International Business Machines Corporation

6710 Rockledge Drive
Bethesda, MD 20817-1826

February 5, 2007

PHCRFI@hhs.gov

U.S. Department of Health and Human Services
Room 434E, 200 Independence Avenue S.W.
Washington, D.C. 20201
Attention: Personalized Health Care RFI

Thank you for the opportunity to provide input on the current and future landscape of personalized healthcare.

Our response reflects IBM both as a major provider of IT solutions for the healthcare and life sciences industry, and also as a purchaser of healthcare for our 300,000 employees. We have tapped the expertise from many parts of IBM to prepare this response.

You can contact me if you would like any additional information or clarification from IBM, or would like to meet with us to discuss our response.

Sincerely,

Nancy Friedland
IT Architect
U.S. Federal Health and Social Services

IBM Global Business Services

February 5, 2007

U.S. Department of Health and Human Services

Personalized Health Care RFI

Submitted to: U.S. Department of Health and Human Services
Room 434E, 200 Independence Avenue S.W.
Washington, D.C. 20201
Attention: Personalized Health Care RFI

Submitted by: IBM Corporation
Nancy Friedland
6710 Rockledge Drive
Bethesda, MD 20817-1826
Telephone – 314-252-5821
Email address – nefried@us.ibm.com

RESTRICTION ON DISCLOSURE AND USE OF DATA

The information in this proposal shall not be disclosed outside the U.S. Department of Health and Human Services organization and shall not be duplicated, used or disclosed in whole or in part for any purpose other than to evaluate the proposal, provided that if a contract is awarded to IBM as a result of or in connection with the submission of this proposal, U.S. Department of Health and Human Services shall have the right to duplicate, use or disclose the information to the extent provided in the contract. This restriction does not limit the right of U.S. Department of Health and Human Services to use information contained in the proposal if it is obtained from another source without restriction.

IBM's products and services are proposed under the terms of the IBM Customer Agreement.

The IBM home page can be found at: http://www.ibm.com.

Table of Contents

Section **Page**

Introduction 1

1. Concepts on anticipated approaches for the use of EHR and population- and community-based health care system databases for longitudinal data collection in addressing: 2

2. Anticipated Applications of Genomic-Based Clinical Testing in medical decision-making, safety assessment, and risk management 5

3. Establishment of biospecimen resources obtained from clinical medical services for application in research, clinical trials, health services planning, clinical effectiveness, and health outcomes evaluations 10

4. Organizational or institutional practices to address ethical, legal, and social implications regarding the use of patient information, including genetic data, to support personalized health care 14

5. Examples of utilizing large clinical data repositories for practical clinical research to discover effective technologies, therapeutics, diagnostics, and prevention strategies for different populations 20

6. Issues and challenges associated with incorporating genomic information as a part of a broad longitudinal data collection 24

11. Examples of the use of disease registries to track specific diseases and response to drug therapies across different subpopulations 24

7. Needs for community-wide standards or best practices that will facilitate large-scale data integration and exchange to benefit personalized health care 28

9. Development of ontologies across different clinical data repositories that will facilitate the utility of the data for answering clinical research questions 31

10. Models for linking clinical data repositories across disparate care providers 34

12. Models for prioritizing analyses to fill gaps in evidence of effectiveness of therapeutic interventions for different populations 44

13. Strategies for accumulating patient data necessary for research that may not be available through EHRs 47

14. Concepts or models on the potential use of clinical data and related resources for research applications 51

15. Models of cost-benefit analysis for integrated data systems, EHR, and clinical resources to inform medical decision-making 55

8. Feasibility and potential benefits for establishing linkages of institutional or organizational data resources with private and publicly available health databases 61

16. Opportunities and challenges for the development of electronic tools to aid in the integration and analysis of large datasets of clinical parameters to assist in outcomes evaluations 61

Use or disclosure of data contained on this sheet is subject
to the restriction on the title page of this proposal.

List of Figures

Figure 1: Ivo Nelson at Congressional Hearing in March, 2006 15
Figure 2: IBM's Member-Centric Experience 16
Figure 3: Mayo Clinical Life Sciences System Architecture 21
Figure 4: Data Volume from Typical Epidemiological Longitudinal Investigation 25
Figure 5: Functional Diagram of Personal Health Care Information Exchange (see Note 1 below) 35
Figure 6: Personal Health Care Information Exchange Architecture 37
Figure 7: Conceptual View of Accumulating Data for Research 48
Figure 8: The Translational Health Roadmap Process 49
Figure 9: Flow of Information Between Clinical and Research Applications 52
Figure 10: The Larger Future Technology Picture Including Unstructured Data 54
Figure 11: Potential Sustainable Business Model Scenario 56
Figure 12: Healthcare Information Exchange Business Model 58
Figure 13: Financial Model Process 60

Introduction

IBM views personalized health care – or information-based medicine – as the convergence of life sciences, healthcare, and information technology that will revolutionize the discovery of new treatments and the practice of medicine. Information-based medicine will improve existing pharmaceutical and medical practices with knowledge generated from the integration of diverse clinical and biomedical data, and it will accelerate new research discoveries into clinical practice. We identified personalized medicine as a strategic business opportunity for IBM and our clients over 4 years ago. We formally designated information-based medicine as an Emerging Business Opportunity and made significant investment in people and programs to develop our capabilities in this arena. In 2007 the information-based medicine EBO has graduated to become a regular part of our healthcare and life sciences business.

IBM has served clients across the healthcare ecosystem in their health information technology needs, including commercial providers and payers, state and local public health agencies, health research organizations, and federal healthcare agencies. IBM provides significant leadership and participation in technology and health IT standards-setting bodies. We have made significant investments in our healthcare practice, from the acquisition of a leading healthcare provider consulting firm, to funding IBM clinical genomics research projects – such as our collaboration with the Mayo Clinic, to the establishment of a Center for Healthcare Management that produces thought leadership and fosters public-private collaboration.

IBM has also invested in healthcare software development. IBM has deep expertise in infrastructure technologies that are fundamental to realize the potential of personalized healthcare, such as relational database, data warehousing, data analysis/mining/modeling/visualization, and management of very large volumes of data. We have built upon our long-time experience within many industries and focused specifically on the unique aspects of healthcare and life sciences data over the past several years. IBM Research has developed algorithms for analysis and mining of data for decades, has focused on genomic data for over 10 years, and has focused on techniques for analyzing clinical with genomic data for several years. We have contributed software assets that comply with healthcare interoperability standards to the open source community, and we have led the development of the HL7 clinical genomics standards activity.

In our work with the Nationwide Health Information Network (NHIN), we have designed an architecture, developed a prototype instantiation of this architecture, and demonstrated this prototype at the recent American Health Information Community (AHIC) meeting and NHIN Forum. The NHIN architecture and prototype support the AHIC use cases and conform to the Health Information Technology Standards Panel (HITSP) standards. The IBM NHIN prototype effort used important interoperability standards for healthcare published by HITSP, key SOA interoperability principles and advanced data management algorithms developed by IBM scientists. Our Health Information Exchange, used to collect and share health data electronically, will help physicians access and view electronic medical records even if those records originate from disparate systems in multiple locations. Our use of the IHE Framework (Integrating the Healthcare Enterprise) played a major role in allowing participants to support this initiative.

We look forward to working with HHS and the healthcare community to exploit our capabilities and bring personalized health care to fruition.

Use or disclosure of data contained on this sheet is subject
to the restriction on the title page of this proposal.

1. Concepts on anticipated approaches for the use of EHR and population- and community-based health care system databases for longitudinal data collection in addressing:

- **Disease Susceptibility**
- **Clinical Course and Outcomes**
- **Treatment Response**
- **Evidenced-based Clinical Decision Support**
- **Optimal Healthcare Delivery Systems**

While EHRs are a first step in providing a comprehensive view of patient information and a set of information-based tools to support care delivery, an EHR is not requisite to the process of aggregating information for population-based analytics and community-based surveillance. Healthcare organizations collect large amounts of electronic data on a daily basis. Dispensed medications, resulted lab tests, dictated physician notes, procedures, diagnoses and often images are all available in various systems. An EHR consolidates access to this information. But this data may be re-purposed to other means in advance of an EHR implementation.

The large challenge in re-purposing this data is the low adoption rates of standard nomenclature for the various electronic data. While HL7 is widely used as a transaction standard, the content of the transactions is left to the individual healthcare organization to define. Even within a single healthcare entity, the same lab test could be defined differently if it is done in a different lab system. So the first hurdle for the use of electronic data for population-based analysis or surveillance is adoption of a set of standard vocabularies to create a common language for the data to be aggregated. For population-based functions to be meaningful there must be a commonality of information to support apples to apples comparison of health information.

With the implementation of EHRs and the adoption of content standards, the EHR can be effectively leveraged as a communication tool to healthcare practitioners, providing information on the most current information-based approach to a specific condition or providing individualized treatment options based on a person's genotype. A positive cycle of bedside to bench to bedside is well supported through the use of EHRs. Data collected through the delivery of care can be used by clinical researchers looking for more effective therapies and processes. That research could prompt some new path of basic research. New discoveries can, in turn, be embedded in the clinical decision support capabilities of the EHR to shorten the time from therapeutic discovery to care delivery adoption.

Based on our experience and observation, IBM anticipates some of the following capabilities:

1. Disease Susceptibility

Aggregated data may be used retrospectively to first understand the factors that determine disease susceptibility (i.e., what combination of phenotypic, genotypic and/or external factors might make a person or a population more or less likely to contract a particular disease) and then, prospectively, to identify the individuals or sub-set of a population that has the identified characteristics.

Once the characteristics that pre-dispose a person to a disease are known, these criteria can be embedded in the clinical decision support capabilities of EHRs to alert practitioners about the susceptibility of an individual patient. Then the practitioner can prospectively discuss options with the patient. Preventive strategies could then be decided on. These strategies could be

reinforced to the patient through the use of a Personalized Health Record (PHR) that provides appropriate alerts and reminders to support patient compliance.

In the event of a disease outbreak, individuals who are most sensitive to complications (e.g., the youngest, oldest and those with suppressed immune systems) could be alerted to take certain precautionary measures (e.g., obtain a vaccination). Further, based on a historical understanding of what types of conditions might indicate a particular disease is on the rise in a specific geography, algorithms can be defined that scan the de-identified data in the local health information network as an early detection mechanism. This information can then help local public health officials take appropriate actions to prevent the uncontrolled spread of the disease.

2. Clinical Course and Outcomes

Once data is aggregated within a healthcare organization, it can be used to:

- Identify the potential population of patients for a given clinical study to help researchers understand if a sufficient number may exist to support the study intention
- Identify potential candidates for an approved study, provided a blanket consent has been obtained from the patient indicating they are interested in study participation
- Identify patients on a specific therapy and track their outcomes over time to determine if unanticipated complications are present post-therapy induction

To assist in candidate identification for current clinical studies, the criteria for study inclusion can be embedded in the EHR clinical decision support capability, alerting practitioners about the availability of a clinical trial for their patient. For example, once a physician has diagnosed and staged a cancer, the EHR could provide a list of the available studies the patient might be eligible for, so that a determination can be quickly made and treatment promptly begun. This can improve the number of candidates included in a trial, the knowledge gained from the trial and potentially patient outcomes. Further, patients on clinical trials can be flagged in the EHR so that other practitioners are alerted to this status and avoid clinical interventions, if at all possible, that might void that patient's further participation in the trial.

Often clinical trials require that specific data elements be captured on each trial participant. Using an EHR, the capture of this data can be included in the normal clinical work flow, increasing both trial efficiency and reliability of the data. The requirements for how the data is captured can be embedded in the EHR and automatically monitored to ensure quality data.

3. Treatment Response

Often, critical data regarding the response of a patient to a particular treatment is locked in progress notes and other text-based documents in the EHR. For the purposes of analytics, this text data must be analyzed and discrete data elements extracted that can then be used in aggregate analysis. The creation of a health analytics environment that appropriately brings together in a common vocabulary discrete data elements (lab results, medications, data extracted from text) that describe a patient is critical to understanding what does and does not work, and why.

Using set theory inquiry techniques, a population of patients with the same condition or procedure can be identified. This cohort group can then be sub-set by difference in outcomes. Using analytic tools, clinical researchers can evaluate the differences between the various patient subsets that influenced the difference in outcomes. As the richness of the data sources grows – to include genomic, proteomic and metabolomic data about patients – so too can the

complexity of the analysis grow regarding the cause and effect of differences in treatment response.

Data mining techniques can also be applied to the aggregated patient data to look for relationships that are not intuitively obvious. While healthcare to date has not extensively used this capability, other industries (e.g., retail) have found great success in finding relationships between data that indicate predictable human behaviors. It is possible that the application of these same techniques in healthcare may uncover causal relationships in treatment response previously not imagined.

4. Evidence-based Clinical Decision Support

First a standard of practice must be proved through clinical research to use the data captured in the course of clinical care in a bedside to bench to bedside continuum. This standard of practice can then be embedded in the clinical decision support (CDS) capability of the EHR to support practitioners in the delivery of healthcare. CDS provides a number of avenues to provide information to physicians and other care providers to assist in treatment decisions. And, once a treatment decision has been made, all practitioners who see a patient may be reminded of the criteria to be followed. This can include medication dosing and timing; lab test intervals and results ranges; nutrition and exercise advice; and follow up visit schedules.

Practitioners record their actual treatment decisions and activities in the EHR. This data can then be aggregated to determine if the outcomes are as expected for those patients treated according to the evidence-based protocol. This can be contrasted with the data for patients treated in a way other than indicated by the protocol. This analysis can inform future refinements to the approach to care, supporting a focus on continuous outcomes improvement.

5. Optimal Healthcare Delivery Systems

Community aggregation of health information can support a number of optimal healthcare delivery analysis scenarios. First, differences in outcomes for the same procedure or disease can be evaluated between healthcare organizations and individual providers. This report card type of analysis can be done at a much more accurate and granular level than current efforts, which rely primarily on administrative data and therefore are sometimes clinically suspect. The inclusion of real clinical outcomes data (e.g., lab results, discrete data extracted from progress notes) improves the reliability of these comparisons and can aid the healthcare consumer in making decisions on where to seek care. This can also help healthcare organizations focus their efforts.

The ability to evaluate healthcare outcomes across a broad population may indicate the optimal type of healthcare provider and healthcare setting for a given type of patient or condition. This analysis could indicate that certain conditions are best treated in a home-centric manner, supported by mid-level practitioners connected to a patient via electronic tools. Or that a specialist should be the first point of contact for certain chronic conditions. Or that the first line of primary care should be a retail clinic. Or that a primary care physician should provide ongoing care for certain diseases once they are stabilized. The current assumptions of where care should be provided and who should provide the care could be adjusted once better aggregate clinical information is available. This understanding could lead to designing a more efficient and effective and less costly healthcare delivery system.

Use or disclosure of data contained on this sheet is subject
to the restriction on the title page of this proposal.

2. Anticipated Applications of Genomic-Based Clinical Testing in medical decision-making, safety assessment, and risk management

The success of the Human Genome Project has rapidly increased the opportunities of using genomic information in clinical applications. Genetic tests exist for hundreds of diseases. Although most are used for diagnosing rare genetic disorders today, this will change within the foreseeable future.

Anticipated Applications of Genomic-Based Clinical Testing

Healthcare practice is currently undergoing a transformation to increasingly use genetic testing for a much broader range of applications such as diagnosis, risk evaluation, drug response variation and prognosis. Recent developments in the molecular field have enabled large-scale genotyping of significant populations. Today this is primarily done in research environments, but it will become increasingly important for clinical care with translational medicine methods. The ability to use genetic markers (and other biomarkers such as proteomics, metabolomics, and imaging) to monitor risk, personalize treatment and predict outcomes will enable much more targeted healthcare delivery.

Targeted Drug Discovery

In 2002, IBM published "Pharma 2010: The Threshold of Innovation," our view of the future of the pharmaceutical and biotech industries. The executive summary of this report is included as Attachment 2. The complete report is available at:

http://www-03.ibm.com/industries/healthcare/doc/content/resource/insight/941673105.html?g_type=rhc.

This report predicted the decline of pharmaceutical company product portfolios, and the consequent growth of mergers and acquisitions, as a consequence of reduced productivity of the blockbuster model of drug discovery. We predicted that companies that focused on targeted therapies would be more successful in the future, and experience since then supports this prediction. In the targeted therapy model, genomic and other patient information may contribute to discovery of the therapy, to selection of clinical trial cohorts that are likely to respond to the new therapy with minimal toxicity, and to optimization of dosing. The best-known example of this approach is herceptin, which was fast-tracked for approval in breast cancer patients with overexpression of the Her-2 receptor, and is quite successful in helping the 25% of patients who have this biomarker. Many other examples are accumulating.

In 2004 IBM published "Pharma 2010: Silicon Reality," our view of computer-based technologies that could help implement our vision of the pharmaceutical industry in 2010. The report from this study is included as Attachment 3. The report is also available at:

http://www-1.ibm.com/businesscenter/smb/ca/en/contenttemplate/gcl_xmlid/27876?&ca=SMBCAENpharma2010&tactic=&me=W&met=inli&re=SMBCAENlifesciintropromo3.

Of the seven technologies identified as transforming drug discovery and development, four will impact genomic-based clinical testing:

- Petaflop and Grid Computing to solve large-scale biomolecular problems
- Predictive Biosimulation to simulate how a biological system works as a whole
- Advanced Storage Solutions to manage the huge amounts of clinical and genomic data

- Web-scale Mining and Advanced Text Analytics to extract important information from the vast stores available on the internet

These technologies are critical to make effective use of the large stores of genomic information in conjunction with clinical trials and patient practice.

Treatment Personalization Through Genomics

Multiple factors affect any single patient's response to a specific treatment, including drug interactions, environmental factors such as diet, allergies, and noncompliance. However, the genetic predisposition of the patient is a major source of variation that influences both the intended therapy effects and adverse ones.

Several genetic markers that predict drug metabolism have been identified, including genes that affect the response to substances such as Codeine, Warfarin and several leukemia and asthma drugs. Clinical assays focus on these genetic markers to distinguish between patients who will respond favorably to a treatment and those who do not respond well or who are at high risk for adverse effects. Using this information, treatment can be personalized, which will lead to higher efficacy (better drug response) and improved patient safety (fewer adverse side effects). Using genetic tests it is possible to target both of these factors and thereby stratify patients into subgroups with different treatment options.

With the increased use of pharmacogenomic methods in clinical applications, the number of treatment options and considerations will increase dramatically. This will increase the need for accurate information for clinical decision support.

For treatment personalization through genomics IBM identifies several challenges:

- Protecting the privacy of the patient, ensuring only authorized access to genetic information
- Enabling clinicians to connect the information collected through genomic assays with more traditional investigations (lab tests, physiological investigations, anamnesis)
- Development of clinically relevant assays so that a better decision can be made than would otherwise be possible. This involves the development of tests with high clinical interpretability and standardized validation
- Storing, and interpreting large amounts of genotypic data
- Managing the increased amount of information needed to make clinical decisions

IBM initiatives and activities related to these challenges:

Patient Privacy

- IBM has completed projects with healthcare clients to promote secure access to patient clinical and genomic data in compliance with HIPAA and other patient privacy regulations. For example, we have assisted the Mayo Clinic in entering and indexing over a million patient records, and associating that data with genomic information for those patients. We have addressed similar issues in work with iCapture, Lineagen/University of Utah, and Translational Genomics Institute among others.
- We have worked with the Karolinska Institute's Biobank Information Management System (BIMS) and with the UK Biobank to provide secure access to patient data for research purposes, and with National Geographic Society for one of the world's largest collections of genetic information (Genographic Project).

- Our work with the National Health Information Network (NHIN) and with several regional health information organizations (RHIO) has enhanced our expertise in addressing the issues associated with securing such sensitive data.
- Using our federated search software products, various distributed databases can be searched as if the data were in a single combined (virtual) database, with security managed across all of the data sources. This assures consistency in access rules and proper audit trails for access to all the data.

Integration of Clinical and Genomic Data

- IBM has implemented data warehouses and federated data collections combining clinical and genomic data for several clients.
- Our efforts to establish and implement a consistent Health Information Exchange (HIE) model is discussed several articles in the Jan. 2007 issue of IBM Systems Journal (see Attachment 1).
- We have implemented a consistent, cross-clinical study data repository (JANUS) for FDA and NCI clinical studies.
- IBM completed a project with a large pharmaceutical company to build a decision support system for cohort selection. This project included a dimensional model based datamart that integrated clinical endpoints (demographics, efficacy, safety and lab results) with genotypic data and structured clinical trial protocol information to support clinical trials.
- For another large pharmaceutical company, IBM built a Service Oriented Architecture (SOA) based data integration platform. This project used data federation to dynamically link genomic and proteomic information from external sources such as GenBank and dbSNP with internal clinical and Single Nucleotide Polymorphism (SNP) databases.
- IBM is working with a large pharmaceutical client to build a regulated collaborative environment that includes images. Images are captured at the trial site, and they are analyzed and annotated by the Contract Research Organizations. The images are further validated by independent reviewers before being delivered to the sponsoring pharmaceutical company for integration into their drug application submission process to FDA.

Development of Better Clinical Relevant Assays

- IBM Research has developed clinical decision support systems based on mining large collections of clinical data. Several of these projects are summarized in the current IBM Systems Journal (see Attachment 1). Novel aspects of this work include mining information sources for new associations and developing clinically relevant rules from these insights. While this work is now primarily at the research stage, the rapid growth of large, consistent, multi-site data stores will soon allow rapid generation of testable hypotheses on drug-drug interactions, drug-environmental associations, and optimal treatments for defined patient cohorts.

Increased Drug Safety Through Genomics

It is not uncommon for new drugs to fail in late phases of clinical trials due to adverse effects or unpredictable behavior. A very recent example is Pfizer's cancellation of clinical trials for torcetrapid, which was expected to be a potential blockbuster product for treatment of heart disease that would increase HDL. During a Phase 3 clinical trial, more people than expected died while taking torcetrapid. By utilizing pharmacogenomic, proteomic, metabolomic,

toxicogenomic, and imaging biomarkers it may be possible to identify adverse effects earlier in development, thereby reducing development costs. Targeting positive responders could rescue drugs that otherwise would fail. This approach offers many potential advantages, including faster approvals with earlier market introductions, increased patient compliance, expansion of treatment to new population groups and prophylactic medication.

Genomics can also improve drug safety when used in Phase 4 clinical trials and in other research on safety and effectiveness. Phase 4 trials usually have a more diverse patient population than Phase 3 trials, as historically the population of cohorts in Phase 3 trials has not been very diverse.

For increased drug safety through genomics, we identify the following challenges:

- Storage of the huge amounts of genomic (and other –omics biomarker) data. The data volumes used in genomics research are normally much larger than those used in clinical practice.
- Development of methodologies for searching the large data volumes and connecting this information to other types of data (e.g., phenotypes, environmental information).
- Gaining regulatory acceptance of data from new genomics-based methodologies

Examples of IBM initiatives and activities related to these challenges:

Storing, Refining, and Interpreting Large Genomic Data Stores

- Our work with the Karolinska Institutet BIMS, the Genographic Project, and the National Digital Medical Archive (for medical imaging and other data) has tested the handling of very large medical records. We have determined that computer grids can handle all anticipated sizes.
- Work at IBM Research has advanced analysis of sparse matrices found in data collections such as the Genographic project.

Data Models and Clinical Genomics Standards

- IBM collaborated with the FDA to create the data model JANUS for storing clinical submissions from sponsors for review and approval of new drug applications. JANUS supports an open and standards CDISC (Clinical Data Interchange Standards Consortium) based design approach for a cross-trial clinical submission data repository that includes genomic data. This approach significantly improves the review process by allowing FDA reviewers to analyze the data more comprehensively and faster. The reviewers can compare data across drug applications for the same therapeutic area of drug family, and they can retrieve data using a unified access layer through a variety of analytical applications. We published a white paper presenting the roadmap and vision of JANUS adoption in the clinical development industry (pharmaceutical companies, FDA, NCI, Contract Research Organizations, trial sites).
- We are working with a number of standards organizations such as HL7, CDISC and others, to enable data collection across clinical studies and across various data sources.

Risk Evaluation and Prophylactic Treatment

New genomic technologies in conjunction with other biomarkers have the potential to significantly improve the accuracy of medical risk assessments with subsequent patient stratification into susceptible subpopulations. The information generated from such genetic investigations will be used more frequently for creating personalized risk profiles. By using such profiles it is possible to initiate prophylactic treatment, stimulate patient life style changes and improve screening efforts. In addition, genomic investigations will increasingly be used to assess risk predisposition for common diseases, not just the rare ones as is most common today.

Although genetic profiling for risk assessment will lead to many opportunities for better public health, it also implies some noteworthy challenges:

- Development of clear and precise ethical guidelines for utilization of genetic information for predictive purposes
- Importance of genetic counseling for decision making by the patient (or healthy individual)
- Ensuring regulatory guidelines and technical methodologies for providing protection against genetic discrimination in the context of employment, insurance and financial transactions
- Development of common best practices for the collection of informed consent.

Use or disclosure of data contained on this sheet is subject
to the restriction on the title page of this proposal.

3. Establishment of biospecimen resources obtained from clinical medical services for application in research, clinical trials, health services planning, clinical effectiveness, and health outcomes evaluations

The transformation of healthcare towards personalized medicine will require proactive leadership and collaboration by an extended ecosystem of stakeholders. Biospecimen resources are a necessary prerequisite for this transformation. Leaders from the bio-pharmaceutical and diagnostics industry, government, and academia recognize that researchers and clinicians need access to molecular information and the associated phenotypic information, often contained in medical records or collected in the course of clinical trials.

Biobanks — also referred to as biorepositories or tissue banks — provide both types of critical information, and thus can serve as an important translational bridge between research and clinical practice. They can accelerate the discovery and development of medical treatments that are targeted to subpopulations characterized by shared genetic and other characteristics to be taken into account when diagnosing, preventing or treating diseases.

Due to the compelling potential benefits of Biobanks as resources for population genetics and biomedical research, there has been a global movement towards establishing them. The Framework Program 7 of the European Union has singled out this area as one of the most important investment areas for life sciences research. Similarly, governments in Taiwan, Singapore, and Japan have started Biobank projects.

As stated by Dr. Anna Barker of the National Cancer Institute (NCI) at IBM's second Biobank Summit II, held in November 2004 at Tarrytown, NY, "Unless substantial action is taken with biobanking and biospecimens, we will delay personalized medicine by 30 to 40 years."

Biospecimen resources and their associated clinical data will be a key enabler of improved genetic analysis and screening. Improved genetic screening will, in turn, result in improved effectiveness and efficiency of clinical trials and should allow for shorter directed trials. For a specific disease, personalized medicine may result in the development of multiple personalized therapeutic options. These cumulative therapeutic approaches, combined with supporting genetic testing, may begin to replace the traditional pharmaceutical industry blockbuster model, assuming combined markets begin to approach traditional market sizes.

Movement towards this information-based medicine will allow for the potential to recover failed drug candidates through elimination of at-risk patients and improved identification of target patients. Directed clinical trials to validate efficacy and safety issues should allow for potential recapture of previously written-off investments and will provide for new revenue opportunities. Biospecimens and associated data collected from clinical trial participants will be critical to identifying the key efficacy and safety factors that can drive product recovery.

As the knowledge built on biobank specimens grows, recognition of the value of biobanks is becoming increasingly apparent. Researchers look to biobanks to advance their scientific efforts. Pharmaceutical companies are finding biorepositories as tremendous resources in advancing drug discovery and development. Government and other funding organizations look to advance long-term healthcare goals. The general public is beginning to see the potential for disease solutions in their lifetimes. As biobank-enabled medical research yields knowledge, forward momentum is accelerating

Advances in technology and efficient application of technology to both normal and disease biospecimens are beginning to uncover the genetic basis for the differential manifestation of

specific diseases. In addition, population genetic studies are adding insights about environmental factors that complement genetic characteristics. Significant efforts are underway in academic research centers and in bio-pharmaceutical research labs to identify biomarkers that can be used as indicators for these differential effects. For some cancer patients, personalized medicine is already a reality. Novartis' leukemia drug, Gleevec, and Genentech's breast cancer drug, Herceptin, are well-known current examples. Continued development of such targeted, effective products will require the availability of significant quantities of high-quality biospecimens from biobanks.

Biospecimens Must be Related to Patients

Due to recent scientific and technological breakthroughs, molecular medicine has emerged as an important new way to deepen our understanding of biochemical pathways and disease mechanisms. These new technologies include:

- High-throughput genotyping, to identify genetic and gene-environment causes of disease
- High-throughput proteomics, allowing the development of early-detection markers
- New tools for the measurement of exposures, improving precision
- A molecular disease taxonomy

These technologies will allow a much better understanding of both the causes and intermediate stages of disease. However, new methods do not obviate the need for good epidemiologic study design. The greatest opportunities for informed medical practice come from the application of new methods to large-scale human cohort studies.

The essential ingredients of the standard epidemiologic cohort study are people and time. Individuals (e.g., from the general population, from a specifically exposed group such as smokers or asbestos workers, cancer survivors, or those at high risk of a specific disease) are recruited. Specific data on behavior, exposures, family history, medical history, etc. are collected from participants. Blood is collected for a variety of biologic measures. And the participants are followed over time to disease endpoints of interest or death. Inferences are made about possible causal relationships by comparing the rates of disease and mortality in the exposed, or the genetically predisposed, with the rates in the unexposed, or non-predisposed. New tools will allow exploration of a wider range of biology in this cohort-study setting and the cohort design itself can be used to expand the range of questions that can be asked. In short, what is essential is to recruit a living population laboratory, in addition to assembling collections of well characterized, but anonymous, tissues in general and specific biobanks,.

It is then the role of IT to acquire, integrate, manage, and analyze the data associated with both blood and tissue samples and with the patients who have provided those samples.

Biobank information and management systems must be designed to integrate biospecimen data with phenotypic patient data, and these systems must incorporate data security and privacy functions to prevent unauthorized access.

Because biospecimens are fundamentally the property of the individual from whom they were obtained, the ability to study and analyze individual samples is limited by the informed consent conditions approved by the donor. Recognizing the rights of the donor as well as the research opportunities (those currently recognized and those yet to be determined), the establishment of standardized informed consent options will be important to ensure that the new medical knowledge can be derived from biobanks.

Biospecimen Resources and Associated Quality Issues

The preservation of biospecimens is not as simple as aliquoting materials into cryotubes and placing them in a -80°C freezer. While DNA would be effectively preserved under these and less-severe conditions, proteins, mRNA, and small molecule metabolites may continue to degrade over time or may be directly affected by the freezing conditions. Because biological samples are complex and potentially self-degrading, it is important to anticipate the types of testing that are to be performed on biospecimen samples.

Basic research has well characterized the appropriate storage conditions required for most, if not all, of the different types of biological materials that can be present in biospecimens. In cases, where specific research requirements are known, it is easy to identify proper collection and storage conditions for specific biospecimens. Alternatively, where biospecimens are broadly collected with the potential for future testing of unidentified biological analytes, multiple storage conditions may be required. Based on the major categories of biological analytes of interest (i.e., DNA, RNA, proteins, lipids, small molecule metabolites), identification of standard storage options should be possible, allowing for maximum flexibility for current and future testing.

In addition to standardization of storage conditions, standardization of extraction and testing procedures will also be needed, based on quality assurance, and data documentation requirements. Identified areas requiring specific standardized processes and procedures include genetics (especially DNA), gene expression, metabolomics, lipidomics, and proteomics.

The ability to differentiate specific variations caused by specific disease conditions requires comparative analysis with appropriate normal reference samples to ensure that the variations are really linked to the disease. As part of their general approach, biobank repositories are a significant resource in providing standardized normal reference samples. With standardized collection procedures ensuring comparability between different samples and appropriate relevance to disease tissues, biobank repositories can ensure that comparative analysis is possible. For example, in a simple comparison of cancerous and non-cancerous tissues, other potential biases (e.g., age, sex, fasting versus non-fasting) may be just as important in analytical testing, resulting in potentially incorrect conclusions.

International and National Standardization Efforts

The P3G (Public Population Project in Genomics – http://www.p3gconsortium.org) is an international consortium for the development and management of a multidisciplinary infrastructure for comparing and merging results from population genomic studies. P3G has as its primary focus the harmonization of biobanks and population studies. In collaboration with international working groups, P3G is working to establish commonality among existing research databases, including identification of comparable variables and establishing the basis for inter-biobank correspondence criteria. P3G has achieved sufficient critical mass to form the principal international body for the harmonization of public population projects in genomics.

In the US, the National Cancer Institute has an integrated effort across all of its divisions to identify and implement quality practices in NCI-funded projects through the Office of Biorepositories and Biospecimen Research. This effort began with collaborative work on the National Biospecimen Network (NBN) Blueprint, a vision for a nationwide biorepository management system to support genomic and proteomic research. The requirements outlined for this system emphasized:

- Use of best practices-based standard operating procedures for the collection, processing, storage, annotation, and distribution of biospecimens
- Harmonization of informed consent procedures
- Development of a common informatics platform

The NBN Blueprint also detailed the importance of upholding the highest ethical and legal standards by establishing a chain of trust beginning with the patient who donates a biospecimen and extending throughout the network to each scientist who uses the biospecimen for research purposes.

The primary goal of the International Society for Biological and Environmental Repositories is to provide information and guidance on the safe and effective management of specimen collections. The society serves to pool and share collective knowledge and experiences with colleagues. This knowledge helps member repositories operate more efficiently in managing and preserving their collections. Careful management ensures the collections are available for study as new biomarkers emerge and more sensitive measurement technologies become pertinent. Now, on the threshold of exponential new developments in molecular biology, nanotechnology, environmental toxicology, and computerized data management, all organizations need an active interchange of ideas and expertise. To that end, subcomponents of government, academia, the private sector, and manufacturers as well as interested individuals are invited to become active participants in ISBER and to exchange knowledge on repository issues.

Translational Medicine

The practice of medicine is in the early stages of transformation from an episodic focus on identifying symptoms of disease towards a more longitudinal understanding of susceptibility, disease etiology and individualized preventive therapies. Longitudinal genetic and treatment information about individuals generates medical insights. These insights can be acted upon to design more targeted therapeutics, more informative diagnostics, and more specific treatments to cure and to prevent disease. The fast growth of diverse genotypic and phenotypic patient data is expected to continue, based on new scientific advances and the growing recognition of the need to invest in nationally and internationally interoperable health care IT infrastructures. Information-based medicine can leverage this growing wealth of available information to support pharmacogenomics and the delivery of new, targeted, personalized treatments and therapies that will ultimately increase the quality and reduce the cost of health care.

The ultimate goal is improved patient care and outcomes, hence the emphasis on translation of research into clinical practice. Translational medicine must proceed in steps way and be encouraged and supported by all healthcare stakeholders in order to succeed. The steps are:

1. Biobanks to biomarkers: conduct first class biomedical and translational research
2. Translation of discoveries from the laboratory to the clinic and the community
3. Creation of multi-disciplinary research teams, including bio-pharmaceutical and diagnostic industry
4. Education of health care practitioners and patients, creation of public awareness
5. Secure financial stability and funding by proving Return on Investment and developing new business models for payers, providers, and industry

4. Organizational or institutional practices to address ethical, legal, and social implications regarding the use of patient information, including genetic data, to support personalized health care

Why Practices and Policies are Necessary

Healthcare is one of the most personal services rendered in our society, yet to support the delivery of this care, many people – both clinical and administrative – across diverse organizations must have access to intimate patient information. Access to patient information is now granted to personnel other than healthcare providers and payers. In order to receive appropriate care, patients must feel free to reveal that personal information. In return, those with access must treat this patient information confidentially.

However, maintaining confidentiality and trust is becoming more difficult. Information systems technology has contributed to this change, allowing instant retrieval of medical information and wider access to a greater number of people. Within healthcare organizations, personal information contained in medical records is reviewed not only by physicians and nurses, but also by professionals in many clinical and administrative support areas.

While the speed of IT advancements has enabled the sharing of patient data, developing standards to ensure privacy has advanced more slowly. Healthcare professionals have a moral and professional obligation to protect the privacy of patients' medical records, and in 1996, a legal obligation was established – the Health Information Portability and Accountability Act of 1996, commonly referred to as HIPAA. Under HIPAA, healthcare institutions must meet strict privacy requirements for maintaining medical records, whether electronic or paper-based, and for exchanging the clinical data they contain between healthcare providers. (The exception is when records are required for treatment purposes.) (1) *Catherine Arnott Smith, Ph.D.; Assistant Professor; School of Information Studies; Syracuse University; Syracuse, NY.*

While the healthcare organization may own the health record, the information in that record remains the patient's personal property. Healthcare organizations must determine the appropriateness of all requests for this information under applicable federal and state law and act accordingly. Employers are acting as patient advocates and must obtain proper employee/patient authorization to release information or follow carefully defined policies on the release of information without consent. Organizations and institutions must develop and enforce policies and procedures to protect this information while realizing the benefit from having access to this information. These policies must comply with more than HIPAA and other legal regulations – there are ethical and social implications to using patient information.

IBM Corporation Practices

On March 15 and 16, 2006, IBM testified before the Health Subcommittee of the Energy and Commerce Committee, which has jurisdiction over the Medicaid programs, FDA, CDC, and NIH. IBM executives:

- Told congressional subcommittees that it is time for a new healthcare paradigm
- Described creating a patient-centric healthcare system built on a standardized IT infrastructure

The testimony of Ivo Nelson, IBM health industry leader, and Dr. Jane F. Barlow, IBM wellness director, at separate subcommittee hearings underscored IBM's commitment to act as a strong voice in the campaign to empower employees in the private and public sectors with electronic personal health records. (see Attachment 4 for the complete transcript of the March 15 Barlow testimony).

Figure 1: Ivo Nelson at Congressional Hearing in March, 2006

IBM has more than 329,000 employees in over160 countries. As a consumer of healthcare services for our employees, and as a supplier of products and services for the entire healthcare ecosystem, IBM has developed and promoted the use online health information systems. IBM has made available to its employees on-line health and wellness tools to improve their health and wellness.

For the 180,000 U.S.-based employees, IBM offers a first-of-its-kind Web-based Personal Health Record (PHR) system, which collects and analyzes employee health information. The PHR must be protected by those who access, maintain and analyze this information. IBM allows employees access to information in a manner that enables them to better manage health and wellness of themselves and their families.

IBM believes we have provided a robust PHR system that enables IBM health plans to provide personalized health care for our employees. The PHR system is part of what IBM considers a Member-Centric Experience, developed by leveraging our technology, IT expertise and our Global Business Services consultants. A Member-Centric experience enables our employees and healthcare providers to:

- Drive increased employee accountability
- Deliver better quality of care to result in lower healthcare costs and more productive, innovative employees

The PHR system is also the basis for analytics to identify gaps in care and to optimize benefits:

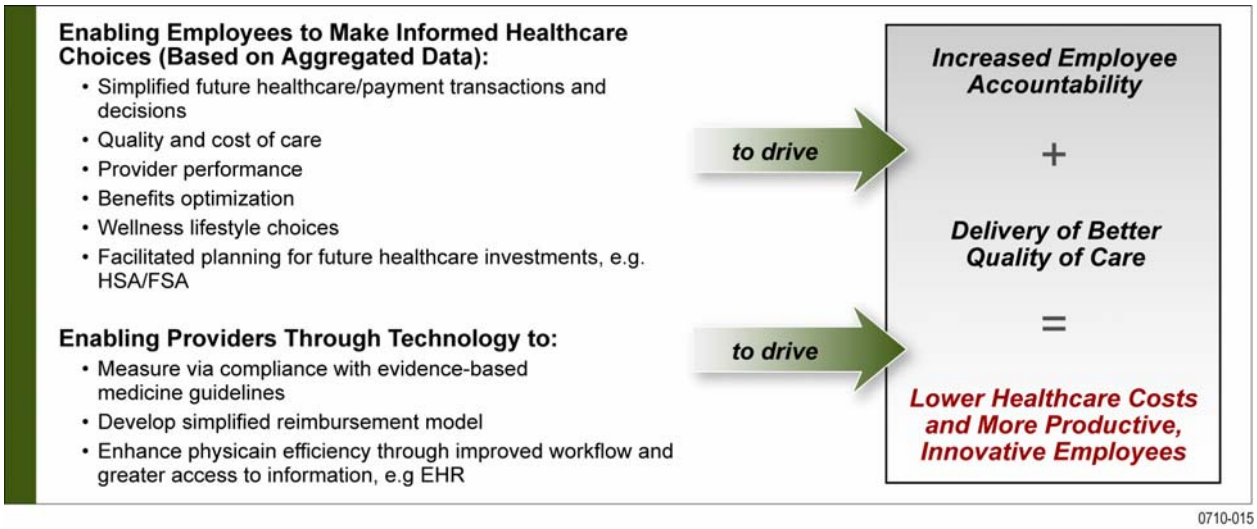

Figure 2: IBM's Member-Centric Experience

(2) Elizabeth J. Dietz; Partner-Healthcare; Vin Tavormina; Partner-Public Sector Healthcare; IBM Global Business Services; March 2006.

To support our Member-Centric Experience, IBM must gather employee/member information. The member must trust IBM in gathering, storing, accessing and sharing this data. Members rely on their employer and healthcare providers to keep their personal information confidential. But trust can be particularly difficult to achieve in the anonymous, virtual environment of the Internet. If members do not trust the organization, they may withhold or misrepresent personal information.

To develop trust and demonstrate that our privacy policies are enforced, IBM adheres to these practices and policies regarding the use of electronic health information:

1. Consideration of these e-health Ethics Initiatives
   - Health On the Net (HON), Code of Conduct (www.hon.ch/HONcode/Conduct.html)
   - American Medical Association, Guidelines for Medical and Health Information Sites on the Internet
   - Health Internet Ethics (Hi-Ethics), Ethical Principles for Offering Internet Health Services to Consumers. (www.hiethics.org/Principles/index.asp)
   - Internet Healthcare Coalition, e-Health Code of Ethics (www.ihealthcoalition.org/ethics/ethics.html)
   - URAC "Health Web Site Standards" (www.urac.org/documents/Health WebSitev1-0Standards040122.pdf).
2. HIPAA Guidelines regarding data that would appear in the intranet, including:
   - Permitted uses and disclosures of personal health information
   - How to create de-identified health information that can be disclosed without consent
   - Standards for Electronic Transactions and Code Sets, including:
     - Standards for Health Care Claims
     - Other Covered Transactions Privacy

- Privacy of Individually Identifiable Health Information
- Privacy Compliance Procedures
- Patient Privacy Rights Security
- Physical Safeguards and Technical Security Mechanisms
- An implemented HIPAA Compliance Program

3. IBM will collect Personal Information on an individual only if the individual provides the information into the IBM-sanctioned intranet tool. Unless an individual gives permission by informed consent, the intranet host will not allow any third party to use personal information collected. Other constraints include these member options:

   - To construct a member's Health Management Center Home Page, the intranet host will use concept unique identifiers (CUI). The host tailors the information on the Health Management Center Home Page to reflect individual interests, concerns and personal health characteristics. A CUI is attached to every piece of information a member provides.
     - For example, a member completes the Health Risk Assessment and states he has diabetes. Each piece of information is tagged with a CUI that is specific to diabetes. Every user that indicates he or she has diabetes receives this CUI tag.
     - Each time a member views his Health Management Center Home Page, the CUI tag is matched to content related to diabetes. If the intranet's algorithms determine that this is likely to be an important topic to the member, it will appear on his personalized page.
   - Opt-In: Members indicate their preference to participate in a program, receive emails, feature, tool, or enhancement on a website. Typically, Opt-in members must provide certain information to the website or otherwise actively indicate their choice or preference to participate in the website program.
   - Opt-Out: Members must take some action to indicate their preference to stop participating in a program, email, feature, tool or enhancement on a website.
   - IBM designates certain information as Protected Health Information (PHI). This is information that can be traced back to an individual that is in electronic, paper or oral form that relates to, among other things, past, present or future physical or mental health or condition. PHI information is protected by:
     - Limiting access to your PHI to you, the Plan, and any third party vendor that provides services on behalf of the Plan
     - Using firewalls to protect PHI held on IBM servers
     - Maintaining a record of the disclosures of member PHI
     - Protecting the physical security of intranet host servers. Physical security is maintained through pass code locked door access and pass code authority.
     - The intranet host server protects member access to their information by providing geographic redundancy. The host servers store identical information at two separate locations.

4. IBM partners with a third-party advisory or watchdog organization to act as a liaison with the web host to resolve employee security concerns (http://www.truste.org/watchdog.html).

   (3) IBM-sponsored Health Management Center website; Privacy and Confidentiality Statements; December 2006

5. Practice the process of de-identifying member records. This process involves stripping numerous types of specific information about a patient, as well as any relatives or other individuals whose medical data may be included, from records before access is permitted for research or other non-treatment-related purposes. Patient or member name is not included with this information.
6. Release of information cannot be made without proper authorization except under limited circumstances. Patients or their legal representatives must consent to the release of information, and IBM maintains records of most disclosures for review upon patient request. IBM will, however, disclose all information as required by law. Other exceptions to patient confidentiality are to promote public health, to protect children and spouses from abuse and to comply with certain laws.
7. Upon formal request from the member, IBM will no longer display a member's PHI via the internet. This information may, however, be retained in back-up media. This PHI may only be disclosed as required by law and when in accordance with the notice of privacy practices furnished to members by their health plan.

Genetic Non-Discrimination

On January 30, 2007, IBM's Chief Privacy Officer, Harriet Pearson, testified before the House Education and Labor Subcommittee on Health, Employment, Labor and Pensions about the subject of genetic non-discrimination. IBM called on Congress to enact laws that prevent discrimination based on genetic information, which is increasingly prevalent in the diagnosis and treatment of many medical conditions, as well as in research to discover the fundamental genetic mechanisms of major diseases.

Pearson, architect of IBM's genetic non-discrimination policy, testified, "The reasons for making genetic privacy part of our broader discrimination protections are clear to us: first, a person's genetic profile or makeup should be treated the same as other innate human characteristics, including one's race, gender, national origin, sexual orientation, age or physical abilities. Simply stated, a person's genetic profile is as natural and as inseparable from who they are as any other physical trait or attribute."

In October 2005, IBM became the first major corporation in the world to establish a genetics privacy policy that prohibits current or prospective employees' genetic information from being used in any employment decisions. IBM has taken the position that genetic information will not be used in hiring; in determining employees' eligibility for healthcare coverage or other benefits; or in other employment decisions to which such information is not relevant. IBM believes genetic data should be secured not only because the privacy of such information is vital to innovation in healthcare and life sciences, but also because protecting such personal data is a natural extension of our nondiscrimination practices.

IBM believes that genetic data or test results are a valuable tool used in health diagnosis and treatments. One practical reason genetic information should be controlled is that in many instances test results only suggest a risk of developing a disease. It is our view that no employee should lose their health insurance or their livelihood because they have a statistical chance of developing a medical disorder.

Instead, such information can be – and should be – used to positive ends: enabling preventative lifestyle changes or tailoring medical regimes to reduce a patient's chance of developing any disease their genes might incline them toward. With genetic tests available for

almost 1,000 diseases today and hundreds more under development, personal genetic data is becoming more commonplace. The danger of not safeguarding information raises the potential for a person with a genetic predisposition toward one or more diseases to be denied healthcare insurance, lose their job or be denied employment altogether.

References

1. ""Electronic Health Records: Sharing Knowledge while Preserving Privacy," IBM White Paper; Catherine Arnott Smith, Ph.D.; Assistant Professor; School of Information Studies; Syracuse University; Syracuse, NY.
2. Elizabeth J. Dietz; Partner-Healthcare; Vin Tavormina; Partner-Public Sector Healthcare; IBM Global Business Services; March 2006.
3. IBM-sponsored Health Management Center website; Privacy and Confidentiality Statements; December 2006.

5. Examples of utilizing large clinical data repositories for practical clinical research to discover effective technologies, therapeutics, diagnostics, and prevention strategies for different populations

Mayo Clinic

The twentieth century saw great advances in healthcare that were made possible by physicians and scientists who aggregated, analyzed, summarized and implemented all data, information and knowledge components required for individual healthcare. The diversity and large size of data gathered within research and now available in daily practice exceeds the capacity of even the most brilliant scientists. Preprocessing of data is required before presenting it in summarized fashion for evaluation by the healthcare provider of the 21st century. Future healthcare will depend on computers to sort, analyze, rank and interpret an ever growing body of data to implement truly individualized, outcomes-based healthcare. Mayo Clinic and IBM agree that realization of the promise of data-driven, outcomes-based healthcare requires a combination of computer and biomedical expertise. A natural collaboration was formed in 2001 to address these issues.

The Mayo Clinic/IBM collaboration has successfully completed numerous phases. The initial focus of the collaboration was to create an integrated repository of patient data for analytical purposes. Unlike clinical operational systems, where database design and access is optimized for fast, complete access to relevant data for a single patient, the data repository is optimized for cross-patient data analysis. Therefore, the repository is built on a model of:

- Extracting patient data from a diverse set of operational systems, including the electronic medical record systems as well as research derived data such as microarray data
- Transforming the data into a consistent format and vocabulary
- Loading the transformed data into an integrated data repository, and
- Providing access to the repository for users to gain new insights for research and ultimately to improve patient care

During Phase I of the collaboration, the initial repository – the Mayo Clinic Life Sciences System (MCLSS) – was created. At the completion of Phase I, it contained more than four million Mayo Clinic clinical records covering patient demographics, diagnostics, and laboratory results. The repository was built using the standard IBM DB2 software product, and was populated using a combination of database replication and extract/transform/load (ETL) procedures. The design for the Phase I architecture was to move the greatest amount of data into the repository given the constraints on time and resources in place at that time. To enable researchers to access the data and meet the security and privacy requirements for access to the data, an innovative web-based query construction tool was developed – Data Discovery and Query Builder (DDQB). DDQB provides a secure web-based query construction and management framework from which complex queries can be created without requiring that the non-technical user deal with the underlying complexities of the physical database.

Phase II added microarray data and unstructured text. Strong adherence to internationally accepted standards was a focus of this phase, as a standards-based architecture is both a method of integrating data and a means for Mayo Clinic to participate in multi-center healthcare or research initiatives in the future. The microarray project included data storage, retrieval and analysis of microarray data. It incorporated the MAGE-ML standard to move data into the warehouse. The unstructured text project included storage, analysis, indexing, and retrieval using state-of-the-art computer science, natural language processing, and language

understanding principles that are required to annotate Mayo Clinic's 15 million electronic clinical notes. The text project consisted of one-million documents in the HL7 CDA (Clinical Document Architecture) format, a messaging infrastructure, IBM's Unified Information Messaging Architecture (UIMA) infrastructure and Mayo Clinic and IBM annotators for text analysis. DDQB provided a secure web-based query construction and management framework. The Phase II architecture is shown in Figure 3 below.

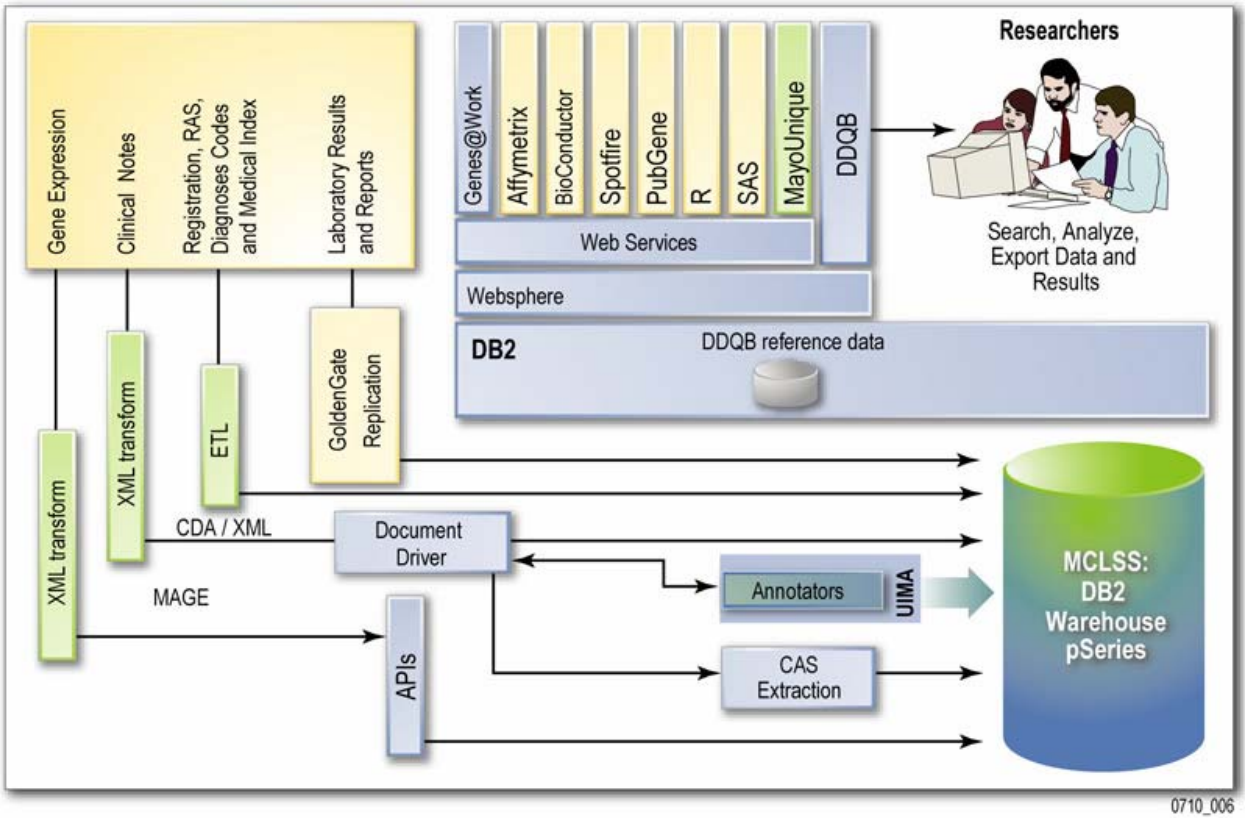

Figure 3: Mayo Clinical Life Sciences System Architecture

The Mayo Clinic Life Sciences System is available to assist investigators with their medical research and clinical studies. Medical research strives to create and contribute to the understanding of medical knowledge to ultimately improve the care and treatment of patients and reduce the burden of human illness. The clinical research process has often been managed by the individual research investigator, but it can be difficult to scale when performed in this manner. A prototypical set of processes for clinical research involves research protocol management/approval, research protocol reporting, research study design, patient cohort identification/data retrieval, patient enrollment, cohort abstraction/study management, study analysis, study results, and study publication(s). A key aspect of the clinical research process is data collection to support the research goals. With electronic access to patient data becoming more available, the collecting of primary data enables a number of opportunities to assist the investigator in the research process. Patient cohort identification can be used for many study purposes: case sampling, incidence or prevalence studies, clinical trials, or human studies. With MCLSS, one set of Mayo researchers were able to find patient cohorts with a specific disease, diagnosis and lab tests living in a particular zip code in minutes, versus months to find those patients without MCLSS.

In addition to the data itself, data about the data, or metadata, within the Mayo data repository can provide critical information to inform investigators browsing the repository and facilitate query formation. Extended metadata can also contain access information and methods for data elements in a federated architecture when these data elements are not stored within the main database. Traditional use of the term metadata implies information about the status and provenance (source and ownership history) of data elements or values – data demographics.

With the generation of a large data repository comes the responsibility of ensuring the validity of this data. Data stewardship is the practice of ensuring that the data and metadata meet quality, accuracy, format, and value criteria, and that it is properly defined and understood (standardized) across the enterprise. Good data stewardship is essential for successful deployment of a medical data repository.

Effective use of this aggregated data requires an infrastructure (software, hardware, data governance, data stewardship) to support data intensive analytic activities. This involves data/process modeling, security, privacy, auditing, data standardization, data movement and transformation. The warehouse architecture needs to support hundreds or thousands of users, be able to store hundreds of terabytes (eventually petabytes) of data and be able to grow and expand with minimal difficulty. Additional storage and processing challenges will be faced as more and more genomic and proteomic data becomes available.

Soon, advanced research, clinical and business analytic methods will be used that incorporate semantic retrieval methods based on ontology-driven query refinement. This includes analysis of structured data, semantic analysis of unstructured or semi-structured data such as clinical notes and image data, and discovery of associations through human discovery or semi-automatic (data mining) techniques.

Successful implementation of a data warehouse for research purposes quickly leads to the desire to expand its use for the entire enterprise. Healthcare organizations like Mayo are under increasing cost, margin, quality, and patient safety pressures. This requires the ability to analyze all aspects of the business of healthcare and drives the need for an enterprise-wide data warehouse.

As a result, IBM and Mayo Clinic are working collaboratively on a Health Information Warehouse model that can be used across the industry to collect and organize information from across a healthcare enterprise. This includes clinical, genomic, financial, educational, administrative and operational data. The goal is to convert this data aggregated from multiple sources, both structured and unstructured, into knowledge, which will improve individualized prognosis, diagnosis and treatment of patients.

Other Examples

Additional examples of the use of large clinical repositories for practical clinical research include projects with iCAPTURE at the University of British Columbia and with the University of Virginia.

The world-renown researchers of the iCAPTURE Centre, a partnership of St. Paul's Hospital, the Vancouver Hospital and Health Sciences Centre and the University of British Columbia collaborated with IBM scientists. In the first phase of this project, the IBM team helped build iQEngine to uniquely assist researchers relate genetic susceptibility of patients with cardiovascular and respiratory diseases to environmental influences such as culture and socioeconomic status, educational backgrounds, inhaled cigarette smoke, pollutants, viruses, allergens, diet, and obesity.

The prototype system provided new insights that may, ultimately, help researchers as well as physicians make patient management decisions. Phase II broadened the patient population and data sets to include genetic sequences, such as single nucleotide polymorphisms and haplotypes, related to atherosclerosis, aortic valvular stenosis, systemic inflammatory response syndrome, asthma, and other diseases.

IBM has carried out high dimensional (with many parameters) data mining of patient records with the University of Virginia. The pioneering study with Virginia addressed 667,000 patient records, by far the largest archive of very comprehensive records constructed in good format at that time. Many thousands of rules for guiding physicians and researchers were generated and tested automatically as queries against the medical literature. 89% of the rules were statistically quantified rules known to medical decision making, 8% were new to human medicine but consistent with findings in the basic biological science literature, and 3% were new to basic science. (Refer to Attachment 6 and our response to topic 14 on page 51 for additional information on this study.)

6. Issues and challenges associated with incorporating genomic information as a part of a broad longitudinal data collection

11. Examples of the use of disease registries to track specific diseases and response to drug therapies across different subpopulations

Longitudinal data collection for epidemiological investigations or clinical registers has traditionally been focused on accumulating phenotypic and environmental information. This data is typically collected through various methods, including interviews, paper questionnaires, clinical health records and various forms of electronic data capture (EDC). Although this data has been collected for a long time, there are still major challenges such as the difficulty of representing complex data such as ontologies and data models. In addition, the data is usually somewhat subjective and it is difficult to measure the quality of this data.

With the development of new molecular technologies, the ability to incorporate genomic data into longitudinal data collections has increased significantly. The many challenges with genomic data collection are somewhat different than for phenotypic and environmental information. For example, data complexity challenges when collecting environmental information are usually not applicable to genomic data, which is typically relatively simple in terms of data structures and ontologies needed to store information.

Genomic Data Characteristics

Genomic data is fundamentally different from data that is traditionally collected in longitudinal studies. Two aspects play important roles in this difference.

Lower Degree of Data Complexity

The complexity of genomic data is usually lower than traditional longitudinal data. In its purest form, genomic information is very simple to describe with only the four available nucleotides A, T, G and C. Although many genetic assays provide much more complex data structures, the complexity, and thus the difficulty of selecting ontologies and data models, for genomic data is lower than for environmental data.

Extreme Data Volumes

Compared to traditional data collected in longitudinal studies, genomic data volumes are much larger. Figure 4 shows the number of variables collected for a single individual in a typical epidemiological longitudinal investigation with genetic data.

Although these are hypothetical numbers, they are representative of what is typically available today. The significant increase in the volume of genomic data introduces new challenges in data management, described below.

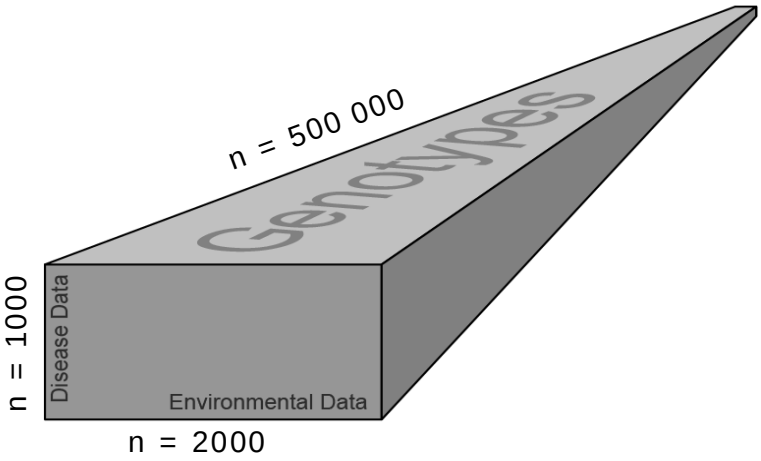

Figure 4: Data Volume from Typical Epidemiological Longitudinal Investigation

Major Challenges

Scientific Challenges

The volume of data in genomic investigations highlights new challenges from a scientific perspective. As the data volumes increase, science will become more data-driven as a complement to the more traditional hypothesis-driven paradigm. New ways of working are required. For example, when 500 000 variables are available for every subject, how would you search – how do you even know which variables are of interest?

Another challenge is the interpretation of data, which can be very difficult when comparing information from different genotyping platforms and source material.

Ethical and Legal Challenges

Of all issues and challenges, the ethical issues are probably the most important ones to address. It is very important that ethical committees are involved and approve all genomic data collection in longitudinal studies. As the public becomes more aware of the opportunities – and threats – associated with genomic data, the process of informing participants becomes a major challenge. An additional challenge is managing the collection of consent forms from the participants where options range from narrow to broad consent models.

Ensuring the privacy of the data is critical. It is important to restrict access to this data from unauthorized organizations such as insurance companies and employers, and also to protect the actual participants from obtaining this data. Since the interpretation of genomic data is very complex, how the information about their genetic profiles is shared with the participants must be carefully considered.

Technical Challenges

Many technical challenges are associated with incorporating genomic information in longitudinal data collection, including choosing a molecular platform and appropriate biospecimen material. However, the overwhelming technical difficulty is the extremely large data volumes that are produced by large-scale genotyping. The data volumes are large both in terms of the number of variables as described above, and also in the physical disk space required to store this data.

The genomic data requires new data models, new tools and new methodologies for handling the data compared to traditional epidemiological investigations.

Financial Challenges

There are some significant financial challenges with integrating genomic data into a large longitudinal study. Large-scale genotyping is today very expensive, although the price will continue to drop. Two models exist – either genotype a lot from the start, or perform targeted genotyping when necessary (on demand genotyping). Depending on the type of research, both of these models must be used. Different biobanks across the world have chosen different strategies to address this challenge.

Approaches for Addressing the Challenges

Karolinska Institutet Biobank BIMS

Key goals of the Karolinska Institutet (KI) Biobank include:

- Creating a link between the actual biobank samples and the large databases containing phenotype and genotype information of the individual donors
- Providing a common comprehensive user interface to access and explore the data

The data integration system – the Biobank Information Management System (BIMS) – has been co-developed by IBM and KI. In its latest release (3.0) this system incorporates significant function for managing extremely large volumes of genotypic information, and it also connects the genotypic data to clinical, lifestyle and register information. Built on IBM technology, it provides the following capabilities:

- Integration of a large number of data sources of different formats, at different locations and with different data models. The information provided spans from clinical records to epidemiological questionnaire information and disease register data.
- Management of enormous genotype data volumes. The solution can store and query up to one million individuals with one million genotype markers for each individual.
- Advanced query interface for all data stored within the system
- Efficient security measures for controlling the access to data

Genomic Messaging System

It is a difficult challenge to efficiently carry annotation and unconditional or fine grained patient consent for research along with the content data, in a way that will survive transfer between many different formats and ontologies, yet pro-actively help ensure compliance. The Genomic Messaging System (GMS), also known as the Clinical Laboratory Messaging System (CLaMS), was developed to address interoperability:

- Efficiently
- Including disaster recovery
- With a variety of healthcare and life science systems
- With different protocols and formats
- Conserving governance and compliance embedded in content data
- With data moved between diverse formats

GMS is based on information theoretic principles to allow conciseness with diverse possible modes of use and novel extra forms of encryption. Both sender and receiver must have the GMS program. GMS is a communication and storage protocol constructed in a way that does not replace existing communications or standards – it can embed or lie within them. It has tools to handle new formats and ontologies that are encountered. It has special data types to transport DNA, protein sequence, clinical record, and medical image information, as well as embedded code to be executed on receipt, which may include upgrades of GMS itself. Other features of GMS include multiple password protection at multiple points in the communication stream and commands capable of directing workflow, so that for example, fine grained patient consent will only permit consented work, and by the appropriate persons.

GMS Language (GMSL) is the format transmitted or stored and is highly compressed, resembling machine code. Every 0/1 bit has meaning, and there is inbuilt error detection. For example, DNA bases AGC and T can optionally be represented by 2 bits. It is a universal language into which a variety of legacy formats and ontologies can be disassembled and reconstructed in original or new format or ontology. GMS contains certain additional character and format compliance tools, notably for miscibility with XML, though it is larger in scope than XML. Passwords and other features can be inserted into DNA streams or images, so that even if the DNA data is removed and inserted into a new format or ontology, it maintains compliance.

GMS can be used in a grid computing architecture, and is constructed to facilitate continuing local operation and recovery even in the case of collapse of the supporting healthcare IT infrastructure.

7. Needs for community-wide standards or best practices that will facilitate large-scale data integration and exchange to benefit personalized health care

IBM believes that standards are crucial in order to realize personalized healthcare. In many cases of personalized practice, the exchange of information is done between totally disparate organizations such as healthcare providers and innovative genetic testing facilities. The exchanged data is complex and consists of two main types of data – clinical and genomics. Existing standards in healthcare and life sciences serve one of these two types of data and thus are not the best fit for the job. There is a need for an overarching standard to integrate both types of patient-specific data.

Genomic data typically consists of raw data (e.g., full sequence of the patient's screened genes) and its interpretation is performed by the genetic facility (e.g., a deleterious somatic mutation has been found). It is fundamental for any selected standard to be capable of holding both the raw data and the interpretations. In this way, the raw data can be interpreted again by different experts, algorithms, and decision support applications, as well as at different points in time when new scientific discoveries become available.

Genomic data varies in its complexity and the extent to which it is used. Simple testing identifies genes and mutations; more complex assays include full DNA sequencing, RT-PCR for the expression level of a small number of genes, and microarrays to identify the expression levels of vast numbers of genes in the individual. The place of genomic data sets within common clinical information constructs can be similar to that of other common health observations, but there are several characteristics that distinguish genomic data sets from typical observations such as blood pressure or potassium level:

- The amount of data. Typically, a single genetic locus is not sufficient.
- The complexity of the data. The DNA sequences (. . . AGCT. . .) need to be represented along with their variations, transcription outcome, and translation to proteins.
- Detailed descriptions of the methods used to obtain the genomic data. These are necessary for the receiver to interpret the data correctly and assess its level of reliability.
- The interpretation of the data. This is constantly evolving as new discoveries are made.
- The emerging common formats used by bioinformatics communities, for example, the Bioinformatic Sequence Markup Language (BSML) and the Microarray and Gene Expression Markup Language (MAGE-ML).
- The semantics of the genotype-phenotype relations, which are represented in a variety of ways, depending on the point of view (clinical research, pharmaceutical, or healthcare).

IBM believes that there is a need for an overarching standard that will address the above challenges of bio-medical integrated informatics. The design principles that the selected overarching standard should implement are:

- Bridge between clinical and genomic data and create a continuum of patient-centric data anchored in the patient EHR (the longitudinal and cross-institutional electronic health record)
- Use the best-of-breed approach and bring together recognized clinical standards and common bioinformatics markups used by the life sciences community.

- Use a clinical standard as the overarching format and use bioinformatics markups to encapsulate the raw genomic data specific to the patient in the clinical standard, which can then be incorporated into the patient's EHR.
- Constrain the selected bioinformatics markups to make sure that they fit the clinical standard requirements, in particular the requirement to uniquely identify the patient, which is not mandatory in most bioinformatics markups.
- Use of XML as the syntactical framework of the proposed overarching standard is recommended, since it is the preferred content framework in IT and it is also very common in bioinformatics.

Based on the above principles, IBM recommends the use of the following standards for personalized healthcare.

The Overarching Standard

IBM recommends the use of HL7 Clinical Genomics Specifications as the overarching standard for personalized healthcare. HL7 is the major standards organization in healthcare and its Clinical Genomics Special Interest Group (CG SIG) has been developing standards for personalized healthcare. These standards follow the above principles, and early adopters are already experimenting with them. Many stakeholders contributed to and participated in the Clinical Genomics standardization process, including providers Partners Healthcare and Mayo Clinic, vendors IBM and GE, pharmaceutical companies Merck and Elli Lilly and US governmental agencies NCI, CDC and FDA.

The Clinical Genomics standard supports the routine use cases in health care such as detection of known mutations, while preparing the information infrastructure for more advanced cases, such as full sequencing and detection of somatic mutation or the use of gene expression methods.

The core model of the HL7 Clinical Genomics specifications is the GeneticLocus model, which consists of placeholders for various types of genomic data relating to a specific locus on the genome (e.g., DNA or RNA) including sequencing, expression, and proteomic data. Within the GeneticLocus model, existing bioinformatics markups were utilized to represent raw data received from genomic facilities, enabling the principle of raw data encapsulation. These markups were constrained to make them compliant with HL7. The GeneticLoci model describes a set of loci, such as a haplotype (several variations of the same chromosome), a genetic profile, and genetic testing results of multiple variations or gene expression panels. The GeneticLoci model uses the GeneticLocus model to describe each of these loci.

Another part of the HL7 CG specification is the FamilyHistory model, to describe a patient's genetic family history (pedigree) with clinical and genomic data. It also includes a message interaction between two disparate pedigree applications, in which a person's pedigree is sent from one application to the other. This model utilizes the GeneticLocus model to carry the genomic data for the patient's relatives. Breast and ovarian cancers are among the use cases where there is a clear need to represent a patient's family history, because these diseases run in families. These cancers are associated with variations in the breast cancer genes. A complete family history can be assessed for potential risk, supporting clinical decisions in genetic counseling and at the point of care.

Genomic Data Standards

Bioinformatics markups are utilized in the HL7 Clinical Genomics standard. Through a consensus process, the HL7 Clinical Genomics SIG considered several bioinformatics markups and decided to recognize a few of them. The schemas of the selected markups are constrained to meet the HL7 criteria, which are mainly concerned with the unique identification of the patient. The only two markups recognized so far are MAGE-ML for gene expression data and BSML for sequencing data.

It is important to note that the goal is not necessarily to select one bioinformatics markup for each type of data. Rather it could be that a number of markups will be selected and constrained to allow more flexibility at the raw data representation. Nevertheless, all recognized markups that represent the same type of data (e.g., sequences) will be mapped to the HL7 Clinical Genomics standard in a way that assures that the same essential data set can be extracted from the raw data and placed in HL7 objects. Those objects can then be associated with phenotypic information, for example, clinical observations of the patient or scientifically known phenotypes such as known drug responsiveness.

Encapsulate & Bubble-up

To summarize the conceptual workflow that the HL7 Clinical Genomics standard enables, the metaphor of bubbling up is used. Bubbling up differs from the notion of summarization or annotation. Summarization implies that the data being summarized is thoroughly understood, whereas it is difficult to summarize clinical genomics data because most of the knowledge in this field is still unknown. Bubbling up is similar to gem mining activities; one does not always know what will be found, and the findings vary depending on the location. Annotation is the result of bubbling-up processes, whereby the data which is input to these processes is annotated and, more important, enriched through the creation of new or annotated genotype-phenotype associations. The encapsulate and bubble-up workflow may lead to gradual distillation of the raw genomic data in the context of diagnosis and treatment provided to a specific patient at a specific time, while making the raw data available within the patient's medical records so that it may be possible to parse it again (for example, by alternative algorithms or when new knowledge becomes available).

Clinical Genomics Level Seven

Clinical Genomics Level Seven (GCL7) is a set of services that can be used by clinical decision support applications. It complies with the HL7 Clinical Genomics standard and provides services to encapsulate raw genomic data (e.g., sequencing or expression data) in HL7 compliant messages. In addition, CGL7 attempts to create new genotype-phenotype associations based on patient data available in medical records along with information found in publicly available resources such as OMIM. Overall, CGL7 realizes the encapsulate and bubble-up workflow paradigm underlying the HL7 Clinical Genomics standard.

CGL7 can be used either as Web Services or as a specialized API for rapid development of decision support applications. If persistency is required, CGL7 can communicate with any HL7 V3 RIM-based repository. For example, it can run on the IBM Clinical Genomics solution.

More details about CGL7 can be found in the paper "The seventh layer of the clinical-genomics information infrastructure" at:

http://www.research.ibm.com/journal/sj/461/shabo.html

9. Development of ontologies across different clinical data repositories that will facilitate the utility of the data for answering clinical research questions

The Ontological Continuum

IBM believes that publicly-available ontologies will play a significant role in promoting clinical research and in bringing research results to practice. Ontologies are metadata and are part of a continuum. The continuum starts with enterprise metadata, continues to biological and medical terminologies, and on to reference databases and registries, as well as information exchange standards. Publicly-available and professionally-acknowledged ontologies should be the main mechanism through which we achieve semantic interoperability. Functional interoperability is not sufficient as it only guarantees the exchange of data between disparate entities but does not provide the full meaning of any received data item. There is an urgent need for a semantic layer that can be provided by ontologies in conjunction with the entire ontological continuum of metadata.

The ontological continuum:

- Starts with actual patient data (observations, images, documents)
- Then uses terminologies to understand the meaning of codes used in the data (e.g., the exact meaning of SNOMED code 230145002)
- Continues to registries of standard specifications to understand the constructs used in the data (e.g., a Complete Blood Count construct embedded in a Lab Results message)
- Finally accesses ontologies that provide a broader context of the data, codes and constructs used in patient records or in research data sets

Ontologies are about the rich description of relationships between concepts that are a part of the information infrastructure needed for decision support applications to infer the full context and meaning of the data.

Ontologies and Standards

Standards facilitate the exchange of data and information. Nevertheless, standards are generic by nature, and localization, customization and extension of a generic standard to fit an organization's requirements are common. These processes challenge the semantic interoperability goal, and ontologies could address this challenge by describing the refinements done to a given standard.

The use of internationally-recognized standards in the IT industry is pervasive and is essential for successful information exchange. Industry standards in healthcare and life science often branch out to specialties and subspecialties. For example, a generic specification for a clinical document is usually insufficient, since each subspecialty in healthcare desires its 'own format for clinical documents (e.g., a hematology discharge summary for chronic leukemia patients). The branching out process should be done in such a way that each of the leaf artifacts can be validated against a central information model, thus achieving semantic consistency. This process would ensure that the lower-level components can be reused by higher-level artifacts. For example, a tissue typing observation could be used in a tissue typing lab report as well as in a bone-marrow transplantation procedure summary.

Standardization of health data is still in its infancy. Only the most common data sets are standardized (e.g., lab and pharmacy), most often at the structural level. It is anticipated that many standard specifications will be fine-tuned for the needs of the various subspecialties/use

cases/scenarios found in healthcare and life science. In addition, for each topic there will be multiple specifications. Even for common data such as blood pressure, there are different ways to represent the relevant data items, as well as their relationships and constraints. Such representations are referred to as Templates or Archetypes in the medical informatics and standardization communities.

Ontologies of standards-derived Templates/Archetypes will register and describe any standard specification submitted by recognized organizations. The National Institute of Standards and Technology (NIST) is developing a web-based repository, which will focus on indexing the Template's metadata. These repositories will evolve to be a major part of healthcare and life sciences ontologies, which could also be considered as part of the Semantic Web vision [1]. In healthcare, the ontologies will be used by healthcare providers to comply with the latest versions of the standards, or to dynamically parse a medical record received from another healthcare provider complying with a different standard specification. True semantic interoperability and full integration of data from various sources will be dependent on these ontologies.

An example of the integration challenge is the emerging concept of the patient-centric, cross-institutional and longitudinal Electronic Health Record. In order to present a coherent and useful EHR at the next point of care, data created by different standards, templates, and terminologies throughout an individual's lifetime must be integrated. The longitudinal EHR is considered the final frontier of health informatics [2-3]. However, only if it is parsed by ontology-aware applications can it become more than just inconsistent health data.

The Structured ←→ Unstructured Information Challenge

Many existing ontologies contain vast amount of data, information and knowledge. Due to the complexity and rapid changes in health information (especially in life sciences), many of these ontologies are unstructured. The following description is from OMIM (On-line Mendelian Inheritance in Man, [4]):

> *Despite the dramatic responses to EGFR inhibitors in patients with non-small cell lung cancer, most patients ultimately have a relapse. Kobayashi et al. (2005) reported a patient with EGFR-mutant, gefitinib-responsive, advanced non-small cell lung cancer who had a relapse after 2 years of complete remission during treatment with gefitinib. The DNA sequence of the EGFR gene in his tumor biopsy specimen at relapse revealed the presence of a second mutation {131550.0006}. Structural modeling and biochemical studies showed that this second mutation led to the gefitinib resistance.*

This anecdote demonstrates what is envisioned for patients in the personalized medicine era. In this study, the researchers detected a second somatic mutation in the patient's tumor tissue. They used structural modeling and biochemical studies to understand the influence of this mutation on the drug the patient was taking. However, the description of this case is unstructured, and even the OMIM entry ids (e.g., identifying mutation entries in OMIM) are not consistent with other conventions and notations. For example, sequence variations are different in OMIM and HGVS (Human Genome Variation Society). Thus, there is a need to raise the level of structuredness and compliance among existing ontologies that are so critical for personalized healthcare.

From SOA to the Semantic Web

Ontologies should be publicly available and accessible through Web Services, to enable them to participate in a service-oriented architecture at the enterprise level. At the same time enterprises can contribute to the vision of the Semantic Web [1], which facilitates the full utilization of ontologies in the distributed environment of health IT systems.

Summary

The critical characteristics of ontologies for personalized healthcare and clinical research are:

- Integral part of a continuum of metadata, terminologies, exchange standards and registries of domain specific standard-based templates
- Publicly available
- Implemented through Web Services with an underlying infrastructure that orchestrates the entire ontological continuum
- Strive to the highest level of structuredness

References:

1. Berners-Lee, T., Hendler, J., and Lassila, O. (2001). “The Semantic Web.” Scientific American, 284(5): 35--43.

2. Shabo, A. (2003). Integrated EHR: The Final Frontier. BioIT World, November 2003. See http://www.bio-itworld.com/archive/111403/strategic_ehr.html.

3. Shabo, A. (2004). Synopsis of the Patient Records Section: Structuring the Medical Narrative in Patient Records – A Further Step towards a Multi-Accessible EHR. The IMIA 2004 Yearbook of

4. Medical Informatics: Towards Clinical Bioinformatics.

   Online Mendelian Inheritance in Man (OMIM). See quoted entry at: http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=131550.

10. Models for linking clinical data repositories across disparate care providers

IBM's model for linking clinical and genomic information across disparate care providers is a standards-based Personal Health Care Information Exchange (PHCIE). Our proposed PHCIE is specifically focused on supporting the following critical activities:

- Retrieving longitudinal patient data from multiple sources for each patient, including genetic profiles, regardless of where or in what format the data is stored. This entails integration of clinical data within and across enterprises, communities, regions, states, and even nationwide.
- Aggregating, normalizing, de- and re-identifying data into large clinical data warehouses, and smaller specialized data marts, to support research into clinical outcomes, new interventions, identifying individual differences in response to therapies.
- Reporting notifiable disease conditions and healthcare-associated infections to public health authorities
- Stringent enforcing of privacy policies as defined by patient consent protocols, and auditing access to clinical information.

The architecture described in this section is modeled after the architecture proposed in our ONCHIT Nationwide Health Information Network architecture prototype. Like the NHIN, the proposed PHCIE is concerned with secure, standards-based integration of clinical data across disparate care providers. Unlike the NHIN, the proposed PHCIE emphasizes research/evaluation of aggregated data, and includes genomics data as well as traditional clinical data.

Architecture

In this section, we describe the components of the proposed PHCIE architecture, explain how the components work together to address the four critical integration activities outlined above, and map them to products from IBM and our business partners.

Figure 5 below provides a functional view of our proposed PHCIE. Stakeholders include clinicians, researchers, patients, and public health surveillance authorities. Data comes into the PHCIE from:

- Hospital Electronic Medical Records
- Patient Personal Health Records
- 3rd party vendors providing information such as medication histories and genetic profiles
- Reference Labs
- Public genomics databases on the Internet
- Ultimately, state, regional, and national Heath Information Exchanges

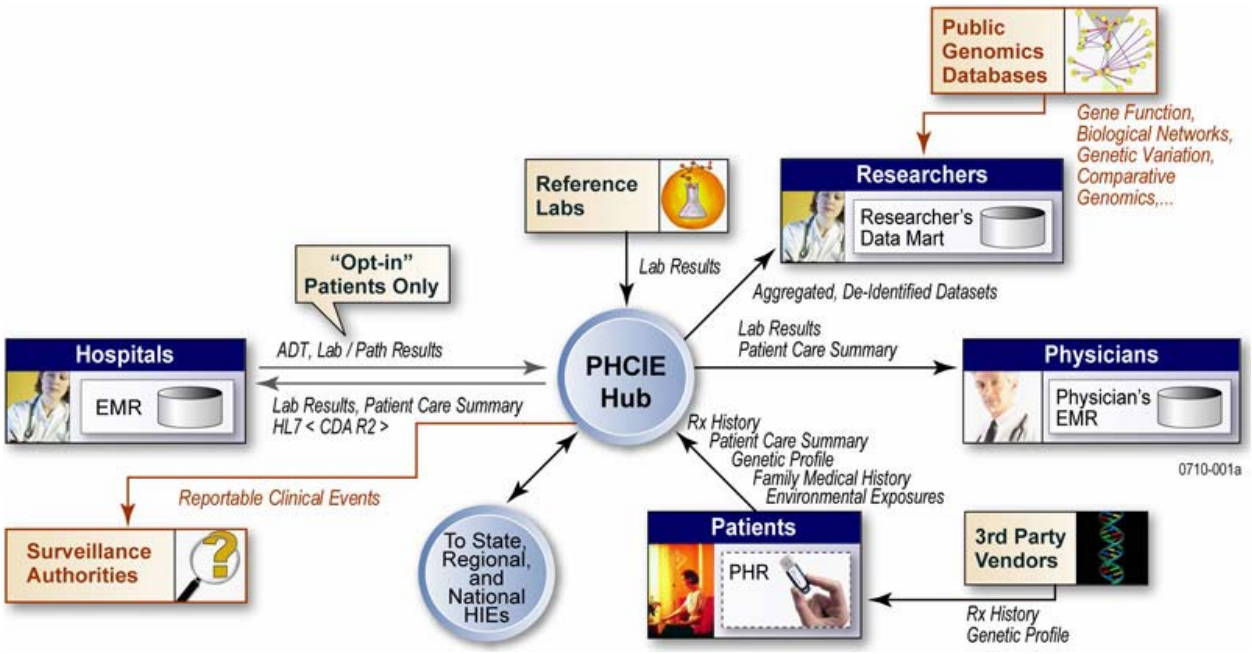

Figure 5: Functional Diagram of Personal Health Care Information Exchange (see Note 1 below)

Standards-based Architecture

One of the key features of our proposed PHCIE architecture is that it is standards-based. Using standard formats and profiles for data exchange that are shared throughout the healthcare community enables stakeholders to share data, both within the PHCIE and across other regional, statewide, and nationwide PHCIE's. The emerging NHIN is an example of a nationwide Healthcare Information Exchange (HIE) that can form the basis of a nationwide PHCIE.

The architecture of our proposed PHCIE is based on the Cross-Enterprise Document Sharing (XDS), Patient Identifier Cross-Referencing (PIX), Audit Trail and Node Authentication (ATNA) and Patient Demographics Query (PDQ) profiles of IHE (Integrating the Healthcare Enterprise, www.ihe.net). Below is a brief summary of the relevant IHE Profile Definitions as specified in the IHE IT Infrastructure Technical Framework, volume 1 (ITI-TF1).

IHE Patient Identifier Cross-Referencing (PIX)

The PIX profile provides cross-referencing of patient identifiers from multiple Patient Identifier Domains. These patient identifiers can then be used by identity consumer systems to correlate information about a single patient from sources that know the patient by different identifiers.

Patient Demographics Query (PDQ)

The PDQ profile provides ways for multiple distributed applications to query a central patient information server for a list of patients, based on user-defined search criteria, and retrieve a patient's demographic (and, optionally, visit or visit-related) information directly into the application.

Cross-Enterprise Document Sharing (XDS)

The XDS profile enables a number of healthcare delivery organizations that belong to a clinical affinity domain (e.g., a community of care) to cooperate in the care of a patient by sharing clinical records in the form of documents as they proceed with patient care delivery activities. This profile is based upon ebXML Registry standards, SOAP, HTTP and SMTP. It describes the configuration of an ebXML Registry in sufficient detail to support Cross Enterprise Document Sharing.

Audit Trail and Node Authentication (ATNA)

The ATNA profile establishes the characteristics of a Basic Secure Node:

1. It describes the security environment (user identification, authentication, authorization, access control) assumed for the node so that security reviewers may decide whether this matches their environments.
2. It defines basic auditing requirements for the node.
3. It defines basic security requirements for the communications of the node using TLS or equivalent function.
4. It establishes the characteristics of the communication of audit messages between the Basic Secure Nodes and Audit Repository nodes that collect audit information.

Cross-Enterprise Document Sharing--Imaging Document Source (XDS-I)

The XDS-I profile is responsible for providing access to DICOM images within a community environment

For more information on the IHE and a detailed description of the XDS, PIX, and PDQ Profiles see Attachment 7, “Overview of IHE Integration Protocols.” For a summary of additional healthcare standards supported by the proposed PHCIE, see Attachment 8.

HL7 Clinical Genomics Standard

The HL7 Clinical Genomics standard specification serves as an overarching framework for clinical genomics IT solutions, to allow complex integration of data to be entirely realized using an HL7 V3 RIM based repository. Additional information on our use of the HL7 Clinical Genomics standard is in the “Shared EHR and PHR Data Repository” discussion in the next section.

Components and Products

The software components comprising the PHCIE architecture are shown in Figure 6.

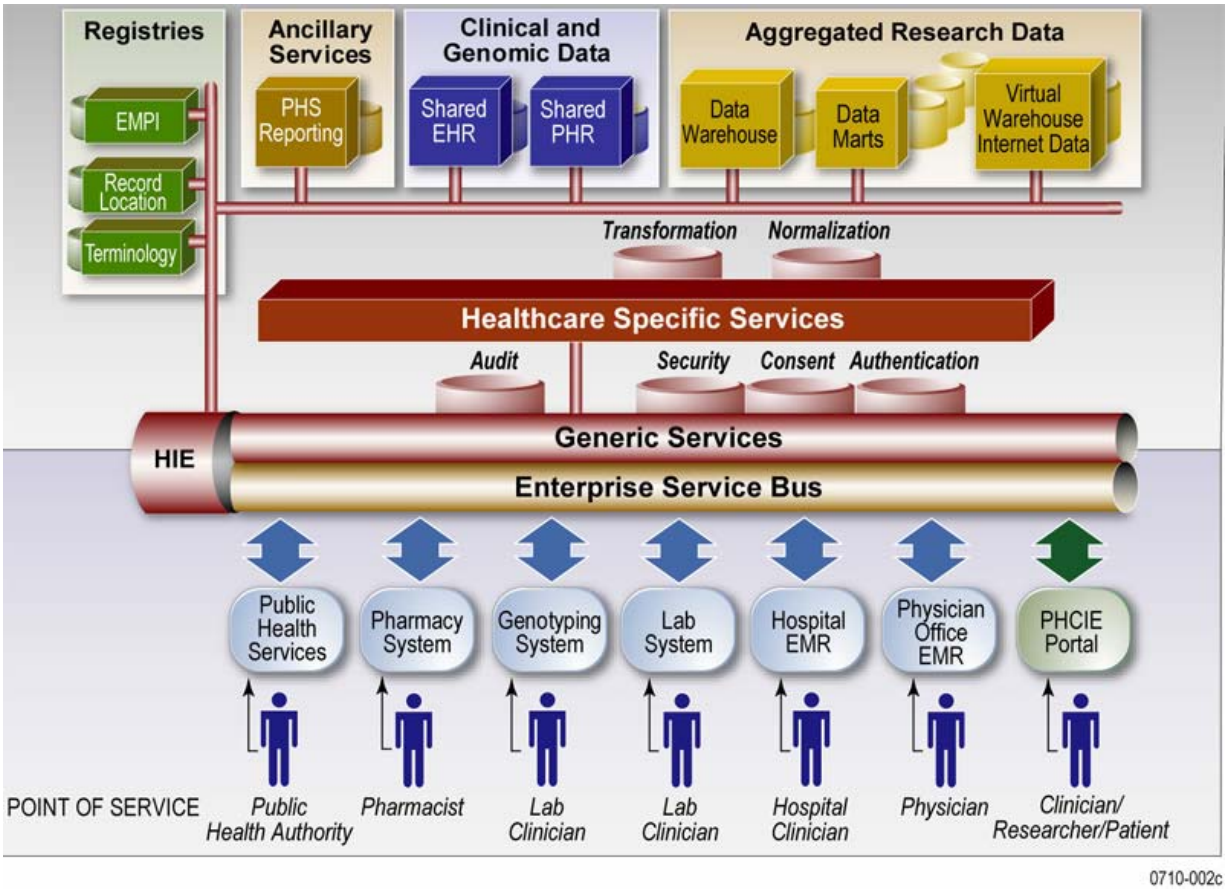

Figure 6: Personal Health Care Information Exchange Architecture

eMPI (Master Person Index)

This service component provides a single, trusted view of each patient by consolidating duplicate and fragmented patient records. It identifies and instantly links the records for each patient or other entity within and across the member organizations of a Regional Health Interchange Organization. The proposed PHCIE system can utilize any eMPI that supports the IHE PIX and PDQ integration profiles. However, we recommend the best-in-class eMPI offered by our business partner Initiate Systems.

Initiate Systems' Identity Hub software compares patient records using a probabilistic algorithm, one of the most proven and field-tested algorithms available. During the past 17 years, the algorithm has been used to analyze over 2 billion records. Importantly, a significant number of those predictions have been field verified during the course of data cleansing/MPI remediation projects. The results of the on-site verification have been channeled back to scientists resulting in a natural continuous improvement loop, resulting in few false positives (mistaken matches) or false negatives (missed matches).

Technology to enable probabilistic matching is imperative if computers are to replicate, consistently and effectively, the evaluation and judgment processes of human clerks attempting to link common records. Ideally, computers would emulate the intuitive thought processes of human beings as they review, judge, evaluate, measure, and score linkage qualifications of records representing commonality. Neither the technique of shared identifiers nor the deterministic matching method is able to match records under conditions of uncertainty. Only

probabilistic matching mimics the human ability to recognize that two slightly dissimilar records are in fact the same object. For more information on our probabilistic matching algorithm, see Attachment 9.

Products that can provide this function:

- Initiate Systems Identity Hub 7.0
- Initiate Systems HL7 Query Adapter 7.0
- IBM WebSphere Application Server V6.0

Record Location Service

Clinical data in XDS Document Repositories is registered with the XDS Document Registry. The Registry contains pointers to the data in the Document Repositories (typically a set of URIs), along with enough metadata to enable retrievals of individual documents in the set (e.g., patient id, record type). For more detail on how the XDS Document Registry interacts with the Document Sources, Consumers, and the PIX/PCS MPIs, see Attachment 7.

Products that can provide this function:

- IBM IHII Document Registry
- IBM DB2 Universal Data Base v 8.2
- IBM WebSphere Application Server V6.0

Terminology Registry

See Normalization Service below.

PHS Reporting

The PHCIE Public Health Reporting service is based on IBM WebSphere® Business Integration for Healthcare Collaborative Network (WBI for HCN). This product provides a fast, security-rich foundation for detecting and responding to adverse health events, including bioterrorism.

WBI for HCN is based on a hub-and-spoke architecture. As the hub of the network, it sits at the center of the data exchange, providing the administrative and control aspects of the network. The hub works with gateways, which are, the spokes. It has two primary components – the administrative server and the message flow server.

The administrative server component of enables participants to identify and configure the types of clinical information (topics) they want to receive (subscribe to) or are willing to provide (publish). It also enables subscribers to set up alerts based on business rules they approve.

The message flow server component transfers messages from publishers to authorized subscribers, enabling the system to perform content-based routing of clinical topics. The message flow server includes a configuration manager, user name server, message broker, and queue manager. It provides a high-speed message broadcast component that can keep track of a system participant's interest in clinical topics.

Products that can provide this function:

- IBM WBI for Healthcare Collaborative Network

Shared EHR and PHR Data Repository

The XDS-compliant EHR/PHR Data Repository performs multiple critical functions. It acts as a point of data aggregation for a community, providing integration of data across the community. It also makes available additional functions that may not otherwise be available from the individual providers' existing systems. The Data Repository is intended to mitigate the scalability, availability, and robustness problems associated with (1) the lack of standards compliance and (2) the heterogeneous hardware and operations support of individual sources. The Data Repository acts as a buffer that helps to prevent single system failures at the local level from propagating into the exchange, decreasing performance and/or preventing critical data from reaching the point of care. The Data Repository functions in the larger architecture as an XDS Document Repository (see Attachment 7 for details).

The HL7 Clinical Genomics standard specification serves as an overarching framework for the Data Repository, so that complex integration of clinical and genetic data may be realized in an HL7 v3 framework. The Data Repository is a relational database. It has a schema is based on the HL7 v3 Reference Information Model (RIM), with extensions for genomics data sources such as gene expression data (MAGE-ML), genetic variation data (BSML/HapMap), CDISC ODM for clinical trial data, and HL7 V3 messages and CDA for clinical data. The Data Repository is therefore a dual XDS – HL7 RIM based repository.

Products that can provide this function:

- IBM DB2 Content Manager V 8.2

Research Data Warehouse and Data Marts

The Research Data Warehouse is the central repository of clinical and genomic data for analysis. It is populated through Extract, Transform, Load methods from the EHR and PHR data repositories. Where necessary, data is cleansed, transformed, normalized, and deidentified by other component services before it is loaded. Once the warehouse has been initially loaded, the schedule for updates from the EHR and PHR repositories is flexible – weekly, nightly, or in real time, depending on the currency needs of the researchers.

The Research Data Warehouse is based on the DB2 Data Warehouse Edition, an integrated platform for developing warehouse-based analytics. It includes core components for warehouse construction and administration, as well as Web-based applications with embedded data mining and multi-dimensional online analytical processing. It provides an infrastructure for warehouse building and maintenance, including tools for application design, deployment, execution and administration.

Specialized analysis may require only small subsets of the full Research Data Warehouse, e.g., cancer patients, women over 50, cardiac patients. In addition, a restructuring of the data might make the analysis much easier to perform. For this we offer state-of-the-art tools to create and populate specialized Data Marts from the full Research Data Warehouse.

Products that can provide this function:

- IBM DB2 Data Warehouse Edition

Virtual Warehouse for Internet Data

To investigate genes, proteins, and biological pathways that may help explain individual differences in health states, disease processes, and outcomes from interventions, researchers require more than clinical data and genetic profiles. They require full, flexible query access to an integrated, up-to-date view of all related genomic information, irrespective of where it is stored (within an organization or across the Internet) and its format (traditional database, flat file, web site, results of runtime analysis).

IBM WebSphere Information Integrator is a readily extensible data integration system. It implements federated search to provide integrated access to life sciences data sources. Rather than simply duplicating and loading all data sources of interest into a common local data warehouse, this federated middleware wraps the actual data sources in place, encapsulating the details of the sources and how they are accessed. Users can write single declarative SQL queries that span multiple data sources, in a variety of formats, distributed over the Internet and within the organization, as if they are components of a single relational database. Data sources currently supported include NCBI Entrez databases (e.g., Nucleotide, Protein, Genome, Gene, SNP, Unigene, PubMed), OMIM, KEGG, the BLAST and HMMER sequence analysis algorithms, regular-expression pattern matching functions, and user-developed scripts in any language (as long as they take inputs as command-line parameters and output results as XML).

As high throughput biology generates large volumes of systems biology data, there is a growing need for robust, efficient systems to manage metabolic and signaling pathways, chemical reaction networks, gene regulatory networks, and protein interaction networks. Network data frequently is best represented as graphs, and researchers need to navigate, query and manipulate this data in ways that are not well supported by standard Relational Database Management Systems. For exploring biological pathway data, the Systems Biology Graph Extender, a research prototype, extends the IBM DB2 Universal Database* software with graph objects and operations to support declarative SQL queries over biological networks and other graph structures. Supported operations include neighborhood queries, shortest path queries, spanning trees, graph transposition, and graph matching. In a federated context, graph operations may be applied to data stored in any format, whether remote or local, relational or non-relational. A single federated query may include both graph-based predicates and predicates on related data sources, such as microarray expression levels and clinical prognosis/outcome.

Products that can provide this function:

- IBM WebSphere Information Integrator
- IBM Systems Biology Graph Extender

Transformation Service

This service component converts between different data format and messaging standards, to permit the exchange of information between disparate systems. It supports data transport and translation among systems via pluggable brokers that transform multiple data formats, including HL7 2.3, XML, X12, ASTM, NCPDP, DICOM 3.0, HL7 CDA R2 and R3, and other structured formats.

This component transforms clinical data from non-XDS-compliant systems so that a comprehensive picture can be made available via the IHE Integration Profiles, to be analyzed at

the subscriber end point. It also verifies that only approved data is securely published in the agreed-upon manner (de-identified, for example, if specified by the sender).

Products that can provide this function:

- IBM WBI for HCN Gateway V1.0
- IBM WebSphere Application Server V6.0

Normalization Service

In populating an aggregated data warehouse from a variety of disparate sources of clinical and genomic information, it is critical for the data to be normalized to a common semantics (common terminology, common units). This service component provides conversion capabilities between different vocabularies, using the mappings specified in the Terminology Registry. Our Terminology component provides code set and medical nomenclature management and mapping, supporting a variety of medical vocabularies, including CPT, ICD, LOINC, SNOMED, NDC, RxNorm, and HCPCS. It is based on the HL7 Common Terminology Service standard, as implemented by the Mayo Clinic Lexgrid project and contributed to the open-source Eclipse Open Health Framework project.

Products that can provide this function:

- Mayo Clinic Lexgrid
- IBM WebSphere Application Server V6.0

Security/Authentication

This service component provides user authentication and access control to information retrieved from source sites. The system utilizes a federated approach to security services wherein security credentials are passed between security domains via defined schemes. Trust relationships are established between participants of the PHCIE and allow sharing of credentials and subsequent exchange of information. The architecture permits multiple layers of data filtering to control access. Information can be filtered initially at the source in order to support and comply with local data access rules and individual patient consent.

Products that can provide this function:

- Tivoli Access Manager

Audit

This component provides a secure means of tracking and logging access, exchanges of information between systems, and user changes within a given system. Audit function is required to meet HIPAA as well as other government regulations. The Audit services allow users with appropriate authorization to view who has accessed the system, the stated reason for access, and the time of access.

Products that can provide this function:

- IBM DB2 UDB v8.2
- IBM IHII ATNA service

Patient Consent

This service component allows Healthcare Providers to manage individual patient consent to health information in order to maintain HIPAA compliance. The system provides the patient with a flexible means to restrict access to sensitive classes of health information both inside and outside the community. Patient Consent is implemented as follows:

1. Each patient specifies, from the list of all physicians in the network, those who may not VIEW that patient's clinical records.
2. Each patient specifies, from the list of all physicians in the network, the list of physicians who should not be NOTIFIED when a record for that patient becomes available.

The PHCIE security modules guarantee that, for a specific patient, the set of physicians who should be notified is a subset of the set of physicians who can view the patient's records.

In addition, each physician indicates whether he wants to be notified when new results for a patient are available (assumes physician is in the Notified list defined by the patient).

Products that can provide this function:

- Initiate Systems Identity Hub v 7.0

PHCIE Portal

Our PHCIE architecture is flexible and allows the healthcare provider to utilize any IHE standards-based portal solution. We recommend the XDS-compliant open-source Practice Management and Electronic Medical Record software from The Possibility Forge. This portal integrates software from the open source ERP package Compiere with customized modules for the healthcare industry. The PHCIE Portal is an XDS Document Consumer (see Attachment 7 for details).

Products that can provide this function:

- Possibility Forge EMR Portal

Enterprise Service Bus (ESB)

An Enterprise Service Bus is used to pass data among the PHCIE participants and to provide the infrastructure for service requests. The ESB is a very flexible architecture that allows the PHCIE to provide immediate value by interconnecting XDS-compliant (or XDS-enabled) systems. In a phased implementation, additional features and functions can be added to the infrastructure, and the infrastructure itself can scale to handle higher transaction and data volumes.

Products that can provide this function:

- IBM WebSphere Message Broker V6.0
- IBM WebSphere Business Integration For Healthcare Collaborative Network (Hub) V1.0

XDS Enabling Toolkits

This component helps existing non-IHE-compliant systems to become PIX/PDQ Consumers, XDS Document Sources, XDS Document Consumers, and ATNA audit clients. It consists of Java APIs and the OHF Bridge, a web service that IHE-enables non-Java applications. IBM recently contributed most of these components to the open-source Eclipse Open Healthcare Foundation for use by the healthcare community.

Products that can provide this function:

- OHF Client Classes
- OHF Bridge

Note 1:

- Public Genomics databases picture: Wosik, Ewa. "Probabilistic Boolean and Bayesian Networks." Connexions. March 29, 2006. http://cnx.org/content/m12395/1.4/.
- DNA picture: DNA double helix. Suhail Islam, Imperial Cancer Research Fund, London. http://www.sciencemuseum.org.uk/on-line/lifecycle/135.asp

12. Models for prioritizing analyses to fill gaps in evidence of effectiveness of therapeutic interventions for different populations

One type of model would identify and prioritize gaps arising when the scientific results related to best therapy selection are subjected to analysis. IBM is developing decision support systems for both the physician and the general researcher or epidemiologist. These are rule-based inference systems. The rules are based on well-founded expertise, on data-mining the clinical and clinical genomic record, and on the use of more general data analytics. The miscible rules describe both association of many qualitative components or events (e.g., female, nonsmoker, African American, high blood creatine, renal dialysis) or multivariance between many numeric values. Four classes of gaps are described below.

Class I Gap of Evidence

New rules can highlight missing evidence. Data mining has rapidly assembled many thousand medical rules, in a convenient statistically quantified form. Data mining has also discovered many new rules that can generate research proposals for validation and further exploration [1]. These new statistically quantified rules allow ranking by strength to help prioritize future study. The new rules have been identified by using the initial rules as queries against the medical literature. This overall highly automatable process can help fill gaps in our current medical understanding.

The need for developing procedures for considering further classes of gaps of evidence has become manifest even when sampling is planned to focus on specific disease states, or is sliced by exploiting special biological situations. An example is a recent eETHICS ONE project between IBM and the National Human Genome Center (Howard University and Hospital, Washington DC) and a Caribbean Medical School (St Matthew's School of Medicine) intended to focus on personalized medicine for African Americans. It addresses the construction and data mining of clinical genomic records of African Americans and their historical migration routes into the US. It includes consideration of the effect of relatively genetically isolated island and plantation communities, where gene pools are dominated by very few founding fathers and mothers. In the Cayman Islands, for example, Cayman ataxia, breast cancer, diabetes, obesity, and possibly learning disability and domestic violence, as well as certain beneficial factors such as a degree of business acumen, are known to, or may have, a genetic component. Cayman ataxia is now known to be correlated with a specific gene defect and hence has a strong biomarker. Different islands and historical plantation regions have different enriched disorders (or benefits), and so represent diverse sample hot spots.

Additional Classes

The additional three classes of gaps in evidence relate to gaps in the rules themselves. They represent a problem to be solved by discovering missing new rules, and thus reduce the difficulty to a Class I gap of evidence, which may be managed as above. In prior experience and pre-simulated studies for the eETHICS ONE project, it has become apparent that rules from the analysis are likely to be missing for several reasons. Examples underlying omission are:

- Lack of sufficient data for more complex rules and the need for more sampling
- Lack of computational power to discover the complex rules in a very large space of possible complex rules (there are circa 2N rules that are possible in principle from N parameters)

The further classes are best defined by resolution method, as follows.

Class II Gap of Evidence

This class of gaps is due to the current absence of rules that drastically modify existing rules. This particular gap of evidence also relates to the well known problem in handling high dimensional genomic data for clinical purposes. The eETHICS ONE study encountered cases where a combination of gene affects is dominant over independent and simple additive contributions. The likelihood is compounded because a full project would imply a constant stream of new genomic marker data, much of which is often of unknown relevance *a priori*.

Two or more genomic biomarkers or sets of biomarkers may independently suggest a particular best choice of therapy. There is no guarantee that the appearance of all these markers together in a patient will make that therapy the best choice. There could be an interaction between the directly or indirectly indicated genes that renders the treatment weaker, useless, or even in worst case lethal. Inference as a basis for further investigation can provide an *a priori* ranking for the importance of such further studies. Similar concepts allow prioritized investigation of other kinds of relationships which may, but do not necessarily, follow from knowledge accumulated so far. In addition, one should perform a more focused study on the rule components. This strategy, however, may be considered as governed by the above *a priori* prioritization. In general, all rules likely to be adopted should be preliminarily validated in a further drill down data analytic study to resolve incompleteness by discovering missing influences.

Class III Gap of Evidence

This type of gap is also pragmatically defined by its resolution method, which is a special probabilistic form of inference and occurs where rules with different content are conspicuous by their absence. It is exemplified by conclusion of the rule “there are no pregnant males.” In fact, such rules are now directly generated as strong negative association rules. Occurrence would be expected on the basis of the high frequencies of the individual components (here the abundance of male patients and of pregnant patients). The space of all things that might be is large, so testing for such is computationally demanding in much less trivial examples with many components.

Class IV Gap of Evidence

A further class of gap is defined in the way it may be resolved – by more sophisticated forms of inference from existing rules. An inference process is capable of deducing, from a set of rules, new rules which necessarily follow, and new rules which might follow and which form the basis of future study. As is the more general intent for all classes of gaps of evidence, one may thus prioritize future sampling and research to identify any relevance of a hypothesized rule D when given the relevance of rules A and B and C. This is most readily achieved by inference when there are points of reference to the content matter of future rules in the existing rule set, in the manner of a syllogism or certain types of logical puzzle.

Other Tools

Other computer based, but not empirical data-centric, methods can assist in prioritization. These methods generally fall into the realm of systems biology. Predictive simulations of *ab inito* quality imply that deduction is made from first basic physicochemical principles, not experimental evidence. Approaches in systems biology may also be semi-empirical, combing experimental data (including measures discussed in the footnote below) as input or for parameterization of the model, or for validation or filtering of output.

References

1. “Data mining and clinical data repositories: Insights from a 667,000 patient data set” Irene M. Mullins, Mir S. Siadaty, Jason Lyman, Ken Scully, Carleton T. Garrett, W. Greg Miller, Rudy Muller, Barry Robson, Chid Apte, Sholom Weiss, Isidore Rigoutsos, Daniel Platt, Simona Cohen, William A. Knaus, Computers in Biology and Medicine, 2006 Dec;36(12): 1351-77. Epub 2005 Dec 22

13. Strategies for accumulating patient data necessary for research that may not be available through EHRs

In considering data necessary for research, and the strategy for accumulating data that may not be available through the Electronic Health Record, it is first important to detail the information that is assumed to be available. Experience has shown that the following data elements are usually accessible in a sophisticated EHR that includes both inpatient and ambulatory data:

- Basic demographic data
- Insurance information for the patient
- Episodes of care including: Begin date, End date, Diagnosis, Associated physicians, Location of care, Episode summary
- Diagnoses/Problems list
- Allergies
- Vital signs
- Subjective and objective assessments
- Treatment plans
- Drugs ordered
- Drugs administered
- Lab orders and results, including anatomic pathology
- Radiology orders and results
- Transcribed reports (History and Physical, Discharge Summary)
- Nursing orders
- Nursing documentation
- Other ancillary orders (including supplies) and documentation (physical therapy, occupational therapy, dietary, respiratory care)
- Physician documentation
- Health Information Management ICD 9 codes, DRG's, APG's. E & M code, and others

Other patient data that may be necessary for research is dependent on the intent of the specific research. Some areas of data that are generally unavailable in an EHR, but could be advantageous in areas of research are:

- Genomic data
- PACS data such as Radiology and Cardiology images
- Videos (such as the video of a pumping heart)
- Cost of care by episode/by diagnosis
- Pervasive health data such as that acquired through patient monitoring devices
- Data acquired through Patient Portals or Personal Health Records which are reviewed and updated by the patient (over the counter drugs, supplements, vitals, exercise)
- Home Health care records
- Reference and referral lab data
- Payer data
- Data in paper-based electronic medical records

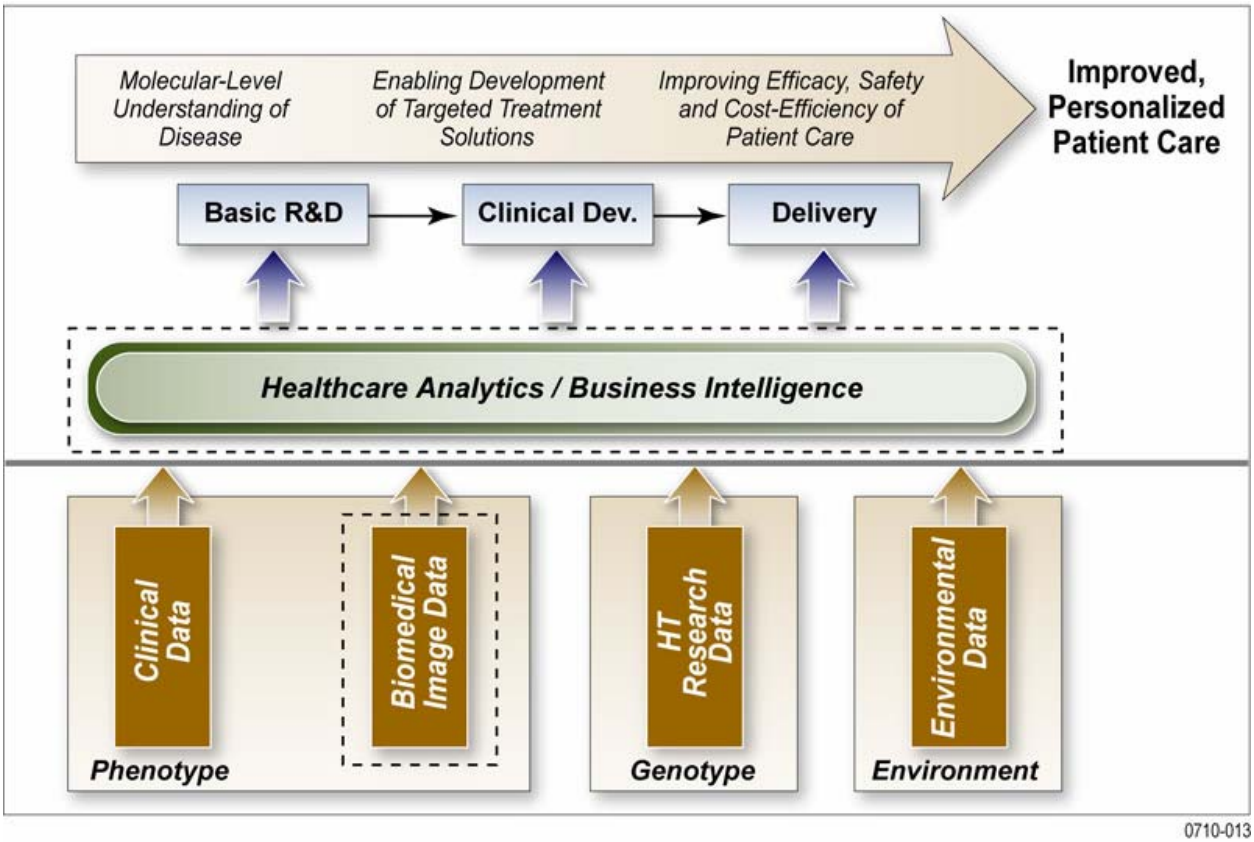

Figure 7: Conceptual View of Accumulating Data for Research

The strategies utilized to gather this data will vary based on the data required for the research, the potential sources of the data, and the format of the data. To assure a successful project, it is helpful to have a specific roadmap (illustrated in Figure 8 below) that involves all the key groups such as Operations, Information Technology, and Medical Informatics. This team needs to clearly understand what the organization is currently executing with regard to informatics and to gain a baseline of data. One methodology is to conduct interviews and/or create questionnaires to confirm the current environment, and to understand any unmet organizational needs. Once the correct participants have been identified, it is valuable to create a high level data flow of core systems and determine the current state of how data is aggregated for clinical and business intelligence. Then with the correct multi-discipline stakeholders in place (clinical research, clinical/business operations, IT) and the current state defined with key issues identified, themes of research opportunities can be considered with the appropriate value proposition. This should include clinical and business priorities with guardrails, assumptions and guiding principles.

Once the vision has been established, a governance structure should be created to guide the process and determine the decision-making criteria. Once the clinical and business needs are well understood, a new high-level future data model should be established and validated. The impact of the new data model should be reviewed and key decisions identified. Data ownership should be defined. Roles for data sharing should be confirmed. A detailed analysis will be required to identify the data required, and, of that, what data is currently available. Then it is time to determine the source and volumes of the data that is not currently available and where and how it can be collected. The relationship of this new data to the currently available data must also be defined, as well as how all the data elements will be accessed.

Actually acquiring the new data may require multiple processes. The simplest may be creating new processes that add the data to the EHR, assuming that the data is useful for purposes other than research. For example, interfaces can be implemented to pull data from medical devices. New screens can be created to collect new elements through direct data entry from EHR users.

Based on the flexibility of the EHR, new data models may be created within that system to house patient-provided data or data from a home health provider or a home health system. Since all of the required data usually cannot be housed in the EHR, a research database or data warehouse will be required with data integrated from multiple systems such as the EHR, billing systems, payer systems, monitors, external labs and others as indicated above.

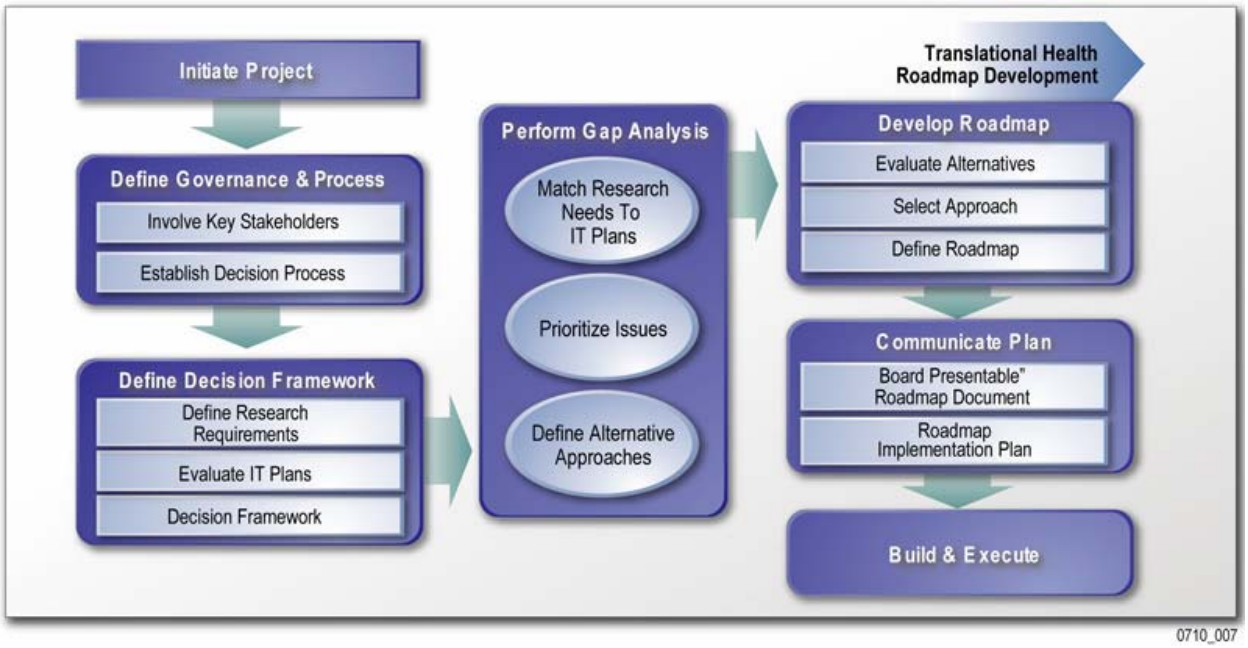

Figure 8: The Translational Health Roadmap Process

Regardless of the data needed for the research, and how it is acquired there are several key issues:

- All key stakeholders must be involved to truly comprehend the clinical/business need and the value to be obtained by the research.
- All key stakeholders must agree to syntactic and semantic data standards.
- Data elements required must be uniquely defined and identified and appropriately modeled with other data elements.
- Change management processes must be defined and understood and documented for adequate and accurate data collection.
- Interface requirements and resources cannot be underestimated or ill defined.
- A good data dictionary is essential for effectively searching a well-organized clinical data repository

14. Concepts or models on the potential use of clinical data and related resources for research applications

Models for efficient use of clinical data for research benefit from considering the larger and evolving context, which is the starting point of an evolutionary path from current solutions. The resulting models need to:

- Promote patient record databases in a structured form, which is suitable for data mining for rules and guidelines for physicians and researchers
- Advance both high dimensional data mining to discover more complex rules and mining of unstructured data such as medical images and text
- Develop centers for the required high performance computing, perhaps including sophisticated inference processes that draw conclusions from rules and ultimately drive workflows in healthcare and biomedical research

Until recently, clinical data was not readily accessible to promote biomedical research. The results of research had little impact on the patient except after a considerable time lag, typically years. Clinical practice and research thus stood clearly apart, despite the now preferred practice of Evidence Based Medicine (EBM) founded by Cochrane[1] and Sackett & Guyatt[2]. Although there have some criticisms of its previous implementations, few would criticize EBM's fundamental aspirations: to reduce medical errors by augmenting traditional physician practices based on subjectivity, hunches, personal experience, and anecdotal evidence, in favor of objective, conscientious, explicit, and judicious use of latest best evidence in clinical practice.

Evidence Based Medicine was not formulated as an Information Technology concept, but it does require IT. The implementation difficulties primarily relate to the difficulty of objectively and rapidly accessing and validating large amounts of the latest knowledge, which is widely dispersed in form and location. Information Technology and rapid methods of biomedical data capture now make possible the appropriate realization of these aspirations through a deep intertwining of clinical observations, clinical decision making, clinical research, and biomedical research in the overall dynamic of medical workflow. They enable a self-catalyzing cycle of medical knowledge that is rapidly accessible for clinical and research applications (Figure 9).

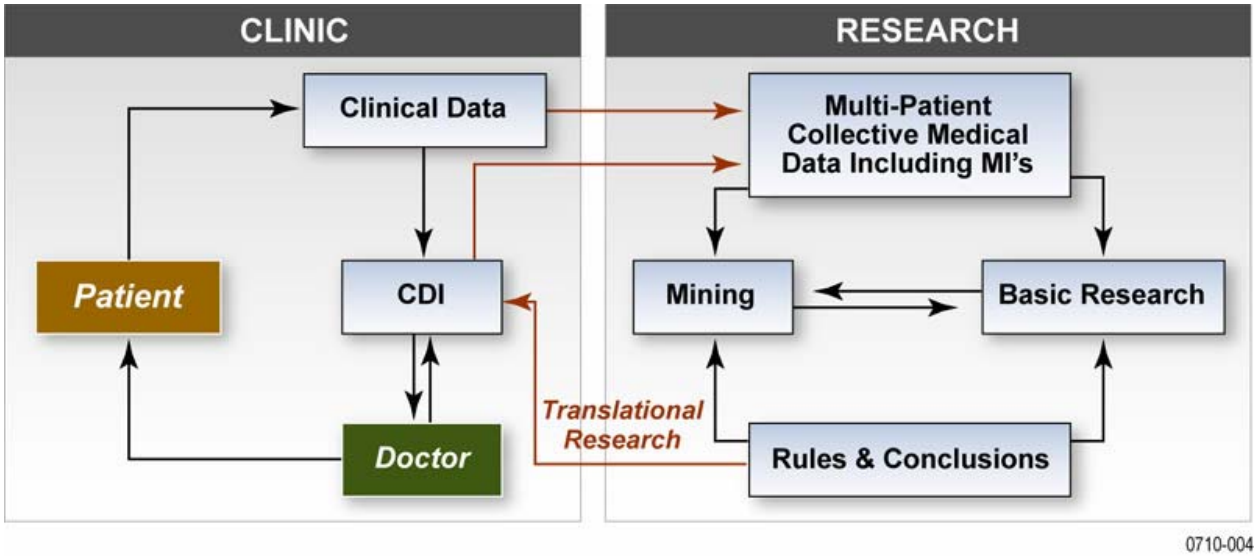

Figure 9: Flow of Information Between Clinical and Research Applications

Figure 9 illustrates how large archives of clinical data (digital patient records including imaging, genomic and other biomedical data) can be subjected to data analytics. Analytics include data mining, generation, exploration, validation, and curation of rules and guidelines through combinations of reference to medical literature analysis and basic research. One example of this type of analysis is a pioneering study of 667,000 patient records by IBM and medical collaborators[3]. This study used the output of statistically weighted rules as input to automatic searches of the medical literature. It was found that:

- 53% were rules already known to physicians (though conveniently rapidly generated and statistically qualified)
- 33% were less well known to the average physician
- 14% were seemingly novel but potentially relevant to diagnosis or therapy selection

Perhaps some 11% of the 14% are being rediscovered as credible in the literature of animal models and basic life science research, and merely require research for extension to healthcare. The remaining *circa* 3% appear wholly novel and require new epidemiological surveys or new experimental research programs. Hence the overall process is an effective means of rationally generating research proposals.

Many thousand rules relating 2, 3 or even 4 items (entries per patient) at a time may be readily generated. In addition, because of the high-dimensionality of the data[4], high performance computing is likely to uncover more complex rules (i.e., with more items). Such rules are important and may even dominate. For example, cardiovascular diseases and cancer may have rules with perhaps 20-70 items or more, which are not always readily reducible into independent or additive contributions from 2-4 item rules. Combinatorial analysis[5] shows that medical records with N entries per patient can imply up to X^N potential rules to be discovered by data analytics (X is 2 or more). With X=2 a mere 100 entries per patient could imply up to 10^{30} rules. Many may have too low a statistical weight to count as rules in the normal sense, but many of them may rest on enough sample data that they should be examined in order to know that they *can* be excluded. This means that high performance computing at selected centers may be critical for extracting maximal knowledge, and the amount of medical discovery may be directly linked

to the computing power implemented. With such potentially large amounts of data and computing capacity, storage and network bandwidth are also issues.

New rules stimulate research, and research not only validates, but returns new rules. Figure 9 also illustrates how *translational research* continuously augments a key element in the medical workflow – Clinical Decision Intelligence (*CDI*). The rules from research can be supplemented by additional rules and guidelines from respected medical information from experts, textbooks, and past literature. They can also be supplemented by rules related to governance, compliance, and, potentially fined grained, or conditional, patient consent. Combined with an inference and decision support process, such rules collectively represent CDI as an *inference engine,* which plays a central role in an ever-growing repository of medical knowledge. The inference process is important since rules – especially for the more complex disease states – rarely stand alone in their effect. For example, many weak and sometimes seemingly contradictory rules may combine to produce a strong weight of evidence. Inference represents advice to physicians to benefit specific patients. When new potential rules or conclusions are uncovered, it again stimulates validation and exploration by biomedical researchers.

Figure 10 illustrates that the above description so far is incomplete, as it considers only the use of *structured data.* Structured data is in well-formed spreadsheets, relational databases, HL7 medical documents, and so forth, which are readily accessible for data mining. However, in commerce in general, some 80% of data is *unstructured data*, such as images, graphs, sound recordings, and text. At present the percentage of data that is unstructured is probably significantly higher in healthcare. Examples of text to be mined are the medical and scientific literature, and hand written medical records. An example of an intermediate state is well-structured medical spreadsheets that include physician notes and surgical and autopsy reports as unstructured, textual blobs in the spreadsheet cells.

Clearly an important step is the conversion of all unstructured to structured data, and the industrialized nations are essentially at an early stage in this process. However, in Figure 10, note that an arrow goes not only from unstructured to simple structured data, but also to grammar. The tools for extracting meaning from language (i.e., basic syntactic and ultimately full grammatical analysis applied to text or speech) are being implemented and are in continuous development. Similar notions can be applied to understanding of images.

Extracted meaning is increasingly represented in the form of graph data structures relating concepts (analogous to towns on a map) by their relationships (analogous to roads between them). Such graphs are known as *Universal Knowledge Representations*, *semantic nets*, or *mind maps.* The graph structures represent the most sophisticated form of CDI as an inference engine. In contrast, inference using only rules derived from data mining of simple structured data usually considers only basic concepts and describes the relationships between them only as associations, or correlations in quantitative value. In the Universal Knowledge Representation, relationships, such as causation, influence, ownership (e.g., a nose is part of the face), position in ontology (hierarchic taxonomic classification), relative spatial positions, and probabilistic relationships provide a fuller description of extracted knowledge. Automatic inference using such structures can lead to insight or decisions by a physician or researcher, or automated switch points in an automated or semi-automated medical workflow.

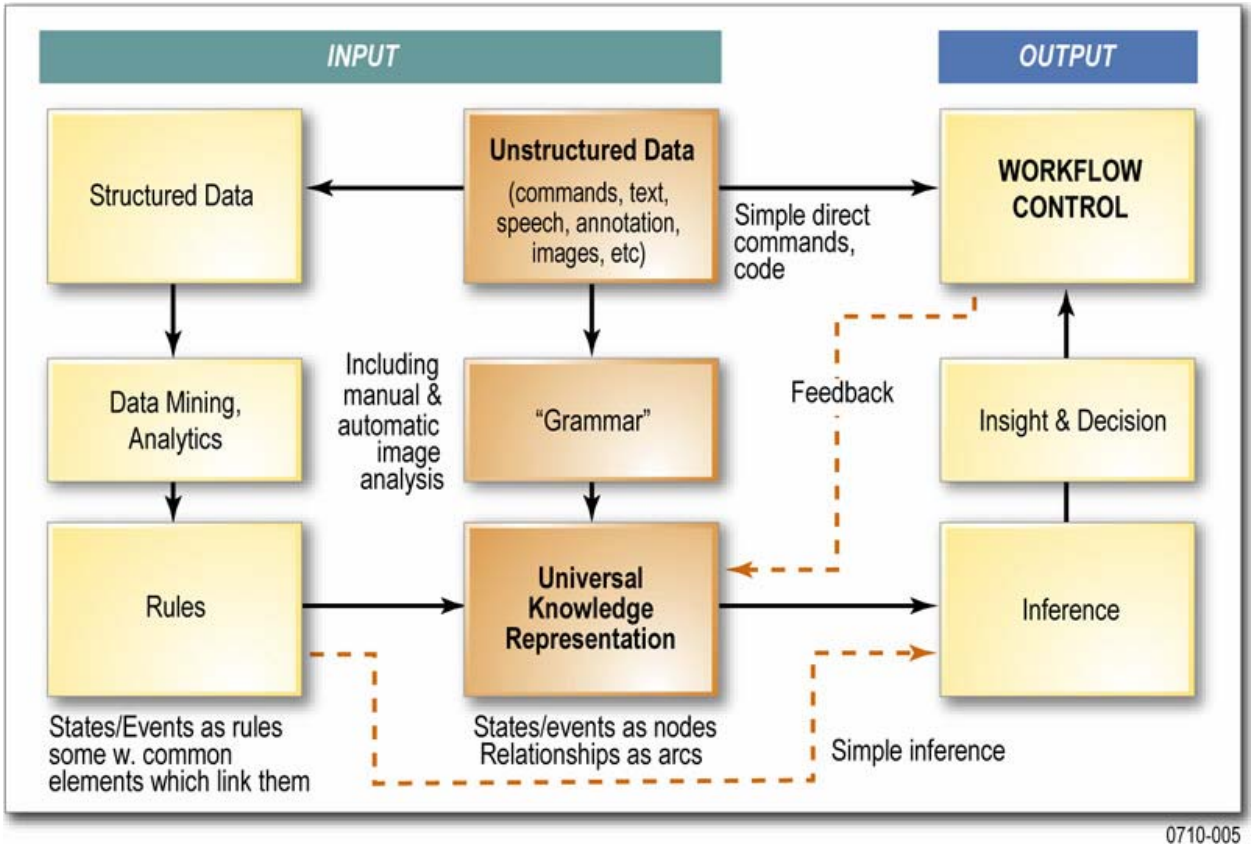

Figure 10: The Larger Future Technology Picture Including Unstructured Data

References

1. Cochrane AL (1972). "*Effectiveness and Efficiency: random reflections on health services.*" London: Nuffield Provincial Hospitals Trust.

2. Sackett, D., and G. H. Guyatt. (1992) "Evidence-Based Medicine: A New Approach to the Teaching of Medicine. *Journal of the American Medical Association* 268: 2420-2425.

3. Mullins, I.M, Siadaty, M.S., Lyman, J., Scully, K., Garrett, C.T, Miller, W. G., Muller, R. Robson, B., Apte, C. Weiss, S., Rigoutsos, I,. Platt, D., Cohen, S., and Knaus, W.A. "Data Mining and Clinical Data Repositories: Insights from a 667,000 Patient Data Set" (2006), Computers in Biology and Medicine, Dec;36(12): 1351-77. Epub 2005 Dec 22

4. "The Dragon on the Gold: Myths and Realities for Data Mining in Biotechnology using Digital and Molecular Libraries," B. Robson (2004) J. Proteome Res. (Am. Chem. Soc.) 3 (6), 1113-9

5. Robson, B. (2005) "Clinical and Pharmacogenomic Data Mining: 3. Zeta Theory As a General Tactic for Clinical Bioinformatics" B Robson J. Proteome Res.) 4(2); 445-455 (2005).

15. Models of cost-benefit analysis for integrated data systems, EHR, and clinical resources to inform medical decision-making

The working assumption for a healthcare ecosystem cost-benefit model is that since it is unlikely to be financed on an ongoing basis as a public utility, the integrated data systems, EHR, and clinical resources should be self-sustainable after the initial funding period. The business model must establish compelling value for each stakeholder, and incorporate mechanisms to evenly redistribute costs, risks, and expected returns, given the known value imbalance for providers. Development of the business model covers multiple segments of the health care system and identifies revenue streams from (and conversely, plan for corresponding offerings to) stakeholders such as payers, employers, Pharmacy Benefits Managers, pharmaceutical companies, and public health agencies.

IBM has a comprehensive business model, already used successfully in several healthcare market planning engagements. Figure 11 depicts the complex potential relationships between the stakeholders in a healthcare market as they try to achieve a sustainable financial model.

Sustainable Business Model Scenario for Interoperability in Healthcare

Beginning at the top left of Figure 11, many separate federal investments in health information technology will be coordinated to achieve more strategic and better scaled results. The other funding that can be considered a source of investment capital is any revolving or low-interest loans to fund overall market development or investment in their stakeholders' required infrastructure.

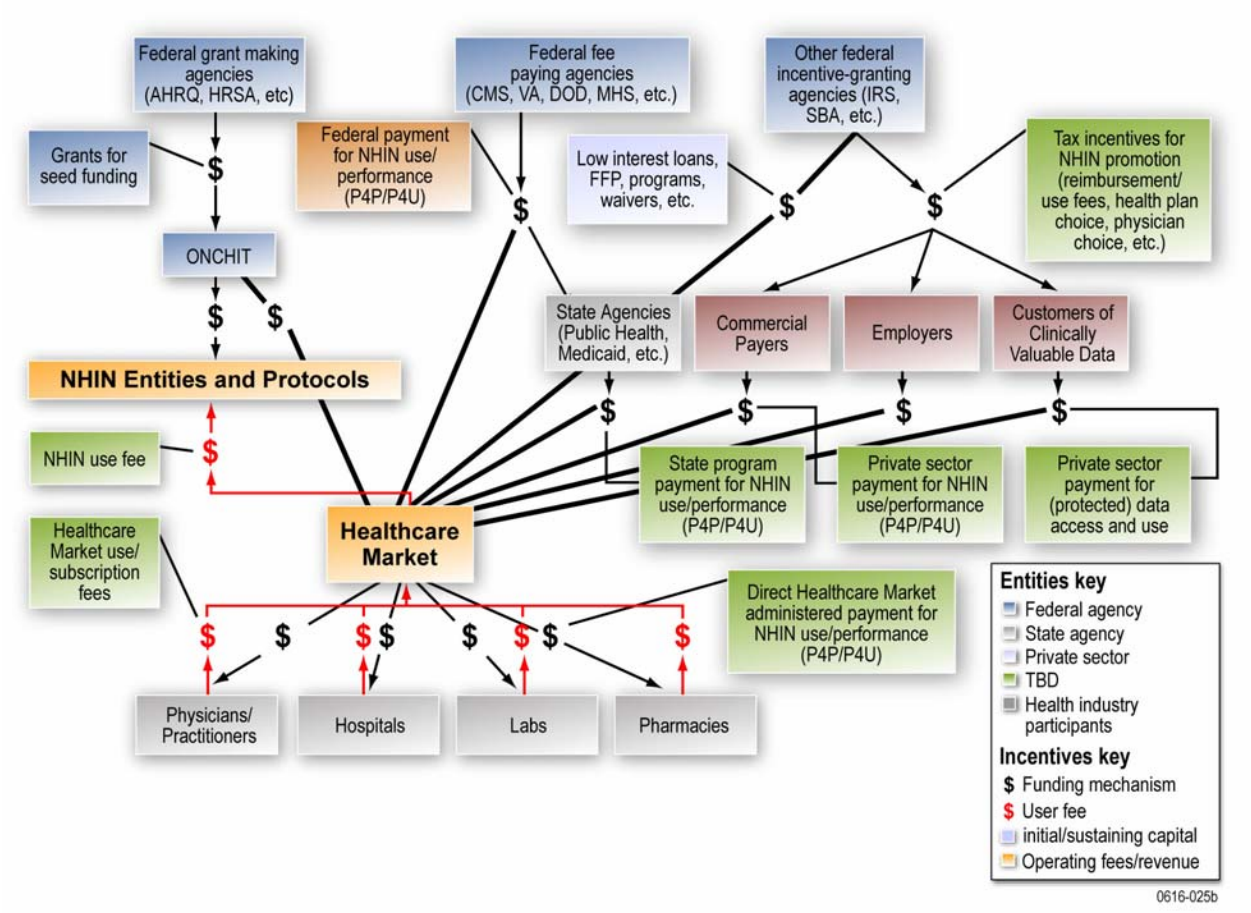

Figure 11: Potential Sustainable Business Model Scenario

Beyond initial capital, the sustaining Financing Plan is designed to take advantage of various financing mechanisms. First, federal purchasing agencies like CMS, Veterans Affairs, and others can develop pay-for-use (P4U) and pay-for-performance (P4P) for providers' adoption of EHRs and integrated data system technologies. With federal and state leadership (for Medicare, Medicaid, and other populations) to provide marketplaces with needed critical mass, the market will likely achieve a tipping point where private sector payers, employers, and others will also fund such programs (for P4U and P4P). To accelerate the process, the Federal Government, at a relatively low cost, can offer the private sector tax incentives for promoting integrated data systems and EHR creation and use. Similarly, payers could be directly involved in integrated data systems and EHR promotion activities, or they could contract with providers who are participating in healthcare markets. The last funding mechanism represented in the diagram is for direct payment by life sciences companies, payers, public health agencies, and others for clinically valuable data. Participation Protocols would spell out how and when release of such de-identified data is appropriate. Healthcare markets would negotiate the value of such data, and thus set the price for these customers.

While the funds flow described above explains the inflows to the healthcare markets, the bottom left of the diagram shows the outflow of fees from healthcare market subscribers to sustain the markets and network operations. We expect that subscribers will achieve varying levels of

return on investment for various services available from the Healthcare Market. Not all participants will achieve returns equally for all services.

This model depicts one possible cost and revenue model and is provided to demonstrate our understanding of the complexities of this task.

Cost and Revenue Model Approach

Our approach relies on four primary principles that we believe necessary for the cost-benefit analysis.

Sustainability: Our premise is that the healthcare markets must develop a financial model for health information exchange and quality improvement that is adopted by and paid for on an ongoing basis by its stakeholders. The model should not assume massive and ongoing federal investment. Our approach is designed to help stakeholders define a self-sustaining financial model, in part by leveraging the cost/revenue templates already developed for over 30 potential information exchange-based initiatives. To achieve this sustainability, our templates cover the entire healthcare ecosystem, i.e., identify revenue streams from (and, conversely, plan for corresponding offerings to) stakeholders such as providers, payers, employers, and public health agencies.

Adoption Rate: The value derived from an exchange depends on the volume of participating stakeholders, with a common issue that initial participants may realize little return until a critical mass of participants is reached and benefits can be realized. Our model takes into account various adoption rates to perform sensitivity analysis(1), and helps to identify the adoption timetable and potential incentives needed to neutralize this first mover disadvantage.

Redistribution and Incentives: In many cases, stakeholders other than the party making the investment are primary beneficiaries. Our business model is designed to first understand each stakeholder's raw return on investment (ROI) projections for each initiative supported by the integrated data systems and EHRs. It then incorporates what if programs to evenly redistribute expected costs and returns. These programs include a variety of pay-for-use and pay-for-performance incentives.

Competitive Landscape: Our experience is that a good ROI projection is critical but not sufficient. The primary participants of the exchange also want to be reassured that the Health Information Network will be competition-neutral – their relative competitive positions will not be affected by the exchange and/or the exchange will allow participants to compete on value-added services and products. Our business model is more than a cost and revenue model. It includes an assessment of each primary participant's pre- and post-exchange relative competitive position, and helps establish compelling value for each stakeholder.

As illustrated below in Figure 12, the business model is developed interactively with the governance, technical, and service models. These other models are produced as part of the deployment and operational planning tasks and provide the cost and revenue elements.

[1] Conversely, the model also helps assess the "network effect," when a high adoption rate generates an increased return to all stakeholders as more participants come forward.

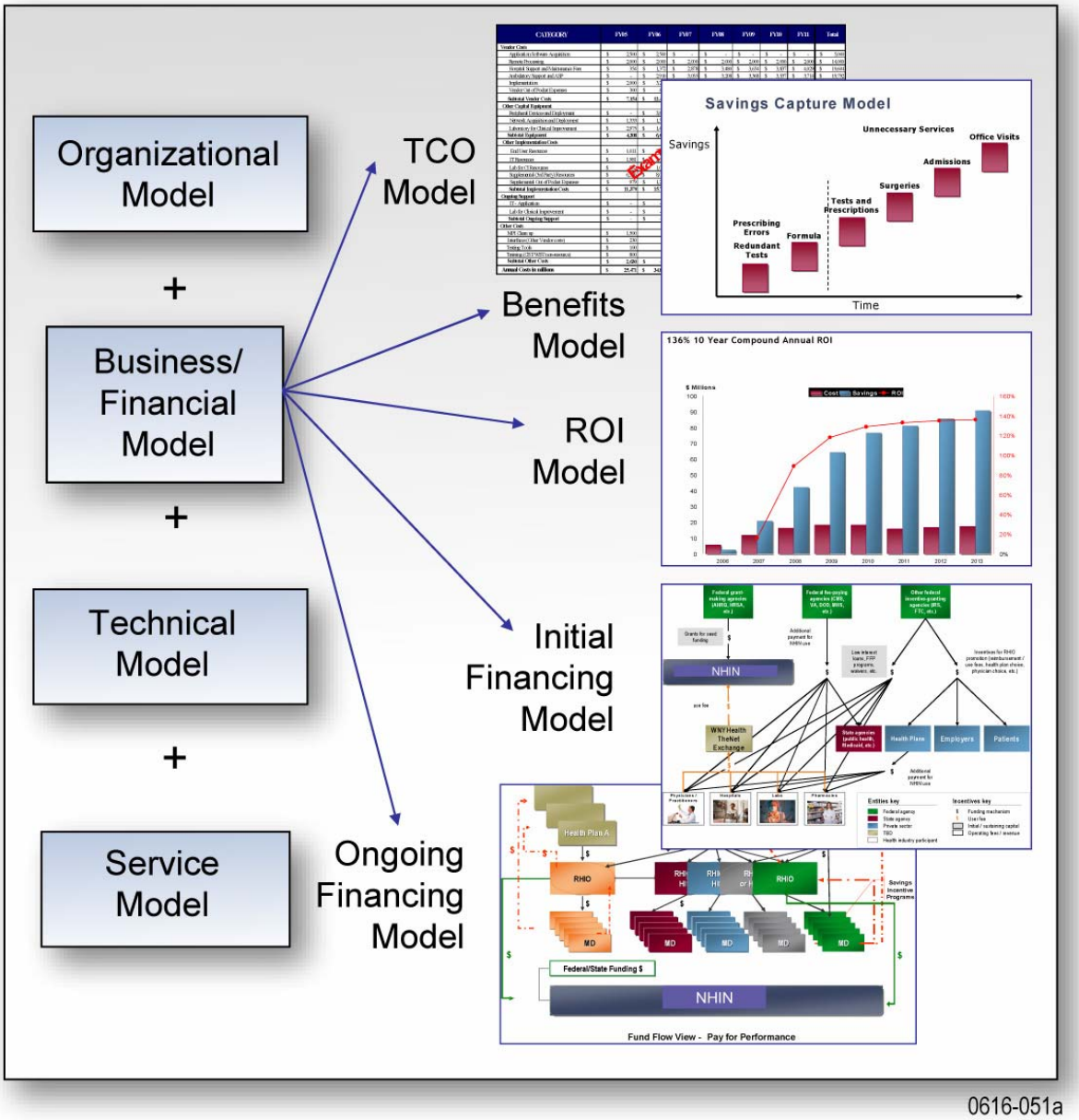

Figure 12: Healthcare Information Exchange Business Model

The cost and revenue model includes interrelated components that incorporate the cost-benefit analysis requirements for costs for rolling out, operating, maintaining, commercializing, and evolving a healthcare market, options for who bears the costs, and a revenue model that could work to achieve a self-sustaining market for widespread health data exchange and health application interoperability.

Total of Cost of Ownership (TCO) calculations. The model integrates both one-time and ongoing cost items, typically calculated over a seven-year period. We believe that our cost model is both realistic and comprehensive. For instance, it includes the costs related to providing the integrated data system and EHR services (patient identification services,

normalization services), as well as the costs related to change management, adoption policy, administering the incentive programs.

Benefit projections. The benefits projections are also calculated over a seven-year period. They leverage a large and continuously updated body of research as well as proprietary data derived from our own experience, both as a consultant and as a large employer. The model includes the possibility of assigning various confidence levels to each projected benefit, and various ramping-up periods (time projected for the benefits to be fully realized).

A comprehensive ROI model. The ROI model produces projections at the healthcare market level, but also per specific initiative, and per specific stakeholder. A key feature is the ability to model the impact of incentive programs to more evenly redistribute expected cost and benefits, both between stakeholders and over time periods.

Initial Funding. The model provides a distinctive view of the initial stages as stakeholders and revenue streams are typically very specific.

Ongoing Financing. This part covers the definition and formalization of the business arrangements that support the financial model envisioned, including the proposed incentive program(s).

Figure 13 summarizes the bottom-up and then top-down process that is followed in developing the financial model.

The bottom-up approach (Steps 1 and 2) is designed to collect the data that will feed the financial model as well as to assess each primary stakeholder's priorities and concerns. This reflects our field experience that, rather than rushing on engineering the financial model, time must first be set aside to understand the dynamics and potential roadblocks in a given ecosystem. This step, and the resulting data, will be performed in the deployment and operational planning phase. This includes information on the strategic and business operations as well as the governance, communications and process deployment activities that form the basis for understanding the business future. Operation and deployment planning activities provide value to each healthcare market by giving them the basis from which to drive business relationships, integrate processes, and align organizations. This may also form the basis for certifying the healthcare market as a participant in the Health Information Network. The work product from this activity will focus on information needed to develop the cost and revenue model.

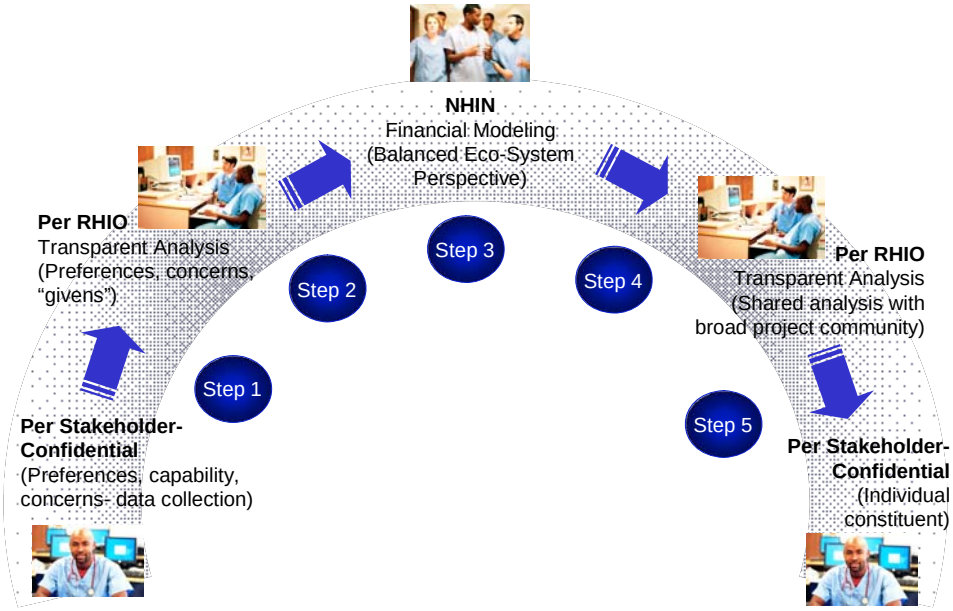

Figure 13: Financial Model Process

The business model is typically constructed in Step 3, and validated iteratively and interactively by the stakeholders – typically in a joint design session where the draft financial model is presented and multiple what-if scenarios are developed on the fly. One premise is that self-sustainability should be achieved on an ongoing basis at both the market and Health Information Network levels. Each health market should have a self-sustainable cost and revenue model for the community-based initiative(s) it has or will deploy. The Health Information Network should also be financially self-sustainable. This activity assesses the feasibility of each participating health market's financial model to test these models' underlying logic and the value proposition(s).

The top-down approach, Steps 4 and 5, is critical, and also stems from our direct field experience. Beyond the overall agreement typically achieved in Step 3, there is a need for developing and validating customized views of the exchange's business case with the various stakeholders' senior management, who may or may not have been directly involved with the planning effort. Our experience is that these primary stakeholders' individual presentations are critical for the project to really take off. Our financial model is designed to be easily tailored to produce these individualized views, from both a financial and a risk/relative competition perspective.

Evolution of integrated data systems and EHRs will be based on market forces, government regulations, investment strategies, consumer needs, and other ever-changing factors beyond the control or predictability of current efforts.

8. Feasibility and potential benefits for establishing linkages of institutional or organizational data resources with private and publicly available health databases

16. Opportunities and challenges for the development of electronic tools to aid in the integration and analysis of large datasets of clinical parameters to assist in outcomes evaluations

Vision, Opportunities and Potential Benefits

What is the vision for integrated, networked, collaborative information systems? What will be the opportunities and benefits of such system for patients, clinicians, hospitals, payers, and society?

We believe that, with proper protection of personal data security and patient privacy, the participants in the ecosystem of healthcare will all enjoy great benefits from integrating and analyzing aggregated private and public health information. New and more powerful electronic tools for data creation, capture, storage, sharing, integration and analysis can yield important new benefits in an era characterized by demands from patients, clinicians, payers, government and the public for safer and better care, delivered more efficiently and at lower overall expense.

Imagine patients with complete and accurate comprehensive medical histories that they can make available wherever they are getting care. With each care episode those records are updated automatically and ready for review by the next provider. Consumers and patients care about their healthcare quality and worry about the potential costs. They want to receive the best care from their doctors at reasonable costs. They want to be involved in their own care and make informed decisions for themselves.

Imagine the improvements in quality and reduced costs that networked databases that provide information on care costs and care service qualities, coupled with patient directed health insurance, may bring. Imagine patients having access to information that is specific to their physiologic makeup and their environment and how that might alter the position of preventative care.

For physicians and other care providers, imagine the reduction in risk and uncertainty that they will experience based on those complete records of patient care. As information about best practices continues to grow with ever more specificity based on molecular medicine and more targeted treatments, imagine the value of decision support systems that can bring to the care providers information on latest vetted best practices, access to information on other like patients including information on their treatments and outcomes. Imagine automated systems that can flag for care providers potential conflicts in treatment/therapeutics as new health information becomes available or can help cope with the combinatorial complexity of drug interactions, especially as more and more drugs come to market with very specific targeted application.

The new system will incorporate best practices and most-current medical knowledge, which will lead to better care that is supported by the most comprehensive medical and research evidence. Physicians and nurses will securely access the latest and complete digital data of the patients under their care. The power of the technology allows them to provide each patient a personalized care plan to treat and prevent clinical problems. Combining this knowledge base with patient medical history, current symptoms, and latest test results (including genetic and genomic tests) leads to much more personalized care. Benefits in increased patient safety, thoroughness of provided care and patient satisfaction are expected outcomes.

Imagine how much quality, effectiveness and efficiency will be improved when provider operations can be instrumented with information systems to measure and evaluate key performance indicators. Enterprise-wide integrated information coupled with best practices from the industry will both drive and enable unprecedented improvements in operations across healthcare systems. Greater efficiency leads to a competitive advantage in a healthcare market. Hospital organizations with care operations can cost hundreds of million each year to support. Many face hardship and challenges to achieve break even or profit. Thus efficiency in operational and financial control is essential for hospitals to survive and prosper. Part of gaining efficiency is to anticipate shifts in disease patterns and demands for certain services, deploy resources and staff to their fullest, prevent costly complications of care, and capture all billable activities. By tracking what is done during a patient stay and facilitating next steps faster, new and advanced electronic tools can eliminate much costly duplication of tests and delays in sequential treatment. These tools can also reduce the length of patient stays and free up limited capacity for other patients.

Imagine the impact on public health quality improvement and cost saving when each community is able to take advantage of the advanced electronic and computerized tools to scan entire populations for the most current disease-state statistics. They will give healthcare organizations better information about the demand for their service area as well as earlier warning signs about potential outbreaks. Early detection and awareness leads to containment and minimizing the health impact on populations, in addition to minimizing overall costs.

The rate of medical and public health research progress will be enhanced through integrated systems that speed identification and selection of cohorts for study or for trials. Modeling and simulation against richer and richer human molecular models and systems combined with access to published literature and patent databases will speed identification of promising therapeutics. Modeling and simulation will reduce lab failure and fallout rates. Sophisticated analytic tools will find and propose relationships that have always been there but that have not been visible for potential study leading to better care.

Advanced and intelligent IT systems can potentially reign in the growing societal healthcare costs via early detection of trends that raise costs and then facilitate sound execution of policies that hold costs down. Payers, such as the government programs and commercial insurers will certainly benefit. The integrated and analytical electronic tools will enable the identification of large-scale public health trends, detection of disease changes, and prediction of spending patterns. Health plans and payers with advanced data integration and analytical capability could practice evidence-based member management, improve compliance with therapy, develop customized benefit packages, and thus provide better services to their customers and help them stay healthier.

Challenges

There are challenges to building out these solutions. Some of them are technical. Some of them are financial. The most significant changes will be social and the culture within the workflows that make up health care delivery. All participants within the ecosystem will be affected.

A key technical challenge is standards. To achieve the value that comes from information-based medicine and the positively reinforcing network effect that members of the ecosystem anticipate, they will need to come together on standards for vocabularies, ontologies, encodings, and formats and protocols for integrating and sharing information. This will be an ongoing process as medicine marches forward discovering new disease types/subtypes, new lab tests, new therapeutics, new procedures and treatment regimes. There are a number of available

standards to begin with today, and early adopters' systems will provide important learning for the industry. Progress will require ongoing efforts to mature and expand standards with progress in medicine.

There are financial challenges as well. The vision calls for connecting and integrating information from all the participants in the ecosystems. Academic medical research centers and integrated delivery networks are leading the way with deployment of systems. The value these systems create for them justifies their leadership investment. However, the most value comes when all the participants are part of the information network. They will need help from partners and from technology providers driving adoption costs down to affordable levels. Complexity will need to be encapsulated and mitigated through integration. Healthcare costs will lead to reconfiguration of the industry and the changes in economics will align financial incentives through time as consumer influence and financial involvement increases and competition among providers increases.

The first step is adoption of electronic health records and clinical information systems that integrate IT into the clinical work flow enabling data capture. We anticipate a tipping point where providers will feel compelled to invest to stay part of the information based health care providing community. The second phase will be unlocking the value of the information in support of clinical process and workflow changes that improve provider quality and operations. This will require further integration of disparate systems using technologies such as metadata management, data warehousing and service oriented architecture to leverage existing systems and to combine them to create new solutions.

Additional significant technical challenges include providing data real time (e.g., imaging, video, digital microscopy). This includes IT support to manipulate, transfer and display in ways most helpful to physicians at the point of care. Automation of analysis will bring supercomputer and data analytics environments to payers and providers. Both data volumes and processing demands will lead to revisiting underlying IT infrastructures. As information technology becomes more important to success, the IT resources will need to be carefully managed for cost and flexibility.

Governance will be an important challenge. At the bottom will be governance of data and management of metadata for the integrated solutions. It will be critical to eliminate any configuration errors in decision support systems and to manage proper establishment, rollout and improvement (i.e., lifecycle management) for order sets, treatments, drug interactions, and rules regarding the care process.

Security and privacy issues must be managed. Society will be working through the tradeoffs between the value that comes from health information being shared and aggregated within the system and the issues surrounding privacy and uncertainties around misuse. Regulation and compliance standards will require monitoring to develop plans for meeting/complying.

Changes in care process and clinical workflow will lead to changes for the extended care team. Physicians may find change more difficult given the premium on their time and the changes in economics they are experiencing. Today we find available systems are sufficiently robust and refined that they can bring value, but they will lead to changes in how people work and provide care. Experienced consultants on governance, strategy, workflow and change management can help providers as they bring aggregated and best practice experience to physicians.

Feasibility

It is feasible to integrate information resources within health care delivery systems, between delivery systems, and further to integrate that information with other public and private health databases and create networked, interoperable solutions for information sharing across participants within the ecosystem.

IBM has a deep heritage in providing the tools and services to help clients capture, integrate, manage and manipulate data to transform it into information and knowledge and insight. These insights enable our customers to improve their operations and create new opportunities. IBM has a full line of information management products and deep services engagement experience to bring to the emerging information-based medicine future. These tools are industry hardened and proven dealing with the massive amounts of information in the media industry and the insurance industry, as well as addressing the very high demands of transaction systems and high performance computing. Years of satisfying customers' aggregate requirements from across industries positions IBM to take on the data management needs of the health care industry. Success in the marketplace confirms our readiness.

New information based medicine will drive needs for new more capable IT infrastructure. New molecular medicine, imaging systems and digital microscopy will bring unprecedented volumes of information to the healthcare provider. Decision support solutions, health analytic solutions and automated manipulation of unstructured information will bring unprecedented demand and dependency on IT systems for providers. IBM stands ready with highly available, scalable system architectures leveraging virtualization technologies and management tools to make them flexible and resilient. As research and practice further depend on information, modeling and simulation, flexible high performance computing needs will increase. IBM offers a full line of supercomputing capability with computing clusters, BlueGene and Cell processor technologies.

IBM is a leader in efforts to create interoperable health information networks. The IBM Interoperable Health Information Infrastructure (IHII) project is targeted at developing state-of-the-art technologies that enable the management and sharing of electronic healthcare records in a distributed heterogeneous environment. The IHII project, a joint collaboration of IBM Research and Development, is a foundation for the solution IBM was selected to build for the ONCHIT Nationwide Health Information Network architecture prototype. As a result, it has created a rich set of building blocks for assembling information networks within complex Integrated Delivery Networks, between providers within RHIOs and as a base for national health information networks.

For the past 18 years IBM has been a leader in addressing industry's desire to integrate information across communities through standards development and creation of industry data models. This work always begins with establishing conventions around common definitions and vocabulary within the community. Then, working with customers, it moves through a series of modeling activities. IBM led the way in the insurance industry and has done similar work leveraging its expertise in finance, retail, telecommunications and other industries. In the healthcare industry, IBM and Mayo Clinic have integrated over 4.4 million patient records that were in non-integrated formats, into a unified system based on a standard technology platform that incorporates robust security and privacy features. This allows physicians and researchers access to a comprehensive set of records that can be analyzed with the security and privacy needed to protect patient confidentiality. IBM is also building a clinical decision support system with Geisinger Healthcare System. We are using this work with Mayo, Geisinger, and other customers to develop data models for healthcare enterprises. IBM is building enterprise models

for both the provider and the payer within a single common framework to further enable communication across this important boundary.

IBM has deep clinical expertise. Hundreds of clinicians joined the IBM consulting team with the acquisition of Healthlink, which has been top-ranked by the KLAS clinical IT implementation study. These doctors and nurses share an understanding of the issues that healthcare providers are facing, and they are able to design information integration solutions that can be developed to ensure real priority problems are solved. Further IBM has developed rich healthcare domain experience within its business intelligence and data warehousing practices. These practices can enable providers with integrated data systems with a single source of truth transcending the various application systems that grew up serving specific departments or functions and/or are different from one another due to mergers and acquisitions. Additionally IBM has technologies for integrating information from many heterogeneous sources either within or beyond the walls of the provider institution. These warehouse solutions are base enablers for the health analytic solutions that can lead to provider improvements in cost and quality.

In summary, collectively IBM has the healthcare domain and technology expertise, health care customer experience in building data integration solutions and health information networks, and the products and services to provide the infrastructure to satisfy the demands today and well into the future. With IBM, integrated and linked information management solutions connecting and relating information inside and outside the healthcare organization are indeed feasible.