Initial high-level model definition of an ACGT-specific Clinico-Genomic EHR

Project Number:	FP6-2005-IST-026996
Deliverable id:	D 5.5
Deliverable name:	Initial high-level model definition of an ACGT-specific Clinico-Genomic EHR
Submission Date:	28/11/2008

COVER AND CONTROL PAGE OF DOCUMENT	
Project Acronym:	ACGT
Project Full Name:	Advancing Clinico-Genomic Clinical Trials on Cancer: Open Grid Services for improving Medical Knowledge Discovery
Document id:	D 5.5
Document name:	Initial high-level model definition of an ACGT-specific Clinico-Genomic EHR
Document type (PU, INT, RE)	RE
Version:	2.0
Submission date:	28/11/2008
Editor: Organisation: Email:	Anca Bucur Philips Research anca.bucur@philips.com

Document type PU = public, INT = internal, RE = restricted

ABSTRACT:

This deliverable proposes a high level data model for incorporating genomic data in a future genomic-enabled clinical information system (e.g. EMR, EHR) in the context of oncology, based on existing and emerging standards. It argues that a genomic-enabled system is becoming highly relevant for clinical practice taking into account the new validated discoveries from clinical research, but it could also have an important role to support new research (i.e., hypotheses building and retrospective validation, cohort studies, etc.) by collecting and integrating large amounts of data from clinical practice and enabling extensive querying.

KEYWORD LIST: genomic-enabled EHR, patient record, genomic data, microarray, HL7, MIAME, MAGE

MODIFICATION CONTROL			
Version	Date	Status	Author
1.0	25/11/08	Draft	Anca Bucur, Jasper van Leeuwen, Richard Vdovjak, Jeroen Vrijnsen
2.0	28/11/08	Final	Anca Bucur

List of Contributors

- Anca Bucur, Philips Research
- Jasper van Leeuwen, Philips Research
- Richard Vdovjak, Philips Research
- Jeroen Vrijnsen, Philips Research
- Vicky Danilatou, FORTH
- Christine Desmedt, Jules Bordet Institute
- Dimitris Kafetzopoulos, FORTH
- Thierry Sengstag, Swiss Institute of Bioinformatics

Contents

1 Introduction

The cancer research and clinical practice are currently faced with an explosion of relevant information, which brings along an unprecedented specialization of expertise. A genetic disease, cancer requires increasing amounts of genomic information for its diagnosis, patient stratification and treatment selection. Much of this currently expensive genomic data should become persistent and accessible through a future post-genomic clinical information system (e.g. EMR, EHR), to be used for both patient treatment and future research.

We have carried out interviews, with experts in the clinical, bio-molecular and bioinformatics domain, as part of a task focusing on analyzing the opportunities for allowing the two currently separated worlds of clinical research and clinical practice in oncology to connect and benefit from each other, and on identifying the needs for integrating genomic data in future clinical information systems (e.g. EMRs/EHRs). Discoveries from clinical research should be transferred to clinical practice as soon as possible to provide better treatment options to a larger number of patients. At the other end, clinical practice needs to provide feedback to research and as most of the patients are treated outside of clinical research, ways (e.g. approaches) and appropriate means (e.g. tools) should be found to use the large amount of data collected in that process for advancing clinical research.

Currently, breast cancer patient management is the most advanced with respect to patient stratification, therapy options, outcome prognosis and response prediction. This set of diseases (it was recently proven that breast cancer is not one, but four molecularly different diseases) can be used as model to envision how complex the management of cancer-related patient information will become in the future, for all cancer types.

Cancer is a genetic disease and genomic information will be essential in the future in all cancer types, but currently most commercially-available genomic tests have been developed in the context of breast cancer. More data is available for breast cancer than for any other type of cancer, and genomic data becomes more and more important as new discoveries are being made. Several microarray-based prognostic tests have been validated and approved, such as MammaPrint and OncotypeDX, and as they become reimbursed their use in clinical practice increases very fast. Those tests make a significant difference in the patient treatment, with respect to outcome, quality of life and costs, due to their better prognostic value compared to classical tests.

As more and more data of various types (clinical, genomic, imaging, pathology, etc.) is being collected for current clinical practice, it becomes increasingly important to preserve that data and to use it for research but also for the future benefit of the patient, in the light of new discoveries. Preserving the data is especially meaningful when storing and maintaining it is significantly cheaper than re-acquiring it, and when new insight can be obtained based on old data, as it is the case with genomic information.

The volume of data collected for a patient in the context of a complex disease such as cancer increases tremendously and a significant part of the medical history can be relevant. In the case of recurring cancer patients, the relevant cancer-related health episodes can go many years back. Co-morbidities are often relevant as well, as they are a very constraining factor for choosing a therapy. For example, many chemotherapy agents are cardio-toxic, and in order to choose for the right therapy prior information concerning cardiac disease is important. The cardiac function will also be monitored during the entire treatment. In this context, a decision support tool focused on the tumor board and the tumor board preparation should be able to extract all the relevant prior health episodes from the patient record, both cancer- and non-cancer-related, while still being able to screen out irrelevant patient information in patient records that may include many episodes and span decades.

2 Rationale

The task of assessing the impact of genomic information on the future treatment of cancer and the need to include such information in a future post-genomic clinical information system (e.g. EMR, EHR) requires a significant amount of domain insight. Also, in order to address the development of a data model for genomic data in, for example, a future EMR (and its implementation), questions such as who the users are and who will fill in the data, what data needs to be stored and how it will be used, and how will research results be introduced in practice are highly relevant and should be answered first.

The need to properly answer these questions based on expert clinical knowledge motivated us to first carry out interviews with domain experts involved in the treatment and research of cancer. Based on the interviews we have defined several relevant clinical scenarios that we used to derive requirements for a genomic-enabled clinical information system (e.g. EMR or EHR).

The interview outcomes, and the amount and scope of research in the oncology area that takes into account genomic data, made us conclude that there is a large body of genomic information that will become part of standard practice in the coming years. As a medical record should provide access to all relevant patient data used in the process of delivering care, all the new genomic information used in standard care, but also in research, needs to make its way into the future patient records.

We have focused on clinical research requirements and present directions because they allow us to envision the context and the needs of the future standard clinical practice. We have developed our scenarios for genomic data mainly based on research results related to breast cancer, as this area provides the most advanced patient stratification and management and the most diverse therapy options and it can contribute with the most complex and diverse requirements to a future genomic-enabled clinical information system. The amount of genomic data generated and used in the context of breast cancer is also significantly larger and more varied than in any other oncology area, which also poses higher requirements on the clinical information system that would store, access and manage such data.

Most current off-the-shelf EMR/EHR solutions do not fit the needs of research (e.g. searching, aggregating, integrating data) and they are also very far from considering the relevance of genomic data for clinical practice.

3 Requirements for a future post-genomic EHR based on interviews with clinical users

In this Section we summarize the main insights concerning relevant data types that should be accessible from a post-genomic clinical information system (e.g. EMR or EHR) in the context of oncology. Based on these requirements we define several scenarios concerning genomic data that are described in Section 4.

Although in the interviews we have focused on breast cancer, because as we previously mentioned this disease is more advanced with respect to decision making and classification, but also with respect to existing treatment options and data collection and management, the issues identified are valid in a much wider context, as the situation is similar in other cancer types. The underlying biology of the disease is different but the key questions and the methodologies are very similar. In the future more detailed patient stratification based on genomic information is needed. Tests should be disease-specific, easy to interpret and should allow reducing costs by providing the best therapy to the patients who respond to it. Extrapolating from breast cancer research and practice allows us to envision future trends concerning information relevant for cancer patient management which should become persistent in a clinical information system.

It is often required that all collected medical information concerning the patient is accessible when a patient is discussed in the tumor board and a treatment plan is selected because it is difficult to discern a priori what information will be needed. In many cases everything is kept, no filtering is carried out, as it is considered dangerous because important information may be lost. Currently decisions are based on morphometric and pathological (histology, hormone receptor status, status lymph nodes) factors. All these help stratify the patients and decide on a suitable treatment plan.

The family history is also taken into account. Known risk models are used in practice (e.g. St. Gallen). Knowing that a patient has a high risk may influence the therapy choice. For example for patients with BRCA1 mutations radiotherapy needs to be avoided, because BRCA1 is involved in cell repair.

Next to the types of data currently used such as classical clinical data (age, hormone receptor status, routinely-assessed biochemical markers, etc.) and survival data when the patient is followed, new types of data will become relevant. Other new types of information are microarrays, SNPs chips for mutations in the genome, chips to identify patterns of transcription connected to variables relevant for the analysis of the genes, etc.

Molecular profiling tests can be very comprehensive, and can identify more groups than currently known. Currently existing tests outperform histopathology, but they are not widely used in clinical practice.

To further improve the outcome, one needs to start from the molecular subgroup of breast cancer and to understand what are the underlying mechanisms causing that those patients belong to that group, what are the pathways involved, and based on that decide the therapy. A quantitative assessment is necessary, to replace the previously used, less accurate, immunohistochemistry.

Tests that will be used in clinical practice may also be different than those from clinical research. One reason is the need to reduce costs in order to be able to scale up to wide clinical practice. Currently, the arrays used in research are often tailor-made and very expensive. Prices may be required to go an order of magnitude down in order to turn these tests into affordable diagnostic tools. Also several tests could be combined on one array to reduce costs.

In the clinical research context the more data will be available the better, which supports the idea of preserving all genomic patient data. People will tend to collect more and more data, as access to as much data as possible is important in order to be able to better stratify patients, refine the research discoveries, generate and investigate new hypotheses.

Old data is currently used as a validation means for fundamental research discoveries, and for hypotheses generated in the context of a clinical trial. Typically a clinical trial addresses a question which is answered formally with the clinical trial data, but the data contains much more information than that targeted by the trial. By mining that data one can suggest new signatures, new targets, and new ideas for clinical trials. New signatures can be retrospectively validated on old data, before starting a new prospective clinical trial. This approach will probably start being routinely used in large university hospitals, who would be able to capitalize on their investments in technology, e.g. micorarrays, proteomics analysis. In general, in large research organizations, no data from clinical practice is discarded as being irrelevant, as it may be relevant for new research. Genomic data from clinical practice will be very valuable for clinical research and should be stored in the EMR.

A typical hospital EMR does not provide yet a complete view on all patient data and there are many legacy systems still in use. A healthcare institution may have for example an integrated system for inpatient environment, but not for outpatient. In general, researchers (oncologists, radiologists) have their own databases for research data. They share data on collaboration basis, or when it is necessary for the patient treatment.

Several healthcare organizations with strong research focus are creating their own EMR because a large percentage of their patients are in clinical research and the EMR products commercially available do not address adequately the research-based care. They put effort into building tools and databases themselves because existing commercial solutions are suboptimal and do not fulfil their needs.

The clinical trials data is often separate from the current care data, it is maintained in a distinct system. A database with all known (e.g. institution-wide) clinical trial protocols and the eligibility criteria may also be maintained. At the moment, the research and the current care systems are not as integrated as they should and several research institutes have identified this as a problem they need to address. There are ongoing efforts to change that situation and to allow transparency between the two systems. The main obstacle to unify the research and the care databases is to get the data in an interoperable state. Currently there is also a lot of paper-based data involved in the process, and it is difficult to mine the data as it is not structured and annotated using common terminologies and it is not in a scenario or a form that users can readily access.

In healthcare organizations there is a need for a unified longitudinal view on all the patient data. Access to data from clinical practice can be beneficial for research: treatment, outcome, etc. Genomic data from clinical practice could be used in clinical research to formulate hypotheses for trials, to generate additional level of depth. Analytical tests to substantiate hypotheses can be carried out on data from clinical practice. Patient data could also be used, including genomics, as there are many more things to observe than those currently available or known. The research could use feedback from practice and also use the data in the EMR to identify eligible patients for trials. Existing data can be used to look for something else, or it can bring new information in the future when new tools may become available to interpret the data, or new signatures may be discovered.

The trend towards individualized medicine is influencing clinical trials with respect to their design, making them more adaptive. Gene association studies could be carried out based on stored genomic data to find risk factors for certain diseases. The validity of the data for genotyping should be the entire life and beyond, as it may affect the descendants of the patient. The validity of microarray data depends on many factors. Currently for genomic tests consent is required and patients decide whether to do the test or not. They are also responsible to inform their relatives when increased risk is found.

Old data (including frozen tissue samples) is very often used for identifying new research hypotheses or validating existing ones. Also for patients that received a certain therapy, if a new discovery takes place that has importance on the outcome of the patient, they would retrieve data from the system or use tissue from the tissue bank to detect the patients that would benefit from the new discovery and the new therapy option. Today’s tests may be run on tissue that was extracted five years before. If in an “old” patient case a certain marker is found that would predict that the patient may benefit from a new therapy, the patient should be notified about the new option.

A lot of genomic information needs to become persistent in the medical record: All the genes, the annotation, the association among them, the clinical covariance, the clinical variables, the context. All raw data should be preserved. Expression data (microarrays), proteomics, genotyping will become part of the wide-spread clinical practice in the future and should be stored in the EMR. The whole profile needs to be stored in order to be used for future hypotheses building and testing, e.g. to discover new subtypes. Hypotheses can be built on a cohort and another set of patients can be used for prospective validation in a clinical trial. One should be able to retrieve the matrix for expression values, together with the annotation and the clinical parameters. A query example is to search for association with clinical parameters. All the data needs to be kept for a long time.

Co-morbidities influence the therapy received, e.g. specific chemotherapy agents or Herceptin that are cardio-toxic cannot be prescribed to patients with heart problems, or when prescribed the patient needs to be individually managed (e.g. may receive a lower dose). The predisposition is also evaluated, family history, cardiac function, etc. The medical history is relevant, e.g. for drug interactions. In clinical trials relevant co-morbidities are part of the eligibility criteria recorded in the Clinical Report Forms. With known cardio-toxic treatments there are always cardiovascular evaluations planned in the protocol. In order to take into account co-morbidities, statistical methods need to be available to assess them. For co-morbidities unrelated to cancer no models exist and there is no data available.

In cancer clinical trials the re-evaluation of the treatment is carried out according to the protocol, or in the case of unexpected events (such as severe adverse reaction). The guidelines of the trial specify what and when to evaluate, when to consider abandoning the treatment and how to address side effects. In clinical practice the treatment is also followed for adverse side effects.

In all clinical trials there is a large list of eligibility criteria. Part of the clinical practice, the clinicians will interview the patients when they see them for the first time and this information will be recorded in the medical record and used later to decide whether there are suitable clinical trials for those patients.

There is a large need for information related to predictive markers and a lot of research is going on in that area. Quite often the oncologist is not sure what treatment would work best, and then access to information about the latest clinical trials is relevant.

What genomic information needs to be stored and managed in a clinical information system depends on the way the information is digested. Genotype information needs to become persistent as it has long validity. Sequencing is very robust and linear, so there is no doubt that it should be maintained. All robust data should be preserved (genotype, sub-groups, etc.). Disease-related information changes faster so it may have shorter validity. Expression profiling generates very vast information and access to large storage resources needs to be available to be able to preserve it.

The molecular pathways are useful for research. Molecular targets are focused on pathways: block or promote certain pathways. Clinicians would prefer the end result of the test (e.g. poor vs. good prognosis, responsive vs. non-responsive), but all the data should be available in an EMR. It would be useful to identify appropriate candidates to register in clinical trials, to choose groups, combine and find new information. Stored data can also be used for the formulation of new hypotheses for clinical research.

Availability of more prognostic tests will be very valuable to help stratify the patients and to decide whether they should receive chemotherapy or not. Any new test that is prospectively validated by a clinical trial should be considered to be used. MammaPrint and OncotypeDX are not used yet in large scale practice, but they will probably be introduced soon.

The volume of available data increases a lot. The discoveries concerning prognostic tests are expected to slow down in the future, as what was relatively easy to find has already been discovered. Researchers will need to focus on the more difficult questions and the underlying information becomes more important.

Raw genomic data is currently not used in the current practice hospital setting. For tests based on microarrays, only the test values or the classification are needed for the clinical decision. At the moment, only the validated test results are stored in practice, but in the future the entire arrays should be stored to be used for research or for future evaluation and treatment. However, this may differ in a small clinic which would have no knowledge to analyze that data.

All existing technologies currently used for genomic research are going to be used in clinical practice, but probably in a more simplified and cheaper form. In the future, guidelines will be necessary to specify what tests to perform and depending on the results what therapy to choose. The rate of change of guidelines will accelerate.

Standardized microarray tests may be transferred to the clinical practice after they are proven prospectively by clinical trials. Microarrays will not be the only technology used in practice. There may be small kits, RT-PCR, or larger chips with more tests. Currently, PCR is more standardized and cheaper to perform and it is also more available.

In practice simpler microarrays than those used to identify the signatures, with smaller number of genes, can be used. It is however harder to make a small chip (normalization is more difficult). The other alternative is to combine several tests on a single array, leveraging the economy of scale, and to perform them at once. The classical clinical grade will be combined with the genomic grade to provide a more accurate prognosis. In practice the validated signatures will be used. The clinicians will have to send the material to a specialized lab which will send them the result back. The result should be easy to interpret. Available commercial tests, such as MammaPrint only provide the test result and not the expression data.

To speed up the transfer of results from research to practice the dissemination should be faster. To reduce the duration of the trial, the recruitment of patients needs to be faster.

The dissemination of the results is not easy. There needs to be a way to learn about new tests, to access the tests, to be able to order them. Clinicians need to have easy access to information. Currently they have to actively search for information, e.g. by attending congresses. Traditionally, dissemination is done via establishing protocols or guidelines. Online services could play a more important role here. Currently there is individual responsibility and the physician should be eager to find out about new research results. Difficult cases do go to academic hospitals in general. Both information pull and push is important, e.g. notify clinicians when new tests are available. There are currently no mechanisms for straightforwardly transferring the best practices to wide-spread clinical practice. A system to integrate the latest guidelines as prescribed by a standard-setting organization, such as ASCO would be very useful to disseminate the best practices and improve care.

In research hospitals the established clinical guidelines are routinely used, and they are extended with criteria based on own research.

In clinical-practice-only hospitals decision support tools to guide through complex protocols are needed, including literature search for complex cases. All relevant results should be retrieved, as one would not want only the processed information from the system. It also needs to be possible to go back to the raw information when desired. An IT solution to keep information up-to-date or to provide results for complex protocols may be needed (decision support), e.g. a tool to suggest tests.

4 Clinical scenarios supporting the need for preserving genomic data in a future clinico-genomic EHR

An important observation is that current EMR and EHR solutions do not support clinical research requirements, despite the fact that the data in clinical practice is regarded as a valuable source of information for new research and for validation of results in many important cancer centers. Not being able to properly query their clinical practice data deprives the healthcare organizations of a valuable resource that could be used for improving the treatment of cancer patients. Research data is maintained in different data sources than the clinical data, where it can be properly managed and queried. This creates unnecessary duplication and inconsistencies, as patients treated in clinical trials will have data in different systems. In this context, large cancer centers see it as a necessity to develop their own EMR systems or their own additions to those systems, as being able to properly access their data from clinical practice and to use it for research will represent a huge competitive advantage for their organization. To achieve this goal, clinical and research data needs to be fully integrated.

With respect to genomic data in a future post-genomic EHR, we have identified several clinical scenarios in which the preservation of previously collected patient data is essential:

1) Genomic data is used for treating a concrete patient for
    a) Assessment of prognosis
    b) Choice of treatment
2) A new discovery enables the use of existing “old” data (accessible from the EMR) for treating a concrete patient
    a) For treating the same disease
    b) For treating other diseases

   c) For indentifying the likelihood of relapse

3) EMR-stored genomic data is used for research when increased population is necessary
   a) When carrying out cohort studies
   b) When researching rare diseases
   c) When side effects for drugs and treatment are only visible in large retrospective studies
4) Existing “old” data is used for new hypothesis building and testing
5) Existing “old” data is used for revalidation (refinement) of research results
6) New techniques for analyzing the data are applied (revalidation from another point of view)
7) Exhaustive large-scale data mining of EMR data, including genomic data (e.g. expression) to find potentially relevant correlations
8) In clinical practice, existing data is mined for quality assurance to verify compliance with guidelines, improving the process of delivery of care and preventing errors
9) In clinical research, existing data can be mined for quality assurance to verify compliance with trial protocols
   a) Mine trial data for quality assuring matching
   b) Mine differences (automatically)
   c) Merge trials, retrospective studies
10) Genomic patient data can be used to indicate potential predisposition to disease (and therefore the need for testing) for other family members
   a) Known relation between data and a certain predisposition at the time of the data collection
   b) New relation between previously stored genomic information and a certain predisposition based on new research results

5 Relevant standards

In this section we discuss several standards relevant in the context of modelling genomic data that should become accessible from a post-genomic clinical information system. These standards will be further used in our proposed model and in the description of the user scenarios.

In Figure 5.1, we present an overview of the standards that we consider will have an impact on the definition of a genomic-enabled health record system, with a special focus on the standards related to the modelling of microarray data. The horizontal layering in the figure presents how the different standards depend on each other, with one exception: the HL7 Reference Information Model layer does not relate to the layers below, it only forms the basis for the layers above. As becomes clear from the figure, we consider HL7 to be the basis of the future clinico-genomic EHR, as this standard is a large international standard and is already used extensively for data exchange within hospitals.

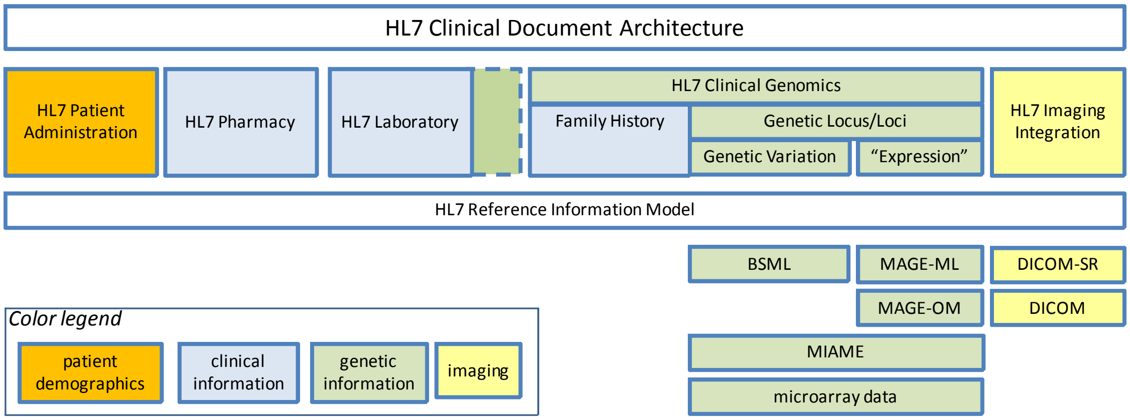


Figure 5.1 Standards that will impact the definition of a genomic-enabled health record

In the rest of this section we describe in more detail the standards we considered relevant for our initial model definition for microarray data (see Section 8).

HL7 Reference Information Model

The main aim of the HL7 messaging standard is to ensure that health information systems can communicate their information in a form which will be understood in exactly the same way by both sender and receiver. Whereas HL7 version 2 was a pure messaging standard for interoperability, version 3 (V3) not only specifies how to send a message, but also what a message can contain. To achieve this goal, V3 makes use of vocabularies and ontologies like SNOMED and LOINC.

At the basis of all HL7 V3 messages is the Reference Information Model (RIM) [1], an abstract model of the concepts which underlie healthcare information. The RIM is defined as an object-oriented model (Figure 5.1), and the following are the definitions of the six core RIM classes:

- **Act**: an action of interest that has happened, can happen, is happening, is intended to happen, or is requested/demanded to happen. An Act instance is a record of such an action.
- **Entity**: a class or specific instance of a physical thing or an organization/group of physical things capable of participating in Acts. This includes living subjects, organizations, material, and places. The Entity hierarchy encompasses human beings, organizations, living organisms, devices, pharmaceutical substances, etc.
- **Role**: establishes the roles that entities play as they participate in an Act. Each role is 'played' by one Entity (the Entity that is in the role).
- **Participation**: an association between a Role and an Act. The Participation represents the involvement of the Entity playing the Role with regard to the associated Act. A single Role may participate in multiple Acts and a single Act may have multiple participating Roles.
- **ActRelationship**: an association between a pair of Acts. This includes Act-to-Act associations such as collector/component, predecessor/successor, and cause/outcome.

The class has two associations to the Act class, one named "source" the other named "target".

- **RoleLink**: a connection between two roles expressing a dependency between those roles

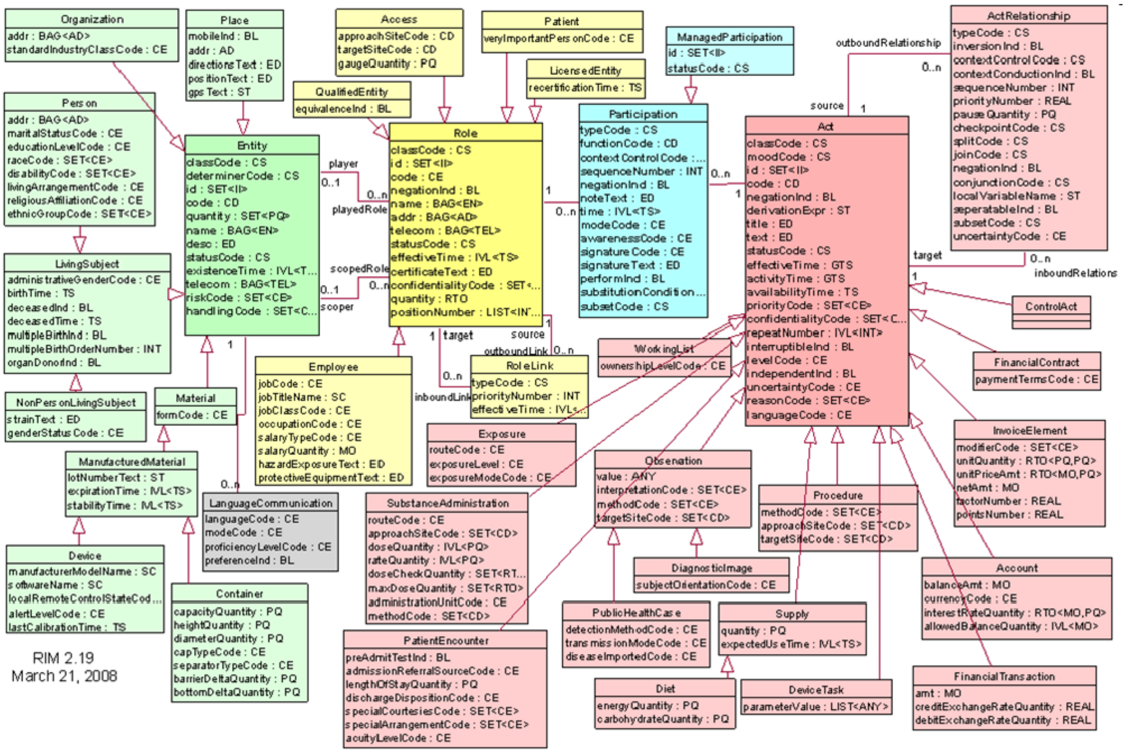


Figure 5.1 The HL7 RIM model

To render the above definitions more concrete, consider the following scenario:

> “Jane McGee visits her GP, John Smith. She complains of dizziness when standing up, therefore John decides to measure her blood pressure.”

Figure 5.2 shows an object diagram of how the information contained in this scenario ends up in a HL7 RIM-based message.

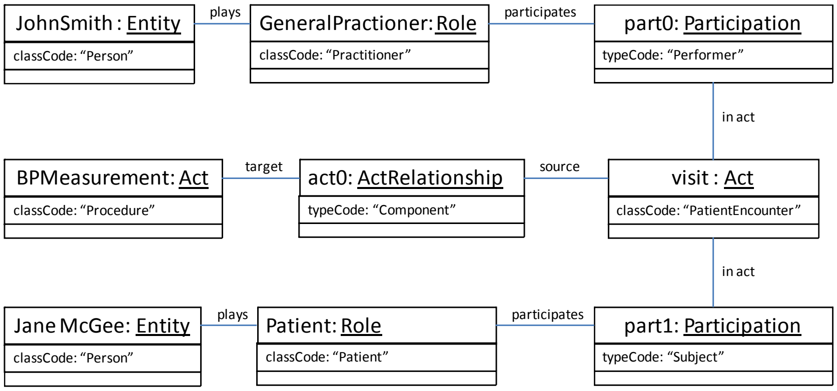


Figure 5.2 Example of an HL7 message based on the RIM model

HL7 Clinical Genomics

In the Clinical Genomics working group, HL7 V3 standards are developed that enable the exchange of interrelated clinical and personalized genomic data between interested parties [2]. Currently, the domain consists of three topics: Genotype, Genetic Variation, and Pedigree (Family History). The latter topic aims at describing a patient's pedigree based on genomic data. As such, it utilizes the models from the Genotype topic to contain the genomic data for the patient's relatives.

The Genotype topic consists of two (HL7 RIM-based) models: the Genetic Locus and the Genetic Loci. The latter model groups several genetic locus instances, e.g. in case of a genetic test of several genes. The first model describes data related to a genetic locus: the position of a particular given sequence in a genome or linkage map. The model includes sequencing and expression data, and can be linked to clinical information or phenotypes. Existing bioinformatics mark-up languages such as MAGE-ML [3] and BSML [4] are utilized to represent the raw genomic data.

The Genetic Variation topic [2] defines a model that is a constraining of the Genotype topic models. The focus of this model is on variations in the DNA of individuals, derived using methods such as SNP probes, sequencing and genotype arrays that focus on small scale genetic changes. However, gene expression analysis, e.g. based on microarray data, is not suitable for the Genetic Variation model and will be addressed (in the near future) by different models within the HL7 Clinical Genomics working group.

HL7 Clinical Document Architecture

The HL7 Clinical Document Architecture (CDA) is a document markup standard, based on the HL7 RIM, that specifies the structure and semantics of clinical documents for the purpose of exchange [5]. Examples of such clinical documents are referral notes, discharge summaries, and clinical summaries. The CDA document is defined such that it can include text, images, sounds, and other multimedia content. A CDA document is encoded in the XML language, and basically consists of a header and a (structured) body. The header classifies the document and provides information on the encounter, the patient, and other involved entities. The body contains the actual clinical report, and often is divided into nestable document sections. Each of these sections can contain a single "narrative block" (i.e., human readable content) and any number of CDA entries. These entries represent structured content and can be an envelope for information described in each of the HL7 domains, as the CDA is RIM-compliant.

MIAME (Minimum information about a microarray experiment)

Although increasingly used for gene expression data analysis at a genome-wide level, microarray technology still has the limitation of insufficient standardization for presentation and exchange of such data. The MIAME standard [6] aims at establishing a common way for recording and reporting microarray-based gene expression data, and proposes the minimum information required to ensure that microarray data can be interpreted and that the results that yield from the analysis of the data can be independently verified. The standard only defines the content and the structure of the information and does not address the actual technical format of storing and communicating the data.

MIAME has also identified the need for controlled vocabularies and ontologies for data representation in order to enable interoperability. As there is a very limited availability of controlled vocabularies, MIAME proposes a representation in lists of ‘qualifier, value, source’ triplets, which authors can use to define their own attributes (i.e. qualifiers) and provide the appropriate values and the source from which the terms were extracted.

A significant amount of context data is necessary to describe a microarray experiment because the results of such an experiment (gene expression) are only meaningful in the context of the conditions in which the experiment was run. Most microarray experiments only report relative changes in gene expression relative to a non-standardized reference, the data is normalized in different ways and is represented in non-standardized formats, and the annotation describing the data is often insufficient. All these factors make comparing data from distinct experiments very difficult. MIAME attempts to alleviate these issues by specifying the annotation necessary to properly interpret the data and the detailed description of the experiment, including the way in which the gene expression level measurements were obtained.

Next to the gene expression matrix, which contains for each gene and sample in the array the measured expression, MIAME advises to provide information about the genes that whose expression was measured and about the experimental conditions under which the samples were taken. The information required can be divided at a conceptual level into three logical parts: gene annotation, sample annotation and a gene expression matrix.

According to MIAME there are at least three levels of data relevant to a microarray experiment:

a) The raw data (scanned images)

b) The quantitative outputs from the image analysis procedure (microarray quantitation matrices)

c) The derived measurements (gene expression data).

As the transformation steps leading from the raw data to the gene expression data are not standardized it is necessary to record a detailed description of how the expression values were obtained. The gene annotation part will provide full and detailed description of each element on the array and the sample annotation will describe the biological samples and the exact conditions in which the samples were taken. In the case of commercial arrays the required information needs to be provided only once, by the array supplier and subsequently referenced by the users. This also holds for standard protocols, which need to be described only once after they are established.

According to MIAME a microarray experiment is defined as a set of one or more hybridizations, each of which relates one or more samples to one or more arrays. The minimum information required for a published microarray-based gene expression experiment includes descriptions of the following six sections:

a) **Experimental design**: the set of hybridization experiments as a whole that addresses a common biological question. This section includes a free text format description of the experiment or a link to an electronically available publication. The minimal information included in this section includes:

   a. The type of the experiment
   b. The experimental variables, including parameters or conditions tested
   c. Quality-related indicators such as usage and types of replicates
   d. The experimental relationships between the array and the sample entities.

   Each of these will be assigned unique identifiers that are cross-referenced with the information in the following sections. This information will allow the user to reconstruct the experimental design and to relate information from the other MIAME sections.

b) **Array design**: each array used and each element (spot, feature) on the array. This section provides a systematic definition of all arrays in the experiment, including the genes and their physical layout on the array. This section contains two parts:
   a. A list of the physical arrays providing for each a unique id and a reference to a particular array design
   b. The description of the array designs: a description of the array as a whole (e.g. platform type, provider and surface type), a description of each type of element or spot used (e.g. synthesized oligo-nucleotides) and a description of the specific properties of each element (e.g. the DNA sequence).

c) **Samples**: samples used, extract preparation and labelling. This section describes the biological material for which the gene expression profile is being generated and is divided into three parts which describe the source of the original sample, the technical extraction of the nucleic acids and their subsequent labelling.

d) **Hybridizations**: procedures and parameters. This section describes the laboratory conditions under which the hybridizations were carried out.

e) **Measurements**: images, quantification and specifications. This section describes the actual experimental results. It consists of three parts: the original scans of the array, the microarray quantification matrices based on image analysis and the final gene expression matrix after normalization and consolidation from possible replicates. The image data should be provided as raw scanner image files and should be accompanied by information that includes relevant scan parameters and lab protocols. The quantification matrices should be accompanied by a description of the software used, the underlying methodology (such as algorithms and statistics), all relevant parameters and the definitions of the quantifications used (e.g. mean or median intensity). The gene expression matrix (summarized information) consists of sets of gene expression levels for each sample. At this point the expression values may have been normalized, consolidated and transformed in any number of ways. Detailed specifications should be provided of all numerical calculations that were applied to the unprocessed quantifications to obtain the expression data.

f) **Normalization controls**: types, values and specifications. This section describes the normalization strategy (e.g. spiking, housekeeping genes, etc.), the normalization and quality control algorithms, the identities and locations of the array elements serving as controls, their type (e.g. spiking, normalization, negative or positive hybridization controls, etc.) and hybridization extract preparation (how the control samples are included in sample targets prior to hybridization).

MIAME requires the storage of vast amounts of information, but often the majority of information is similar for many experiments. The goal of this standard is to describe the data in sufficient detail to allow for the understanding of how conclusions were reached. This standard only refers to the conceptual content of the data and other standards, such as MAGE-OM, MAGE-ML and MAGE-TAB, needed to be defined to encode the data for standardized acquisition, storage and interoperable communication.

MAGE

While MIAME focuses on the conceptual content of the data, specifying what information is needed in order to be able to interpret and reproduce a microarray experiment, MAGE provides data exchange standards to facilitate the exchange of gene expression data. The core of MAGE is the MAGE-OM, which provides an object model for the exchange of gene expression data. MAGE also provides two data exchange formats, MAGE-ML -which provides a mark-up language- and MAGE-TAB – which provides a tabular format (which is the current recommendation).

MAGE-OM (see [1]) defines the object model for Gene Expression data and it is modelled using UML. Table 5.1 illustrates the main packages of MAGE-OM.

Table 5.1 The main packages of MAGE-OM

Package name	**Description**
BioSequence	Specifies classes that describe the sequence information for a BioSequence.
QuantitationType	This package defines the classes for quantitations, such as measured and derived signal, error, and pvalue.
ArrayDesign	Describes a microarray design that can be printed and then, in the case of gene expression, hybridized.
DesignElement	The classes of this package are the contained and referenced classes of the ArrayDesign and describe through the DesignElements what is intended to be at each location of the Array
Array	Describes the process of creating arrays from array designs.
BioMaterial	Specifies classes that describe how a BioSource is treated to obtain the BioMaterial (typically a LabeledExtract) used to create a BioAssay.
BioAssay	Specifies classes that contain information and annotation on the event of joining an Array with a BioMaterial preparation, the acquisition of images and the extraction of data on a per feature basis from those images.
BioAssayData	Specifies classes that describe the data and information and annotation on the derivation of that data
Experiment	Represents the container for a hierarchical grouping of BioAssays.
HigherLevelAnalysis	Describes the results of performing analysis on the result of the BioAssayData from an Experiment.
Protocol	Provides a relatively immutable class, Protocol, that can describe a generic laboratory procedure or analysis algorithm, for example, and an instance class, ProtocolApplication, which can describe the actual application of a protocol.

Description	The classes in this package allow a variety of references to third party annotation and direct annotation by the experimenter.
AuditAndSecurity	Specifies classes that allow tracking of changes and information on user permissions.
Measurement	The classes of this package provide utility information on the quantities of other classes to each other.
BioEvent	An abstract class representing an event that takes sources of some type to produce a target(s) of some type.

The model can express microarray designs, microarray manufacturing information, microarray experiment setup and execution information, gene expression data, and data analysis results, satisfying the MIAMI requirements. MAGE-OM tries to be generic and as complete as possible. Users typically use a subset of the provided classes and relations, which would fulfil their needs.

MAGE-ML captures MAGE-OM in an xml notation, explicit mapping rules map the MAGE-OM model to xml. Although MAGE-ML is supported in various tools as import and export format, it is a cumbersome format to use in a laboratory when no appropriate tooling or software development expertise is available.

MAGE-TAB [2] fills this gap by providing a simple format, still capturing the requirements of the MIAMI standard. MAGE-TAB is a tabular format, can be easily manipulated with various tools (even spreadsheet programs). MAGE-TAB specifies 4 types formats: I*nvestigation Description Format* (general information about the investigation), *Array Design Format* (array design description for each array type), *Sample and Data Relation Format* (relationships between samples, arrays, data and other objects), and *raw and processed data files* (data-matrix or the native formats of the data).

6 Relevant data types and technologies

Recent advances in cancer research and clinical practice have led to an explosion of information available for diagnosis, patient stratification and treatment selection of cancer patients. The central dogma of molecular biology - a framework for understanding the transfer of sequence information between sequential information-carrying biopolymers (see [9]) - leads us to the types of information that can be expected.

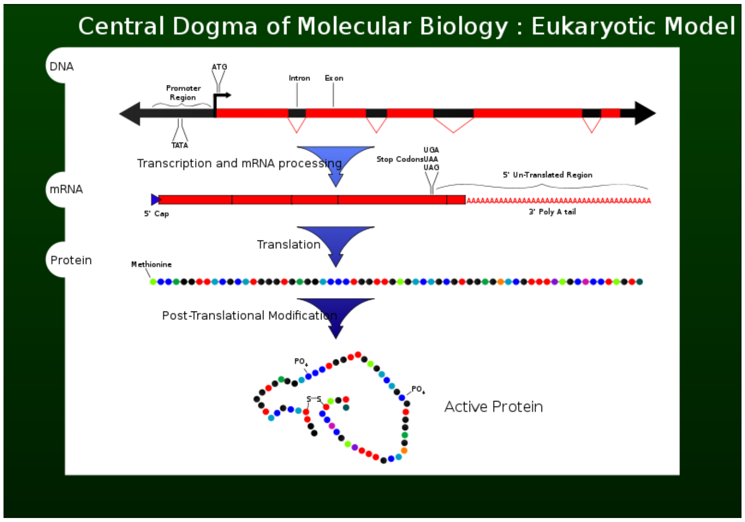


Figure 6.1 - Central dogma (see [9])

DNA is transcribed into mRNA, which is subsequently translated into proteins - the workhorses of the cell. Tests can potentially be made on every level (DNA, RNA, protein), including tests on the mechanisms that regulate the amount of expression (e.g. methylation). For example overexpression of the HER2 receptor (in breast cancer associated with worse prognosis and increased recurrence) is in clinical practice measured at the protein level (using immunohistochemistry or ELISA) or at the DNA level (Using FISH, detecting an amplification of the gene).

Difference with traditional lab tests

One could argue that these *omic tests can be treated as "ordinary" lab tests for use in an EHR. A significant difference between "standard" lab test and *omic tests is the amount of relevant data returned. For example for a microarray, each probe(set) reports on a specific mRNA sequence representing each a separate test value. In the light of the quickly evolving understanding of the molecular mechanisms of cancer, it is possible that the interpretation of test values might change (see the scenarios in Section 4), pleading for retaining the raw datasets.

Breast cancer tumor markers – clinical practice and state-of-the-art

In order to understand the types of information that might become available in a Clinico-Genomic EHR breast cancer is chosen as "model disease". Breast cancer is one of the most common type of cancer and research successes have lead to a revolution in the amount of personalized information available, characterizing tumors at the molecular level. A summary of the recommendations for use of Tumor Markers in Breast cancer by the American Society of Clinical Oncology, updated in 2007, is given in Table 6.1 (see [10]).

Table 6.1 Summary of Guideline Recommendations

Recommended Tumor Markers in Breast Cancer	
Specific Marker	2007 Recommendation

CA 15-3 and CA 27.29 to contribute to decisions regarding therapy for metastatic breast cancer

For monitoring patients with metastatic disease during active therapy, CA 27.29 or CA 15-3 can be used in conjunction with diagnostic imaging, history, and physical examination. Present data are insufficient to recommend use of CA 15-3 or CA 27.29 alone for monitoring response to treatment. However, in the absence of readily measurable disease, an increasing CA 15-3 or CA 27.29 may be used to indicate treatment failure. Caution should be used when interpreting a rising CA 27.29 or CA 15-3 level during the first 4-6 weeks of a new therapy, since spurious early rises may occur.

CEA to contribute to decisions regarding therapy for metastatic breast cancer

For monitoring patients with metastatic disease during active therapy, CEA can be used in conjunction with diagnostic imaging, history, and physical examination. Present data are insufficient to recommend use of CEA alone for monitoring response to treatment. However, in the absence of readily measurable disease, an increasing CEA may be used to indicate treatment failure. Caution should be used when interpreting a rising CEA level during the first 4-6 weeks of a new therapy, since spurious early rises may occur.

ERs and PgRs

ER and PgR should be measured on every primary invasive breast cancer and may be measured on metastatic lesions if the results would influence treatment planning. In both pre-and postmenopausal patients, steroid hormone receptor status should be used to identify patients most likely to benefit from endocrine forms of therapy in both the early breast cancer and metastatic disease settings. In patients with DCIS who are candidates for hormonal therapy, data are insufficient to recommend routine measurement of ER and PgR for therapy recommendations.

HER2 evaluation in breast cancer

HER2 expression and/or amplification should be evaluated in every primary invasive breast cancer either at the time of diagnosis or at the time of recurrence, principally to guide selection of trastuzumab in the adjuvant and/or metastatic setting. Other utilities for HER2 evaluation are also discussed separately above.

HER2 to select patients for anti-HER2–based therapy

High levels of tissue HER2 expression or *HER2* gene amplification should be used to identify patients for whom trastuzumab may be of benefit for treatment of breast cancer in the adjuvant or metastatic disease settings.

The utility of HER2 for predicting response to specific chemotherapeutic agents

Level II evidence (prospective therapeutic trials in which marker utility is a secondary study objective) suggests that overexpression of HER2 (3+ by protein or > 2.0 FISH ratio by gene amplification) identifies patients who have greater benefit from anthracycline-based adjuvant therapy. If a clinician is considering chemotherapy for a patient with HER2-positive breast cancer, it is recommended that an anthracycline be strongly considered, assuming there are no contraindications to anthracycline therapy. In the context of trastuzumab therapy, there is Level I evidence (single, high-powered, prospective, randomized, controlled trials specifically designed to test the marker or a meta-analyses of well-designed studies) that a nonanthracycline regimen may produce similar outcomes. At present, the Update Committee does not recommend that HER2 be used to guide use of taxane chemotherapy in the adjuvant setting.

uPA and PAI-1 as a marker for breast cancer

uPA/PAI-1 measured by ELISAs on a minimum of 300 mg of fresh or frozen breast cancer tissue may be used for the determination of prognosis in patients with newly diagnosed, node negative breast cancer. IHC for these markers is not accurate, and the prognostic value of ELISA using smaller tissue specimens has not been validated. Low levels of both markers are associated with a sufficiently low risk of recurrence, especially in hormone receptor–positive women who will receive adjuvant endocrine therapy, that chemotherapy will only contribute minimal additional benefit. Furthermore, CMF-based adjuvant chemotherapy provides substantial benefit, compared with observation alone, in patients with high risk of recurrence as determined by high levels of uPA and PAI-1.

Multiparameter gene expression analysis for breast cancer (Note: This topic is new to the guideline)

In newly diagnosed patients with node-negative, estrogen-receptor positive breast cancer, the Onco*type* DX assay can be used to predict the risk of recurrence in patients treated with tamoxifen. Onco*type* DX may be used to identify patients who are predicted to obtain the most therapeutic benefit from adjuvant tamoxifen and may not require adjuvant chemotherapy. In addition, patients with high recurrence scores appear to achieve relatively more benefit from adjuvant chemotherapy (specifically (C)MF) than from tamoxifen. There are insufficient data at present to comment on whether these conclusions generalize to hormonal therapies other than tamoxifen, or whether this assay applies to other chemotherapy regimens. The precise clinical

	utility and appropriate application for other multiparameter assays, such as the MammaPrint assay, the "Rotterdam Signature," and the Breast Cancer Gene Expression Ratio are under investigation.

The recommendations for use of Tumor Markers in breast cancer give an overview of biomarkers currently available for clinical practice. Especially the use of multi-gene or multi-protein assays is the focus of very active research. Table 6.2 gives an overview of recent developments. Noticeably is the large number of genes typically involved in the tests.

Table 6.2. Multi-Gene or Protein Assays in Development or Commercially Available (see [11])

Signature	Commercially available	# of Genes/Proteins	Technology
Amsterdam 70 Gene Signature* (MammaPrint)	X	70	Microarray (Agilent)
Recurrence Score* (Oncotype DX)	X	16 (21)	Realtime RT-PCR
H/ITM (HOXB13/IL17BR)*	X	2	RT-PCR
Sensitivity to Endocrine Therapy (SET)		200	Microarray
Rotterdam Signature		76	Microarray (Affymetrix)
5-protein assay		5	IHC
5-protein assay	X	5	IHC
Key 2 BC Prognostic Test		3	FISH
Invasiveness Gene Signature		186	Microarray (Affymetrix)
Wound-Response Signature		512	Microarray (Custom cDNA)
Genomic Grade		97	Microarray (Affymetrix)
p53 Signature		32	Microarray (Affymetrix)
Cancer Death Signature		11	Microarray (Affymetrix)

Technology overview

In this section, an overview of the technologies driving the multi-gene and protein assays is given.

FISH

Fluorescent in situ hybridization is a test at the DNA level and uses fluorescent probes to bind to the parts of the chromosome with which the probe has a high similarity. With a microscope, the locations of the probes are subsequently observed.

IHC

Immunohistochemistry is a test at the protein level and uses a fluorescent or stained antibody to bind to a specific antigen in a tissue section.

PCR

polymerase chain reaction is a method to amplify ("clone" or "multiply") DNA. ***RT-PCR***: Reverse Transcription PCR is a method to first convert RNA to cDNA (using reverse transcriptase), after which PCR is used to amplify the cDNA.

Realtime PCR: Realtime PCR is a method to amplify ("clone" or "multiply") DNA, during each amplification step the amplified DNA is quantified (by intercalation of fluorescent dye or the use of modified flourescent DNA oligo-nucleotide probes).

Microarray test

A microarray test is a test where a chip is used which contains an array of oligonucleotide probes. The cDNA (or cRNA) under investigation is fragmented, labeled with a fluorescent probe and flushed over the microarray. A fragment will bind to a probe when it has a high similarity. This technique can be used to measure variations in expression levels and SNP's.

Genomic variation

Besides diagnosis, patient stratification and treatment selection of cancer patients, genomic information is also used to determine predisposition for diseases. As DNA is inherited from the parents, parts that can cause a predisposition can be passed on. An example of this are the BRCA1 and BRCA2 genes, in which a mutation might lead to an increased risk for breast cancer. The most comprehensive test is the complete sequencing of these genes, as it detects all changes in the genes. To reduce costs, alternative tests only test for known, often occurring mutations at a particular location (e.g. the mutations known as 185delAG, 538insC, and 6174delT).

7 Implementation of the HL7 Genetic Locus Model

In a previous report [12], we argued that the HL7 Clinical Genomics special interest group is one of the few initiatives on the standardization of clinic-genomic information [13]. The core model in this initiative is the Genetic Locus model. This model describes data related to a genetic locus, i.e., a fixed position on a chromosome such as the position of a biomarker that may be occupied by one or more genes.

To get a better understanding of the Genetic Locus model, we decided to implement it using the VAMPIRE framework [14]. The main idea of this framework is to capture domain knowledge by means of models and to develop (parts of) applications by instantiating these models and generating the code, documentation, and other artefacts automatically.

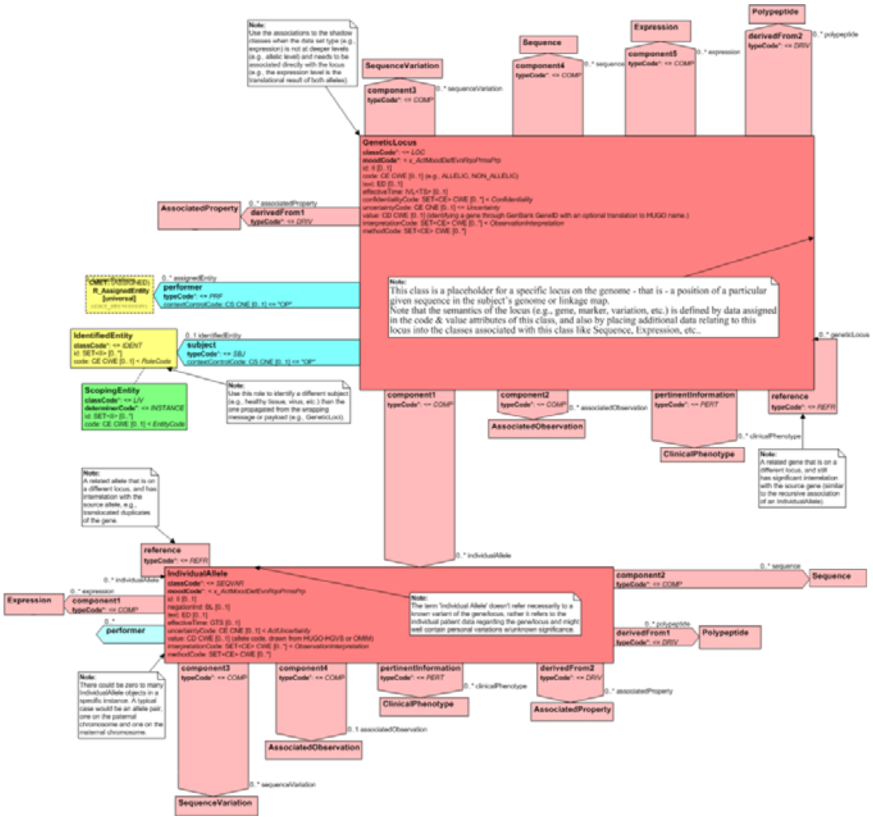
Figure 7.1 shows a fragment of the Genetic Locus domain message information model (D-MIM). A D-MIM is based on the HL7 reference information model (RIM) [1]. Although the RIM is defined in the UML notation, the D-MIM is defined using a diagramming convention developed by HL7, which enables the diagrams to be smaller and to convey more information. However, translating the model to an object-oriented implementation is not straightforward.

In the RIM, Act classes are related via the class ActRelationship, as shown in Figure 7.2. In the derived models, like the Genetic Locus model, derived classes of Act (e.g., GeneticLocus and IndividualAllele in Figure 7.1) are connected with specific instantiations of the class ActRelationship (e.g., component1). These instantiations of ActRelationship override some of the attributes with fixed values, like the source- and target Act classes. There are multiple ways to model these associations between Acts:

1. The GeneticLocus class has an ActRelationship in its list of outboundRelationships, where this ActRelationship has as source the GeneticLocus class and as target the IndividualAllele class
2. The GeneticLocus class has an additional attribute called component1, which is a subclass of ActRelationship (shown if Figure 7.2)
3. Forget about the ActRelationships, and let the outbound- and inboundRelationships attributes be subtypes of Act

We have chosen to follow option 2, mainly because it makes the domain model more explicit: “a genetic locus is linked to an individual allele”, as opposed to “a genetic locus is linked to an Act”. Secondly, it remains possible to add more information to the ActRelationship (sub-)classes, if decided at some point (which primarily is the reason why HL7 defined this class). Finally, during application development, code completion will give more hints and model checking becomes easier. The consequence of this choice is, that the outbound- and inboundRelationships attributes of the Act classes become redundant. Option 1 would require adding lots of constraints to the model, which is error-prone and difficult to maintain as the models are still under development; option 3 would not be compliant with the HL7 RIM anymore.

Figure 7.3 shows the fragment of the implemented Vampire-model corresponding to the Genetic Locus model of Figure 7.1. It can be seen that each of the ActRelationships classes of the GeneticLocus class are modeled explicitly. Using this model, we are able to accept and produce HL7 messages that contain GeneticLocus XML instances. However, this model seems to be inefficient to store or communicate microarray data, as each spot of the microarray would result in a GeneticLocus instance.



Note:
Use the associations to the shadow classes when the data set type (e.g., expression) is not at deeper levels (e.g., allelic level) and needs to be associated directly with the locus (e.g., the expression level is the translational result of both alleles).

Note:
This class is a placeholder for a specific locus on the genome - that is - a position of a particular given sequence in the subject's genome or linkage map.
Note that the semantics of the locus (e.g., gene, marker, variation, etc.) is defined by data assigned in the code & value attributes of this class, and also by placing additional data relating to this locus into the classes associated with this class like Sequence, Expression, etc..

Note:
Use this role to identify a different subject (e.g., healthy tissue, virus, etc.) than the one propagated from the wrapping message or payload (e.g., GeneticLoci).

Note:
A related gene that is on a different locus, and still has significant interrelation with the source gene (similar to the recursive association of an IndividualAllele).

Note:
A related allele that is on a different locus, and has interrelation with the source allele, e.g., translocated duplicates of the gene.

Note:
The term 'Individual Allele' doesn't refer necessarily to a known variant of the gene/locus, rather it refers to the individual patient data regarding the gene/locus and might well contain personal variations w/unknown significance.

Note:
There could be zero to many IndividualAllele objects in a specific instance. A typical case would be an allele pair, one on the paternal chromosome and one on the maternal chromosome.

Figure 7.1 Fragment of the Genetic Locus model

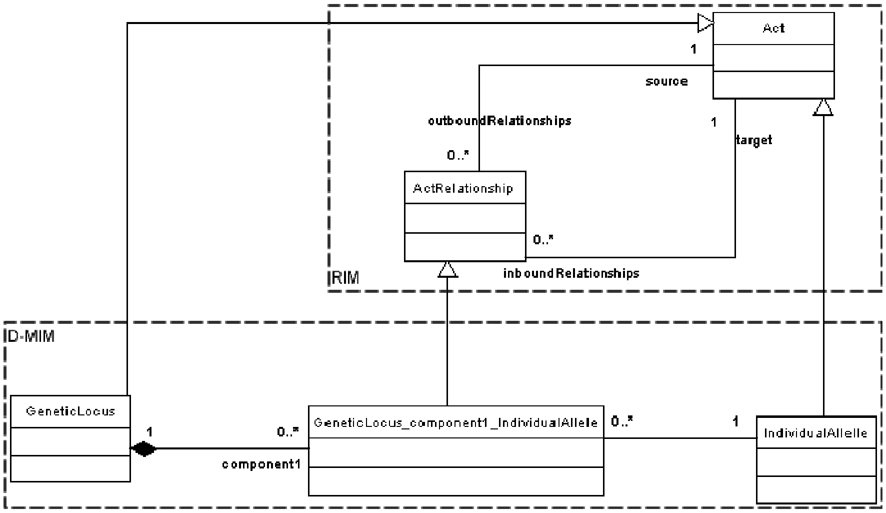


Figure 7.2 The mapping of the D-MIM notation on the RIM model

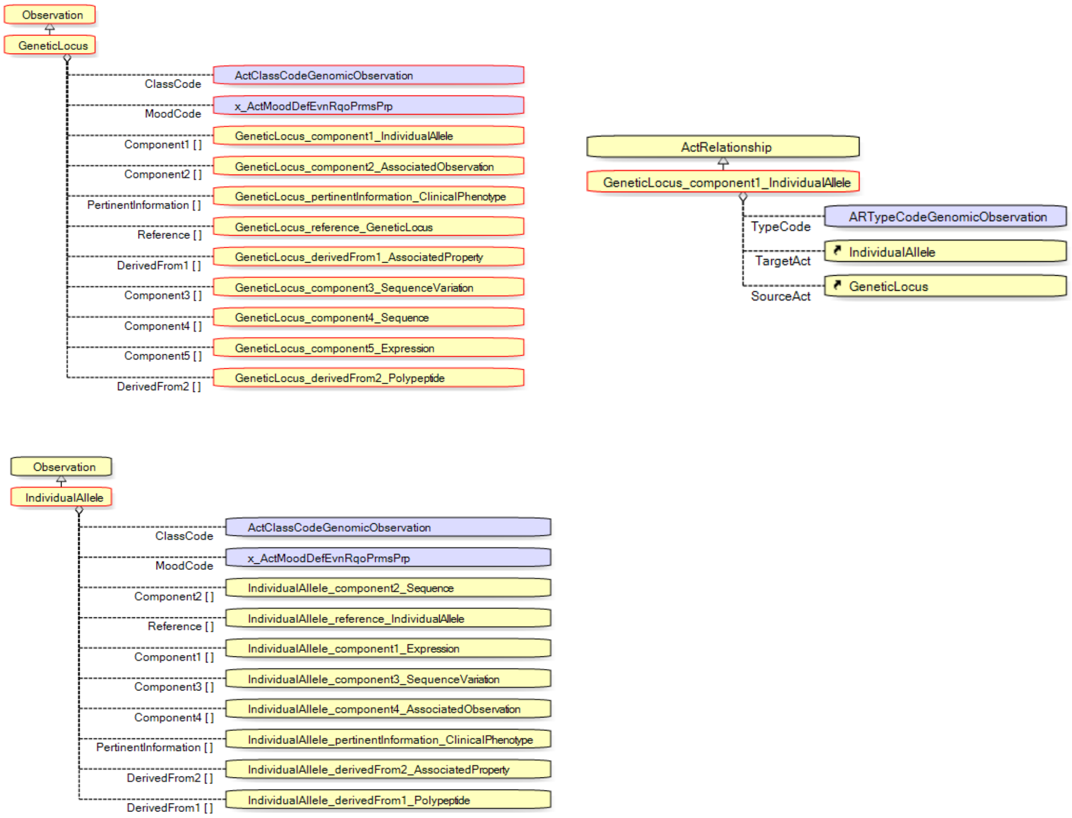

Observation
GeneticLocus
ClassCode
ActClassCodeGenomicObservation
MoodCode
x_ActMoodDefEvnRqoPrmsPrp
Component1 []
GeneticLocus_component1_IndividualAllele
Component2 []
GeneticLocus_component2_AssociatedObservation
PertinentInformation []
GeneticLocus_pertinentInformation_ClinicalPhenotype
Reference []
GeneticLocus_reference_GeneticLocus
DerivedFrom1 []
GeneticLocus_derivedFrom1_AssociatedProperty
Component3 []
GeneticLocus_component3_SequenceVariation
Component4 []
GeneticLocus_component4_Sequence
Component5 []
GeneticLocus_component5_Expression
DerivedFrom2 []
GeneticLocus_derivedFrom2_Polypeptide
ActRelationship
GeneticLocus_component1_IndividualAllele
TypeCode
ARTypeCodeGenomicObservation
TargetAct
IndividualAllele
SourceAct
GeneticLocus
Observation
IndividualAllele
ClassCode
ActClassCodeGenomicObservation
MoodCode
x_ActMoodDefEvnRqoPrmsPrp
Component2 []
IndividualAllele_component2_Sequence
Reference []
IndividualAllele_reference_IndividualAllele
Component1 []
IndividualAllele_component1_Expression
Component3 []
IndividualAllele_component3_SequenceVariation
Component4 []
IndividualAllele_component4_AssociatedObservation
PertinentInformation []
IndividualAllele_pertinentInformation_ClinicalPhenotype
DerivedFrom2 []
IndividualAllele_derivedFrom2_AssociatedProperty
DerivedFrom1 []
IndividualAllele_derivedFrom1_Polypeptide

Figure 7.3 The implementation of the Genetic Locus model in Vampire

8 Proposed high-level data model for incorporating microarray data into an EMR

Although there are several existing HL7 CDA and Clinical Genomics standards addressing the issues of communication of clinico-genomic data, we consider them not applicable for the actual storage of the data. The MIAME [6] and MAGE [3] standards provide models for the storage and exchange of microarray-based gene expression data, but are mainly tailored towards research purposes: The underlying data models are too complex and too elaborate to be directly used in an EHR or EMR. For that reason, we define an initial simplified model for the storage of genomic information in a medical record, which combines elements from the existing standards with the requirements derived from the interviews with clinical experts. The aim of this model is to provide a higher abstraction to genomic data and microarray data in particular such that it shields the user from the underlying complexity of the involved standards, while clearly identifying their use and application. The resulting high-level model is depicted in Figure 8.1.

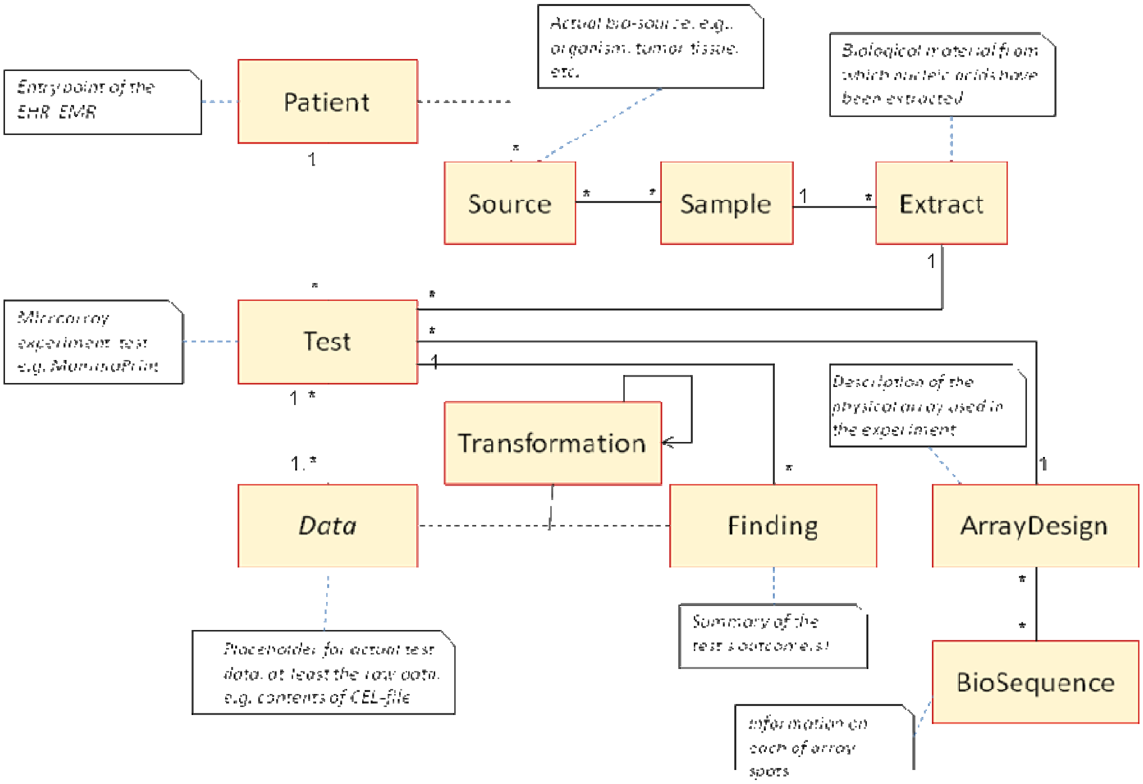

Entry point of the EHR EMR
Patient
1
Actual bio-source, e.g., organism, tumor tissue, etc.
Biological material from which nucleic acids have been extracted
*
Source
*
*
Sample
1
*
Extract
1
Microarray experiment test e.g. MammaPrint
*
Test
*
*
1
1 *
Transformation
Description of the physical array used in the experiment
1.*
*
1
Data
Finding
ArrayDesign
*
*
Placeholder for actual test data, at least the raw data, e.g. contents of CEL-file
Summary of the test outcome(s)
BioSequence
Information on each of array spots

Figure 8.1 Initial data model described as UML class diagram

In this model, we assume a patient-centric medical record, where all clinical data as well as the demographic data is encapsulated by the Patient class. The scope of the high-level model is limited to microarray data; all other relevant medical as well as administrative data, such as the purpose of ordering a microarray test or an overview of the patient's medical history, is considered to be reachable from the abstract Patient class.

Similar to other clinical tests, a doctor can order multiple microarray tests for a patient, for example the MammaPrint or Oncotype DX tests. The Test class stores all metadata on the performed test, such as the type and the purpose of the test, the date of testing, and other experimental factors.

The actual microarray used for the test is described in the ArrayDesign class. As more and more standardized microarray tests become available in the future, ArrayDesign can simply be an external reference to the design as published by the manufacturer, or it can be the complete array's blueprint, including information on each of the sequences (BioSequence class) placed on the spots of the array.

Because the microarray test only has a meaning in the context of a particular extract that was hybridized onto the array, we describe this extract, i.e., the biological material for which the genetic profile is determined, in the Extract class. In this class, it is possible to describe how the extract was derived from the physical extracted specimen, the sample (Sample class). The source from which the sample originates (e.g., from the lung or left breast) is described in the Source class.

The outcome of the test is described in the Finding class. This can be a single finding, for example whether there is a good or bad prognosis, but also multiple findings are possible, depending on the type of array used. For example, the progesterone, oestrogen and HER2 expression levels could be measured along the lines of the research described in [15], generating several findings in a single test. In this initial model we define the findings simply as key-value pairs, as each finding is test-specific and no standardization of microarray test results exists yet.

Each of the clinical findings is derived from a particular dataset of the test, by applying a specific (algorithmic) transformation(s) on that data. This can be a simple transformation, e.g., a threshold function, or a complex transformation consisting of a sequence of smaller transformations. The Transformation class is used to encapsulate each of these data processing steps.

For research purposes, it is necessary to store the actual (raw) data of the test, possibly with some additional description or annotation. The Data class is specified here as an abstract class, as there are multiple types of data that can be stored. Examples of the data to store are the intensity values and the scanned images. The MAGE standard defines how to store the data in its BioAssayData package [16]. As this package actually defines more than we intend with the abstract Data class, Figure 8.2 shows the mapping from this package to our model.

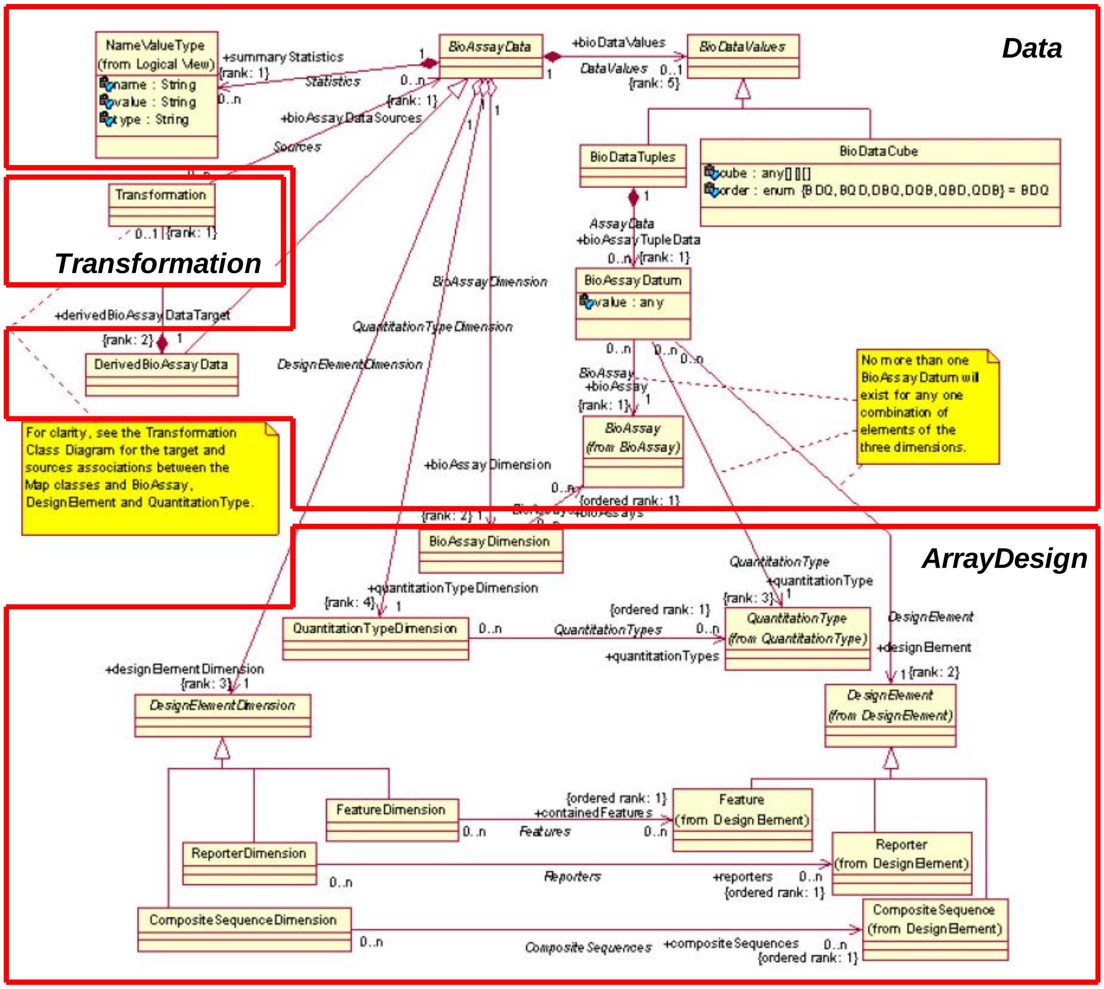


Figure 8.2 Mapping from MAGE BioAssayData package to our data model ([MAGE_spec])

A different approach for storing the actual data forms the core of the caArray approach [15]. In that data management system, which was originally based on the MAGE model, a more serialized way of storing the data is chosen. According to caArray, this simplifies the model and makes the data storage more efficient. Figure 8.3 shows the mapping from that approach to our high-level model.

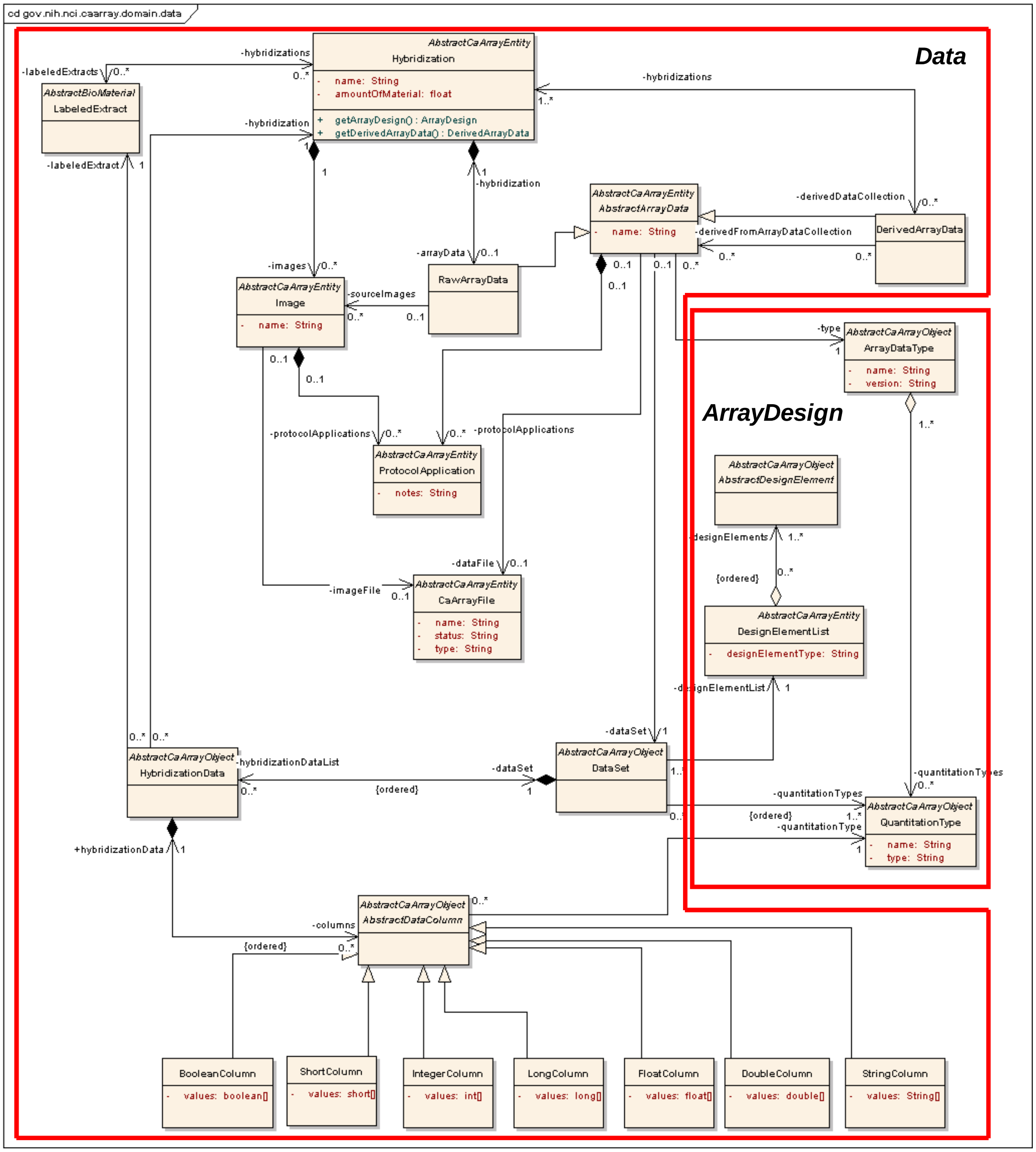


Figure 8.3 Mapping from the caArray Data classes to our data model ([17])

Please note that, there is one important prerequisite to enable the deployment and use of a clinico-genomic HER that also supports clinical research: The microarray test manufacturers, as well as the labs doing the actual testing, will have to provide the raw data to the health record system, such that they can be stored and re-used later either for the same patient (different findings) or for research purposes, like cohort studies. In the current practice, this is not the case: the only results clinical practice gets back for a microarray test are the findings of the test.

Scenarios model Walkthrough

In Section 4 we identified several clinically relevant scenarios which pose their own requirements on the data model. In this section we revisit these scenarios mapping them onto our initial data model, identifying the relevant classes and their role in a particular scenario. The classes that are used in a particular scenario are highlighted by rounded rectangles. We adopted the following color convention: Orange color indicates that we refer to previously created class instances (e.g. previously stored data); green color indicates newly created instances (e.g. a new test or transformation). In some scenarios there might be need for pseudo anonymization of patient data, this is indicated by red color.

1) Genomic data is used for treating a concrete patient for
a) Assessment of prognosis
b) Choice of treatment

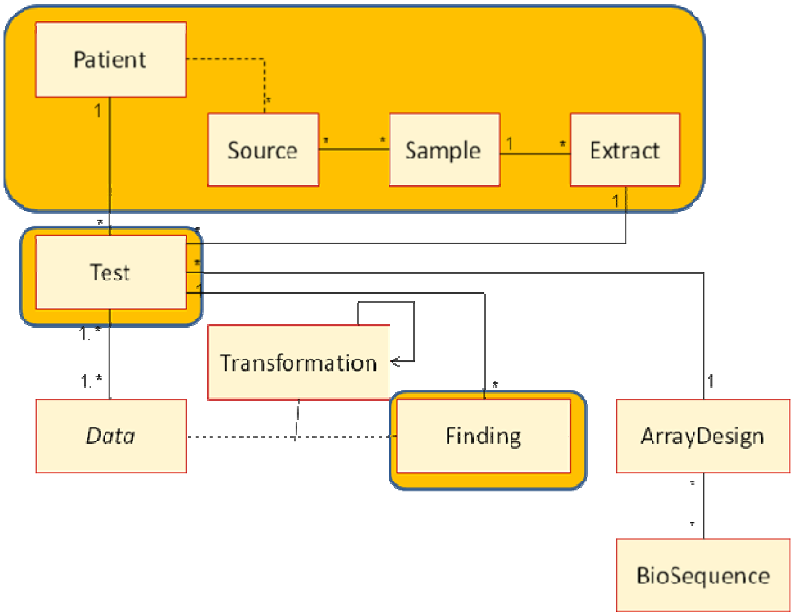


Figure 8.2 Model walkthrough for Scenario 1

In this scenario we assume that an available genomic test (e.g. MammaPrint) is ordered for a particular patient in order to answer a particular clinical question (e.g. breast cancer prognosis assessment). This scenario might be considered the simplest use case from the model perspective because it does not directly necessitate the permanent storage of raw data (unless combined with other use cases). Classes that are required (highlighted by rounded rectangles) in this use case include Patient, Source, Sample, and Extract as well as the actual Test and Finding classes. The Patient class serves both as a placeholder for patient demographics as well as an encapsulation of the non-genomic data that was collected for the patient (lab results, imaging, clinical reports, etc.). This class also contains the actual order of the required genomic test. The Test class indicates the type of the genomic test that was performed (e.g. MammaPrint). The Finding class contains the conclusions of the performed genomic test (e.g., good prognosis with respect to disease free survival).

It is worth noting that this is the only use case which assumes the physical presence of the (biopsy) sample and subsequently the extract which will be placed on the actual microarray. All other use cases require only the metadata describing these items.

2) **A new discovery enables the use of existing "old" data (accessible from the EMR) for treating a concrete patient**
   a) **For treating the same disease**
   b) **For treating other diseases**
   c) **For indentifying the likelihood of relapse**

In this scenario, a new discovery is made which refers to the same genomic data that has been already collected in the past for a previously ordered test. The relevant patient clinical data is accessed via the Patient class. Classes Source, Sample and Extract are only accessed to verify the metadata about the physical entities they represent – if the metadata conforms to the newly designed test, there is no need to redo the sample extraction/biopsy. The newly designed test represented by the Test class (highlighted green) can access the previously stored data via the Data class (highlighted by an orange rounded rectangle). The new test introduces a sequence of new transformations represented by the Transformation class (marked green), which ultimately produces a new finding captured by the Finding class (highlighted green). In order to be able to understand and process the stored data, in some cases the microarray design of the previous test needs to be consulted.

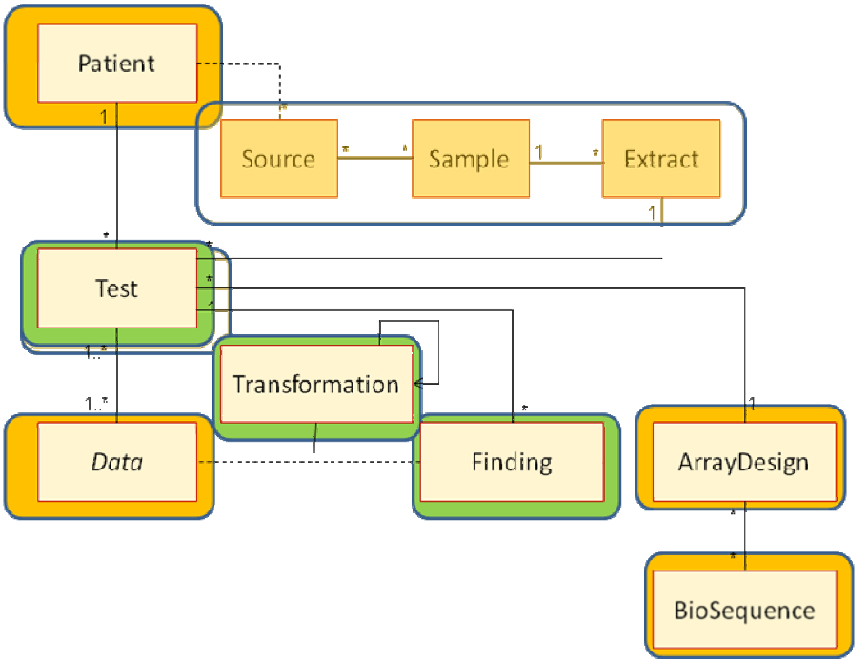


Figure 8.3 Model walkthrough for Scenario 2

3) **New techniques for analyzing the data are applied (revalidation from another point of view)**

This scenario is based on the fact that the state of the art of microarray processing is evolving. For instance, a new normalization technique can improve the signal to noise ratio and can be re-applied on already collected data, potentially yielding a change in the finding. This scenario is similar to the previous one, the main difference being that here we do not deal with a new genomic discovery (new test), instead the main focus is on improving an existing test by refining its transformations. Also, the value of the resulting finding can be different from the one stored previously, but the type of this finding is assumed to be the same as the test itself still aims at answering the same clinical question, while in the previous scenario the type of finding might change completely.

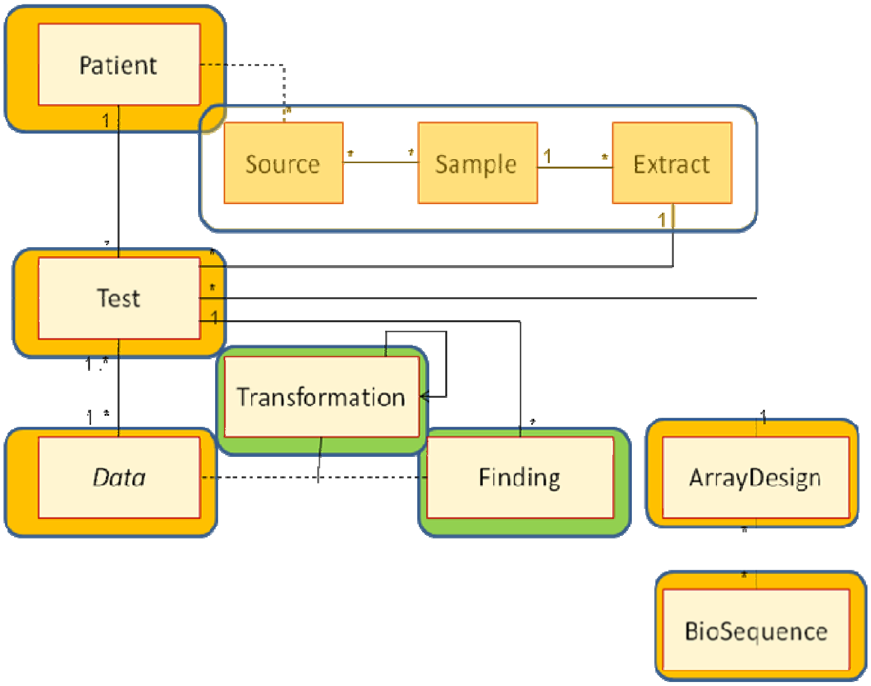


Figure 8.4 Model walkthrough for Scenario 3

4) **EMR-stored genomic data is used for research when increased population is necessary**
   a) **When carrying out cohort studies**
   b) **When researching rare diseases**

c) **When side effects for drugs and treatment are only visible in large retrospective studies**

5) **Existing “old” data is used for new hypothesis building and testing**

6) **Existing “old” data is used for revalidation (refinement) of research results**

From the data model point of view, the above scenarios can be grouped into one as they all interact with the data model in a similar fashion. They all require (pseudo) anonymised access to patient data. This usually requires the change or removal of the patient demographic data or some parts thereof. These use cases assume access to a multitude of patient records and stored genomic data in order to be able to derive statistically significant results. These results are not reflected directly in terms of a new finding for a particular patient (until a new genomic test is devised, which is covered in use case 2), but they aim at advancing clinico-genomic research. These use cases typically involve access to raw genomic data, performing their research-tailored data transformations which aim at new discoveries. Therefore, the original transformations and findings associated with the original test are not of high interest here.

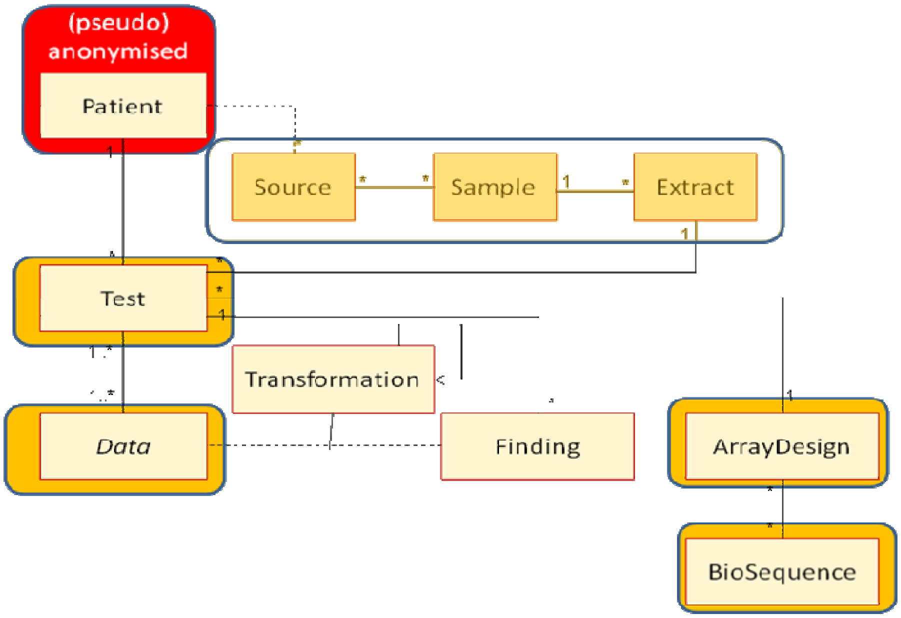


Figure 8.5 Model walkthrough for Scenario 4, 5, 6

7) **Exhaustive large-scale data mining of EMR data, including genomic data (e.g. expression) to find potentially relevant correlations**

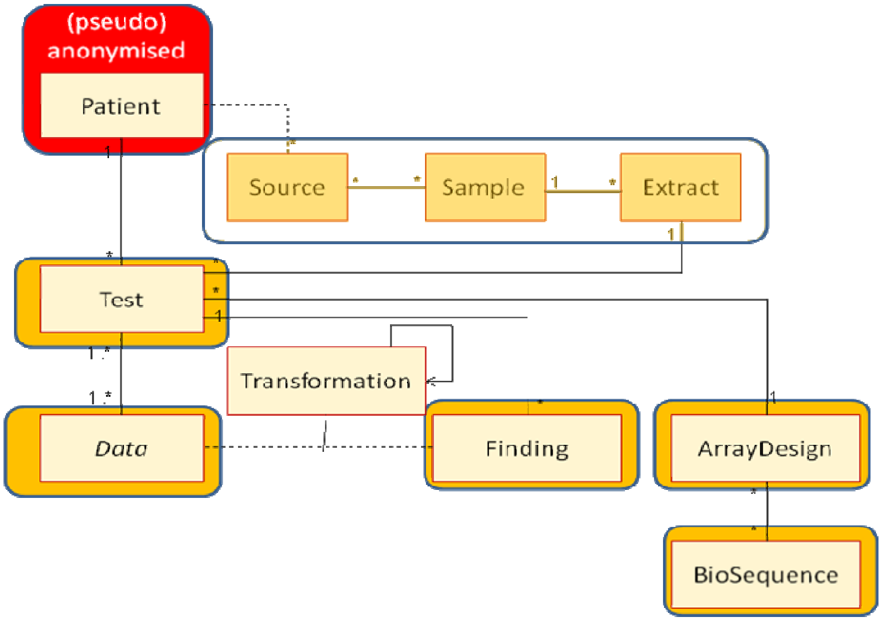


Figure 8.6 Model walkthrough for Scenario 7

This scenario assumes a large-scale mining effort on all available data, both genomic and clinical. Given a large population sample for mining, new previously unknown correlations among various clinical parameters, demographic data and genetic parameters can be discovered. Access to raw genomic data is required that allows the data mining process to treat (if necessary) even an individual gene as an independent variable. Findings are also considered as relevant clinical data that can potentially contribute to new correlations. Note that this would typically be a large-scale research experiment and the patient data needs to be properly anonymised.

8) In clinical practice, existing data can be mined for quality assurance to verify compliance with guidelines, improving the process of delivery of care and prevent errors

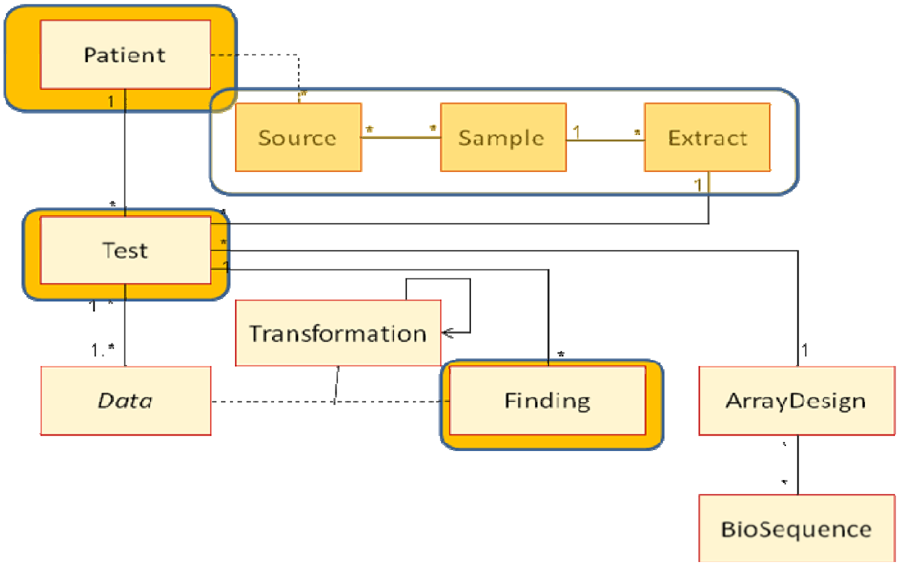


Figure 8.7 Model walkthrough for Scenario 8

In this scenario we envision a retrospective audit of performed procedures for a particular patient assuring compliance to a particular protocol as described in the clinical guidelines. In this context it is important to have access to the clinical data represented by the Patient class, to the information about the biopsy source and about the sample, to the description of the genomic test which was performed as well as to the finding obtained. The raw data is not accessed in this use case, as from the clinical guidelines perspective the genomic test is considered a black box.

9) In clinical research, existing data can be mined for quality assurance to verify compliance with trial protocols

a) Mine trial data for quality assuring matching

b) Mine differences (automatically)

c) Merge trials, retrospective studies

This is a clinical research equivalent of Scenario 8. Unlike in the previous case, here we assume that all parts of the data can be scrutinized including the raw data and the transformations performed. All classes in the data model can be accessed for validation purposes. Patient data is assumed to be properly anonymised.

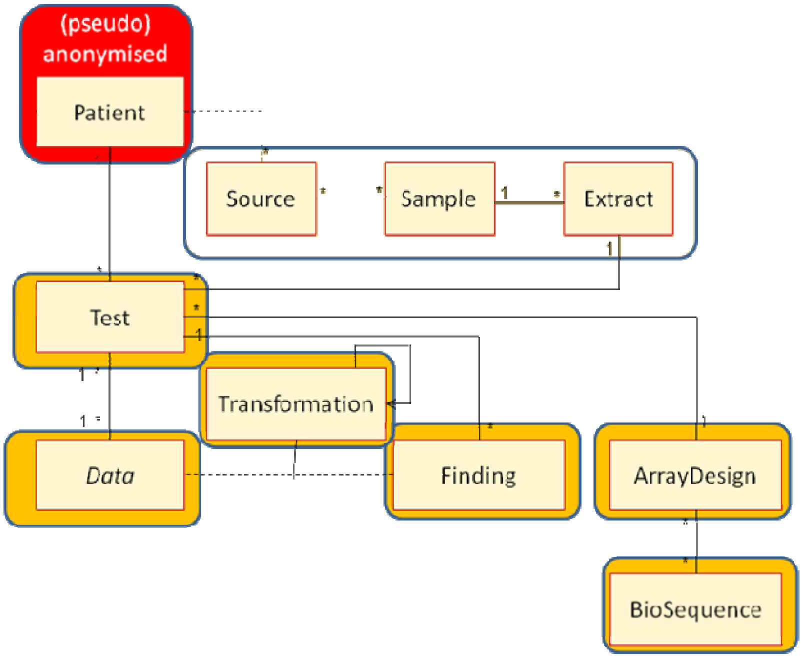


Figure 8.8 Model walkthrough for Scenario 9

10) Genomic patient data can be used to indicate potential predisposition to disease (and therefore the need of testing) for other family members

a) Known relation between data and a certain predisposition at the time of the data collection

b) New relation between previously stored genomic information and a certain predisposition based on new research results

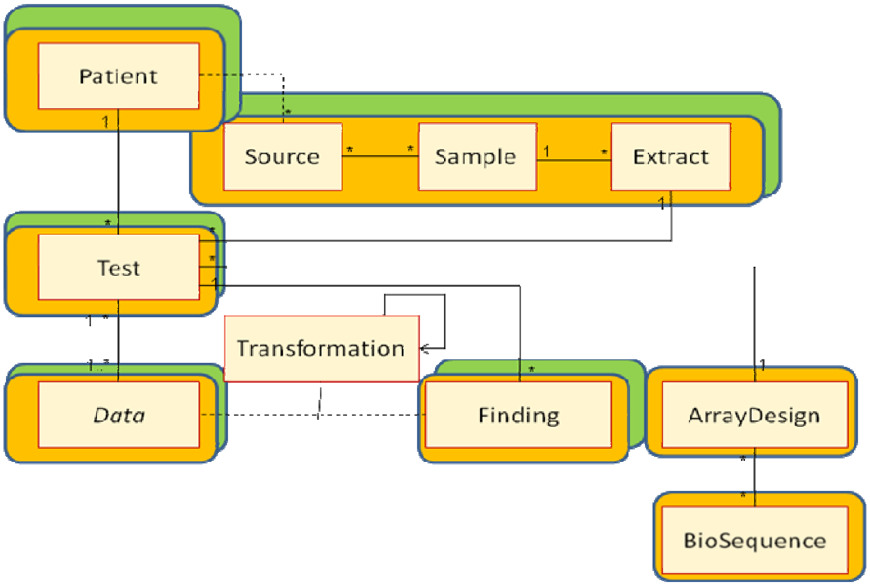


Figure 8.9 Model walkthrough for Scenario 10

In this scenario available genomic data and/or test findings of one patient can trigger a test order or a test recommendation for patient's direct relatives. For instance in case of a detected BRCA1 mutation, the patient's relatives may be recommended to undergo the same genetic test. In this use case we actually have two (or more) instances of the Patient class with their own test, data, and findings.

9 Conclusions

With the new discoveries in the cancer research, increasing amounts of genomic data are starting to be used in the context of the cancer patient management and should become part of the patient record. For clinical research the data collected in the clinical practice is a valuable resource allowing for new hypotheses generation and testing, cohort studies and research in rare diseases and selection of patients for clinical trials. In this context, storing all genomic information available and not only the processed test results, even when that information is not directly used in the care-providing process, is meaningful.

As the costs of generating genomic data, such as microarray-based gene expression, and extracting information from that data is still quite high it does not make economic sense to discard all the collected data and to preserve only the test result instead of storing and re-using it, especially that it is assumed that there is much more information in the data than what can be obtained in a single test or experiment. New standards for storing and exchanging genomic data such as MIAME and MAGE also require that all data should be preserved and annotated, including the raw microarray image. Of course, MIAME's and MAGE's main focus is research, but the reasoning behind the storage and full annotation of all genomic data is precisely based on the fact that the data encapsulates information far beyond the scope of a single experiment and should be preserved and shared. In this context, there is no reason to miss on such a valuable source of data which is represented by clinical practice. Many research-focused healthcare organizations have already identified this opportunity, and they invest in infrastructure to support this approach.

Also, for the clinical practice preserving the raw genomic data is valuable, as based on new validated discoveries from clinical research they could run tomorrow's tests on yesterday's data for the benefit of their patients.

Of course, at this point a genomic-enabled EMR is less relevant for a practice-only healthcare organization where genomic data is not yet used and often there is not enough knowledge to analyse this type of data. However, when genomic tests will be commonly used, such an organization could see it relevant to preserve and fully annotate (according to existing standards) its genomic data to provide it for research (e.g. at a cost).

Several microarray-based genomic tests are currently commercially available. When these tests are ordered only the test result is returned (e.g. good or bad prognosis). To support the use of the microarray data according to the scenarios we have defined, the raw file and/or the expression matrix should be provided as well. As it becomes increasingly clear that genomic data can generate much more knowledge than what can be extracted in a single experiment or test and the understanding of the expression data evolves we can expect that this process will change in the future.

The commercially available EMR and EHR systems do not currently take into account the requirements of clinical research and they do not enable a proper integration of data from both clinical research and clinical practice. They also do not provide support for storing, managing and sharing genomic data. On the other hand, the research world addresses these challenges by building standards and investing in infrastructure.

In this document we describe several scenarios supporting the need for a genomic-enabled clinical information system and propose an initial high-level data model for genomic data in a future system, based on established standards. As some of the standards currently have a research focus and are very complex, in our high-level model we propose a simplified solution that is suitable to be used in the context of a healthcare organization.

10 Bibliography

[1] HL7 Reference Information Model, Specification available from http://www.hl7.org/v3ballot/html/infrastructure/rim/rim.htm

[2] HL7 Clinical Genomics Domain, http://www.hl7.org/v3ballot/html/domains/uvcg/uvcg.htm

[3] MicroArray Gene Expression, http://www.mged.org/Workgroups/MAGE/introduction.html

[4] Bioinformatics Sequence Markup Language, http://xml.coverpages.org/bsml.html

[5] R.H. Dolin et al., “HL7 Clinical Document Architecture, Release 2”, J. Am. Med. Inform. Assoc, Vol. 13(1):30-39, Jan/Feb. 2006

[6] A. Brazma et al., “Minimum information about a microarray experiment (MIAME) – toward standards for microarray data”, Nature Genetics, vol. 29, 365-371, 2001

[7] Gene Expression Specification, Object Management Group, http://www.omg.org/cgi-bin/doc?formal/03-02-03

[8] T.F. Rayner et al., “A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB”, BMC Bioinformatics, vol. 7, 2006

[9] http://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology

[10] L. Harris et al., American Society of Clinical Oncology 2007 Update of Recommendations for the Use of Tumor Markers in Breast Cancer, http://tinyurl.com/5fym4c

[11] L. van 't Veer, “Clinical use of prognostic expression classifiers”, The Netherlands Cancer Institute, http://tinyurl.com/6gy6r3

[12] A. Tesanovic et al., “Deliverable 5.3: Clinico-Genomic Electronic Health Record”, May 2008

[13] A. Shabo and D. Dotan, "The Seventh Layer of the Clinical-genomics Information Infrastructure", *IBM Systems Journal*, vol. 46(1):57-67, 2007. Available at http://www.research.ibm.com/journal/sj/461/shabo.html

[14] H. Jonkers, M. Stoucken and R. Vdovjak, “Bootstrapping Domain-Specific Model-Driven Software Development within Philips”, Proc. 6th OOPSLA Workshop on Domain-Specific Modeling, Oct. 2006, Portland, Oregon, USA

[15] P. Roepman et al., “Microarray-based readout of ER, PR, and HER2 expression in breast cancer tissue”, Breast Cancer Symposium, 2008

[16] Gene Expression Specification, Version 1.1, October 2003

[17] caArray 2.1 Technical Guide, National Cancer Institute, Center for Bioinformatics, August 2008. Available from http://caarray.nci.nih.gov/