U. S. DEPARTMENT OF HEALTH & HUMAN SERVICES

Identifying Opportunities To Maximize the Utility of Genomics Research Data Through Electronic Health Information Exchange

Meeting Summary

Dorfmann, Andrew
Nadler, Jessica

10/15/2009

This report presents a summary of the presentations and discussions at the workshop held by the Office of the Assistant Secretary for Planning and Evaluation, the National Cancer Institute, the Food and Drug Administration and the Office of the National Coordination for Health Information Technology on October 15, 2009. The report presents the views and opinions of the workshop participants and does not necessarily represent the views, positions, and policies of the Department of Health and Human Services or its agencies.

Table of Contents

Executive Summary

Health care has become a knowledge industry and, importantly, a growing sector of a knowledge economy. The basis of the healthcare knowledge industry is the increasing use of molecular, genetic and genomic information to tailor preventive strategies and better characterize and treat disease. It is therefore critical to maximize the ability to develop and deliver this knowledge for utilization in health care.

The discovery, interpretation and development of new information through clinical research, and delivery of improved information and knowledge through clinical decision support are interdependent, and must be connected to facilitate a rapid learning health care system linking new knowledge with knowledge delivery to the system. In a learning health care system, clinical information would be used to assess current standards of care and to generate hypotheses for basic and clinical research. In turn, new knowledge from clinical research would be delivered to health care providers through electronic systems, such as the electronic health record, to facilitate clinical decisions based on cutting edge information with the aid of electronic tools. Currently, data for clinical research and clinical care records are relatively separate due to historical and cultural aspects of these endeavors which resulted in independent and disconnected systems. The inability to use information for both purposes is a significant barrier to the development of new knowledge and its implementation in health care delivery. Incorporation of common data standards, architecture, platforms and networks for both clinical research and clinical care, with appropriate methods for maintaining information integrity, security and quality control can overcome this barrier. While the integration of information from both clinical care and clinical research is critical to advance the knowledge base for health care, it is not without potential risks. Clinical information from individual patients must be kept secure to maintain information integrity, ensure the privacy of the individual and respect their wishes about access and use of their data. These aspects of privacy and security can and must be built in to the standards, architecture, platforms, networks and governance associated with the use of personal health information. Furthermore, the results of clinical research should not be integrated into health care delivery before data is adequately replicated, reviewed and clinical relevance and utility understood. Experimental protocols cannot be carried out in the health care system without the express and informed consent of the patient. As with clinical care information, the standards and architecture of clinical research systems can and do support the appropriate consenting of research participants. Moreover, when the knowledge generated by clinical research is ready for implementation in clinical care, common information systems and networks can rapidly provide that information to clinicians in useful ways through clinical decision support tools.

Clinical genomics is a paradigm for an information intensive area that leverages the ability to integrate a rapidly evolving knowledge base into clinical care as well as the ability to provide investigators with clinical and outcomes information for research. As a relatively new discipline, it is an ideal area for harmonizing information standards supporting clinical care and underlying clinical research. As new as clinical genomics may seem to some, the application of genetic, genomic and molecular information in

clinical care is occurring now, making this the ideal time to address these issues. Furthermore, sophisticated algorithms (informatics) may be the only practical way to integrate data intensive clinical genomic information in order for it to be interpretable and clinically relevant. The development and usefulness of these informatics tools will require standardized data and information flow. In many cases there are existing standards and architectures that can support clinical genomics that will need to be implemented by the community. In some key areas, such as biospecimens, standards development needs to occur.

Ultimately, the goal of maximizing the utility of genomics research data through electronic information exchange is not the data itself, but the potential value of accelerating improvements in health care. In clinical medicine and clinical research important questions become apparent from well-organized, queriable, transparent data, accessible to knowledgeable scientists, clinicians, experts. In addition, emerging clinical guidelines can be readily communicated, in an actionable format, during clinical workflow, thus providing the opportunity to narrow the current time lag for clinical adoption of new guidelines. A system facilitating visibility, translation and transportability will more rapidly give rise to clinically relevant research based on real life data as well as better-informed application and implementation of new discoveries and emerging clinical guidelines.

Background

Genomics research is one of the most promising, fastest moving and extensive areas in biomedicine today. Substantial resources have been dedicated to genomics throughout the health agencies of the U.S. Department of Health and Human Services (HHS), in other departments of the Federal Government, through federally supported efforts in academia, and in privately supported research, including the pharmaceutical and biotechnology sectors. These investments reflect both the opportunity surrounding a new realm of biological knowledge and the expectation of high potential returns for improved effectiveness in health care and health outcomes. This expenditure has resulted in the production of a vast amount of data, previously unseen in most areas of clinical research, contributing to the evolving and dynamic health care knowledge industry.

Genomics includes broad domains of research; genome-wide association studies, population genomics, whole genome expression, protein and metabolism studies, companion diagnostic and pharmacogenomic applications. Genomic information is already being applied across a spectrum of public health and health care delivery activities. With such variety, it is not surprising that different research approaches and different data-reporting methodologies have arisen. What is less apparent are the opportunities that may be missed when data developed under different protocols or for different purposes are not captured in ways that could make the information useful for broad application across the health care enterprise—in research, in medical product development, evaluation and review, and ultimately in clinical care delivery.

The timing and potential for integrating genomic data exchange with the adoption of electronic data technologies is especially significant. In part, this is because both areas have simultaneously been enabled as a result of the decrease in the cost of computing power and data storage in the past decade.

The importance of data exchange for genomic research also stems from the large amount of information needed to discover and validate genomic findings and applications. In the longer term, information technology will be crucial in the delivery of genomic-supported medical care.

A key prerequisite for all these purposes is data standards harmonization adequate to support electronic exchange and analysis of genomics information. Although standards must allow for growth and change in the field, harmonization of terminologies and standards for minimal data input are essential for optimizing and leveraging this information rich resource. With these purposes in mind, HHS sponsored the stakeholder workshop "Identifying Opportunities To Maximize the Utility of Genomics Research Data Through Electronic Health Information Exchange." The workshop brought together leaders in government, academic, biotechnology, pharmaceutical, health information technology, and clinician communities to consider near-term opportunities for strengthening the foundation of clinical genomics and to identify areas for standards harmonization. Participants reviewed current processes of genomics research and standards in use, assessed the readiness and completeness of existing standards, identified gaps and barriers, and worked toward a consensus pathway for harmonization of existing standards, as well as a plan for coherent development of future standards.

In particular, the workshop examined activities and needs in the clinical research area, especially the need for data standards adequate to capture important elements in the workflows of genomic research. In identifying opportunities, the workshop focused not on uniformity in research approaches or analytical methods, but rather the ability to adequately capture, transmit and translate information about processes in a standardized manner. The workshop participants discussed consensus regarding a pathway for implementation of existing standards and technologies along with harmonization efforts that could be achieved within approximately a 3-year timeframe.

Clinical genomics is believed to have high potential for improving precision, effectiveness and outcomes in health care, as part of a movement toward more individualized, evidence-based or "personalized" health care tools and therapies. The role of data exchange is especially significant for clinical genomics, both as an important element of research and its potential for more rapidly integrating genomic knowledge and clinical information for better informed treatment decision making. This HHS stakeholder workshop was a step toward recognizing and consolidating early efforts, working together toward harmonization of data recording and exchange in the clinical research setting, and building toward the longer term goals of improved patient care through the use of genomic information.

Summaries of Keynotes, Panel Presentations and Discussion

Keynotes: Jesse Goodman, Daniel Masys, John Halamka, Elizabeth Trachtenberg

Jesse Goodman, MD, MPH, Chief Scientist and Acting Deputy Commissioner for Scientific and Medical Programs at the U.S. Food and Drug Administration (FDA)

Dr. Goodman opened the meeting by emphasizing the unique opportunity the scientific and medical community currently has to revolutionize the way health care is delivered by applying basic science discovery to clinical care intervention. The key to this paradigm shift is integration of genetics in medical practice through connections between biological and clinical information with genomic data. **Goodman** stated that we have the opportunity to "transform how we develop products and how we use medicines… genomic and more broadly systems biology information will help us do that." He went on to highlight two domains of clinical genomics information that are critical to the FDA mission: safety and product development.

In the safety domain, **Goodman** discussed the presence of polymorphisms in the population that can indicate the response to specific clinical interventions and the subsequent 'personalization' of using these interventions for appropriate patients. Enabling this level of understanding in the safety domain requires the development of an evidence base from clinical and genomic information currently available.

In product development, **Goodman** commented on the flood of genomic data generated in clinical trials and the need to be able to analyze it efficiently. Advances in genomic technology have provided vast amounts of data that cannot reasonably be managed without computational tools. For FDA reviewers to incorporate this data effectively into their work, they need to have access to both the tools and data in a format that can be analyzed. This necessitates the need for data standards for clinical genomic information.

Goodman noted that standards development organizations need to consider the end user of the information in order to produce useful standards. Over-proliferation of data standards can defeat the purpose of being able to exchange and analyze data effectively. Finally, particularly for the rapidly developing area of clinical genomics, data standards need to be adaptable over time. He ended with an appeal to the standards development community to aid those analyzing clinical genomics data in tapping the vast amount of information available to apply it to the complex chronic diseases that currently burden our health care system.

Daniel Masys, MD, Professor and Chair of the Department of Biomedical Informatics and Professor of Medicine at the Vanderbilt University School of Medicine

Dr. Masys described the Electronic Medical Records and Genomics (eMERGE) Network which ties phenotypic data extracted from electronic health records (EHR) with genotype information across five different clinical areas from five medical centers. The integration of data across these centers had both scientific and technological challenges. Scientifically, it was necessary to determine if patient populations at different centers were sufficiently similar to draw significant biological conclusions. Technically, data exchange standards were necessary for the sharing of genotypic and phenotypic data. **Masys** emphasized that this project is an *immediate* consumer of standards for this type of information.

eMERGE highlighted several issues in data management, including the ability to represent change in clinical features over time, the handling of uncertainty in clinical data, integration of other types of data,

such as billing data, and the potential for re-identification of individuals from this information. The network has been active in working on these issues, and while they are not yet resolved, the process has been instructive about standards development.

A major challenge for eMERGE is the consistent and specific definition of disease phenotype. Structured data, in its current state, is not sufficient to define a phenotype for use with genomic data. For example, billing codes are often not clinically specific and can be applied variably to increase reimbursement, leading to false negatives and false positives in the data. Natural language processing (NLP) has been critical for the identification of specific clinical phenotypes in EHR records. The implementation of NLP is a difficult process, but the computer selection logic (pseudocode) can be implemented at various sites with unique systems.

Masys also discussed the introduction of genomic and proteomic data as an exemplar for complexity in medicine. He emphasized that the human cognitive capacity is exceeded by the thousands of parameters introduced into clinical decision making by genomic information. He argued that computer aided clinical decision support will become necessary to use this information practically in a clinical setting. He further pointed out that the current way this information is reported by laboratories performing genetic and genomic assays is functional for a single clinical decision at one point in time, but potentially important information for the future may be lost. To resolve this issue for the eMERGE network standards have been implemented.

The most utilized standards are information exchange standards, such as TCP/IP internet protocol suite and Health Level 7. Knowledge representation standards are less widely used and adopted. International Statistical Classification of Diseases and Related Health Problems (ICD9) and Systematized Nomenclature of Medicine (SNOMED) are two of the most well known representation standards. These are typically designed by the government to standardize some aspect of health care delivery. In **Masys'** experience, the common and usable standards always persist over the theoretically accurate, but difficult to implement. A current case is HL7 V2.x versus V3. V3 is more powerful, but too complicated to understand and use. As with systems, **Masys** paraphrased Dr. Donald Lindberg, Director of the National Library of Medicine, saying that "standards that get used, get better". They do not need to be perfect at their inception, they need to be sufficient to perform, published and used.

John Halamka, MD, MS, Chief Information Officer of Beth Israel Deaconess Medical Center, Chief Information Officer of Harvard Medical School

Dr. Halamka described impact of the American Recovery and Reinvestment Act on health information technology (HIT) and its standards. He noted that 400 of its 1000 pages were devoted to the Health Information Technology for Economic and Clinical Health Act (HITECH). These priorities are as follows:

1. Privacy is foundational

Patients and consumers need trust the systems implemented to handle their health and genomic information or they will not be willing to share their data. In order to maximize the utility of data collected during the course of clinical care, data must be contributed by patients and consumers. In

Halamka's experience, the vast majority of people, if they have confidence in the information systems, will consent to the use of their data.

2. Health information must enabled for both use and exchange

Physicians will capture phenotypic and genotypic information using EHR systems and be able to exchange it for the purposes of clinical care and research. This exchange is enabled by standards and data architecture for health information. The National Health Information Network (NHIN), which will be a system of standards, policies and architecture, combined with great governance, will provide a secure environment for the exchange of this data between providers, payors and patients. **Halamka** emphasized that the National Coordination for HIT, David Blumenthal, has indicated the NHIN will be consumer-centric and present an avenue for the setting of preferences for data sharing by the patient.

3. Certified EHR for each person

Every clinician, by 2014, will have a system that can capture information for each patient.

4. Accounting of disclosures

HITECH requires the ability to provide full disclosure of who has viewed a health record to the patient. This will facilitate the privacy and security of health information by recording if the information has been viewed by unauthorized entities.

5. Improve the quality of the data

Data will be captured in a structured format to enable use and exchange beyond what would be possible if records are merely an electronic version of their paper incarnations, such as a PDF file.

6. Encryption of the data for exchange

Rendering the data unusable or indecipherable when moving it from place to place to ensure that data can only be viewed by authorized users.

7. Comprehensive collection of patient demographic data

EHR systems will need to capture race, ethnicity, primary language and level of education.

8. Address the needs of the vulnerable

HITECH emphasizes the use of HIT to reduce health disparities and to protect the needs of vulnerable populations, such as children and the elderly.

Halamka outlined the work by the Health Information Technology Policy Committee requires EHR systems to be capable of electronic reporting of full lab results, including genomic testing results, by 2011. In addition, e-prescribing, medication lists, reporting of quality measures, immunization records and administrative transactions will also be required. The HIT Standards Committee informs the Policy Committee about what standards are implementable for these data exchanges. There is a mechanism by which the Standards Committee can learn from the HIT community which standards have been implemented successfully. Standards are available or under development for these functions and can be found at the Health Information Technology Standards Panel (HITSP) website. Work is ongoing by

groups, such as CDISC and HL7, to fill gaps in the existing standards to connect phenotypic and genotypic information.

Halamka also talked about his experience as a participant in the Personal Genome Project, which makes public his phenotypic and genotypic information on Google. The analysis of his genome uncovered an increased risk for glaucoma, for which he had never exhibited symptoms. As a result, he had his inter-ocular pressure measured, discovered it was high and sought treatment to lower it. On further discussion with his family, **Halamka** learned that both his father and grandfather experienced vision issues at early ages. He related that the knowledge of his genotype and family history has allowed him to have a better understanding of his health. This is an example of the promise of clinical genomics to support disease prevention in addition to treatment.

Elizabeth Trachtenberg, PhD, MS, Director of the Center for Applied Genomics, Children's Hospital & Research Center Oakland

Dr. Trachtenberg described her work with human leukocyte antigens (HLA) and the transplant research and care community that uses this information. HLA are highly polymorphic loci that are difficult to genotype unambiguously due to their high level of sequence similarity; to date over 4000 variants have been identified. HLA genotype is critical for matching transplant donors and recipients for successful transplant outcomes. This community, therefore, had to come together to develop and implement standards to allow the international exchange of information for both research and donor matching. **Trachtenberg** described the groups and consortia working with HLA informatics, which has over 40 years of cooperative data. This rich database includes multiple types of information, including cellular, genetic and reagent information, outcomes analysis, population analyses, meeting proceedings and research results. Other resources include the American Society of Histocompatbility and Immunogenetics which provide good models for developing data standards and accreditation bodies, proficiency testing and nomenclatures.

Trachtenberg discussed the scientific challenges of her work and the opportunities presented by new sequencing technologies. These new technologies, however, will create more issues with nomenclature for these loci. She concluded by emphasizing the need for better data consistency, consistent application of analytical methods and better data portability, all of which require consistent and agreed upon standards.

Keynote Discussion

Questions about both data curation and public availability of data were addressed to Dr. Trachtenberg. There was a general discussion of the need for patient involvement in the exchange of information. Finally, it was pointed out that "data is not medicine." Therefore, it is important that the evidence base is developed with regard to genetic and genomic information in the context of health care delivery. When discussing this type of information with patients, there is a danger of overemphasizing the clinical value of the information when the medical and biological significance is not yet clear.

Panel 1: Clinical Genomics and Standards Development

Rebecca Kush (CDISC) addressed the questions of 'what are standards, what do they mean and who needs them?' She gave the example of the universal serial bus (USB) drive as a common standard that does not restrict the data they contain or inhibit creativity in their contents, rather they make moving the data/information far more efficient. **Kush** described standards in use today that pertain to clinical genomics information, such as the lab data standards, the FDA data eSubmission standards, the CDISC Clinical Data Acquisition Standards Harmonization (CDASH) standard for data collection in research studies, and the Biomedical Research Integrated Domain Group (BRIDG) model. **Kush** emphasized the need to align clinical research with the rest of clinical care. This can be accomplished through the development of common platforms and standards that can be used by both communities as evidenced through the joint CDISC-HL7 (Health Level 7) effort to develop such standards within the HL7 Clinical Genomics Workgroup. She also discussed the HITSP Clinical Research Interoperability Specification, which defines standards for the exchange of a core dataset from EHR systems into clinical research systems. This serves as a foundation for the exchange of more information types between these two systems, such as genomic data.

Kush called for standards development organizations (SDOs) to work together on standards development and emphasized the need to align and harmonize standards among groups using common terminologies for content. She stated that there are genomic standards being developed and they need to be harmonized and implemented by the community.

Philip Pochon (Covance) reiterated that data structure is foundational, but argued that the efficiencies are gained through the implementation of the semantic structures for communicating common information. He also spoke about HL7 V3 standards, which are more elegant then necessary, while HL7 V2 is in wide use.

Pochon then focused on genomic information, which can be stratified into levels of interpretation or analysis into raw data, significant findings and clinical interpretation of variation. There are standards for the encapsulation and exchange of raw data. The strength of this data is the ability to use a wide variety of analytical tools. Significant findings include genetic and genomic variations, including mutations, polymorphisms and gene expression changes. Finally, clinical interpretation of variation includes the biological consequences associated with a detected genetic or genomic variation. Standards for these levels of interpretation include HL7 Clinical Genomics models, genetic variation common message element type (CMET), gene expression CMET and the CDISC Study Data Tabulation Model (SDTM) pharmacogenomics domains. **Pochon** concluded that success is dependent not only on the availability of standards, but that Implementation guides are critical to their utilization.

Mollie Ullman-Cullere (Partners HealthCare Center for Personalized Genetic Medicine) began by outlining the requirements for the use of genomic data in a clinical setting. She argued that the genetic/genomic data must be:

a. Structured in the EHR

b. Available to clinical decision support (CDS)
c. Integrated into data warehouses
d. Supported by an intelligent interface engine
e. Annotated with current clinical interpretations

HL7 and LOINC have collaborated to produce a standard for reporting fully qualified clinical genetic test results and their associated interpretation. HITSP has recognized this standard as supporting the American Health Information Community (AHIC) Personalized Healthcare Use Case. Partners HealthCare has integrated clinical genomics from their Laboratory for Molecular Medicine into its EHR, including associated CDS. In addition, these data were also appropriately integrated into their institutional review board approved research patient data registry. This first implementation highlighted the need to better codify the data and to use more robust and consistent structure. These lessons learned were incorporated into the current HITSP approved HL7/LOINC standards for clinical genomics. **Ullman-Cullere** described the ability to exchange clinical genomic information, using HL7 V2 messaging, with the Intermountain Healthcare System in Utah. This enables data sharing and care delivery as an integrated provider system with repositories, lab information systems and the Partners molecular genetics laboratory.

Ullman-Cullere described the clinical information workflow incorporating clinical genomic information, beginning with the ordering, performance and reporting of a genetic test. Upon results reporting, identified variants and associated clinical interpretations are displayed to the clinician. CDS and a genomic knowledgebase alert the clinician to additional or new information not in the report (e.g. association with an additional drug/disease or reclassification of the variant). For example, over time we learn more about the clinical implications of variants of unknown significance. This new information can be useful for patient care if it is integrated into the care delivery process. This integration is only feasible with appropriate IT support. **Ullman-Cullere** emphasized that personalized medicine will require significant IT support due to the volume and complexity of information and the extent of data of unknown or changing clinical significance. Standards for genomic data will be required for data exchange and clinical decision support since data will be coming from multiple sources. There are data model extensions underway to cover the following aspects of this information integration, including the representation of large sets of DNA markers with no clinical interpretation, toxicogenomic results and interpretation, and support for complex genetic disease.

John Quackenbush (Harvard University) opened the panel by describing the Tower of Babel problem associated with data standards, referencing the adage that “the great thing about standards is that there are so many to choose from”. He went on to describe his work with the Microarray Gene Expression Data Society (MGED), which began in response to the Genome project. They found that expression data was not useful without metadata. MGED developed a “standard” for genomic data, but did not develop an implementation, which merely made the data larger and more cumbersome. So they revisited their work and made the standard useful with more common tools.

Quackenbush summarized the lessons learned from MGED. First, to develop a good standard, you need to demonstrate its utility. Second, he echoed the sentiment that standards will be imperfect at the outset and need to be designed to enable evolution. Third, he reiterated the need to be able to mine unstructured data. He concluded with the forecast that direct to consumer genomics is on the horizon and will generate vast amounts of data. Standards and architecture need to be able to handle it.

Martin Will (Microsoft) discussed the need to shorten the lifecycle of knowledge from discovery to clinical practice. He described the BioIT Alliance as a community committed to facilitating this through the use of information technology.

Frances Schrotter (ANSI) described the HITSP process to develop interoperability specifications that select standards for specific functions of an information exchange. She also reiterated that HITSP work products feed directly into the HIT Standards Committee in its role under HITECH.

Betsy Humphreys (NLM/IHTSDO) emphasized the necessity that standards for healthcare and research data converge. She argued that the end-data consumer need not be aware of standards embedded in tools that they use regularly to do their work; standards can be implemented into programs "under the hood." **Humphreys** strongly advised against more standards, since there are many extant standards that can be improved through use and refinement. She further advocated for data standardization upon collection of raw data. Data itself must be structured for maximal utility, not just interpretation of the data's meaning. **Humphreys** discussed several key terminology standards for clinical genomics, including SNOMED, Logical Observation Identifiers Names and Codes (LOINC), RxNorm, the newborn screening coding and terminology guide, the National Center for Biotechnology Information Reference Sequence Gene, the ClinicalTrials.gov and PhenX (phenotypes and exposures) and DBGaP (genotypes and phenotypes) databases.

Mike Glickman (International Standards Organization) described ISO, an organization founded ten years ago, involving 40 countries. The experience with ISO indicated the need to harmonize work early in standards development, rather than after standards have been published and implemented. **Glickman** advocated for open, usable, consensus based standards. He also mentioned that the BRIDG model is currently going through the process of becoming a harmonized standard through the Joint Initiative Council (HL7, CEN, ISO, CDISC, IHTSDO), which is devoted to global harmonization of standards.

Panel 1 Discussion

Susan Shurin asked the panel about the information and standards needs of long-term projects like the National Children's Study. This project will be collecting data for a quarter of a century, and will need to be able to reanalyze data in the future for questions that may not have been foreseen at the outset. **Shurin** wondered how to align this type of study design with the future of standards activities. The panelists agreed that information integration is a major problem in large scale projects. Unfortunately, funding mechanisms are heavily skewed towards data generation and often do not include sufficient resources for standards development and data architecture. Currently, standards and tools are an intellectual exercise undertaken as a public service. Another suggestion was the use of open and

available standards and software to have a better chance at having longevity in the standards. It was reiterated that groups should refrain from generating new standards, instead working to extend those being implemented.

The panel also discussed the need for genetic data to be available throughout the lifespan of the individual and may be reinterpreted multiple times throughout that lifespan. Indeed, one challenge is that research is about ten years ahead of standards development.

A question was asked about some of the standards that have grown out of the MGED work, specifically Minimum Information for Biological and Biomedical Investigations (MIBBI) and Minimum information about a microarray experiment (MIAME). **Quackenbush** described these as an effort to collect all the standards together and gather up the core data for various kinds of genetic analysis (fluorescence *in situ* hybridization (FISH), metabolomics, rtPCR, etc). The effort is focused on the metadata for these experiments and is still under development.

Panel 2: Information Technology Infrastructure to Support Genomics Research

Jeffrey Elton (KEW Group LLC) began the panel by discussing how genetics is utilized in both clinical care and clinical research. He predicted that molecular diagnosis will be the new basis for diagnosis and treatment of many diseases and that the distinction between data for clinical decisions and data for clinical research will continue to blur. With that future in mind, it is imperative that standards are sufficiently flexible to accommodate new technology and methodology. Increasingly, clinical decisions are being driven on the basis of genetic information; lung cancer is a primary example. A genetic test for the epidermal growth factor receptor (EGFR) can predict who will respond to specific chemotherapeutics (Iressa or Tarceva). Over time these patient stratifications will become more refined as we gain knowledge about the molecular nature of disease.

Elton also forecast that with the decreasing cost of genome sequencing for patients, the bottleneck in clinical research will become storing, mining, analysis and interpretation of these data. Discovery will become an information management problem, not a DNA sequencing problem. This is particularly true for the delivery of clinical care and presentation of clinically actionable information to care providers

Robert Plenge (Harvard/BWH) outlined the need to efficiently collect enough DNA and detailed clinical data that can be used for clinical studies. To do these studies using data collected during the course of clinical care requires the ability to glean data from electronic medical records. He described a project using the Partners HealthCare EHR system to identify rheumatoid arthritis (RA) patients from a 4 million patient population, 31,000 of which received an ICD-9 code for RA. The definition of RA as a phenotype needed to be specific enough to enable genome-wide association studies.

At Partners, the EHR and clinical research systems were separate, so the group initiated Informatics for Integration Biology and the Bedside (I2B2) with National Institutes of Health funding, to develop and distribute open source software that would integrate these two data sources. The RA study collected

discarded blood samples from RA identified patients to extract DNA and connect to clinical information about the phenotype of these patients.

Martin Will (Microsoft)discussed how to bring genomics and the IT worlds together in a meaningful way. He asserted that data should be independent of the physical manner in which it is stored. Furthermore, data itself has no purpose, it is only useful as information. So the use of the data should drive how it is handled and stored from the outset.

Ken Buetow (National Cancer Institute) stated that genomics is a component of a complex information ecosystem that encompasses many types of data, including data for clinical research, clinical pathology and imaging. Data is produced in multiple sites and requires integration for its utilization. **Buetow** said that many different communities generate data about disease (phenotypic information) and the challenge is connect all these domains of data. This requires semantic and syntactic interoperability. Semantic interoperability allows systems to understand the data exchanged via shared data models depending on standard vocabularies and common data elements. Syntactic interoperability allows systems to exchange data through shared interfaces.

Buetow echoed the sentiment of previous speakers when he exhorted users not to proliferate more standards. He suggested they work with standards currently being implemented and extend them; he commented that they use HL7 and CDISC standards and contribute to these development processes. He also described the Cancer Bioinformatics Grid (caBIG) platform which has developed standards-based open source tools to enable clinical research methodologies, such as adaptive clinical trials

Panel 2 Discussion

Robert Plenge was asked what the key was to getting that data out of the clinical system. He replied that natural language processing (NLP) was critical for their success. The core problem was how clinical pathology reports are dictated and the variation between clinicians. Pathologists themselves will interpret these reports differently. The question was asked whether NLP pseudocode can be shared across EHR systems and across organizations. **Buetow** answered that this capability is under investigation. A follow up question was whether genetic data is less problematic because it is clean, whereas pathology data is variable and unstructured. The response was that many situations may never be structured, so NLP will be necessary to glean useful information from the database. **Richard Cotton** added that there is a system in Australia called “biogrid” which can interrogate clinical systems, but there is no similar system in the United States. That could be a useful method to pursue, however, privacy policies in the two countries differ.

Ken Kawamoto asked if it was possible to insist on some standards like SNOMED as an incentive through the Federal Government. **Plenge** replied that adoption of standards is most likely to come from improvements to clinical workflow.

Panel 3: Genomic Experiment Platforms and Mutation Test Registry

Michael Christman (Coriell Institute) moderated the panel and opened with a description of the Coriell Personalized Medicine Collaborative, a project to collect genome information on participants to determine best practices for using it in clinical care. There are already about 45,000 participants in this prospective study. It contains three population arms, one of the general population, one specific for cardiovascular disease and on specific for cancer. The project directly involves clinical partners and the information can be used for discovery and returned to the patient. A subset of the data generated will be clinically useful; this subset is the only information that will be returned to the patient. A chief goal of the study to find out what information is indeed clinically actionable. Decisions regarding the information returned to the patients are made by the Informed Consent Oversight Group, which functions like a study section. The clinical and genomic information will help develop quantitative risk reports in combination with lifestyle data. This is the basis of a personalized manner of reporting clinical risks. This is a long-term longitudinal study, which is extremely important for the generation of an evidence base for genomic medicine.

David Carey (Geisinger Health System) described Geisinger as a large health network in central Pennsylvania, which is both non-academic and non-profit. Geisinger Medical Center is the central hub (five inpatient; 40 outpatient clinics) of a hub and spoke model for the network. **Carey** asserted that Geisinger was a good place to test out new technologies because it cares for a population of 3 million patients that is largely non-transient, older and poorer than most urban centers. The network has also made a specific commitment to HIT for both clinical research and the improvement of care.

Geisinger initiated the "My Code" biobanking project, which consists of two patient cohorts, one representing the general population and one disease-specific. Consent for participation is an opt-in model, so their patients are actively interested in participating in the study. The disease-specific cohort is recruited from specialty clinics where patients have already received a diagnosis. This cohort has enrolled 11,000 patients. The population cohort is recruited from outpatient clinics and has broad inclusion criteria. This cohort will have longitudinal accrual of patients and phenotypes, and has enrolled 12,000 individuals.

Recruitment has been done using EHR data. De-identified records are retrieved from the EHR system through a data repository. Upon identification of specific candidates, researchers can gather full records from the EHR. With systems in place, this will produce a great deal of genomic-phenotypic information over time.

Clay Marsh (Ohio State University Hospital) focused on the support of wellness rather than care of disease. He argued that health solutions should be portable and scalable across a variety of delivery systems and forecast a future where more health care is performed at home than in a provider's office. This state will be enabled by programs to make health care easy to use. This highlights the need for in a format that supports clinical decision-making.

Richard Cotton (Genomic Disorders Institute Melbourne) discussed the Human Variome Project, which strives to codify all the mutations present in the human population. This information is intended to meet specific needs of the clinical genomics community by defining variations and their clinical interpretations. For inherited diseases, mutations that can be used for prognosis are defined. For common diseases, polymorphism associations can be connected with risk assessment. Coordination and information flow are key to determining a sufficient level of evidence to make clinical decisions based on the identification of genetic variation. The Human Variome Project is part of a network of information sources and governing bodies that have the purpose of relating phenotypes to genotypes.

Sharon Terry (Genetic Alliance) spoke of the organization of the Genetic Alliance Bio Bank. This project was the direct result of activity by patient advocacy groups of parents of children with rare Mendelian disorders. The biobank developed a sophisticated privacy protocol for their data, to ensure patients and their families of the security of their information. Genetic Alliance formed a public-private partnership with Private Access, to develop a privacy layer for the Bio Bank. This privacy layer is similar to the Consumer Preference model from Google. Patients using it can define access permissions to their data in accordance with their comfort level and desire for its usage. **Terry** argued that people should have access to their own data, despite objections from physicians and other professionals. She stated that policies are moving towards greater personal control of health information. She concluded by emphasizing that the opt-in model for consent in clinical research demands that patients are comfortable with their level of control over their own data.

Panel 3 Discussion

Michael Christman commented that genetic information will likely be compromised in a manner similar to identify theft, so laws must be created to protect the individual from the damage that can ensue.

A two-part question was directed to the panel, the core of Personalized health care will depend upon the use of this information, so 1) what is being done now at various institutions, and 2) what should be the model for data availability, a national resource like the NLM or a more federated model? **Christman** responded that at Coriell, access begins with data that is de-identified and then made available to researchers.

The panel commented on the need to validate data to progress towards evidence based medicine, while balancing the need to provide data to primary care physicians in a useful way. Primary care providers are not geneticists, and have only limited time with patients. The need of the care provider can be met through CDS comprised of more than alerts. This CDS would instead have dashboards and more integrated information for clinicians. Currently, only about 6% of hospitals have the infrastructure for even rudimentary CDS. The introduction of clinical genomics will further exacerbate this problem, highlighting a need to change the culture of the health care delivery system.

Upon being asked how Geisinger got such high opt-in rates for its study, **David Carey** responded simply, “trust.”

Panel 4: Statistical/Biological Analyses and Clinical Trials

Joyce Hernandez (Merck) presented two Merck bioinformatics oncology studies, both using the CDISC standards in order to standardize data for analysis. Rather than focusing on a single gene, the studies expanded to the entire gene signature. This required various sources of information, including public databases, experimental cell line panels and patient profiling. This data all had to be integrated to reach a biological conclusion. **Hernandez** cited standards development done through the HL7-CDISC Clinical Genomics Workgroup. There is an SDTM domain in progress for clinical genomics data.

Weida Tong (FDA) related lessons from the FDA's voluntary genetics data submission program (VGDS). This program was initiated to start receiving and analyzing genomic data independent of FDA review in advance of large amounts of this data being submitting during the review process. Data submitted to the VGDS has changed over time, early data was mostly from expression arrays; later submissions were primarily from genotyping data. The FDA developed an informatics tool to handle all the incoming data, called ArrayTrack, which has been expanded to handle proteomic, metabolomic and other data types. The project has worked toward addressing quality control issues, such as consistency of microarray data. The VDGS has also attempted to reach consensus on analysis methods for submitted data. Finally, the FDA receives data from a wide variety of sponsors, and thus a wide variety of technical platforms, leading to difficulty using common analytical tools and an inability to aggregate data across studies. The experience at the FDA led to a guidance document regarding pharmacogomic data submissions. **Tong** asserted that standards will help the aggregation and analysis of data and the FDA seems to be promoting CDISC as a data standard for submissions.

Panel 4 Discussion

The panel was asked what the actual level of combinability was across technical platforms. **Tong** replied that it depends on the level of aggregation to be achieved. It is unlikely that there will be a perfect method for combining data across studies, however an imperfect method that is implementable is a possibility. A start is to maintain data in a format that is standards compliant. Data for ArrayTrack is less problematic to combine across platforms, but other data types, such as metabolomic and proteomic lag.

The FDA is now receiving over 90 submissions a year containing various kinds of genetic data, not only microarray data. FDA is working on a database to compile all the data, since it will be relevant to multiple clinical phenotypes. This compilation is not being done at present, but it is a significant need for the FDA.

Another question asked was how to model knowledge, instead of just data? How is information made accessible. **Elizabeth Tracthenberg** replied that City of Hope has a model system related to toxicology. This is suggestive of a service oriented architecture for the clinical knowledge base.

Panel 5: Biospecimens

Helen Moore (NCI)presented take home messages from her work at the Office of Biorepositories and Biospecimens Research. Biospecimens are obtained with varying specimen collection methods, varying

specimen data, varying clinical information and varying restrictions on the use of the specimen. All of these pose problems for the use of the biospecimens for clinical genomics research and cause variation and/or inaccuracies in the data generated. **Moore** argued for more standardization of specimen handling and information gathering. **Steve Gutman** (University of Central Florida) also presented examples of variability in data generated as a result of biospecimen handling and metadata.

Lynn Bry (Partners HealthCare) described three challenges to Genomic Research, high throughput phenotyping, high throughput genotyping, and high throughput sample collection. To address the last challenge, Partners developed a sample collection protocol called Crimson. Crimson identifies specimens, such as discarded blood samples, generated during the course of clinical care and matches them to various research projects in the Partners network. Samples are identified for studies from clinical information in the Partners EHR system, using both structured data and natural language processing.

The process can work through either a de-identified protocol, or the sample can be archived for a short period to allow patient consent to be obtained. The use of Crimson has produced significant cost savings and efficiencies contributing to the quality of the studies and the ability to do studies that would otherwise have been cost-prohibitive.

Panel 5 Discussion

A question was posed regarding how Crimson introduced efficiencies to reduce costs to such an extent. **Bry** responded that some of the efficiencies are gained from gleaning the specifics of the research study needs up front. Once the needs are understood, some economies of scale are achieved. Furthermore, investigators often receive higher quality specimens for their particular study.

Panel 6: Applications in Clinical Genomics Databases Today and Tomorrow

Leslie Glenn Biesecker (NHGRI) presented the ClinSeq study, a high throughput genome sequencing project. This project generates huge quantities of both sequence data and clinical data, more than 16 TB per subject, and it is necessary to develop a robust infrastructure to perform analyses to answer biological questions. The first biological question revolves around atherosclerosis. One thousand subjects are to be enrolled and their genomes sequenced. Large numbers of variants will be identified in the population of 1000. Some of these will be involved in the development of atherosclerosis, or other clinical conditions, and can be validated.

ClinSeq is an example of hypothesis generating research. This type of research is important for clinical genomics, since the data may provide answers to questions that have not yet arisen. This is the antithesis of the classical approach to research, which is to identify phenotype, then try to find the genotype to match the phenotype. **Biesecker** emphasized that this is a paradigm shift in medicine and medical research. This approach generates genomic scale data, stratifies patients on the basis of genotypic not phenotype data, then associates the genotype with the phenotype.

The project is already generated interesting data. **Biesecker** has found that rare variants, found in a single individual, are common, lots of individuals have different variants. He also warns that this approach will not be well received by physicians or patients, because much of the genomic information will not be clinically actionable.

Mark Allard (FDA) described a project generating high-coverage sequencing of *Salmonella* strains for the rapid identification of food borne pathogens. The project is developing the knowledge to be able to detect species, identify serotype and track sub-type to perform molecular epidemiology on outbreaks of food borne illness. In-depth sequencing data is necessary to identify marker polymorphisms in bacterial genotypes to positively identify pathogens from the food back to the farm. This project generated a large amount of sequence data and required a bioinformatics team to analyze it because standardized tools were not available for this type of research.

The Molecular Methods and Subtyping Branch of the Center for Food Safety and Applied Nutrition is also interested in using the sequencing capacity developed in the *Salmonella* project to develop methods for sequencing a "meta-genome". This is defined at the suite of organisms present in a particular food or soil sample that may, as a whole, contribute to an outbreak of food borne illness or to clinical variation among outbreaks.

Martin Maier (National Bone Marrow Donor Program) presented the process of matching of HLA alleles between unrelated bone marrow donors and recipients. Only 30% of patients needing bone marrow transplant have a donor match, typically within their family. The National Bone Marrow Donor Program helps match the remaining 70% of patients in need of a bone marrow transplant with donors. HLA matching begins by querying archived donors at a low resolution and gets more fine tuned at higher resolution on a selected group. Genetic variant alleles of the HLA loci are being discovered at an extremely rapid rate. Matching at HLA and the major histocompatibility complex (MHC) are both difficult and the loci are so complex that it is difficult for clinicians and patients to understand the clinical implications of the molecular information. Thus, the necessity for an informatics approach.

Data exchange standards are under development to support the informatics approach, but HLA genotyping is an inherently ambiguous system and therefore very difficult to standardize. It presents a good model for other highly polymorphic loci. Two groups have been formed to address the standards and informatics problem, Immunogenetics Data Analysis Workgroup (IDAWG) and A Growable Network Information System (AGNIS).

Panel 6 Discussion

The panel was asked if there are generalizable lessons to be applied to other genomic systems. They responded that it is important to present complex data in a consumable form for physicians and patients without losing the information. Also, nomenclature is crucial to this kind of data, which means the standards for nomenclature are important as well.

Direct to consumer testing was discussed by the group. Whole genome data released to consumers has the potential to be problematic. Currently, this presents too much data without clinically actionable

information. Clinicians will be more comfortable with this type of data when there is evidence supporting clinical care choices based on the genomic information.

Summarizing the Workshop

The workshop planning committee co-chairs, **Rebecca Kush** and **Mollie Ullman-Cullere** summarized the discussions of the workshop and suggested potential next steps for the clinical genomics community.

Common themes highlighted from the workshop:

- Augment existing standards, rather than contribute to the proliferation of new ones.
- Complete and implement submission standards for FDA to support reviews that can utilize common tools and aggregate data across studies.
- Standards must be able to evolve and support long-term longitudinal studies, such as the NICHD National Children's Study.
- Standards are required for polymorphisms and mutations and that can be used in CDS.
- Standards must be developed and implemented for biospecimen collection and handling.

Kush and **Ullman-**Cullere summarized opportunities discussed during the workshop:

- An SDO could undertake the development of a lab order message for EHR systems.
- Systems could implement the new HITSP interoperability specification for clinical research and augment the core dataset with genomics information, as well as interoperability specifications supporting the AHIC Personalized Health Care Use Case.
- Groups can participate in the development of the SHARE database of controlled terminologies.

It was also concluded that standards development work and investment in data architecture are required to support and advance improvements in clinical care and an effective and efficient clinical research endeavor and should be resourced appropriately.