



The Biomedical Informatics Grid (BIG): A Platform for 21st Century Biomedicine

A Community-based Personalized Health Care Project Summary

Submitted by:

Kenneth H. Buetow, Ph.D.

Associate Director, Bioinformatics and Information Technology
Director, Center for Biomedical Informatics and Information Technology
National Cancer Institute

September 2008

Table of Contents

- I. **Overview**
- II. **Personalized Medicine Requires a 21st Century Systems Approach**
- III. **The cancer Biomedical Informatics Grid (caBIG®): Proof of Concept Platform for Personalized Medicine**
- IV. **caBIG® Enterprise: Platform for Networking the Global Biomedical Community**
- V. **The BIG Health Consortium: 21st Century Model for Biomedicine**

I. Overview

The realization of Personalized Medicine requires biomedicine to embrace a new 21st century systems-based approach. This system blends the research and health care sectors and intimately involves consumers. It realigns incentives. It reduces costs. It becomes a learning system.

The use of information is key to achieving this new personalized medicine ecosystem. Information binds the individual components and connects the community. The vast volume and complexity of information will require the ubiquitous adoption of state-of-the-art information technology throughout biomedicine.

The new Personalized Medicine paradigm will not emerge fully formed. It will require progressive steps of iterative and incremental development that can be assembled into a whole. Successive cycles of prototyping and development will reach larger communities and broader market penetration.

The National Cancer Institute (NCI) and the cancer community represent an ideal test bed for the development and deployment of such a Personalized Medicine paradigm. Cancer is a disease of genes, and the cancer community has performed pioneering work in Personalized Medicine. Its standing platforms provide unique opportunities for experimentation.

The NCI began efforts to create the new Personalized Medicine ecosystem early in this decade by launching the cancer Biomedical Informatics Grid (caBIG®). In its prototype phase from 2003-2007, caBIG® developed the collaborative models and information technology necessary to join the cancer community through the NCI-designated Cancer Centers. More than 1,000 individuals from 200+ institutions participated during those years.

Now in its “enterprise” phase, caBIG® is being deployed to connect consumers, the care delivery system, and the research community. Close to 60 NCI-designated Cancer Centers are adopting caBIG® infrastructure and tools, as are 16 Community Cancer Centers that in the aggregate touch 20 million lives. There are to date over 100 grid nodes currently online at a variety of U.S. government, academic and commercial organizations, linking those entities in a data-sharing network -- an unprecedented feat of cultural and technical change for the biomedical enterprise.

caBIG® tools and infrastructure are also reaching beyond cancer as the information technology backbone for other disease research communities; are being applied internationally; and are being repurposed to support electronic healthcare transactions in the Nationwide Health Information Network.

Building on that foundation, the NCI has launched the BIG Health Consortium™. This broader ecosystem will leverage the caBIG® connectivity platform and join government, academic, non-profit and industry efforts to realize the Personalized Medicine paradigm.

II. Personalized Medicine Requires a 21st Century Systems Approach

Personalized Medicine as a New Biomedical Paradigm

Personalized Medicine is a new paradigm in biomedicine. Its successful implementation requires integration of unprecedented amounts of information and diverse communities. The ability to collect, analyze, share, and integrate massive quantities of biological and clinical data in real time is a prerequisite for Personalized Medicine.

Biomedicine is a complex system. There are key interdependencies between the sectors that compose this complex system. Personalized Medicine's goal is to transform this system and must therefore recognize and embrace its complexity. Key opportunities to create a self-sustaining Personalized Medicine ecosystem come from understanding resource and information flows within the larger system.

Strategies for Addressing Complexity. Industry provides best practices for active design of complex systems. First, best practice requires one to recognize the system as a whole. Next, it identifies the interfaces between the components. Within the boundaries of the interfaces, individual components are developed and manipulated iteratively and incrementally. It is also important that initial development occur in a limited context, but one with sufficient complexity that it faithfully captures the complexity of the system component. Finally, additional complexity is also added incrementally with the controlled expansion of scope. This approach permits rapid incremental success without being stymied by the complexity of having to “boil the ocean.”

The Essential Role of Information Technology. The daunting complexity of the personalized medicine ecosystem makes the use of information technology critical. But information technology within the biomedical enterprise has been slow to develop and is rarely connected between laboratories even within a single institution, much less between different institutions. In contrast with other national efforts, such as in defense or federally-funded physics research, the U.S. biomedical research enterprise has never had any such information technology system.

Thus, to address the complexities of cancer and these discontinuities of the research process, a 21st century biomedical enterprise requires *interoperability*; that is, access to integrated tools to collect, analyze, and share data in standardized formats. This interoperability is a means to link together all the scientists, clinicians, patients, and other participants so that they can share such standardized information rapidly.

The current generation of internet and world wide web technologies makes information technology approachable to biomedicine. Information technology is critical to the interface connecting the components of the biomedical ecosystem. It enables efficient operations within components.

A Systems View of Personalized Medicine

Multidimensional Stakeholder Ecosystem. The full ecosystem of Personalized Medicine encompasses members of the axes of biomedicine. It includes researchers, physicians, and consumers as participants. The researcher category includes discovery, translational, and clinical arenas. In an alternative axis, the ecosystem includes the academic, not-for-profit, commercial, and government sectors. A complete survey of the ecosystem also contains gatekeepers, such as regulators and payers.

Connectivity Through Information Technology. The needs of Personalized Medicine for information-sharing are accommodated by best practices in information technology. Applications of information technology are arbitrarily segmented between approaches used to connect information and approaches to connect people.

Best practices to connect information call for the use of a services-oriented architecture. Services should interoperate through well-defined interfaces. The architecture defining the interface should include Enterprise, Information, Computational, and Engineering viewpoints, and be technology platform neutral. The information should be represented utilizing internationally accepted standards where available.

Communications using information technology is rapidly evolving. Tremendous opportunities exist in utilizing web technologies, especially the emerging Web 2.0 approaches to community organization and business.

Personalized Medicine Ecosystem as a Learning System. A key benefit of conceptualizing the complete Personalized Medicine ecosystem is the capacity to convert biomedicine into a learning system. More specifically, by capturing the entire biomedical life cycle, it is possible to synergistically combine research, care delivery, effectiveness measurement, quality assessment, and safety.

Cancer as the Pioneering Field in Personalized Medicine

Cancer researchers have been at the leading edge of the Personalized Medicine revolution, and many of the first-generation personalized medicine products have been developed for cancer indications. There are three obvious reasons for this early focus:

- First, cancer is a complex set of diseases, for which molecular medicine approaches predate even the Human Genome Project. It has been known for decades that cancers are caused by genetic changes – either inherited or acquired – that result in abnormal cell proliferation, cell division or cell death. As early as 1971, a cancer geneticist analyzed individual retinoblastoma cases and proposed that the individual tumors were caused by the combination of an inherited mutation plus a mutation that was acquired, the “two-hit” hypothesis. We have since discovered that mutations in oncogenes and tumor suppressor genes affecting several defined stages of the

cellular life cycle can contribute to the evolution of an individual cancer. As Nobel laureate Michael Bishop famously concluded, “the seeds of cancer are within us”.

- Second, cancer is a serious, often deadly condition, for which the efficacy rates of therapeutics have traditionally been extremely low. Since selection of the most efficacious treatment for the patient can be an urgent life-or-death decision, personalized medicine approaches vs. time-consuming “trial and error” are compelling. For example, misapplying the treatment for one type of lymphoma (diffuse large B-cell lymphoma) to an ostensibly very similar lymphoma (Burkitt’s lymphoma) reduces the survival rate from 80% to less than 20%. (Staudt, et.al, N. Engl. J. Med. 354: 2431-42, 2006). Molecular sub-typing is therefore necessary to differentiate between patients, and to select appropriate treatment early and accurately.
- Third, the adverse effects of cancer therapeutics are extremely unpleasant, disfiguring and potentially fatal, thereby making it even more important to select the optimum therapeutic choice the first time, to avoid the doubly-negative impact of adverse effects from futile treatment.

It is not surprising, then, that the “poster children” for Personalized Medicine have arisen in oncology: Herceptin® (trastuzumab) in the treatment of breast cancer and Gleevec® (imatinib mesylate) in the treatment chronic myelogenous leukemia (CML).

Herceptin, a monoclonal antibody that blocks the human epidermal growth factor receptor 2 (HER2) protein is found to be effective for the approximately 25% of breast cancers associated with a gene amplification and over-expression of HER2. As a result, Herceptin treatment is now regularly preceded by a companion diagnostic for HER2 overexpression.

Gleevec (imatinib mesylate) was designed to target the mutant Bcl-Abl protein formed by an abnormal genetic fusion in chronic myeloid leukemia. Gleevec was found to help 98% of patients with therapy-resistant CML in its first clinical trial. Since its development, Gleevec has been found to target other cellular proteins and is being tested for its effectiveness in other forms of cancer.

Joining the ranks of tests for HER2 overexpression in breast cancer and BCL-ABL kinase mutation in CML, are tests for the enzyme EGFR in non-small cell lung cancer (NSCLC) and KRAS in colorectal cancer. EGFR amplification has been associated with improved outcomes for treatment with Iressa (gefitinib), an EGFR inhibitor. The presence of KRAS in its normal (non-mutated) form has recently been associated with improved outcomes in the treatment of first line metastatic colorectal cancer with Erbitux (cetuximab).

In addition to diagnostics based on enzymes with strong associations to cancer (which include RAF, KIT, JAK2, and DNA repair enzymes in addition to those mentioned above), candidate diagnostics are also being developed based on changes in expression patterns of larger sets of genes analyzed. Changes in epigenetic DNA modifications, such as markers of methylation, are also being studied for their predictive utility.

The National Cancer Institute's 21st Century Biomedical Test-bed

The NCI's Unique Research and Care Delivery Platforms. The NCI has a unique collection of administrative platforms that capture the entire lifecycle of biomedicine development, and hence it supports a unique environment in which the Personalized Medicine paradigm can be prototyped. For over 30 years, NCI has supported Comprehensive Cancer Centers, which blend research, care delivery, and prevention. There are more than 60 of these centers, distributed nationally and housed at the most prestigious research and care delivery institutions throughout the United States. More specialized programs include more than 50 NCI Specialized Programs of Research Excellence (SPOREs) that support translational research, and 10 NCI Cooperative Group programs that conduct multi-institutional clinical trials. Most recently in the care delivery area, the NCI has launched a Community Cancer Center Program (NCCCP) with 16 sites that cover 20 million lives.

III. The cancer Biomedical Informatics Grid (caBIG[®]): Proof of Concept Platform for Personalized Medicine

“What is required in cancer research to find definitive answers is a system to share data and leverage all the events in the cancer world. It is impossible to succeed without embracing that notion. The concept of caBIG[®] is, therefore, right on target.”

Kim Lyerly, M.D.

Director, Duke Comprehensive Cancer Center

Origins and Development of caBIG[®]

The National Cancer Institute (NCI) identified the need in 2003 for an informatics initiative of unprecedented scope for the biomedical community in recognition of three factors: the growing clinical and economic burden of cancer; the transformation of research catalyzed by the molecular revolution and multiple genomics technologies that were generating massive amounts of data; and the recognition that the “essential unity” of research and clinical care had powerful potential to improve the outcomes of all cancers, as it had done in the field of pediatric oncology.

As a first step in building an informatics infrastructure that would enable Personalized Medicine, the NCI officially launched the caBIG[®] (cancer Biomedical Informatics Grid) initiative in 2004 as a pilot program. Its initial objective was to develop capabilities that would meet the self-identified needs of the NCI Cancer Center community. (For more information on the history of caBIG[®], see the **caBIG[®] Pilot Phase Report** at <http://cabig.cancer.gov/resources/report.asp>)

caBIG[®] Strategic Principles.

Four fundamental principles underlie the activities of caBIG[®] and guide all of its operations:

- **Open Access:** Participation in caBIG[®] and the products delivered by caBIG[®] are open to all, enabling access to tools, data, and infrastructure by the cancer and greater biomedical research communities.
- **Open Development:** Software development projects are assigned to particular participants, but are informed iteratively with multiple opportunities for review, comment, further modification, and development by the caBIG[®] community.
- **Open Source:** The software code underlying caBIG[®] tools developed with the support of the NCI is available to software developers for use and modification. This software is licensed as open source to promote the reuse of existing code, hence optimizing the full benefit of the research dollars spent. Nonetheless, caBIG[®] recognizes the need for and importance of commercial software to the biomedical enterprise, and accommodates it through caBIG[®] interfaces. The open source license is industry-friendly, allowing commercialization of derivative products and

fostering industry interest and innovation, while still adhering to the principle of open source for caBIG[®]-funded activities.

- **Federation:** caBIG[®] software and standards enable local organizations, such as Cancer Centers, to share data resources with the larger cancer care and research community and to use resources contributed by others. On the grid, these resources can be aggregated from multiple sites to appear as an integrated research dataset, while the individual resources remain under the control of the local organizations.

“I view caBIG[®] as being absolutely essential to our strategic mission. We have very ambitious plans for the implementation of an integrated clinical and molecular database that can ultimately guide personalized medicine.”

Louis M. Weiner, M.D.

*Director, Lombardi Comprehensive Cancer Center
Georgetown University*

caBIG[®] Philosophy and Culture: The Predominance of Community. The concept of a “caBIG[®] community” has driven its underlying strategy and much of its organization and culture. To provide limited, but meaningful scope and focus, the caBIG[®] effort was launched in the NCI-designated Cancer Centers. Each Cancer Center was initially visited to discover its capabilities and needs. Community engagement was maintained by creating workspaces with open, transparent participation around the areas of need identified by the community. These workspace developed specific plans of and priorities for the component of the community they represented. Over-arching workspaces were created to define the architecture, define the interfaces, and provide coordination between components.

The caBIG[®] program leveraged novel communications approaches to support its open national community. Virtual meetings were held regularly via teleconference and supported by real-time internet sharing of presentations. It utilized e-mail lstrsv technology to support asynchronous communications. It also made very heavy use web technology to join the active community, sharing the agendas and results of virtual meetings as well as the technical artifacts of the development efforts.

caBIG[®] has also encountered several obstacles along the way. Among the most daunting was the cultural shift required for caBIG[®] success, in a biomedical enterprise that was unaccustomed – and not incentivized – to connect and share cross-institutionally.

The caBIG[®] initiative in its early years did underestimate the need for continual communications not only about its progress overall, but the specific details and timing of each program to the broader cancer community. Generally, caBIG[®] expected less interest than it actually did receive at each stage, and hence did not always prepare for what was at times an overwhelming desire for participation by large and diverse groups. There were also

many legal, technical, and cultural issues surrounding the objective of data-sharing that have since been addressed and specifically written into agreements.

caBIG® as the World Wide Web of Cancer Research. As noted above, the mission of the caBIG® initiative is to link the entire cancer research community in a World Wide Web of research. caBIG® provides infrastructure for creating, communicating, and sharing bioinformatics tools, data, and research results, while using shared applications, shared data standards, and shared data models, all operating on a cancer community network (caGrid).

caGrid is underlying service oriented architecture that provides universal mechanisms for enabling interoperable programmatic access to data and analytics in caBIG®. caGrid also creates a self-described infrastructure wherein the structure and semantics of data can be programmatically determined, and provide a means by which services available in caBIG® can be programmatically discovered and leveraged.

There are to date over 100 grid nodes currently online at a variety of U.S. government, academic and commercial organizations, enabling those entities to share data.



Figure 1. At the caBIG® website, a “hot map” shows the organizations on the caGrid network as they come online for data sharing.

Use-driven Capabilities – Real Solutions to Real Problems

“We see caBIG[®] supporting our ability to deliver personalized medicine by allowing us to share our knowledge and our understanding and our data with others in a way that creates datasets that are much larger than could ever be generated by a single institution or a single program.”

Robert Clarke, Ph.D., D.Sc.
 Professor and Co-Director Breast Cancer Program
 Lombardi Comprehensive Cancer Center
 Georgetown University

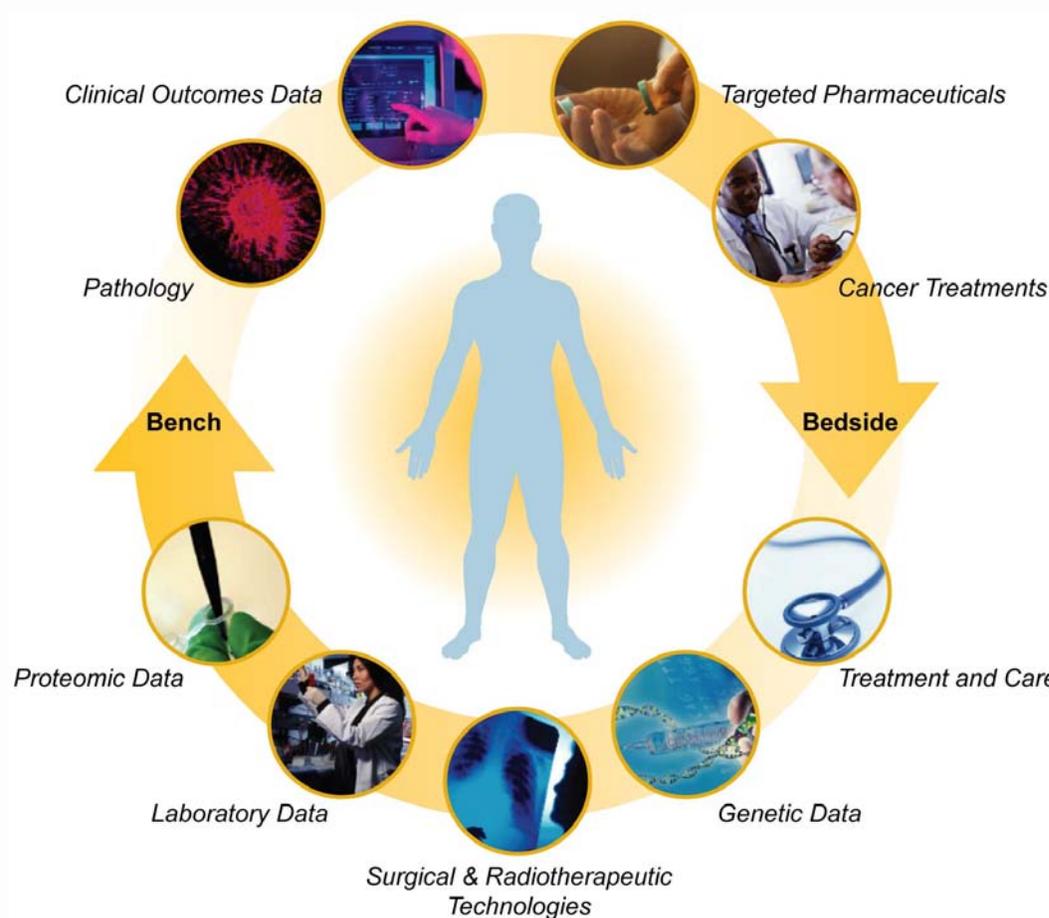


Figure 2. Central to the fulfillment of Personalized Medicine is the ability to link previously disparate functions into a seamless continuum.

caBIG[®] provides more than 40 software tools, as well as the connecting network called caGrid, by which every function required in the molecular-based discovery and clinical research continuum can be performed and linked together.



Figure 3. The caBIG[®] portfolio of software tools facilitate every function of translational medicine.

The extensive and continually evolving portfolio of caBIG[®] capabilities can be reviewed at the website (www.cabig.nci.nih.gov) and freely downloaded for use.

Described below are representative examples of how tools developed in the early stages of the caBIG[®] initiative are being used to solve actual challenges in translational research within the NCI Cancer Center community.

Imaging Sharing and Analysis. A key goal of the caBIG[®] Imaging program is to facilitate sharing of DICOM (Digital Imaging and Communications in Medicine) images in a wide variety of research settings by leveraging caBIG[®] technology, including caGrid, and by

creating medical image-specific applications for annotation, visualization, and image analysis. Key applications and capabilities include:

- **NCIA:** The National Cancer Imaging Archive (NCIA) application is a searchable national repository that integrates cancer images with genomic and clinical information. Users can access the NCI-hosted instance of the NCIA, or they can install a local copy for their own use. NCIA supports federation of data at multiple sites while still permitting single query searching across those sites. NCIA is adaptable to other therapeutic areas beyond cancer.
- **caIMAGE:** A publicly-accessible database of cancer images maintained by the NCI that allows researchers to search for images or submit their own.
- **XIP:** The eXtensible Imaging Platform (XIP) provides an open-source toolkit to speed development of medical imaging applications built from an extensible set of modular components, making it simpler and less expensive to provide post-processing applications onsite, and increasing the uniformity of imaging and analysis.
- **AIM:** A software application and associated ontologies and object models to create, validate and render image annotations and markups, as part of the XIP toolkit.

Optimizing Clinical Trials Scheduling with caBIG®. With increased focus on individualized care, and a growing number of patients enrolled in multiple clinical trials, there is increased need to track specific patient activities and outcomes. When a patient is enrolled in multiple trials, tracking events, treatments, and potentially dangerous side effects is critical, in order to attribute these outcomes to the correct trial, and facilitate adjustments to the treatment schedules.

The caBIG® Patient Study Calendar (PSC) was developed to ease the burden for clinicians and coordinators. In use for clinical trials at Thomas Jefferson University, The Mayo Clinic, and Northwestern University, the PSC assists trial coordinators and clinical researchers by creating reusable templates that can be personalized for each patient in a trial, and by enabling sharing of calendars between sites. By linking the calendars with electronic data capture (EDC) of the lab results and pathology data, for example, the PSC allows the coordinators to adjust treatment schedules as needed and help assign particular outcomes to specific treatments in a trial. Because the PSC does not use a proprietary calendar format, the data can be exported to existing calendaring systems in use at the clinical trial center, simplifying integration into the existing infrastructure. The calendar can also be shared with the patients, enabling them to track their own progress and feel less uncertainty about their future activities in the trial.

Duke University - Department of Defense Breast Cancer Trial. Duke University is conducting a genomics-guided adaptive clinical trial on metastatic breast cancer, with caBIG® informatics support. The goal of the study is to compare the results of genomic-guided treatment selection versus conventional arm randomization for two different chemotherapeutic agents, based on a 60% likelihood of response to a particular drug.

The trial is enabled by a wide collection of caBIG® technology, including the caBIG® CDMS system, caTissue for breast tissue sample management, and caArray to handle the gene expression microarray data, all connected by caGrid. The trial is currently enrolling patients, with ongoing improvements planned to simplify and streamline the informatics connectivity. This basic trial informatics framework will be used for additional trials in prostate and lung cancer that are scheduled to start in coming months.

REMBRANDT (Repository of Molecular Brain Neoplasia Data). Genomic technologies -- such as gene expression profiling, SNP data, gene sequence data, and proteomics profiles - - can complement traditional data (such as pathology information and outcomes data), giving researchers and clinicians a much more complete characterization of tumors. This knowledge has resulted in the sub-grouping of many cancers into discrete subtypes, each with a different predicted outcome, and with different optimal treatment strategies.

Traditionally, researchers had to access each of these data types separately, making the discovery of synergies and relationships very difficult. The REMBRANDT (Repository of Molecular Brain Neoplasia Data) web portal provides users with a user-friendly interface that facilitates *ad hoc* querying across multiple diverse data domains. The goal of the REMBRANDT project is to molecularly characterize a large number of adult and pediatric primary brain tumors and to correlate those data with extensive retrospective and prospective clinical data. By standardizing the collection of tumor samples and their annotation, along with uniform methodologies for generating the genomic data, comparisons across tumors and varying data types becomes possible. This ability to query comparable data across multiple domains allows physician-scientists to understand subtle differences between sub-classes of brain tumors, and thus make the right decisions during patient treatment. Over 890 cases have been examined to date, with additional samples added monthly. More than 300 researchers have registered to use the REMBRANDT web portal.

The REMBRANDT portal (<http://caintegrator-info.nci.nih.gov/rembrandt>) is enabled by caIntegrator, a service-oriented architecture that permits pluggable web-based graphical user interfaces and uses the Clinical Genomics Object Model (CGOM) to provide standardized programmatic access to the data in the warehouse.

VASARI (Visually AccessSable Rembrandt Images). Increasingly, oncologists recognize that the genomics, MR imaging features, and biologic behavior of tumors vary even within the same histological subtype of tumor. VASARI (Visually AcesSable Rembrandt Images) is a post-facto assembly of clinical MRI images obtained at the time of diagnosis on the same samples used in the REMBRANDT program. The goal of VASARI is to develop reproducible methods to classify MRI images of glioma tumors and provide linkages between those images, histology, and genetic data obtained from brain cancer specimens (REMBRANDT). Ultimately researchers are exploring whether there are imaging features that might be better predictive markers of biologic behavior than histology.

The caBIG® NCIA (National Cancer Imaging Archive) toolkit provides the mechanism for image storage and analysis and caIntegrator provides the underlying query infrastructure.

caBIG® and BreastCancerTrials.org. One of the difficulties faced by cancer patients is identifying state-of-the-art treatment available in clinical trials while still managing the demands of their lives. Unlike juvenile cancer, where about 60% of patients take part in a clinical trial, it is estimated that fewer than 10% of eligible adults with cancer do so. The much higher rate of cure in children may be at least partly associated with their increased participation in clinical trials.

caMatch was developed to help identify patients who are potentially eligible to participate in clinical trials, by examining patient data from their electronic health records and matching them to eligibility criteria for specific trials. A web-based user interface simplifies program use for the patient. The technology has been tested successfully as part of the BreastCancerTrials.org website, a pilot program run by the UCSF Center of Excellence for Breast Cancer Care. The success of this program has prompted an expansion from the San Francisco area pilot to nationwide availability in 2008.

“The real impact of caBIG® is going to be when large groups of institutions start using these tools to acquire tissues, standardize tissue, and data acquisition. When that happens, we will be able to do research much quicker, and have results that are clinically valid and transferable to the patient, with individualized, personalized care.”

Gustavo Ayala, M.D.
Fulbright Professor of Pathology
Baylor College of Medicine

IV. caBIG® Enterprise: Platform for Networking the Global Biomedical Community

*“The progress we have made has opened up a vision of a future personalized cancer medicine, when doctors will determine prognosis and treatment options by understanding each patient’s unique genetic makeup and the genetic aberrations that have led to his or her cancer... The National Cancer Institute has not only embraced but **is leading the way** to that future, and is dedicated in all it does to ushering it in as rapidly as possible.”*

*National Cancer Institute
An Annual Plan and Budget Proposal for FY 2009*

Unifying Research and Care

Beyond providing the informatics needed for molecular based research, there is a need in Personalized Medicine to link the research endeavor back to health care delivery. Specifically, caBIG® is providing the ability to integrate molecular profiling, family history and molecular diagnostics into the Electronic Health Record, as well as to share back clinical outcomes data and clinical trial results into the discovery enterprise to achieve a ‘rapid learning’ system.

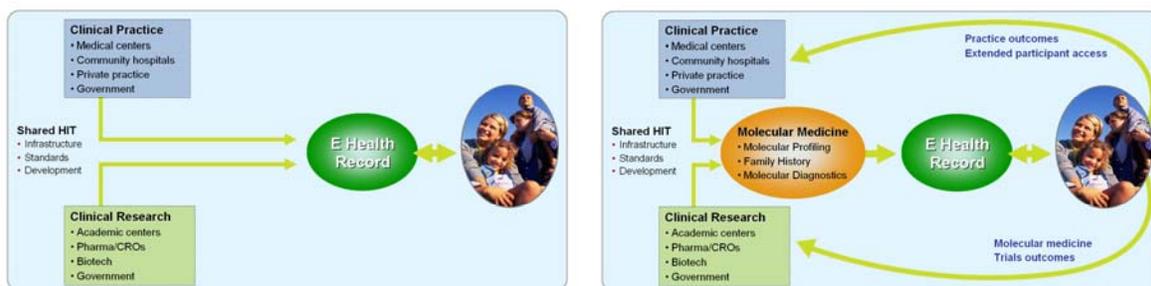


Figure 4. caBIG® facilitates the integration of molecular information into Electronic Health Records.

Following the completion of the pilot phase of the caBIG® initiative, the NCI took the next step towards an infrastructure for Personalized Medicine by extending caBIG® to an “enterprise phase”, with expanded capabilities to network the larger cancer community and beyond.

Today, caBIG® is a network of interconnected data, individuals, and organizations, designed to share data and knowledge, simplify collaboration, speed research to move new diagnostics and therapeutics from bench to bedside faster and more cost effectively, and ultimately to realize the potential of Personalized Medicine to improve patient outcomes.

A total of 56 NCI-designated Cancer Centers across the nation are working to connect their research and clinical care capabilities into a caBIG[®]-enabled information network. Through the NCI's Community Cancer Centers Program (NCCCP), 16 Community Cancer Centers that in the aggregate touch 20 million lives are also becoming a part of this network. caBIG[®]-enabled connectivity enables these Centers to participate in clinical research studies and to bring the benefits of Personalized Medicine to their patient population in real time.

More than 1,000 individuals from over 200 organizations have participated in caBIG[®] activities since the initiative's inception. Moving forward, however, it will be difficult to count the participants, since research users are increasingly applying caBIG[®] tools automatically as part of their studies without even noticing that they are "powered" by caBIG[®] infrastructure. In addition, as more and more software becomes caBIG[®]-compatible, countless users will benefit from its interoperability features without awareness of its presence.

caBIG[®] in Action

In the "enterprise" phase, caBIG[®] infrastructure and tools are becoming ubiquitous among NCI intramural and extramural programs, as it enables and accelerates basic and clinical research. Representative examples of such caBIG[®]-enabled activities are:

Inter-SPORE Prostate Biomarker Study (IPBS). The SPORES (Specialized Programs of Research Excellence) are NCI-sponsored clinical research groups each specializing in a particular type of cancer. While each SPORE conducts its own trials, when biomarkers have been compared between centers, there has been a high degree of variability in the clinical significance of biomarkers screened from one center to another. The IPBS study was designed to assess ways to unify the data collection and analysis of samples, improving consistency of results. The IPBS study leverages caGrid to connect all participating centers, and applies caTissue to track the samples and manage the analysis results.

caBIG[™] and Mutational Analysis. The International HapMap project is a continuing effort to compare the genetic sequences of groups of different individuals to identify chromosomal regions where genetic variants are shared. The first two phases were completed in 2007 and opened the door to wider use of Genome Wide Association Studies (GWAS), where DNA markers are scanned across the genomes of many individuals to find genetic variations associated with a particular disease. In the past year, GWAS studies have found genetic associations for coronary heart disease, Type I diabetes, and breast cancer, among others.

However, researchers need sophisticated tools in order to make sense of the potentially millions of data points generated in a single GWAS study. To make these studies both simpler to interpret and more productive to find disease associations, caBIG[®] has created several tools to analyze data from GWAS and other mutational studies. The cancer Genome-Wide Association Studies (caGWAS) model allows researchers to integrate, query, report, and analyze significant associations between genetic variations and disease, drug

response or other clinical outcomes, helping researchers to find the “needle in a haystack”. Originally developed for use in cancer research, the caGWAS model was extended to accommodate the specific study needs of the cardiovascular research community as well.

In addition, the Cancer Genetic Markers for Susceptibility (CGEMS) project represents the first public release of a GWAS study for cancer. Accessible by the CGEMS data portal (<http://cgems.cancer.gov>), over 500,000 SNPs have been analyzed so far, facilitated by caGWAS to produce and upload pre-computed results tables rapidly.

The data generated as part of the CGEMS program has already helped identify variations in FGFR2, associated with increased risk for breast cancer, and multiple loci associated with increased risk for prostate cancer.

The Cancer Genome Atlas and the Cancer Molecular Analysis Portal. One of the biggest challenges to researchers of high throughput genomics technologies is how to collect and work with the large quantities of diverse experimental data. The caBIG[®]-enabled Cancer Molecular Analysis (CMA) Portal (<http://cma.nci.nih.gov>) provides powerful tools and resources that enable cancer researchers across the world to explore, visualize, and integrate genomic characterization, sequencing, and clinical data from a variety of data sets.

The Portal exemplifies the caBIG[®] core principles of open development and federation. The CMA Portal allows researchers to use analysis programs developed at three different organizations, and to access data produced by more than 10 different institutions, all by a unified web interface. The tools available on CMA Portal allow researchers to access clinical characteristics such as survival data and tumor staging, and correlate those with mutation and other genomic data. This capability enables researchers to conduct cross-platform queries, helping them to find correlations between research and clinical data that would be difficult, if not impossible, to find using conventional means.

The first data set accessible from the CMA Portal is from The Cancer Genome Atlas (TCGA). TCGA is a comprehensive and coordinated effort to improve understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing. TCGA is an integrative, multidisciplinary effort to develop and assess a framework for systematically identifying and characterizing the genomic changes associated with three cancer types: glioblastoma multiforme, squamous cell carcinoma of the lung, and serous cystadenocarcinoma of the ovary. Together, TCGA and CMA advance the opportunities for scientists and clinicians to analyze and employ TCGA data, to develop a new generation of targeted diagnostics, therapeutics, and preventives for cancer, and pave the way for more personalized cancer medicine.

FIREBIRD. To participate in FDA-sanctioned clinical trials, all investigators must fill out a variety of certification documents; key among them is the FDA registration Form 1572. Until recently these were paper-based forms, but the Federal Investigator Registry of Biomedical Informatics Research Data (FIREBIRD) application is changing that process. FIREBIRD is the first module implemented toward the vision for a Clinical Research Information Exchange (CRIX) infrastructure. FIREBIRD will leverage legally enforceable digital signatures

compliant with Title 21 Regulations using an Identity Assurance infrastructure, Secure Access for Everyone (SAFE).

FIREBIRD enables investigators to register online with the National Cancer Institute and other sponsors, including medical product companies. Through a single web-based interface to a secure central repository, investigators will be able to maintain their profile containing the accreditation information required for their participation in biologic, drug, or medical device trials. Investigators electing to participate in government, academic, or industry trials can access and apply their profile information to regulatory submission documents automatically, thus removing paper-based latencies and infrastructure costs. FIREBIRD is already in wide use across the clinical research community.

National Lung Screening Trial. Medical images play a critical role in cancer diagnosis and treatment, and the DICOM (Digital Imaging and Communications in Medicine) image standards allow technical interoperability between various medical imaging hardware and software systems. These standards, however, do not address workflow issues or how to integrate medical images with other types of biomedical information, such as genomic data, or clinical outcomes information. In addition, a standard part of the DICOM format includes the patient's name within the structure of the image file, complicating de-identification of the images for later population studies.

The caBIG® Imaging program has several collaborations underway:

- The National Lung Screening Trial uses caBIG® imaging tools to integrate radiology and pathology data.
- The Grid-enabled MAX project involves integration of caBIG® tools with all the current cooperative group quality assurance activities for imaging and radiation therapy from the Quality Assurance Review Center (QARC) and the University of Massachusetts Medical School and NCI Advanced Technology Consortium (ATC).
- An additional project expands the application of caBIG® imaging tools to optical images generated by digital histology imaging tools.

I-SPY. Unlike the treatments provided for most other diseases, cancer therapeutics are virtually all toxic compounds. To minimize the side effects and improve efficacy of these treatment with these agents, it is vital to identify biomarkers to predict which agents will be most effective for a particular cancer.

The I-SPY 1 (Investigation of **S**erial Studies to **P**redict **Y**our **T**herapeutic **R**esponse with **I**maging **A**nd **m**o**L**ecular analysis) trial is a national study to identify these biomarkers that may be predictive of response to therapy for women with late stage breast cancer.

Informatics support for the I-SPY trial includes integrating and analyzing clinical, MRI imaging, gene expression, CGH, Immunohistochemistry, and other data types. By correlating MRI image data with this collection of molecular characterization data from the tumors, researchers hope to identify biomarkers predictive for outcomes, ultimately

resulting in more effective patient treatments. The integration for I-SPY comes from caIntegrator, providing data warehousing and data mining access to researchers via a web portal, and provides an excellent example of cross-study integration and analysis in support of translational research. Over 300 women with stage II and III breast cancer have been enrolled to date. The study has also established standards for MR imaging and developed novel tools for data sharing, tissue tracking, common information repositories and clinical trial automation.

The TRANSCEND project (TRANslational Informatics System to Coordinate Emerging Biomarkers, Novel Agents, and Clinical Data) is a follow-on to the I-SPY 1 trial. The goal of TRANSCEND is to develop the next generation of clinical trials data collection by the use of web-based case-report forms (CRFs) to simplify data collection, improve collection of clinical data in support of the CRF forms at 2 I-SPY trial sites, demonstrate integration with an electronic health record system (Tolven eCHR) with the bioinformatics infrastructure in place for the I-SPY 1 trial, and develop common data elements (CDEs) for breast cancer. In addition to the caBIG® tools used in I-SPY 1, caTissue and NCIA are part of the informatics infrastructure being developed for TRANSCEND.

Clinical Data Management System (CDMS). An overarching goal of caBIG® is to increase collaboration between basic and clinical researchers by encouraging the adoption of standards-based tools and data collection. One area where the lack of standards seriously inhibits large-scale data comparison is in multisite clinical trials. This issue was recognized by the Clinical Trials Working Group of the National Cancer Advisory Board report “Restructuring the National Cancer Clinical Trials Enterprise”, which recommended creating an interoperable information technology platform for clinical trials. Broad use of standards-based electronic data capture systems improves the quality and comparability of data obtained at the different sites, facilitates multicenter trials, reduces trial administration overhead, and provides significant cost and time savings when compared with paper-based systems.

The NCI recently announced that it had acquired licensing rights from Medidata to distribute the Rave® Clinical Data Management System (CDMS) software package, with related installation, support, and maintenance services free to any interested NCI-funded organizations conducting oncology clinical trials. The new software will interoperate with other caBIG®-compatible software tools, and will enable data sharing and collaboration within each research organization, between diverse research organizations, and with NCI itself.

Providing Support to Organizations and Individual Users

The current caBIG® portfolio encompasses more than 40 end-user software applications of relevance tools across the spectrum of basic and clinical research. To simplify user adoption of these software applications, they have been grouped into “suites” that support common functions and analytical goals.

The **Life Sciences Distribution** is a collection of software applications designed to manage biospecimens and work with genomic and gene expression data, and provides valuable support to scientists conducting basic biomedical research. Key software applications include:

- **caArray:** A microarray data management system that supports data annotation and exchange, handles a wide variety of industry-standard array data formats, and provides both browser-based and programmatic access.
- **caGWAS:** A tool that allows easy management of genome-wide association data.
- **caTissue:** A biobanking management tool for the collection, tracking and management of biospecimens and the samples derived from them.
- **geWorkbench:** A desktop bioinformatics program that provides a wide variety of tools for analyzing gene expression data, sequence data, and biological pathway information.

The **Clinical Trials Compatibility Framework** provides a collection of tools that support the management of clinical trials, patient participation and data collection throughout the clinical trial lifecycle. Key software applications include:

- **caAERS:** A tool that captures and manages patient adverse event data, supporting regulatory compliance.
- **c3PR:** The Cancer Central Clinical Participant Registry helps clinical researchers track subject registrations to the trials across multiple studies and sites.
- **caXchange:** caXchange handles the translation of multiple source data formats into standards-compliant HL7 version 3 format.
- **PSC:** The Patient Study Calendar enables clinical trials managers to schedule treatment and care events for each participant in the trial.
- **CTODS:** The Clinical Trials Object Data System provides a database for storing and sharing clinical information, including de-identified data.
- **CDMS:** The Clinical Data Management System provides standardized electronic data capture by interacting with clinical data systems, reducing data entry errors and facilitating workflows.

Data Sharing and Intellectual Capital (DSIC). While much of the work of caBIG® is focused on information technology to facilitate data exchange and collaboration, technology *per se* provides no guidance about the appropriateness of such sharing, nor does it consider less-tangible factors such as protection of intellectual property, both of which can become roadblocks to collaboration.

To address these concerns, the caBIG[®] Data Sharing and Intellectual Capital (DSIC) Workspace has organized. Composed of a diverse group of stakeholders across the biomedical research community, including biomedical researchers, clinicians, technology transfer experts, intellectual property and regulatory attorneys, policy specialists, patient advocates, bioethicists, and bioinformaticists, the DSIC Workspace recognizes that there are varying levels of sensitivity of health information and that many data exchanges require agreements, validation of users, authorization of intended uses, etc.

The Data Sharing and Security Framework (DSSF). The Data Sharing and Security Framework is a collection of processes and guidelines that address legal, ethical, regulatory, policy, proprietary, and contractual barriers to data exchange. These guidelines are based on the sensitivity of data rather than its use, and provide access controls appropriate to the different levels of sensitivity.

Beyond Cancer

The tools and infrastructure of caBIG[®] can be generalized and applied in a variety of biomedical settings beyond the initial cancer community, as follows:

- *Beyond cancer*, the tools and infrastructure of caBIG[®] are being used to enable Personalized Medicine approaches in other therapeutic areas, such as in cardiovascular disease at the National Heart Lung and Blood Institute (NHLBI).
- *Beyond research*, caBIG[®] is linking discovery, clinical research and care delivery, in order to achieve the essential unity of research and care.
- *Beyond the National Institutes of Health*, caBIG[®] is being integrated into the federal health architecture to connect the Nationwide Health Information Network.
- *Beyond U.S. borders*, caBIG[™][®] tools and infrastructure are being adopted to enable biomedical enterprises in the United Kingdom, India, Singapore, China, and some countries in Latin America to achieve Personalized Medicine.
- *Beyond the “silos” of the traditional health care enterprise*, the caBIG[®] infrastructure is being applied to link a complex ecosystem of constituencies in the BIG Health Consortium (see Section V below), to demonstrate Personalized Medicine in real settings, in real time.

Joining Cardiovascular Research. One example of the adaptation of caBIG[®] to other therapeutic areas is the Cardiovascular Research Grid (CVRG) project, currently under development at Johns Hopkins University (JHU), Ohio State University (OSU) and the

University of California at San Diego (UCSD). According to Raimond Winslow, director of the Institute for Computational Medicine at Johns Hopkins:

"The Cardiovascular Research Grid will enable us to assemble large, geographically scattered research teams and bring together the leading experts in the world to focus on a common problem, regardless of their location. This grid will enable experimentalists to share their data with computational scientists, who will analyze and model the data. The computational scientists will then share their results with their experimental colleagues, who use it to refine their experiments. In this fashion, we believe the creation of the Cardiovascular Research Grid will accelerate the discovery of new approaches for treating heart disease."

At the heart of this effort is caGrid, the underlying software infrastructure that promotes caBIG[®] interoperability across the biomedical research community through the use of caBIG[™] applications, tools, information standards, and data and analytical resources. The JHU, OU, and UCSD teams are adapting caGrid to connect tools and data from the cardiovascular research community with the goal of enabling collaboration and shared discovery between cardiovascular researchers internationally. Although still in the early stages of a four year grant from the National Heart, Lung and Blood Institute, progress has already been made developing standardized vocabularies for describing biomedical data, models and data analysis applications in cardiovascular research.

caBIG[®] Architecture and Health

Effective communication and collaboration between the clinical research and clinical care communities requires the use of common standards-based systems for data collection and management. Unfortunately, it is often *competing standards* rather than a *lack of standards* that inhibits interoperability between these communities.

To address this problem, the stakeholders from the Clinical Data Interchange Standards Consortium (CDISC), the HL7 Regulated Clinical Research Information Management Technical Committee (RCRIM TC), the National Cancer Institute (NCI), and the US Food and Drug Administration (FDA) have worked together to produce a shared view of the dynamic and static semantics that collectively define the domain of clinical and preclinical protocol-driven research, and the associated regulatory "metadata" that describes the clinical trial.

The caBIG® policies and tools that specify controlled terminologies, data element structure, data models, and computable metadata about those data elements are all openly developed, made freely available, and provide a pre-made framework for an effort like BRIDG. The caBIG® program has been a key partner and supporter of BRIDG and was instrumental in bringing the interested parties together at the outset. caBIG® continues to play a critical role in future plans to produce a similar data standards model for the non-clinical research space.

V. The BIG Health Consortium: 21st Century Model for Biomedicine

As the next step in its strategy to achieve Personalized Medicine, the NCI is pro-actively moving to working to break down the traditional silos of the biomedical enterprise and work collaboratively with all the key stakeholders that must be empowered in this new paradigm.

On September 10, 2008, the NCI convened 25+ leaders from academe, government, advocacy, policy, and commerce, to grapple with the daunting challenge of transforming the biomedical enterprise to achieve the benefits of Personalized Medicine and demonstrate that the “disconnected islands” of the 20th century can be reconfigured to improve health care. A new group, known as the BIG Health Consortium, was formally launched that day.

Mission and Goals

The BIG Health consortium is a partnership comprised of all the key stakeholders in health care: patients, providers, payers, product innovators, advocates, investors, and information technologists. Conceived by the National Cancer Institute (NCI), its mission is to show – in real settings, in real time – how and why personalized medicine works. Through a series of demonstration projects, BIG Health is modeling a new approach in which clinical care, clinical research, and scientific discovery are linked.

The goals of BIG Health are to:

- Demonstrate feasibility of implementing a new model of translational medicine
- Create an “ecosystem” of participants that seamlessly integrate research, care delivery and consumer health information
- Break down traditional silos that are barriers to rapid discovery and learnings
- Accelerate and enhance research productivity and improve clinical outcomes

Assembling a New, Integrated Ecosystem

The BIG Health Consortium™ is designed to foster an integrated and interactive ecosystem (or “mega-community”) of previously unlinked sectors within life sciences and health care, who gather to conduct demonstration projects to make Personalized Medicine a reality. Each participating organization is expected to share its capabilities, as well as to derive benefit, in order to have a self-sustaining endeavor.



Figure 5. Successful adoption of Personalized Medicine requires an ecosystem of stakeholders from every sector of life sciences and health care.

Among the organizations that are participating in the BIG Health Consortium[™] are cancer centers; integrated healthcare providers; academic centers; medical schools; diagnostic laboratories and product developers; personal genomics firms; patient advocacy and action-tank organizations; venture capitalists; biopharmaceutical companies; and government programs.

The informatics infrastructure of caBIG[®] will be generalized to “BIG” (Biomedical Informatics Grid) and applied as the underlying connectivity or “electronic glue” for BIG Health.

Demonstration Projects

BIG Health Consortium participants have agreed initially to conduct two kinds of demonstration projects, in cancer and other therapeutic areas, to show the essential unity of research and clinical care:

“Virtual” Clinical Research. Infrastructure and processes will be engineered to enable patients to be molecularly-profiled and pre-enrolled in clinical research, so that trials can be conducted without re-inventing the entire infrastructure for every new therapy. Such a

system would expedite the development of new diagnostics and therapeutics, as well as driving new knowledge rapidly into the regulatory system, the scientific literature, and the care delivery community.

“Learning Health Care System”. Activities will be undertaken to create a learning health care system in which data on health care outcomes are analyzed and correlated with molecular profiling data, in order to identify and validate biomarkers as drug targets and as tools for sub-grouping of populations to optimize treatment. Conversely, validated biomarkers can then be shared with the care delivery system to inform treatment and optimize clinical outcomes, in a ‘virtuous’ circle of discovery, knowledge, and practice.

Through these real-world projects, BIG Health can overcome many of the key obstacles to Personalized Medicine:

- **Cost/inefficiency of Screening:**
Builds screening into clinical care
- **Access to Study Populations:**
Draws existing patient base into clinical research
- **Acceptance in the Community:**
Provides a proactive role for the consumer
- **Misaligned Incentives:**
Provide alternative business models and partners
- **Data Disconnects:**
Provides the IT infrastructure to link entire process
- **Lack of Interoperability of Research and Clinical Systems:**
Provides a ready-made system of interoperability
- **Requirements for Systems-Level Effort that Daunt an Individual Company:**
Shares the “burden” of transformation
- **Lack of Knowledge Among Patients and Physicians:**
Provides a pathway for education

New Models of Interactivity

Just as BIG Health is developing 21st century biomedical models that unify basic research, clinical research and clinical care, it is important for this ecosystem to overcome the traditional limitations of time and space that slow research progress, in order to facilitate rapid and productive interactions among the participants. Thus, a BIG Health website (www.BIGHealthConsortium.org) is under development to provide Web 2.0 interactive tools that enable “networking” for community development of shared documents; organizational interface and coordination; and project management capabilities.

BIG Health Moving Forward

Plans are underway for convening dedicated working groups; for developing consensus around the Guiding Principles; for designing and implementing the Demonstration Projects; and for reaching out to a broader set of participants. All BIG Health activities will be conducted with the maximum transparency in order to network the largest possible participation within the biomedical enterprise.

##