Clinical genomics data standards for pharmacogenetics and pharmacogenomics

Amnon Shabo (Shvo)
IBM Research Lab in Haifa,
Haifa University Campus,
Mount Carmel,
Haifa, 31905, Israel

This special report concerns a talk on data standards given at a workshop entitled 'An International Perspective on Pharmacogenetics: The Intersections between Innovation, Regulation and Health Delivery', which was held by the Organization for Economic Co-operation and Development (OECD) on October 17–19, 2005, in Rome, Italy. The worlds of healthcare and life sciences (HCLS) are extremely fragmented in terms of their underlying information technology, making it difficult to semantically exchange information between disparate entities. While we have reached the point where functional interoperability is ubiquitous, we are still far from achieving true semantic interoperability where a receiving system can use incoming data as though it was created internally. The critical enablers of semantic interoperability are information standards dedicated to HCLS data, spanning all the way from biological research data to clinical research and clinical trials, and finally to healthcare clinical data. The challenge lies in integrating various data standards based on predetermined goals, thereby improving the quality of care provided to patients.

Keywords: clinical genomics, healthcare, information standards, life sciences, pharmacogenetics, pharmacogenomics, personalized medicine

Existing standards

In healthcare, the most common use of standards is for medical terminology, some of which became ubiquitous, and even mandatory in several use cases. For example, International Coding of Diseases (ICD) [101] is maintained by the World Health Organization and is used in many countries around the world for reporting purposes and in some cases within the healthcare enterprise. Systematized Nomenclature of Medicine (SNOMED) [102] was recently licensed by the US and UK governments. It is now mandatory in the new program for health information technology (IT) in the UK and free of charge to any healthcare provider in the US. The SNOMED Clinical Terms (CT) core terminology contains over 357,000 healthcare concepts with unique meanings and formal logic-based definitions organized into hierarchies. As of January 2004, the fully populated table with unique descriptions for each concept contained more than 957,000 descriptions. Approximately 1.37 million semantic relationships exist to enable reliability and consistency of data retrieval.

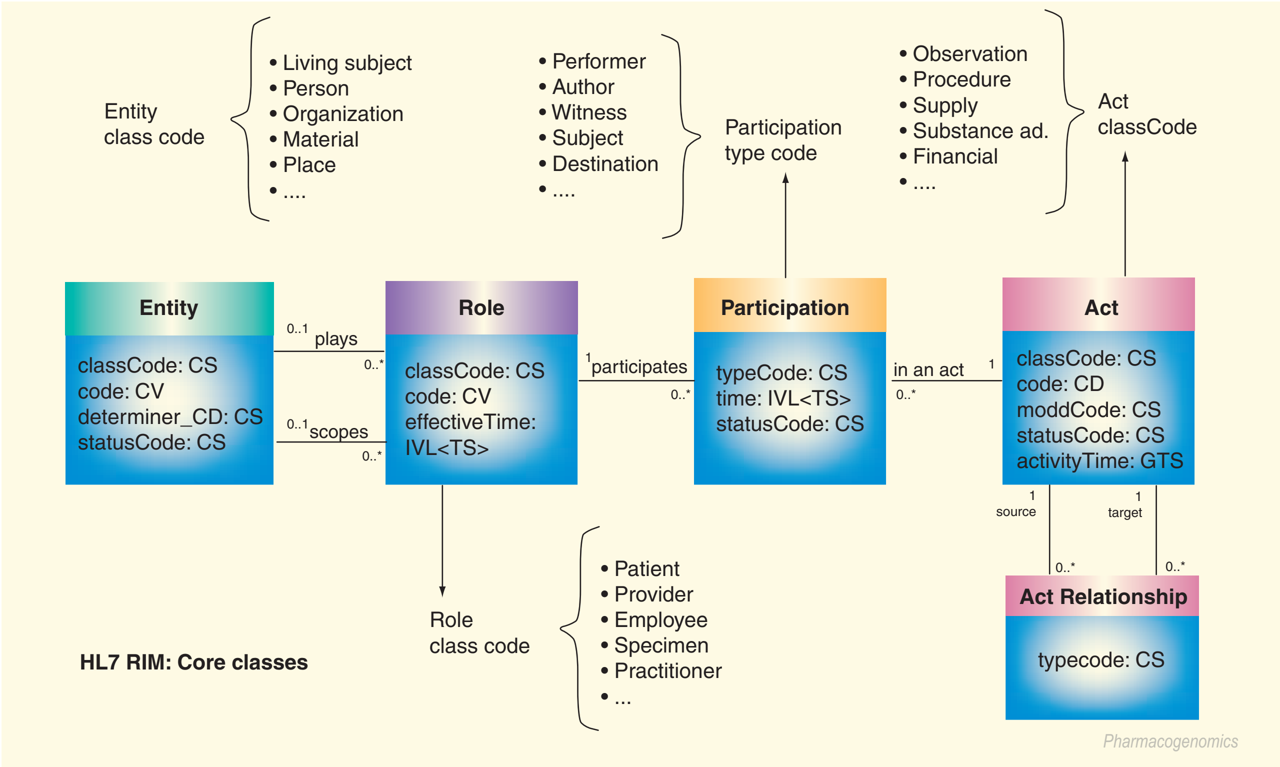
While SNOMED has evolved from a pathology coding scheme, logical observation identifiers names and codes (LOINC) [103] is still considered the major taxonomy of laboratory terms. LOINC is a terminology that facilitates the exchange and pooling of testing results, such as blood hemoglobin, serum potassium, or vital signs used in clinical care, outcomes management and research. In order to relay these results back to the ordering system, most laboratories and other diagnostic services use Health Level Seven (HL7) [104] messages to send their results electronically from their reporting systems to care information systems. However, most laboratories and other diagnostic care services identify tests in these messages by means of their internal and idiosyncratic code values. Thus, the care system cannot fully decipher and properly file the results they receive unless they either adopt the producer's laboratory codes (which is impossible if they receive results from multiple sources), or invest in mapping each result producer's code system to their internal code system. LOINC codes are universal identifiers that solve this problem for laboratory and other clinical observations and can be conveyed by HL7 messages.

HL7 is the most dominant standardization body in the area of health messaging. Its scope covers the next level of standardization built on the standardized terminologies described above. HL7 develops standards for the specification of messages and documents in areas such as:

- Clinical: laboratory, medical records and clinical documents, patient care, pharmacy, public health reporting, regulated studies and clinical studies
- Administrative: account and billing, claims and reimbursement, patient administration, personnel management scheduling
- Infrastructure: transmission infrastructure, message control infrastructure, query infrastructure, master file infrastructure

10.2217/14622416.7.2.xxx © 2006 Future Medicine Ltd ISSN 1462-2416

Figure 1. The backbone of the HL7 Reference Information Model.



An entity plays a role that participates in an action. The RIM backbone is a generic set of classes from which health specific classes are derived, for example, observation, procedure and substance administration are all subclasses of act.
CD: Concept descriptor; CS: Coded simple value; CV: Coded value; GTS: General Timing Specification; HL7: Health Level Seven; IVL: Interval; RIM: Reference Information Model; TS: Point in time.

Another messaging standard is Digital Imaging and Communications in Medicine (DICOM) [105], which is ubiquitous in the area of medical imaging. Most of the imaging modalities output the images in DICOM format, though with slight differences. Thus, imaging modalities, image archiving systems, and radiology information systems are able to exchange images and view them in a standard way.

The third level of standardization builds on the two lower levels (terminologies, messages/documents/images) and targets the final frontier of medical informatics – the electronic health record (EHR). EHRs refer to the patient records in the healthcare enterprises, but recently, and more importantly, pertain to the emerging concept of a patient-centric longitudinal and cross-institutional record, a complex information entity that poses great challenges for the standardization of medical information. To date, there is no EHR information standard that is as widely implemented as those for terminologies and messages. The European Standardization Body (CEN) developed the EN 13606 specifications [106] for EHR, and together with its new archetypes technology, which addresses specific requirements of the various clinical domains, it offers a viable EHR standard that is likely to be implemented in national EHR projects in various countries.

In the life sciences realm, efforts have been made to devise common formats to represent biological and research data, mainly in genomics. For example, microarray and gene expression markup language (MAGE-ML) [107] was developed to represent gene expression assays, and the bioinformatic sequence markup language (BSML) was developed for sequencing data. BSML [108] was the first extensible markup language (XML) application in the life sciences field. It was developed under a grant from the National Human Genome Research Institute (NHGRI) to provide standard methods for communicating genomic research information. The BSML format is a mixture of: the definitions section, which encodes the bioinformatics data; the research section, which encodes queries, analysis, and experimentation; and the display section, which encodes information for graphic representation of the bioinformatics data. While this mixture has the merit of being a multipurpose design, it also points to the fact that early efforts in the standardization of biological data were not designed in a modular and scaleable way.

The Clinical Data Interchange Standards Consortium (CDISC) [109] develops the industry standards that support the electronic acquisition, exchange, submission, and archiving of clinical trial data. CDISC has a number of working groups that are developing standards, including the laboratory standards (LAB), operational data model (ODM), and submission data standards (SDS) models.

CDISC and the HL7 Regulated Clinical Research and Information Management (RCRIM) Technical Committee (the clinical trial group) work together to create clinical trial data standards that could be recognized internationally (American National Standards Institute [ANSI], International Organization for Standardization [ISO], and so on) since CDISC is not an ANSI-accredited organization. Regulatory agencies such as the US FDA encourage the adoption of standards developed by CDISC within the HL7 organization to make it better aligned with clinical data provided by the hospitals. The HL7 RCRIM group uses the HL7 methodology and tooling to express the CDISC data formats in the 'HL7 language'.

There are many more existing standards and common data formats in the world of HCLS. However, this report is concerned with bridging the gap between types of personal information in healthcare and life sciences. The next section describes a 'bridge' standard fusing genomic data with clinical and administrative data.

Data standards for clinical genomics

As mentioned above, several HCLS data standards are already in place, but they typically serve a limited scope, either in healthcare or in the life sciences. The emerging use of genetic data in healthcare poses a challenge in integrating those existing standards, as well as supporting new work flows that eventually aim at the use of patient-specific genetic data at the point of care. One such work flow envisioned by the HL7 Clinical Genomics Special Interest Group is the 'encapsulate & bubble-up' paradigm underlying the design of the HL7 Clinical Genomics Standard Specifications. It demonstrates a method for encapsulating raw genomic data resulting from genetic testing of the patient and then bubbling-up the most clinically-significant items while associating them with observed phenotypes in the patient, as well as scientifically known phenotypes from publicly available reference databases such as Online Mendelian Inheritance in Man (OMIM) [110] and based on ontologies such as Gene Ontology (GO) [111] and Clinical Bioinformatics Ontology™ (CBO) [112].

The HL7 Clinical Genomics Standard Specifications are part of the new HL7 V3 family of standards. All HL7 V3 message and document standard specifications are derived from a new Reference Information Model (RIM) by means of refinement, thus achieving better semantics consistency across the various specifications. The RIM was developed mainly for clinical and administrative data and the challenge is to extend it to life sciences data so that it can become a comprehensive health reference information model that will bridge the gap between clinical and biological data (in particular, genomic data).

The HL7 RIM is the core of the new HL7 version 3 and includes dozens of classes that represent the building blocks from which each healthcare message/document is built. The heart of the HL7 RIM comprises four fundamental classes: entity, role, participation, and act. Most of the classes are specializations of one of these fundamental classes. For example, a procedure class representing a medical procedure is a specialization of the act class. Observations are also considered acts, and represent lab orders and results, diagnoses, and more. Patients and healthcare providers are represented through the associations of entity-role-participation. For example, a person is an entity that has the role (competency) of a physician and has the participation of an attending physician in the act of admitting the patient to the hospital. A role can be played by an entity and be scoped by an entity. For example, the role of a physician is played by a person and scoped by the healthcare enterprise (entity) that assigned the responsibility to the physician.

The clinical genomics domain addresses requirements for the interrelation of clinical and genomic data at the individual level. Much of the genomic data is still generic. For example, the human genome is in fact the DNA sequences believed to be the common sequences in every human being. The vision of 'personalized medicine' is based on those correlations that make use of personal genomic data such as the single nucleotide polymorphisms (SNPs), which differentiate any two individuals and occur approximately every thousand bases. Aside from general differences, health conditions such as drug sensitivities, allergies and others could be attributed to the individual SNPs or to differences in gene expression and proteomics. The clinical genomics domain emphasizes the personalization of the genomic data and the 'intelligent' linking to relevant clinical information. These links are most likely the main source from which geneticists and clinicians could benefit.

Pharmacogenetics & pharmacogenomics

The cases where genomic data are used in healthcare practice vary in the complexity and extent of the data used, since the current testing methods are still very expensive and not widely used. It is common to find simple testing, such as identifying genes and mutations. However, full sequencing of alleles and the use of new methods to identify the expression of a number of genes (for example, a panel of a few dozens of genes) of an individual are becoming available for the clinical practice. The HL7 Clinical Genomics Special Interest Group has been focusing on the genetic testing routinely performed in healthcare, while preparing the information infrastructure standard for more futuristic cases. For example, regarding medications, the term pharmacogenetics refers to the more routine use of genetic information, while the term pharmacogenomics refers to the discovery phases of the drug development process. In the future, these could also be personalized as part of the broader vision for personalized medicine.

Discussions held at the OECD Workshop on Pharmacogenetics [113] elaborated on the distinction between pharmacogenetics and pharmacogenomics. Generally speaking, pharmacogenetics is concerned with identifying the best medicine for a specific disease occurring in a patient population with a particular genotype. It deals with the differences between individuals in their responses to medicines in terms of their metabolism (pharmacokinetics) or action (pharmacodynamics). Pharmacogenomics deals with the genetic markers that may diagnose or stage a disease in the context of drug responses. It attempts to optimize the identification of those drugs in the discovery process that bring about the most appropriate pharmacological responses. Pharmacogenomics uses emerging genomic technologies to explore how interacting groups of genes may influence pharmacological function and therapeutic drug response. While more closely examining the differences between pharmacogenetics and pharmacogenomics, it becomes apparent that they both deal with drug responses for clinical purposes but in different phases of the drug development cycle and with different scopes. Methods of developing new drugs could be applied in the future to existing medications in order to fine tune them to a specific patient in the spirit of personalized medicine. Even today, new research has implications on existing medicines, and therefore, clinical genomic bridge standards should take into account the data representation requirements of the entire continuum between pharmacogenetics and pharmacogenomics. In the HL7 Clinical Genomics Specifications, the term genomics refers to the entire continuum as discussed here.

The main part of the HL7 Clinical Genomics Specifications is the Genotype model, a data payload component that can be used in any message conveying genomic data. It consists of various types of genomic data relating to a specific genetic locus, including sequencing, expression, and proteomic data, thus includes functional genomic data as well (note that the term Genotype is used in its broad sense). Within the Genotype model, existing bioinformatics mark-up languages are utilized to represent data received from genetic/genomic facilities.

Design principles of the HL7 Clinical Genomics Specifications

- The main goal is to bridge and associate personal genomic data and clinical data using HL7 messages so that personalized medicine can be facilitated.
- Data representations such as MAGE-ML for gene expression data and BSML for sequencing data are perceived as the 'raw genomic data'.
- The main paradigm underlying the design is 'encapsulate & bubble-up' – a conceptual workflow envisaged to take place in the following way: The encapsulation phase is a static process where certain ‘encapsulating objects’ in the data model are being populated with portions of raw genomic data, based on predefined constrained bioinformatics schemas such as MAGE-ML. The constraining process is part of the standardization effort and aims at leaving out portions such as the display elements in the BSML markup and others that seem irrelevant to the clinical practice. It also makes sure the data refers to only one patient and one genetic locus in accordance with the basic data model of the standard (the Genotype). The 'bubbling-up' phase is a dynamic

process where genomic-oriented decision-support applications parse the raw genomic data already encapsulated in the HL7 message. This phase brings to the surface those portions that seem to be the most clinically significant to the patient's clinical history and current treatment plans, based on the most updated scientific knowledge. The results of this bubbling-up process are held in other HL7 objects in the Genotype model that can also be associated with clinical data in the patient's medical records (represented in the Genotype model as the 'clinical phenotype').

- These static and dynamic phases lead to a gradual 'distillation' of the raw genomic data in the context of diagnosis and treatment provided to a specific patient. They also hold essential parts of the raw data within the HL7 objects so that it would be possible to parse the data again, for example, when new knowledge becomes available. The complete raw genomic data could be made accessible by referencing, for example, by using the Life Sciences Identifier (LSID) [114].

Although describing the Genotype model in detail is beyond the scope of this report, **Figure 2** illustrates a portion of the model related to the Sequence class. The 'value' attribute of this class can hold raw data in bioinformatics markup, such as BSML. There is a recursive association (at the bottom left) that allows the representation of another biological sequence derived from a source sequence (for example, mRNA sequence derived from a DNA sequence). An association with clinical phenotype (at the bottom right) allows the link to clinical data most likely residing in the patient medical records. Other associations are related to sequence variations and proteomic data. The entire model is described in detail in the HL7 V3 ballot package available from the HL7 site [104].

The challenge of sustaining clinical genomic data

The exact constellation in which clinical genomic data belonging to a patient is stored and maintained has various implications on all parties with an interest in the data. The patient wishes to obtain the best care possible, while at the same time maintaining the privacy of the data. The healthcare providers need the data not only for direct care purposes but also for their operations and quality control. Research institutes and pharmaceutical companies need the data in their innovative efforts when trying to base their experiments on complete histories of the patients enrolled in their studies. It is common to find the data scattered across different enterprises and represented by various incompatible and proprietary formats. Thus, if data is kept in the sources where it was created, it becomes difficult to compile a complete clinical genomic history of a patient.

Many nations around the world are undertaking preliminary efforts to create national and regional infrastructures for health information interoperability. The two extreme approaches can be found in the US and the UK. The US is promoting the notion of a regional health information organization, which keeps a registry with the location of all medical records pertaining to a patient and is capable of resolving these locators and compiling an EHR on the file [115]. The records themselves continue to reside at the location where they were created. In the UK, the National Health Service is promoting a more centralized approach where patient data is sent to the so-called spine [116], which is, in fact, playing the role of a national repository for the citizens' EHRs, though definitely not complete EHRs. While the regional and distributed approach in the US has serious problems regarding availability, the centralized approach of the UK has the obvious issues of any centralized operation on that scale run by a government.

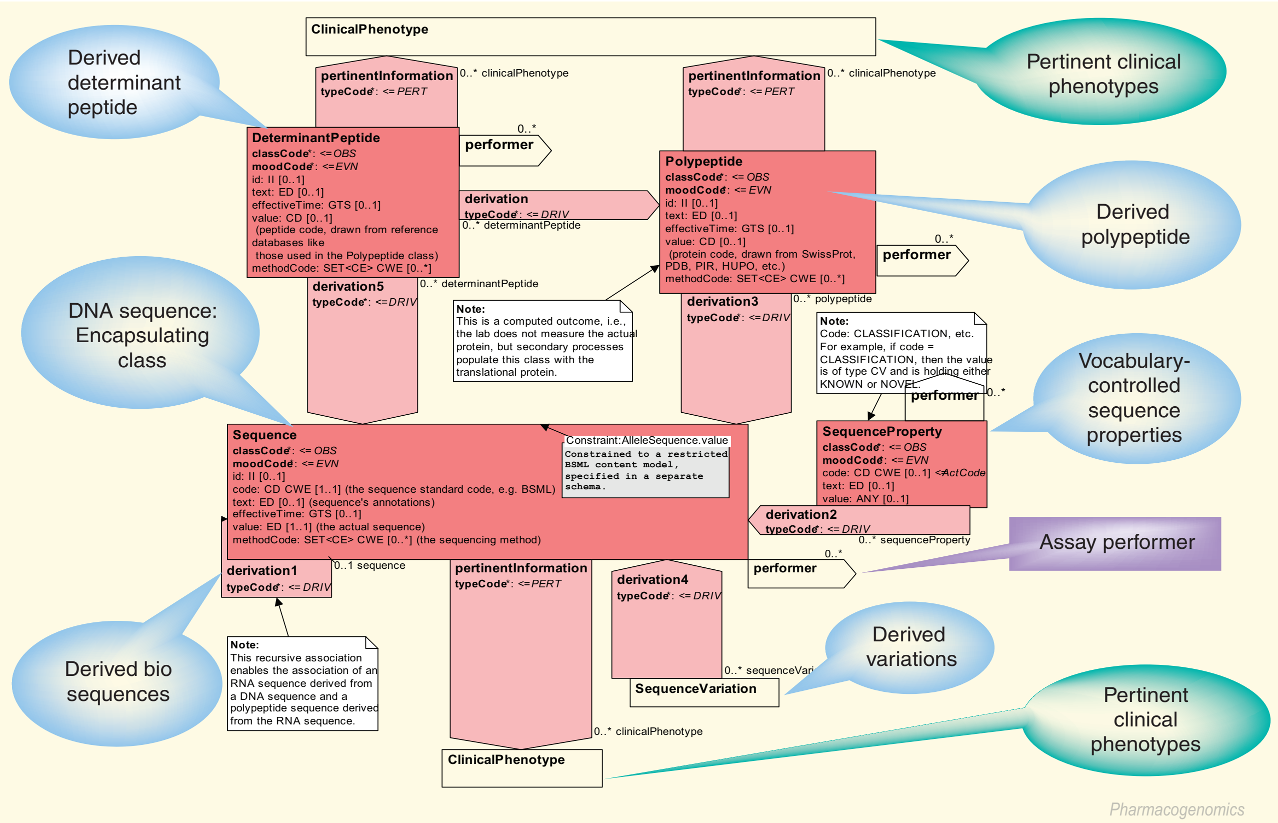
An alternative to the problem of EHR sustainability is the vision of independent health record banks (IHRB) [1,117]. The idea is to have lifetime EHRs sustained by independent banks whose main business will be to maintain the lifetime EHR of their customers. There will be multiple competing banks that will be regulated only by governments. Each bank will have to be independent of all other interested parties – it will have to be independent of healthcare providers, health insurers, government agencies and even the patient themselves, in order to avoid any conflict of interest and provide an objective service.

Figure 3 shows the various approaches being experimented with around the world and compares them based on their focal point.

When a person is enrolled in a clinical trial, all information (clinical and genetic) that is related to the person and collected during the trial should be sent to the person's longitudinal and cross-institutional EHR, ideally maintained by an IHRB. Clinical trial data is also personal data that should be part of the person's health history.

Research institutes and pharmaceutical companies will then be able to obtain full clinical genomic histories from IHRBs upon consent of

Figure 2. A portion of the Genotype data model with the Sequence class as a focal point.



Note:
This is a computed outcome, i.e., the lab does not measure the actual protein, but secondary processes populate this class with the translational protein.

Note:
Code: CLASSIFICATION, etc.
For example, if code = CLASSIFICATION, then the value is of type CV and is holding either KNOWN or NOVEL.

Constrained to a restricted BSML content model, specified in a separate schema.

Note:
This recursive association enables the association of an RNA sequence derived from a DNA sequence and a polypeptide sequence derived from the RNA sequence.

The Genotype model is a major part of the HL7 Clinical Genomics specifications.
HL7: Health Level Seven.

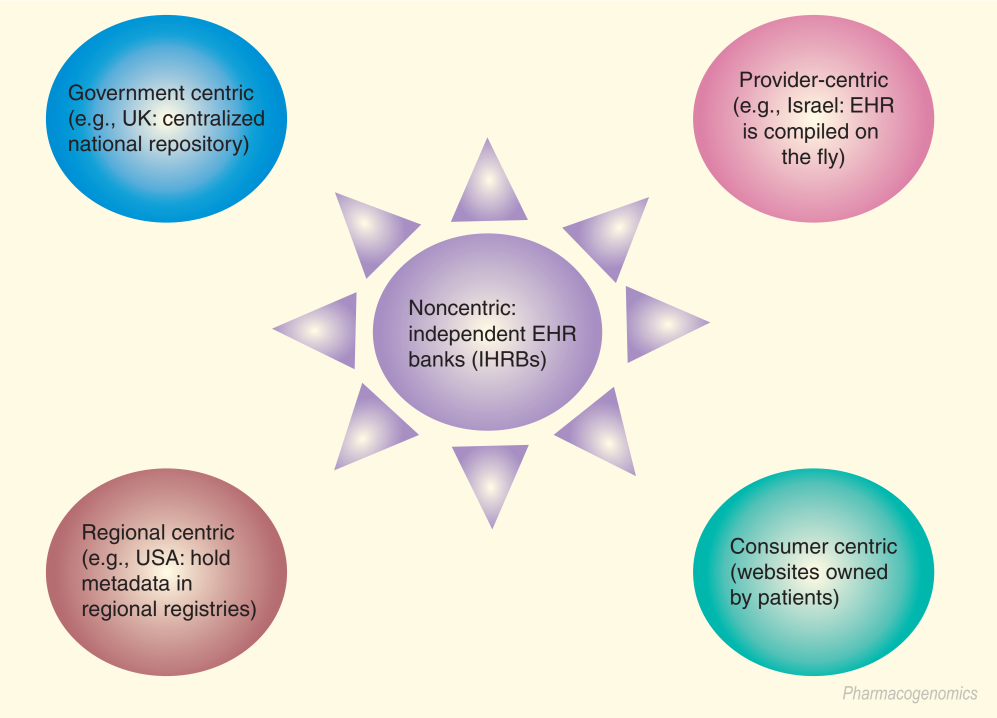
the patients enrolled in their studies. This will greatly benefit all parities as it will resolve the issue of how to accumulate historical data and will allow each new study to rely on all the data collected so far, including other clinical trials in which the patients may have participated. Today, most of the published papers on clinical trials do not consist of the raw data and certainly not in the EHR framework which would enable researchers to base their current study on semantically interoperable histories of the patients enrolled in that study.

Current laws in most countries require providers to keep the medical records for a certain number of years. This period is typically much less than the average lifetime, and is generally approximately 7 years for adults. There is a need for new legislation for the IHRB vision to be realized, namely establishing IHRBs as the sole archivers of medical records, while requiring their independence as described above. The IHRB databases could then become the ultimate source for new research in pharmacogenetics and the most viable source of personalized medicine. Obviously, all of this can only be realized if internationally-recognized data standards for clinical genomic data are in place.

Highlights

- Interoperable data standards in clinical genomics are essential for bridging the information gap between healthcare and life sciences data formats.
- Such interoperable standards could enable more effective use of pharmacogenetics in clinical practice.
- Storing clinical genomic patient data should eventually be performed using the longitudinal and cross-institutional patient electronic health records.
- Storing individual health records can be done using several different models currently being experimented with at the regional and national level in many countries.
- The independent health records banks, an alternative to the current models proposed for health record sustainability, can also include genetic data and could thus provide an invaluable source of data for pharmacogenetic and pharmacogenomic studies.

Figure 3: Various EHR sustainability models are compared based on the stakeholder at the center of the model.


Government centric (e.g., UK: centralized national repository)
Provider-centric (e.g., Israel: EHR is compiled on the fly)
Noncentric: independent EHR banks (IHRBs)
Regional centric (e.g., USA: hold metadata in regional registries)
Consumer centric (websites owned by patients)
Pharmacogenomics

These models are contrasted with the model of independent health record banks, which is not centered on any interested party in the health arena.
EHR: Electronic health record; IHRB: Independent health record banks.

Bibliography

1. Shabo A: A global socio-economic-medico-legal model for the sustainability of longitudinal electronic health records. *Methods Inf. Med.* (2006) (In press).

Websites

101 ICD – International Classification of Diseases. www.who.int/classifications/icd/en/.

102. SNOMED – Systematized Nomenclature of Medicine www.snomed.org/.

103. LOINC – logical observation identifiers names and codes www.regenstrief.org/loinc/

104. HL7 (Health Level Seven), an ANSI-accredited standards developing organization in healthcare www.hl7.org.

105. DICOM – Digital Imaging and Communications in Medicine – Standard Specification http://medical.nema.org/dicom.html.

106. Health Informatics – Electronic healthcare record communication – Part 1: extended architecture. ENV13606-1, Committee European Normalisation, CEN/TC 251 Health Informatics Technical Committee. www.openehr.org/standards/t_cen.htm.

107. MAGE-ML – Microarray and gene expression markup language www.mged.org/Workgroups/MAGE/mage.html.

108. BSML - Bioinformatic sequence markup language www.bsml.org.

109. CDISC – The Clinical Data Interchange Consortium www.cdisc.org.

110. OMIM – Online Mendelian Inheritance in Man. www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

111. GO – Gene Ontology www.geneontology.org/

112. Hoffman M, Arnoldi C, Chuang I: The clinical bioinformatics ontology: a curated semantic network utilizing RefSeq information. *Pac Symp Biocomput.* 139-150 (2005). http://helix-web.stanford.edu/psb05/hoffman.pdf

113. An international perspective on pharmacogenetics: the intersections between innovation, regulation and health delivery. Organization for Economic Co-operation and Development (OECD). October 17–19, 2005, in Rome, Italy. www.oecd.org/document/16/0,2340,en_2649_34537_35517584_1_1_1_1,00.html.

114. LSID – Life Sciences Identifiers Specification from OMG – The Object Management Group www.omg.org/docs/lifesci/03-12-02.doc.

115. The US National Health Information Infrastructure (NHII) http://aspe.hhs.gov/sp/NHII/.

116. The UK National Health Service (NHS) – Connecting for Health. The Care Record Service of the Spine. www.connectingforhealth.nhs.uk/delivery/programmes/spine.

117. Shabo A, Vortman P, Robson B. Who's afraid of lifetime electronic medical records? TEHRE 2001. Proceedings of Towards Electronic Health Records Conference, London, UK, November 14 (2001). www.haifa.il.ibm.com/projects/software/imr/papers/WhosAfraidOfEMRfinal.pdf.