knowledge library. To this end, we actively solicit feedback from user groups and facilitate communication in order to share their experiences and improve performance. This group is the primary source of new requirements to alter configurations or add new tools. The user community also maintains a web-based FAQ repository to retain and share knowledge.

Security Training: As part of the global security management plan, our administrative team provides security training for all users before they are granted access. A signed workforce authorization serves as an internal contract and clearly defines the understanding of rules and sanctions. We also leverage and track required training imposed by the academic institution and the health care system. For example, the school of medicine requires continuous HIPAA training for all employees. Research personnel at the university are required to complete an ethics training in order to participate in research requiring IRB review. Despite these institutional requirements, best practices or curriculum regarding data security have not been integrated into departmental curriculums. In an era of increasing data security concerns, requiring in-person training provides a critical teaching opportunity.

Conclusion/Next Steps

Planning, financing, building, and operating a secure computing and analytic system can be a long process. In an academic environment, this also requires collaboration between diverse multidisciplinary stakeholders who may not usually interface or have the same priorities. It is always challenging for stakeholders and customers to remain objective and evaluate whether a technical platform will meet the requirements, especially while commercial vendors are promoting products as being optimal and complete solutions. It is important not to fall into a "if we build it, they will come" fallacy, and instead to build with careful collaboration and extensive discussions with users. This way, investments can be maximized and expectations are clearly communicated. An important aspect of understanding our user expectations was the implementing of the data management benchmarks. During this process we were able to test our performance expectations, and to clearly and objectively set expectations for end users. Implied in the building on existing infrastructures versus building an enclosed and dedicated system are that the availability and stability are dependent on the leveraged infrastructures. A system architecture enabling scalability supports unexpected growth and staging of cost throughout the life cycle of the environment.

This framework was developed for a large data analytics program and was key to the development of the SeDAP computing environment. SeDAP adoption by a user group representing multiple health care data researchers and data managers from multiple schools across UNC has been rapid.

Next steps for this system include standardization of information and data structure normalization to improve access to the data. We hope to accelerate cohort discovery and to shorten the life cycle of grant development and research projects by deploying a consolidated Observational Medical Outcomes Partnership (OMOP) model within a relational database system. The framework and processes of benchmarking and soliciting user inputs will continue to help with implementation and evaluation.

References

1. de Montjoye, Y.A., et al., *Identity and privacy. Unique in the shopping mall: on the reidentifiability of credit card metadata*. Science, 2015. 347(6221): p. 536-9.
2. US Department of Health and Human Services, *HHS Announces Proposal To Update Rules Governing Research On Study Participants*. 2015. Web. 14 Sept. 2015.
3. *Modifications to the HIPAA Privacy, Security, Enforcement, and Breach Notification rules under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act*; other modifications to the HIPAA rules. Fed Regist. 2013 Jan 25;78(17):5565-702.
4. Meyer, A.M., et al., *Big data for population-based cancer research: the integrated cancer information and surveillance system*. N C Med J, 2014. 75(4): p. 265-9.
5. Stokols D, Misra S, Moser RP, Hall KL, Taylor BK. *The ecology of team science: understanding contextual influences on transdisciplinary collaboration*. Am J Prev Med. 2008;35(2 suppl):S96-S115.
6. *An Introductory Resource Guide for Implementing the Health Insurance Portability and Accountability Act (HIPAA) Security Rule*. http://csrc.nist.gov/publications/nistpubs/800-66-Rev1/SP-800-66-Revision1.pdf