Clinico-genomic electronic health record: state-of-the-art review and initial specification

Project Number:	FP6-2005-IST-026996
Deliverable id:	D 5.3
Deliverable name:	Clinico-genomic electronic health record: state of the art review and initial specification
Submission Date:	22/02/2008

Information Society Technologies

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 5.4
Document name:	Clinico-genomic electronic health record: state of the art review and initial specification
Document type (PU, INT, RE)	RE
Version:	7.0
Submission date:	07.05.2008
Editors: Email: Organisation:	Aleksandra Tesanovic aleksandra.tesanovic@philips.com Anca Bucur anca.bucur@philips.com Philips Research

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

ABSTRACT:

Scientific advances in the genomic area have changed the concept of cancer diagnosis and treatment. Gene expression analysis alone has had a profound impact on cancer classification and diagnosis, and has contributed to the development of new targeted treatment approaches. It is therefore to be expected that genomic information will play a decisive role during the cancer care cycle of a patient, including screening, diagnosis, treatment, and follow-up. For this reason it is of essence to know (i) what genomics information should be used in each stage, and (ii) in what way that information could be used for targeted cancer care over the care cycle efficiently and meaningfully. Of course, the complexity of the domain does not allow for a definite answer to either of the two questions. This report investigates how to addresses these two challenges based on the current knowledge in the domain.

Our goal is to elaborate a specification and a data model for the genomic information to be accessible from a future EHR/EMR system, with focus on the oncology domain and from the

ACGT perspective. To this end, we need to investigate the technology, the state-of-practice and the state-of-research in these areas, based on literature study and on our experiences within ACGT (and the ACGT clinical trials), and to envision how the research results will be translated and used into the clinical practice, what is the relevant genomic information that is most likely to be used in the future clinical practice, what technologies are the most promising, and what systems need to be developed and deployed to allow for this evolution.

In this report we also exemplify how the genomic information could be used for cancer classification and therapy, risk factor assessment, and phamacogenomics. We show that an electronic health record (EHR) has an increasingly important role in longitudinal management of patient's health information, and as such will have a prominent position to support the cancer care cycle of a patient. In that context, an EHR constitutes for most clinicians the primary way of access to healthcare information, and in the future that may include genomic information. We define a *clinico-genomic EHR* as an EHR that besides the usual types of data also provides access to relevant information regarding genomic biomarkers, pathways, and genes that a healthcare professional may need for targeted diagnosis and personalized treatment.

We study current EHR standards, available implementations of patient record management systems (CPR, EMR, EHR, PHR, etc.), and their adoption, to conclude that the task of extending an EHR system with genomic data should be approached in a generic way, as the solution should apply to a variety of system implementations and approaches. The focus therefore should be on what type of data needs to be integrated in a future system and on how that data should be modeled taking into account that this should apply to various existing systems.

KEYWORD LIST: Electronic health record, genomic information, genes, pathways, DNA, HL7, *open*EHR

MODIFICATION CONTROL			
Version	Date	Status	Author
1.0	16/06/07	EHR state of the art survey	Aleksandra Tesanovic
2.0	07/09/07	Revision of EHR state of the art	Aleksandra Tesanovic, Anca Bucur, Lefteris Koumakis
3.0	06/01/08	EHR scope	Anca Bucur
4.0	06/02/08	Genomic information	Dimitris Kafetsopoulos, Anca Bucur, Andreas Persidis
5.0	21/02/08	Conclusions	Anca Bucur, Lefteris Koumakis
6.0	22/02/08	Abstract, introduction, references, paper flow	Aleksandra Tesanovic, Anca Bucur
7.0	07/05/08	Final version	Anca Bucur

List of Contributors

- Aleksandra Tesanovic, Philips Research
- Anca Bucur, Philips Research
- Dimitris Kafetzopoulos, FORTH
- Lefteris Koumakis, FORTH
- Andreas Persidis, BioVista

Contents

1. Introduction

The influence of genes on cancer development and predisposition has been a subject of extensive research effort over the years. For different cancer types a number of relevant genes and related biological processes have been identified. As an example, for breast cancer some studies show that 50 to 85 percent of women carrying the breast cancer susceptibility genes (BRCA) mutation will develop breast cancer by age 70 [2]. The expression of certain genes is also found important in determining the clinical activity of breast tumors [2]. Prominent examples include estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), etc.

The knowledge about the relevant genes and their role in biological processes responsible for cancer is steadily increasing. With that, so is their use in clinical practice for targeted diagnosis and treatment. In 2006 the Food and Drug Administration (FDA) approved the first targeted therapy for use in early-stage HER2-positive breast cancers [3]. The drug, trastuzumab (Herceptin), is claimed to offer hope of a cancer-free future to thousands of women diagnosed worldwide each year with HER2-positive early-stage breast cancer (Herceptin was previously only approved for women with a metastatic HER2-positive breast cancer).

Given the benefits and the potential of using genomic information for cancer diagnosis and treatment, it is to be expected that this type of information will become a valuable part of our medical health records in the near future, as our demographics, health history, medical images, lab data and medications currently are. However, to assess what genomic information should become part of an EHR, the following questions need to be answered:

1. What is the relevant genomic information that is likely to be used in clinical practice?
2. How can this information be easily accessed, managed, and used efficiently in everyday clinical practice for the cancer care?

An electronic health record (EHR) shows promise to become a primary access system to clinical and genomics information for most care givers, decision support systems, clinical researchers, and other healthcare stakeholders (see also Figure 1). Currently an EHR is playing a role in supporting the overall care cycle of a patient [4] as it provides means for accessing all relevant medical data of a patient, including [3]:

- observations (about what has occurred), opinions (what should occur), and care plans (how to ensure what should occur in the future);
- all clinical data, including diagnostic and other test data;
- an abstraction of specialized data, e.g., standard interfaces to images, guidelines, or clinical decision support; and
- the master patient index, workflow support, controlled medical vocabulary, order entry, computerized provider order entry, pharmacy, and clinical documentation applications.

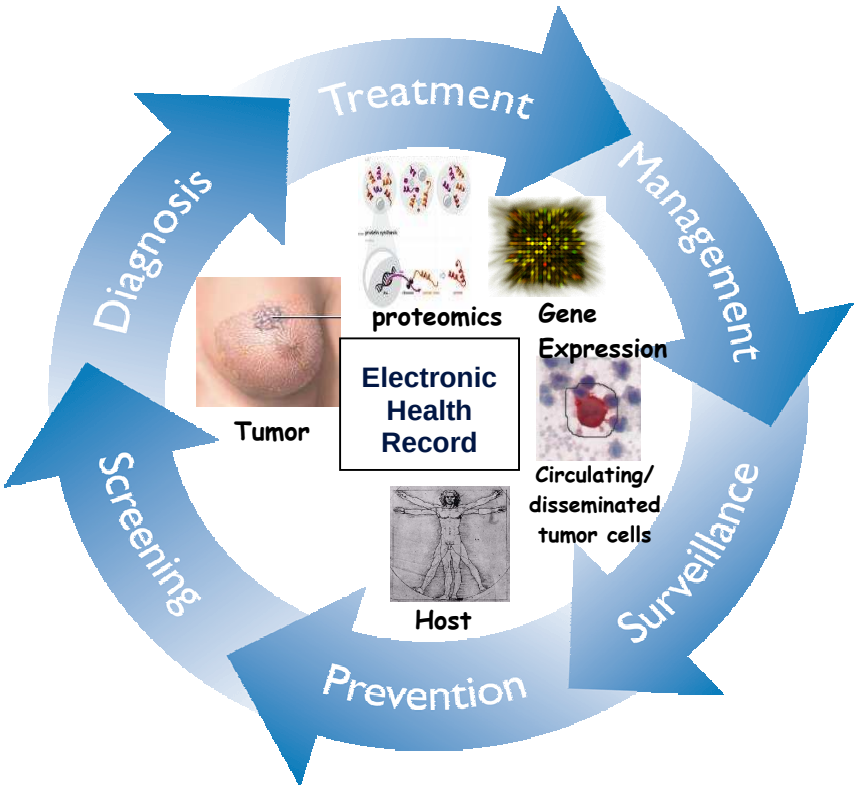


Figure 1: Care cycle for the breast cancer, where an EHR plays a pivotal role in integrating and representing genomic information needed for personalized diagnosis and treatment.

The key properties of an EHR are as follows. An EHR is:

- Patient centered, implying that an EHR contains the patient-specific information, i.e., the records are not anonymized or summarized for health research, and ideally include information relevant to all kinds of care givers, including allied health, and emergency services as well as patients themselves. This is in contrast to provider-centered or purely episodic records.
- Health related, implying that all health-related information for a patient is stored in an EHR. This includes not only data from the healthcare providers, but also other health-related information provided by the patient, such as off-the-shelf medicine, nutrition, sport activities, etc.
- Longitudinal, implying that an EHR is envisioned to contain information from birth until death, and is therefore not restricted to any particular episode of care.
- Cross-enterprise, implying that an EHR contains the information that can be created (and used) by different health providers.

However, if EHRs are to become the ultimate vehicle for individualized medical treatments there is a clear necessity to ensure that also the appropriate genomic information is represented there. Taking into account the large body of research, standardization and software development in the EHR area, our goal is not to duplicate that work or to propose new standards. Based on our experiences within ACGT, and the ACGT clinical trials, we aim to identify relevant genomic information that should be part of an EHR of the future.

With this goal in mind we organized the report as follows. In chapter 2 we discuss personalized medicine and show its importance for medicine in general, and cancer care in particular. The basis for personalized medicine is genomic information. We therefore

highlight genomic information that would be of interest for targeted care of cancer patients. We believe that enabling targeted patient treatment requires a large effort, to integrate vast amount of information at all levels (clinical, imaging, genomics, etc.), and that the EHR needs to be part of these developments. It is thus necessary to ensure that the clinically relevant genomic information is represented in the EHR and that it is properly structured and managed. Therefore, in chapter 3 we introduce the concept of an EHR and outline its scope. In this context and for this report, we denote an EHR that can access relevant genomic information to be used by healthcare professionals for targeted diagnosis and personalized treatment as a *clinico-genomic EHR*. Currently a number of EHR standards exist and we elaborate on the most important ones. In chapter 4 we give an example of a possible approach to building a clinico-genomic EHR and of services that could be provided using the genomic information. In chapter 5 we present our conclusions from this exploratory study and future research steps.

Although our target is to capture genomic information for oncology in general, in this report we often use breast cancer as the prominent example. Further in the report we omit detailed definitions of well-known terms in the domain. Interested readers can find these in Appendix 2.

2. Clinico-genomic information

In this chapter we first introduce and define personalized medicine and its potential impact. We then identify relevant genes that are likely to be used in personalized medicine for cancer. In that context we use the BioLab Experiment Assistant (BEA) tool [5] to exhaustively search the research literature for genes and biological processes identified as relevant for oncology. From that list, based on discussions with the clinical and the bioinformatics partners in ACGT, we emphasize the genes involved in the ACGT trials. Finally, we give examples of the genomics information in action, by showing its application for phamacogenomics, gene expression profiling of diseased tissue, and risk factor assessment.

2.1. Personalized medicine: a vision of the future

The traditional way of treating cancer patients is to prescribe the conventional therapy to all patients. It is has been shown, however, that patients respond very differently to the conventional therapies. Some patients have severe, life-threatening side effects, other respond to therapy as desired, while some do not even respond. While a drug may shrink a tumor for the responsive patient, it may have no effect for the non-responsive patient. Patients that do not respond to therapy or have toxic responses do not fit the standard drug therapies and need to be treated differently. Similarly, conventional drug doses can only work well with the patients that are not pre-disposed to the toxicity.

Personalized medicine is a term that denotes the possibility to give "*the right treatment at the right dose for the right person at the right time*". The promise of personalized medicine is essential for cancer diagnosis, treatment, and follow-up as it enables therapies to be tailored to each patient based on their genetic makeup and other relevant medical measurements. The ultimate aim of personalized medicine is to select the best treatment and determine the right dosage for every patient based on factors such as the patient's unique physiology and genetic profile, the physiology and the molecular biology of the disease as well as the patient's ability to metabolize particular drugs [6].

In general, expected benefits of personalized medicine include [7]:

- earlier detection of a disease, at a time when treatment is more effective,
- selection of optimal therapy and reduction of trial-and-error prescription,
- reduction of adverse drug reactions,
- increased patient compliance,
- drug discovery,
- reduction of time, cost and failure rate of clinical trials,
- avoidance of withdrawal or marketed drugs,
- shift the emphasis in medicine from reaction to prevention, and
- reduction in the overall cost of healthcare.

In the last decade, scientific advances in the genomic technology have already changed the concept of diagnosis and treatment of diseases, going toward personalization. Developments like DNA sequencing, gene expression analysis, genotyping, SNP

mapping, and pharmacogenomics have been the foundation for the vision of personalized medicine [7].

Gene expression analysis has already had a profound impact on classification and diagnosis in oncology, and has contributed to the development of new targeted treatment approaches [8]. It is thus reasonable to expect that genomic information will become an irreplaceable part of diagnosis and treatment for cancer and thereby contribute valuable information to the patients' medical records.

Given this vision of the future, that the move toward using clinico-genomic information for cancer treatment is inevitable, it is of essence to assess what this information should comprise, and in what way the healthcare professionals could use it for targeted diagnosis and treatment.

Some of the applications of genomic technology that could be incorporated in the future medical records and could improve the clinical outcome of the patients are discussed below:

1) *Gene expression* profiling of diseased tissues, e.g. tumors, has introduced a new insight into the classification and prognostic stratification of diseases. For example, diffuse large B-cell lymphoma (DLBCL) is a heterogeneous disease with different clinical outcomes. Molecular profiling with cDNA microarrays has uncovered distinct subclasses with prognostic significance (7). Another example of a disease which has similar clinico-pathological characteristics but is very heterogeneous in terms of biology and clinical outcome is breast cancer [9].
2) *Identification of predisposition risk factors* to certain diseases with genotyping methods and genome-wide association studies could help prevent disease by altering dietary habits and lifestyle or even by early intervening with preventive therapeutic agents [10]. Such risk information could be stored in a future EHR/EMR.
3) *Pharmacogenetics* is the study of variability in drug response that is regulated by hereditary factors. Drug response includes the processes of drug absorption and disposition (e.g. pharmacokinetics, (PK)), and drug effects (e.g. pharmacodynamics (PD), drug efficacy and adverse effects of drugs) [11]. Pharmacogenomics is a broader term which encompasses the study of the whole genome in order to identify variants that determine variability in drug response [12]. Commercially available diagnostic tests, like AmpliChip CYP450, are an example of pharmacogenetic application in the prevention of adverse drug reaction and the improvement of drug efficacy [13].

Current technology for genomic testing includes not only genomic sequencing, SNP mapping and gene expression profiling but also functional testing like examination of protein products (proteomics), metabolic profiles and epigenetics (methylation or acetylation of DNA) that influence gene expression (see also Figure 2)**.**

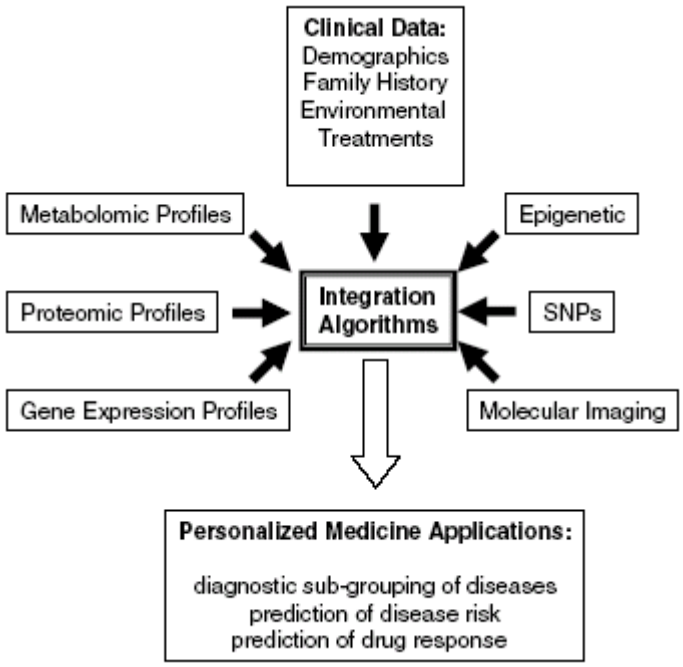


Figure 2: Integration of molecular, clinical and other data to enable personalized diagnostics and personalized risk predictions [14].

In summary, genomics will generate vast quantities of data, much of it unintelligible to the healthcare provider [14]. Diagnosis and therapy of diseases will be individualized with the support of genetic knowledge based systems and molecular expert systems. Specific drug therapy will be chosen for every patient based on his genotype and metabolic phenotype. Moreover, people will be stratified based on their genetic profile as "group-at-risk" for certain diseases and conditions and information will be provided for preventive measures as well as for frequent screening methods for early diagnosis and intervention.

2.2. Relevant Genes: an Overview

As mentioned, we use the BEA tool to exhaustively search the research literature for genes and biological processes identified as relevant for oncology. BEA is an advanced life sciences literature analysis application that supports visually driven search of very large literature databases, experiment design and research hypothesis generation [5].

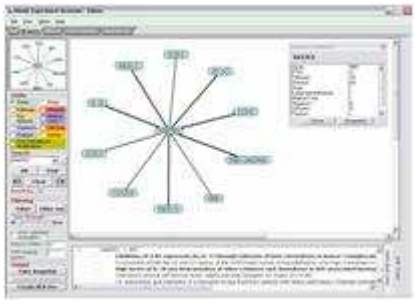


Figure 3: BEA snapshot, taken from [5]

BEA correlates information on parameters such as experimental methods, genes, pathways, diseases, and reagents from thousands of published experiment descriptions in the top science journals as well as over 12 million abstracts. The way BEA operates is to

- Find published experiments that are related to a specific area of interest based on multiple parameters, including genes/proteins, pathways, diseases, drugs, reagents and others.
- Capture and organize literature search and analysis work and easily creates reports.

We posed the following questions to BEA:

- What genes are described in the research literature as relevant in the context of oncology (detection, prediction, prognosis, etc)? If possible, classified on the cancer type.
- What patways are described in the research literature as relevant in the context of oncology (detection, prediction, prognosis, etc)? If possible, classified on the cancer type.

BEA's exhaustive search returned a list of genes relevant for oncology and a list of relevant processes, both classified by a cancer type (in fact we obtained two sets of six tables for breast, lung, ovarian, skin, and head and neck cancer as well as leukemia). The tables with raw results are given in Appendix 3.

The raw results, as can be seen from tables in Appendix 3 contain duplicates and the same genes that occur under different aliases. We removed duplicates, and found official gene names using HUGO gene nomenclature [15] and NCBI gene database [16]. Then, we refined the results further by talking to clinicians involved in ACGT project to obtain the relevant subset, which is presented in Table 1 for breast cancer. As the genes have many aliases we format the results based on the approved gene symbol and name, and also list key properties of that gene. We expect that the gene in an EHR should be uniquely identifiable, and in that context identification via an approved symbol might be an option.

Table 1: The list of genes relevant to breast cancer, based on the official gene nomenclature from [15]

Approved Gene Symbol	Approved Gene Name	Location	Sequence Accession IDs	Previous Symbols	Aliases
ESR1	Estrogen receptor 1	6q24-q27	X03635	ESR	NR3A1, Era
ESR2	Estrogen receptor 2 (ER beta)	14q21-q22	X99101		NR3A2, Erb
BRCA1	Breast cancer 1, early onset	17q21-q24	U14680		RNF53, BRCC1
TP53	Tumor protein p53	17p13.1	AF307851 NM_000546		P53, LFS1
BRCA2	breast cancer 2, early onset	13q12-q13	U43746 NM_000059	FANCD1, FACD, FANCD	FAD, FAD1, BRCC2
PGR	progesterone receptor	11q22-q23	M15716		PR, NR3C3

EGFR	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)	7p12	NM_005228	ERBB	ERBB1
BCL2	B-cell CLL/lymphoma 2	18q21.3	M14745 NM_000633, NM_000657		Bcl-2
CCND1	cyclin D1	11q13	Z23022	BCL1, D11S287E, PRAD1	
VEGFA	vascular endothelial growth factor A	6p12	AB021221 NM_001025366	VEGF	VEGF, VEGF-A, VPF
AKT1	v-akt murine thymoma viral oncogene homolog 1	14q32.32-q32.33	M63167 NM_005163		RAC, PKB, PRKBA, AKT
ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	17q11.2-q12	X03363	NGL	NEU, HER-2, CD340
CDKN1A	cyclin-dependent kinase inhibitor 1A (p21, Cip1)	6p21.1	U03106 NM_078467	CDKN1	P21, CIP1, WAF1, SDI1, CAP20, p21CIP1
COL1A1	collagen, type I, alpha 1	17q21.3-q22	Z74615		OI4
CDH1	cadherin 1, type 1, E-cadherin (epithelial)	16q22.1	L08599 NM_004360	UVO	uvomorulin, CD324
IGF1	insulin-like growth factor 1 (somatomedin C)	12q23.2	X00173 NM_000618		
EGF	epidermal growth factor (beta-urogastrone)	4q25	X04571		
TNF	tumor necrosis factor (TNF superfamily, member 2)	6p21.3	X02910	TNFA	TNFSF2, DIF
ABCG2	ATP-binding cassette, sub-family G (WHITE), member 2	4q22-q23	AF103796 NM_004827		EST157481, MXR, BCRP, ABCP, CD338
KERSMCR	cytokeratin, Smith Magenis syndrome chromosome region	17p11.2			
TGFB1	transforming growth factor, beta 1	19q13.1	X02812	TGFB, DPD1	CED
BAX	BCL2-associated X protein	19q13.3-q13.4	NM_138763		
ACTB	actin, beta	7p15-p12	M28424 NM_001101		
MYC	v-myc myelocytomatosis viral oncogene homolog (avian)	8q24			c-Myc
NKRF	NF-kappaB repressing factor	Xq24	AJ011812 NM_017544		ITBA4, NRF
PTGS2	prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	1q25.2-q25.3	D28235 NM_000963		COX2
IGF1	insulin-like growth factor 1 (somatomedin C)	12q23.2	X00173 NM_000618		
CDKN1B	cyclin-dependent kinase inhibitor 1B (p27, Kip1)	12p13.1-p12	AF480891		KIP1, P27KIP1
CASP3	caspase 3, apoptosis-related cysteine peptidase	4q34	BC016926 NM_004346		CPP32, CPP32B, Yama, apopain
CCNE1	cyclin E1	19q12	M73812	CCNE	
AR	androgen receptor (dihydrotestosterone receptor; testicular feminization; spinal and bulbar muscular atrophy; Kennedy disease)	Xq12	M20132 NM_000044	DHTR, SBMA	AIS, NR3C4, SMAX1, HUMARA
PRAP1	proline-rich acidic protein 1	10q26.3	AF421885 NM_145202		UPA
PIK3C2A	phosphoinositide-3-kinase, class 2, alpha polypeptide	11p15.5-p14	Y13367 NM_002645		PI3K, PI3K-C2alpha
DACT1	dapper, antagonist of beta-catenin, homolog 1 (Xenopus laevis)	14q22.3	AF251079 NM_016651		DAPPER1, THYEX3, HDPR1, DAPPER, FRODO
PTEN	phosphatase and tensin homolog (mutated in multiple advanced cancers 1)	10q23	U92436 NM_000314	BZS, MHAM	MMAC1, TEP1, PTEN1
BAX	BCL2-associated X protein	19q13.3-q13.4	NM_138763		
MAPK3	mitogen-activated protein kinase 3	16	M84490	PRKM3	ERK1, p44mapk, p44erk1
MAPK1	mitogen-activated protein kinase 1	22q11.2	M84489	PRKM2,	ERK, ERK2,

				PRKM1	p41mapk, p38, p38, MAPK2
MMP9	matrix metallopeptidase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase)	20q12-q13		CLG4B	

2.3. Genes in Action: Examples

In this section we discuss examples of the current applications of genomics in cancer care, focusing on gene expression profiling and disease classification, risk assessment, and pharmacogenomics.

2.3.1. Gene Expression Profiling and Disease Classification

Recent advances in the technology of gene expression analysis and the completion of the human genome project have changed the approach of diagnosis, classification and treatment of cancer. Traditionally, disease taxonomy was based on clinical, morphological, and molecular parameters. New information on the genetic basis of the diseases has clarified the wide heterogeneity in prognosis, clinical course and treatment response, and in some cases has led to a reclassification. Here we give two representative cancer examples, such as breast cancer.

- **Cancer classification**

It has been shown that breast cancer is a heterogeneous disease according to the estrogen receptor (ER), tumor grade, and age [22]. Other prognostic factors are tumor size, the histologic type and the status of axillary lymph nodes [23]. Using cDNA microarrays distinct patterns of gene expression in 40 different breast cancers were identified, and classification into four subtypes of the disease suggested: luminal-like, basal-like, HER-positive (ErbB2), and normal-like.

Basal-like tumors express genes which are characteristic of myo-epithelial cells of the basal part of the normal breast epithelium and they are ER negative, while *luminal-like* tumors express genes which are characteristic of the normal luminal cells and they are ER positive. Later, the same researchers have identified three distinct subtypes of the luminal like, subtype A, B and C (see Figure 4). Luminal A type cancers have the most favourable long-term survival and tend to respond better to endocrine therapy, whereas basal-like and HER-positive are more aggressive and have a higher likelihood of being grade III [24]. Basal like tumors have a high frequency of p53 and BRCA1 mutations. However, in [25] a higher pathologic complete response with neo-adjuvant chemotherapy has been reported. In [26] breast cancer according is distinguished according to ER status. Samples ER (-) were subdivided into basal-1, basal-2 and HER2/neu whereas ER(+) into luminal-1, -2 and -3 [27].

Aside from this molecular classification, microarrays have resulted in various gene expression signatures with prognostic implications [24]. Two conceptually different supervised approaches for prognostic marker discovery have been applied so far: the

"top-down" and the "bottom-up" approach. The top-down approach simply is looking for gene-expression patterns associated with clinical outcome without any biological assumption, whereas the bottom-up approach first uses a hypothesis that links a gene expression profile with a specific biological phenotype and subsequently correlates these findings to survival [27].

Researchers from the Netherlands Cancer Institute in Amsterdam, using a top-down approach, have reported a 70-gene signature (Mammaprint ®) using the Agilent platform in 78 samples from young untreated lymph node negative patients [23]. Subsequently, they validated this signature on 295 patients with both negative and positive node, treated and untreated disease, and they showed that this signature was the strongest predictor for distant metastases-free survival, independent of adjuvant treatment, tumor size, histological grade and age.

Later, another group from Rotterdam, using the Affymetrix platform, identified a 76-gene signature in 115 untreated node-negative breast cancer patients. Subsequent validation in 171 patients has shown that this signature was also predictive for distant metastases-free survival.

Interestingly, these two signatures were validated by the TRANSBIG consortium as predictive of the clinical outcome, although they had only 3 genes overlapping. Another interesting finding of the TRANSBIG study was the time-dependence of the distant metastasis. Both of the signatures were able to predict early metastases within 5 years, but not metastases developed later. This observation suggests that the mechanisms involved in early and late metastases are different and needs clarification. There are already studies that measure the presence of circulating tumor cells or detect tumor cells in bone marrow and correlate their presence with poor outcome and disease recurrence. Other gene expression signatures with prognostic information have also been published, including genomic grade [28],_wound response, invasiveness, and p53 status [24]. Despite different genes in their lists, they all carry similar information regarding prognostication and they are useful for determining the risk of recurrence in the ER+ subgroup. The Stanford group, using a bottom-up approach has identified a wound response signature which had a markedly worse clinical outcome. Recently Sotiriou et al [27] have focused on the histological grade and have tried to identify distinct gene expression. They developed a gene expression grade index based on 97 genes that were consistently differentially expressed between low and high grade breast cancers. These genes were associated mostly with cell-cycle progression and proliferation. Proliferation genes appear to be the most important prognostic factors in all these different signatures, as illustrated in Figure 7.

- **Therapy response prediction**

The next application of the microarrays is the identification of therapy response. Several studies have tried to predict drug sensitivity based on the gene expression, thus providing also new information on the biology of breast cancer as well as on resistance mechanisms. As a consequence, new therapeutic targets have been discovered, e.g., microtubule-associated protein (MAPT) which is associated with paclitaxel resistance.

Figure 6 summarizes the evolution of the treatment strategies based on the new technology and the identification of specific molecular targets.

It is clear that gene-expression profiling has great potential for improving breast cancer management. The Breast International Group (BIG), with the European Organization for Research and Treatment of Cancer (EORTC) Breast Cancer Group recently launched the 'Microarray for Node Negative Disease may Avoid Chemotherapy' (MINDACT) study to prospectively validate the use of the Amsterdam gene-expression signature as a tool to better select good prognosis patients who would not benefit from adjuvant chemotherapy.

6,000 women will participate in the study, and it is estimated that 10–15% fewer women will be treated with chemotherapy in the experimental arm.

A similarly ambitious initiative is taking place in the US: the Oncotype DX RS is currently under evaluation in the TAILORx trial, which is part of the US National Cancer Institute (NCI) Program for the Assessment of Clinical Cancer Tests (PACCT). This trial randomizes patients with intermediate recurrence score (RS) to receive either hormonal therapy alone or hormonal therapy plus chemotherapy. Patients with low RS will be treated with hormonal therapy alone and patients with a high score will receive chemotherapy plus hormonal therapy.

These two trials should provide level I evidence about the clinical relevance of applying gene-expression predictors to daily breast cancer patient management [27].

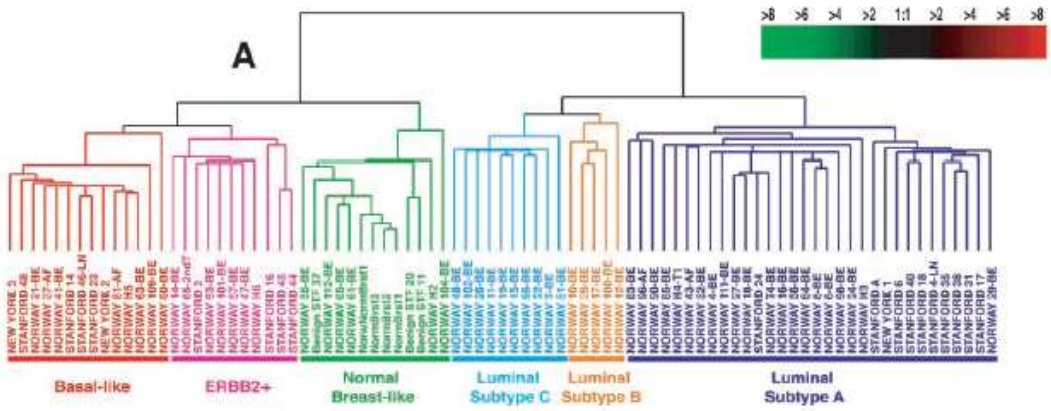


Figure 4: Subtypes of breast cancer based on the differences in gene expression. The cluster dendrogram showing the six subtypes of tumors are colored as: luminal subtype A, dark blue; luminal subtype B, yellow; luminal subtype C, light blue; normal breast-like, green; basal-like, red; and ERBB21, pink

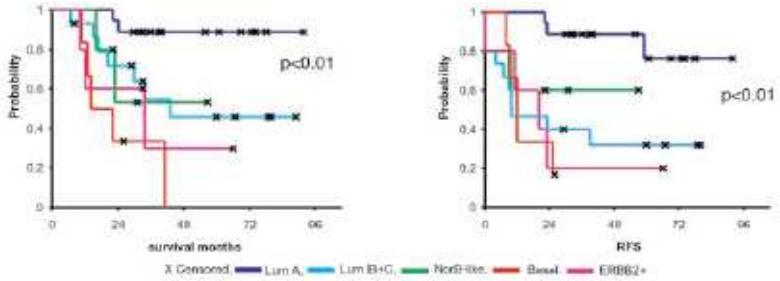


Figure 5: Overall and relapse-free survival analysis of the 49 breast cancer patients, uniformly treated in a prospective study, based on different gene expression classification [29]. (*A*) Overall survival and (*B*) relapse-free survival for the five expression-based tumor subtypes based on the classification presented above (luminals B and C were considered one group).

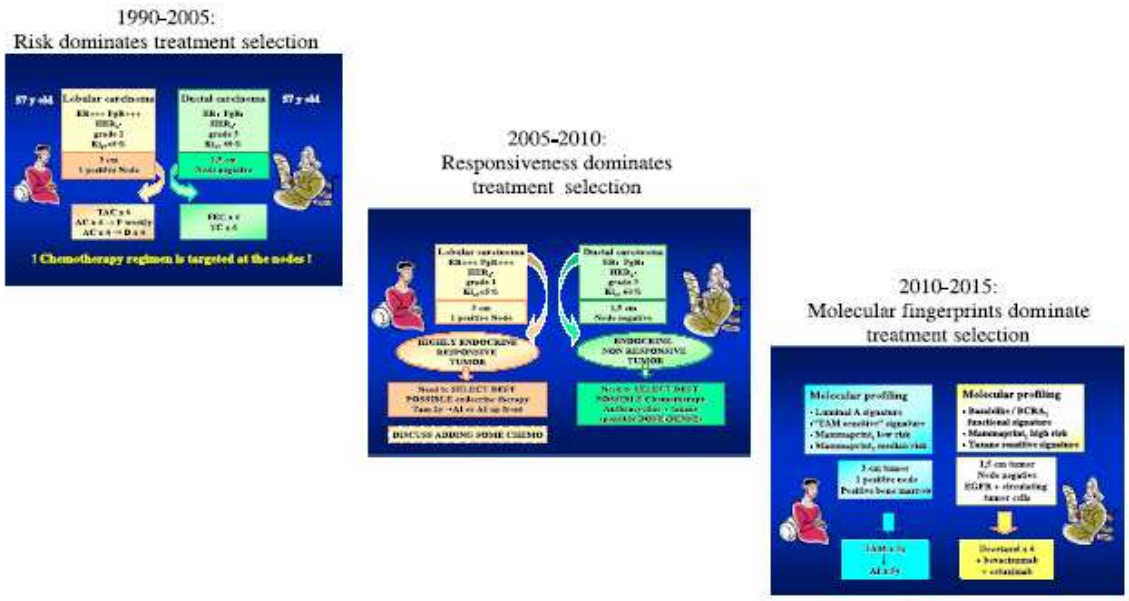


Figure 6: Evolution of treatment strategies: "Aiming at the target" [24]

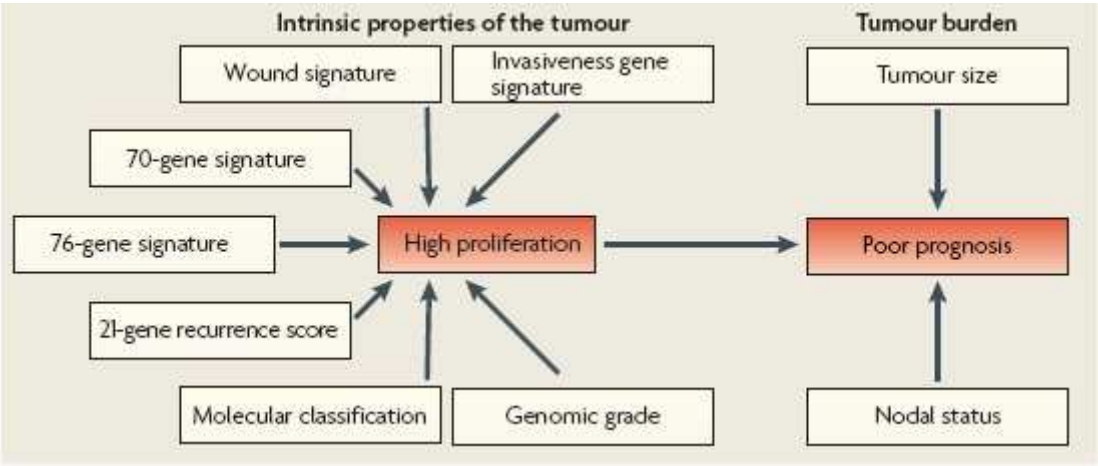


Figure 7: Integrative analysis of gene-expression profiles [27].

- **New applications of microarrays**

In [30] a number of important contributions of DNA microarrays to experimental pathology in the oncology area have been presented, and several challenges of bringing this technology to clinical practice discussed. Profiling gene expression has initially been the predominant application of DNA microarrays. In the last few years, the platform has been adapted to some new applications as well, such as measuring germline genetic variation, occurring as SNPs or copy number variations. Genome-wide association studies using SNP arrays have identified genetic loci conferring increased cancer risk. Microarrays have also been used to identify altered patterns or imbalance in DNA methylation in cancers. Arrayed antibodies and antigens have been used to quantify levels of tumor proteins and antitumor humoral response, respectively. Tissue microarrays have been used to

measure a single gene's expression across many samples, accelerating the evaluation and validation of initial DNA microarray discoveries across large patient cohorts. Several multigene tests derived from microarray data are being evaluated in prospective clinical trials (e.g. Agendia's MammaPrint).

Although microarray datasets or genomic sequences are currently used in a clinical decision process only to identify a certain characteristic or deviation from standard (e.g. mutation, insertion, overexpression of a certain sequence), which would imply that instead of storing the entire dataset an yes/no answer to a clinical question may be sufficient, the storing of the entire dataset would be beneficial long term as it can be used for future characterization of the same disease (e.g. when new research results are available), for other clinical decisions concerning the patient but not necessarily related to the current study, for future clinical research and for epidemiological studies. When that alternative will become available, entire genomes of patients could be stored, and the EHR should be able to link to those.

Much effort is currently directed towards the standardization of the operating procedures for specimen collection and processing, RNA isolation and labeling, and microarray hybridization, imaging and analysis. More rigorous statistical analysis has also led to an increased likelihood of validation. Recently, high reproducibility of findings, when standard operated procedures are followed, and improved inter-platform concordance have been demonstrated. Also it has been shown that in the case of discordant findings among different laboratories, the different reported gene signatures reported might reflect the same underlying biology or provide comparable clinical utility.

There is currently no agreement as to whether microarray tests should assay only a focused set of diagnostic genes, or a wider set of genes that may provide additional information but also involve additional risks, like unanticipated diagnoses [30]. Microarrays are expected to be increasingly used for clinical diagnosis.

Microarrays provide additional information or outperform standard histopathological markers

- Many genes provide more information than a few
- Adequate performance characteristics are demonstrated
- Testing impacts the patient management decision
- There has been appropriate validation of clinical utility (ideally including prospective clinical trials)
- The cost of the tests is justified by their clinical utility

Based on current research results, we can assume that microarrays are likely to have clinical utility in cancer classification and subtyping and in classifying the anatomical sites of origin of tumors to select the optimal treatment regimen and to help in diagnostically challenging cases. Microarray analysis can also provide useful prognostic information to identify patient groups with better outcome to decide whether they can be spared therapy, and in treatment response and toxicity monitoring.

2.3.2. Genotyping and Risk Association

Historically, genetic predisposition to complex diseases has been determined using techniques of genome-wide scanning with microsatellites (short tandem repeat sequences). A more contemporary approach in genotyping involves the use of high-throughput identification of single nucleotide polymorphisms (SNPs) [31]. The HapMap project [32] is developing a map of the SNPs in the human genome. The Phase II HapMap project has already genotyped over 3.1 million human SNPs in 270 individuals

from four diverse populations (2007 International HapMap Consortium) [32]. The identification of SNPs in individuals with disease and in unaffected controls can give information about the susceptibility to a variety of diseases [31]. Such studies are called *association studies*. The identification of various haplotypes (collections of SNPs inherited together) which are present within a population, limits the number of the SNPs needed to be studied into a smaller number of "tag" SNPs. Table 2 shows known associations between common polymorphisms and common diseases.

The development of high-throughput economic commercial genotyping platforms has facilitated genome-wide association studies. The first disease where an association study has shown a common variant is the complement factor H gene as a risk allele in age-related macular degeneration [33]. The identification of biological pathways of the diseases will further contribute to the development of new targeted treatments. Populations will be sub-classified in risk groups according to their genotype and decision making support information will be given to the physician in order to prevent the disease.

Table 2: Associations between common polymorphisms in genes and common diseases [33]

Associations between common polymorphisms in genes and common diseases or dichotomous traits

Disease/trait	Gene (ref)	Gene (ref)	Gene (ref)	Gene (ref)
Cancer				
Acute leukemia	CYP1A1 (45)	CYP2D6 (45, 46)	GSTM1 (45)	GSTT1 (45)
	MTHFR (20)	NAT2 (47)		
Bladder cancer	GSTM1 (48)	GSTP1 (49)	GSTT1 (50)	
Breast cancer	COMT (51)	CYP17 (52)	CYP19 (53)	CYP1A1 (54)
	CYP1B1 (55, 56)	ERBB2 (57)	ESR1 (58)	GSTM1 (59)
	HRAS (60)	HSPA8 (61)	NAT1 (62)	NAT2 (63)
	PGR (64)	SHBG (65)	SOD2 (66)	TP53 (67)
	VDR (68)			
Cervical cancer	GSTT1 (69)	MTHFR (70)	TP53 (71)	
CLL	ETS1 (72)	TNF (73)		
Colorectal cancer	ALDH2 (74)	APC (75)	CYP1A1 (76)	DIA4 (77)
	GSTM1 (78)	GSTT1 (79)	LTA (80)	MSH3 (81)
	MTHFR (18)	NAT1 (82)	NAT2 (83)	XRCC1 (84)
Endometrial cancer	CDKN1A (85)	CYP1A1 (86)	MMP1 (87)	MTHFR (86)
	TP53 (88)			
Gastric cancer	ALDH2 (74)	GSTM1 (89)	GSTT1 (90)	IL1B (91)
	MYC (92)			
Glioblastoma	PPARG (93)			
Head/neck cancer	ADH1B (94)	ALDH2 (94)	CDKN1A (95)	CYP1A1 (96)
	CYP2D6 (97)	CYP2E (98)	FCGR3A (99)	GSTM1 (100)
	GSTM3 (101)	GSTP1 (102)	GSTT1 (101)	LTA (103)
	MYCL1 (104)	NAT1 (48)	NAT2 (102, 105)	TP53 (106)
Hodgkin's lymphoma	HSPA8 (61)	TNF (61)		
Liver cancer	CYP2D6 (107)	CYP2E (108)	EPHX1 (109)	
Lung cancer	ALDH2 (74)	CDKN1A (110)	CYP1A1 (111)	CYP1B1 (55)
	CYP2A6 (112)	CYP2E (113)	DIA4 (114)	DIA4 (115)
	EPHX1 (116)	GPX1 (117)	GSTM1 (118, 119)	HRAS (120)
	LTA (121)	MGMT (122)	MPO (123)	NAT1 (124, 125)
	NAT2 (126)	TF (127)	TP53 (128)	
Melanoma	HRAS (129)	MC1R (130)	XRCC3 (131)	
Non-Hodgkin's lymphoma	EPHX1	ETS1 (132)	PGR	
Oral leukoplakia	GSTM1 (133, 134)	GSTT1 (133, 134)		
Oligoastrocytoma	ERCC1 (99)			
Ovarian cancer	HRAS (135)	TP53 (136)		
Prostate cancer	AR (137, 138)	CYP17 (139, 140)	CYP1A1 (141)	CYP1B1 (142)
	CYP3A4 (143)	ELAC2 (144)	GSTP1 (49)	SRD5A2 (145)
	VDR (146)			
Renal cell cancer	CYP1A1 (147)	GSTT1 (148)		
Testicular cancer	GSTP1 (49)			
Cardiovascular disease				
CAD/MI	ACE (149)	ADRB3 (150)	AGTR1 (151)	APOA1 (152)
	APOB (153)	APOE (154)	CD14 (155)	CYBA (156)
	F13A1 (157)	F2 (158)	F5 (159)	F7 (160)
	FGB (161)	GP1BA (162)	GSTM1 (163)	HTR2A (164)
	IRS1 (165)	ITGA2 (166)	ITGB3 (167)	LPL (168)
	MMP3 (169)	MTHFR (13, 14)	NOS3 (170, 171)	NPPA (172)
	PLAT (173)	PON1 (174)	PON2 (175)	PPARG (176)
	SELE (177)	SELP (178)	SERPINA8 (179, 180)	SERPINE1 (181)
	TGFB1 (182)	THBD (183)	WRN (184)	
DVT	F13A1 (185)	F2 (186)	F3 (187)	F5 (7)
	MTHFR (19)	PLAT (188)	PON1 (189)	
Dilated cardiomyopathy	ACE (190)	EDNRA (191)	PLA2G7 (170)	SOD2 (192)
HTN	ACE (193)	ADD1 (194)	AGTR1 (195)	CYP11B2 (196)
	DIA4 (197, 198)	DRD1 (199)	GCK (200)	GNAS1 (201)
	GNB3 (202)	GYS1 (203)	HSD11B2 (204)	INSR (205)
	MTHFR (206)	NPPA (172)	REN (207)	SAH (208)
	SCNN1B (209)	SERPINA8 (210)	TGFB1 (182)	TH (211)
Survival post-CHF	ADRB2 (212)	AMPD1 (213)		
Dermatology				
Acne	MUC1 (214)			
Contact dermatitis	NAT2 (215)			
Eczema	CMA1 (216)			
Psoriasis	C4A (217)	CDSN (218)	LTA (219)	OTF3 (220)
	SERPINA8 (219)	TAP1 (221)	TNF (222, 223)	VDR (224)

Disease/trait	Gene (ref)	Gene (ref)	Gene (ref)	Gene (ref)
Endocrinology				
Addison's disease	CTLA4 (225)			
Gestational DM	INSR (226)			
Graves' disease	CTLA4 (227)	IFNG (228)	IL4 (229)	TAP1 (230)
	THRB (231)	TRHR (232)	VDR (233)	
Hyperparathyroidism	VDR (234)			
Male infertility	AR (235)	LHB (236)		
Obesity	ABCC8 (237)	ADRB2 (238)	ADRB3 (239)	APOB (240)
	APOD (241)	GNB3 (242)	LDLR (243)	LEP (244)
	LIPE (245)	NMB (246)	NPY5R (247)	PPARG (248)
	TNF (249)			
Osteoporosis/fracture	COL1A1 (250)	TGFB1 (251)	VDR (252)	
PCOS	CYP11A (253)	CYP17 (254)	FSHB (255)	FST (256)
	INS (257)	LHB (258)		
Short stature	DRD2 (259)	VDR (260, 261)		
Type 1 diabetes	BCL2 (262)	C4A (263)	CCR2 (264)	CD3D (265)
	CD4 (265)	CTLA4 (266)	GCK (267)	ICAM1 (268, 269)
	IFNG (270)	IGHV2-5 (271)	IL6 (272)	INS (273)
	LTA (274)	NEUROD1 (275)	PSMB8 (276)	VDR (277)
	WFS1 (278)			
Type 2 diabetes	ABCC8 (279)	ACE (280)	ADRB2 (281, 282)	CD4 (283)
	FRDA (284)	GCGR (285, 286)	GCK (287, 288)	GYS1 (289)
	HFE (290)	INS (291)	INSR (292, 293)	IPF1 (294)
	IRS1 (295)	KCNJ11 (296)	PCSK2 (297)	PPARG (37)
	PPP1R3 (298)	RRAD (299)	SLC2A1 (300)	SLC2A2 (301)
	TCF1 (302)	UCP3 (303)		
Gastroenterology				
Celiac disease	CTLA4 (304)	TNF (305)		
Cholelithiasis	APOB (306)	CETP (307)		
IBD	BDKRB1 (308)	F5 (309)	IL10 (310)	IL1RN (311)
	MLH1 (312)	MTHFR (313)	MUC3A (314)	TNF (315)
	VDR (316)			
Pancreatitis	IL1RN (317)			
Primary biliary cirrhosis	CTLA4 (318)	VDR (319)		
Infectious disease				
Cerebral malaria	CD36 (320)	ICAM1 (321)	NOS2A (322)	TNF (323)
HIV infection/AIDS	CCR2 (324)	CCR5 (325, 326)	CX3CR1 (327)	MBL2 (328)
	SDF1 (329)	SLC11A1 (330)		
Leishmaniasis	TNF (331)			
Leprosy	TNF (332)	VDR (333)		
Meningococcal disease	FCGR2A (334)	SERPINE1 (335)	TNF (336)	
Parasitic infections	ADRB2 (337)	NOS2A (338)		
RSV bronchiolitis	IL8 (339)			
Severe sepsis	IL1RN (340)			
Trachoma	IL10 (341)	TNF (342)		
Tuberculosis	SLC11A1 (343)			
Viral hepatitis	MBL2 (344)	TNF (345)		
Miscellaneous				
Athletic endurance	ACE (346)			
Benzene toxicity	DIA4 (347)			
Fair skin, red hair	MC1R (348)			
High altitude HTN	ACE (349)			
Lead poisoning	ALAD (350)			
Longevity	ACE (351)	APOA1 (352)	APOB (353)	APOE (354)
	SERPINE1 (355)			
Macular degeneration	APOE (356)	EPHX1 (357)	SOD2 (357)	
Tobacco use	DRD2 (358)	SLC6A3 (359)		
Trichloroethylene toxicity	GSTM1 (360)	GSTT1 (360)		
Neonatal disease				
Cleft lip/palate	BCL3 (361)	MSX1 (362)	RARA (363)	TGFA (364)
	TGFB2 (365)	TGFB3 (362)		
Neural tube defect	MTHFR (16, 17)	MTR (366)	T (367)	
Pyloric stenosis	NOS1 (368)			
RDS	SFTPA1 (369, 370)			

Disease/trait	Gene (ref)	Gene (ref)	Gene (ref)	Gene (ref)
Neurology				
Absence seizures	GABRB3 (371)	OPRM1 (372)	SLC6A3 (373)	
Alzheimer's disease	A2M (374, 375)	ACE (376)	APBB1 (377)	APOA4 (378)
	APOC1 (379)	APOC2 (380)	APOE (381)	BCHE (382)
	BLMH (383)	IL1A (386)	CTSD (384)	HTR6 (385)
	LRP1 (387)	NOS3 (388)	PSEN1 (389)	SERPINA3 (390)
	SLC6A4 (391)	TF (392)	TFCP2 (393)	TGFB1 (394)
	TNFRFS6 (395)	VLDLR (396)		
Creutzfeldt-Jakob disease	PRNP (397)			
Epilepsy	CHRNA4 (398)			
Guillian-barré syndrome	TNF (399)			
Head injury outcome	APOE (400)			
Hydrocephalus	APOE (401)			
Intracranial aneurysms	ACE (402)	ENG (403)	MMP9 (404)	
Ischemic stroke	ACE (405)	APOE (406)	CYBA (407)	ENG (408)
	F13A1 (409)	F2 (410)	FGB (411)	GP1BA (162)
	ITGA2 (412)	MTHFR (413, 414)	NOS3 (415)	NPPA (416)
	PLA2G7 (417)	PON1 (418)		
Migraine headache	DBH (419)	MTHFR (420)	SLC6A4 (421)	
Multiple sclerosis	CTLA4 (422)	IL1RN (423)	MBL2 (424)	PTPRC (425)
Myasthenia gravis	FCGR2A (426)	IL1B (427)	TNF (428)	
Otosclerosis	COL1A1 (429)			
Parkinson's disease	A2M (430)	ADH4 (431)	CCK (432)	COMT (433)
	CYP1A1 (434)	CYP2D6 (435)	DLST (436)	DRD2 (437)
	EPHX1 (438)	GSTP1 (439)	MAOA (440)	MAOB (441)
	MAPT (442)	NAT2 (443)	NOS3 (444)	SERPINA3 (445)
	SERPINA3 (445)	SLC6A3 (446)	SLC6A4 (447)	SNCA (448)
	UCHL1 (449)			
Obstetric disease				
Endometriosis	ESR1 (450)			
Fetal loss	ACP1 (451)	CTLA4 (452)	EPHX1 (453)	F2 (454)
	F5 (455)	MTHFR (456)		
Preeclampsia	AGTR1 (457)	F2 (458)	F5 (459)	LPL (460)
	MTHFR (461)	NOS3 (462)	SERPINE1 (463)	TNF (464)
Pharmacogenetics				
Albuterol response	ADRB2 (465)			
Antidepressant response	GNB3 (466)			
Aspirin response	ITGB3 (467)			
Azathioprine toxicity	TPMT (468)			
Beta-blocker response	GNAS1 (201)			
Clozapine response	DRD3 (469)	HSPA1A (470)	HSPA2 (470)	HTR2A (471)
	HTR2C (472)	HTR6 (473)	TNF (474)	
Drug-induced tardive dyskinesia	CYP2D6 (475, 476)	DRD2 (477)	DRD3 (478)	HTR2C (479)
	SOD2 (480)			
Fluvastatin response	APOB (481)			
Fluvoxamine response	SLC6A4 (482)			
Irinotecan toxicity	UGT1A1 (483)			
Leukotriene Inhibitor response	ALOX5 (484)			
Lithium response	IMPA1 (485)			
Menadione-associated urolithiasis	DIA4 (486)			
Omeprazole response	CYP2C19 (487, 488)			
Pravastatin response	CETP (489)	MMP3 (490)		
Tacrine response	APOE (491)			
Tricylic antidepressant response	CYP2D6 (492)			
Warfarin response	CYP2C9 (493)			

Disease/trait	Gene (ref)	Gene (ref)	Gene (ref)	Gene (ref)
Psychiatry				
Anorexia	HTR2A (494)			
ADHD	COMT (495)	DRD4 (496)	DRD5 (497)	SLC6A3 (498)
	HTR2A (499)	SNAP25 (500)		
Autism	ADA (501)	EN2 (502)	FMR1 (503)	
Bipolar disorder	APOE (504)	ATP1A3 (505)	COMT (506)	DDC (507)
	DRD3 (508)	GABRA5 (509)	HTR5A (510)	HTR6 (511)
	MAOA (512)	MAOB (513)	PLA2G1B (514)	PLCG1 (515)
	SERPINA8 (516)	SLC6A4 (517)	TPH (518)	
Compulsive gambling	DRD2 (519)	DRD4 (520)		
Depression	ACE (521)	COMT (522)	DRD3 (523)	DRD4 (524)
	GNB3 (466)	HTR5A (510)	SLC6A4 (525)	TPH (526)
OCD	DRD4 (527)	HTR1B (528)	HTR2A (529)	SLC6A4 (530)
Panic disorder	ADORA2A (531)	CCK (532)		
Schizophrenia	APOE (533)	CCK (534)	CCKBR (535)	COMT (536)
	DRD2 (537)	DRD3 (538)	DRD4 (539)	DRD5 (540)
	GNAL (541)	HMBS (542)	HRH2 (543)	HTR2A (544)
	HTR5A (510)	HTR6 (545)	KCNN3 (546)	NTF3 (547)
	OPRS1 (548)	PLA2G4A (549)	PLA2G7 (550)	YWHAH (551)
Pulmonary disease				
Asthma/atopy	ACE (552)	ADRB2 (553)	CCR5 (554)	CFTR (555)
	GSTP1 (556)	HNMT (557)	IL10 (558)	IL13 (559)
	IL4 (560)	IL4R (561)	IL9R (562)	LTA (563)
	MS4A1 (564)	NOS1 (565)	NOS3 (566)	PLA2G7 (567)
	SCYA5 (568)	SERPINA8 (569)	TAP1 (570)	TAP2 (571)
	TBXA2R (572)	TNF (563)	UGB (573)	
COPD/emphysema	CFTR (574)	EPHX1 (575)	GC (576)	GSTP1 (577)
	SERPINA1 (578)	SERPINA3 (579)	TNF (580)	
Pneumoconiosis	TNF (581)			
Pulmonary fibrosis	TGFB1 (582)			
Pulmonary embolism	FGA (583)			
Sarcoidosis	ACE (584)	CCR2 (585)	CCR5 (586)	SLC11A1 (587)
	VDR (588)			
Renal/urologic disease				
IgA nephropathy	TRA@ (589)			
Nephrotic syndrome	SERPINA1 (590)			
Renal failure	BDKRB1 (591)	DCP1 (592)	HSD11B2 (593)	KLKB1 (594)
	NOS3 (595)	SERPINA8 (592)		
Urolithiasis	DIA4 (486)			
Rheumatology				
Behcet's disease	ICAM1 (596)			
Intervertebral disc disease	COL9A2 (597)			
Juvenile chronic arthritis	IL6 (598)	TAP2 (599)		
JRA	SLC11A1 (600)			
Osteoarthritis	COL2A1 (601)	VDR (602)		
Rheumatoid arthritis	CRH (603, 604)	ESR1 (605)	HSPA1A (606)	IFNG (607)
	SLC11A1 (608)	TAP2 (609)	TRD@ (610)	XRCC3 (611, 612)
Sjogren's syndrome	GSTM1 (613)			
SLE	ACE (614)	ADPRT (615)	BCL2 (262)	C4A (427)
	C4B (616)	CTLA4 (617)	CYP2D6 (618)	FCGR2A (619)
	HSPA2 (620)	IGHV3-30-5 (621)	IL10 (622)	MBL2 (623)
	TNF (624)	VDR (625)		
Wegener's granulomatosis	CTLA4 (626)	PRTN3 (627)		

For each disease or trait, the number(s) in parentheses identifies the first reference(s) reporting a significant association with a polymorphism in the gene indicated by its official symbol. Citations can be found at *www.geneticsinmedicine.org*. Full gene names and OMIM numbers are listed in Table 4. CLL, chronic lymphocytic leukemia; CAD/MI, coronary artery disease/myocardial infarction; HTN, hypertension; CHF, congestive heart failure; DM, diabetes mellitus; PCOS, polycystic ovary syndrome; IBD, inflammatory bowel disease; RDS, respiratory distress syndrome; ADHD, attention deficit hyperactivity disorder; OCD, obsessive compulsive disorder; COPD, chronic obstructive pulmonary disease; JRA, juvenile rheumatoid arthritis; SLE, systemic lupus erythematosius; RSV, respiratory syncytial virus; DVT, deep vein thrombosis; IgA, immunoglobulin A.

2.3.3. Pharmacogenomics

Standard drug treatment is efficient for only a limited percentage of patients as shown in Table 3**.** Many of the drugs fall in the range of 50-60% efficacy. The lowest range of response is 25% for cancer chemotherapy. An example of targeted treatment that has dramatically improved survival is imatinib mesylate, a tyrosine kinase inhibitor in chronic myelogenon leukemia (CML). CML is characterized by an abnormal chromosomal translocation [17] and constitutive expression of an abnormal fusion protein bcr-abl that has a tyrosine kinase activity which promotes cell division and blocks apoptosis, leading to unregulated growth of hematopoietic stem cells. The discovery of a specific inhibitor of bcr/abl function, imatinib mesylate, has tremendously increased the survival of these patients [18]. The development of point mutations in the bcr-abl gene has resulted in the presence of imatinib resistance. An individualized approach in the treatment of CML has already detected resistant mutants of bcr-abl and discovered alternative treatment with novel tyrosine kinase inhibitors like dasatinib and nilotinib [19]. Table 4 summarizes some of the diagnostic tests and treatments that already have been used in a personalized setting.

Table 3: Response rates of patients to a major drug for a selected group of therapeutic areas [20].

Therapeutic area	Efficacy rate (%)
Alzheimer's	30
Analgesics (Cox-2)	80
Asthma	60
Cardiac Arrythmias	60
Depression (SSRI)	62
Diabetes	57
HCV	47
Incontinence	40
Migraine (acute)	52
Migraine (prophylaxis)	50
Oncology	25
Osteoporosis	48
Rheumatoid arthritis	50
Schizophrenia	60

Table 4: Selected personalized medicine drugs, treatments and diagnostics [21]. This list is not intended to be comprehensive, but reflects commonly used products as of September 2006. Chart is based on research and industry sources. *Entries in which diagnostic tests received formal FDA approval, or drugs that have a reference to pharmacogenomic selection in their label, are shaded yellow. Abrreviations used in the table include:* ***BCR-ABL*** *= breakpoint cluster region – Abelson;* ***BRCA 1,2*** *= breast cancer susceptibility gene 1 or 2;* ***c-KIT*** *= tyrosine kinase receptor;* ***CYP*** *= cytochrome P450 enzyme;* ***HER2*** *= human epidermal growth factor receptor 2;* ***TPMT*** *= thiopurine S-methyltransferase; and* ***UGT1A1*** *= UDP-glucuronosyltransferase 1A1.*

Therapy	Biomarker/Test	Indication
Anti-retroviral drugs	TruGene®-HIV1 Genotyping kit	Guides selection of therapy based on genetic variations that make the HIV virus resistant to some anti-retroviral drugs
Cancer treatment regimens	Oncotype DX 21- gene assay	Quantifies the expression of 21 genes linked to the likelihood of breast cancer recurrence in women, and the magnitude of benefit from certain types of chemotherapy and hormonal therapy
Camptosar® Irinotecan	UGT1A1	Colon cancer: variations in the UGT1A1 gene can influence a patient's ability to break down irinotecan, which can lead to increased blood levels of the drug and

[1] U.S. Food and Drug Administration, FDA Clears Genetic Test That Advances Personalized Medicine Test Helps Determine Safety of Drug Therapy 22 August 2005, http://www.fda.gov/bbs/topics/NEWS/2005/NEW01220.html (accessed 15 August 2006).

		a higher risk of side effects[1]
Drugs metabolized by cytochrome P450	AmpliChip® CYP2D6/CYP2C19	Aid in determining treatment choice and individualizing treatment dose for therapeutics that are metabolized primarily by the specific enzymes[2]
Gleevec® (imatinib mesylate)	BCR-ABL	Chronic myelogenous leukemia (CML): treatment of patients with Philadelphia chromosome positive CML in blast crisis, accelerated phase or chronic phase[3]
Gleevec® (imatinib mesylate)	c-kit	Gastrointestinal stromal tumor (GIST): treatment of patients with kit (CD117) positive unresectable and/or metastatic malignant GIST
Herceptin® (trastuzumab)	HER-2/neu receptor	Breast cancer: treatment of patients with metastatic breast cancer that overexpresses HER2 protein and who have received one or more chemotherapy regimens[4]
Immunosuppressive drugs	AlloMap® gene profile	Monitors patient's immune response to heart transplant to guide immunosuppressive therapy
Pharmaceutical and surgical prevention options and surveillance	BRCA 1,2	Guides surveillance/preventive treatment based on susceptibility risk for breast and ovarian cancer
Pharmaceutical and lifestyle prevention options	Familion® 5-gene profile	Guides prevention and drug selection for patients with inherited cardiac channelopathies such as Long QT syndrome (LQTS), which can lead to cardiac rhythm abnormalities
Pharmaceutical and surgical treatment options and surveillance	P16/CDKN2A	Guides surveillance/preventive treatment based on susceptibility risk for melanoma
Purinethol ® (mercaptopurine)	TPMT	Guides adjustment of dose in treatment of acute lymphoblastic leukemia[5]
Tamoxifen	Estrogen receptor	May predict whether adjuvant tamoxifen therapy is beneficial

Pharmacogenomics is one of the first clinical applications of the postgenomic era. In the recent years, commercially available pharmacogenomic tests have been approved by the FDA, but their application in patient care remains very limited [34], due to several reasons:

- Individual response to a drug is influenced not only by genetic factors but also by environmental factors like poor compliance, drug-drug or food-drug interactions. Moreover, the age and the metabolic status of the patient (e.g., renal, pulmonary, cardiovascular, and liver function) play a significant role in the efficiency of a drug.
- Metabolism of drugs is complex and in some cases not only the parent drug but also the metabolites have clinical effects [35].
- In some cases there are alternative pathways of elimination of a drug. Most of the drugs are metabolized by more than one enzyme, so even in the case that one of the enzymes is not active the other will successfully metabolize the drug.

Pharmacogenetic variations can occur on the level of drug transporter, drug metabolizing enzymes (DME), drug targets, and other biomarker genes [36]. In terms of polymorphism

[2] U.S. Food and Drug Administration, FDA classification 21 CFR 862.3360 1 April 2005, http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=862.3360 (accessed 15 August 2006).

[3] U.S. Food and Drug Administration, FDA Oncology Tools Approval Summary for imatinib mesylate for Accel. Approv 18 April 2003, http://www.accessdata.fda.gov/scripts/cder/onctools/summary.cfm?ID=326 (accessed 15 August 2006).

[4] U.S. Food and Drug Administration, FDA Oncology Tools Product Label Details in Conventional Order for trastuzumab, http://www.accessdata.fda.gov/scripts/cder/onctools/labels.cfm?GN=Trastuzumab (accessed 15 August 2006).

[5] U.S. Food and Drug Administration, Purinethol® drug label, http://www.fda.gov/Medwatch/SAFETY/2004/jul_PI/Purinethol_PI.pdf (accessed 15 August 2006).

and individual differences in drug response, drug metabolizing enzymes are most important.

Phase I DMEs participate in the oxidative metabolism of xenobiotics, which leads to pharmacological inactivation or activation, facilitated elimination and/or addition of reactive groups for subsequent phase II conjugation. Examples of phase I DMEs are cytochrome p450 (CYP) enzymes, flavin-containing monooxygenases, reductases, esterases, and alcohol dehydrogenases.

Phase II DMEs conjugate drugs using organic donor molecules, e.g., glutathione, UDP-glycuronic acid, and acetyl-coenzyme A. Examples of phase II DMEs are glutathione-S-transferases, N-acetyltransferases, UDP-glucuronosyltransferases, epoxide hydrolases and sulfotransferases.

CYP450 metabolizes 70-80% of prescribed drugs (see also Figure 8). CYP450 enzymes are highly polymorphic resulting in abolished, reduced, altered or increased activity [37]. Thus, the populations can be classified into four major groups according their phenotype:

a) ultrarapid metabolizers (UM), with more than two active genes encoding a certain CYP450;
b) extensive metabolizers (EM), with two functional genes;
c) poor metabolizers (PM), lacking functional enzyme due to defective or deleted genes; and
d) Intermediate metabolizers (IM), carrying one functional and one defective allele or two partially defective alleles [36].

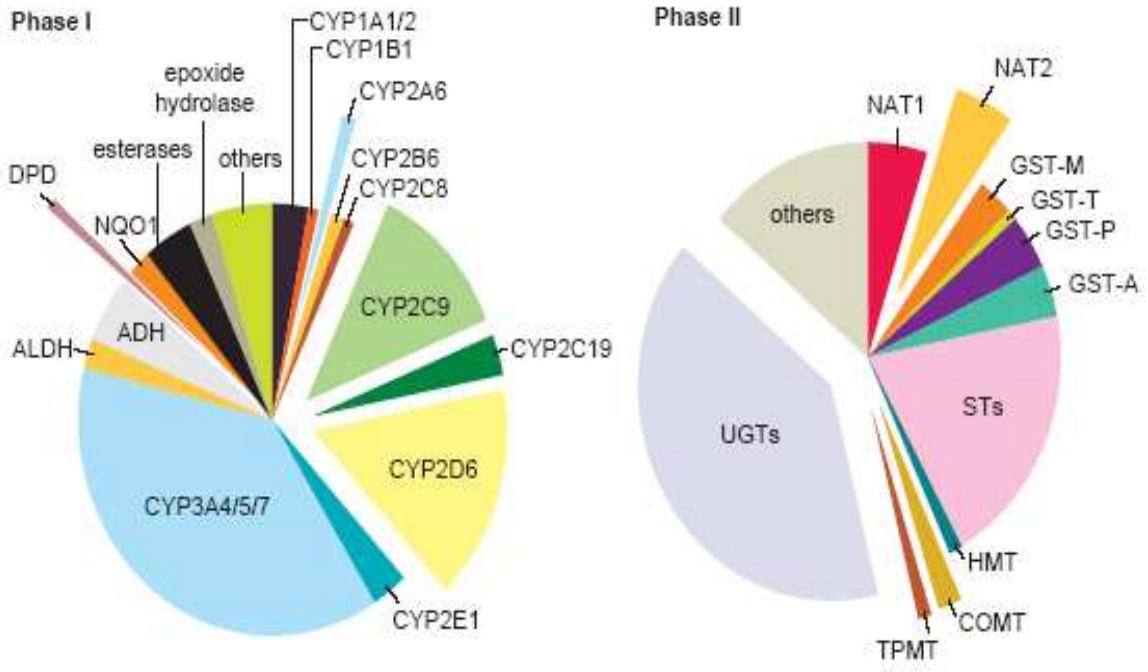


Figure 8: Polymorphic drug metabolizing enzymes [43].

CYP450 enzymes are a group of heme proteins found predominantly in liver, but also in the small intestine, lungs, kidney, and brain [37]. There are currently 58 human CYP genes [38] and more than 350 functionally different CYP alleles recognized [39]. The most

important genes in humans are *CYP1A2, CYP2C9, CYP2C19, CYP2D6, CYP3A4, and CYP3A5*. The functional CYP polymorphisms consist of gene deletions, gene duplications, and deleterious mutations creating inactive gene products, e.g., small insertions and deletions causing frame shift mutations. Furthermore, aminoacid changes can change substrate specificity. Another important aspect is the copy number variation where multiple functional gene copies of one allele can result in increased drug metabolism and reduced drug efficacy in standard doses [36]. Recently, epigenetic regulation of CYP genes has been recognized. Namely, hypomethylation in the promoter and enhancer regions of *CYP1B1* has been observed in prostate cancer and it is associated with enhanced expression of *CYP1B1*. Moreover, promoter methylation of *CYP1B1,* which is a tamoxifen and estradiol-metabolizing CYP450, predicts response in tamoxifen-treated and non-tamoxifen treated patients [40]. Micro-RNA (Mi-RNA) regulation is another epigenetic mechanism of mRNA expression. Mi-RNA regulation of *CYP1B1* has been reported and has been associated with increased expression in malignant tumors [36]. Further research is needed to validate these epigenetic mechanisms and their relevance with inter-individual variations of CYP450. These epigenetic markers could also be used as diagnostic tools for early detection of tumors, as prognostic markers for treatment response as well as in therapeutics [41].

Adverse drug reactions (ADRs) are a significant problem in drug treatment [42]. It is estimated that in USA more than 100,000 annual deaths and 100 billion USD annual cost are associated with ADRs. Polymorphisms in CYP450 enzymes have been associated with increased ADRs and reduced efficacy of a drug treatment. Simplistically, standard drug doses will achieve normal concentrations and effect in homozygous EMs but might be toxic in PMs and ineffective in UMs [37]. Therapeutic window of the drug is also of importance. Drugs with narrow therapeutic window given in PMs are associated with increased ADRs, but drugs with wide therapeutic window given in PMs are associated with increased efficacy, as recommended doses are more than enough. On the contrary, drugs with wide therapeutic window when given in UMs are associated with poor efficacy as standard doses are not enough. We present next few examples of CYP450 polymorphisms with clinical importance and we emphasize on CYP2D6 and CYP2C19 as there is already a diagnostic test for these two genes.

- **CYP2D6**

CYP2D6 or debrisoquine-4-hydroxylase is the first drug metabolizing enzyme reported to be polymorphic [44] and one of the more intensively studied [45]. It is involved in the elimination of 25% of drugs [37]. There are racial differences in function-altering polymorphisms. Table 5 shows the geographical distribution of CYP2D6 allele frequencies.

Table 5 : Geographic differences in CYP2D6 allele frequencies [13].

Allele	Predicted Enzymatic Activity	Japan	China	Caucasian EU	Caucasian US	Black American	Black African	Amerindian	Saudi Arabia	Turkey
*1	Normal	42-43%	23%	33-37%	37-40%	29-34%	28-56%	66%	*	37%
*2	Normal	9-13%	20%	22-33%	26-34%	20-27%	11-45%	19%	*	35%
*3	None	*	1%	1-4%	<2%	<1%	<1%	0%	*	0%
*4	None	<1%	0-1%	12-23%	18-23%	7-9%	1-7%	4%	4%	11%
*5	None	5-6%	6%	2-7%	2-4%	6-7%	1-6%	4%	<1%	15%
*6	None	*	*	<2%	3%	<1%	0%	1%	*	7%
*9	Reduced	*	*	0-3%	7%	<1%	0%	0%	*	<1%
*10	Reduced	39-41%	50-70%	1-2%	4-8%	3-8%	3-9%	1-17%	<1%	6%
*17	Reduced	*	*	<1%	*	15-26%	9-34%	*	<1%	<1%
*41	Reduced	*	*	20%	*	*	*	*	*	*
*1XN	Increased	<1%	*	<1%	<1%	1%	3%	*	*	<1%
*2XN	Increased	<1%	1%	<2%	<1%	1%	3%	*	10%	<1%
*4XN	None	*	*	<1%	<1%	2%	1%	*	*	<1%

Tamoxifen has been the most important drug worldwide for the prevention and treatment of hormone receptor positive breast cancer [46]. So far the only validated markers which predict response are estrogen and progesterone receptors. Tamoxifen is a prodrug which is metabolized to 4-hydroxy-tamoxifen, N-desmethyltamoxifen and endoxifen, a potent metabolite. Lower plasma levels of endoxifen have been shown in CYP2D6 PMs. In a recent phase III clinical trial of 256 postmenopausal women with breast cancer, impaired CYP2D6 metabolism was associated with two fold higher risk of cancer recurrence [46]. In a recent Italian randomized trial of chemoprevention with tamoxifen in 5,408 hysterectomized women the frequency of CYP2D6 *4/*4 allele was higher in women with breast cancer (p=0.015). If this observation will be confirmed by further studies, CYP2D6*4/*4 analysis could be incorporated in a screening panel to better tailor breast cancer treatment and prevention. Genotyping patients for CYP2D6 appears to be valuable to exclude the suboptimal use of tamoxifen in select individuals.

Antiemetics 5-HT3 receptor antagonists have dramatically improved quality of life in postoperative patients and patients receiving chemotherapy. Nevertheless, a significant proportion of patients (up to 30%) do not respond to the antiemetic treatment. Several studies have shown that CYP2D6 UMs experience more vomiting [47]. Another study [48] did not show any difference in ADRs because of the high therapeutic index of the drug. The development of other non-CYP2D6 dependent antiemetics like granisetron as well as new receptor families will probably overcome the need for CYP2D6 genotyping.

Opioids analgesics have long been used to treat acute pain (such as post-operative pain). Codeine is partly metabolized to morphine via CYP2D6, with most being glycuronated to codeine-6-glycuronide and the remainder being metabolized by CYP3A4 to norcodeine [37]. The analgesic effect of codeine is probably mediated by the morphine metabolite. Failure of analgesia has been reported in PMs healthy volunteers. Because of its high therapeutic index the frequency and intensity of ADRs is similar between PMs and EMs. Occasionally, case reports of codeine intoxication with UM have been reported. However,

it is unlikely that genotyping for CYP2D6 will occur routinely in practice, as physicians are familiar with codeine and the main issue is lack of effect in PMs. Alternatively, treatment with oxycodone and hydrocodone which are CYP2D6-non-dependent is an option for these patients.

- **CYP2C19**

CYP2C19 (S-mephenytoin hydroxylase) is responsible for metabolizing 8% of marketed drugs. Three major phenotypic groups exist, PMs, homozygous EMs and heterozygous EMs. A novel variant with UM phenotype has also been recognized. To date, 15 variant alleles of CYP2C19 have been identified, but two alleles CYP2C19*2 and *3 are the most common alleles being associated with affected drug metabolism [43]. Table 6 shows the geographic distribution of the most common alleles.

Table 6: Geographic differences in CYP2C19 allele frequencies [13].

Allele	Predicted Enzymatic Activity	Chinese	Black	Caucasian
*1	Normal	58%	83%	86%
*2	None	35%	16%	13%
*3	None	7%	<1%	<1%

Cyclophosphamide is one of the most widely used anticancer drugs. It is a prodrug that requires activation by multiple CYP450 enzymes, including CYP2C19, 2B6, 2A6, 2C9, 3A4 and 3A5. In a study of 62 patients with lupus nephritis those with CYP2C19*2 allele had a significantly lower risk of premature ovarian failure which is a frequent side effect with cyclophosphamide [49].

Thalidomide has been used in the treatment of multiple myeloma. Recently Li et al have genotyped patients with multiple myeloma for CYP2C19 and have shown that the response rate is higher for EMs than for PMs [50].

- **CYP2A6**

CYP2A6 is responsible for the oxidation of nicotine, which is the major pathway of nicotine metabolism. Some CYP2A6 alleles have been associated with nicotine dependence and therefore with the risk for lung cancer. CYP2A6 oral inhibitors could modulate smoking behaviors and its medical consequences (46).

Thiopurine methyltransferase (TPMT) is the drug-metabolizing enzyme that probably has the strongest case of all for prospective pharmacogenetic testing. TPMT is involved in the metabolism of azathioprine and 6-mercaptopurine, drugs widely used in the treatment of leukemia, autoimmune diseases and in transplantation patients. Approximately 1% of the general population has TPMT variants with enzyme deficiency. These patients will develop life-threatening myelosuppression if they receive standard doses. Some clinical trials have already incorporated the prospective TPMT genotyping and dose individualization of thiopurines [37].

Dihydropyrimidine dehydrogenase (DPD) metabolizes the pyrimidine analogue 5-fluorouracil (5-FU) which is widely used to treat solid tumors. Some patients are more susceptible to develop severe side effects, like gastrointestinal symptoms, myelosuppression and neurotoxicity. Case reports show that partial and absolute DPD deficiency increases the risk. Genotyping for DPD could identify those with absolute DPD deficiency and select them for alternative chemotherapy or those with partial deficiency and treat them with a lower dose. Moreover, high intratumoral DPD activity has been associated with poor outcome [37].

Uridine diphosphate glycuronosyltransferase UGT1A1 variants have been associated with increased toxicity with irinotecan. Irinotecan is widely used in the treatment of colorectal and lung cancer. UGT1A1*28 is the variant most frequently implicated in defective glycuronidation, which is the primary cause of Gilbert syndrome. A significant proportion of the population up to 25% is homozygous for this variant. UGT1A1 testing is the first pharmacogenetic FDA approved test for use in conjuction with irinotecan [37].

3. EHR: State of the Art

From the previous chapter it is clear that we are currently witnessing a significant increase in available genomic information, which follows the general trend of information explosion in all healthcare-related research areas (clinical research, bioinformatics, etc.). However, so far, this has had very little impact on the clinical practice and on the actual management of individual patients. Some important causes for this low impact are the lack of infrastructure to deliver the information, the lack of common, agreed-upon ways to represent, process, store and manage the information, and the unavailability of end-user tools to make this information efficiently accessible to healthcare professionals, while filtering irrelevant data to avoid information overflow, and to allow clinical research communities to share their results and to bring them within reach for healthcare organizations. When cooperation is supported across organizations, each research result can be evaluated based on large amounts of patient data. Also, in the context of personalized medicine, a meaningful and thorough patient stratification requires the comparison of large amounts of data from many patients and different patient populations.

As we show in the following sections, with the penetration of EHR systems the healthcare domain is moving towards fully paperless storage and management of patient data. As genomic data plays an increasingly important role in the cancer care cycle and in the cancer-related clinical research, any future EHR should be able to model, manage, store and provide access to that type of data.

In this section we first define an EHR and draw its scope. Since EHRs need to be cross-institutional, longitudinal sets of records that enable continuity of care, it is clear that their adoption and long-term use is depends on the existence of appropriate standards for information representation and exchange. Therefore, we also give an overview of the most prominent EHR standards, and conclude the section with a discussion on the way in which genomics information should be modeled given the current state of the art.

3.1. Definition and scope

The importance of EMR and EHR systems for achieving high quality and efficient patient care and for reducing medical errors is widely recognized in the healthcare area.

However, what electronic health and/or medical records (EHR, EMR, respectively), or their predecessor, the computerized patient record (CPR), actually are and what they involve (requirements, functions, features, etc.) are still subjects for debate, as there are many different and conflicting views at the level of industry vendors, government and regulatory bodies, and standardization organizations. Even when a same definition is agreed upon, different vendors can provide different sets of features under the same name. In fact, many vendors use these two terms interchangeably.

This makes the task of selecting and purchasing a patient record management system extremely difficult and confusing for a healthcare organization, as the definition is unclear and the functions and features that should be provided by an EHR or EMR system are not standardized. This is problematic because a mistake, e.g. purchasing a system that does not provide the desired functionality or choosing one of the many vendors that are not going to survive in the next years, can be extremely costly.

The main idea of all these concepts (EHR, EMR, CPR) is principally the same, which is to provide electronic access to the medical record of the patient. People started talking about something called the electronic health record in the 60s, but at the time the idea could not take off, as computers were practically nonexistent [51]. The concept reappeared 30 years later when the Institute of Medicine introduced a more precise concept of the electronic patient record and its importance to future medicine. The report promoted the concept of a

computer-based, longitudinal, life-long, integrated patient record including entries from all healthcare providers where a patient was known.

The benefits of moving from a paper-based system of patient record to an electronic patient record were obvious at the time, as they are now:

- Simultaneous, remote access to (distributed) patient data by all authorized providers.
- Faster access to data, more efficient and at lower costs.
- Seamless communication among providers.
- Reduction of medical errors, which results in better health care and lower cost.
- Improved patient privacy and confidentiality and easier monitoring of access, e.g., to detect abuses.
- Flexible data layout and therefore easier integration with other information resources.
- Incorporation of various related electronic data and records, which may be continuously processed and updated.
- Easier searching and finding of diverse relevant clinical data.
- Scalable, cheaper storage and easier maintenance and management of the data.

However, no system providing all these features exists yet.

The Computerized Patient Record (CPR) was the first attempt at electronic records. A CPR is defined a computer-based medical record system that includes all information (clinical and administrative) for a patient and for all healthcare organizations and physicians who have treated that patient. The concept and vision initially failed, mainly due to the lacking of electronic data standards and to the fact that information systems were disparate. Even today, seamless communication among information systems of distinct healthcare providers is very unlikely and integration is very difficult. Another sensitive issue still unsolved is related to privacy and access rights to the data. Providing a unique identifier to each person to be used for all health episodes is also not feasible as it brings more problems than it solves (protect that information from id theft, avoid abuses, etc.). In [1] a CPR is defined as a record about an individual patient stored in a healthcare provider's computer, in a database that is typically the property of the provider. It will usually only contain the patient's demographic data and medical information collected when the patient visited that provider.

An Electronic Medical Record (EMR) is a newer concept that extends the capabilities of a CPR. An electronic medical record system is an organized collection of all records about an individual patient stored in the computer systems and databases of all the providers who have provided care to that patient within one enterprise. The EMR is not stored on any one individual computer, but is assembled dynamically, in real time, from various systems when needed. According to [51], the EMR eliminates the optimistic goal of the original CPR concept of universal access and aims at a more achievable approach of sharing patient information from different systems among authorized professionals within an organization. The new term was introduced because the all-comprising original idea of the CPR was lost and the term had become synonymous with old technology. However, the term EMR is used very inconsistently in practice, as it often refers just to medical records created and stored in an electronic format, being quite similar to the CPR.

EMRs are defined by software manufactures, and they define them by features, not as an approach to patient documentation. A medical records manager's view of an EMR is considered the closest to what it should be: a system that securely stores patient information, may be accessed by the staff from various locations, and can import / export or consolidate information to and from other systems such as practice management

systems, scheduling, billing, imaging, etc. Most clinicians look at it as a possible time-saver because many EMR systems compliment dictation with point-and-click entry, prescription writing, and some automated report generation. The definition of an EMR within today's market place varies as much as there are products available. There is no EMR standard today.

The Electronic Health Record (EHR) avoids some of the controversial aspects of the CPR concept as it was initially introduced, such as the unique identifier. It also has a more realistic approach towards the exchange of information among various organizations while promoting full interoperability. However, despite the effort of many institutions and bodies, the process of defining data standards required by an EHR is complex and still not within reach.

As defined by [51], an EHR (electronic health record system) describes a longitudinal record of patients' health healthcare - from cradle to grave. It combines both the information about patient contacts with primary healthcare as well as subsets of information associated with the outcomes of periodic care whether held in EMRs, CPR's or other information systems.

In [52] it is recognized that the views on EMR/EHR systems are often inconsistent. However, they advocate a clear differentiation between the two concepts. EHRs are considered reliant on EMRs being in place, while EHRs would never reach their full potential without interoperable EHRs in place. The EMR is defined as the legal record created in hospitals and ambulatory environments. The EMR is also the source of data for EHR. In more detail, they define an EMR as an application environment composed of the clinical data repository, clinical decision support, controlled medical vocabulary, order entry, computerized provider order entry, pharmacy, and clinical documentation applications. This environment supports the patient's electronic medical record across both inpatient and outpatient environment and is used by healthcare practitioners to document, monitor and manage healthcare delivery within a healthcare organization. This definition fits with the one introduced in [51]. On the other hand, HIMSS Analytics define an EHR as comprising a subset of each healthcare organization's EMR, assumed to be summaries like ASTM's Continuity of Care Record (CCR) or HL7's Continuity of Care Document (CCD), it is owned by the patient and has patient input and access that spans episodes of care across multiple healthcare organizations within a community, region, or entire country. According to [52], the EHR represents the ability to easily share medical information among stakeholders and to have a patient's information available at all points of care visited by the individual. The stakeholders are composed of patients/consumers, healthcare providers, employers, and/or payers/insurers including the government. This definition creates a clear separation between EMR and EHR, but blurs the difference between EHR and yet another (newer) concept, the Patient Health Record (PHR). Moreover, in this case the EHR environment relies on functional EMRs that allow healthcare organizations to exchange data/information with each other, or with other stakeholders as defined above. This puts the burden of achieving interoperability on the EMR and assumes that the healthcare organizations would be willing to pay for it and the EMR vendors will provide it, which is not achievable in the current context.

In such a complex and heterogeneous landscape, the advice in [51] for healthcare organizations is to move to an electronic record system, but in small steps. The focus should be on internal communications of systems, while looking at interoperability with outside systems later. And as things will change rapidly and many sellers may no longer be around in a few years, they should spend as little money as possible and not consider extravagant features.

The approach suggested in [51] would suit the needs of healthcare organizations and help them avoid spending a lot of money on a system that will disappear from the market, but at the same time does not help with achieving the vision of an all-comprising interoperable EHR system. Other evidence also supports such an approach, despite the fact that it does not fit in the long tern EHR vision.

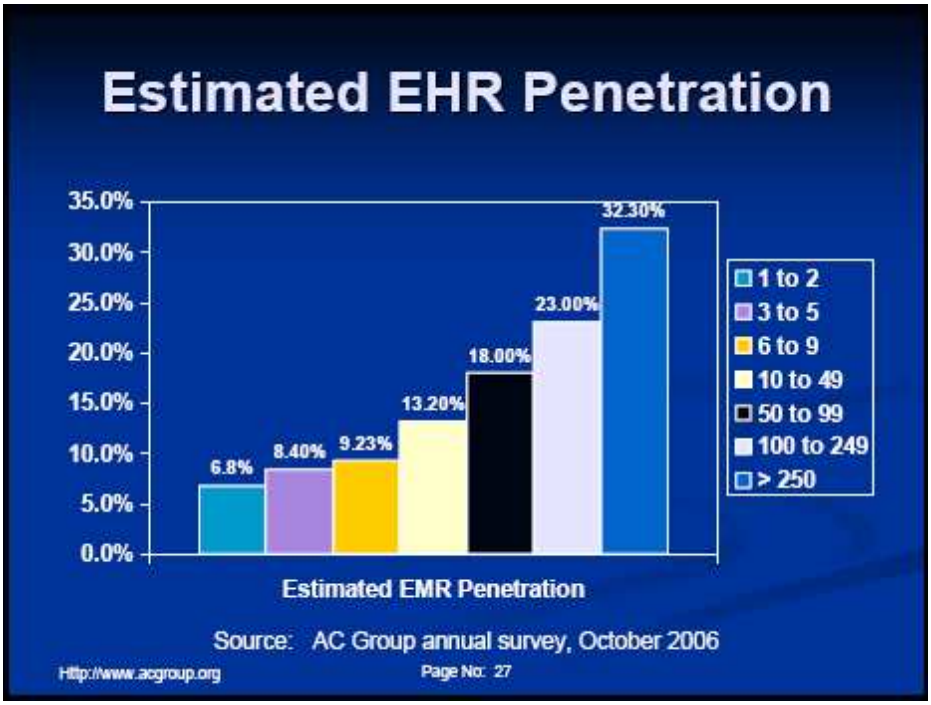


Figure 9: EHR/EMR penetration in big hospitals according to the AC Group Survey, October 2006

The AC Group annual survey [53] estimates (see Figure 9) that the EMR/EHR penetration in large hospitals (>250 clinicians) in the US is 32.3%, for hospitals with 100 to 249 clinicians is 23%, for 50 to 99 is 18%, while for small sites with 3 to 5 clinicians the penetration rate is only 8.4%. (October 2006). The same survey estimates that the deployment of EMRs will continue but not earlier than 2012 will penetration reach 80% of the healthcare organizations in the US. A similar trend can be expected in Europe as well.

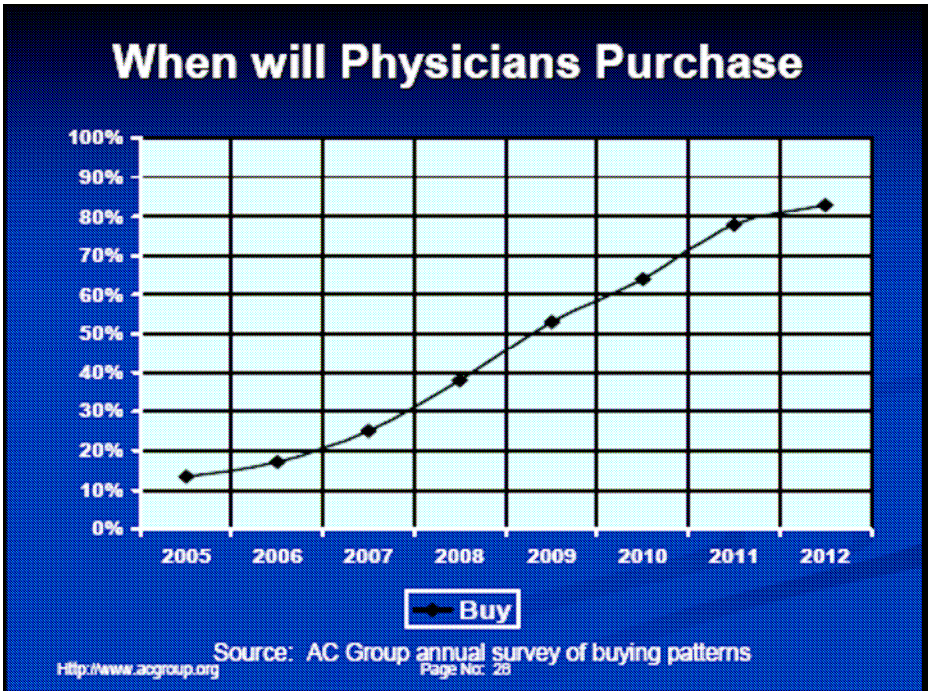


Figure 10: Estimated adoption of EHR/EMR solutions

The study also estimates that there will be an implementation gap between the demand for EHRs and the sales of EHR solutions, as shown in Figure 10. Still, the offer of EHR solutions on the US market is rather large, with over 380 providers claiming to sell EHR. However, as indicated in Figure 11, it is expected that a large majority of those providers will disappear before 2012.

Many EHR/EMR implementations fail for reasons such as poor planning, unrealistic expectations, lack of physician or healthcare provider support, high entry barrier, complexity, wrong selection of product, and workflow incompatibility.

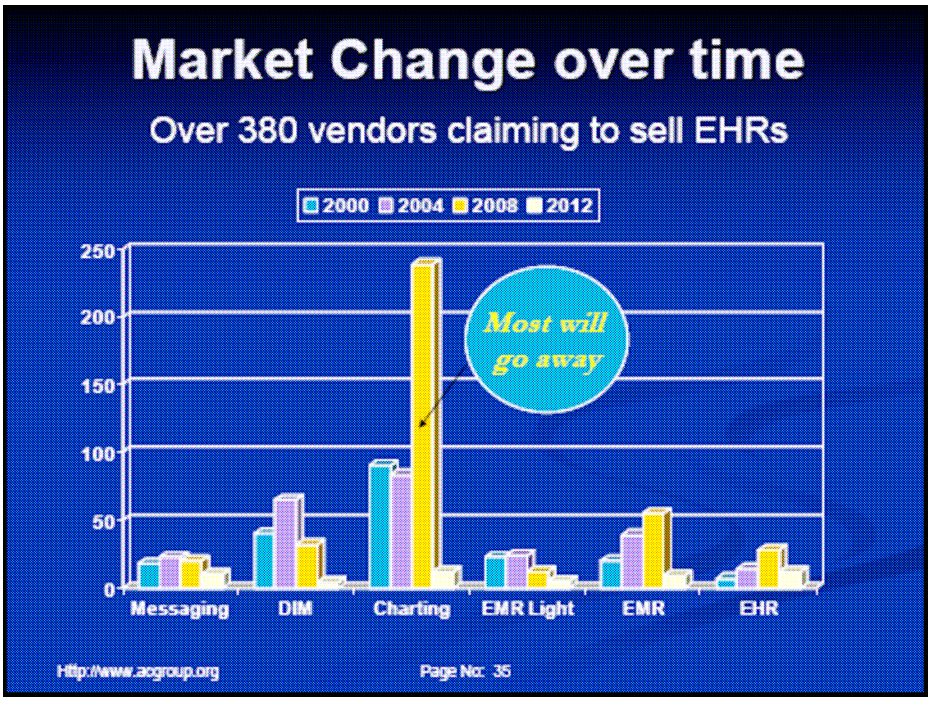


Figure 11: Distribution of the volume of vendors selling EHR/EMRs over time

There are several types of products on the market covering a different set of functions, and each organization should select the desired system based on the set of functions it needs. In this context, only a relatively small percentage of organizations will deploy a full-fledged EHR or EMR, the rest selecting a simpler, lower cost solution such as an EMR Lite; see also Figure 12 and Figure 13. Even more than the availability of many vendors, this diversity of products with respect to the supported functions makes standardization difficult and integration into an all-comprising EHR unrealistic.

Market Segmentation

Function	PMS	Secured Message	DIM	Charting	EMR Lite	EMR	EHR
Billing	X						
Scheduling	X	X	X	X	X	X	X
Labs		X	X	X	X	X	X
Transcription		X	X	X	X	X	X
Paper Doc			X	X	X	X	X
E-Prescribe				X	X	X	X
E & M Coding					X	X	X
Standards/CCR						X	X
National Alerts						X	X
Chief Complaint						X	X
Health Maint.							X
PHR							X

Figure 12: Market segmentation based on selling features of an EHR/EMR

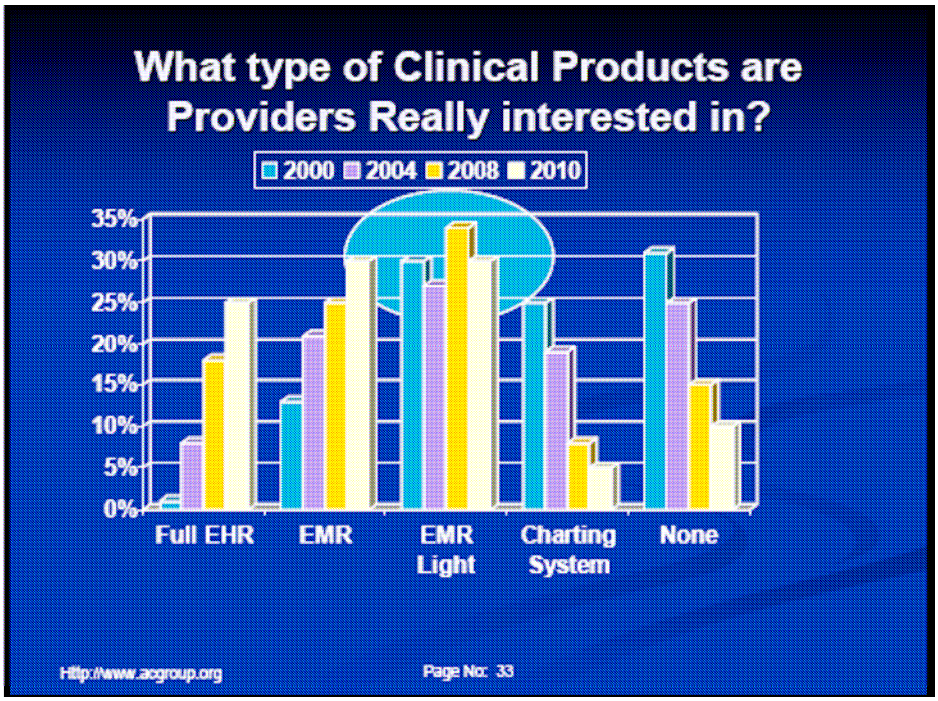


Figure 13: The estimation of desired features of an EHR or an EMR over time

3.2. State of the art EHR standards

A number of initiatives have been established [54-56] with the goal to define key characteristics of an EHR. Although a lot of effort is being invested in standardization activities in the domain of health information systems, an agreement upon common standards by the active standard development organizations has not been reached yet [4]. Standardization efforts differ not only on the international level, e.g., the United States and Europe have different standardization bodies, but also on a national and hospital level. Moreover, vendors often have their own clinical data models [57].

A prominent standard development organization relevant for EHRs in the United States is Health Level Seven (HL7), which is one of the accredited standards developing organizations of American National Standards Institute (ANSI) operating in the healthcare arena [4, 58]. The domain of the HL7 is both clinical and administrative data. HL7 is an international standard and as such it is also relevant in Europe for hospital, radiology, and cardiology information systems.

The European Standardization of Health Informatics (CEN/TC 251) organization has been active in defining a standard EHR architecture (international health information standard CEN ENV 13606 [59]).

Another influential international (non-profit) organization that is working toward open, interoperable, long-life EHRs is *open*EHR. Founded by stakeholders in the UK and Australia, *open*EHR publishes all its specifications and their corresponding reference implementations as open source software [60].

The Digital Imaging and Communications in Medicine (DICOM) Structured Reporting (SR) standard [61-63] also plays a role in how EHR information is structured as it defines a document structure for encoding and exchanging information.

As HL7, *open*EHR, and DICOM are often mentioned in literature as the relevant EHR standards [64], we briefly describe the key characteristics of these standards.

3.2.4. Messaging Paradigm: The HL7 Standard

The HL7 standard has initially been developed to address the problem of healthcare information exchange. As such, HL7 consists of a set of pre-defined logical formats for packaging healthcare data in the form of messages, which are transmitted among computer systems. An HL7 message is essentially a collection of data containing information about an event in a healthcare enterprise. The HL7 committee has compiled a collection of message formats and related clinical standards that loosely define an ideal presentation of the clinical information. Together, the HL7 standards provide a framework for data exchange. The name HL7, Health care Level 7, indicates that the HL7 standards are located at the highest level (seventh) of the International Organization for Standardization (ISO) [65] communications model for Open Systems Interconnection (OSI). The OSI seventh layer addresses definition of the data to be exchanged, the timing of the interchange, and the communication of certain errors to the application. The seventh level supports functions such as security checks, participant identification, availability checks, exchange mechanism negotiations and, most importantly, data exchange structuring.

Reference Information Model

The HL7 Version 3 (HL7 V3) prescribes the message development process based on astatic model of clinical data (domains) called the Reference Information Model (RIM). RIM explicitly represents the connections that exist between the information carried in the fields of HL7 messages.

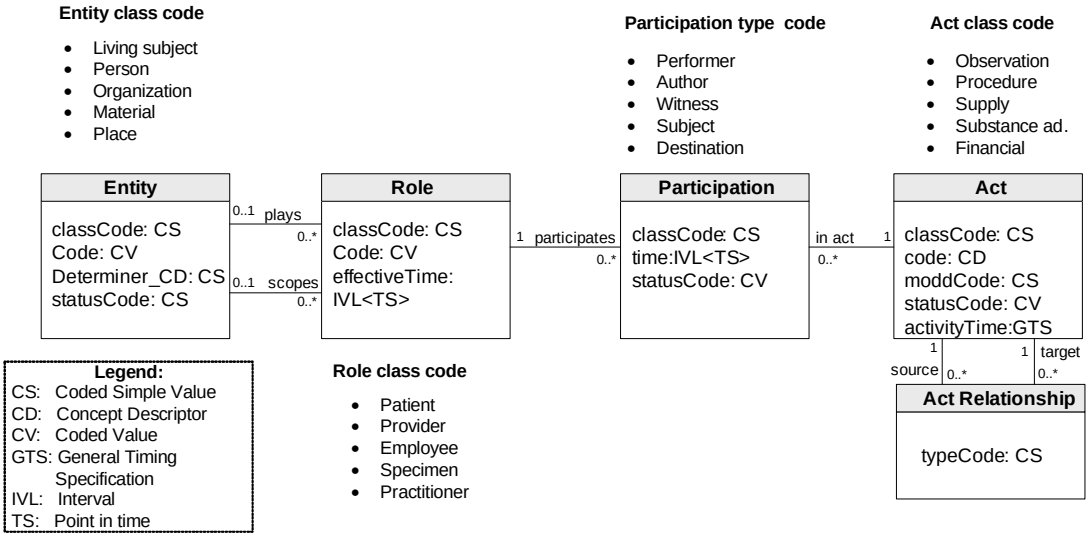


Figure 14 The Most relevant classes of the HL7 Reference Information Model (RIM) [12]

RIM defines the two main classes of healthcare interactions, Entity and Act (see Figure 14). Entity (also referred to as the player, actor, or stakeholder in healthcare) can include all living subjects, formal and informal organizations, various materials, as well as the places that may be of interest in a healthcare messaging context. Acts represent all

actions involving entities in either a clinical or administrative context. In RIM, as illustrated in Figure 14, two additional classes, Role and Participation, are placed between Act and Entity. Role captures the fact that an Entity, e.g., patient, primary care physician, and registered nurse, may temporally assume one or more Roles in a particular healthcare context. Secondly, the concepts of "capability", e.g. Advanced Cardiac Life Support, and "certification", e.g. Licensed Practical Nurse, are also modelled using instances of the Role class.

The two remaining classes in the RIM, Act_relationship and Role_link (not shown in the Figure), are used to associate or link instances of the class with which it is associated. The class Role_link is also used to establish a dependency-based link, such as accountability, chain-of-trust, etc., between two instances of an Entity-in-a-Role.

RIM also includes vocabulary tables of over 5,000 data attributes. For clinical content needs, the HL7 vocabulary tables refer to external terminology sources, such as SNOMED, ICD-10, and MEDCIN.

Functional Model

The HL7 standardization body has also developed a functional model for EHRs. The functional model defines a standardized set of functions that might be present in an EHR. The functions that are considered essential in at least one care setting are included in the specification. Each function is specified on a high level, in an implementation-independent manner. The functions are organized into three layers: direct care functions, supportive functions, and information infrastructure, as shown in Figure 15. Orthogonal to these layers is a functional profile, which defines the context of the functions in the layer and their priorities for a specific instance (e.g., ambulance care vs. clinical care). The functional profile is the one that is implemented in an EHR or a system of EHRs; the standard does not make a difference between an EHR and the system of connected EHRs.

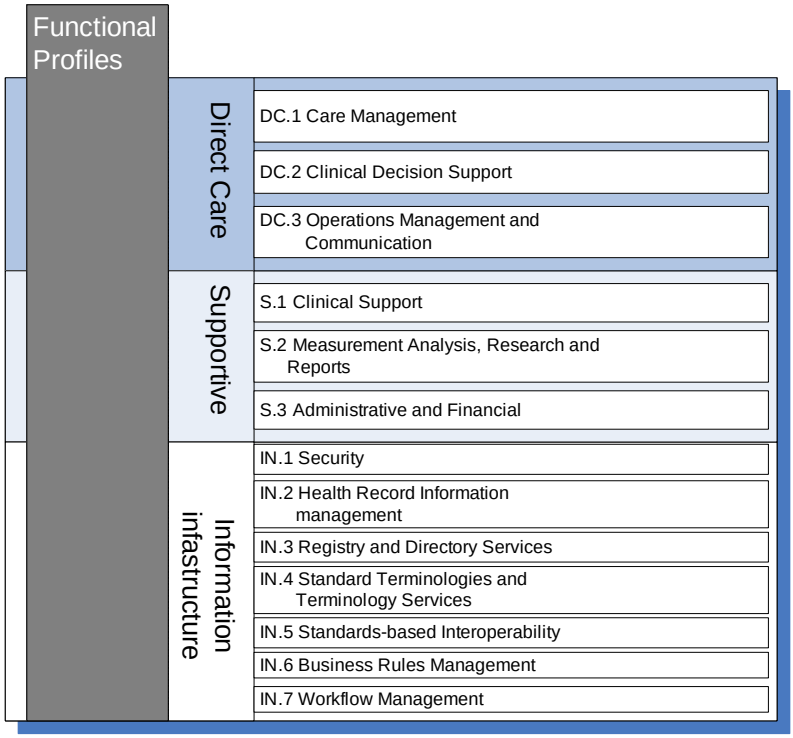


Figure 15. HL7 Functional model overview

Within each layer and each header in a layer in Figure 15 (DC.1-DC.3, S.1-S.3, and IN.1-IN.7), a number of functions are defined in the Functional Reference model documentation. For a full description of the HL7 defined functions we refer interested readers to [56].

Clinical Document Architecture

A Clinical Document Architecture (CDA) refers to the document markup standard that specifies the structure and semantics of a clinical document, e.g., a discharge summary or progress note, for the purpose of exchange [66]. CDA documents use RIM HL7 V3 data types. A CDA document can include text, images, sounds, and other multimedia content, and it can be transferred within a message or can exist independently, outside the transferring message.

CDA documents are encoded in eXtensible Markup Language (XML) as follows. Each document consists of a header and a body. The *header* identifies and classifies the document and is consistent across all clinical documents. The header also provides information about patient, provider, and encounter. The *body* part is divided in section and can contain narrative text and multimedia content, optionally augmented by coded equivalents (see Figure 16). Arrows in Figure 16 are Act Relationships, such as has component, derived from, etc. Entries are coded clinical statements, e.g., Observation, Procedure, Substance administration, etc.

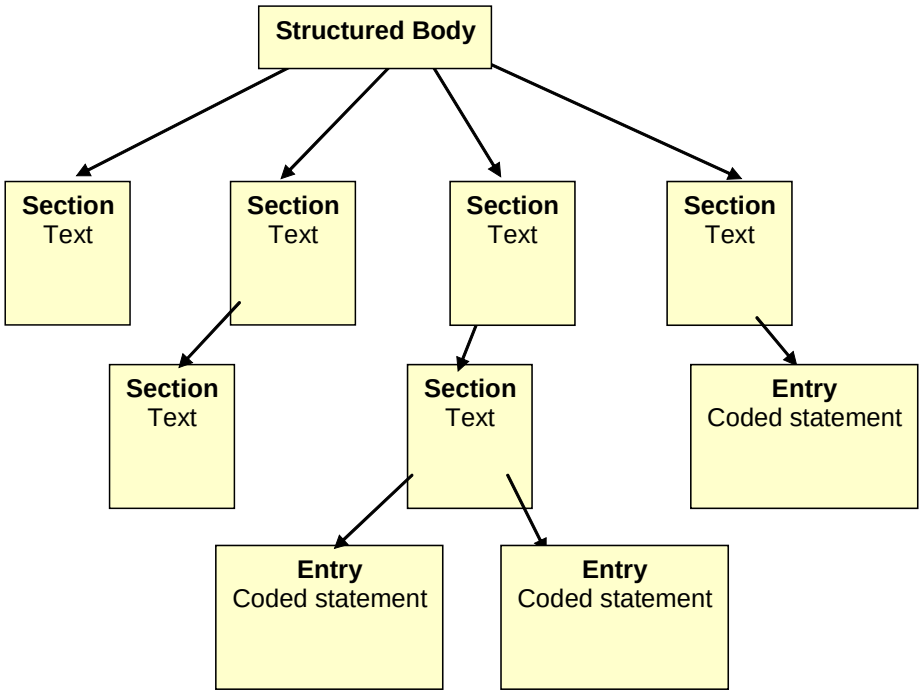


Figure 16 The Structured body of the CDA

A simplified example of a CDA document is illustrated in Figure 17. As can be noted, the header lies between the <ClinicalDocument> and the <structuredBody> tags. The body of the document contains the clinical report and can be entered either as unstructured text, or can be comprised of structured markup. In Figure 17 a structured body is wrapped by the <structuredBody> tag. This structured body is divided into recursively nestable document sections. Each section (wrapped by the <section> tag) contains one narrative block and a number of CDA entries and external references. CDA requires human-readable narrative blocks, as that is needed to reproduce the legally attested clinical content. Hence, the narrative block is a critical component of CDA and must contain the human readable content that can be rendered. The ‘‘originator’’, i.e., the Role responsible for creation of a conformant CDA document, must ensure that the attested portion of the document body is conveyed in narrative blocks such that a recipient, adhering to recipient rendering rules, will correctly render the document. This process ensures human readability, and enables a recipient to receive a CDA document from anyone and faithfully render the attested content using a single style sheet.

CDA entries in a <section> are structured content provided for further computer processing, e.g., in decision-support applications. CDA entries typically encode content present in the narrative block of the same section. Two <observation> CDA entries are also depicted in Figure 17. These entries are derived from classes in the RIM and enable formal representation of narrative clinical statements.

```
<ClinicalDocument>
   …CDA header …
   <structuredBody>
      <section>
         <text> (narrative block comes here) </text>
         <observation>…</observation>
         <substanceAdministration>
      <supply>…</supply>
         </substanceAdministration>
         <observation>
            <externalObservation>...
            </externalObservation>
        </observation>
      </section>
   </structuredBody>
</ClinicalDocument>
```

Figure 17 The Major components of the CDA, adopted from [66]

Guidelines for creating CDAs are documented in so-called Implementation Guides, which are a part of HL7 Informative Documents. Each Implementation Guide has a Template ID attribute that is included in the root element of the conforming document.

In conclusion, CDA is being regarded as one of the primary candidate formats for diagnostic reports and medical summaries due to its characteristics, such as allowing a smooth transition from free text to coded content, and referencing external documents such as images, signals, films, and enabling easy display (simple style sheet).

3.2.5. Archetypes paradigm: The *open*EHR Standard

The key concept introduced by *open*EHR is the concept of *archetypes* [67]. The archetype is defined as [67]: *a model defining some domain concept, expressed using constraints on instance structures of an underlying reference model*. The notion of archetypes relies on separating knowledge and information levels in the information system.

An archetype is normally used to describe the domain knowledge, which in turn defines valid information structures. An archetype can be considered to be a type of a template [68] as it represents the original pattern or model of which all other things of the same type are representations or copies.

Without archetypes the knowledge concepts are directly encoded into models and are used to build software and databases. Consequently, the knowledge concepts become names and attributes of classes, tables, and columns. Hence, the domain knowledge is hard-coded, via technical development, into an application system, as shown in Figure 18. EHRs built with such a process are typically not stable, as changes in the knowledge definition typically require the system to be re-built, re-deployed, and re-tested, with extensive data validation and migration [67].

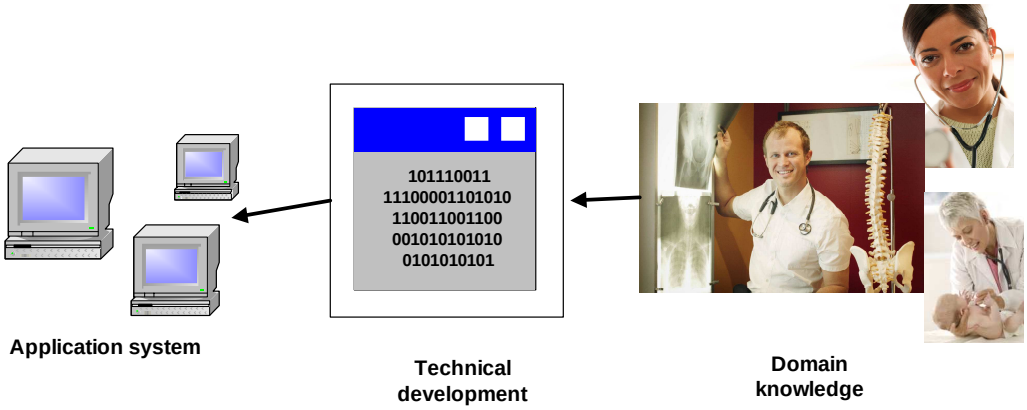


Figure 18 Traditional ERH approach

With the introduction of archetypes, the complete separation of the information model, e.g., object models of software or models of database schemas, from the domain model is achieved (see Figure 19). The key benefit of archetypes is that they enable the domain knowledge to evolve without direct coupling to the technical development and vice versa.

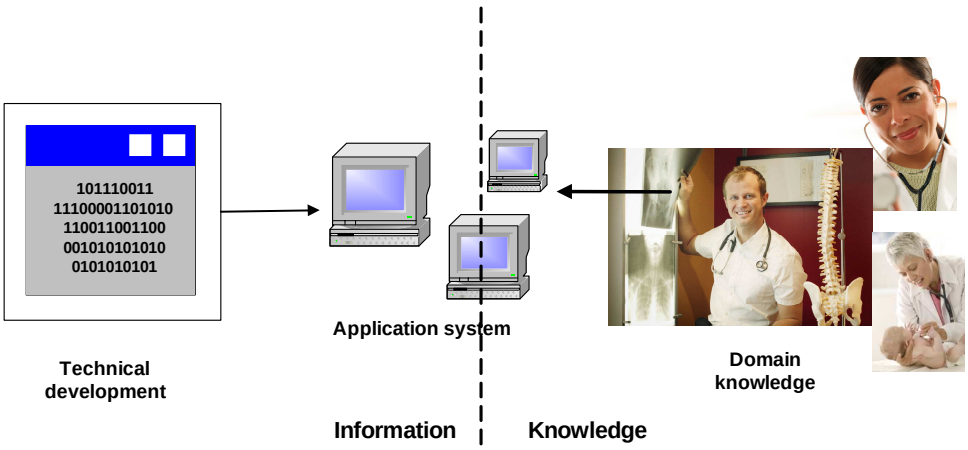


Figure 19 The two level *Open*EHR approach

The (semi-final) list of the available archetypes is given in Figure 20. Note that with the two-level modelling approach, the usability of an EHR system critically depends on the availability of suitable archetypes. Hence, representing a disease or a domain, e.g., breast tumour, can include many archetypes and also could require extensive mapping to existing archetypes and possibly adding new archetypes.

A formal language, called Archetype Definition Language (ADL), for expressing archetypes has been introduced by the *Open*EHR initiative [69]. Three other syntaxes within ADL are used to describe constraints on data that are instances of an information model (see Figure 21):

- cADL, which is a constraint from ADL,
- dADL, which is a data definition form of ADL, and
- a version of first-order predicate logic (FOPL).

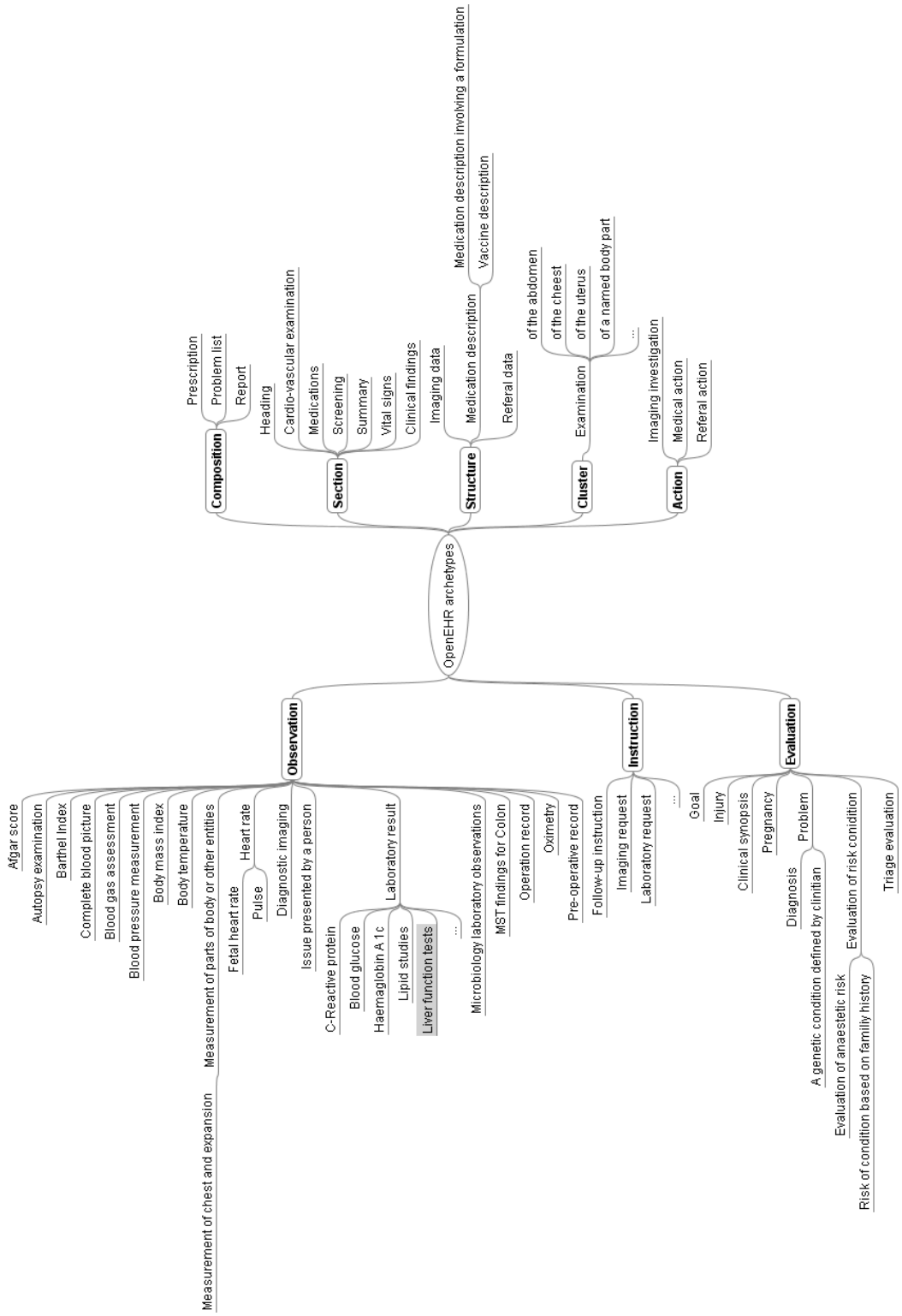


Figure 20 The OpenEHR archetypes

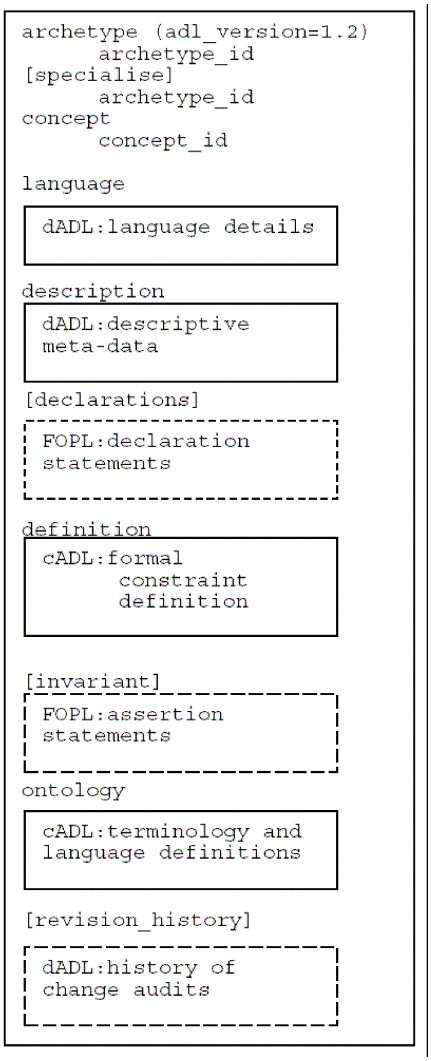


Figure 21 ADL archetype structure, taken from [70]

As shown in Figure 21, an archetype is divided into three main parts: descriptive data, constraint rules (definition part), and ontology definitions [70, 71]. The *descriptive part* contains a unique identifier, a machine-readable code describing the clinical concept modeled by the archetype, and various metadata including the version, the author's name or the organization, the language used or the purpose of the archetype, etc. The *definition part* defines constraint rules and represents the main part of an archetype. Constraint rules define restrictions on the valid structure, cardinality, and content of EHR component instances complying with that archetype. The *ontology* part defines the controlled vocabulary (i.e., machine readable codes), which can be used at specific places in the instance of an archetype. An example of the ADL description of the observation archetype related to imaging is shown in Figure 22. The descriptive data marked with the first box in the figure indicates the author and the purpose of archetype, and also information to which imaging devices this observation could relate. The constraint rules are then defined for various imaging devices. The third box shows the ontology part that defines dictionary for imaging, including category of the images, descriptions related to observations, etc.

```
archetype (adl_version=1.4)
          openEHR-EHR-OBSERVATION.imaging.v1

concept
          [at0000]  -- Diagnostic imaging
language
          original_language = <[ISO_639-1::en]>
description
          original_author = <
                    ["name"] = <"Sam Heard">
                    ["organisation"] = <"Ocean Informatics">
                    ["date"] = <"26/03/2006">
                    ["email"] = <"sam.heard@oceaninformatics.biz">
          >
          details = <
                    ["en"] = <
                              language = <[ISO_639-1::en]>
                              purpose = <"For recording findings found at diagnostic
imaging">
                              use = <"">
                              keywords = <"Xray", "X-ray", "radiology", "scan",
"ultrasound", "MRI", "CT", "CAT", "nuclear">
                              misuse = <"">
                    >
          >
          lifecycle_state = <"Initial">
          other_contributors = <>
definition
          OBSERVATION[at0000] matches {         -- Diagnostic imaging
            data matches {
              HISTORY[at0001] matches {         -- history
                    events cardinality matches {1..*; unordered} matches {
                    EVENT[at0002] occurrences matches {0..*} matches-- Any event
                      data matches {
                         ITEM_TREE[at0003] matches {      -- Tree
                           items cardinality matches {0..*; unordered} matches {
                             ELEMENT[at0008] occurrences matches {0..1} matches {
-- Clinical value matches
                                                  DV_TEXT matches {*}
                                                                    }
                                                            }

          CLUSTER[at0016] occurrences matches {0..1} matches {     --Imag. details
            items cardinality matches {0..*; unordered} matches {
              ELEMENT[at0014] occurrences matches {0..1} matches { -- Test name
                    value matches {
                      DV_TEXT matches {*}
                                        }
                                        }
              ELEMENT[at0004] matches {         -- Category

                     DV_CODED_TEXT matches {

                              [local::

                              at0006,   -- Ultrasound
                                        at0007,   -- Nuclear medicine
                                        at0012,   -- CT-Scan
                                        at0013]   -- MRI

                                        }
                                                  }
                                                  }


                                                  ...
```

```
ontology
          term_definitions = <
            ["en"] = <
                    items = <
                    ["at0000"] = <
                    description = <"Findings associated with diagnostic imaging">
                                                text = <"Diagnostic imaging">
                                >
                                ["at0001"] = <
                                          description = <"@ internal @">
                                          text = <"history">
                                >
                                ["at0002"] = <
                                          description = <"*">
                                          text = <"Any event">
                                >
                                ["at0003"] = <
                                          description = <"@ internal @">
                                          text = <"Tree">
                                >
                                ["at0004"] = <
                                          description = <"The category of imaging">
                                          text = <"Category">
                                >
                                ["at0005"] = <
                                          description = <"Imaging performed using Xray">
                                          text = <"Xray">
                                >
                                ["at0006"] = <
                                          description = <"Imaging performed using
ultrasound">
                                          text = <"Ultrasound">
                                >
                                ["at0007"] = <
                                          description = <"Imaging using radio-isotope
scans">
                                          text = <"Nuclear medicine">
                                >
                                ["at0008"] = <
                                          description = <"Description of the clinical
findings">
                                          text = <"Clinical">
                                >
```

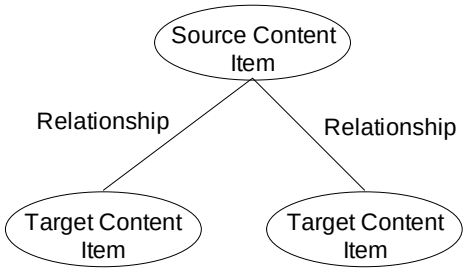
Figure 22: An example of the ADL for the observation.imaging archetype

3.2.6. DICOM Structured Reporting

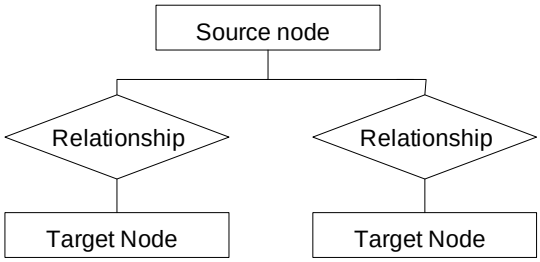
Examinations that involve modalities, including examinations in cardiology, oncology, orthopaedics, surgery, etc. normally use quantitative measurements and descriptive observations during examination and diagnosis, e.g., spatial coordinates of important findings, flow measurements, and volumes. Such quantitative metrics are typically objective and reproducible, and often included in the final report. Traditionally the majority of clinicians handled such information on paper, transcribing hand-written worksheets, and occasionally with proprietary standards or text transferred via a serial port [63].

Digital Imaging and Communications in Medicine (DICOM) Structured Reporting (SR) is a standard that defines document structure for encoding and exchanging information using the DICOM hierarchical structure and services [61-63]. DICOM is the standard that has

been developed to enable interoperability of medical imaging equipment. With DICOM SR a structured document is defined more by how it is constructed than by what it contains. Hence, a DICOM SR document can convey any kind of structured content, not just reports [78]. SR documents can be used wherever there is a need for lists or hierarchically structured content, or a need for coded concepts or numeric values, or a need for references to images, waveforms, or other composite objects. Hence, the key area of development in SR is evidence data from imaging procedures, such as catheterization laboratory modalities, ultrasound, echocardiography, and computer-aided diagnostics (CAD) [63]. To that end, DICOM SR introduces DICOM Information Object Definitions (IODs) and services used for the storage and transmission of structured reports. DICOM IODs define the data structures describing real-world objects, e.g., patients, images, and reports that are involved in radiology operations.



(a) An SR tree diagram



(a) An SR Entity-Relationship diagram

Figure 23 Diagrams of SR trees [62]

The structure of DICOM SR, i.e., the specification of IODs, in essence consists of a directed acyclic graph, i.e., sequence of `Content Item` nodes linked with `Relationships`, as shown in Figure 23. Each `Content Item` is represented with a name/value pair. The name refers to a single `Concept Name` that is defined by code (not with free text) to enable indexing and searching. Hence, each concept name is a coded entry that uses triplet encoding attributes as follows [61, 62].

- The code value, e.g., `Code=121071`, which is computer readable and searchable identifier.
- The code scheme designator, e.g., `Coding Scheme=DICOM`, which is the identifier of the coding organization.
- The code meaning, e.g., `Concept Name=Finding`, in which human-readable text is entered.

As an example, consider the following textual outcome of the examination:

"A mass of round shape and irregular margin, 15mm in diameter, was found in the central region of the breast. The mass is classified as malignant due to the irregular margin."

The physician could store the same report in DICOM SR format, as shown in Figure 24. DICOM SR would form a graph of 7 nodes and 7 named links, with 1 block of free text (`Malignant mass`), a number (`15mm`), and 13 machine-readable codes taken from four different vocabularies. For example, the names of the items, e.g., `Finding`, `Shape`, `Size`, and `Margin`, in Figure 24 could be encoded as follows:

Finding	Shape	Size	Margin
Code=121071 Coding Scheme=DICOM Concept Name=Finding	Code=M-020F9 Coding Scheme=SNOMED3.0 Concept Name=Shape	Code=112025 Coding Scheme=DICOM Concept Name=Size Descriptor	Code=G-A428 Coding Scheme=SNOMED-RT Concept Name=Margin

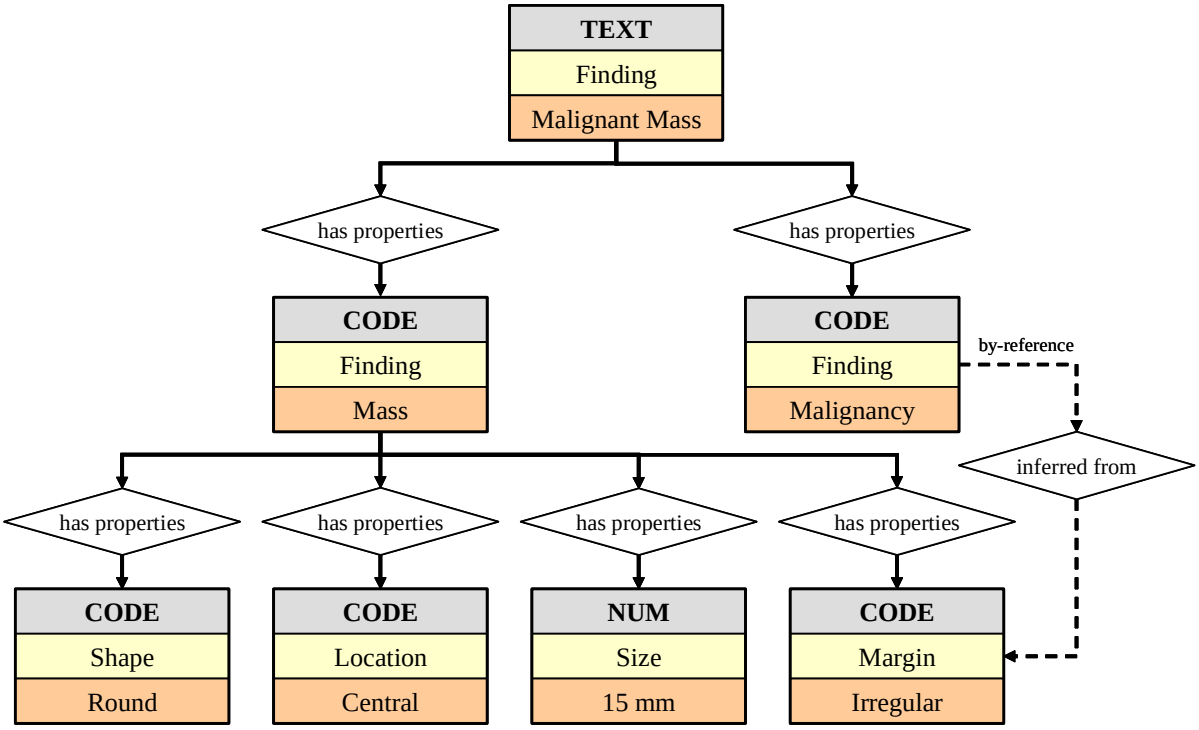


Figure 24 An example of a DICOM SR

Content codes, e.g., Malignant Mass, Mass, Malignancy, Round, Central, 15mm, Irregular in Figure 24, describe the "value" of each node. In the case from the Figure 24, the "value" of `Round, Central, 15mm`, and `Irregular` could be encoded as follows:

Round	Central	15mm	Irregular
Code=M-02100 Coding Scheme=SNOMED 3.0 Concept Name=Round shape	Code=F-0178F Coding Scheme=SNOMED- RT Concept Name=Central region of breast	Code=MM Coding Scheme=UCUM Concept Name=mm	Code=G-A402 Coding Scheme=SNOMED 3.0 Concept Name=Irregular

The report as shown in Figure 24 is now fully machine processable. Note, however, that producing a coded report manually is an enormous amount of work for the reporting physician who operates under high workload, and who has no obvious immediate benefit from the coded report. It is clear that coded, machine processable reports find user acceptance *only* if the users perceive a benefit from the change in their workflow and the creation of the coded report is automated as much as possible with templates for each type of documents. Conducive to adoption of DICOM SRs would be development of

applications that can fill in codes automatically wherever possible and smart user interfaces that present codes in decreasing order of likeliness (most likely result first) based on background information. These applications would then be able to automatically create a human readable version of the report from the codes. [6]

One of the goals of DICOM SR standardization-related activities along this path is to use object-oriented representations and web-based interfaces to enable physicians to use off-the-shelf technologies such as browsers to access patient information [79]. Use cases where this type of access would be useful is the retrieval by a radiologist of images stored in a Picture Archiving and Communication System (PACS) and their display for diagnostic interpretation or post-processing, with demographic and study information originally obtained from a hospital information system or radiology information system. At the workstation, the radiologist can then create structured reports that can be mapped from DICOM to open technologies such as XML.

To cover a variety of applications for DICOM SR, the DICOM SR standard defines Service Object Pair (SOP) classes. A SOP class pairs the description of the data (an IOD2) with a service such as network or media storage, and a SOP Instance is an actual occurrence of a data set [78]. Hence, DICOM SR SOP definitions enable users to link textual and other data to particular images or waveforms, as well as to store and coordinate the findings *so that users can see exactly* what is being described in a report [79]. These DICOM SR IODs and corresponding DICOM SR Storage SOP Classes enable the query and retrieval of SR SOP Instances as Instance-level entities, following the DICOM Query/Retrieve model.

There are three general purpose SR Service Object Pair (SOP) classes:

- The basic text SR IOD, which is used to represent a minimal amount of coded entries, where references are allowed only at the leaves level of the SR textual tree with no by-reference content items, i.e., the by-value is the only allowed relationship.
- The enhanced SR IOD, which is a superset of basic text SR IODs, and is used to represent simple reports that include spatial and temporal regions of interest.
- The comprehensive SR IODs, which is a superset of previous two categories of SR IODs and is used to represent complex reporting without any restrictions on references.

As previously mentioned, in addition to report creating and encoding, DICOM SR also defines how to manage documents, i.e., store and retrieve them. Specifically, the rules define where the documents come from, their destination, version, and the procedure how to get relevant documents. To simplify SR documents, the SR meta-data for document management are stored outside of the SR tree, in a separate module called SR Document General Module.

Different patterns of SR applications are referred to as SR Templates, which are described by a DICOM Content Mapping Resource (DCMR). An SR template is simply a pattern of SR content that suggests or constrains concept names, relationship types, value types, and value sets for a particular application. A template can be as simple as a small pattern to describe the characteristics of a lesion, or as complex as an entire SR document tree for a specific reporting application. In some cases, a template that applies

[6] In general, this is an active area of DICOM SR standardization.

to an entire SR document may form the basis for the definition of an application specific SOP class. The template (see Figure 25) describe and constrain the content items, value types, relationship types, and value sets that might be used in either part of the SR document trees or the overall tree for a specific reporting application. A number of DICOM working groups have been established, which provide templates for specific applications in separate DICOM supplements. For example, DICOM Working Group 15 provides the CAD templates for mammography in supplement 50, the chest CAD templates in supplement 65, etc. DICOM Working Group 12 provides the catheterization laboratory SRs in supplement 66, the ventriculography SRs in supplement 76, etc. DICOM Working Group 8 works on providing XML schema flexible enough for representing all data contained in DICOM SR.

TID XX00
BREAST IMAGING REPORT
Type: Non-Extensible

	NL	Rel with Parent	VT	Concept Name	VM	Req Type	Condition	Value Set Constraint
1			CONTAINER	EV (yyy000, 99SUP79, "Breast Imaging Report")	1	M		
2	>	HAS CONCEPT MOD	INCLUDE	DTID (1204) Language of Content Item and Descendants	1	M		
3	>	HAS OBS CONTEXT	CODE	EV (yyy003, 99SUP79, "Baseline screening mammogram")	1	U		DCID (230) Yes-No
4	>	HAS OBS CONTEXT	CODE	EV (yyy004, 99SUP79, "First mammogram ever")	1	U		DCID (230) Yes-No
5	>	HAS OBS CONTEXT	INCLUDE	DTID (XX01) Breast Imaging Report Procedure Context	1-n	M		
6	>	CONTAINS	INCLUDE	DTID (XX02) Breast Imaging Report Section	1-n	M		

Row number
Nesting level
Relationship with parent
Value type
Value multiplicity
Requirement type

Figure 25 A template for the breast imaging report as given in DCMR

DICOM SR is a powerful concept that essentially covers everything needed for an EHR (when combined with other DICOM services). A complete set of services for document management, including two-level information model is based on SR Templates. However, DICOM SR assumes complex binary encoding and a communication protocol not well-known outside the imaging domain (no XML, no HTTP). The specification is not developed with cross-enterprise applications in mind (no master patient index concept, complex network set-up). For these reasons, it has gained low acceptance outside radiology and cardiology.

3.2.7. Integrated Electronic Health Record (I-EHR)

A fundamental requirement for achieving continuity of care is the seamless sharing of multimedia clinical information. Different technological approaches can be adopted for enabling the communication and sharing of health record segments. In the context of the emerging global information society, the creation of and access to the Integrated Electronic Health Record (I-EHR) of a citizen has been assigned high priority in many

countries. This requirement is complementary to an overall requirement for the creation of a Health Information Infrastructure (HII) to support the provision of a variety of health telematics and e-health services.

Any successful I-EHR realization requires, from a technological point of view, the existence of certain supporting features. Those features impose specific requirements that ought to be met, in order to achieve user acceptance and meet the foreseen benefits. Certain technological requirements for the I-EHR service, imposed by end-user needs and expectations are listed below:

- Round the clock availability;
- Provision of fast responses even at high workload periods (therefore workload balancing and redirection have to be considered);
- Restricted access to information;
- Easy maintenance (remotely in some cases – automatic notification in place);
- Low usage cost;
- Role-based access to information;
- Secure communication of information;
- Activity monitoring;
- Access to reliable, and up-to-date information;
- Native user interface;
- Support direct access to multimedia clinical data communication;
- Scalable (new IT systems should be easily incorporated in the federation);
- Support for standardized coding (semantic unification is a real need);
- Customisable user interface (both adaptive and adaptable to the expertise level of the end user – allow for the isolation and identification of clinical significant information); and
- Highly available (i.e. across various networks and platforms).

The data populating the I-EHR reside in a variety of highly heterogeneous, autonomous and decentralized information systems. The realization of an I-EHR depends on the ability to provide integrated access to the different segments of one's EHR. Physical integration is not necessarily required.

There are a number of envisaged benefits from the development and deployment of an I-EHR service, provided that the need for citizen consent, user authentication, and the required levels of security is properly addressed. Envisaged benefits include the following:

- Vital health information would be available and accessible 24 hours a day, seven days a week, regardless of where the person requiring care happens to be.
- Healthcare practitioners would be better positioned to offer more effective and efficient treatment, and could spend more quality time with the patient as they would be able to view a patient's relevant medical history. Contrast this with the current situation, where medical practitioners have access, if at all, to a partial or inaccurate patient history and may recommend a course of treatment that could potentially be life-threatening.
- The number of redundant procedures would be reduced as physicians would have access to previous medical or lab examinations. Certain procedures may also pose a health risk to patients, if repeated unnecessarily, and ought to be avoided.
- Improved quality of care could be achieved based on an enhanced ability of health planners and administrators to develop relevant healthcare policies for the future. Population health statistics, developed from the information contained in the I-EHR, can be instrumental in the formulation of such policies.

- An I-EHR would greatly empower individuals by giving them access to their own personal health records. It will enable them to make informed choices about options available to them and give them the opportunity to exercise greater control over their own health.

From the viewpoint of standardization, the single most important characteristic of the EHR is the ability to share EHR information between different authorized users.

In technical terms, this requires interoperability of information in the EHR and interoperability of EHR systems which exchange and share this information. There are two main levels of shareability or interoperability of information:

- functional interoperability – the ability of two or more systems to exchange information (so that it is human readable by the receiver), and
- semantic interoperability – the ability for information shared by systems to be understood at the level of formally defined domain concepts (so that information is computer processable by the receiving system).

3.3. Discussion

Although HL7 is on a path to become a de-facto standard in the US and world-wide, a number of concerns have been raised with respect to the standard (including its version 3), and in particular the instability of the RIM [72-74]. A recent study [73] has shown a number of ambiguities in RIM and a variety of logical and ontological flaws, which pose severe obstacles for those who are developing concrete implementations of the standard.

Despite growing criticism, HL7 has successfully promoted its version 3 world-wide. In England (Connecting for Health [75]) and Canada a nation wide-effort on HL7 support has is being carried out. Also, in the Netherlands (NICTIZ [76]) a national health care infrastructure has been started up on the basis of the HL7v3 RIM and the HL7-specific method for creating messages [74]. However, experiences from the use of HL7v3 implementations are not very positive. On one hand HL7 message specifications are considered too rigid to be implemented in a flexible form and, on the other hand, a uniform implementation is labeled too laborious and costly [74]. The same source points to a growing concern that the planned UK healthcare infrastructure would not be able to cope with the enormous volume of the HL7v3 XML messages and would eventually get clogged. In this context Australia, for example, has outlined a strategy for healthcare standards where a transition to a European standard ENV 13606 and *Open*EHR solutions have been advised [77].

As mentioned previously, HL7 V3 using XML to encode CDAs provides well-structured hierarchies to patient records and facilitates the integration of image and non-image medical information into the broader health care context. Many HL7 CDAs usages are by primary care physicians, for which the report is the principal deliverable. However, other care giver stakeholders, such as referring physicians, are frequently interested in seeing the images mentioned in the report. Measurement results are also of interest for trend analysis or other processing. Therefore there is an interest in ensuring that DICOM SRs can be transformed in HL7 format and represented within an EHR. A DICOM Workgroup 20 is therefore established to contribute additional classes and attributes needed for representing DICOM information model in HL7 RIM, and develop methods for supporting imaging integration within HL7 V3.

Taking into account all the presented facts about EHR scope and adoption from the start of this section and the variety of standards and their implementations, we conclude that

the task of extending the information available in EMR/EHR systems with genomic data should be approached in a generic way, as it should apply to a variety of EHR systems. The focus should be on what type of data needs to be integrated in an EHR of the future and on how that data should be modeled taking into account that this should apply to a variety of existing systems, as the EMR/EHR landscape is and may remain heterogeneous in the predictable future, with respect to both products (implemented features and customization) and vendors. At this point no dominant single solution can be selected that is expected to take over.

To exemplify our solution to the inclusion of genomic data in a future EHR we may choose *Open*EHR (discussed in section 3.2.5), as it is a promising standard in the area.

4. Envisioned solutions and services

To the best of our knowledge there is only one initiative on standardization of clinico-genomic information that can be used in the context of an EHR. Namely, the HL7 Clinical Genomics Special Interest Group (CGL7) is currently defining standards that would enable exchange of clinical and personalized data between interested parties [80]. As CGL7 is an initiative is driven by IBM within the IBM Healthcare & Life Sciences Information Integration activities, CGL7 is, naturally, envisioned to be used in conjunction with IBM Clinical Genomics solutions or as a stand-alone service hosted by IBM or integrated within the enterprise information systems.

According to its creators, CGL7 is designed to serve as a missing layer of genotype-phenotype associations. In that context CGL7 is viewed as a middleware that provides a dedicated Application Programming Interface (API) for rapid development of decision support applications and stand-alone web service solutions.

The key characteristics of the CGL7 solutions under development are:

- The association of patient genetic data with clinical data using HL7 messages by defining *a genotype model*.
- The use of already existing and established *data representations for genomic information* and biological sequences, e.g., MAGE-ML [81] and BSML. MAGE-ML stands for MicroArray and Gene Expression Markup Language and it is an XML-based language explicitly designed to describe and communicate information about microarray based experiments, including microarray designs, microarray manufacturing information, microarray experiment setup and execution information, gene expression data and data analysis results [81]. BSML, a Bioinformatics Sequence Markup Language, is also an XML-based language primarily focused on describing DNA, RNA, and protein sequences [82].
- Application of the so-called *encapsulate and bubble–up* conceptual workflow.

In the next sections we give a brief overview of CGL7 architecture and outline the named three core characteristics of the CGL7 standard.

CGL7 Architecture

Figure 26 shows the CGL7 architecture and illustrates how the standard is going to be employed. The raw genetic data is produced by the genetic testing facilities and encapsulated into a message conforming to the CGL7 standard.

The *encapsulation phase* therefore refers to obtaining raw genetic data from and encapsulating them into patient records, based on pre-defined, constrained bioinformatics format. The constrained bioinformatics format is necessary to ensure that only elements of genetic information relevant for clinical practice are present. This phase of encapsulation and storing data in EHR is static and encapsulation is performed based on a static pre-defined BSML schema.

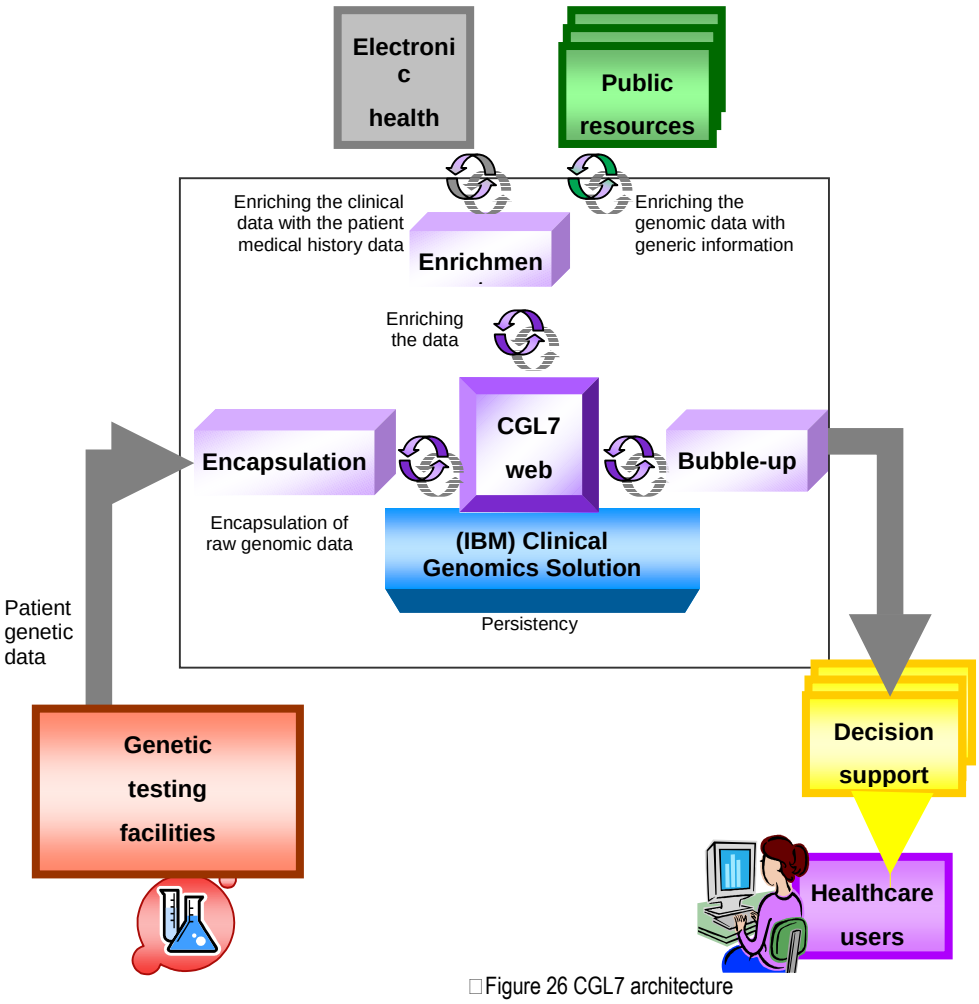


Figure 26 CGL7 architecture

The bubble-up phase is a phase in which in an iterative manner the raw genetic material is parsed and those portions of the material that seem to be most interesting from the clinical perspective are made prominent [80]. Hence, in the process dynamic bubble interactions between the clinical data stored in an EHR and data stored in other publically available public sources of genomics information. During the bubbling up phase the network of interactions of clinical data and genomics data is established such that it can be utilized by a decision support system, which would then recommend the (personalized) treatment at a point of care. The idea behind this workflow is to provide a process in which gradual distillation of raw genomic can be done in the context of diagnosis and treatment provided to a specific patient at a specific time, and at the same time making the raw data available in the EHR for the purposes of parsing in again in another context, when, for example, new knowledge becomes available (see Figure 27).

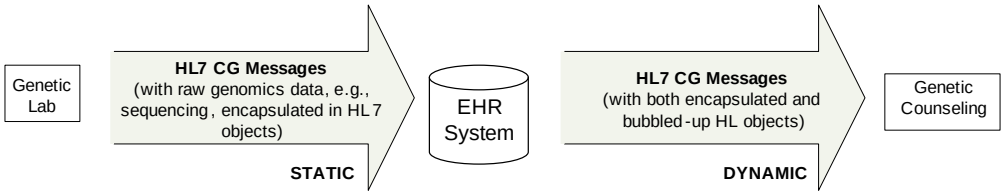


Figure 27 Encapsulate and bubble-up workflow

Genotype Models in CGL7

The core model of the CGL7 is the `GeneticLocus` model, which provides placeholders for various types of genomics data relating to a specific locus on the genome, e.g., DNA or RNA, including sequencing and proteomics, the locus and its alleles, the expression data sequence variations, and clinical phenotypes. Figure 29 gives a birds-eye view of the focal areas of the model.

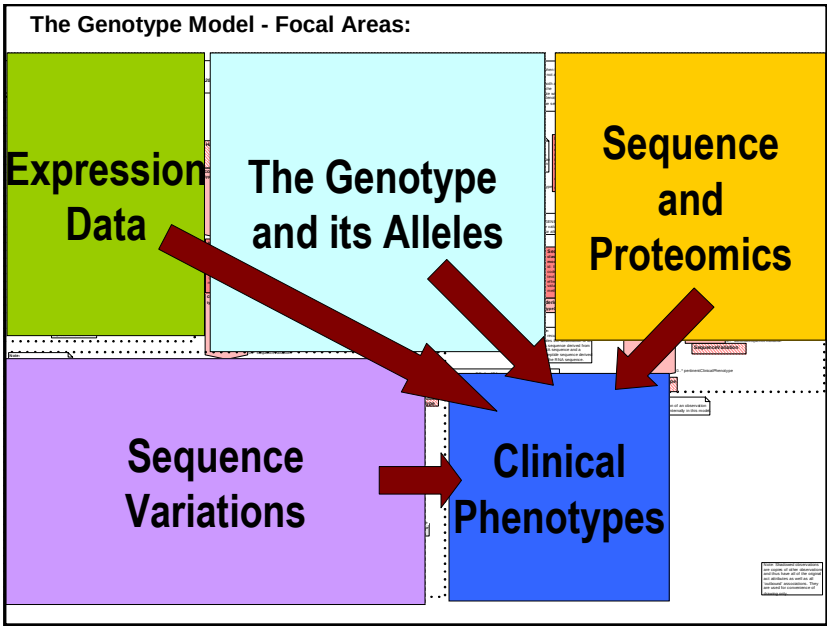


Figure 29 The key focal areas of the CGL7

The conventions used in the model are based on HL7 RIM. A portion of the `GeneticLocus` model that describes expression data is shown in Figure 30. As seen from the portion of the model shown in the figure, the `Expression` class is associated with the individual allele class, which in turn is associated with the entry point, the GeneticLocus class. Similarly, Figure 30 shows the portion of the `GeneticLocus` model related to the `Sequence` class, where it is visible that the `Sequence` class is also associated with the `IndividualAllele` class, which is in turn associated with the `GeneticLocus` class (see also Figure 30). The value attribute of the `Sequence` class can hold raw data in bioinformatics markup format, such as BSML. The `DerivedFrom` association is recursive and enables representation of a biological sequence derived from a source sequence, e.g., an mRNA sequence derived from a DNA sequence. `pertinentInformation` is the association with `ClinicalPhenotype` which enables genomic data to be linked with clinical data that most likely resides in a EHR. `ClinicalPhenotype` stands for phenotype model whose classes are described elsewhere in the `GeneticLocus` model.

Another model that is a part of the HL7 CG specification is the `FamilyHistory` model (see Figure 31), which is developed to describe a patient's genetic family history with clinical and genomics data. Breast and ovarian cancer are among the use cases where there is a clear need to represent the family history.

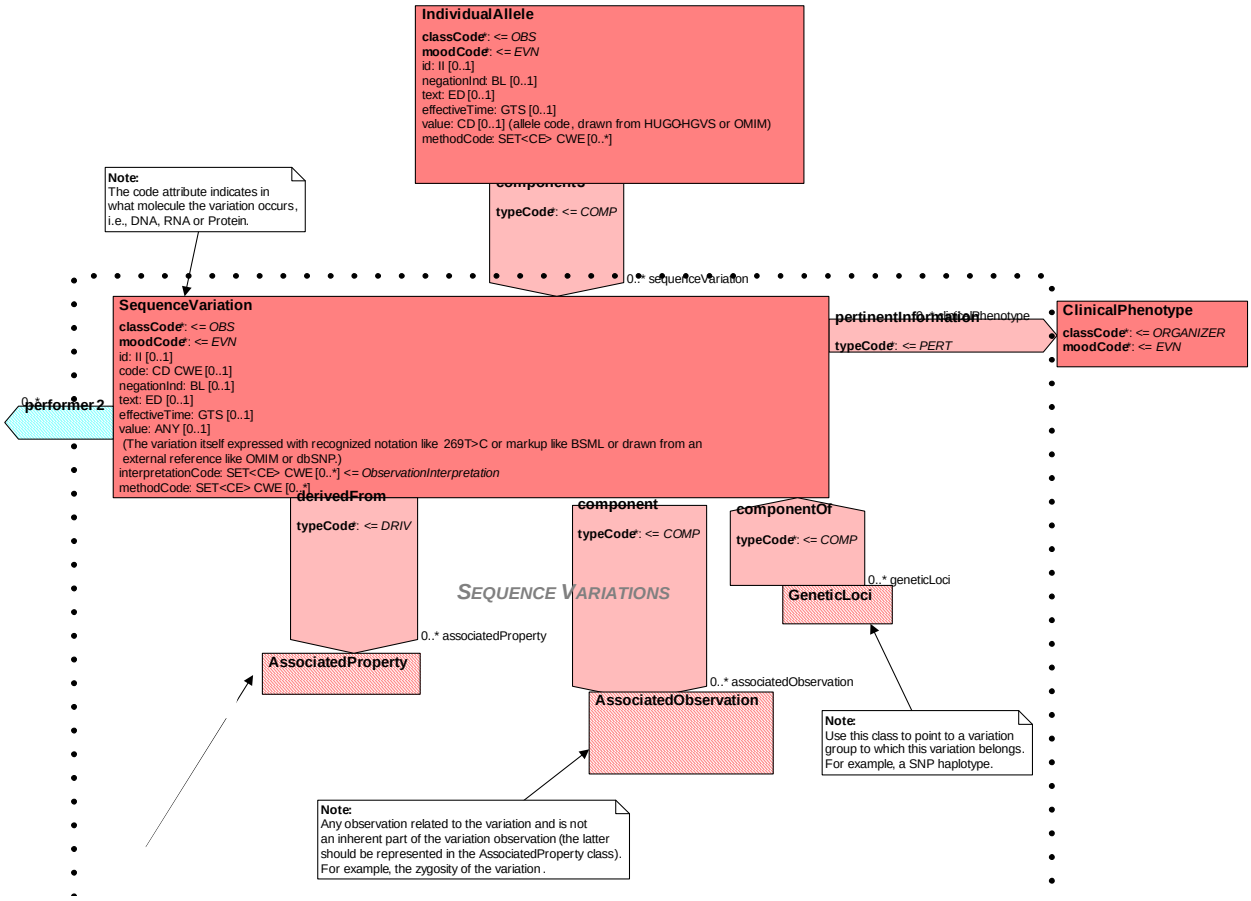


Figure 30 The portion of the GeneticLocus model related to the Sequence class

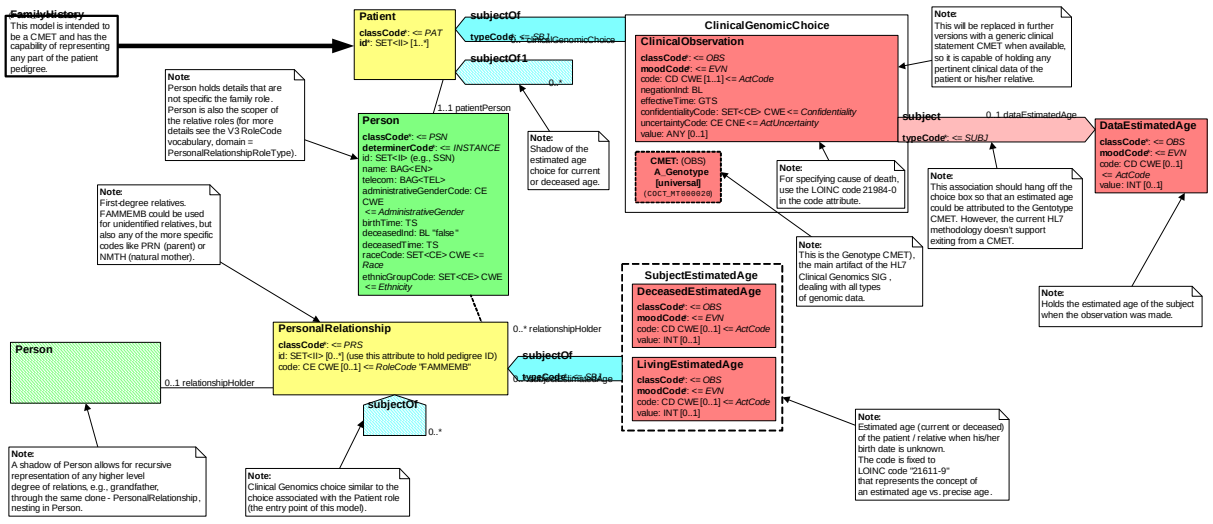


Figure 31 The family history (pedigree) model

5. Conclusions and future work

In this document we have carried out a state-of-the-art review of the most prominent EHR initiatives and standards and we have briefly discussed the impact of those initiatives on the state-of-practice with respect to EMR/EHR products currently deployed by healthcare organization. Next, in the context of oncology, we have assessed the role that genomic information should play in the clinic, and therefore also in an EHR of the future, based on current results in the research literature.

As explained in chapter 3, a universally accepted definition of the EHR does not exist at this time. Definitions currently in use range in level of detail and scope, and many of them have originally referred to and evolved from various concepts preceding the EHR including the EHCR (Electronic Health Care Record), EPR (Electronic Patient Record), CPR (Computerized Patient Record), and EMR (Electronic Medical Record). To emphasize this heterogeneity, several definitions proposed by different organisations are listed below. They are extracted from the draft [83] technical Specification "Requirements for an Electronic Health Record Reference Architecture".

- An electronic longitudinal collection of personal health information usually based on the individual, entered or accepted by health care providers, which can be distributed over a number of sites or aggregated at a particular source. The information is organised primarily to support continuing, efficient and quality health care. The record is under the control of the consumer and is stored and secured securely.
- A longitudinal collection of personal health information of a single individual, entered or accepted by health care providers, and stored electronically. The record may be made available at any time to providers, who have been authorized by the individual, as a tool in the provision of health care services. The individual has access to the record and can request changes to its contents. The transmission and storage of the record is under strict security.
- A collection of data and information gathered or generated to record clinical care rendered to an individual
- A comprehensive, structured set of clinical, demographic, environmental, social, and financial data and information in electronic form, documenting the health care given to a single individual.
- A healthcare record in computer readable format.
- An electronic patient record that resides in a system designed to support users through availability of complete and accurate data, practitioner reminders and alerts, clinical decision support systems, links to bodies of medical knowledge, and other aids.
- A virtual compilation of non-redundant health data about a person across a lifetime, including facts, observations, interpretations, plans, actions, and outcomes. Health data include information on allergies, history of illness and injury, functional status, diagnostic studies, assessments, orders, consultation reports, treatment records, etc. Health data also include wellness data such as immunization history, behavioural data, environmental information, demographics, administrative data for care delivery processes, health insurance, and legal data such as consent forms.

To summarize, EHRs may include (or access), without being limited to, information regarding subject of care identification, demographics, health history, clinical summaries, problem lists and diagnoses, diagnostic values and interpretation, care plans and decision support, treatments, consent, vital signs and alerts, provider identification, clinical documentation for chronic diseases, encounters, immunizations, primary care and community care, quality and safety. Next to locally residing information EHRs also contain

links to external sources of information that are used in the context of providing care (e.g. accession numbers to images residing on PACS). As the trial and error paradigm in healthcare is increasingly replaced by an evidence-based approach, also the amount of relevant information that needs to be accessible through the EHR increases massively.

In the context of moving from symptom-driven generic disease-centric therapy to early-detection (or prevention) targeted patient-specific therapy, genomic information is likely to play an important role, in combination with other, more traditional types of information such as clinical data and imaging. There is significant evidence in the research literature to support the vision that genetic and epigenetic markers will contribute to detection, diagnosis, risk stratification and therapy selection and planning in cancer, and will constitute important fragments of information to be stored in or accessed from a clinico-genomic EHR.

Cancer is a disease characterized by genome dysfunction. However several genetic or epigenetic changes need to happen to allow for cancer to occur. For epigenetic modifications, it is assumed that a genome-wide imbalance in methylation levels may indicate a state of disease (while in normal tissue the methylation is localized mainly in the CpG-poor, coding areas, in tumour tissue the CpG-rich promoter areas are methylated), and that the methylation state of some genes can be used as biomarker for tumorigenesis. Therefore, it is likely that in order to characterize a disease in a specific patient, large amounts of data will need to be analyzed and the future genomic tests will not target single genes but groups of genes, or genome-wide modifications. New algorithms and approaches will be required to vertically integrate the various sources of data and to link the different types of genomic data to each other and to external (public) sources of genomic data and annotation. Genomic data, including the results of simple genomic markers detection (e.g., HER-2 test), could also be stored in the EHR as lab results. Other types of data, characterized by large volume and complex analyses, such as microarrays and large genomic sequences, could be stored externally in separate systems, in a similar way to storing imaging data on PACS while only maintaining the report in the RIS and in the EMR.

All these aspects pose strong scalability and flexibility requirements on a future EHR/EMR system, as large amounts of data will need to be stored, managed and integrated, external sources of reference data (e.g. public databases) will need to be accessible, and the system should be easily extensible with new types of data to enable its users to leverage on the new developments from clinical research, molecular biology and bioinformatics. Another important feature of a future EHR/EMR currently not provided by most systems deployed in practice is to enable research on the data, facilitating complex queries and data mining.

This report presented the results of an initial exploration of the genomic information to be included in a future clinico-genomic EHR/EMR, based on literature. We could conclude that genomic information has the potential to play an important role in the various stages of the cancer care cycle, and that the data to be stored, analyzed and managed is both heterogeneous and in large volume. Reasoning about this type of information in the context of cancer detection and therapy is very complex, and to support adoption in the clinical practice, tools able to extract the relevant information and to present the right view on the data to each healthcare professional involved in the treatment decision, planning, and follow-up need to be developed.

To elaborate a detailed specification, and subsequently build a data model for the genomic information to be accessible from an EHR, we will first select a domain, breast cancer, and carry out interviews with clinicians, molecular biologists and bioinformaticians in ACGT involved in breast cancer clinical trials to select the pieces of genomic information that are relevant for their research and need to be managed. We will also consider how results from the current clinical research may be translated and applied into clinical practice. Based on this ACGT-specific information we will propose a data model

for genomic data specific to breast cancer, considering both the research and the clinic. Further, this data model can be extended to include other cancer types.

This additional data will not necessarily become a direct part of an EMR/EHR, but it should be accessible through the hospital EMR (or EHR when in place), even when stored and managed by a separate system. As the EHR/EMR landscape is currently very heterogeneous and its future evolution is hard to predict, we will not focus on a specific technology or EHR/EMR product, but aim to build a data model that is generic enough not to depend on the chosen EHR/EMR solutions.

Appendix 1 - Abbreviations and acronyms

ADL	Archetype Definition Language
ADR	Adverse Drug Reactions
API	Application Programming Interface
ANSI	American National Standards Institute
BEA	BioLab Experiment Assistant
BSML	Bioinformatics Sequence Markup Language
CAD	Computer-Aided Diagnostics
CDA	Clinical Document Architecture
CPR	Computerized Patient Record
DCMR	DICOM Content Mapping Resource
DICOM	Digital Imaging and Communications in Medicine
DICOM SR	DICOM Structured Reporting
DME	Drug Metabolizing Enzymes
EHR	Electronic Health Record
EMR	Electronic Medical Record
ER	Estrogen Receptor
FDA	Food and Drug Administration
HER2	Human Epidermal growth factor Receptor 2
HL7	Health Level Seven
HL7 RIM	Health Level Seven Reference Information Model
MAGE-ML	MicroArray and Gene Expression Markup Language
PHR	Personal Health Record
PR	Progesterone Receptor
RIM	Reference Information Model
SOP	Service Object Pair
XML	eXtensible Markup Language

Appendix 2 – Terminology

EHR	An EHR is an electronic system that captures and manages episodic and longitudinal health information by encompassing relevant medical data for a patient. Medical data include clinical data, clinical decision support, master patient index, workflow support, controlled medical vocabulary, order entry, computerized provider order entry, pharmacy, and clinical documentation applications [54].
EHR node	The physical location where EHRs are stored and maintained is often referred to as an EHR node [55].
EHR system	A set of components that collectively form the mechanisms by which electronic health records are created, used, stored, and retrieved are normally denoted an EHR system [55]. The EHR system, besides health information includes also people, data, rules and procedures, processing and storage devices, and communication and support facilities.
EHR interoperability	Interoperability and information sharing is required on two levels as follows [55]. *Functional interoperability* is the capability of two or more (EHR) systems to exchange information (in a human readable form). *Semantic interoperability* is the property of the information shared by an EHR system to be understood at the level of formally defined domain concepts, in order to ensure that that it can be processed by the receiving information system. Semantic interoperability is essential for facilitating the real value-added EHR clinical applications such as intelligent decision support and care planning.
EMR	EMR is often defined as an electronic system that represents a care givers clinical data repository, controlled medical vocabulary, and clinical documentation applications. The data in the EMR is the legal record of what happened to the patient during their encounter at the care giver and is owned by the care giver organization [84]. The data from an EMR can be a source for the data contained in an EHR [54].
PHR	A personal health record (PHR) is an electronic application through which individuals can access, manage and share their health information in a private, secure, and confidential environment [85]. It allows people to access and coordinate their lifelong health information and make appropriate parts of it available to those who need it, e.g., caregivers. PHRs normally contain optional tools specifically intended for individuals (consumers) who do not have legal and professional obligations for health record-keeping [86]. Hence, the health information contained in a PHR is specifically intended to be managed by an individual with him/her determining access rights to the PHR and being responsible for its content. The information contained in a PHR comes both from health providers and the individual. As such PHRs can, for example, be used as vehicles for transparency about treatment options and transactions, ranging from the evidence base for various treatments to the costs of medical services [86]. Note, however, that a PHR does not replace the legal record of any provider [85].
Clinical terminology	C*linical terminology* represents a structured collection of descriptive terms, i.e., a vocabulary of medical terms, specifically intended for use in the clinical practice [87]. In this context, a clinical (terminology) vocabulary contains terms that describe the care and treatment of patients and cover clinically relevant areas, such as diagnoses, symptoms, surgical procedures, treatments, drugs, etc. The aim of defining such a structured vocabulary is to use it in computer applications and thereby (i) enable clinical staff to record patient information in a consistent manner, and (ii) provide means to communicate recorded clinical information in a standard way between healthcare systems and individuals. A number of clinical terminologies exist, including: Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) [87]; International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10) is vocabulary of diagnostic codes developed by the Canadian Institute for Health Information (CIHI) [88]; and MEDCIN® [89] is a commercial product that contains a nomenclature for coding the clinical encounter.
Genomics	Studies a makeup of individuals DNA, which is set at conception and remains the same.
Proteomics	Focuses on constellation of proteins within the cell, which can be altered by health environmental and physical changes.
Pharmacogenomics	Applies the findings of genomics and proteomics toward determining how individual genomic differences will interact or respond to a particular drug.
Pharamacogenetics	Same as pharmacogenomics, this implies the study of genetic variation that gives rise to differing response to drugs, focusing only on few candidate genes, as opposed to the entire genome.
Pharmacology	The science of drugs, including their composition, uses, and Effects.

Appendix 3 – Raw results of the BEA tool analysis

- **Breast Cancer**

Gene name	Citations
ESTROGEN RECEPTOR	2781
ERBB-2 RECEPTOR	2531
BRCA1	1774
P53	1541
AROMATASE	1289
BRCA2	1274
PROGESTERONE RECEPTOR	1235
NODAL	1132
ESTROGEN RECEPTOR ALPHA	1093
EPIDERMAL GROWTH FACTOR RECEPTOR	1080
TUMOR SUPPRESSOR PROTEIN P53	1022
HER2	916
HER-2/NEU	692
BCL-2	642
CYCLIN D1	597
EGFR	591
HER-2	575
EPIDERMAL GROWTH FACTOR	559
PROTO-ONCOGENE PROTEINS C-BCL-2	527
VEGF	509
OESTROGEN RECEPTOR	507
AKT	503
SLN	496
VASCULAR ENDOTHELIAL GROWTH FACTOR A	475
E2	469
P-GLYCOPROTEIN	468
KI-67 ANTIGEN	462
VASCULAR ENDOTHELIAL GROWTH FACTOR	462
ERBB2	411
P21	403
KI-67	397
PROTO-ONCOGENE PROTEINS C-AKT	395
1-PHOSPHATIDYLINOSITOL 3-KINASE	382
COLLAGEN	381
BREAST CANCER RESISTANCE PROTEIN	366
E-CADHERIN	365
KERATINS	364
INSULIN-LIKE GROWTH FACTOR I	362
ESTROGEN RECEPTOR BETA	357
EGF	344
TUMOR NECROSIS FACTOR-ALPHA	333
SD	315
ABCG2	313
CA-15-3 ANTIGEN	301
LUCIFERASE	298
APRIL	296
BCRP	288
CYTOKERATIN	283
TRANSFORMING GROWTH FACTOR BETA	282

ERBETA	279
BAX	274
ACTIN	272
C-MYC	272
NF-KAPPA B	267
COX-2	259
MAPK	255
CYCLIN-DEPENDENT KINASE INHIBITOR P21	254
IGF-I	248
GRANULOCYTE COLONY-STIMULATING FACTOR	247
P27	238
CDKN1A	232
CYCLOOXYGENASE 2	232
CASPASE 3	231
PHANTOM	227
CI	223
INSULIN	222
ERK	220
CASPASE-3	217
AKT1	212
NF-KAPPAB	208
GLUTATHIONE TRANSFERASE	203
TP53	201
RT	198
CARCINOEMBRYONIC ANTIGEN	196
CEA	193
PS2	190
G-CSF	189
MUC1	189
CYCLIN E	189
CASP3	186
ANDROGEN RECEPTOR	184
UPA	182
PI3K	182
SN	180
BETA-CATENIN	177
URINARY PLASMINOGEN ACTIVATOR	176
PTEN	176
CYCLOOXYGENASE-2	176
LAMININ	175
BCL-2-ASSOCIATED X	174
PROTEIN KINASE C	171
MITOGEN-ACTIVATED PROTEIN KINASE 3	171
ER-ALPHA	170
CYTOCHROME C	170
TGF-BETA	170
ERK1/2	169
MMP-9	168
MITOGEN-ACTIVATED PROTEIN KINASE 1	168
PROLACTIN	167

Biological Process	Citations
APOPTOSIS	3065
TRANSCRIPTION	2142
INDUCTION	2043
CELL PROLIFERATION	1842
CELL CYCLE	1565
CELL GROWTH	1312
ANGIOGENESIS	1062
LOCALIZATION	1049
CELL DEATH	850
MENOPAUSE	690
PATHOGENESIS	551
METHYLATION	480
TRANSPORT	436
DNA REPAIR	424
SECRETION	420
CELL CYCLE ARREST	362
DRUG RESISTANCE	343
CELL ADHESION	341
INDUCTION OF APOPTOSIS	335
CELL MIGRATION	326
IMMUNE RESPONSE	264
MENARCHE	247
TRANSLATION	240
BONE RESORPTION	207
S PHASE	198
CELL MOTILITY	190
RNA INTERFERENCE	184
LACTATION	183
MITOSIS	172
DNA METHYLATION	171
ONCOGENESIS	161
CIRCULATION	160
HOMEOSTASIS	157
GLAND DEVELOPMENT	154
MAMMARY GLAND DEVELOPMENT	152
MENSTRUAL CYCLE	148
AGING	148
HYPERSENSITIVITY	145
EXCRETION	138
PROTEOLYSIS	118
SEGMENTATION	116
CELL DIFFERENTIATION	115
CELL DIVISION	114
MORPHOGENESIS	107
SENSITIZATION	104
ESTROGEN METABOLISM	87
GENE SILENCING	86
G1 PHASE	84
CELL KILLING	82
CASPASE ACTIVATION	81

- **Lung Cancer**

Gene name	Citations
EPIDERMAL GROWTH FACTOR RECEPTOR	1421
P53	959
EGFR	933
TUMOR SUPPRESSOR PROTEIN P53	608
NODAL	536
VASCULAR ENDOTHELIAL GROWTH FACTOR A	423
VEGF	420
VASCULAR ENDOTHELIAL GROWTH FACTOR	419
CT	390
BCL-2	323
CARCINOEMBRYONIC ANTIGEN	299
RT	299
SD	284
APRIL	283
EPIDERMAL GROWTH FACTOR	281
PROTO-ONCOGENE PROTEINS C-BCL-2	278
CEA	260
COX-2	253
TUMOR NECROSIS FACTOR-ALPHA	244
ERBB-2 RECEPTOR	239
CYCLOOXYGENASE 2	230
KERATINS	223
AKT	222
COLLAGEN	213
K-RAS	211
PROTO-ONCOGENE PROTEINS C-AKT	203
P16	200
P21	198
CYCLOOXYGENASE-2	195
FDG	194
CASPASE-3	189
GLUTATHIONE TRANSFERASE	187
KI-67	180
INTERFERON TYPE II	173
TELOMERASE	172
NF-KAPPA B	167
IFN-GAMMA	164
KI-67 ANTIGEN	156
PTGS2	154
INTERLEUKIN-2	153
CASPASE 3	153
BAX	152
CI	151
TNF-ALPHA	150
CYCLIN D1	149
IL-2	143
P-GLYCOPROTEIN	142
MMP-9	142
CYCLIN-DEPENDENT KINASE INHIBITOR P21	142
GRANULOCYTE COLONY-STIMULATING	140

FACTOR	
COPD	139
CYCLIN-DEPENDENT KINASE INHIBITOR P16	137
MATRIX METALLOPROTEINASE 9	136
1-PHOSPHATIDYLINOSITOL 3-KINASE	135
EGF	134
CDKN1A	133
MMP-2	130
CD4	129
NF-KAPPAB	129
PHANTOM	128
TTF-1	127
CD8	127
TRANSFORMING GROWTH FACTOR BETA	126
CYTOKERATIN	124
VIMENTIN	123
MATRIX METALLOPROTEINASE 2	122
E-CADHERIN	121
SOMATOSTATIN	121
HEMOGLOBIN	121
THYROID NUCLEAR FACTOR 1	120
ERK	118
TTP	115
N2	115
GSTM1	114
CYTOCHROME C	113
SCC	112
CASP3	112
PROLIFERATING CELL NUCLEAR ANTIGEN	111
PROTEIN KINASE C	111
TP53	108
IL-6	107
IL-12	105
ACTIN	104
BETA-CATENIN	100
LUCIFERASE	99
DLT	99
AKT1	99
INTERLEUKIN-6	98
LAMININ	97
INTERLEUKIN-12	96
SMR	96
CYP1A1	96
BCL-2-ASSOCIATED X	95
LACTATE DEHYDROGENASE	94
INTERFERON	94
RETINOBLASTOMA	94
HISTONE DEACETYLASE	94
G-CSF	94

P27 93

Biological Process	Citations
APOPTOSIS	1825
INDUCTION	1610
ANGIOGENESIS	861
TRANSCRIPTION	840
CELL PROLIFERATION	746
CELL CYCLE	736
CELL GROWTH	599
PATHOGENESIS	529
METHYLATION	467
LOCALIZATION	457
CELL DEATH	440
DNA REPAIR	289
SECRETION	273
INDUCTION OF APOPTOSIS	242
CELL ADHESION	210
IMMUNE RESPONSE	204
CELL CYCLE ARREST	187
DRUG RESISTANCE	184
TRANSPORT	151
DNA METHYLATION	144
CIRCULATION	121
TRANSLATION	113
COAGULATION	108
CELL MOTILITY	104
CELL MIGRATION	103
CELL DIFFERENTIATION	93
MITOSIS	83
S PHASE	83
RNA INTERFERENCE	83
HYPERSENSITIVITY	81
ONCOGENESIS	79
INFLAMMATORY RESPONSE	75
CELL KILLING	71
AGING	70
CASPASE ACTIVATION	59
GENE SILENCING	59
G1 PHASE	53
PROGRAMMED CELL DEATH	52
HOMEOSTASIS	52
SENSITIZATION	51
SEGMENTATION	50
CYTOKINE PRODUCTION	44
PROTEOLYSIS	44
GLUCOSE METABOLISM	44
CELL DIVISION	44
EXCRETION	41
BONE RESORPTION	38
LUNG DEVELOPMENT	34
INTERPHASE	33
MISMATCH REPAIR	32

- **Skin Cancer**

Gene name	Citations
P53	408
NODAL	368
TUMOR SUPPRESSOR PROTEIN P53	280
COLLAGEN	247
SCC	245
INTERFERON	238
INTERLEUKIN-2	211
INTERFERON-ALPHA	201
SLN	193
TYROSINASE	176
BCL-2	173
CD4	173
TUMOR NECROSIS FACTOR-ALPHA	162
KERATINS	152
PROTO-ONCOGENE PROTEINS C-BCL-2	141
KI-67 ANTIGEN	128
IL-2	127
SPITZ	127
VIMENTIN	127
CD34	127
CDKN2A	126
INTERFERON ALFA-2B	124
MONOPHENOL MONOOXYGENASE	123
S100	116
CYCLIN-DEPENDENT KINASE INHIBITOR P16	116
CD8	111
CD30	110
KI-67	109
BRAF	109
CTCL	108
INTERFERON TYPE II	101
CD56	99
P16	98
ACTIN	96
MART1 ANTIGEN	95
CYCLOOXYGENASE 2	95
VASCULAR ENDOTHELIAL GROWTH FACTOR	95
CYTOKERATIN	93
PROTO-ONCOGENE PROTEINS B-RAF	93
VEGF	91
VASCULAR ENDOTHELIAL GROWTH FACTOR A	88
COX-2	86
CYCLIN D1	86
HAIRLESS	83
CD3	79
ORNITHINE DECARBOXYLASE	78
TNF-ALPHA	78
MART-1	74
CYCLOOXYGENASE-2	73
EPIDERMAL GROWTH FACTOR RECEPTOR	72
CD68	71
MELAN-A	70

PROLIFERATING CELL NUCLEAR ANTIGEN	70
BETA CATENIN	70
AKT	69
BETA-CATENIN	69
APRIL	68
AP-1	68
NF-KAPPA B	67
CDK4	66
SD	65
MC1R	64
PROTO-ONCOGENE PROTEINS C-AKT	64
CD20	64
PUNCH	64
TRANSCRIPTION FACTOR AP-1	63
IFN-GAMMA	63
CD31	60
RAS	60
DESMIN	59
NF-KAPPAB	58
P21	58
TRANSFORMING GROWTH FACTOR BETA	57
GP100	56
INTERLEUKIN-12	56
PATCHED RECEPTORS	56
MMP-2	56
EPITHELIAL MEMBRANE ANTIGEN	56
IL-10	56
MATRIX METALLOPROTEINASE 2	55
KERATIN	55
PROTEIN KINASE C	54
MELANOCORTIN TYPE 1 RECEPTOR	54
ODC	53
E-CADHERIN	52
TNF	52
PTEN	52
GRANULOCYTE-MACROPHAGE COLONY-STIMULATING FACTOR	51
SN	51
CD10	50
CYCLIN-DEPENDENT KINASE INHIBITOR P21	50
CTNNB1	49
IL-12	48
VITILIGO	48
INTERFERON-GAMMA	48
INTERLEUKIN-10	48
BAX	48
CASPASE-3	48
LAMININ	47

Biological Process	Citations
APOPTOSIS	696
INDUCTION	668
PATHOGENESIS	470

TRANSCRIPTION	383
PIGMENTATION	284
LOCALIZATION	267
CELL PROLIFERATION	264
IMMUNE RESPONSE	263
ANGIOGENESIS	261
CELL CYCLE	240
DNA REPAIR	207
CELL DEATH	146
CELL GROWTH	137
AGING	107
HYPERSENSITIVITY	100
SECRETION	96
WOUND HEALING	83
CELL ADHESION	82
PHOTOPROTECTION	75
INDUCTION OF APOPTOSIS	69
METHYLATION	67
HOMEOSTASIS	62
CIRCULATION	50
CELL MIGRATION	49
MISMATCH REPAIR	46
INFLAMMATORY RESPONSE	44
CELL CYCLE ARREST	44
CELL DIFFERENTIATION	42
ONCOGENESIS	42
TRANSPOSITION	37
TRANSPORT	36
KERATINOCYTE PROLIFERATION	35
MITOSIS	33
CYTOKINE PRODUCTION	30
KERATINIZATION	29
KERATINOCYTE DIFFERENTIATION	26
TRANSLATION	26
S PHASE	24
PROTEOLYSIS	24
SENSITIZATION	23
DNA REPLICATION	23
LYMPHANGIOGENESIS	22
RESPONSE TO UV	22
CELL-CELL ADHESION	22
INTERPHASE	21
CELL MOTILITY	20
MELANOCYTE DIFFERENTIATION	19
EMBRYONIC DEVELOPMENT	19
CELL DIVISION	19
CASPASE ACTIVATION	19

- **Head and neck Cancer**

Gene name	Citations
EPIDERMAL GROWTH FACTOR RECEPTOR	319
P53	256
EGFR	219
TUMOR SUPPRESSOR PROTEIN P53	163
SCC	105
CYCLIN D1	91
VASCULAR ENDOTHELIAL GROWTH FACTOR	84
PHANTOM	83
VEGF	83
P16	82
VASCULAR ENDOTHELIAL GROWTH FACTOR A	80
EPIDERMAL GROWTH FACTOR	73
FDG	58
COX-2	57
BCL-2	56
CYCLOOXYGENASE 2	53
AKT	51
HEMOGLOBIN	50
KERATINS	49
PROTO-ONCOGENE PROTEINS C-BCL-2	49
KI-67 ANTIGEN	48
T2	45
ERYTHROPOIETIN	45
COLLAGEN	45
PTGS2	44
CYCLOOXYGENASE-2	43
TP53	41
EGF	40
CYCLIN-DEPENDENT KINASE INHIBITOR P16	39
KI-67	38
MLC	38
T3	37
T4	37
IL-2	37
VIMENTIN	37
CYTOKERATIN	37
INTERLEUKIN-2	35
IFN-GAMMA	35
P21	34
TRANSFORMING GROWTH FACTOR BETA	34
STAT3	34
CRT	34
TUMOR NECROSIS FACTOR-ALPHA	33
GLUTATHIONE TRANSFERASE	33
SLN	33
ALPHA SUBUNIT HYPOXIA-INDUCIBLE FACTOR 1	33
NF-KAPPA B	32
FLAP	32
IL-6	32
PROTO-ONCOGENE PROTEINS C-AKT	32
CI	30

CASPASE-3	30
INTERFERON	30
CYCLIN-DEPENDENT KINASE INHIBITOR P21	29
SUCCINATE DEHYDROGENASE	29
HIF1A	29
TELOMERASE	28
SDHD	28
STAT3 TRANSCRIPTION FACTOR	28
ERBB-2 RECEPTOR	27
BAX	27
P27	27
CDKN1A	27
CD34	27
P63	26
CD44	26
MMP-9	26
CYCLIN	26
E-CADHERIN	25
INTERLEUKIN-6	25
NF-KAPPAB	25
1-PHOSPHATIDYLINOSITOL 3-KINASE	25
SDHB	25
HIF-1ALPHA	25
G1	24
MATRIX METALLOPROTEINASE 9	24
INTERFERON TYPE II	24
CASPASE 3	23
RETINOBLASTOMA	23
CCND1	23
GM-CSF	23
MAPK	22
INK4A	22
P16(INK4A)	22
ACTIN	22
IL-8	21
CYCLIN-DEPENDENT KINASE INHIBITOR P27	21
PROLIFERATING CELL NUCLEAR ANTIGEN	21
CD31	21
MMP-2	21
DLT	20
CD8	20
CD4	20
INTERFERON-ALPHA	20

Biological Process	Citations
INDUCTION	328
APOPTOSIS	322
CELL CYCLE	197
ANGIOGENESIS	176
LOCALIZATION	149
TRANSCRIPTION	136
CELL PROLIFERATION	134
PATHOGENESIS	124

CELL GROWTH	110
CELL DEATH	75
METHYLATION	75
SECRETION	72
DNA REPAIR	68
CIRCULATION	47
CELL CYCLE ARREST	45
WOUND HEALING	44
IMMUNE RESPONSE	40
INDUCTION OF APOPTOSIS	39
TRANSLATION	39
CELL ADHESION	30
TRANSPORT	23
CELL MIGRATION	21
S PHASE	20
HYPERSENSITIVITY	19
SENSITIZATION	18
SEGMENTATION	17
DRUG RESISTANCE	16
DNA METHYLATION	16
EXCRETION	15
DEHISCENCE	15
RESPONSE TO RADIATION	14
CELL DIFFERENTIATION	13
REGENERATION	12
ONCOGENESIS	12
COAGULATION	11
METAPHASE	11
MITOSIS	11
CHEMOTAXIS	10
CELL KILLING	10
GLUCOSE METABOLISM	10
RNA INTERFERENCE	10
INFLAMMATORY RESPONSE	9
MISMATCH REPAIR	9
PROGRAMMED CELL DEATH	9
TRANSPOSITION	9
PIGMENTATION	9
CELL-CELL ADHESION	9
GENE SILENCING	9
INVASIVE GROWTH	8
INTERPHASE	8

- **Ovarian Cancer**

Gene name	Citations
BRCA1	935
BRCA2	679
P53	501
CA125	362
TUMOR SUPPRESSOR PROTEIN P53	351
VASCULAR ENDOTHELIAL GROWTH FACTOR A	236
VASCULAR ENDOTHELIAL GROWTH FACTOR	226
VEGF	225
ERBB-2 RECEPTOR	224
BCL-2	169
EPIDERMAL GROWTH FACTOR RECEPTOR	166
PROTO-ONCOGENE PROTEINS C-BCL-2	149
AKT	129
P-GLYCOPROTEIN	114
PROTO-ONCOGENE PROTEINS C-AKT	113
ESTROGEN RECEPTOR	109
COLLAGEN	103
P21	101
KI-67 ANTIGEN	100
EGFR	100
1-PHOSPHATIDYLINOSITOL 3-KINASE	99
HER-2/NEU	98
TUMOR NECROSIS FACTOR-ALPHA	94
VIMENTIN	93
CASPASE 3	92
CASPASE-3	91
MMP-2	85
BAX	85
PTEN	83
COX-2	83
CYCLOOXYGENASE 2	83
EPIDERMAL GROWTH FACTOR	82
LPA	81
MATRIX METALLOPROTEINASE 2	80
CARCINOEMBRYONIC ANTIGEN	78
KI-67	76
PROGESTERONE RECEPTOR	75
KERATINS	75
MMP-9	75
CYCLIN-DEPENDENT KINASE INHIBITOR P21	75
CYTOKERATIN	74
GONADOTROPIN-RELEASING HORMONE	74
CEA	72
TP53	71
CASP3	71
E-CADHERIN	71
TELOMERASE	71
CDKN1A	71
HER-2	69
NODAL	68
AKT1	67
SD	65

FSH	64
MATRIX METALLOPROTEINASE 9	64
IFN-GAMMA	63
HER2	63
UPA	61
LAMININ	61
LUCIFERASE	61
INTERFERON TYPE II	60
TRANSFORMING GROWTH FACTOR BETA	60
BETA-CATENIN	59
CYCLOOXYGENASE-2	59
EGF	58
URINARY PLASMINOGEN ACTIVATOR	58
ERK	58
MLH1	57
PTGS2	57
ERK1/2	56
LH	55
CI	55
PI3K	55
SURVIVIN	55
CYCLIN D1	54
BCL-2-ASSOCIATED X	54
APRIL	53
PROLIFERATING CELL NUCLEAR ANTIGEN	53
ACTIN	53
CALRETININ	53
GRANULOCYTE COLONY-STIMULATING FACTOR	52
BETA CATENIN	52
PTEN PHOSPHOHYDROLASE	52
P16	51
NF-KAPPA B	51
ESTROGEN RECEPTOR ALPHA	51
INTERLEUKIN-6	50
IL-6	50
P27	50
MMP	49
CD95	49
FAS LIGAND	49
INTEGRIN	49
TUBULIN	48
CD44	47
CA-15-3 ANTIGEN	47
ANDROGEN RECEPTOR	46
GLUTATHIONE TRANSFERASE	46
TOPOISOMERASE I	46
TNF-ALPHA	45

Biological Process	Citations
APOPTOSIS	913
INDUCTION	520
TRANSCRIPTION	433

CELL PROLIFERATION	404
CELL CYCLE	370
CELL GROWTH	329
ANGIOGENESIS	304
PATHOGENESIS	261
LOCALIZATION	205
CELL DEATH	195
DRUG RESISTANCE	185
METHYLATION	178
SECRETION	166
DNA REPAIR	139
INDUCTION OF APOPTOSIS	111
OVULATION	109
MENOPAUSE	104
CELL ADHESION	93
IMMUNE RESPONSE	92
CELL MIGRATION	86
CELL CYCLE ARREST	71
HYPERSENSITIVITY	71
TRANSPORT	61
RNA INTERFERENCE	60
MISMATCH REPAIR	58
DNA METHYLATION	57
CIRCULATION	52
TRANSLATION	52
S PHASE	48
ONCOGENESIS	47
CELL MOTILITY	43
CELL KILLING	42
MITOSIS	40
SENSITIZATION	37
MENARCHE	34
FERTILIZATION	34
MENSTRUAL CYCLE	33
PROTEOLYSIS	24
GENE SILENCING	24
HOMEOSTASIS	23
AGING	22
MENSTRUATION	21
COAGULATION	21
CELL DIFFERENTIATION	20
PROGRAMMED CELL DEATH	20
CASPASE ACTIVATION	19
TROPISM	19
CHEMOTAXIS	18
CONJUGATION	17
DNA REPLICATION	17

- **Leukemia**

Gene name	Citations
LEUKEMIA INHIBITORY FACTOR	788
LIF	666
INTERLEUKIN-6	579
CASPASE-3	394
BCL-2	356
CASPASE 3	353
PROTO-ONCOGENE PROTEINS C-BCL-2	333
P53	289
CASP3	277
CYTOCHROME C	267
TUMOR NECROSIS FACTOR-ALPHA	263
IL-6	242
MLL	241
P-GLYCOPROTEIN	205
TUMOR SUPPRESSOR PROTEIN P53	205
CD4	198
BAX	195
MYELOID-LYMPHOID LEUKEMIA	194
STAT3	193
CD34	187
CML	185
GRANULOCYTE COLONY-STIMULATING FACTOR	178
G-CSF	161
TNF-ALPHA	160
GP130	156
STAT3 TRANSCRIPTION FACTOR	156
REVERSE TRANSCRIPTASE	152
CORE BINDING FACTOR ALPHA 2 SUBUNIT	151
NF-KAPPA B	150
LIFR	146
CASPASE-8	137
PARP	137
NF-KAPPAB	137
HISTONE DEACETYLASE	137
GRANULOCYTE-MACROPHAGE COLONY-STIMULATING FACTOR	136
ERK	136
PROTEIN KINASE C	132
P21	131
AKT	129
PROMYELOCYTIC LEUKEMIA PROTEIN	127
GM-CSF	124
1-PHOSPHATIDYLINOSITOL 3-KINASE	123
OSM-LIF RECEPTORS	123
CD95	120
LEUKEMIA INHIBITORY FACTOR RECEPTOR ALPHA SUBUNIT	118
INTERFERON TYPE II	117
IFN-GAMMA	117
RUNX1	116
C-MYC	114
PROTO-ONCOGENE PROTEINS C-AKT	112
CD3	112

VEGF	112
CYTOCHROMES C	112
WT1	111
BCL-2-ASSOCIATED X	111
TELOMERASE	110
CASPASE 8	108
CASPASE 9	107
CASPASE-9	107
CD8	105
CYTOKINE RECEPTOR GP130	103
FLT3	102
VASCULAR ENDOTHELIAL GROWTH FACTOR	101
IL-2	98
TNF	97
BCL-X	97
INTERLEUKIN-2	96
TUMOR NECROSIS FACTOR	95
JNK	95
CYCLIN-DEPENDENT KINASE INHIBITOR P21	95
VASCULAR ENDOTHELIAL GROWTH FACTOR A	91
HEMOGLOBIN	90
STEM CELL FACTOR	90
IL6ST	90
CDKN1A	90
ONCOSTATIN M	89
CASP9	88
CD56	86
PKC	85
MONOCYTIC LEUKEMIA	84
CD11B	83
CILIARY NEUROTROPHIC FACTOR	83
ERYTHROPOIETIN	82
INTERFERON	82
MAPK	82
CASP8	81
JAK2	80
POLY(ADP-RIBOSE) POLYMERASE	79
P-GP	79
CD45	78
GLUTATHIONE TRANSFERASE	78
P38	77
EPIDERMAL GROWTH FACTOR	76
BCL-XL	76
C-KIT	76
CD14	75
INTERFERON-ALPHA	75
HSC	74
BID	73

Biological Process	Citations
APOPTOSIS	2194
INDUCTION	1516
TRANSCRIPTION	1260

CELL CYCLE	708
CELL DEATH	636
CELL PROLIFERATION	536
PATHOGENESIS	473
CELL GROWTH	445
INDUCTION OF APOPTOSIS	382
LOCALIZATION	285
CELL DIFFERENTIATION	254
DRUG RESISTANCE	218
SECRETION	201
TRANSPORT	194
METHYLATION	189
IMMUNE RESPONSE	158
ANGIOGENESIS	157
CELL CYCLE ARREST	154
CASPASE ACTIVATION	139
DNA REPAIR	106
HOMEOSTASIS	97
TRANSLATION	96
S PHASE	82
DNA METHYLATION	70
REGENERATION	70
CELL ADHESION	68
CELL ACTIVATION	67
ONCOGENESIS	64
PROGRAMMED CELL DEATH	61
G1 PHASE	59
INTERPHASE	59
MITOSIS	59
CELL DEVELOPMENT	58
RNA INTERFERENCE	58
VIRAL REPLICATION	55
HYPERSENSITIVITY	55
CELL DIVISION	53
DNA REPLICATION	52
EMBRYONIC DEVELOPMENT	51
COAGULATION	51
GENE SILENCING	48
CYTOKINE PRODUCTION	47
CELL KILLING	47
MEMBRANE FUSION	45
TROPISM	45
CHEMOTAXIS	45
CIRCULATION	43
AGING	42
PROTEOLYSIS	41
SENSITIZATION	40

References

[1] HL7 Clinical-Genomics SIG: Genotype Shared Model, http://hl7.org/library/committees/clingenomics/docs/HL7-Clinical-Genomics-Genotype-0.15.zip, 2005.

[2] S. R. Huang, "Personalized Medicine for Cancer," 2007, http://ben-may.bsd.uchicago.edu/bmi2/news/huggins_lectures/huggins_lectures.html.

[3] Herceptin: Novel therapy targets HER2-positive breast cancer, http://www.mayoclinic.com/health/Herceptin/BR00012, 2006.

[4] A. Shabo, "The Implications of Electronic Health Records for Personalized Medicine," *Personalized Medicine,* vol. 3 pp. 251-258, 2005.

[5] BioLab Experiment Assistant (BEA), http://www.biovista.com/pub/doc_view/default.asp?doc_id=6&msgId=74&categ=on&tree=99&ext_doc_id=46&start_doc_id=6&lang_id=0, 2008.

[6] A. Pollack, "Special Drug Just for You, At the End of a Long Pipeline," in *New York Times*, 2005.

[7] E. P. Bottinger, "Foundations, promises and uncertainties of personalized medicine," *Mt Sinai J Med,* vol. 74, pp. 15-21, 2007.

[8] H. Moon, H. Ahn, R. L. Kodell, S. Baek, C. J. Lin, and J. J. Chen, "Ensemble methods for classification of patients for personalized medicine with high-dimensional data " *Artif Intell Med,* vol. 41, pp. 197-207, 2007.

[9] A. A. Alizadeh, D. T. Ross, C. M. Perou, and M. v. d. Rijn, "Towards a novel classification of human malignancies based on gene expression patterns," *Journal of Pathology,* pp. 41-52, 2001.

[10] J. N. Hirschhorn and M. J. Daly, "Genome-wide association studies for common diseases and complex traits.," *Nat Rev Genet,* vol. 6, pp. 95-108, 2005.

[11] http://www.emea.europa.eu/pdfs/human/ich/43798606en.pdf, 2008.

[12] S. S.A., "Clinical Genomics Data Standards for Pharmacogenetics and pharmacogenomics," *Pharmacogenomics,* vol. 7, pp. 1-7, 2006.

[13] AmpliChip, http://www.amplichip.us/documents/CYP450_P.I._US-IVD_Sept_15_2006.pdf.

[14] E. P. Bottinger, "Foundations, promises and uncertainties of personalized medicine," *Mt Sinai Journal of Medicine,* vol. 74, pp. 15-21, 2007.

[15] HUGO Gene Nomenclature Committee, http://www.genenames.org/index.html.

[16] National Center fo Biotechnology Information (NCBI), http://www.ncbi.nlm.nih.gov/sites/entrez.

[17] J. N. Hirschhorn and M. J. Daly, "Genome-wide association studies for common diseases and complex traits," *Nat Rev Genet,* vol. 6, pp. 95-108, 2005.

[18] H. M. Kantarjian, M. Talpaz, and S. O'Brien, "Survival benefit with imatinib mesylate versus interferon-based regimens in newly diagnosed chronic-phase chronic myelogenous leukemia," *Blood,* vol. 108, pp. 1835–1840, 206.

[19] E. Jabbour, J. Cortes, F. Giles, and H. Kantarjian, " Current Perspectives on the Treatment of Patients with Chronic Myeloid Leukemia: An Individualized Approach to Treatment.," *Cancer Journal,* vol. 13, pp. 357-365, 2007.

[20] B. B. Spear, M. Heath-Chiozzi, and J. Huff, "Clinical application of pharmacogenetics," *Trends on Molecular Medicine,* vol. 7, pp. 201-204, 2001.

[21] The Case of Personalized Medicine, http://www.personalizedmedicinecoalition.org/communications/TheCaseforPersonalizedMedicine_11_13.pdf, 2008.

[22] F. André, J. Domont, and S. Delaloge, "What can breast cancer molecular sub-classification add to conventional diagnostic tools? ," *Ann Oncol,* vol. 18 pp. ix33-6, 2007.

[23] M. J. v. d. Vijver, Y. D. He, L. J. v. t. Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. v. d. Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards, "A gene-expression signature as a predictor of survival in breast cancer," *The New England Journal of Medicine (NEJM),* vol. 347, pp. 1999-2009, 2002.

[24] P. Dinh, C. Sotiriou, and M. J. Piccart, "The evolution of treatment strategies: Aiming at the target," *Breast,* vol. 2, pp. 10-16, 2007.

[25] R. Rouzier, K. Anderson, and K. R. Hess, "Basal and luminal types of breast cancer defined by gene expression patterns respond differently to neoadjuvant chemotherapy," in *San Antonio Breast Cancer Symposium* San Antonio, TX, 2004.

[26] C. Sotiriou, S. Y. Neo, L. M. McShane, E. L. Korn, P. M. Long, A. Jazaeri, P. Martiat, S. B. Fox, A. L. Harris, and E. T. Liu, "Breast cancer classification and prognosis based on gene expression profiles from a population-based study," *The Proceedings of the National Academy of Sciences Online (US),* vol. 100, pp. 10393-8, 2003.

[27] C. Sotiriou and M. J. Piccart, "Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care?," *Nature Reviews Cancer,* vol. 7, pp. 545-53, 2007.

[28] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M. J. v. d. Vijver, J. Bergh, M. Piccart, and M. Delorenzi, "Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis," *Journal of the National Cancer Institute,* vol. 98, pp. 262-72, 2006.

[29] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. v. d. R. M, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, and A. L. G. Børresen-Dale, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *The Proceedings of the National Academy of Sciences Online (US),* vol. 98, pp. 10869-74, 2001.

[30] J.R. Pollack, "A Perspective on DNA Microarrays in Pathology," *The American Journal of Pathology,* vol. 171, 2007.

[31] J. N. Hirschhorn, K. Lohmueller, E. Byrne, and K. A. Hirschhorn, "Comprehensive review of genetic association studies," *Genetics in Medicine,* vol. 4, pp. 45-61, 2002.

[32] HapMap Project, http://www.hapmap.org/, 2008.

[33] R. J. Klein, C. Zeiss, E. Y. C. EY, J. Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable, and J. Hoh, "Complement factor H polymorphism in age-related macular degeneration," *Science,* vol. 308, pp. 385-394, 2005.

[34] J. J. Swen, T. W. Huizinga, H. Gelderblom, E. G. d. Vries, W. J. Assendelft, J. Kirchheiner, and H. J. Guchelaar, "Translating pharmacogenomics: challenges on the road to the clinic," *PLoS Med,* vol. 4, 2007.

[35] J. Kirchheiner, K. Brøsen, M. L. Dahl, L. F. Gram, S. Kasper, I. Roots, F. Sjöqvist, E. Spina, and J. Brockmöller, "CYP2D6 and CYP2C19 genotype-based dose recommendations for antidepressants: a first step towards subpopulation-specific dosages," *Acta Psychiatrica Scandinavica,* vol. 101, pp. 173-192, 2001.

[36] M. Ingelman-Sundberg, S. C. Sim, A. Gomez, and C. Rodriguez-Antona, "Influence of cytochrome P450 polymorphisms on drug therapies: Pharmacogenetic,pharmacoepigenetic and clinical aspects," *Pharmacology & Therapeutics,* vol. 116, pp. 496-526, 2007.

[37] S. J. Gardiner and E. J. Begg, "Pharmacogenetics, drug-metabolizing enzymes, and clinical practice," *Pharmacological Reviews,* vol. 58, p. 3, 2006.

[38] Cytochrome P450 Homepage, http://drnelson.utmem.edu/CytochromeP450.html, 2008.

[39] Home Page of the Human Cytochrome P450 (CYP) Allele Nomenclature Committee http://www.cypalleles.ki.se/.

[40] M. Widschwendter, K. D. Siegmund, H. M. Müller, H. Fiegl, C. Marth, E. Müller-Holzner, P. A. Jones, and P. W. Laird, "Association of breast cancer DNA methylation profiles with hormone receptor status and response to tamoxifen. Cancer Res. 2004," *Cancer Research,* vol. 64, pp. 3807-3813., 2004.

[41] M. Widschwendter, "5-methylcytosine--the fifth base of DNA: the fifth wheel on a car or a highly promising diagnostic and therapeutic target in cancer?," *Dis Markers,* vol. 23, pp. 1-3, 2007.

[42] M. Ingelman-Sundberg, "Pharmacogenetics: an opportunity for a safer and more efficient pharmacotherapy," *Journal of Internal Medicine,* vol. 250, pp. 186-200, 2001.

[43] B. D. Juran, L. J. Egan, and K. N. Lazaridis, "The AmpliChip CYP450 Test: principles, challenges and future clinical utility in digestive diseases," *Clin Gastroenter & Hepatol,* vol. 4, pp. 822-830, 2006.

[44] A. Mahgoub, J. R. Idle, L. G. Dring, R. Lancaster, and R. L. Smith, "Polymorphic hydroxylation of Debrisoquine in man," *The Lancet,* vol. 2, pp. 584-590, 1977.

[45] R. Weinshilboum, "Inheritance and drug response," *The New England Journal of Medicine (NEJM),* vol. 348, pp. 529-566, 2003.

[46] M. P. Goetz, S. K. Knox, V. J. Suman, J. M. Rae, S. L. Safgren, M. M. Ames, D. W. Visscher, C. Reynolds, F. J. Couch, W. L. Lingle, R. M. Weinshilboum, E. G. Fritcher, A. M. Nibbe, Z. Desta, A. Nguyen, D. A. Flockhart, E. A. Perez, and J. N. Ingle, "The impact of cytochrome P450 2D6 metabolism in women receiving adjuvant tamoxifen," *Breast Cancer Research and Treatment* vol. 101, pp. 113-21, 2007.

[47] P. K. Janicki, H. G. Schuler, T. M. Jarzembowski, and M. Rossi, "2nd. Prevention of postoperative nausea and vomiting with granisetron and dolasetron in relation to CYP2D6 genotype," *Anesthesia and analgesia* vol. 102, pp. 1127-33, 2006.

[48] M. K. Kim, J. Y. Cho, H. S. Lim, K. S. Hong, J. Y. Chung, K. S. Bae, D. S. Oh, S. G. Shin, S. H. Lee, D. H. Lee, B. Min, and I. J. Jang, "Effect of the CYP2D6 genotype on the pharmacokinetics of tropisetron in healthy Korean subjects," *European Journal of Clinical Pharmacology* vol. 59, pp. 111-6, 2003.

[49] K. Takada, M. Arefayene, Z. Desta, C. H. Yarboro, D. T. Boumpas, J. E. Balow, D. A. Flockhart, and G. G. Illei, "Cytochrome P450 pharmacogenetics as a predictor of

toxicity and clinical response to pulse cyclophosphamide in lupus nephritis," *Arthritis & Rheumatism,* vol. 50, pp. 2202-10, 2004.

[50] Y. Li, J. Hou, H. Jiang, D. Wang, W. Fu, Z. Yuan, Y. Chen, and L. Zhou, "Polymorphisms of CYP2C19 gene are associated with the efficacy of thalidomide based regimens in multiple myeloma," *Haematologica,* vol. 92, pp. 1246-9, 2007.

[51] http://www.avazmd.com/resources/emr_cpr_ehr.html.

[52] D. Garets and M. Davis, http://www.himssanalytics.org/docs/WP_EMR_EHR.pdf.

[53] ACC Group, http://www.acgroup.com.

[54] HIMSS Electronic Health Record Definitional Model, http://www.himssanalytics.org/docs/WP_EMR_EHR.pdf

[55] Electronic Health Records: Definition, Scope and Context, http://secure.cihi.ca/cihiweb/en/downloads/infostand_ihisd_isowg1_mtg_denoct_contextdraft.pdf

[56] HL7 Electronic Health Record Specification download.

[57] "Open Exploration on Electronic Health Records," 2007.

[58] Health Level Seven (HL7) - ANSI Accredited Standards Developing Organization.

[59] European Standardization of Health Informatics (CEN/TC 251), Electronic Healthcare Record Communication.

[60] open EHR.

[61] DICOM Standard Documents, http://medical.nema.org/dicom/2003.html.

[62] R. Hussein, U. Engelmann, A. Schroeter, and H.-P. Meizer, "DICOM Structured Reporting," *RadioGraphics,* vol. 24, pp. 891-896, 2004.

[63] D. Sluis, K. P. Lee, and N. Mankovich, "DICOM SR – integrating structured data into clinical information systems," *MEDICAMUNDI,* vol. 46, pp. 31-36, 2002.

[64] R. Vogl, C. Laucher, R. Penz, P. Schirmer, T. Schabetsberger, and E. Ammenwerth, "A Survey on Shared Electronic Health Record Architectures in Europe," 2007, http://www.telemed-berlin.de/telemed2007/Beitraege/TELEMED-2007-02-Vogl.pdf.

[65] International Organization for Standardization, www.iso.org.

[66] Dolin R.H., Alschuler L., Boyer S., Beebe C., Behlen F.M., Biron P.V., and Shabo S.A., "HL7 Clinical Document Architecture," *J Am Med Inform Assoc,* vol. 13, pp. 30-39, 2006.

[67] T. Beale, "Archetypes: Constraint-based Domain Models for Future-proof Information Systems," in *OOPSLA 2002 Workshop on Behavioural Semantics* 2002.

[68] S. B. T. F. G. M. A. R. P. O. Heeard, "Templates and Arcehtypes: how do we know what we are talking about?," 2003.

[69] "Archetype Definition Language (ADL)," OpneEHR Foundation 2004, http://www.openehr.org/drafts/ADL-1_2_draftF.pdf.

[70] M. B. S. a. A. V. M. a. K. G. Oliveras, "Survey of Electronic Health Record Standards," DEIM-RR-06-004, 2006, http://deim.urv.es/recerca/reports/DEIM-RR-06-004.html.

[71] M. Eichelberg, T. Aden, J. Riesmeier, A. Dogac, and G. B. Laleci, "Electronic Health Record Standards - A Brief Overview."

[72] HL7 Version 3: An impact assessment, www.nhsia.nhs.uk/hl7/pages/HL7_impass_v1.0.pdf 2001.

[73] B. Smith and W. Ceusters, "HL7 RIM: An Incoherent Standard " in *Medical Informatics Europe*, Maastricht, Netherlands, 2006, pp. 133-138.

[74] Archetype paradigm: an ICT revolution is needed, http://www.eurorec.org/files/filesPublic/GF%20Archetype%20Paradigm%20February%2020071.pdf, 2007.

[75] Connecting for Health, www.connectingforhealth.nhs.uk.

[76] Nationale knooppunt en kenniscentrum voor ICT en innovatie In de Zorg (NICTIZ), www.nictiz.nl/.

[77] Standards for E-Health Interoperability: An E-Health Transition Strategy, http://www.nehta.gov.au/index.php?option=com_docman&task=doc_download&gid=252&Itemid=139, 2007.

[78] D. A. Clunie, *DICOM Structured Report*. Bangor, Pennsylvania: PixelMed Publishing, 2000.

[79] A. TIRADO-RAMOS, J. HU, and K. P. LEE, "Information Object Definition–based Unified Modeling Language Representation of DICOM Structured Reporting: A Case Study of Transcoding DICOM to XML," *Journal of the American Medical Informatics Association,* vol. 9, pp. 63-72, Jan/Feb 2002.

[80] A. Shabo and D. Dotan, "The Seventh Layer of the Clinical-genomics Information Infrastructure," *IBM Systems Journal,* vol. 46, pp. 57-67, 2007.

[81] "MicroArray and Gene Expression Markup Language (MAGE-ML) ". vol. 2007, http://www.mged.org/Workgroups/MAGE/mage-ml.html.

[82] BSML Tutorial, http://www.bsml.org/i3c/docs/BSML3_1_Tutorials.pdf, 2007.

[83] "ISO/TC 215, Health Informatics, Health Informatics Profiling Framework," 2005.

[84] Electronic Medical Records vs. Electronic Health Records: Yes, There Is a Difference, http://www.himssanalytics.org/docs/WP_EMR_EHR.pdf

[85] "The Role of the Personal Health Record in the EHR," *Journal of AHIMA,* vol. 76, July-August 2005 2005.

[86] Connecting Americans To Their Health Care: A Common Framework for Networked Personal Health Information, http://www.connectingforhealth.org/commonframework/docs/P9_NetworkedPHRs.pdf , 2006.

[87] SNOMED CT - the language of the NHS Care Records Service, http://www.connectingforhealth.nhs.uk/systemsandservices/data/snomed/snomed-ct-a.pdf.

[88] The International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10), http://www.cihi.ca/cihiweb/dispPage.jsp?cw_page=codingclass_icd10_e, 2007.

[89] MEDCIN: a nomenclature for coding the clinical encounter, http://www.medicomp.com/.