# Bioinformatics Linkage of Heterogeneous Clinical and Genomic Information in Support of Personalized Medicine

**L. J. Frey[1], V. Maojo[2], J. A. Mitchell[1]**
[1]Department of Biomedical Informatics, University of Utah, Salt Lake City, USA
[2]Biomedical Informatics Group, Universidad Politecnica de Madrid, Spain

## Summary

***Objectives:*** Biomedical Informatics as a whole faces a difficult epistemological task, since there is no foundation to explain the complexities of modeling clinical medicine and the many relationships between genotype, phenotype, and environment. This paper discusses current efforts to investigate such relationships, intended to lead to better diagnostic and therapeutic procedures and the development of treatments that could make personalized medicine a reality.

***Methods:*** To achieve this goal there are a number of issues to overcome. Primary are the rapidly growing numbers of heterogeneous data sources which must be integrated to support personalized medicine. Solutions involving the use of domain driven information models of heterogeneous data sources are described in conjunction with controlled ontologies and terminologies. A number of such applications are discussed.

***Results:*** Researchers have realized that many dimensions of biology and medicine aim to understand and model the informational mechanisms that support more precise clinical diagnostic, prognostic and therapeutic procedures. As long as data grows exponentially, novel Biomedical Informatics approaches and tools are needed to manage the data. Although researchers are typically able to manage this information within specific, usually narrow contexts of clinical investigation, novel approaches for both training and clinical usage must be developed.

***Conclusion:*** After some preliminary overoptimistic expectations, it seems clear now that genetics alone cannot transform medicine. In order to achieve this, heterogeneous clinical and genomic data source must be integrated in scientifically meaningful and productive systems. This will include hypothesis-driven scientific research systems along with well understood information systems to support such research. These in turn will enable the faster advancement of personalized medicine.

## Keywords

Personalized medicine , genomic information models

**Geissbuhler A, Haux R, Kulikowski C, editors. IMIA Yearbook of Medical Informatics 2007. Methods Inf Med 2007; 46 Suppl 1: 98-105**

## Introduction

Crick's dogma of biology, DNA to RNA to Proteins [1], whatever recent caveats reveal, did provide researchers over almost four decades with an information processing model which led to numerous breakthroughs and helped facilitate the rapid completion of the Human Genome Project. Biomedical Informatics as a whole faces a much more difficult epistemological task, since there is no similar foundation to explain the complexities of modeling clinical medicine and the many relationships between genotype, phenotype, and environment that might help. This paper discusses current efforts to investigate such relationships, intended to lead to better diagnostic and therapeutic procedures and the development of treatments that could make personalized medicine a reality.

To achieve this goal there are a number of issues to overcome. Primary are the rapidly growing numbers of heterogeneous data sources which must be integrated to support personalized medicine. Solutions involving the use of domain driven information models of heterogeneous data sources are described in conjunction with controlled ontologies and terminologies. A number of such applications are discussed that are currently being developed and used in Europe and the United States. These include genotype and phenotype information models, clinical and genetic concept linkages, grid technology for sharing and analyzing data, and combined approaches for biology and medical systems. A combination of this data-driven and other, more classical, hypothesis-driven approaches, should provide a stronger basis for scientific research in these recent efforts of personalized medicine [2].

Over the last few years, researchers have realized that many dimensions of biology and medicine aim to understand and model the informational mechanisms that support more precise clinical diagnostic, prognostic and therapeutic procedures. As long as data grows exponentially, novel Biomedical Informatics approaches and tools are needed to manage the data. Although researchers are typically able to manage this information within specific, usually narrow contexts of clinical investigation, novel approaches for both training and clinical usage must be developed. Most clinicians are not familiar with this kind of information; how to use it can be controversial, so informatics tools must be developed to help in this process. If not, it is likely that the inclusion of genomic information in clinical practice will be delayed for years given the reluctance of physicians to use complex, unclearly and incompletely tested, and partly connected biological information.

Additionally, the large and increasing number of "omics" disciplines reflects

the proliferation of informatics-based biological disciplines during the last decade in the biomedical area. While it is a good indicator of growth and may attract many students and researchers to the field, the many disconnected subdisciplines also reflect the lack of basic unifying informational principles and focus, which is similar to what is seen in biomedical informatics as a whole.

## Personalized Medicine

After the Human Genome was completed and other "omics" technologies began to proliferate and have an impact, biomedical scientists and practitioners proposed to use the new wealth of biological information, such as SNPs, microarray data, biomarkers, with their ever-increasing data sets into clinical practice. Researchers have expressed the hope that new findings could have immediate clinical application, leading to new visions of medicine, such as "genomic medicine" or "personalized medicine" [3,4,5,6,7,8]. Personalized medicine aims to adapt these mechanisms to individual patients [9]. yet, each patient will not be treated differently from every other patient [10] but instead divided or stratified into groups by genetic and other markers that predict disease progression and treatment outcome. So, personalized medicine, based on genetic factors or markers and even specific pharmacogenomic drug effects [11], cannot be a magical solution for every individual. Instead, drugs will be tailored for groups of people with similar or related genetic characteristics.

In this scenario of personalized medicine, pharmacogenetics/genomics is at the center of research and practice [10,12,13,14]. Pharmacogenetics explains the different responses of individuals to the same drugs. To validate preliminary findings and models, experiments (e.g., clinical trials) must be carried out with large number of patients [9]. Therefore, classical clinical studies must be redesigned to adapt to these new situations.

Weatherall [15] has emphasized the different steps to be taken with caution before personalized medicine can be applied in clinical medicine. Personalized medicine can prove that diseases such as cancer or diabetes type II, formerly classed as a unique and isolated category, can be reconsidered and reclassified, proving that they are due to different causes. This reclassification leads to new diagnostic and therapeutic procedures. Once this process is established, researchers must show that it is cost-effective to test specific markers and use specific drugs. Otherwise, health systems may not afford the enormous demands of this new shift in medicine, particularly if there is not enough evidence of their significant impact in changing medical outcomes in a large enough part of the population.

## Heterogeneous Data Sources

Over the last decade a whole area of "data integration" has evolved and expanded. It includes approaches such as data warehousing, information linkage, data translation and query translation [16] as well as techniques such as ontologies [17] to enable their standardized knowledge definition. We have earlier proposed that genomic data could be integrated into health information systems [18,19,20] but recent research [2,21,22,23,24] suggests that this bridging process will not be easy.

At the time of writing this paper, around 900 biological public databases (e.g., genomic, proteomic, metabolomic, and others) are available to researchers and other professionals. These databases have been designed and maintained as result of many biological research projects that have produced a huge amount of heterogeneous information about genes, proteins and genetic diseases. Public databases usually include different data, ranging from biochemical to public health data. A larger number of organizations maintain their own databases, restricted to public access for different reasons (e.g., socioeconomic, confidentiality, strategic), often focused in one specific area or topic.

Within biomedical databases information is often inconsistent or missing. In systems biology, for instance, we find problems related to functional annotations of genes and proteins, genotype-phenotype relationships, kinetic values for enzymes or components of pathways [12]. In clinical systems, spanning patient information over decades, the inconsistencies can be even higher. Analysing old medical records in paper, in settings where computerized medical records were not available until recently, the rate of missing or inconsistent data can be quite variable and often very high [25].

From a scientific perspective, such an approach increases the problems of data integration and analysis, due to the frequent variability across different settings regarding experiments, techniques, procedures, theoretical approaches, cognitive biases, among others [2,26,27]. Given this lack of consistency, using data obtained from heterogeneous sources to advance scientific research presents different problems, especially in biomedicine. In the "omics" areas, there has been a predominant data-driven research approach.

Multiple sources of genomic-scale data must be integrated to develop more precise descriptions of clinical phenotypes. For instance, gene expression data reflects the effect of oncogenes on

metabolic pathways leading to oncological disease. We know now that cancer is not a precise and unique disease, but a number of pathologies, with different causes and therapies. In fact, the differences in all these databases, considering hardware, software applications (e.g., operating systems, database management systems), semantics as well as the differences mentioned above in scientific approaches, cultural environments, cognitive biases and others, (often unclearly and even subtly hidden in the stored data) must be solved if the researchers want to really integrate information and extract useful patterns. Frequently, data must be normalized [2] and some common data models and coding systems must be used or developed to standardize the representation of genotype-phenotype information.

## Information Modeling

In this effort to enhance information storage and data exchange, bioinformaticians can draw on work that the Object Management Group (OMG http://www.omg.org) has done in the development of the Model Driven Architecture approach. This is an approach that represents systems with a graphical object models. They propose the development of platform independent models using the Unified Modeling Language (UML). For modeling complex systems that combine clinical and genomic data, a special kind of UML model, called a domain model, should be used [28]. Such a domain model incorporates the scientific domain knowledge that is necessary in using clinical and genomic data for personalized medicine. The models can be used to communicate information about the system to both developers and users of the system. For example when creating a system about transcription and translation, the objects of DNA, RNA and Proteins should be modeled along with their associations.

Models that use domain information to represent knowledge about the data are needed in order for the field to create systems that have better information representations and data exchange. Bioinformaticians are in a position to communicate the underlying body of knowledge to developers in order to create such domain models. In bridging the technology and scientific communities they can help in the creation of domain models that communicate semantically meaningful information about the data. Semantically meaningful information models allow developers to construct objects that have meaningful counter parts in use in their area of application. Since many systems are being developed independently, if developers are enabled to represent objects in a meaningful way, then the chances of having reusable objects or at least objects that are easier to map between systems is increased. Reuse of objects between systems will help improve information storage, data exchange and the development of interoperable systems.

## Ontologies and Terminologies

An interesting effort to normalize and reuse vocabularies and knowledge over different projects and groups is related to ontological development. Ontologies provide the semantics needed to bridge the existing gaps between heterogeneous data sources and a formal language for information retrieval.

The underlying vocabularies that are currently being used to support Model Driven Architecture (MDA) development at the National Cancer Institute (NCI) are those provided by the Enterprise Vocabulary Services (EVS). They support a number of terminologies needed by the NCI. The two products being used in MDA at NCI are the NCI Thesaurus [29,30,31,32] and the NCI Metathesaurus [33]. The former is a reference terminology that has a vocabulary for use in clinical care, translational research and basic research. It contains information on 10,000 types of cancer and related diseases and 8,000 therapies. The NCI Metathesaurus is a mapping between multiple terminologies. It includes a specialized version of the UMLS [34,35]. This version is specialized in order to focus on terminologies that can be related to cancer terminologies [28]. Some of the terminologies that the Metathesaurus includes are LOINC [36], SNOMED/CT [37], Veterans Health Administration National Drug File (VA NDF), Gene Ontology (GO) [38] and MGED Ontology [39]. The NCI Metathesaurus contains 1,200,000 concepts mapped to 2,900,000 terms with 5,000,000 relationships. Mapping these terminologies supports the goal of representing and combining clinical and genomic information. These terminologies help the developers of domain UML models by giving them a broad range of terminologies. In this way the projects are combining bioinformatics and clinical informatics concepts in data models that support interoperable systems for the field. This is a key component for building systems that support the development of personalized medicine.

In the medical domain, vocabulary and coding systems such as the ICD 9 and 10, SNOMED, LOINC, MeSH, UMLS, ICNP, GALEN, the NCI thesaurus, and others are now used for ontology-related tasks. Although it can be assumed that they are not “true” ontologies from a formal computing perspective, they have been used in a number of specific applications (e.g., the UMLS) [17]. The Foundational Model of Anatomy

(FMA) is an extended framework for representing classes and relationships describing anatomical structures in a format that is understandable to humans and also navigable by computers. Specifically, the FMA is a domain ontology that represents a coherent body of explicit declarative knowledge about human anatomy.

In genetics, Gene Ontology (GO) has become a great success, leading ontological development in the genetic area. It is a collaborative effort among different organizations and professionals to create a controlled vocabulary of gene and protein roles in cells, to consistently describe gene products in different databases. GO includes three structured, controlled vocabularies (or "light" ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions.

There are no current world-wide standards that represent genotype-phenotype data with specific data models that can be used to enhance information storage and exchange. Likewise, different coding and vocabulary systems have been used in clinical medical domains for electronic health records, HIS, epidemiological surveillance systems, and so on, but only recently have some proposals been made to link clinical and genetic concepts. The UMLS has recently included genetics terms [40] and Gene Ontology [41] within its medical vocabularies and nomenclatures. These efforts have been proposed and investigated by Medical Informatics professionals over the past decade, more recently attracting the attention of bioinformaticians. In a semantic mediation system, the users (humans or machines) do not care about the specific format of the information source, but only about the terms contained in the ontology for building the query in a consistent way that will yield valid computational results. In this sense, ontologies can be understood and used by both humans and machines.

## Applications

In a research environment where thousands of devices, databases, Web-based documents and other sources are used in research, effective software tools from clinico-genomics, semantic mediation and novel computing techniques for such computationally demanding tasks will be needed. For instance, "data grids" can be used in the short term to enhance access to computationally demanding clinico-"omics" applications from remote sites [42, 43]. Bioinformatics web and grid services can be orchestrated to organize intelligent work flows of different applications, organized by using program managers and semantic mediators. Semantic mediation will be needed to intelligently organize such "choreography" [42,44].

ACGT (Advancing Clinico-Genomics Trials on Cancer) [45] is an European Commission supported integrated project funded to design new methods and resources for cancer research. Twenty five partners from Europe and Japan participate in ACGT. The goal of the project is to identify technological gaps and barriers in cancer research and create novel techniques for diagnosis and prevention, as well as to design new models of clinico-genomic trials that will facilitate the creation of new drugs and therapies in the context of personalized medicine.

From a biomedical informatics perspective, ACGT aims to develop a Biomedical Grid infrastructure at a European level to conduct research on two different kinds of cancer: pediatric nephroblastoma and breast cancer. In this Grid-based scenario, research on systems interoperability (based on the development of the "ACGT master ontology for nephroblastoma" and a semantic mediator to organize the choreography of different web and Grid services), in-silico simulation of drug design, data mining, and clinico-genomic information modeling and management are being developed to enable novel approaches to clinico-genomic trials.

In this framework, heterogeneous data from three on-going clinical trials are linked with "omic" information into a common virtual repository, by using an ontology-based approach. Such a repository includes different types of clinical and genomic information, such as including numerical data, text and images from patients participating in the trials, external public databases and in-silico simulation. This project, ending in 2010, aims to solve some of the problems that arise in this kind of research. For instance, to build an efficient biomedical Grid infrastructure or carry out in-silico simulations to design and test new drugs in the context of personalized medicine.

In the United States, the National Cancer Institute has taken up the challenge of combining genomic and clinical data for cancer research and treatment. The approach has been one of building a well specified infrastructure to support the development of interoperable tools built upon that infrastructure. The effort has been undertaken by the Cancer Biomedical Informatics Grid (caBIG™) community. Their approach based on binding controlled terminology is described below. Additionally some of the tooling developed for caBIG is described.

The UML models in caBIG™ contain class diagrams that represent the scientifically relevant objects that are part of running software systems. These are objects like DNA sequence, RNA se-

quence and protein. The classes are composed of the class name itself (e.g., Protein) and the attributes that are in the class (e.g., uniProtKB, name, symbol, etc.) Between the classes there are association links that convey the relationships between the classes. A description is required for each class and attribute and is stored in the UML model.

Using UML domain models to instantiate the MDA framework gives a conceptual representation of the underlying scientific objects via the classes, attributes, associations and descriptions. This model does not provide an unambiguous representation because developers at different sites, while using the same scientific domain knowledge, can create similar classes and attributes, but with different names, configurations and descriptions. The binding of controlled terminology to the model can mitigate the latter problem of different semantics associated with the scientific concepts in the model. The binding does this by unambiguously specifying the description in the model with controlled terminology.

The binding is accomplished by mapping concepts from the EVS NCI Thesaurus and Metathesaurus to the classes and attributes in the UML model. These are bound using the Concept Codes or Concept Unique Identifiers (CUIs) which are maintained by EVS. The descriptions for the classes and attributes are used to determine what CUIs from the EVS terminology should be used for the mapping. This mapping is then used in the creation of the metadata for the data elements that represent the model. This generation of metadata is supported by a suite of tools and infrastructure developed at the NCI. The NCI has implemented the ISO11179 standard for representing metadata in their Cancer Data Standards Repository (caDSR). This incorporates the data structures and format of ISO11179 to store a data element (DE) using a combination of a data element concept (DEC) and a value domain (VD). The DEC maintains the set of CUIs associated with the DE and the VD specifies the data type and permissible values associated with the DE. Since ISO11179 does not have associations in its representation, the caDSR extends upon the standard to store the metadata for the associations between the UML classes. This is in order to more fully support the MDA approach and incorporate the associations that are in the UML information models [28].

The potential benefit of binding UML models to terminology is the ability to more readily support reuse of model elements. Once the attributes are mapped to the terminology, it has a defining series of concept codes. This along with the value domain specifies the data element. Hence, if another model has an attribute mapped to the same series of concept codes and value domain, then the two models share an identical attribute. This is regardless of differences in the naming conventions in the two models and any differences in the descriptions in the two models. The scientific meaning is encapsulated in the series of concept codes and the value domain which is identical for both models. The point of this system is to help developers and users reach consensus and converge to common models for their systems. In this way tooling can be built in an interoperability fashion.

Three tissue banking and pathology tools have been developed using the caBIG™ infrastructure to coordinate the underlying information models. These are caTISSUE Core, caTISSUE Clinical Annotation Engine (CAE) and the cancer Text Information Extraction System (caTIES). The caTISSUE Core system is used to inventory and track biospecimens. This includes searching for specimens and requesting them for studies. The CAE annotation system is based on the College of American Pathologist (CAP) cancer protocols [46]. It has the functionality to import data from anatomic pathology laboratory systems, cancer tumor registries and clinical pathology laboratory systems. The information is clinically oriented and is tightly tied in with the caDSR, enabling the definitions from the EVS terminology to be displayed for field titles in the interface. The tool caTIES extracts structured text from free text surgical pathology reports and encodes it in caBIG™ compliant terminology. This enables researchers to search for annotated tissue over structured terminology instead of free text in order to obtain relevant biospecimens.

The Cancer Translational Research Informatics Platform (caTRIP) system (https://cabig.nci.nih.gov/tools/caTRIP) utilizes the EVS terminology and metadata in caDSR to run distributed queries on federated data resources. This includes the caTISSUE tools that have their metadata registered in the caDSR. This is an example of the UML models bound to controlled terminology enabling systems to interoperate more effectively.

The caIntegrator system (https://cabig.nci.nih.gov/tools/caIntegrator) combines a variety of biomedical data related to clinical trials together with bioinformatics experimental data. These latter data types include Immunohistochemistry (IHC), microarray-based gene expression and SNPs. The tools support the analysis of these data in an integrated system.

Some recent proposals aim to link genotype and phenotype information. One example is the Polymorphism Markup Language (PML), to represent and store SNP (single nucleotide polymorphism) information [41]. This project,

launched by a broad international consortium, aims to model the variation of genetic information, including a whole range of mutations.

Following a different direction, the IBM Haifa Group has led the HL7 Clinical Genomics special interest group (SIG) to create standards for exchanging clinical and genomic data [47]. By using genomic data in health care to support personalized medicine, or integrating genomic data into classical electronic health records, linked to emerging bioinformatics formats such as MAGE-ML for gene expression or BSML for sequencing data. Its genotype model includes various types of genomic data such as sequencing, expression and proteomics data. Developers have tested this model on cystic-fibrosis data, bone-marrow transplantation and pharmacogenetics projects [47].

From a European perspective, the European Commission has launched several initiatives since 2001 in the specific contributions that biomedical informatics can make to personalized medicine. A preliminary conference, called "Synergy between Research in Medical Informatics, Bioinformatics and Neuroinformatics. Knowledge Empowering Individualized Healthcare and Well-Being" was held in Brussels in 2001. In June 2006, another meeting, "ICT for BIO Medical Sciences 2006", analyzed the results obtained in these five years. In this time various conferences, projects, and different activities were carried out. The BIOINFOMED study [18] was delivered in 2002, defining various significant challenges in biomedical informatics at a European level. These challenges were related to linking clinical and genomic information for biomedical research and practice, in issues such as biobanking, genomic-based computerized medical records, pharma-informatics, integrated ontologies or integrated access to clinical and genomic databases. In summary, the goal was to introduce a strong scientific biological basis to clinical medicine that could lead to better diagnostic and therapeutic procedures.

This initiative started in 2004 to promote biomedical informatics in Europe in support of personalized healthcare. The "Network of Excellence" (NoE) [48] was created to launch different directions and ideas, aiming to establish a common meeting place within the European Union in the biomedical informatics area. In this NoE, several work packages were envisioned to deal with dissemination, training and mobility, data modeling, integration and mining as well as four clinical pilots, all of them linking different types of biomedical information. These pilots were designed to create new biomedical informatics approaches in (1) pharmacoinformatics, (2) genomics and infectious diseases, (3) periodontitis, and (4) genetics and colon cancer.

These four pilots are examples of the different approaches and problems that biomedical informatics can face in the new environment of genomic medicine, previously described [49]. For instance, the periodontitis pilot project, led at the VU University Medical Center in Amsterdam, aimed to develop informatics methods to store and analyse clinical and genomic information. Periodontitis is an example of chronic infectious and inflammatory disease caused by multiple factors (genetics, infectious, geographical and environmental) that affect the teeth-supporting tissues. It affects more than 10% of the adult population and nearly 30% of elderly people, increasing the risk of cardiovascular diseases in this group of people. Periodontitis was selected since it seems to have a small number of triggering factors, facilitating clinico-genomic research. A database has been built including clinico-genomic information. PML and other models are being used for representing genotype-phenotype links. In such datasets, data mining methods are being used to discover links between clinical information and genetic traits.

Recently, an initiative called the i2b2 (Informatics for Integrating Biology and the Bedside), an NIH-funded National Center for Biomedical Computing, is "developing a scalable informatics framework that will bridge clinical research data and the vast data banks arising from basic science research in order to better understand the genetic bases of complex diseases". The objective of this center is the research of new diagnostic and therapeutic approaches in the framework of personalized medicine, in areas such as diabetes, hypertension, or Huntington's disease [53].

Researchers in Iceland are gathering genetic data from a genetically isolated population. The hypothesis behind is that such historically restricted group of people will show significant relationships between genetic and clinical data in some specific genetic diseases [54]. Similarly, the Mayo Clinic/IBM Computational Biology Collaboration, led by Prof. De Groen, is developing a comprehensive computerized system for access to and interpretation of clinical, genomic and proteomic data. Data from over four million patients are stored to link genomic and phenotypic information. Mining such large data sources might not increase the feasibility of extracting breakthrough knowledge, as compared to smaller datasets, but they include a large range of data where new hypothesis can be elaborated and tested —particularly linking genomic and clinical traits [55].

Several health sciences centers and individuals have begun to investigate how personalized medicine might be actu-

alized from an informatics standpoint. These include Harvard Partners Health Care at Harvard University [24,56], the Mayo Clinic in Minnesota [57], Duke University [58] and the University of Utah [59,60], among others. These efforts include considerations of how to store genomic data as part of a personalized health record [24,20,21,61,57], an important consideration since the current method has information loss [59] which promotes the potential of repeat testing. Since genetic analyses and sequencing are expensive, the storage and retrieval and reanalysis of genomic data is essential for the field to move forward, especially with the rapid expansion in molecular diagnostics [62,63]. Security concerns arise frequently when genomic data are discussed because of the threat of discrimination from insurance and employers and the ability to link genotype and phenotype data [64,65]. Clinical decision support is an essential component of personalized medicine since the complexity of the genomic data and its interaction with laboratory or pharmacy data requires computer assistance for alerts and reminders [61,66,67]. This underlies the overall complexity of the complex path from the gene or protein sequence data into various genetic testing scenarios and into actionable items in the Electronic Medical Record.

## Conclusion

Like all interdisciplinary areas, educational needs are demanding and complex. In the case of Medical Informatics, basic training included topics related, of course, to medicine (including clinical medicine and decision making, public health, or health services research) and computer science (including AI, probability or statistics). Personalized medicine will also demand new knowledge and expertise. For instance, physicians will need to learn concepts related to genetics or systems biology, whereas biologists and bioinformaticians will have to deal with clinical data and issues that have been unknown to them until now. Such complexity may even increase more dramatically if nanotechnology begins to have a significant impact on clinical practice beyond current laboratory research. Large initiatives will be necessary to create the tooling, interoperability and scientific domain driven knowledge base to effectively advance personalized medicine.

After some preliminary overoptimistic expectations, it seems clear now that genetics alone cannot transform medicine [50,51]. Research on biomarker discovery, that can be detected before clinical onset, has signalled molecular profiling as a great challenge for personalized medicine, but biomarkers with adequate specificity and sensitivity values are still scarce for most diseases. Biomarkers must be evaluated in order to demonstrate their medical significance and cost-effectiveness [52]. In order to achieve this, heterogeneous clinical and genomic data source must be integrated in scientifically meaningful and productive systems. This will include hypothesis-driven scientific research systems along with well understood information systems to support such research. These in turn will enable the faster advancement of personalized medicine.

**Acknowledgements**

The authors thank George Komatsoulis for discussions on domain modeling and interoperability for caBIG™, as well as Dianne Reeves for discussion on data element reuse and Denise Warzel for discussions on the cancer data standards repository.

## References:

1. Crick FHC. The Biological Replication of Macromolecues. Symp Soc Exp Biol 1958; XII:138.
2. Searls DB. Data integration: challenges for drug discovery. Nat Rev Drug Discov 2005 Jan;4(1):45-58.
3. Collins F. S. Shattuck Lecture - Medical and Societal Consequences of the Human Genome Project. N Engl J Med 1999 341:28-37.
4. Abrahams E, Ginsburg GS, Silver M. The Personalized Medicine Coalition: goals and strategies. Am J Pharmacogenomics 2005;5(6):345-55.
5. Meadows M. Genomics and personalized medicine. FDA Consum 2005 Nov-Dec;39(6):12-7.
6. Nicholson JK. Global systems biology, personalized medicine and molecular epidemiology. Mol Syst Biol 2006;2:52.
7. Haselden JN, Nicholls AW. Personalized medicine progresses. Nat Med 2006 May;12(5):510-1.
8. West M, Ginsburg GS, Huang AT, Nevins JR. Embracing the complexity of genomic data for personalized medicine. Genome Res 2006 May;16(5): 559-66.
9. Gurwitz D, Lunshof JE, Altman RB. A call for the creation of personalized medicine databases. Nat Rev Drug Discov 2006 Jan;5(1):23-6.
10. Davies SM. Pharmacogenetics, pharmacogenomics and personalized medicine: are we there yet? Hematology Am Soc Hematol Educ Program 2006;:111-7.
11. Dietel M, Sers C. Personalized medicine and development of targeted therapies: The upcoming challenge for diagnostic molecular pathology. A review. Virchows Arch 2006 Jun;448(6):744-55.
12. Klein TE, Chang JT, Cho MK, Easton KL, Fergerson R, Hewett M, Lin Z, et al. Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenomics J 2001;1(3):167-70.
13. Sadee W, Dai Z. Pharmacogenetics/genomics and personalized medicine. Hum Mol Genet 2005 Oct 15;14 Spec No. 2:R207-14.
14. Woodcock J. The prospects for "personalized medicine" in drug development and drug therapy. Clin Pharmacol Ther 2007 Feb;81(2):164-9.
15. Weatherall D. Sir David Weatherall reflects on genetics and personalized medicine. Interviewed by Ulrike Knies-Bamforth. Drug Discov Today 2006 Jul;11(13-14):576-9.
16. Sujansky W. Heterogeneous database integration in biomedicine. J Biomed Inform 2001 34(4):285-98.
17. Alonso-Calvo R, Maojo V, Billhardt H, Martin-Sanchez F, Garcia-Remesal M, Perez-Rey D. An agent- and ontology-based system for integrating public gene, protein, and disease databases. J Biomed Inform 2007 Feb;40(1):17-29.
18. Martin-Sanchez F, Iakovidis I, Norager S, Maojo V, de Groen P, van der Lei J, et al. Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. J Biomed Inform 2004;37(1): 30-42.
19. Del Fiol G, Williams MS, Maram N, Rocha RA, Wood GM, Mitchell JA. Integrating Genetic Information Resources with an EHR. AMIA Annu Symp Proc 2006; 904.
20. Mitchell JA. The impact of genomics on E-health. Stud Health Technol Inform 2004; 106:63-74.
21. Sax U., Schmidt S. Integration of genomic data in Electronic Health Records – opportunities and dilemmas. Methods Inf Med 2005;44 (4):546-50.
22. Mitchell DR, Mitchell JA. Status of clinical gene

sequencing data reporting and associated risks for information loss. J Biomed Inform 2007 Feb; 40(1):47-54.
23. Mitchell JA, McCray AT, Bodenreider O. From phenotype to genotype: issues in navigating the available information resources. Methods Inf Med 2003;42(5):557-63.
24. Adida B, Kohane IS. 2006. GenePING: secure, scalable management of personal genomic data. BMC Genomics 7(93):1-10.
25. Sanandres-Ledesma JA, Maojo V, Crespo J, Gómez de la Cámara A, García-Remesal M. A performance comparative analysis between rule-induction algorithms. Application to rheumatoid arthritis. Lecture Notes in Computer Science 3337:224-34;2004
26. Pazzani, M. Knowledge discovery from data? IEEE Intelligent Systems 2000;15(2):10-13
27. Maojo, V. Domain-specific particularities of data mining: Lessons learned. Lecture Notes in Computer Science. 3337: 235-42;2004.
28. Komatsoulis GA, Warzel DB, Hartel FW, Shanbhag K, Chilukuri R, Fragoso G. et al. caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability. J Biomed Inform. In press 2007 (doi:10.1016/j.jbi.2007.03.009)
29. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI thesaurus: A semantic model integrating cancer-related clinical and molecular information. J Biomed Inform2007 Feb;40(1):30-43.
30. Hartel FW, de Coronado S, Dionne R, Fragoso G, Golbeck J. Modeling a description logic vocabulary for cancer research. J Biomed Inform 2005;38(2): 114-29.
31. Fragoso G, de Coronado S, Haber M, Hartel F, and Wright L. Overview and Utilization of the NCI thesaurus. Comparative and Functional Genomics, Vol 5(8); 2004. p. 648-54.
32. de Coronado S, Haber MW, Sioutos N, Tuttle MS, Wright LW. NCI Thesaurus: using science-based terminology to integrate cancer research results. Medinfo 2004. 11(Pt 1). p. 33-7.
33. Covitz PA, Hartel F, Schaefer C, De Coronado S, Fragoso G, Sahni H, et al. caCORE: a common infrastructure for cancer informatics. Bioinformatics 2003;19(18):2404-12.
34. Lindberg C. The Unified Medical Language System (UMLS) of the National Library of Medicine. J Am Med Rec Assoc 1990;61(5):40-2.
35. Tuttle MS, Sperzel WD, Olson NE, Erlbaum MS, Suarez-Munist O, Sherertz DD, et al. The homogenization of the Metathesaurus schema and distribution format. Proc Annu Symp Comput Appl Med Care 1992. p. 299-303.
36. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem 2003;49(4):624-33.
37. Kudla KM, Rallins MC. SNOMED: a controlled vocabulary for computer-based patient records. J Ahima 1998;69(5):40-4; quiz 45-6.
38. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25(1):25-9.
39. Whetzel PL, Parkinson H, Causton HC, Fan L, Fostel J, Fragoso G, et al. The MGED Ontology: a resource for semantics-based description of microarray experiments. Bioinformatics 2006;22(7): 866-73.
40. Yu H, Friedman C, Rhzetsky A, Kra, P. Representing genomic knowledge in the UMLS semantic network. Proc AMIA Symp 1999. p. 181-5.
41. Bodenreider O, Mitchell JA, McCray AT. Evaluation of the UMLS as a terminology and knowledge resource for biomedical informatics. Proc AMIA Symp. 2002. p. 61-5.
42. Konagaya A. Trends in life science grid: from computing grid to knowledge grid. Pharmacogenetics Research Network and Knowledge Base. BMC Bioinformatics 2006 Dec 18;7 Suppl 5:S10.
43. Saltz J, Oster S, Hastings S, Langella S, Kurc T, Sanchez W, et al. caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid. Bioinformatics 2006; 22(15):1910-6.
44. de Knikker R, Guo Y, Li JL, Kwan AK, Yip KY, Cheung DW, Cheung KH. A web services choreography scenario for interoperating bioinformatics applications. BMC Bioinformatics 2004 Mar 10;5:25.
45. ACGT. http://www.eu-acgt.org
46. Tobias J, Chilukuri R, Komatsoulis GA, Mohanty S, Sioutos N, Warzel DB, et al. The CAP cancer protocols—a case study of caCORE based data standards implementation to integrate with the Cancer Biomedical Informatics Grid. BMC Med Inform Decis Mak 2006 Jun 20;6:25-40. PML, 2007. http://stdsnp.genes.nig.ac.jp/index.html
47. HL7 SIG. http://www.haifa.ibm.com/projects/software/cgl7/specifications.html
48. INFOBIOMED. http://www.infobiomed.org
49. Kulikowski C. The Micro-Macro Spectrum of Medical Informatics. Challenges: From Molecular Medicine to Transforming Health Care in a Globalizing Society. Methods Inf Med 2002;41:20-4.
50. Kiberstis P, Roberts L. It's not just the genes. Science 2002;296:685
51. Lunshof JE, Pirmohamed M, Gurwitz D. Personalized medicine: decades away? Pharmacogenomics 2006 Mar;7(2):237-41.
52. Collins CD, Purohit S, Podolsky RH, Zhao HS, Schatz D, Eckenrode SE, et al. The application of genomic and proteomic technologies in predictive, preventive and personalized medicine. Vascul Pharmacol 2006 Nov;45(5):258-67.
53. https://www.i2b2.org/
54. Annas GJ. Rules for research on human genetic variation—lessons from Iceland. N Engl J Med 2000;342:1830–3.
55. de Groen, PC. A healthy database: IBM creating a system for millions of Mayo Clinic patient files. Post-Bulletin. Rochester, MN; 2002, Mar 25. p. 1A.
56. Murphy SN, Mendis ME, Berkowicz DA, Kohane IS, Chueh HC. Integration of clinical and genomic data in the i2b2 architecture. AMIA Annu Symp Proc 2006. p. 1040.
57. Husser CS, Buchhalter JR, Raffo OS, Shabo A, Brown SH, Lee KE, et al. Standardization of microarray and pharmacogenomics data. Methods Mol Biol 2006;316:111-57.
58. Willard HF, Angrist M, Ginsburg GS. Genomic medicine: genetic variation and its impact on the future of health care. Philos Trans R Soc Lond B Biol Sci 2005 Aug 29;360(1460):1543-50.
59. Mitchell DR, Mitchell JA. J Biomed Inform 2007;40(1):47-54.
60. Leppert MF, Singh NA. Nonsyndromic seizure disorders: epilepsy and the use of the internet to advance research. Annu Rev Genomics Hum Genet 2003;4:437-57.
61. Hoffman MA. The genome enabled medical record. J Biomed Informatics 2007;40(1):44-6.
62. de Leon J. AmpliChip CYP450 test: personalized medicine has arrived in psychiatry. Expert review of molecular diagnostics. May 2006;6(3):277-86.
63. Mukherjee TK, Mishra AK, Mukhopadhyay S, Hoidal JR. High concentration of antioxidants N-acetylcysteine and mitoquinone-Q induces intercellular adhesion molecule 1 and oxidative stress by increasing intracellular glutathione. J Immunol 2007 Feb 1;178(3):1835-44.
64. Rind DM, Kohane IS, Szolovits P, Safran C, Cheuh HC, Barnett GO. Maintaining the confidentiality of medical records shared over the Internet and the World Wide Web. Annals of Internal Medicine 1997;127(2):138-41.
65. Malin BA. An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future J Am Med Inform Assoc. 2005 Jan-Feb;12(1):28-34.
66. Deshmukh V, Hoffman MA, Arnoldi CA, Bray BE, Mitchell JA. Efficiency of CYP2C9 Genetic Test Representation for Automated Pharmacogenetic Decision Support. AMIA Annu Symp Proc 2007; Manuscript under review.
67. Louis B, Mork P, Martin-Sanchez F, Halevy A, Tzarchy-Hornock P. Data integration and genomic medicine. J Biomed Infor 2007 Feb; 40(1):5-16

**Correspondence to:**
Lewis J. Frey, PhD
26 South 2000 East
5700 HSEB
Salt Lake City, Utah 84112
USA
Tel: +1 801 585-9428
E-mail: lewis.frey@hsc.utah.edu