# The North Carolina Translational and Clinical Sciences Institute:

## ACTIVITIES OF THE BIOMEDICAL INFORMATICS CORE

JAVED MOSTAFA, PH.D., M.A. AND CARLTON MOORE, M.D., M.S.

The environment for clinical and translational research is rapidly evolving. Driven by an accelerating rate of scientific discovery, there is a growing imperative to markedly reduce the time from discovery to patient benefit. It is clear that previously successful approaches to translational and clinical research need to be transformed to accommodate this challenging environment.[1]

In May 2008, the University of North Carolina-Chapel Hill (UNC-CH) joined a select group of nationally recognized universities in receiving a Clinical and Translational Science Award (CTSA). A major component of the services provided by the North Carolina Translational and Clinical Sciences Institute (NC TraCS) to help facilitate clinical and translational research projects is the Biomedical Informatics Core (BMC). BMC is dedicated to support research needs of investigators. The current BMC projects fall under three broad categories: (1) Discovery, (2) Management, and (3) Dissemination. The three categories match the "life-cycle" of a research project, starting with the problem definition phase, moving on to the data collection phase, and finally ending in the findings and impact phase. In this article, we will focus on our main tool for supporting the discovery and problem refinement phases of a research project: the Carolina Data Warehouse for Health (CDW-H). We will conclude the article with an overview of some of the other projects that support the management and dissemination phases.

## Carolina Data Warehouse for Health

### Data

Currently the CDW-H contains all inpatient and outpatient clinical data from 2004 onward, as well as administrative data, such as billing codes and appointment schedules. The first phase of the CDW-H has been created by extracting data elements from the two major operational sources of information in the UNC Health Care System. The first source is the electronic medical record, WebCIS, which supports the clinical operation of our inpatient and outpatient enterprise. The second is our Siemens Decision Support System that stores transactions from our registration, billing, scheduling, and physician ordering systems. On a once daily basis, data extraction takes places from these two systems and is transformed into the core of the data warehouse called the Atomic Data Store (ADS). This integrated and standardized resource forms the basic content of the CDW-H system (*Figure 1*). As of October 2009, the core of the data warehouse contains records for 1.8 million people in 202 tables encompassing 2,840 unique fields.

The CDW-H became operational in March 2009. The technical implementation of the CDW-H took 12 months and leveraged IBM data warehousing technology. IBM InfoSphere™ Information Server is the decision-support platform used that



**Figure 1.** Major functions and resources supported by the Carolina Data Warehouse for Health.
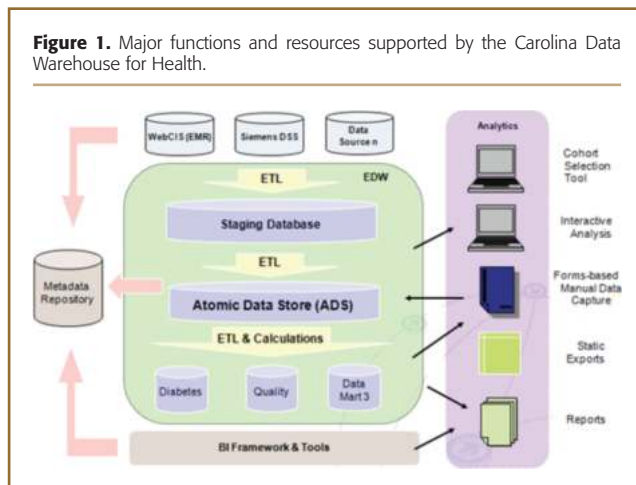
runs on the IBM System z10™ using the IBM DB2® for z/OS® database software.
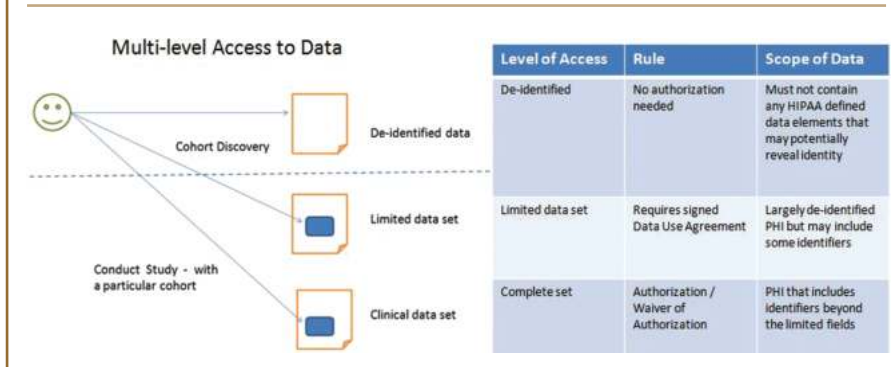
Major subject areas in the CDW-H include: patient, contact information, billing information, payer, organization, patient visit provider, hospitalization, allergies, core measures, diagnosis, medications, health maintenance, immunizations, lab results, medical problems, procedure, and vital signs. Also, included in the CDW-H are: cardiology reports, clinical notes, ECG reports, GI reports, pathology reports, pulmonary reports, radiology reports, and respiratory reports.

### Applications

One CDW-H application allows queries to be defined and executed against the core data set in the ADS. The application, known as the Research Portal, is aimed at researchers in the discovery phase of a project who are interested in identifying and verifying if a patient cohort exists based on a specific set of search criteria. The portal allows query formulation using parameters, such as diagnosis codes, birth year, race, gender, lab results, and drugs. Users can query, retrieve, print, and save outcomes produced in each session. The researcher at this early phase may or may not have institutional review board (IRB) approval, and hence the portal only displays results in aggregated and de-identified format. The Research Portal can only be accessed by UNC staff and students who secured approval and acquired an account. The system requires users to acquire Health Insurance Portability and Accountability Act (HIPAA) training before permission is granted.

The governance of data presentation in the Research Portal application was carefully established such that the decision-logic follows current HIPAA and IRB guidelines (*Figure 2*). It should be emphasized here that our focus was on expanding the

University of North Carolina at Chapel Hill, North Carolina, USA
Correspondence: J Mostafa (jm@unc.edu)

**Figure 2.** Multiple categories of use and the decision-logic for data provisioning from the CDW-H.

Multi-level Access to Data

| Level of Access | Rule | Scope of Data |
|---|---|---|
| De-identified | No authorization needed | Must not contain any HIPAA defined data elements that may potentially reveal identity |
| Limited data set | Requires signed Data Use Agreement | Largely de-identified PHI but may include some identifiers |
| Complete set | Authorization / Waiver of Authorization | PHI that includes identifiers beyond the limited fields |

Cohort Discovery — De-identified data

Conduct Study - with a particular cohort — Limited data set / Clinical data set

secondary use of medical records by minimizing the barriers to access, but not necessarily by eliminating all restrictions. The logic for the use case scenarios as represented in *Figure 2* was developed in close collaboration with the UNC IRB.

In addition to the Research Portal application, CDW-H also has two data marts with specific analytic functions. Our diabetes data mart is disease specific, established for the population-based care of diabetics. The second data mart in the CDW-H is for inpatient quality assessment and focuses on the particular organizational need for quality metrics for hospitalized patients. Within data marts, clinical data can be linked with research data or survey data designed to improve patient quality of care.

### Ongoing & Future TraCS Biomedical Informatics Projects

The TraCS Institute has been attracting significant attention from researchers around campus and according to the most recent data, 98 project requests have been submitted in the last 11 months that require direct support from the BMC (primarily CDW-H access and usage). A logical next step for many projects involving patient cohort identification and recruitment is the creation of a data management system to systematically track patients and collect additional study-specific data. Based on current technology,* we developed a clinical study data management service, which has received 29 requests since going live in May 2009, and is currently serving 22 active research studies.

To avoid duplication of effort and to further streamline the data management procedure, a generic, open-source, highly modularized platform is being built called the Translational Clinical Research Data Management System (TCRDMS). The platform can be used to create a clinical study data management system consisting of modules for form generation, data aggregation, security/authentication, query, and reporting. The goal is to reuse the modules whenever possible and customize them as needed to meet individual study requirements. The TCRDMS system will launch in a few months.

Inclusion of image data in the CDW-H, particularly radiological images, is a future goal. One key challenge will be to seamlessly integrate search, browse, and display functions into the Research Portal. Toward that end, we developed a prototype system called ViewFinder-Medicine (VfM) that is capable of automatically organizing radiological scans into clinically meaningful categories and permitting users to search image collections (e.g., from PACs) based on both preannotated textual descriptors and visual clues. We have recently tested VfM with a curated MRI collection from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and achieved high classification accuracy and retrieval performance.[2]

Finally, a key objective for TraCS BMC is to facilitate the dissemination of accurate and understandable medical information to the community that our institution serves. A major gap in the community is availability of current and accurate medical information, particularly in a form that is easy to access and understand. We know from national surveys that medical information seeking is quite common[3] and most people rely on generic resources (i.e., search engines). In collaboration with our TraCS Community Core, we are currently testing a consumer health information application known as MedSIFTER that is capable of automatically aggregating consumer-centric medical information (e.g., NIH sponsored Medline Plus system) and providing the information based on health profiles in a highly focused and personalized way. We are working with North Carolina's Family Network Community—a community of family members with disabled children—as our first user-group to help us evaluate the utility of the MedSIFTER system. If successful, we plan to broaden the scope of the MedSIFTER system and target similar user groups.

## REFERENCES

1. Williams RL, Johnson SB, Greene SM, Larson EB, Green LA, Morris A, Confer D, Reaman G, Madigan R, Kahn J. Signposts along the NIH roadmap for reengineering clinical research: lessons from the Clinical Research Networks initiative. *Arch Intern Med.* 2008; 168: 1919–1925.

2. Agarwal M, Mostafa J. Image retrieval for Alzheimer's disease detection. In: *Proceedings of The Medical Content-Based Retrieval for Clinical Decision Support (MCBR-CDS) Workshop in conjunction with MICCAI*. London, UK, 2009.

3. Fox S. Health Sites: Some are more equal than others, 2010. Available at: http://www.pewinternet.org/Commentary/2010/January/Health-Sites-Some-Are-More-Equal-Than-Others.aspx. Accessed February 4, 2010.

* We are using REDCap software, which is suitable for small to medium GCRC-type studies.