SINGLE NUCLETIDE POLYMORPHISM (SNP) DATA INTEGRATED ELECTRONIC HEALTH RECORD (EHR) FOR PERSONALIZED MEDICINE

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS INSTITUTE
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

TİMUR BEYAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
HEALTH INFORMATICS

JANUARY 2014

SINGLE NUCLETIDE POLYMORPHISM (SNP) DATA INTEGRATED ELECTRONIC HEALTH RECORD (EHR) FOR PERSONALIZED MEDICINE

Submitted by **Timur Beyan** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in the Department of Health Informatics**, **Middle East Technical University by**,

Prof. Dr. Nazife Baykal _____
Director, Graduate School of Informatics

Assist. Prof. Dr. Yeşim Aydın Son _____
Head of Department, Health Informatics

Examining Committee Members

Prof. Dr. Ünal Erkan Mumcuoğlu _____
Department of Health Informatics, METU

Assist. Prof. Dr. Yeşim Aydın Son _____
Head of Department, Health Informatics, METU

Assist. Prof. Dr. Aybar Can Acar _____
Department of Health Informatics, METU

Assist. Prof. Dr. Kemal Hakan Gülkesen _____
Department of Biostatistics and Medical Informatics,
Akdeniz University

Assist. Prof. Dr. Tuğba Taşkaya Temizel _____
Department of Information Systems, METU

Date: 15.01.2014

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: Timur Beyan

Signature :

ABSTRACT

SINGLE NUCLETIDE POLYMORPHISM (SNP) DATA INTEGRATED ELECTRONIC HEALTH RECORD (EHR) FOR PERSONALIZED MEDICINE

Beyan, Timur
Ph.D., Department of Health Informatics
Supervisor: Assist. Prof. Dr. Yeşim Aydın Son

January 2014, 217 pages

The digital age is revolutionizing the old and historical population-based healthcare paradigm towards personalized medicine. Traditional diagnostic approaches fail to define treatment response or prognosis. Focusing on manifest symptoms often hides risk factors and, so prevention opportunities of diseases disappear.

Today, it's known that most of the complex diseases are result of interaction of genomic, environmental and behavioral factors and personalized medicine is defined as the use of these data to determine individual patterns of disease. Personalized medicine aims to deliver a more accurate representation of medical conditions that are multidimensional, predictive, preventive, pharmacologically effective, person-centered, and individualistic services. However, to reach personalized medicine opportunities, it's an obligation to extend current Electronic Health Record standards and capabilities to support genomic data in healthcare settings.

In this thesis, we developed genomic sequence variation data integrated capabilities for personalized medicine practices based on an integrative approach which will allow us to use genome-wide SNP profiling data and disease models within electronic/personal health records to support knowledge based systems. Finally, developed capabilities were represented and assessed for prostate cancer in a data set as a pilot study.

Keywords: National Health System of Turkey, Personalized Medicine, SNP Genotyping Data, Disease Risk Assessment, Cumulative Models

ÖZ

TEK NÜKLEOTİD POLİMORFİZM (TNP) VERİSİ BÜTÜNLEŞİK ELEKTRONİK SAĞLIK KAYDI (ESK)

Beyan, Timur
Ph.D., Department of Health Informatics
Supervisor: Assist. Prof. Dr. Yeşim Aydın Son

Ocak 2014, 217 sayfa

Sayısal çağ, eski ve tarihsel topluma dayalı tıbbi bakım paradigmasını kişiselleştirilmiş tıbba doğru devrimsel şekilde dönüştürmektedir. Geleneksel tanısal yaklaşımlar tedaviye tepki ve prognozu tanımlamada başarısız kalmaktadır. Apaçık belirtilere odaklanmak sıkça risk faktörlerini gizlemekte ve böylece hastalıkları önleme fırsatları gözden kaybolmaktadır.

Bugün bilinmektedir ki, karmaşık hastalıkların çoğu genomik, çevresel ve davranışsal faktörlerin etkileşimi ile ortaya çıkmaktadır ve kişiselleştirilmiş tıp bu tür verilerin bireysel hastalık örüntülerini belirlemek üzere kullanımı olarak tanımlanmaktadır. Kişiselleştirilmiş tıp, tıbbi durumların daha doğru bir görünümünü yani kişi merkezli, çok boyutlu, öngörücü, önleyici, farmakolojik olarak etkin ve bireysel hizmetleri sunmaktadır. Bununla beraber, kişiselleştirilmiş tıp fırsatlarına erişmek için, güncel ESK standart ve yeteneklerini tıbbi bakım ortamlarında genomik veriyi destekleyecek şekilde genişletmek bir zorunluluktur.

Bu tezde, bilgiye dayalı sistemleri desteklemek üzere, elektronik/kişisel sağlık kayıtlarında bütünsel genom TNP profilleme verisi ve hastalık modellerini kullanmaya izin verecek bütünleşik bir yaklaşıma dayalı kişiselleştirilmiş tıp uygulamaları için TNP verisi bütünleşik yetenekler geliştirdik. Sonunda, geliştirilen yetenekler pilot bir araştırma olarak prostat kanseri için bir veri setinde sunulmuş ve değerlendirilmiştir.

Keywords: Türkiye Ulusal Sağlık Sistemi, Kişiselleştirilmiş Tıp, TNP Genotipleme Verisi, Hastalık Risk Değerlendirme, Birikimli Modeller

to my Family

ACKNOWLEDGEMENTS

The author wishes to express his sincere gratitude to **Assist. Prof. Dr. Yeşim Aydın Son** for her limitless patience, professional mentoring and scientific insight throughout the research.

The author would also like his deepest appreciation to **Prof. Dr. Ünal Erkan Mumcuoğlu, Assist. Prof. Dr. Aybar Can Acar, Assist. Assist. Prof. Dr. Kemal Hakan Gülkesen, Prof. Dr. Ergin Soysal,** and **Assist. Prof. Dr. Tuğba Taşkaya Temizel** for their sincere suggestions, scientific leading and invaluable support.

At the end, the author would like to express appreciation to his family (his beloved wife Oya Deniz and daughter Irmak) for their patience and support throughout the entire period.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A	: Adenine
ASTM	: American Society for Testing and Materials
BioXM™	: BioXM™ Knowledge Management Environment
BMI	: Body Mass Index
BPH	: Benign Prostate Hyperplasia
C	: Cytosine
Cancer GAMAdb	: Cancer Genome-wide Association and Meta Analyses Database
Cancer GEM KB	: Cancer Genomic Evidence-based Medicine Knowledge Base
CBO	: Clinical Bioinformatics Ontology
CCR	: Continuity of Care Record
CCRS	: Central Civil Registration System
CDA	: Clinical Document Architecture
CDA R2	: CDA Release Two
CDCV	: Common Disease–Common Variant
CDRV	: Common Disease-Rare-Variant
CDSS	: Clinical Decision Support System
CG-ASSOC.	: Clinicogenomic Associations
ClinGenKB	: Clinicogenomic Knowledge Base
ClinGenWeb	: Clinicogenomic Web Application
CR-SNP	: Clinically Relevant SNP
dbGaP	: Database of Genotypes and Phenotypes
dbSNP	: Database of Single Nucleotide Polymorphisms
dbVar	: Database of Single Nucleotide Polymorphisms
DGVa	: Database of Genomic Variants Archive
DNA	: Deoxyribonucleic Acid
DTC	: Direct-to-consumer
EGAPP	: Evaluation of Genomic Applications in Practice and Prevention
EHR	: Electronic Health Record
EMBL	: European Molecular Biology Laboratory
EMR	: Electronic Medical Record
EWAS	: Environmental Wide Association Studies
G	: Guanine
GWAS	: Genome Wide Association Studies
HCRS	: Health Coding Reference Server
HGMD	: Human Gene Mutation Database
HGNC	: Human Gene Nomenclature Committee
HGP	: Human Genome Project
HGVS	: Human Genome Variation Society
HL7	: Health Level 7
HL7 CG-SIG	: HL7 Clinical Genomics Special Interest Group
HL7 V3	: HL7 Version 3
ICD	: International Classification of Diseases
ID3	: Iterative Dichotomiser 3
IPHR	: Interactive Preventive Health Record
JSNP	: Japanese Single Nucleotide Polymorphism

LOINC	: Logical Observation Identifiers Names and Codes
LSDB	: Locus Specific Databases
MHDS	: Minimum Health Data Sets
mRNA	: Messenger RNA
NCBI	: National Center of Biotechnology Information
NHDD	: National Health Data Dictionary
NGS	: Next Generation Sequencing
NHIS-T	: National Health Information System of Turkey
NHRI	: The National Human Genome Research Institute
OMIM	: Online Mendelian Inheritance in Man
PACS	: Picture Archiving and Communication
PCPGM	: Partners HealthCare Center for Personalized Genetic Medicine
PharmKB	: Pharmacogenomics Knowledgebase
PHR	: Personal Health Record
RNA	: Ribonucleic Acid
SNOMED	: Systematized Nomenclature of Medicine
SNP	: Single Nucleotide Polymorphism
SSI	: Social Security Institution (of Turkey)
SVM	: Support Vector Machine
T	: Thymine
T2DM	: Type 2 Diabetes Mellitus
WES	: Whole Exome Sequencing
WGS	: Whole Genome Sequencing
XML	: Extensible Markup Language

CHAPTER 1

INTRODUCTION

1.1 Motivation

The digital age is revolutionizing the old and historical population-based healthcare paradigm towards personalized medicine. Traditional medical approaches are not sufficiently predictive and preventive, as they focus on the manifest symptoms that often hide risk factors. Whereas determining risk factors would allow prevention through early diagnosis. Personalized medicine provides new opportunities based on person-centered, predictive, preventive, and effective health care services (Downing, 2009).

Genomic data and its derivatives (transcriptomes, proteomes, metabolomes, etc.) are the most important elements of personalized medicine (Arnold & Vockley, 2011), (Ginsburg & Willard, 2009). Every individual has around four million variations in their own genome, when compared to the reference sequence. Genomic variations can range from single nucleotide changes to gain or loss of whole chromosomes. Single nucleotide polymorphisms (SNPs), where a single nucleotide in the genome alter between individuals or paired chromosomes, are about %90 of genomic variants and some already are important markers in clinic, while others are on the way (Barnes, 2010).

The rapid developments in **Next Generation Sequencing (NGS)** technologies have substantially reduced both the cost and the time required to sequence an entire human genome, and it's expected that NGS-based analyses e.g. **Whole Genome Sequencing (WGS)** and **Whole Exome Sequencing (WES)** will be available for routine use in healthcare and prevention of disease in the near future (Berg, et al., 2011). Providing genomic data to medical professionals will facilitate clinical decisions based on individual's genome and allow tailoring health care services to patients' specific needs and characteristics (Scheuner, et al., 2009). In parallel, **direct-to-consumer (DTC)** genome wide profiling tests are being developed to assess individual disease risks for many common polygenic diseases (Bloss, et al., 2011). DTC genomic companies, e.g. 23andMe, GenePlanet, and DNA DTC generally perform a gene-chip analysis of SNPs using Deoxyribonucleic Acid (DNA) extracted from saliva or serum sample (Helgason & Stefánsson, 2010), (Chua & Kennedy, 2012), (Gullapalli, et al., 2012).

In clinical decision processes, genomic variant data would be used for assessing disease risks, predicting susceptibility, early clinical diagnosing, following the course of the disease, targeted screening, and planning treatment regimens (Ginsburg & Willard, 2009) (Chan & Ginsburg, 2011). A reasonable way to carry this personalized approach into routine for medical practices would be integrating genotype data and its clinical interpretation within the electronic healthcare record systems (Belmont & McGuire, 2009), (Scheuner, et al., 2009), (Hudson, 2011).

Today, in many developed and developing countries, use of health information systems is inevitable for healthcare providers for reimbursement of services and tracking the quality of the healthcare provided (Garets & Davis, 2006), (Häyrinen, et al., 2008). Recently, several

systems and networks have been constituted in many countries of the world, including National Health Information System (NHIS-T) of Turkey (HIMSS Global Enterprise Task Force, 2010). These systems and networks have high potential for integrating genomic data in healthcare practices for personalized medicine.

1.2 Problem Statement

Although, various initiatives was emerged to use genomic data in clinical process e.g. risk assessment, pharmacotherapy, molecular diagnosis etc., examples of genome enabled EMR/EHR systems in routine daily medical practice are so rare.

Today, numerous GWAS studies were performed, a great number of disease associated genomic variation were discovered and some predictive models were proposed to use in clinical settings. But yet, these efforts mostly were not implemented in real clinical environment. Lack of conversion of research data into daily practice may depends on several technical, medical and socio-cultural reasons, but one of the main reasons is definitely the complex and continuously evolving nature of the problem area.

Regarding Turkey, with Health Transformation Program, a national level EHR was implemented and integration to this system became an obligation for caregiver organizations. Despite the lack of clinical genomics researches except a few academic attempts, existing national EHR has a great potential for personalized medicine practices.

In our thesis, we have investigated how NHIS-T can be transformed into a genome enabled national EHR and what are the obstacles, possible solutions and additional requirements of this attempt would be.

1.3 Contributions

- Our thesis is one of the first attempts of genome enabled EHR. There are various examples in different scopes and sizes but the difference of our work is that it's the unique effort to incorporate SNP data into a national level EHR.

- For our study, we have reviewed literature and produced a comprehensive document for requirements of a genome enabled EHR.

- We have developed a methodology to improve existing NHIS-T as a SNP enabled NHIS-T. In our analysis, we have also determined and discussed additional requirements and capabilities, potential problem areas and obstacles.

- We have established knowledge base content for prostate cancer and determined possible decision models, and compare their performances in a case study.

- An example of clinicogenomic knowledge base for predictive medicine was designed and a prototype was developed.

- We designed and developed a simple decision support tool for the end user level, that is capable of applying different clinical interpretation and assessment approaches for SNP data based prostate cancer risk assessment.

- The critical point of study is that our system ensure to process both SNP and external parameters (i.e. family health history and lifestyle data which couldn't been recorded routinely in EHR e.g. BMI, smoking, alcohol consumption etc.) due to bipartite structure of our approach (i.e. conversion of CR-SNP into clinicogenomic associations based on knowledge base and final clinical interpretation of these associations in the end-user application).

1.4 Organization of the Thesis

The main goal of our study is to develop complementary capabilities as prototypes for SNP enabled NHIS-T i.e. clinicogenomic knowledge base and end user decision support applications, which specifically focus on disease risk assessment. Organization of this study is summarized in Figure 1.

First, NHIS-T was analyzed using official technical documentations and the findings were summarized regarding architecture, messaging infrastructure, and terminology standards. In parallel, to determine requirements for a genome enabled EMR/EHR, a comprehensive literature survey was performed. In this work, findings were presented as standards and messaging, clinicogenomic knowledge bases, clinicogenomic decision support, and examples of genome enabled EHR/EMR. All these information and complementary concepts about personalized medicine and clinical use of genomic data were presented as the literature review in Chapter 2.

General methods of these processes and enabling technologies were explained in Chapter 3.

Then, in Chapter 4, developing architecture for SNP enabled NHIS-T were proposed and discussed. As a part of this process, possible ways of incorporating SNP data files into messaging infrastructure using HL7 v3 CDA R2 standard, were explained. In architectural analysis, it was determined that, we need to develop two complementary capabilities for SNP enabled NHIS-T i.e. a knowledge base and clinical decision support capabilities for end users.

In Chapter 5, design and development principles were explained. In this chapter, after discussion of possible architectures, clinicogenomic associations were studied and standardized association and model definition tables were described. Finally, development methods of knowledge base and decision support application were argued and prototype capabilities were presented.

Then, we focused on to prostate cancer risk assessment. We extracted content of complementary capabilities (SNP-disease associations for knowledge base and existing risk assessment models for end user decision support capabilities) from publicly available databases and scientific literature. These subjects explained in Chapter 6 and 7.

At the end (in Chapter 8), we evaluated complementary components using real data from personal genome project which is publicly available resource for genomic, environmental and human trait data (https://my.personalgenomes.org/public_genetic_data) based on different assessment and reporting approaches.

Finally, we discussed our principle results and limitations in Chapter 9, and future works in Chapter 10.

In our study, as an initial attempt through development of much sophisticated infrastructure, we concentrated on the SNP variant data interpretation for prostate cancer as a proof of concept experiment and excluded other types of variants and diseases.

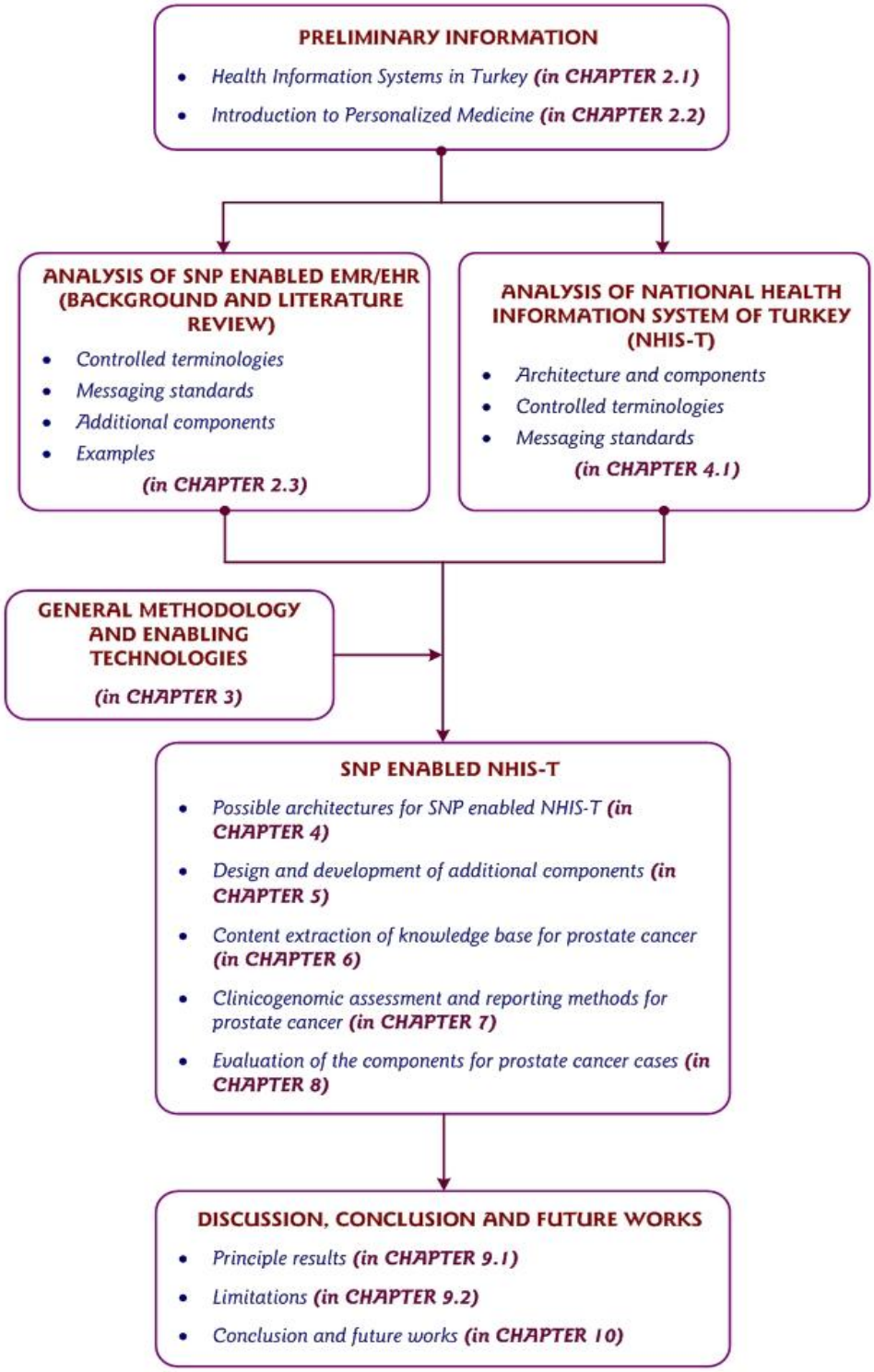


Figure 1: Organization of the thesis.

In addition, security and privacy issues and constraints about hardware and infrastructure are also excluded. Also, the use of personal clinicogenomic information to determine disease risk of patient's family members is considered as out of scope.

CHAPTER 2

BACKGROUND AND LITERATURE REVIEW

2.1 Health Information Systems in Turkey

2.1.1 Information Systems in Healthcare

Current healthcare systems are widely based on standalone or integrated information system infrastructures. **Health Information Systems** capture, store, share, transmit and manage data about health of the individuals or the transactions of the healthcare organizations. This concept includes integrated hospital and primary care information systems, clinical, laboratory, pharmacy, radiology and nuclear medicine information systems, patient administration, human resources, logistics and accounting management systems, and Picture Archiving and Communication Systems (PACS), etc. (Winter, et al., 2011).

Regarding the capabilities of health information systems three terms come into prominence; Electronic Medical Record (EMR), Electronic Health Record (EHR) and Personal Health Record (PHR) (Häyrinen, et al., 2008).

EMR is composed of clinical data repositories, clinical decision support systems, standard medical terminologies, computerized order entry, and documentation applications. Erroneously, **EHR** is sometimes used interchangeably with EMR but essentially they are two different concepts. An EHR comprise the ability for sharing healthcare data among partners and is acceptable as an extraction of healthcare organization's EMR based on some standards e.g. Continuity of Care Document of Health Level 7 (HL7), Continuity of Care Record (CCR) of American Society for Testing and Materials (ASTM), etc. (Garets & Davis, 2006).

PHR provides individuals to access their own medical data and engage in the healthcare (Tran & Gonzales, 2012). **Interactive preventive health record (IPHR)** is a new type of PHR primarily aiming health promotion and disease prevention via educational support, disease risk calculation, reminders and other types of decision support approaches (Krist, et al., 2011).

2.1.2 National Health Information System of Turkey (NHIS-T)

Turkey Health Transformation Program has been gradually implemented since 2003. This program aimed to transform all aspects of national healthcare system e.g. organization, services, responsibilities, finance, etc. As declared on e-health strategy, using information and communication technologies is essential and obligatory component of this transformation (OECD, 2007).

Today, there are many integrated systems for different aspects of Turkish national healthcare system e.g. NHIS-T for patient records, claim processing and reimbursement systems of

Social Security Institution (SSI), Turkish Drug and Medical Device National Databank for medical material identification, etc. (Table 1).

NHIS-T is a national level infrastructure which has centralized service oriented architecture in order to produce and share medical records between stakeholders (HIMSS Global Enterprise Task Force, 2010), (Dogac, et al., 2011).

Table 1: Major electronic healthcare systems in Turkey (except NHIS-T)

Claim processing and reimbursement systems of SSI (MEDULA in Turkish)	Secondary and tertiary level caregiver organizations must use these systems via web services for reimbursement of SSI.
Turkish Drug and Medical Device National Databank	Running to trace and control medical devices at national level with the cooperation of governmental authorities.(http://ubb.iegm.gov.tr)
Family Medicine Information System	Implementing to monitor the performance and activities of family practitioners. Since 2012, integrated to NHIS-T.
Core Resource Management System of Ministry of Health	Combination of Material Resource Management System, Human Resources Management System, Investment Surveillance System, Private Health Institutions Management System.
112 Emergency Service Information Management System	An integrated system that enables tracking from notice of medical emergencies to being archived electronically along with coordinating the emergency units in the process and realizes corporate sources' planning and management.
Turkish Pharmaceutical Track & Trace System	An infrastructure for units belonging to each pharmaceutical product in Turkey, to guarantee and provide the reliable supply of every single drug unit from production to consumption. (http://itsportal.saglik.gov.tr/)
Central Hospital Appointment System	An application that enables people to make appointments by themselves according to their choice of secondary and tertiary care hospitals, oral and dental health centers and doctors. (http://www.mhrs.gov.tr/Vatandas/)

Data elements like name, address, marital status, main diagnosis, treatment method, diastolic blood pressure, healthcare institution, etc. used in the NHIS-T are defined, and then **Minimum Health Data Sets (MHDS)** are generated combining relevant data elements. Both the data elements and MHDS are published as a **National Health Data Dictionary (NHDD**). The last version of NHDD, which includes 418 pieces of data elements, and 64 pieces of data set, is version 2.1 and accessible from its official web site (Republic of Turkey Ministry of Health, 2013). It is mandatory for healthcare providers of Turkey to conform to the NHDD data definitions and MHDS. New MHDS are produced by existing data elements or the NHDD is improved by identifying new data elements when required.

The data elements are coded using medical terminology systems which are accessible from the **Health Coding Reference Server (HCRS)** or locally defined categorical values, such as gender or marital status. There are 294 code systems in HCRS and the current version of the HCRS is 3.0 and is available online via web services. A tabular representation is also accessible in official web page (Republic of Turkey Ministry of Health, 2012) and allows users to query through web browsers. The healthcare professional identities are stored in central Doctor Data Bank and citizen identification is stored in Central Civil Registration

System (CCRS). Both identities are validated against their original sources at storing in central repositories (Dogac, et al., 2011).

HL7 Clinical Document Architecture (CDA) a document markup standard, is produced to exchange information as part of the HL7 Version 3 (V3) standards, and aim to specify the structural and semantic aspects of clinical documents (Benson, 2010). In NHIS-T, MHDS are produced as aggregated clinical document elements named as **transmission data sets** or **episodic EHRs** and then serialized into XML based on the HL7 CDA R2 structure to create **transmission schemas**. Before storing in the NHIS-T central repositories, incoming messages are validated regarding syntax semantics and messages passed these two steps are stored in the central NHIS-T repositories (Figure 2) (Kose, et al., 2008), (Dogac, et al., 2011).

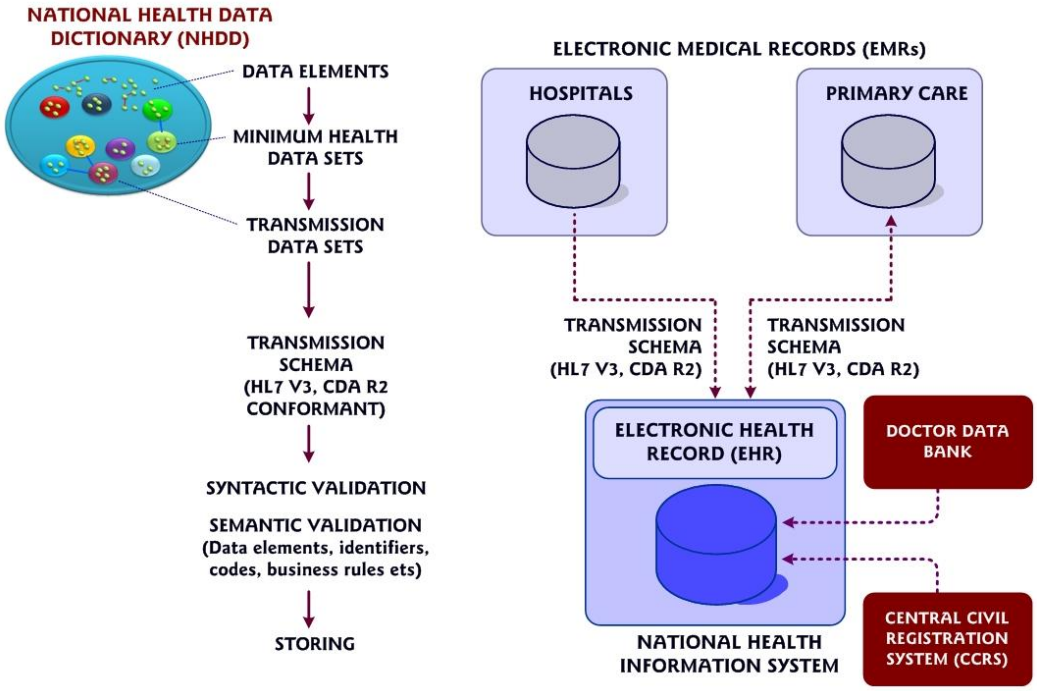


Figure 2: Schematic Representation of NHIS-T (depicted based on current literature).

Current version of NHIS-T allows transfer of the medical data from care providers' (hospital and family practitioner) information systems to central servers via web services. It has the infrastructure that will provide access to patient's records for authorized healthcare professionals within the hospital, and that will allow patients to reach their own medical data i.e. PHR. But, the legal regulations have to be completed before both type of access, authorized or self, is available. Then, the establishment of a PHR system will allowed (Dogac, et al., 2011).

2.2 Introduction to Personalized Medicine

Personalized medicine is a healthcare paradigm that aims to use individual's unique clinical, genomic, environmental, behavioral and sociocultural characteristics to predict disease susceptibility, determine molecular characterization of disease for early diagnosis, tailor treatment regimens, and monitor prognosis. This emerging approach is based on new

discoveries in bioinformatics i.e. *omics revolution* and provide us new opportunities for precise, preventive, and effective medical care based on omics data (Downing, 2009), (Schneider & Orchard, 2011).

Some authors proposed three phases for omics data based medical approach. In the first phase, polymorphism data and mutations of germline cells were used to realize personalized medicine. In the second step i.e. post-genomic omics based medicine, the essential target is to deliver predictive and preventive services based on broad molecular profiles of somatic cells. And finally, omics based systems medicine aim to provide comprehensive (personalized, predictive and preventive) medicine based on molecular pathways and networks and its variations by diseases (Tanaka, 2010), (Shimokawa, et al., 2011). Regarding systems medicine, human health and disease states are defined by interconnected molecular pathways that can be defined by the omics data e.g. genome, epigenome, transcriptome, proteome, and metabolome (Chan & Ginsburg, 2011).

The great idea behind the personalized medicine is to incorporate clinical and laboratory phenotype data and molecular profiling to provide precisely tailored health interventions. For this reason, the term of precision medicine is sometimes used as synonyms of personalized medicine (Mirnezami, et al., 2012)

2.2.1 Omes and Omics

Omics revolution, at first, emerged based on the genome sequencing studies. *Genome* is the whole set of genetic material involved in a nucleated cell. The term of genome was introduced in 1920 by the German botanist Hans Winkler, combining *gen*e and chromos*ome*.

Chromosomes are built from DNA sequences which contain about 3×10^9 nucleotides. Nucleotides are the basic building blocks of the genetic material and every nucleotide contains phosphate, carbon, sugar structures and an organic base (Figure 3).

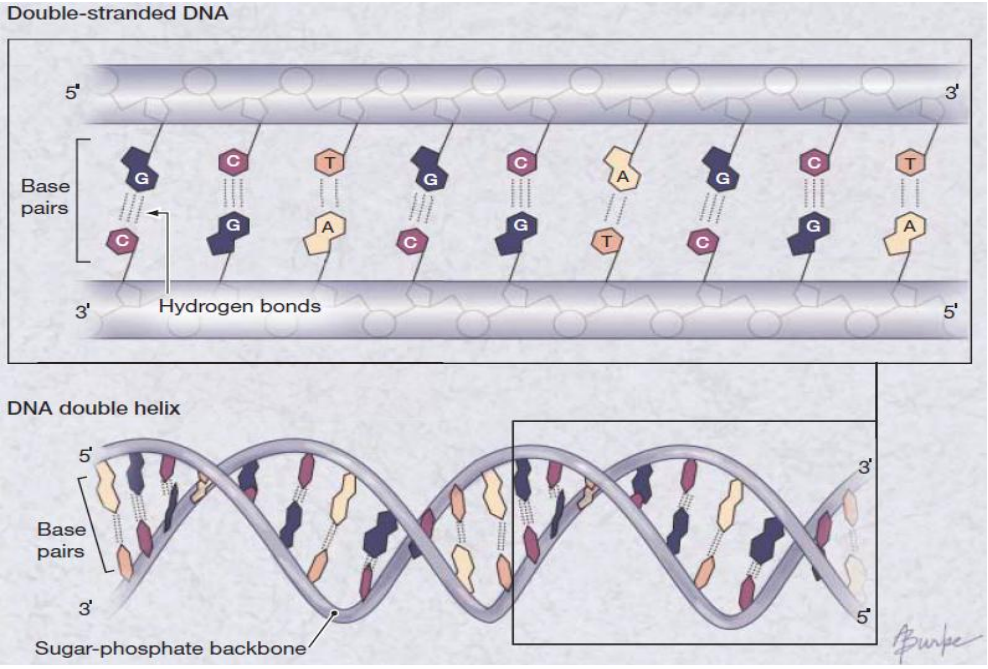


Figure 3: DNA Structure (Attia, et al., 2009A).

Nucleotide is entitled according to the base e.g. Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). DNA has a double helix structure formed by two strands of nucleotides. In this structure, specific base pairs must be matching e.g. A with T, and C with G. An Adenine-Thymine (A-T) connection has two and a Cytosine-Guanine (C-G) connection has three hydrogen bonds (Attia, et al., 2009A), (Schaaf, et al., 2012).

DNA part coding biological information of the organisms is named as **gene** (Figure 4). DNA has two types of sequences i.e. exon and intron. **Exon** is the functional part and accountable in genetic expression. **Introns** are the non-coding parts between exons and not responsible to produce mature mRNA (Brown, 2009), (Pierce, 2010). About one percent of the human genome is exons and 99 percent is non-coding regulatory elements, RNA processing elements, and regions of unknown function. Most of the disease-causing mutations (>85%), are in exonic parts (Bamshad, et al., 2011), (Majewski, et al., 2011).

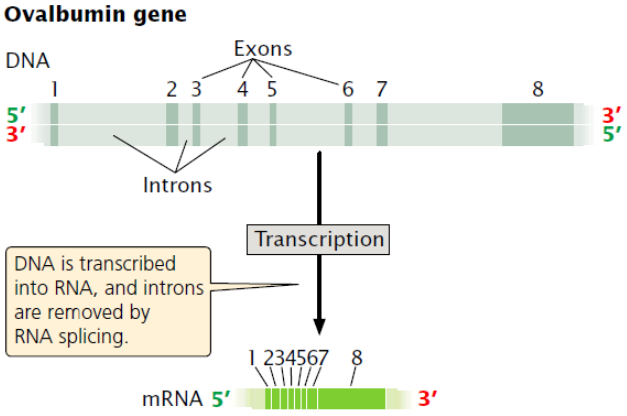


Figure 4: Structure of a Gene (Pierce, 2010).

DNA is a double stranded stretch (plus and minus or forward and reverse respectively), and every nucleated somatic cell has 22 pairs autosomal and one pair sex chromosome. This means for autosomal chromosomes we have two version of DNA strands inherited via maternal and paternal sex cells. Different forms or variations of a particular polymorphism is called as **allele** (Attia, et al., 2009A).

Gene expression is a sequence of subcellular complex reactions aiming to convert inherited data (i.e. gene) into functional chemical molecules. mRNA (transcriptome) is synthesized from DNA strand in cell nucleus (transcription) (Figure 5).

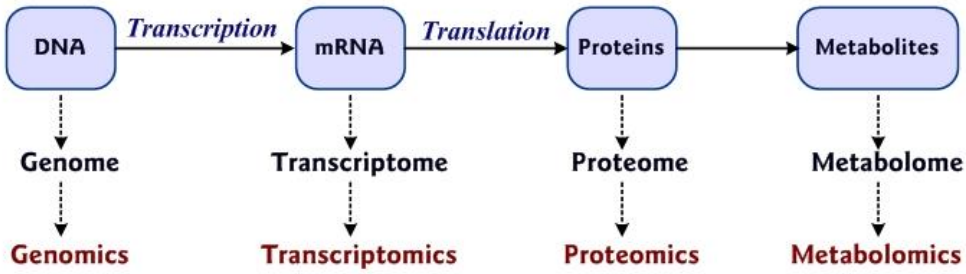


Figure 5: Conceptual Representation of Gene Expression, and Corresponding -Omes and –Omics (Dziuda, 2010).

Then, in ribosomes, proteins are produced from the amino acids using mRNA as a template (translation). Proteins are converted several metabolic products with enzymatic bioreactions (Brown, 2009).

Today, "*ome*" suffix is used to define several subcellular chemicals produced after these reactions. "*Omics*" are scientific research areas dealing with -omes using high-throughput screening techniques and producing big data (Dziuda, 2010), (Gubb & Matthiesen, 2010), (Schneider & Orchard, 2011).

2.2.2 Components of Diseases

2.2.2.1. Monogenic and Polygenic Diseases

Almost all medical conditions have a genetic basis. In monogenic diseases, DNA variations of one single gene are predominantly or completely responsible from pathogenesis of diseases. Monogenic diseases are frequently inherited in one of various patterns, depending on the involved gene e.g. autosomal, X-linked, mitochondrial etc. (Table 2) (Janssens & van Duijn, 2008)**,** (National Library of Medicine (US), 2013).

Table 2: Inheritance patterns of monogenic diseases (National Library of Medicine (US), 2013).

Inheritance pattern	**Description**	**Examples**
Autosomal dominant	One changed (mutation) gene copy in each cell is adequate to manifest clinical condition.	Huntington disease,

		Neurofibromatosis type 1
Autosomal recessive	Two mutated gene copies are required for disease manifestation.	Cystic Fibrosis, Sickle Cell Anemia
X-linked dominant	Mutational genes are in X chromosome and two mutated gene copies are required for manifestation. Thereby, disease risk is more in females than males.	Fragile X Syndrome
X-linked recessive	Mutational genes are in X chromosome, but for manifestation of clinical condition, it's required to have only mutated copies. Because fathers cannot pass X-linked traits to their sons, males are more often vulnerable than females.	Hemophilia, Fabry disease
Co-dominant	Two types of alleles of a gene can be expressed. Both alleles affect the inheritance and identify the features of the genetic situation.	ABO blood group, Alpha-1 Antitrypsin Deficiency
Mitochondrial	Because mitochondrial gene transferred only maternal ovum, fathers cannot pass these types of conditions to their children.	Leber Hereditary Optic Neuropathy

Genetic origin of common complex or multifactorial diseases is more complicated from monogenic diseases. Common medical conditions, such as heart disease, diabetes, schizophrenia, many types of cancers, and obesity are complex and multifactorial conditions which are caused by combination of multiple mutations on different genes, lifestyle and environmental components. Complex diseases don't follow the strict inheritance pattern as in monogenic diseases (Janssens & van Duijn, 2008), (National Library of Medicine (US), 2013).

There are millions of common variations in every population. Each common variation may play a small role in the pathogenesis of a complex disease, but collectively all variations may be a strong factor behind the molecular etiology of the disease. In the presence of specific variation patterns, with the involvement of environmental and behavioral causes clinical conditions may be triggered (Figure 6). In such cases, if people with high risk based on their genotypic profiles can avoid the risk factors, they can prevent themselves from possible manifest of clinical conditions (National Cancer Institute, 2013A).

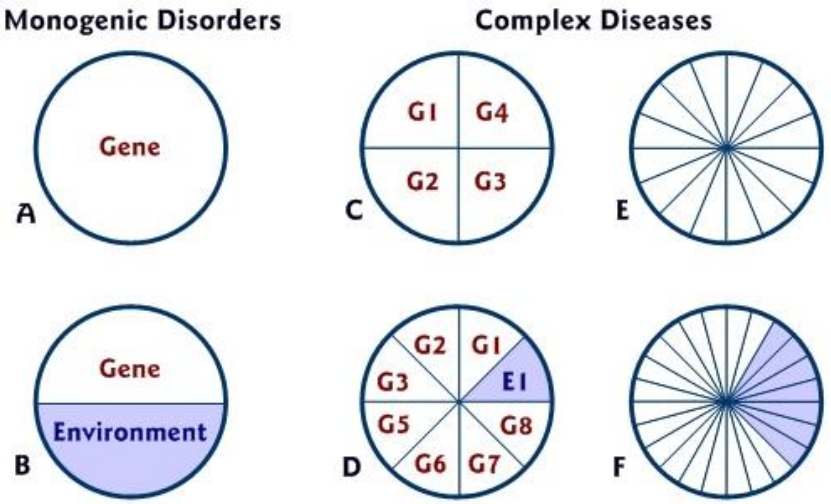


Figure 6: Pathogenic Models of Diseases (Janssens & van Duijn, 2008).
A (Huntington Disease) and B (Phenylketonuria) for monogenic conditions. C-F for polygenic complex diseases. Genetic factors are represented as white areas and environmental factors are represented as grey areas.

2.2.2.2. Genomic Variants

Genomic DNA sequence is about 99.9% identical among humans (Dziuda, 2010). Compared with the reference sequence generated by the Human Genome Project (HGP), any single individual's genome has about 3-4 million variants (Drmanac, 2012).

Genomic variations can range from single nucleotide changes to gain or loss of whole chromosomes. In single nucleotide polymorphism (SNP), a single nucleotide in the genome sequence is different between individuals. Genomic variation can also be caused by insertion or deletion of nucleotides (indels) e.g. variable or simple number tandem repeat polymorphisms, block substitutions, inversion variants and, copy number variations (Frazer, et al., 2009), (Barnes, 2010) (Figure 7).

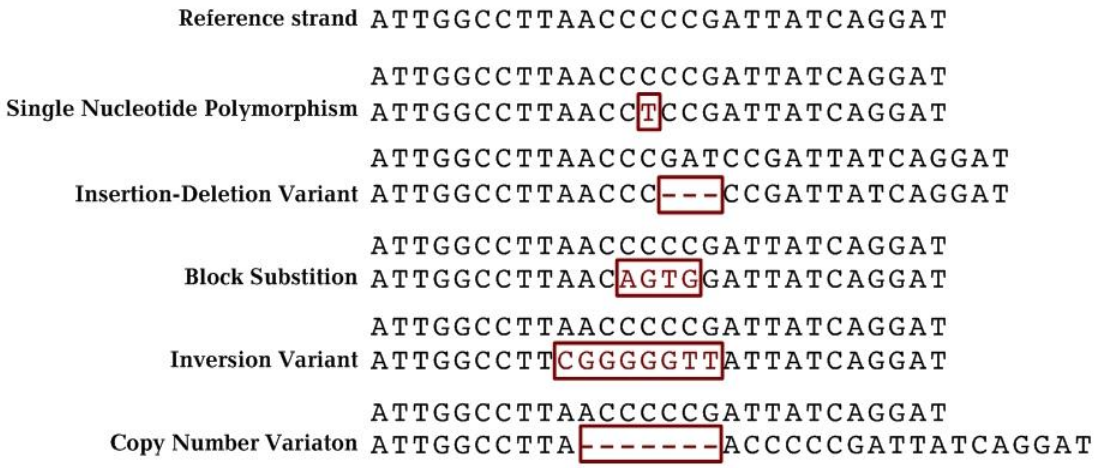


Figure 7: Various genomic variation types (Frazer, et al., 2009).

SNPs are about 90% of all the genomic variations. Although most of are harmless, some of them have great values for disease risk assessment, medical diagnostics and pharmaceutical products (Poo, et al., 2011), (Aronson, et al., 2012), (Drmanac, 2012).

With the advent of NGS technologies, it's also possible to accomplish rapid and cheap WGS. Researchers and clinicians expect that WGS will be one of the most important tools in the personalized medicine era (Berg, et al., 2011), (Scheuner, et al., 2009), (Wright, et al., 2011).

2.2.3 Genomic Tests and Personal Genomics

Conventionally, genetic examinations have been divided into two distinct but connected disciplines: molecular genetics (goal is to test specific small variations) and cytogenetics

(holistic analysis to identify big structural variation). It's expected that, WGS will fill this chasm providing genome-wide molecular feature explaining both small and large variation (Wright, et al., 2011).

2.2.3.1. Genome Sequencing Techniques

Describing the allele at a specific location in the genome is identified as **genotyping**. Various genotyping techniques are developed in time for both diagnostic and research purposes. These techniques contain a broad range from whole genome scanning to analysis of particular sequence variations. Study of particular sequence variations can be performed using polymerase chain reaction (PCR), multiplex ligation-dependent probe amplification, fluorescence in situ hybridization, DNA microarrays and mass spectrometry among other techniques. In last thirty years, with the help of automation and development of high throughput techniques, provided huge improvements for DNA sequencing (Figure 8). (Wright, et al., 2011).

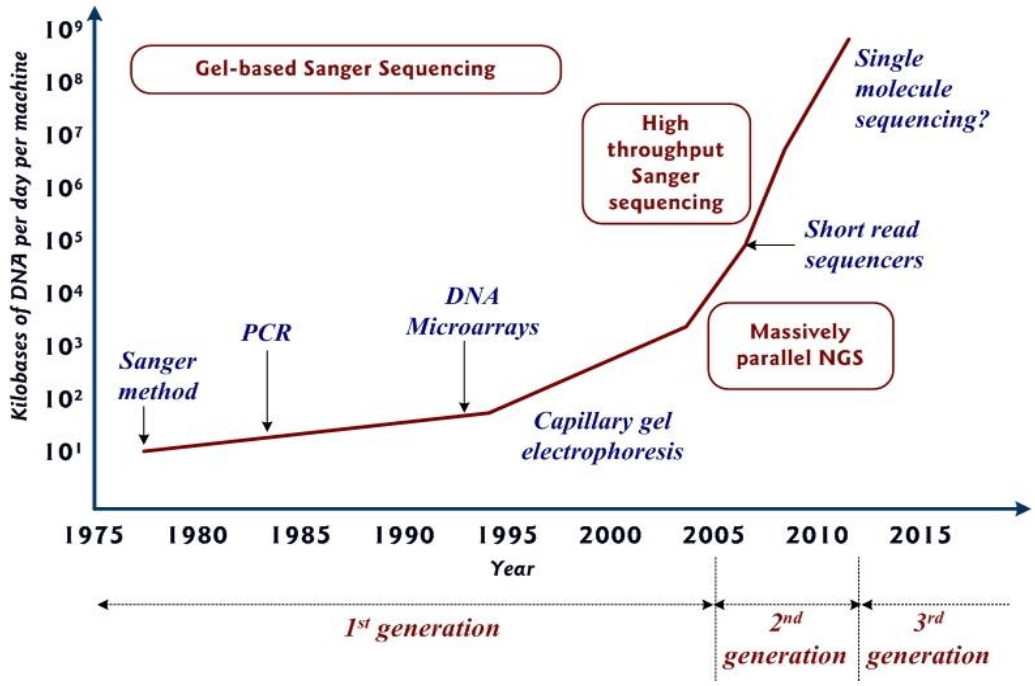


Figure 8: DNA sequencing generations and processing capabilities (Wright, et al., 2011).

With the older DNA sequencing approaches, it was not feasible to produce a whole genome sequence. Due to time and cost limitations, using genomic data in clinical processes have been characteristically restricted. Today, new high-throughput and massively parallel NGS technologies can analyze millions of DNA fragments simultaneously. NGS-based analyses contain WGS and WES. These have considerably decreased both the time and cost restrictions to analyze whole genome sequence (Wright, et al., 2011), (American Medical Association, 2012).

In a characteristic human genome, WGS will determine more than 3 million variations. After the filtering processes, hundreds to thousands of clinically relevant variations, which have

the potential to be underlying reason, could be extracted. WES analyzes all of the exon sequences i.e. exome. In WES, at first, the exons are separated from the non-coding DNA, and then analyzed. Usually, around 20,000 individual variations are determined. As costs continue to decrease, it's expected that WES will be progressively interchanged with WGS (Bamshad, et al., 2011), (Bick & Dimmock, 2011), (Biesecker, 2012), (Raffan & Semple, 2011).

2.2.3.2. The Informatics Pipeline for Genome Sequencing

The informatics pipeline for NGS can roughly be separated into three methodical steps; primary, secondary, and tertiary analytic phases (Figure 9).

At the beginning of sequencing processes, thousands of images are captured as **raw (level-0) data**. Raw data are the base signals in one tile for a certain sequence location (Röhm & Blakeley, 2009), (Nielsen, et al., 2011), (Wright, et al., 2011).

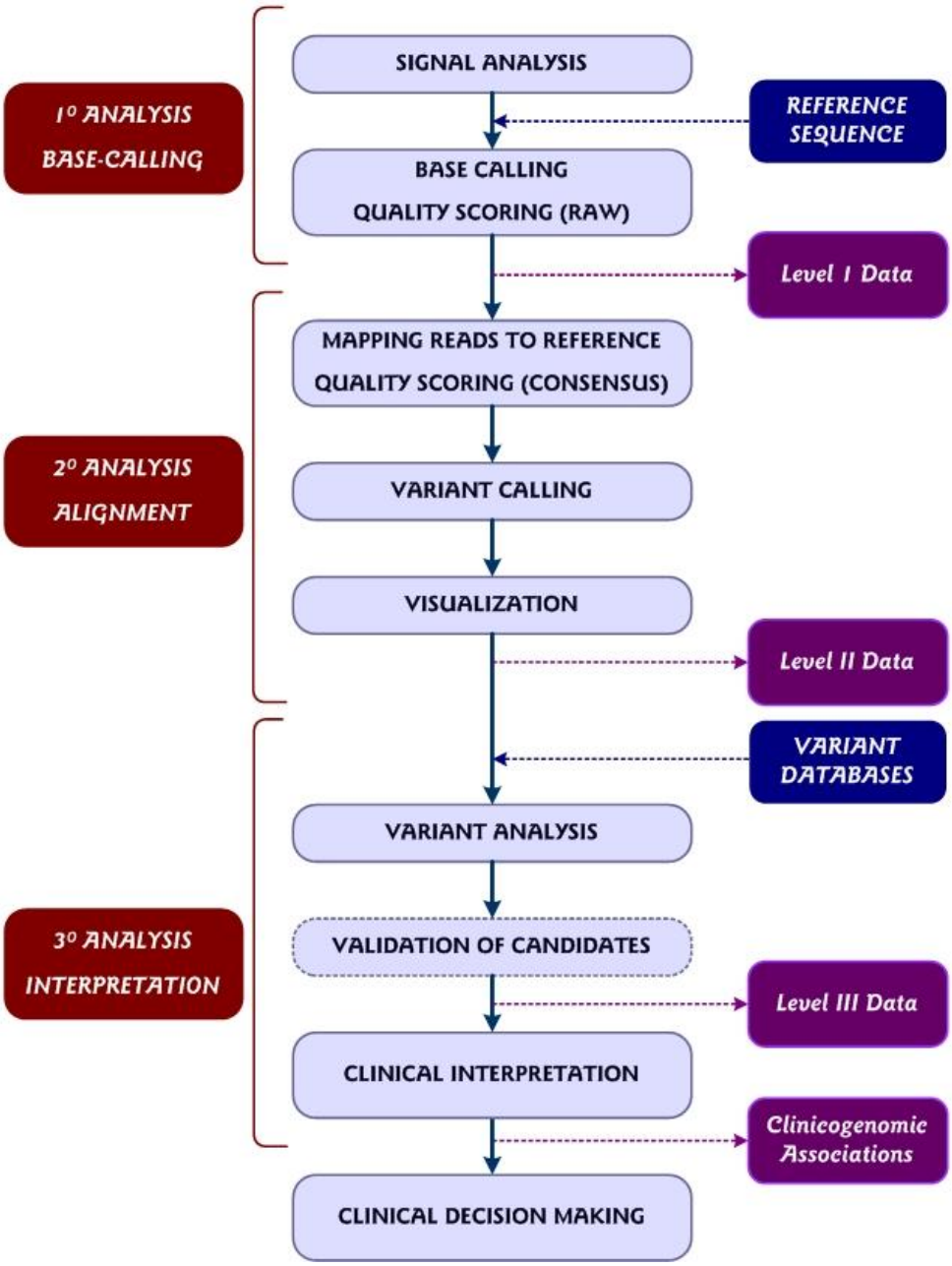


Figure 9: Phases of NGS (Wright, et al., 2011).

In the **primary analysis** phase, captured images are analyzed and base-calling are carried out. In this phase, intensities of light signals are transformed into nucleotide sequences. The output of this phase i.e. **level-1 data** is a file comprising millions of short sequence fragments for each lane and some metadata about each read. Size of every produced level-1 file includes several gigabytes of data. Some examples of file formats generated in this stage are FASTA, FASTQ, SCARF, QSEQ, SRA, RAW, and TXT formats (Röhm & Blakeley, 2009), (Nielsen, et al., 2011), (Wright, et al., 2011).

In the **secondary analysis** phase, DNA reads map to a reference sequence and the variations are determined. The results of this phase are a sorted lists of matches and named as **level-2 data**. Some examples of file formats generated in this stage are SAM, BAM, and vendor specific formats (Röhm & Blakeley, 2009), (Nielsen, et al., 2011), (Wright, et al., 2011).

In the **tertiary analysis,** variations are analyzed to evaluate their origin, uniqueness and likely functional impact using various databases, algorithms and software packages. This phase is based on statistical analysis. The output is **level-3 data (human readable text) file** and some example file formats generated in this stage are VCF, BCF, GVF, and GFF formats (Röhm & Blakeley, 2009), (Nielsen, et al., 2011), (Wright, et al., 2011).

After determining variants, in the clinicogenomic analysis, pathogenic mutations for specific phenotypes (e.g. medical conditions, drug interactions, etc.) are identified. In clinicogenomic analysis, in first, non-pathogenic variations are excluded using different filters, and clinically relevant variants are gathered. In final step of clinicogenomic analysis, relevant variants are interpreted associating with relevant phenotypic and clinical information and **clinicogenomic associations** are extracted (Wright, et al., 2011).

In NGS tests (i.e. WES and WGS) millions of individual variations are extracted. A single organization cannot curate this type of data file (Aronson, et al., 2012). Many online genomic variation resources are developed to filter and interpret personal genomic data (Table 3).

Table 3: Some online genomic variation sources.

Genomic Variation Databases	
dbSNP (Database of Single Nucleotide Polymorphisms)	Simple genetic polymorphisms database (NCBI). http://www.ncbi.nlm.nih.gov/snp
dbVar (Database of Single Nucleotide Polymorphisms)	Structural variation database (NCBI). http://www.ncbi.nlm.nih.gov/dbvar
DGVa (Database of Genomic Variants Archive)	Structural variation database (European Bioinformatics Institute, EBI). http://www.ebi.ac.uk/dgva
1000Genomes	A catalog of SNPs, structural variants, and their haplotype contexts around the world. http://www.1000genomes.org/
Genotype-Phenotype Research Databases	
dbGaP (Database of Genotypes and Phenotypes)	Store and share the results of genotype-phenotype studies (NCBI). http://www.ncbi.nlm.nih.gov/gap
European Genome-Phenome Archive	Store and share the genotypic and phenotypic data resulting from research projects (EMBL). https://www.ebi.ac.uk

Table 3 (cont.): Some online genomic variation sources.

Clinicogenomic (Disease-Variation Associations) Databases	
OMIM (Online Mendelian Inheritance in Man)	Catalog of human genes, genetic disorders and traits, regarding the molecular basis. http://www.omim.org
HGMD (Human Gene Mutation Database)	Published gene variations responsible for human inherited diseases. http://www.hgmd.cf.ac.uk
CDC HuGENavigator	A knowledge base in human genome epidemiology (prevalence of variations, gene-disease, gene-gene and gene-environment interactions, and evaluation of genetic tests). http://www.hugenavigator.net
GWAS Central	Summarized findings of human GWAS. https://www.gwascentral.org
Locus-specific databases (LSDB)	Many LSDBs which are typically curated by gene experts without centralized editing. http://www.hgvs.org/dblist/glsdb.html
ClinVar	Associations among human variations and phenotypes with supportive evidences. http://www.ncbi.nlm.nih.gov/clinvar
AlzGene Database	A broad field summary of GWAS studies performed in Alzheimer's disease. http://www.alzgene.org
PDGene Database	A broad field summary of GWAS studies performed in Parkinson's disease. http://www.pdgene.org
SzGene Database	A broad field summary of GWAS studies performed in schizophrenia. http://www.szgene.org
Other type of sources	
PharmKB (Pharmacogenomics Knowledge Base)	A knowledge source including clinically relevant genotype-phenotype and gene-drug relationships. http://www.pharmgkb.org
SNPedia	A summarized wiki resource of human genetic variation as published in peer-reviewed studies. http://www.SNPedia.com

Today, various tools and techniques are developed for all three phases of NGS analyzing. In the first two phases, analyzes are becoming progressively automated and reliable. However, clinical interpretation is still a major challenge (Wright, et al., 2011).

2.2.3.3. Direct to Consumer (DTC) Genomic Testing

DTC genomic testing does not need the support of a medical doctor or other kinds of healthcare professional to acquire. These sorts of genomic tests comprise carrier testing,

pharmacogenomic testing, and predictive testing for multifactorial complex diseases e.g. hereditary cancers; cardiovascular disease, and depression (American Medical Association, 2010).

As sequencing and genotyping technologies gets cheaper and faster, DTC companies emerged (e.g. 23andMe, GenePlanet, and DNA DTC, etc.), who markets DTC personal genomic services analyzing SNPs to assess polygenic disease risks (Helgason & Stefánsson, 2010), (Bloss, et al., 2011).

2.2.4 Environmental Components of Diseases

Since the HGP completed in 2003, the epidemiological researchers focused on to determine causative polymorphisms as genetic determinants of diseases (Lioy & Rappaport, 2011). With GWAS numerous variants have identified associated with diseases (Balshaw & Kwok, 2012), but findings cannot explain the variability of diseases by only genetic polymorphism. Essentially, in 1980s, molecular epidemiologists had discovered several biomarkers as reflection of interaction between genetic and environmental factors. Currently, some authors propose that, in chronic diseases between 70 to 90% of disease risks are due to environmental factors (Rappaport & Smith, 2010), (Swan, 2012). Because modifying genetic determinants of risk are not feasible except gene therapy, in order to target prevention efforts, it is critical to determine and assess changeable enviro-behavioral factors that interact with genetic determinants to cause disease. Therefore, to study and analysis of environmental factors, in a manner analogous to a GWAS, *Environmental Wide Association Studies (EWAS)* have started (Balshaw & Kwok, 2012).

Today, a number of groups have undertaken efforts to determine the environmental causes of diseases. These factors concerning with mechanisms of human diseases can be categorized as sociodemographic parameters (age, ethnicity, race, gender, family health history), environmental causes (tobacco smoke, pollution, hazardous chemicals, occupational agents, microbial agents, radiation, etc.), behavioral factors (diet, physical activity, use of supplements, drugs, etc.), and internal environment of individual (ageing, body morphology, metabolism, hormones, microflora, inflammation, lipid peroxidation, oxidative stress, etc.) (Rappaport & Smith, 2010), (Wild, 2012). Regarding public health and clinical medicine, these factors can be classified as risk and protective factors according to effects on disease mechanism and prognosis.

To identify lifelong environmental, behavioral, and endogenous exposure history of the human body as an important complementary of disease etiology, the term of *exposome* proposed, inspired from the term of genome. To analyze internal exposures of the body, several omic techniques and tools are also proposed used in genomics, metabonomics, lipidomics, transcriptomics and proteomics (Rappaport & Smith, 2010), (Lioy & Rappaport, 2011), (Wild, 2012).

2.2.5 Sociodemographic Data for Health Records

Age, gender, ethnicity, and race are major sociodemographic factors affecting personal health status. Age is related with almost all medical conditions and often used to categorize patients for comparative studies. In daily life, the terms sex and gender are often used interchangeably, despite they have different meanings. Sex is defined as biological characteristics based on chromosomes, physiology, etc., while gender refers to the sociocultural construction of masculinity and femininity in a society (Verdonk & Klinge,

2012). This distinction is very important in some medical conditions. For example, in prostate cancer; as the male-to-female gender reassignment surgery generally does not involve prostatectomy, a female patient by gender can have prostate cancer (Miksad, et al., 2006).

Race is a socioeconomic construct of human variability based on differences in biological characteristics, physical appearance, social structures, shared worldview, and behavior. This definition contains intertwined cultural and biological factors and sometimes used synonymously with ethnicity, ancestry, nationality, and culture. But in practical life sometimes race and ethnicity accepted as different concepts. All of these terms are valuable predictors to assess disease risk (National Research Council, 2009).

2.2.6 Family Health History

Family health history (family history, family medical history, and family medical tree) is an aggregation of information about health status affecting a person and his/her family members. The scope of family members typically involves three generations of relatives by birth, person, his/her children and his/her siblings (parents, maternal and paternal grandparents, and maternal and paternal aunts and uncles) (Alspach, 2011).

In personal disease risk assessment, family health history is accepted as the most efficient tool to solve complex interactions between genes and environmental factors for a great number of disease e.g. arthritis, asthma, cancer, diabetes mellitus, hypertension, hypercholesterolemia, single-gene disorders (Mendelian inheritance), etc. (Guttmacher, et al., 2004), (Ginsburg & Willard, 2009), (Alspach, 2011).

Today, there are web-based tools to collect and assess family health history in an easy fashion, and patient-completed tools to collect family history are also developed (Weitzel, et al., 2011). Still the EMR/EHR is used to record and store family health history data in narrative format (Hoffman & Williams, 2011). The American Health Information Community's (AHIC) Family Health History Multi-Stakeholder Workgroup proposed a structured data set for family health history within EMR/EHR (Feero, et al., 2008), (Glaser, et al., 2008), (Ginsburg & Willard, 2009).

2.2.7 Clinical Use of Genomic Data

In medical care processes, genomic data and its derivatives can be used on risk assessment, to predict disease susceptibility, targeted screening, clinical diagnosis, to predict the course of the disease, to create a treatment plan and follow-up (Ginsburg & Willard, 2009), (Chan & Ginsburg, 2011) (Figure 10).

2.2.7.1. Genome Wide Association Studies (GWAS)

In an extensive term, examining the genomic variations to detect the variances between individuals is known as GWAS. With GWAS, numerous SNPs can be recognized and examined for their associations regarding the pathogenesis of complex diseases. Today around 12 million SNPs identified and with GWAS nearly 40 multifactorial complex diseases are found to be linked with specific SNPs (Pearson & Manolio, 2008).

In the beginning of the GWAS, the dominant approach was based designed to find direct impacts of common variants on the disease mechanisms (Cirulli & Goldstein, 2010). This approach was named as common disease–common variant (CDCV) hypothesis. These hypothesis has originated great number of the very common gene variants (minor allele frequency (MAF) > 5%) with minor effect sizes (odds ratios <1.5) in the human genome (Khoury, et al., 2009), (Cirulli & Goldstein, 2010).

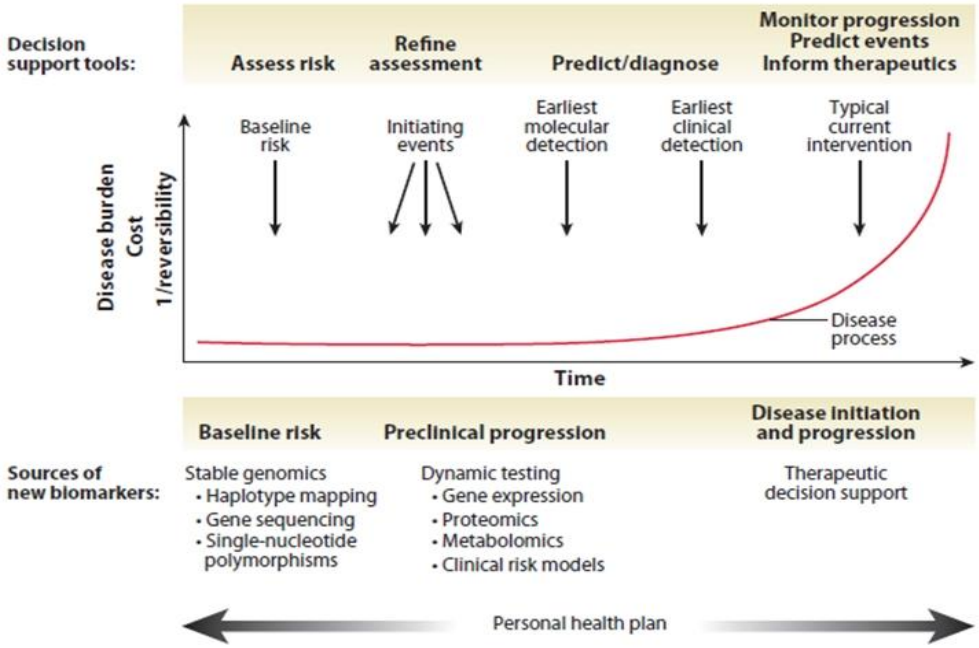


Figure 10: The role of genomics in clinical processes (Chan & Ginsburg, 2011).

CDCV approach didn't become successful to explain all genetic mechanism behind most complex multifactorial diseases, and clinicogenomic associations with strong effects didn't found. Hence, common disease-rare-variant (CDRV) hypothesis was constructed to discover possible explanations of genetic mechanisms of complex disease. Rare SNP is identified as the variation which has major allele frequency of <1% (Cordero & Ashley, 2012).

Today, GWAS sequencing approach is extending to study role of rare variations on disease mechanisms and develop a rare variant-disease catalog (Roden & Tyndale, 2013).

2.2.7.2. Predictive Models and Population Based Screening

For predictive evaluation, combination of limited number of well-known genetic variations can be used to categorize population regarding potential risk of disease. But the number of possible genotype combinations exponentially increase with the number of contained variations (Janssens & van Duijn, 2009).

A person has three different risk allele combinations for every locus i.e. homozygote healthy (zero risk allele), heterozygote risky (one risk allele), and homozygote risky (two risk alleles). For example, in a prostate cancer case with thirty-one risk loci, the number of possible risk alleles can be ranged from zero (all alleles healthy) to 62 (all alleles risky) and

there are 3^{31} distinct possible combinations of these 31 alleles (Pashayan & Pharoah, 2012). For this reason, different type of collective predictive models are developing.

Due to the complex genetic construction of many common diseases, it's hard to explain the associations and interactions between genetic and non-genetic risk factors. Thus, developing analytic models to integrate genetic and non-genetic factors for disease risk assessment is still a critical problem (Salari, et al., 2012), (Khoury, et al., 2013).

Presently, predictive risk models based on the identified common susceptible variations have small values to assess disease risk. Recently, it's proposed that these susceptible common variations can be used as screening tests for population level risk stratification (Pashayan & Pharoah, 2012).

Screening tests have high false-positive rates. Therefore, these tests are not ideal to predict given medical condition in a population and typically definitive diagnostic tests are used to precise diagnosis. The real advantage of population screening is to discover all possible cases of clinical conditions in the population (maximum sensitivity). Suspected individuals i.e. positive individuals regarding screening test, usually undergo subsequent procedures, interventions, and tests (Khoury, et al., 2013).

In genomic medicine, risk-stratified population screening can be applied as only polygenic risk profiling or combined with conventional risk factors (e.g. race, age, family history, etc.). By this approach, standard public health interventions could be applied more effective than conventional screening to each population stratum (Chowdhury, et al., 2013), (Pashayan, et al., 2013).

2.2.7.2.1. Cumulative Risk Models

Typically, most of the clinically relevant SNPs have minor effect (Odds Ratio <1.50-2.00) and there are only limited number of different examples (Stranger, et al., 2011), (Kalf, et al., 2013). Despite the small impact degree of single clinicogenomic association, the combinations of various SNP alleles may be declarative in the pathogenesis of diseases. Some investigators attempt to improve models and multi panels assigning values for various SNP alleles and estimates entire risk of disease for more effective risk prediction (Manolio, 2010).

To assess the cumulative effect of genetic variations, investigators may use several approaches e.g. Risk Allele Scores (RAS), logistic regression analyses (LR), and Cox proportional hazards regression analyses (Cox PH). RAS calculates risk scores by counting the number of risk alleles but ignores the different effects of the individual variations. This model may be rational for polygenic complex conditions which have small magnitude of impacts. In LR and Cox PH methods, risks are predicted using weighted risk scores (Janssens & van Duijn, 2009), (Salari, et al., 2012).

In the literature, several cumulative prediction models have been proposed but most of these are criticized regarding comprehensive evaluation especially for clinical utility (Table 4) (Janssens & van Duijn, 2009), (Little, et al., 2012).

2.2.7.2.2. Combination of SNP Data and Environmental Factors

As explained in the "Components of Diseases" especially common medical conditions e.g. heart disease, obesity, diabetes, schizophrenia, and many types of cancers have complex and multifactorial interactions between genomic, lifestyle and environmental components. For this reason these types of diseases don't follow the strict inheritance pattern as in monogenic diseases (Janssens & van Duijn, 2008), (National Library of Medicine (US), 2013).

Today, several researches carry out studies to develop enviro-genomic risk models. A statistical approach, and software were developed using enviro-genomic parameters and determining individual disease risk (Crouch, et al., 2013).

In a study about colorectal cancer, using this statistical risk model and software, it's found that, disease risk prediction of colorectal cancer could be possible tracking and managing enviro-genomic profile (selected SNPs, alcohol intake, smoking, exercise levels, BMI, fibre intake and consumption of red and processed meat) and prevention of disease could be accomplished changing risky lifestyle factors (Yarnall, et al., 2013).

Table 4: Methodological characteristics of recent studies on the prediction of complex diseases using multiple genes (Janssens & van Duijn, 2009)

First author, (year)	Cases	Analyses
Cauchi (2008)	Type II Diabetes Mellitus	LR
Harley (2008)	Women with Systematic Lupus	LR
Humphries (2007)	Coronary heart disease	Cox PH, weighted
Kathiresan (2008)	Myocardial infarction, ischemic stroke and	Cox PH, RAS
Lango (2008)	Type II Diabetes Mellitus	LR, RAS
Lyssenko (2005)	Type II Diabetes Mellitus	Cox PH
Lyssenko (2008)	Type II Diabetes Mellitus	LR, RAS
Maller (2006)	Advanced Age Related Macular	LR
Meigs (2008)	Type II Diabetes Mellitus	LR, RAS
Morrison (2007)	Coronary heart disease	Cox PH, RAS
Podgoreanu (2006)	Myocardial infarction	LR
Van der Net (2009)	Coronary heart disease	Cox PH, RAS
Van Hoek (2008)	Type II Diabetes Mellitus	Cox PH, LR, RAS
Vaxillaire (2008)	Type II Diabetes Mellitus	LR
Wang (2008)	Severe hypertriglyceridemia	LR
Weedon (2006)	Type II Diabetes Mellitus	LR, RAS
Weersma (2008)	Chronic inflammatory bowel disease	LR, RAS
Yeh (2007)	Colorectal cancer	LR
Zheng (2008)	Prostate cancer	LR, genotype score

2.2.7.2.3. Hybrid Model Based Risk Prediction

Among risk assessment tools besides cumulative models, there are other ongoing efforts utilizing different data mining algorithms to interpret GWAS data for building various predictive models. In studies of "Yücebaş and Aydın Son", several combined parameters were discovered through a hybrid approach combining Support Vector Machine (SVM) and ID3 decision tree based on "A Multiethnic Genome-wide Scan of Prostate Cancer" data set from dbGaP database (study accession no: phs000306 and version 2). First hybrid model (only SNP model) includes 33 SNPs and their alleles and the accuracy, precision, and recall values of this model are %71.6, %72.69 and %68.96 respectively. The second hybrid model was originally developed for African American cohort and contained 28 SNPs, Body Mass Index (BMI), alcohol and cigarette usage. The accuracy, precision, and recall values of this model for African-Americans are %93.81, %96.55 and %90.92 respectively (Yücebaş & Aydın Son, 2014).

2.2.7.4. Clinical Usefulness and ACCE Frameworks

Evaluation of Genomic Applications in Practice and Prevention (EGAPP) working group is an independent group that tests and evaluates existing models regarding validity and utility, and prepare evidence-based recommendations. EGAPP-WG formalized an ACCE (analytic validity, clinical validity, clinical utility, and ethical, legal, and social implications) framework for these purposes (Khoury, et al., 2009).

But still, the absence of central validation of genetic tests is an essential barrier to integrate genetic data with EMR/EHRs in the efficient, effective and ethical manner (Shoenbill, et al., 2013).

2.3 Integration of SNP Data into EMR/EHR

As explained above, the informatics pipeline for genome sequencing can be divided into several analytical steps e.g. base calling, alignment, variant analysis, interpretation, and in all levels different file formats are generated (Röhm & Blakeley, 2009), (Nielsen, et al., 2011), (Wright, et al., 2011). Currently, tools and techniques are developed for automated and reliable analysis, but clinical interpretation of variant data is still a major problem (Wright, et al., 2011).

After WGS tests, a file which contains a huge amount of variant data is acquired (Aronson, et al., 2012). WGS data involve about 3 billion base pairs and entire genome sequence is about 3.2 Gb. Storing and sharing of personal raw genomic sequence exceeds the transmission and storage capacity in many healthcare organizations (Kahn, 2011). Due to the technical limitations, raw genomic data is stored the outside of the EMR similar to PACS for medical images and clinical interpretation of data is preferable sent to the EMR database (Starren, et al., 2013), (Masys, et al., 2012), (Green, et al., 2013).

Today, most of the EMR/EHRs were designed to store and retrieve the transient diagnostic values, such as laboratory or clinical findings, but don't have the ability to manage genomic data (Hoffman, 2007), (Sethi & Theodos, 2009), (Jacob, et al., 2013). The initiatives of integrating a patient's genomic data into EMR/EHRs is of a preliminary nature (Jing, et al., 2012), (Ury, 2013) and until recently, only a few successful systems are established such as Cerner's Genomics Solutions, McKesson's Horizon Clinicals and GeneInsight (Aronson, et al., 2012), (Ginsburg & Willard, 2013).

In the literature, basic requirements of genomic enabled EMR/EHRs were listed as incorporating genotype data and its clinical interpretation into EMR/EHRs, developing accurate and accessible clinicogenomic interpretation resources (knowledge base), interpretation and reinterpretation of variant data, and immersion of clinicogenomic information into the medical decision processes (Figure 11) (Manolio, et al., 2013).

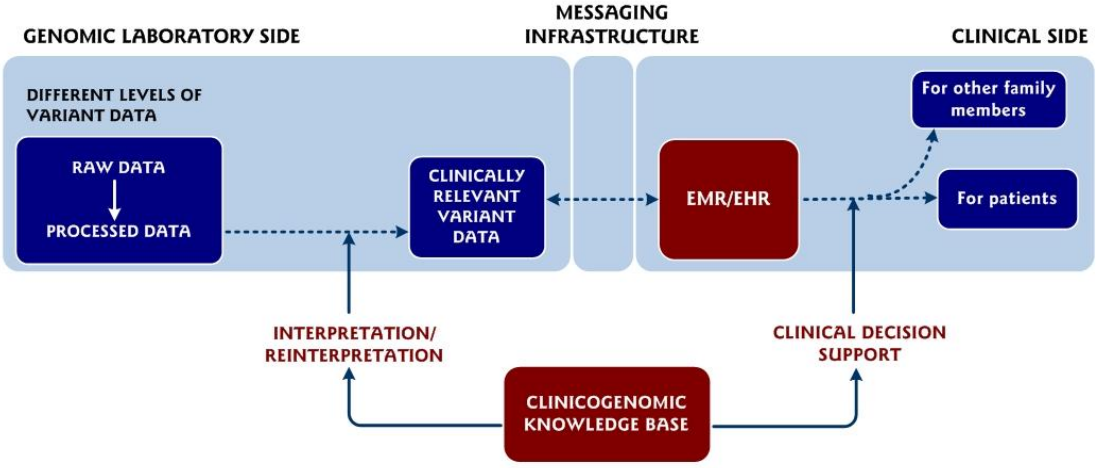


Figure 11: Main Components of a Genome Enabled EMR/EHR (depicted based on current literature). In the genome laboratory side, several levels of sequence data are produced. Since the clinicians need actionable clinical interpretation of variant data, it's sufficient to share clinically relevant data between laboratory and clinical systems. Development of clinicogenomic knowledge base is an obligation to extract clinical meaning from variant data. In clinical side, it's needed to use decision support systems due to the amount of variant. Sometimes, clinicogenomic information may be useful to manage of health status of other family members and other close relatives.

2.3.1 Standards and Messaging

In order to integrate structured genotype and phenotype data into any system, first requirement is to determine data components, terminology standards and identifiers of clinicogenomic information i.e. genotype data and its associated clinical interpretation.

2.3.1.1. SNP Identifiers

In genomic terminology, identifying gene symbols and identifiers are standardized by Human Gene Nomenclature Committee (HGNC), and variant nomenclature defined by the Human Genome Variation Society (HGVS). A candidate identifier for SNP is variant nomenclature from the Human Genome Variation Society (HGVS) i.e. "<Accession Number>.<version number> (<Gene symbol>): <sequence type>.<mutation>". Nevertheless, this one is not extensively accepted as a common standard, since it is more complicated, and rs number is widely adopted and used in the biomedical literature (Poo, et al., 2011). "Rs number, rs#" or refSNP is used to identify every single SNP entry in dbSNP which is the largest database maintained by the National Center for Biotechnology Information (NCBI), dbSNP is interconnected with many other resources, e.g. EntrezGene,

GenBank, the Universal Protein Resource (UniProt), the International HapMap Project, and PharmGKB, AlzGene, PDGene, SzGene and Japanese Single Nucleotide Polymorphism (JSNP) by that rs number (Thomas, et al., 2011). Additionally, in some types of personal genomic file formats (e.g. 23andMe, deCODEme, and Navigenics), SNPs are identified by rs number.

Because different alleles of SNPs may have different degrees and kinds of clinical impact, rs number is insufficient alone to identify the clinicogenomic significance of SNPs. For example, to have a heterozygote allele may not change the risk for the disease but homozygote allele of the same SNP variant may change the risk for a disease dramatically. Consequently, to identify clinically relevant SNP, we need to use a combination of rs number and allele data as a minimum requirement (Attia, et al., 2009A).

Due to double stranded (plus and minus or forward and reverse respectively) nature of DNA, every SNP, can be identified using either of these strands. Sometimes, in various genomic databases, same SNPs alleles are defined with different alleles based on the orientation discrepancy (Attia, et al., 2009A). For clinical researches, both identification approaches are correct but it's required to use and declare a standard.

2.3.1.2. Clinical Terminologies

Integration of variant data and clinical relevancies bring out the issue of terminological standardization. Unfortunately, conventional health information terminologies do not completely support genetic diseases accordingly. There is a critical gap between the databases which involve many terms defining genetic diseases and **Systematized Nomenclature of Medicine (SNOMED)** (Ullman-Cullere & Mathew, 2011).

In order to address the chasm between medical vocabularies and bioinformatics resources, the **Clinical Bioinformatics Ontology (CBO)** is developed and implemented. CBO is a curated semantic network trying to combine different kind of clinical vocabularies (SNOMED-CT, and LOINC), and NCBI bioinformatics resources (Hoffman, et al., 2005), (Hoffman & Williams, 2011).

In addition, the International Classification of Diseases (ICD) codes, like in Turkey, is also preferred to use for identifying clinical conditions, but released versions of ICD does not fully support genomic medicine (Ullman-Cullere & Mathew, 2011).

2.3.1.3. Interoperability Standards

HL7 is a global organization developing health information standards. HL7 Clinical Genomics (CG) Work Group developed standards intended to interoperability issues in genomic medicine (Table 5) (HL7 Clinical Genomic Work Group, 2013).

HL7 suggests the sharing of the essential part of raw genomic when it's needed via encapsulation and bubble-up extracting clinically relevant data based on genomic decision support application (Shabo, 2006).

The **HL7 Genotype model** determined a genotype related data, which is proposed as an essential unit of genomic information exchange in healthcare. This model contains a subset of the overall Clinical Genomics Domain Information Model (HL7 Clinical Genomics SIG, 2005).

Table 5: Some standards of HL7 CG Work Group (HL7 Clinical Genomic Work Group, 2013)

HL7 CGPED, R1, HL7 Version 3 Pedigree Topic - Family History, Last Ballot: Normative Ballot 1 - May, 2007
HL7 IG CG_GENO, R1, HL7 Version 3 Genotype, Release 1, Last Ballot: DSTU Ballot 1 - January 2009
HL7 CG_GV, R1, HL7 Version 3 Standard: Genetic Variation, Release 1, Last Ballot: Normative Ballot 2 - January 2009
HL7 IG LOINCGENVA, R1, HL7 Version 2 Implementation Guide: Clinical Genomics; Fully LOINC-Qualified Genetic Variation Model, Release 1 (US Realm)

The US Department of Health and Human Services, Office of the National Coordinator for Health IT published the **Personalized Healthcare Detailed Use Case**. This use case concentrated to provide a secure transmission of genetic laboratory results, in addition to family history and associated risk assessments. (Office of the National Coordinator for HIT, 2008).

The HL7 Version 3 Domain Information Model, Clinical Sequencing, Release 1 detailed the Personalized Healthcare Use Case. This model produce a variety of additional use case scenarios for clinical genomic, e.g. testing of an individual's hereditary or germline genome, cancer genomics/tumor profiling, early childhood developmental interruption, neonatal testing, and newborn screening (Ullman-Cullere & Mathew, 2011).

HL7 Version 3 Genetic Variation Model specifies the syntaxes and semantics of genetic test transmission. This model is additionally restricted to genetic variation analyses methods.

Also, HL7 organization has published an implementation guide (HL7 Version 2 Implementation Guide: Clinical Genomics; Fully LOINC-Qualified Genetic Variation Model) based on both the HL7 Version 2 Implementation Guide Laboratory Result Reporting to the EHR, and the HL7 Version 3 Genetic Variation data model. This guide covers the reporting of test results for sequencing and genotyping tests and includes testing for DNA variants associated with diseases and pharmacogenomic applications, (Ullman-Cullere & Mathew, 2011).

HL7 Version 2 Implementation Guide: Clinical Genomics; Fully LOINC-Qualified Genetic Variation Model was the first example used by The Partners HealthCare Center for Personalized Genetic Medicine (PCPGM) and the Intermountain Healthcare Clinical Genetics Institute to gather genetic test results and transmit them to a patient's EHR (Shabo, et al., 2009), (Ribick, 2010). GeneInsight Suite (GeneInsight Lab, GeneInsight Clinic and GeneInsight Network) is a platform where clinical variant data sharing was based on HL7 standards (Aronson, et al., 2011), (Aronson, et al., 2012), (National Research Council, 2012), (Masys, et al., 2012).

2.3.2 Clinicogenomic Knowledge Bases

Clinicians can't extract clinical interpretation of variants directly from the medical sources due to temporal and cognitive limitations (Oetting, 2009), (Starren, et al., 2013). So, instead of incorporating all sequence data into medical records, integration of the clinical interpretations of variant data will be more efficient for clinical decision making (Marian,

2011), (National Research Council, 2012). To gain this capability clinically relevant variants must be selected and presented with their clinical meaning, i.e. clinicogenomic associations, and an action plan for clinicians. Since HGP, researches have been discovering new clinicogenomic associations increasingly, it is critical to reinterpret variants and integrate new clinical interpretations into clinical processes (Aronson, et al., 2012).

2.3.2.1. Sources for Clinicogenomic Associations

Clinicogenomic associations which are acquired via researches based on the candidate gene investigation or agnostic screening of complete genome, are published in the scientific literature (Attia, et al., 2009B). Some clinicogenomic knowledge bases collect, curate, interpret and categorize these published associations between genomic variations and clinical conditions. Cancer Genome-wide Association and Meta Analyses Database (Cancer GAMAdb) is a part of Cancer Genomic Evidence-based Medicine Knowledge Base (Cancer GEM KB) and provides GWAS researches and meta-analysis about clinicogenomic associations (http://www.hugenavigator.net/CancerGEMKB/caIntegratorStartPage.do) (Schully, et al., 2011). ClinVar (http://www.ncbi.nlm.nih.gov/clinvar) provides reports for variations and related phenotypes with evidences. AlzGene (http://www.alzgene.org), PDGene (http://www.pdgene.org), and SzGene (http://www.szgene.org) are resources which contain manually curated PubMed articles using systematic methods for Alzheimer disease, Parkinson's disease, and schizophrenia, respectively. SNPedia (http://www.SNPedia.com) is a wiki resource of human genetic variation as published in peer-reviewed studies (Cariaso & Lennon, 2012). PharmGKB (http://www.pharmgkb.org) is a knowledge source containing clinically relevant genotype-phenotype and gene-drug relationships.

However many of existing knowledge bases for the clinical interpretation of variant data have different conventions. Also, they are not error proof and are not sustainable due to funding issues (National Research Council, 2012). Especially for polygenic complex diseases, impact degrees of clinicogenomic association may be different according to race, ethnicity and environmental factors (Stepanov, 2010). Therefore, in personalized risk assessment, it will be an ideal approach to use population specific clinicogenomic results or at least findings from similar communities. If these not possible, it might be conceivable to use other scientific resources with a confidence range. Eventually, experts have been advocating for generating centrally-curated national repositories of clinically significant variants for the interpretation of individual's genomic information (Kawamoto, et al., 2009), (Starren, et al., 2013). To develop a national level clinicogenomic knowledge base is critical to consider consistency of clinicogenomic associations with the sociodemographic characteristics of citizens and overcome the issues about sustainability.

2.3.2.2. Magnitude of Impact and Quality of Evidence

Regarding published results of clinicogenomic associations, two major points are critical i.e. evidence quality of study and effect size of these associations (Attia, et al., 2009B), (Van Allen, et al., 2013).

In clinical practice, absolute risk value of genetic variations are important, but in most of the disease-variation researches, absolute risk cannot be calculated due to lack of information about disease incidence (Janssens & van Duijn, 2009). To measure magnitude of impact for clinicogenomic associations, researchers usually prefer to use conventional approaches, e.g. odds ratios and relative risks for case control studies and cohort studies respectively. These values are presented with confidence interval (Attia, et al., 2009C). In GWAS, several

defects and biases due to the study design, genotyping or collected data quality, will affect the clinical value of results (Pearson & Manolio, 2008), (Attia, et al., 2009B), (Little, et al., 2009). The quality of evidence is scored based on the type of study and how well the study is conducted (Riegelman, 2010) and some guidelines are proposed to calculate evidence degree (Ioannidis, et al., 2008).

For the significance of clinicogenomic association, some of the knowledge sources contain additional data fields that define the magnitude of clinical effects and strength of the relationship between variants and diseases. In ClinVar, clinical significance is defined as a combination of impact and clinical function (e.g. benign, pathogenic, protective, drug response, etc.), and evidence for clinical significance is categorized regarding study count and type such as in vitro studies, animal models, etc. (http://www.ncbi.nlm.nih.gov/clinvar/intro) The PharmGKB uses a systematic categorization for quality of evidence depending on several parameters about methods and results of references (http://pharmgkb.org/page/clinAnnLevels), but impact value is not emphasized as a different criteria. In SNPedia, magnitude is constructed as a subjective measure of interest for magnitude of impact and repute (good, bad) for quality of evidence, but these concepts are not well established. In GET-Evidence (http://evidence.personalgenomes.org/about) clinicogenomic references are categorized according to their evidence degree (high, moderate, or low) and clinical significance (high, medium, or low) are used to produce impact score (Ball, et al., 2012).

2.3.2.3. A Brilliant Example: Pharmacogenomic Knowledgebase (PharmGKB)

The PharmGKB (http://www.pharmgkb.org/) is a pharmacogenomics knowledge resource that target to store and improve clinicogenomic information about drug effects and interactions e.g. drug dose guidelines and drug labels, potentially clinically relevant gene-drug relationships and genotype-phenotype associations (Thorn, et al., 2010).

PharmGKB collects, stores, curates and shares this sort of knowledge. The PharmGKB Knowledge Pyramid is in Figure 12. In the development of PharmGKB content, first scientific pharmacogenomic sources are manually collected by the domain experts. To capture relevant information faster and in an effective manner, natural language processing techniques are utilized. In extraction stage, curators find drug-variation associations, improve drug specific pharmacokinetic and pharmacodynamic pathways and extract critical gene-drug interactions summaries.

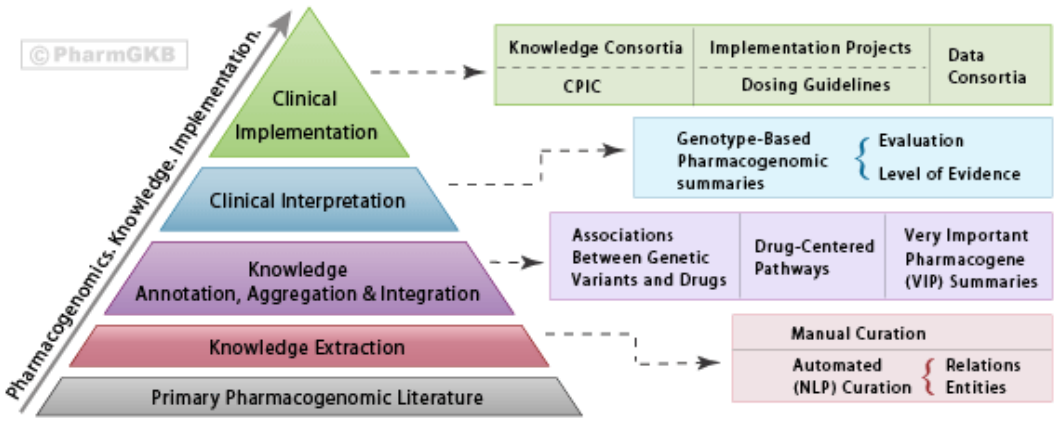


Figure 12: The PharmGKB Knowledge Pyramid (Whirl-Carrillo, et al., 2012).

In clinical interpretation step, curators aggregate variant annotations regarding a specific genetic variant-drug association, and write standardized clinical annotations. These clinical annotations are given a level of evidence depending pre-defined criteria, including study size and statistical relevance of the association.

And finally, in clinical implementation stage, clinically relevant information (drug labels with pharmacogenomics information, genetic tests for pharmacogenomics, drug dosage guidelines) are prepared and published (Whirl-Carrillo, et al., 2012).

2.3.3 Clinicogenomic Decision Support

Conventional medical paradigm compelled clinicians to rely on the indication lists for every advanced examination. But NGS genotyping technologies is a candidate to transform this approach as "firstly examine, than reinterpret repetitively" due to changing nature of genomic information. Individual genetic structure is mostly stable in all lifetime, but its clinical interpretation and effect on medical processes will be changing in time. The basic reason of this effect of genomic data depends on two essential causes i.e. changed scientific information about the role of genetic data, and newly discovered disease-variation associations. It's impossible to track, learn and apply all of these dynamical movement of information for a healthcare professional. (Starren, et al., 2012).

Also, size of genomic data files are huge. An individual can have more than 3,000,000 genetic variants. Globally, in omics domain, produced data exceeds any individual's mental competence (Starren, et al., 2012).

For example, if SNP data i.e. the simplest type of variation can be presented as numerous variation displayed in laboratory report, clinicians cannot be interpret or evaluate these information stack. The volume of variation data integrated into clinical practice exceeds the boundaries of unsupported human cognition and interpretive capacity. Additionally the rapidly growing literature about clinicogenomic associations make it more complicated to stay current for even professionals (Masys, et al., 2012).

Also, it's not reasonable to expect the interpretation of all clinicogenomic data by limited number of genetics experts, and we need more automated solutions to overcome these obstacles (Welch & Kawamoto, 2013). With the growing data load in the genomic era, in order to make informed decisions in a timely manner, the healthcare systems need to shift from expert-based practice to systems-supported practice (Figure 13) (National Research Council, 2008).

To provide the clinical decision support for genomic variations, it is required to integrate genomic data and patient's EMR/EHR, construct an infrastructure allowing interaction with the data over the lifetime of the patient and clinical decision support system (Ginsburg & Willard, 2013).

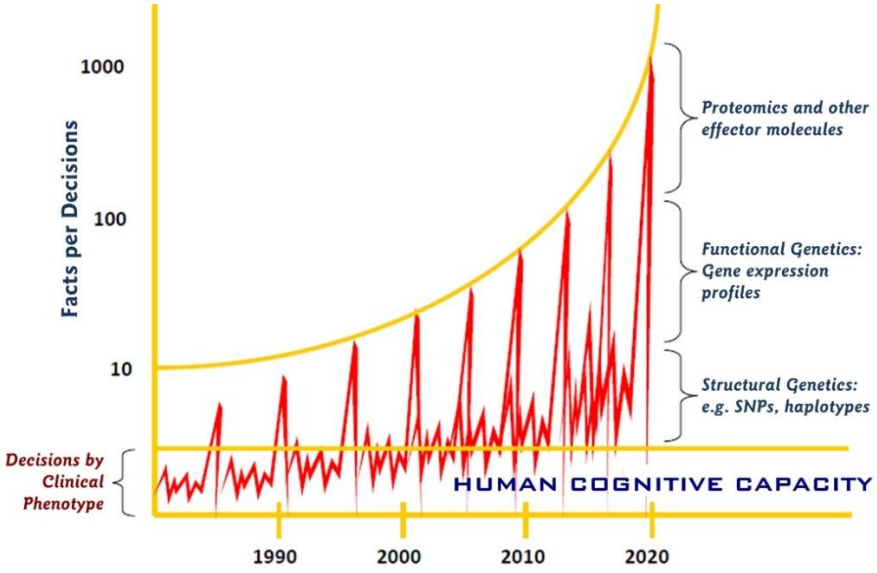


Figure 13: Omic Data and Clinical Decision Making (National Research Council, 2008).

Efficiency of clinical decision support systems are mostly determined by knowledge base and rule engine. Therefore, both of these components must be updated regarding changed scientific literature. To produce a more flexible clinical decision support system, knowledge bases and rule engines are should not be embedded into EMR/EHRs, but developed as externally integrated components (Kannry & Williams, 2013).

Eventually, it's critical to develop a national clinical decision support system infrastructure that allows centrally-curated, accredited, and authoritative clinicogenomic knowledge to the clinical practices through the nation (Kawamoto, et al., 2009).

2.3.4 Using Genomic Information for Consanguineous

Genomic information has lifelong value and one's genomic findings can reveal others' within families (Hoffman, 2007). If a patient is found to have a disease associated variant, possibly other blood relatives would carry the same risk and this new clinical information could be utilized by the patient's health-care provider (Aronson, et al., 2012). This is especially important, not only because of the medical perspective but also for security and privacy issues.

Family history is an important tool for personalized medicine. But structured family history is not a mandatory part of EMR/EHRs, and because of its dynamical characteristics, it's reasonable to collect and confirmed by the patient at each visit. It's clear that, similar to clinicogenomic associations, collection and reinterpretation of family history is critical to capture effective results from these types of predictive models.

2.3.5 Examples of Genomic Data Integrated EMR/EHR

2.3.5.1. GeneInsight Suite

The first successful electronic transmission of genetic test results were realized between Partners HealthCare Center for Personalized Genetic Medicine (PCPGM) and Intermountain

Healthcare in Salt Lake City, Utah in 2009 (Shabo, et al., 2009), (Ribick, 2010), (Aronson, et al., 2011).

PCPGM developed a platform where clinical variant data sharing was based on HL7 standards i.e. GeneInsight Suite and using these systems genetic results are shared by stakeholders electronically (Aronson, et al., 2011), (Aronson, et al., 2012), (National Research Council, 2012), (Masys, et al., 2012).

The workflow of GeneInsight is as presented in Figure 14. The innovation of the PCPGM/Intermountain transmission was made possible by PCPGM's development of medical laboratory report message gateway (VariantWire). VariantWire is a clinicogenomic networking system based on InterSystems Ensemble (Aronson, et al., 2011).

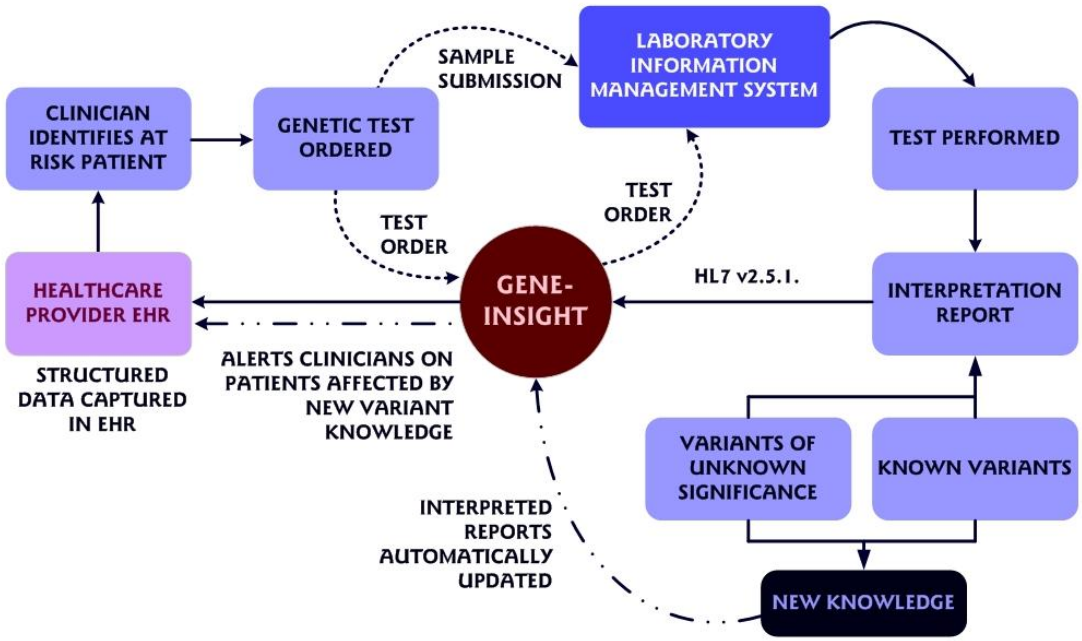


Figure 14: Workflow of GeneInsight.

The PCPGM accomplished a knowledge base containing more than 10,000 unique variants and actively manage and update in time. This knowledge base involves several different type of variations e.g. clinically significant (%20), unknown significant (%10) and likely benign (%30), and unclassified (%40) (Aronson, et al., 2012).

If a variation is changed in the mutation database, patient laboratory test results are automatically updated. Clinicians can see the change of previous category and also receives the e-mail alerts about relevant clinical information (National Research Council, 2012).

2.3.5.2. Cerner's Genomics Solutions

Cerner's Genomics Solutions module allows to data capture from the clinical molecular diagnostic laboratory and incorporate into the EMR. It accommodates mutation and DNA methylation results, numeric results associated with short tandem repeat markers, and viral load data (Gerhard, et al., 2013).

A structured molecular vocabulary (CBO, Clinical Bioinformatics Ontology) is available for simplifying database development and offer reliable reporting of results (Hoffman & Williams, 2011).

2.3.5.3. Genomic Data Integrated CCR based EHR

In this project, integration of genomic variations and Continuity of Care Record (CCR) based EHR was studied and a prototype developed. This system based on an external knowledge base named as OntoKBCF which had been developed to represent clinicogenomic information. OntoKBCF was used to transform individual EHR data to clinical conclusions for clinical decision support. This research evaluated in cystic fibrosis based on simulated patient data (Jing, et al., 2012).

CHAPTER 3

GENERAL METHODOLOGY AND ENABLING TECHNOLOGIES

To develop the SNP data incorporated NHIS-T, we need to produce a general NHIS-T architecture by extending existing infrastructure and developing additional complementary capabilities. Therefore, the current characteristics and capabilities of NHIS-T were studied in the first step and the NHIS-T architecture is criticized regarding main components of genome enabled EHR.

Based on the requirement analysis and results of NHIS-T analysis, we focused how to develop a clinicogenomic knowledge base and a clinicogenomic web application. In this phase we studied some scientific sources to determine and design the structure of clinicogenomic associations and their assessment methods. Then, we have constituted standardized definition tables for clinicogenomic associations and predictive models to design ClinGenKB and ClinGenWeb.

Next, to evaluate our complementary capabilities as a whole, we selected prostate cancer as an ideal clinical condition. In parallel, we prepared the content for knowledge base and assessment and reporting approaches for decision support application. In this phase, we studied medical literature and knowledge sources to extract clinicogenomic associations between SNP alleles and increased prostate cancer risk. Additionally, we searched predictive genomic models assessing individual prostate cancer risk.

In parallel, to evaluate our system with real data, we gathered personal SNP data (24andMe files) of individuals who have been diagnosed with prostate cancer and age matched controls. In the evaluation phase, using these data files, we inferred personal clinicogenomic associations based on ClinGenKB. Finally, we evaluated prostate cancer risk assessment approaches using real personal clinicogenomic data and external data e.g. BMI.

Detailed organization of the whole study is summarized in Figure 15.

3.1 SNP Data Incorporated NHIS-T Architecture

The current characteristics and capabilities of NHIS-T are studied initially, and the NHIS-T architecture is evaluated regarding main components of a genome enabled EHR.

In terms of interoperability, the existing data elements, MHDS, messaging schemas, and terminology standards of NHIS-T were investigated through its official documentation (http://www.e-saglik.gov.tr/SaglikNet/SaglikNetDokumanlari.aspx).

Also, different types of personal genomic file formats, data types and interoperability standards for genomic data were reviewed from the literature. Overall requirements to extent the NHIS-T infrastructure integrating SNP data and associated clinical information were determined and proposed architectures were criticized regarding interoperability issues.

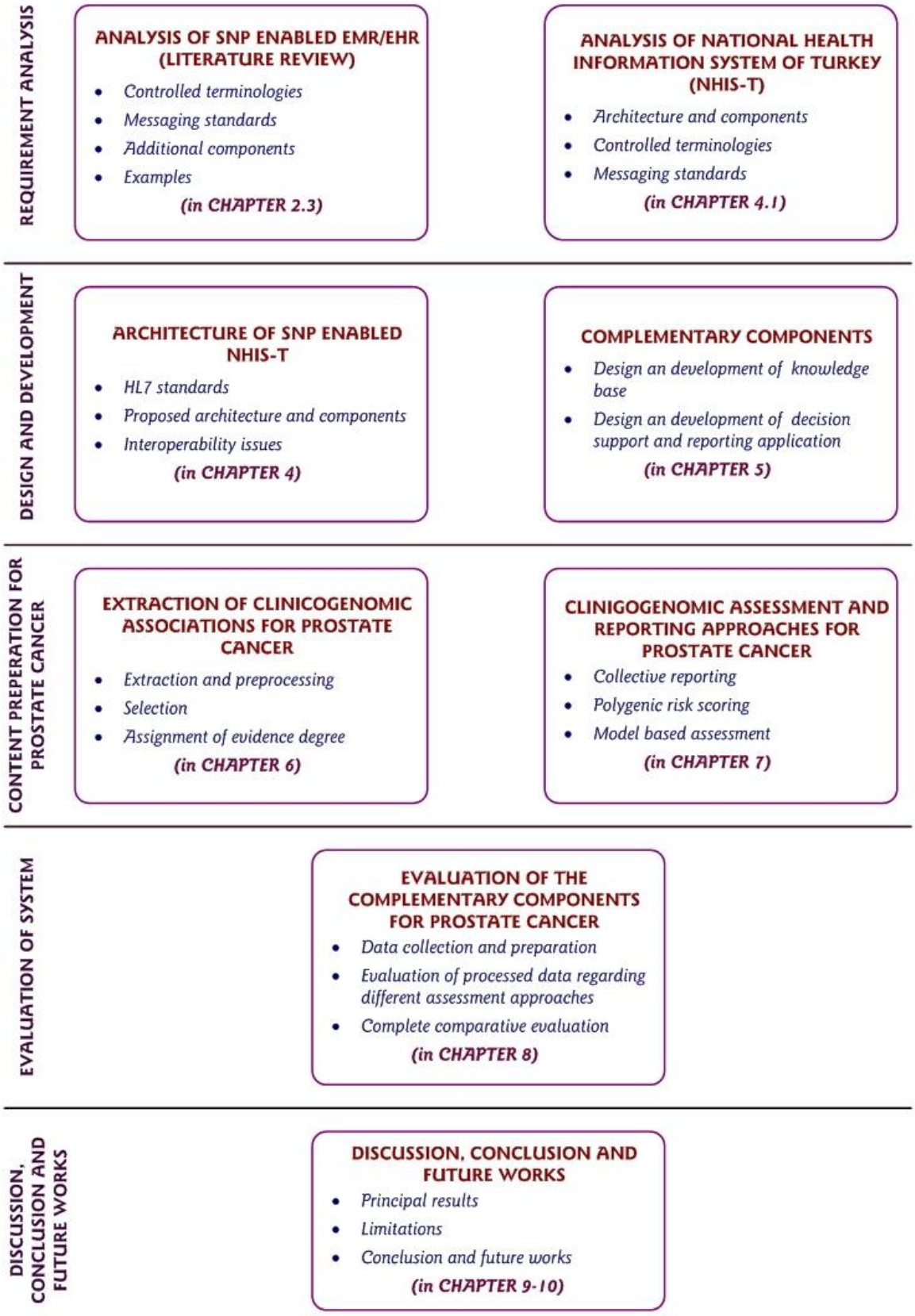


Figure 15: Detailed organization of the study.

3.2 Design and Development of Complementary Components

In the phase of literature survey, we determined the need for two complementary components i.e. knowledge base and decision support application, which are capabilities that supports each other. Therefore, the standardized design elements and models are designed and then developed as the components of the proposed system.

3.2.1 Design of Standardized Association and Model Definitions

To design and develop complementary capabilities (i.e. knowledge base and decision support and reporting applications), we needed to extract the structure of clinicogenomic associations (namely rules).

We analyzed SNPedia and studied scientific literature in a detailed manner and determined the data fields and structure of associations for knowledge base and decision support applications. At the same time, in the scientific literature, we discovered several types of assessment and reporting approaches i.e. collective assessment and reporting of independent associations, polygenic scoring based on all independent associations and model based assessment and reporting methods.

The aim of knowledge base is to collect whole information (independent, complete and model based) about clinicogenomic associations. First, we have analyzed all types of clinicogenomic associations regarding required data fields. Later, we have extracted associations (independent associations and model based associations), that will allow us to define a standard definition. Then, the ClinGenKB is designed using the proposed standard definition and its data parameters.

Additionally, because final interpretation of clinicogenomic associations will be completed at the end user side (ClinGenWeb) using predictive models, we generated a standardized model definition table involving reference values for variants and their corresponding disease risk.

3.2.2 Development of Knowledge Base

Knowledge bases are repositories which help to collect, organize, share, search and utilize information. Developing an accurate and accessible, structured clinicogenomic knowledge source (ClinGenKB) is an essential component of proposed clinicogenomic information integrated EHR. Raw genomic variant data is not appropriate to support clinicians' decision due to its high-dimension. The clinical association of the variant is convenient for clinical decision support where the interpretation of the variant and its associated clinical meaning is periodically updated in the knowledge base. Such systems will allow reinterpretation of variant data throughout dynamic updates.

3.2.2.1. Knowledge Representation and Management Approaches

Clinicogenomic associations, discovered via researches are published in the scientific literature. Some clinicogenomic knowledge bases collect, curate, interpret and categorize these clinically relevant genotype-phenotype associations such as Cancer Genome-wide

Association and Meta Analyses Database (Cancer GAMAdb), AlzGene, PDGene, SzGen, SNPedia and PharmGKB etc.

Despite the fact that those beneficial knowledge bases are publicly available, utilizing them in daily practice is still uncommon for clinicians and scientists. One of the reasons is the diversity of each knowledge base and the information contained within. Knowledge bases are widely distributed, maintained by various institutions or projects and represented differently to serve local needs. Retrieving and integrating information from those sources is time consuming in daily practice (Beyan, et al., 2013).

Semantic web technologies can be applied for identifying genomic risk factors across heterogeneous multiple knowledge bases. A primary aspect of semantic web technologies such as Resource Description Framework (RDF), ontologies, federated queries etc. can be useful to ensure a mechanism for defining and linking heterogeneous data using web protocols and a flexible data model. RDF was introduced since 1998 and now has become a standard for exchanging data in the web. At present, huge amount of data has been converted to RDF and published to general public. Apart from that, those datasets are interlinked each other and formed Linked Open Data. Such condition attracts researchers to develop applications that merge data from multiple medical data sources. SPARQL endpoint is an interface to execute SPARQL query which is a standard language to retrieve RDF data like SQL in the relational databases. Federated SPARQL query is a systems that consists of a federated engine as a mediator and a group of SPARQL endpoints. The federated engine plays a critical role to receive a query from the client and distribute the query to the relevant SPARQL endpoints (Rakhmawati, et al., 2013).

But, when we studied relevant clinicogenomic knowledge sources we faced with various problems areas to use these publicly available databases via semantic technologies e.g. terminology standards, data representation standard, availability and usability of information etc.

Firstly, data types of the included knowledge sources, mostly don't match with each other. For example, in both CancerGAMAdb and SNPedia for allele data which is needed for clinical interpretation, different standards are used. In SNPedia, a standardization based on plus strand is used for allele identification. But allele data of CancerGAMAdb is not consistent and standardized. Also, disease definitions of studied databases are not based on a standard classification or code system. Data type conflictions of databases will be complicate to produce complex queries and inferences. Finally, impact and evidence degrees of associations are not standardized and all are (if exist) based on different approaches.

These knowledge sources ensure clinicogenomic associations as web page or .xls and does not support a standard RDF schema. If we will use these type of data with semantic technologies, we would need to convert these data to RDF.

Eventually, not every clinicogenomic association is applicable for other humans. Especially for polygenic complex diseases, impact degrees of clinicogenomic associations may be very different between different races, based on ethnicity and environmental factors (Stepanov, 2010). We either need population specific researches to define these clinicogenomic association or adjust them at individual level. Additionally, experts have been advocating for generating centrally-curated, national repositories of clinically significant variants for the interpretation of individual's genomic information (Kawamoto, et al., 2009), (Starren, et al., 2013).

Therefore, we decided to develop our knowledge base to store, update and use clinicogenomic associations to support our approach. Actually, clinicogenomic associations are typical examples of logical knowledge. On the contrary of the factual (facts, or data) and procedural knowledge (knowledge about how to perform some task), logical knowledge is the knowledge of relationships between entities. There are various approaches to model and implement these types of knowledge. A logical relationship can be coded as an "if-then" rule inserting into procedural code. But the semantics of procedural code are designed to express a sequence of operations and once encoded procedurally, logical knowledge is no longer easily accessible.

If the relationships can be represented in a tabular form, it is possible to insert these relationships into a database. However, database approach is restricted for exact and clear logical relationships.

Sometimes a mixed approaches can be useful i.e. of both the database and procedural methods. Logical relationships stored in database tables, and the relationships are coded as procedural if-then statements. This can simplify the coding task, but it makes maintenance harder.

The meaning, or semantics, of the logical knowledge is best captured in a pattern-matching sense and the rule engines are designed to use pattern-matching search to find and apply the right logical knowledge at the right time. Today, there are many choices for these types of tools, and they are mostly vendor specific (Merritt, 2004).

3.2.2.2. Implementation of Knowledge Base

For this study, we have preferred to develop our prototype using BioXM™ Knowledge Management Environment (BioXM™) which is a distributed software platform providing a central inventory of information and knowledge (http://www.biomax.com/home/home.php). With BioXM™, we easily generate, manage and visualize scientific models as an extendible network of interrelated concepts.

To build a knowledge base with BioXM, we designed the domain-specific data model with semantic objects (elements, annotations, ontologies and databanks) and the connections (relations) using BioXM™ graph viewer based on our clinicogenomic association definitions.

Next, we defined importing scripts to transfer extracted independent and model based clinicogenomic associations and personal CR-SNP data to knowledge base. BioXM™ supports the data import and export as XML, HTML, excel or plain text format.

Finally, we prepared views, queries and smart folders to manage our data model and inferring processes.

3.2.3 Development of Decision Support and Reporting Application

After the transferring the personal clinicogenomic associations data file to the end users' (specialists, family practitioners and patients) application, another critical issue namely the final interpretation and reporting is emerged. Reporting presents itself here as a critical point for maximizing the effectiveness of the overall system in translating clinicogenomic data into clinic. High-dimensional variant data and its clinical associations along with its

interpretation have to be reported and visualized in a simplistic and holistic manner both for healthcare professionals and patients.

Regarding clinicogenomic decision support, our approach aims to divide two phases of clinicogenomic interpretation i.e. conversion of variant SNP into clinicogenomic association and clinical interpretation of these associations. Final interpretation is completed on the client side. This approach ensures us an opportunity to add external parameters which will be monitored or collected by end users. For example, in some cumulative prostate models, positive family history augments the total risk value in addition to clinically relevant SNPs. Family history is not a constant parameter and may be changed in time. Effective tracking of changes in family history ideally is accomplished by individuals. Similarly, clinical, environmental, behavioral or sociodemographic factors can be involved to assess the total risk with variant data in end user level.

Accordingly, we developed the practical reporting approaches and a simple prototype system using Zoho Reports™ on the client side. Zoho Reports™ (https://reports.zoho.com) is an on-demand reporting and business intelligence tool which supports several report generation capabilities e.g. chart/graph, tabular views, summary views, pivot tables, dashboards and SQL driven querying. Most importantly, it's possible to embed generated reports within external web sites and web applications.

The decision support application developed in this thesis (specifically named as ClinGenWeb for prostate cancer) is a web application processing genomic associations, clinical and environmental risk parameters. In this application, it's possible to report relevant clinicogenomic SNPs or to assess independent risk based on some models with the combination of conventional health data and clinicogenomic associations.

In ClinGenWeb, personal predictive risk can be analyzed in three main category i.e. detailed reporting of independent associations, the complete assessment of total clinically relevant SNPs (polygenic scoring), and model based interpretation of clinicogenomic associations. Some types of models are based on assessing only relevant SNPs. But a few models involve external data (family history, BMI, etc.). If collected, corresponding risk factors for prostate cancer can be used to calculate the model based risk. Also, external personal data about clinical and some environmental risk factors for prostate cancer can be reported.

3.3 Preparation of Systems for Evaluation Phase

To evaluate our complementary components, we need some contents and individual data including SNP variant data for a specific medical conditions. Because of its complicated nature and burden on public health, we preferred to choose prostate cancer to evaluate our complementary capabilities i.e. knowledge base (i.e. ClinGenKB) and decision support application (i.e. ClinGenWeb) (Figure 16).

Prostate cancer, which is the most common malignancy affecting men in Western countries is highly heterogeneous and multifactorial polygenic disease. This heterogeneous characteristics of prostate cancer could be a partially explained by genetic factors (Boyd, et al., 2012). In additional to genetic factors, age, race, family history, endogenous hormones, diseases, some environmental exposures and several behavioral features are proposed in literature as confounders of prostate cancer (National Cancer Institute, 2013B), (Sartor, 2013). This characteristics make prostate cancer as an ideal case for researching benefits of incorporations individual SNP data into EHR regarding personalized medicine.

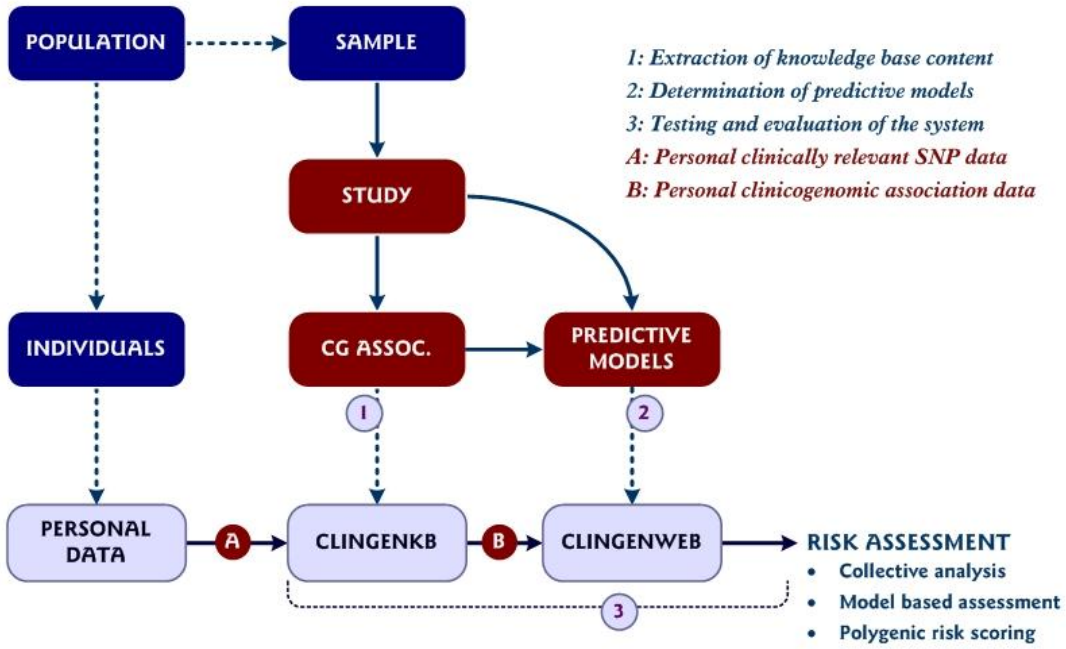


Figure 16: Main components of the evaluation preparations. 1: Extraction of SNP-prostate cancer risk associations (knowledge base content) from sources, 2: Determination of predictive models, 3: Testing and evaluation of the system. A: Personal clinically relevant SNP data, B: Personal clinicogenomic associations data.

3.3.1 Extraction of Clinicogenomic Associations

Knowledge base component includes clinicogenomic associations, namely associations between a specific clinical conditions and a specific genomic variant. To extract SNP-prostate cancer risk associations as the content of our knowledge base, we developed a layered approach i.e. extraction, preprocessing, selection and assignment of evidence degree. The details of our methods and results are presented in Chapter 6 as a whole. After that, we transferred these extracted associations and converted to suitable structure for developed knowledge base.

3.3.2 Assessment and Reporting Approaches

The second component of complementary capabilities of SNP integrated NHIS-T is a decision support application. In our study, input data (genomic data e.g. SNP file) is firstly processed using knowledge base and then assessed by decision support applications.

Therefore, we need to analyze assessment and reporting approaches from scientific literature and add these approaches to our application. In literature, there are various types of risk assessment and reporting approaches. Detailed analysis of these methods are explained in "Ch.7: Assessment and Reporting Approaches".

After we analyzed and determined the meaningful and useful assessment and reporting methods from literature, we constructed our decision support application to exploit these approaches and evaluate case and control data.

3.4 Evaluation of the System

To evaluate our complementary capabilities (i.e. ClinGenKB and ClinGenWeb), we gathered real data (24andMe files) from personal genome project (https://my.personalgenomes.org/public_genetic_data).

Then we prepared our data to process in knowledge base. After processing of these data, we acquired personal CR-SNP data file. After that, we transferred these data into clinical decision support application and we assessed and reported using various methods gathered from scientific literature. Finally, the results are studied and compared regarding clinical usefulness.

Various tests are used to calculate the performance of diagnostic and screening tests e.g. sensitivity, specificity, PPV, NPV, LR+, LR-, accuracy, AUC etc. We can use these metrics to determine and compare the value of our models (Fardy, 2009) (Okeh & Ogbonna, 2013).

The sensitivity and specificity of a model may be useful to explain how well the test was carried on, but they ensure limited information on the impact of a positive or negative test for a person.

Theoretically, the best test for both screening and diagnosis is the one with the highest sensitivity and specificity. However, these types of tests are often complex, expensive, invasive and impractical for screening population. Therefore, for high-risk population screening test, sensitivity is preferred to evaluate performance, while specificity is preferred for low-risk population screening tests.

High PPV makes the model disease quite likely in a subject with a positive test. A test with a high negative predictive value makes the disease quite unlikely in a subject with a negative test.

LR+ and LR- may be good indicators to determine a disease risk. If LR+ is more than 10, the presence of risk has critical for disease risk. And also, if LR- is less than 0.1, the absence of disease risk is prominent.

The area under the curve (AUC) may be another indicator to determine disease risk. For diagnostic tests, the relationship between the AUC and diagnostic accuracy can be classified as excellent (0.9-1.0), very good (0.8-0.9), good (0.7-0.8), sufficient (0.6-0.7), bad (0.5-0.6) and not useful (< 0.5).

To measure accuracy of a test, the overall accuracy can be used but this value is highly dependent on the prevalence of the disease. Another option is the diagnostic odds ratio (DOR) which has relationships with likelihood ratios (DOR=LR+/LR-).

CHAPTER 4

SNP DATA INCORPORATED NHIS-T ARCHITECTURE

4.1 Existing NHIS-T

4.1.1 Data Elements and Data Sets

Data elements like name, address, marital status, main diagnosis, treatment method, diastolic blood pressure, healthcare institution, etc. used in the NHIS-T are defined, and then **Minimum Health Data Sets (MHDS)** are generated combining relevant data elements. Both the data elements and MHDS are published as a **National Health Data Dictionary (NHDD**). The last version of NHDD, which includes 418 pieces of data elements, and 64 pieces of data set, is version 2.1 and accessible from its official web site (Republic of Turkey Ministry of Health, 2013).

It is mandatory for healthcare providers of Turkey to conform the NHDD data definitions and MHDS. New MHDS are produced by existing data elements or the NHDD is improved by identifying new data elements when required.

4.1.2 Codes and Identifiers

The data elements are coded using medical terminology systems which are accessible from the **Health Coding Reference Server (HCRS)** or locally defined categorical values, such as gender or marital status.

There are 342 code systems in HCRS and the current version of the HCRS is 3.0, which is available online via web services. A tabular representation is also accessible in official web page (Republic of Turkey Ministry of Health, 2012) and allows users to query through web browsers.

The healthcare professional identities are stored in central Doctor Data Bank and citizen identification is stored in Central Civil Registration System (CCRS). Both identities are validated against their original sources at storing in central repositories (Dogac, et al., 2011).

4.1.3 Messaging of Data Sets

4.1.3.1. HL7 Standards

HL7 is one of the American National Standards Institute (ANSI) accredited Standards Developing Organizations (SDOs) operating in the healthcare arena. The HL7 version 2.x (HL7v2.x) is the most widely used in the world and the HL7 version 3 (HL7v3), the latest versions.

HL7 v2.x has not a clear information model and involves many optional data fields. To overcome this vagueness problem, HL7 v3 is developed, which is based on an object oriented data model, called Reference Information Model (RIM). HL7 v3 RIM keep information as has classes (e.g. people, places, roles, things, events) and relationships. Every class has a name, description, sets of attributes, relationships, and states. The core RIM classes are Act, Entity, Role, Participation, ActRelationship, and RoleLink. (Figure 17) (Benson, 2010).

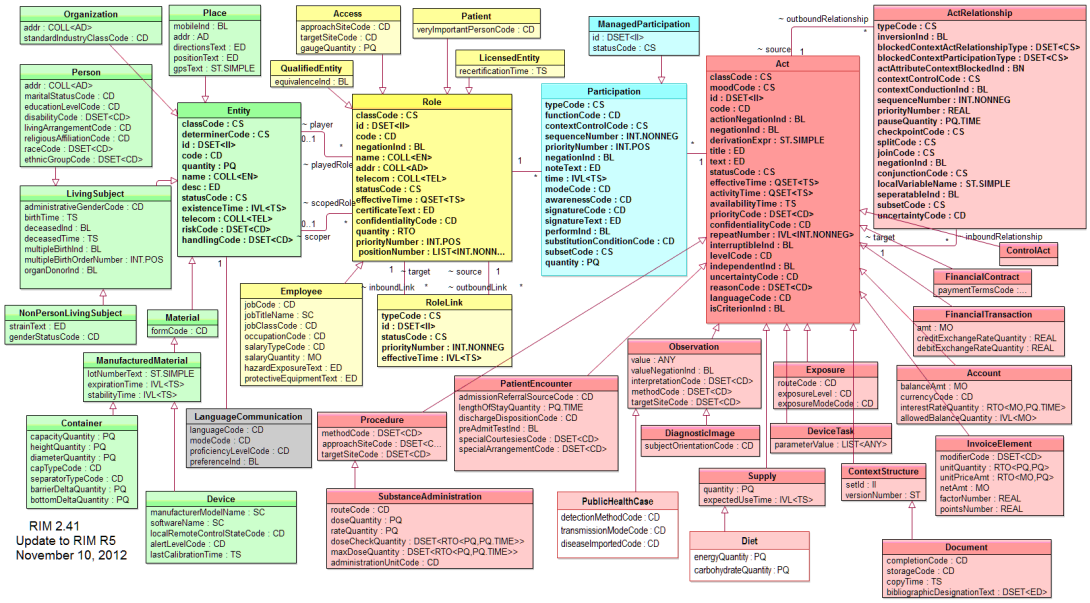


Figure 17: RIM Version 2.41 (HL7 International, 2012).

RIM classes may be used in a RIM-derived model. HL7 v3 specifications are derived from the RIM, thus ensuring a better level of semantic interoperability. Main idea of the HL7 v3 approach is that of constraining RIM, for specific usages. There are several specific constrained models e.g. Domain Message Information Model (DMIM), Refined Message Information Model (RMIM), Hierarchical Message Description (HMD) and message type etc.

DMIM is a general model of a domain. RMIM, which is most used constrained information model, only have one point of entry and can be expressed as a tabular format, i.e. a HMD. A message type can be exchanged as a linear string of XML and validated using an XML schema.

To develop the HL7 v3 message specifications, Message Development Framework (MDF) is used. The MDF is a complete, fully documented, model-based methodology and based on object oriented methodologies. HL7 V3 specification is built around subject domains, for each of which it provides storyboard descriptions, trigger events, interaction designs, DMIM derived from the RIM, RMIM derived from the DMIM, HMD and its resulting message types. Finally, those are encoded in XML. Based on the MDF, each Technical Committee and Special Interest Group in HL7v3 is contributing to the standard in its domain of expertise.

HL7 Clinical Document Architecture (CDA) a document markup standard, is produced to exchange information as part of the HL7 v3 standards, and aim to specify the structural and semantic aspects of clinical documents (Benson, 2010).

A HL7 CDA document is an XML file containing header and body sections. A CDA template is defined as a complete set of constraints of a CDA document and can be explained as detailed and human-readable document based on xPath based tabular approach. The generic CDA specification can be constrained through the document-level, section-level and entry-level templates. The unconstrained CDA specification is called "CDA Level One". When section-level templates are applied to an unconstrained CDA document, it is called "CDA Level Two". "CDA Level Three" is the CDA specification with entry-level templates applied (Boone, 2011).

4.1.3.2. HL7 Standards in NHIS-T

In NHIS-T, MHDS are produced as aggregated clinical document elements named as **transmission data sets** or **episodic EHRs** and then serialized into XML based on the HL7 Clinical Document Architecture (CDA) R2 structure to create **transmission schemas** (Figure 18) (Kose, et al., 2008), (Dogac, et al., 2011).

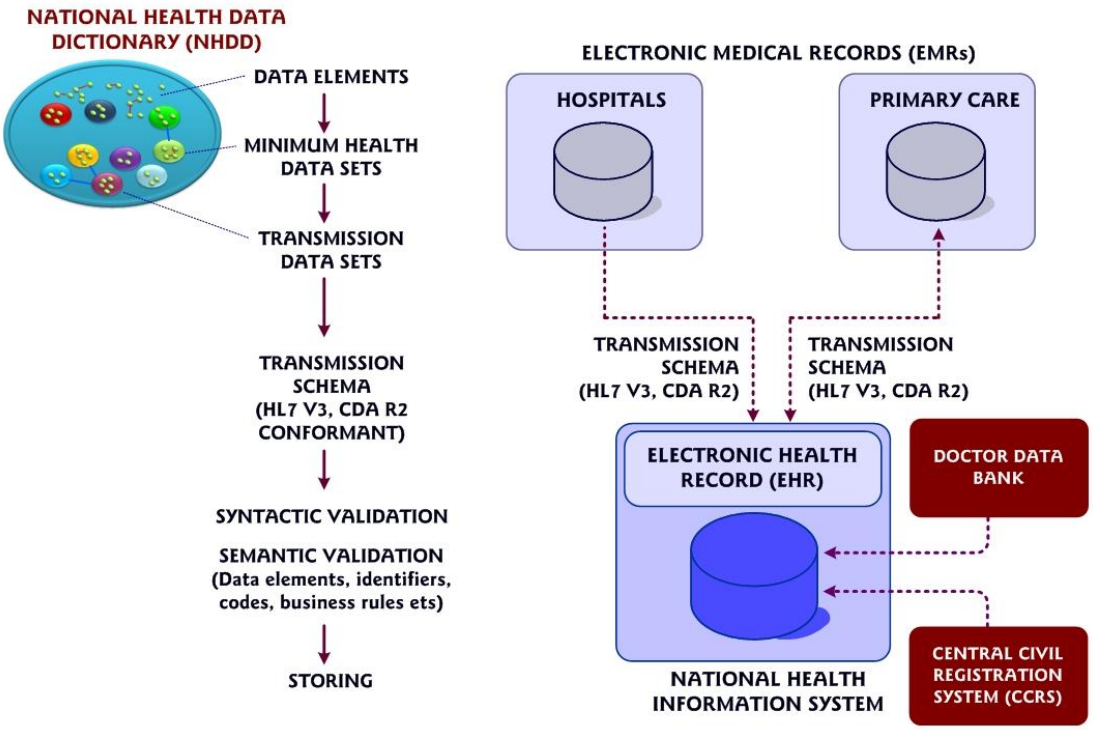


Figure 18: Schematic Representation of NHIS-T (depicted based on current literature).

In the current version of the NHIS-T, the transmission schema instances are localized according to Turkey's HL7 Profile. During this process, the rules which are set in the "HL7 Refinement, Constraint and Localization" are applied. However, the original HL7 CDA schemas are modified which breaks the CDA conformance of NHIS-T transmission schemas, since a conformant CDA document should at a minimum validate against the CDA

Level One Schema. Yet, since the messages are derived from CDA RMIM, the current version of NHIS is HL7 v3 CDA R2 compliant.

Then, each "Transmission Schema" is wrapped with a root element named after the main data set in the transmission. The "Data Sets" in the "Transmission Schemas" correspond to the "Sections" in the CDA Documents. And, the data elements are represented by nesting new "section" elements in the data set's "section" elements (Figure 19).

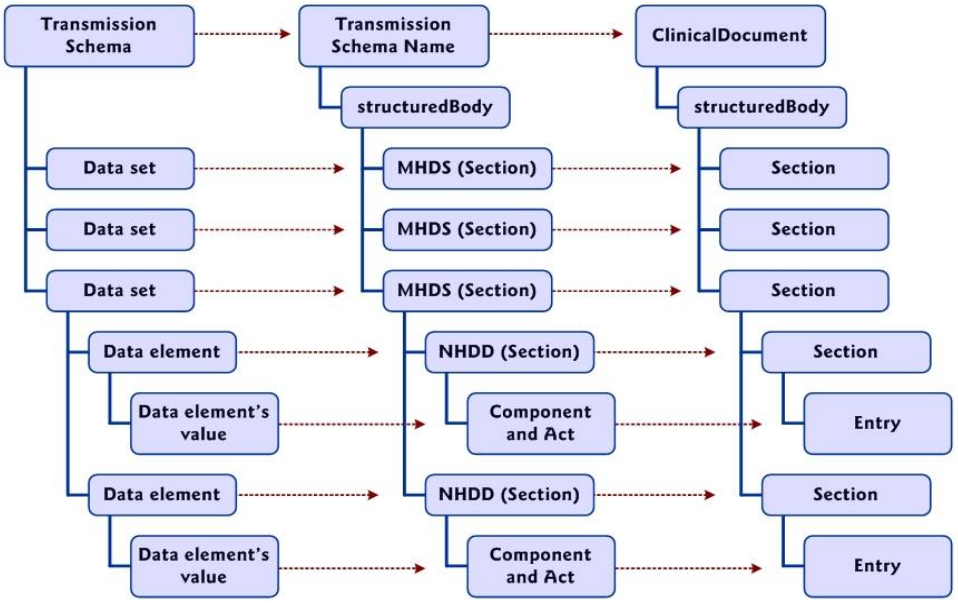


Figure 19: Relationships between the artifacts of NHIS-T, the "Transmission Schemas" and the HL7 v3 CDA R2 (Kabak, et al., 2008).

4.1.4 Validation and Storing

Before storing in the NHIS-T central repositories, incoming messages are validated regarding syntax, semantics and messages passed these two steps are stored in the central NHIS-T repositories (Kose, et al., 2008), (Dogac, et al., 2011).

Current version of NHIS-T allows the transfer of medical data from care providers' information systems to central servers via web services. It has the infrastructure that will provide access to patient's records for authorized healthcare professionals within the hospital, and that will allow patients to reach their own medical data i.e. PHR. But, the legal regulations have to be completed before both type of access, authorized or self, is available. Then, the establishment of a PHR system will allowed (Dogac, et al., 2011).

4.2 HL7 Standards for Clinical Genomic Domain

HL7 Clinical Genomics (CG) Work Group developed standards intended to interoperability issues in genomic medicine (HL7 Clinical Genomic Work Group, 2013).

4.2.1 HL7 v2 Standard for Genomic Data Sharing

HL7 organization has published an implementation guide based on HL7 Version 2 standards (i.e. HL7 Version 2 Implementation Guide: Clinical Genomics; Fully LOINC-Qualified Genetic Variation Model) using both the HL7 Version 2 Implementation Guide Laboratory Result Reporting to the EHR, and the HL7 Version 3 Genetic Variation data model. This guide covers the reporting of test results for sequencing and genotyping tests and includes testing for DNA variants associated with diseases and pharmacogenomic applications (Ullman-Cullere & Mathew, 2011). In Figure 20, an example of HL7 v2 message for genetic variation is presented.

```
MSH-->As according to HL7 VERSION 2.5.1 IMPLEMENTATION GUIDE: ORDERS AND
OBSERVATIONS; INTEROPERABLE LABORATORY RESULT REPORTING TO EHR (US
REALM), RELEASE 1, ORU^R01, HL7 Version 2.5.1, November, 2007.
OBR|1||PM-08-J00094^HPCGG-LMM^2.16.840.1.113883.3.167.1^ISO|lm_DCM-
pnlB_L^Dilated Cardiomyopathy Panel B (5 genes)^99LMM-ORDER-TEST-
ID||20080702000000|20080702100909|||||||||234567891^Pump^Patrick^^^^^^NPI
^L||||||20080703000000|||F||||||00000009^Cardiovascular^99HPCGG-GVIE-
INDICATION^^^^^^Clinical Diagnosis and Family History of
DCM|&Geneticist&Gene&&&&&NPI^^^^^^^HPCGG-
LMM&2.16.840.1.113883.3.167.1&ISO|||||||||||||||55233-1^Genetic analysis
master panel ^LN
SPM|1|||119273009&Peripheral blood&SNM3&&&&0707Intl&&Blood,
Peripheral||||||||||||20080702000000
OBR|2||PM-08-J00094-1^HPCGG-LMM^2.16.840.1.113883.3.167.1^ISO|55232-
3^Genetic analysis summary
panel^LN|||20080702000000|||||||||||||20080703000000|||F||||^PM-08-
J00094&HPCGG-LMM&2.16.840.1.113883.3.167.1&ISO
OBX|1|CWE|51967-8^Genetic disease assessed^LN||399020009^DCM-Dilated
Cardiomyopathy^SNM3^^^0707Intl||||||F|20080702100909||||||||||Laboratory
for Molecular Medicine^L^22D1005307^^^CLIA&2.16.840.1.113883.4.7&ISO|1000
Laboratory Lane^Ste. 123^Cambridge^MA^99999^USA^B
```

Figure 20: An example of HL7v2 message for genetic variation: MSH segment maps to the ClinicalDocumentModel; OBR (ObservationGroupModels) represent a set of Observations; SPM is SpecimenModel; OBX (ObservationModel) segment can be laboratory result.

"HL7 Version 2 Implementation Guide: Clinical Genomics; Fully LOINC-Qualified Genetic Variation Model" was the first example used by The Partners HealthCare Center for Personalized Genetic Medicine (PCPGM) and the Intermountain Healthcare Clinical Genetics Institute to gather genetic test results and transmit them to a patient's EHR (Shabo, et al., 2009), (Ribick, 2010).

4.2.2 HL7 v3 Standard for Genomic Data Sharing

The HL7 v3 Genetic Variation specification is based on the HL7 RIM. It uses the HL7 data types, vocabulary binding mechanisms built into the RIM and Bioinformatic Sequence Markup Language (BSML) to model the sequence information.

The root class in the Genetic Variation model is "GeneticLoci". The GeneticLoci model describes a set of loci, such as a haplotype, a genetic profile, and genetic testing results of multiple variations or gene expression panels. The GeneticLoci model uses the GeneticLocus model to describe each of these loci.

A Genetic Locus represents a single gene or coding region, and may have its own interpretation. A Genetic Locus is composed of one or more individual alleles, sequences and observed sequence variations.

Within the GeneticLocus model, HL7 suggests the sharing of the essential part of raw genomic via "encapsulation" and extracting clinically relevant data via "bubble-up" based on genomic decision support application (Figure 21) (Shabo, 2006).

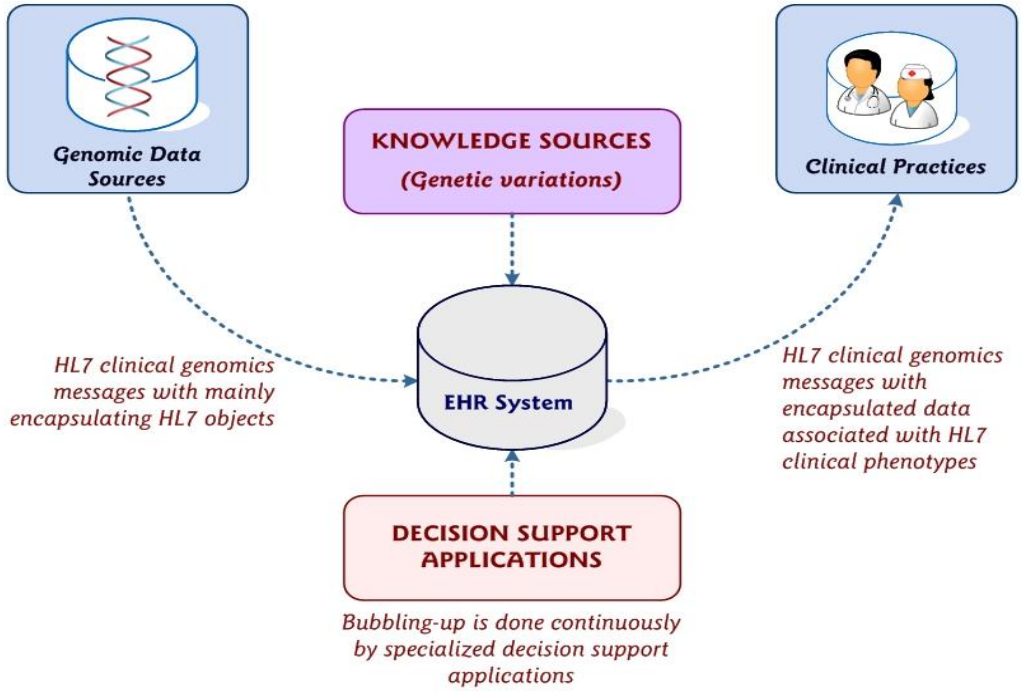


Figure 21: Encapsulation and bubble-up workflow with a focus on enterprise EHR systems accompanied by decision-support applications: In the static phase of this workflow, encapsulation is performed based on a static predefined BSML schema. In the dynamic phase, clinically relevant SNP data is bubbled up into HL7 SequenceVariation objects, and these objects are linked with clinical data from the patient EHR, thus ensuring the disease and therapeutic risk assessment.

An example XML code segment for GeneticLocus is presented in Figure 22. ***The encapsulation phase*** contains the incorporation of raw genomic data sent from genomic data sources to clinical settings, based on a predefined, constrained bioinformatics format i.e. BSML. Constraining the bioinformatics markup schemas ensure us to dismiss clinically irrelevant data elements and to refer genomic data to one patient only with the patient identifiers.

The bubble-up phase is an iterative process wherein various clinical genomic decision support applications parse the encapsulated raw genomic data and make prominent the clinically relevant data based on the most up-to-date knowledge available. The results of this phase are held in genotype-phenotype associations supported by the standard's specification.

```
<GeneticLocus>
        <individualAllele moodCode="EVN">
            <text>EGFR receptor gene</text>
            <value code="EGFR"/>
                <sequence moodCode="EVN">
                    <value mediaType="text/xml">
                        <bsml:Sequences>
                            <bsml:Sequence id="seq1" molecule="dna" title="EGFR..."
                                length="5616">
                                <bsml:Seq-data>
                                gcgcggccgc agcagcctcc gccccccgca cggtgtgagc gcccgacgcg
                                ccggagtccc gagctagccc cggcggccgc cgccgcccag accggacgac
                                    ...
                                </bsml:Seq-data>
                            </bsml:Sequence>
                        </bsml:Sequences>
                        <bsml:Isoforms>
                            <bsml:Isoform-set>
                                <bsml:Isoform id="variation1" seqref="seq1" location="2240"
                                    change="" replaces="cctcttcatg cgaaggcg"/>
                                <!-- possibly more isoform tags denoting other variations -->
                            </bsml:Isoform-set>
                        </bsml:Isoforms>
                    </value>
                    <sequenceVariation moodCode="EVN">
                        <code code="DNA.MUTATION"/>
                        <text>A somatic mutation in the active site of the EGFR receptor gene
                                is found in about 10% of non-small cell lung cancer tumors</text>
                        <value xsi:type="CE" code="131550.0001" displayName="18-BP DEL,
                                NT2240" codeSystemName="OMIM"/>
                        <interpretationCode code="DELETERIOUS"/>
                            <clinicalPhenotype classCode="ORGANIZER">
                                    <observationGeneral>
                                        <code/>
                                        <statusCode/>
                                        <effectiveTime value="20010101"/>
                                        <value xsi:type="CE" code="D2-F1007"
                                                codeSystemName="SNOMED CT"
                                                displayName="Non-small cell lung cancer"/>
                                    </observationGeneral>
                            </clinicalPhenotype>
                            <clinicalPhenotype classCode="ORGANIZER">
                                    <observationGeneral>
                                        <text>Iressa (gefitinib) responder</text>
                                        <effectiveTime value="20010101"/>
                                        <value xsi:type="CS" code="gefitinib-responder"/>
                                    </observationGeneral>
                            </clinicalPhenotype>
                            <associatedProperty>
                                <code code="TYPE"/>
                                <text>
                                    <reference value="#variation1"/>
                                </text>
                                <value xsi:type="CV" code="SNP"/>
                            </associatedProperty>
                            <associatedProperty>
                                <code code="REFERENCE"/>
                                <value xsi:type="URL"
                                        value="http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?
                                        db=nucleotide&val=41327737"/>
                            </associatedProperty>
                    </sequenceVariation>
                </sequence>
        </individualAllele>
</GeneticLocus>
```

Encapsulated data

Bubble-up data

Figure 22: Sample code segment from GeneticLocus XML instance.

4.3 Architectural Extension of SNP Data Incorporated NHIS-T

To incorporate clinicogenomic information into medical records, as explained in the "Background and Literature Review", it is necessary to build an infrastructure providing clinicogenomic information and subsequent updates to physicians, a curated knowledge base

extracting clinical information from relevant SNP data, and supporting systems processing up-to-date data for clinical decisions over the patient's lifetime (Aronson, et al., 2012) (Ginsburg & Willard, 2013).

In the light of literature, for a SNP data incorporated NHIS-T, we need some improvements in three components; 1) Enhancement of existing messaging infrastructure to share personal SNP data and clinicogenomic associations between stakeholders, 2) Development of a national level clinicogenomic knowledge base for transforming personal SNP data to clinicogenomic associations, 3) Advancement of end user applications (EMR, PHR, etc.) for reporting of clinicogenomic interpretation of clinically relevant SNP data.

Regarding technical capabilities (e.g. network bandwidth, storing and processing capacities, etc.) different types of architectures can be developed, but development of two additional components (knowledge base and reporting capability) is inevitable. Clinicogenomic knowledge base must be constructed at the national level as a manually curated and continuously updated source which contains clinical information and its possible associations with SNP variants. In the end users applications (EMR, PHR, etc.), clinicogenomic associations and external data (e.g. family history, environmental and behavioral data) must be interpreted independently or based on predictive models to support decision making. In this section, possible architectures are overviewed. Knowledge base, decision support and reporting capabilities will be analyzed in the next chapters.

4.3.1 Sharing Raw Data

After next generation genotyping tests (e.g. WGS), a file which contains a huge amount of variant data is acquired (Aronson, et al., 2012). WGS data involve about 3 billion base pairs and entire genome sequence is about 3.2 Gb. Storing and sharing of personal raw genomic sequence exceeds the transmission and storage capacity in many healthcare organizations (Kahn, 2011). Due to the technical limitations, raw genomic data is stored the outside of the EMR similar to PACS for medical images and clinical interpretation of data is preferable sent to the EMR database (Starren, et al., 2013), (Masys, et al., 2012), (Green, et al., 2013). In a characteristic human genome, WGS will determine more than 3 million variations. After filtering processes, hundreds to thousands of clinically relevant variations, which have the potential to be underlying reason, could be extracted. (Bamshad, et al., 2011), (Bick & Dimmock, 2011), (Biesecker, 2012), (Raffan & Semple, 2011).

In our study, we have several restrictions and additional focuses that are not provided throuh HL7 v3 clinical genomics presumptions. In our design we propose to;

- use only SNP variation data (not other types of variations and omics data),
- present SNP data as a preprocessed input file e.g. direct to consumer genomic variation file (not as raw data or nucleotide sequence),
- develop and exploit a central, accredited and national level knowledge base,
- concentrate on clinical practice (not research domain),
- interpret the clinical effects of SNPs using predictive models (not only independently),

- exploit additional behavioral and family history data.

For these reasons, although it's possible to share raw data in HL7 v3 standards, we preferred to reduce raw data into clinically relevant data and accepted to collect raw data in genomic laboratory side and share clinically relevant data in EMR/EHR/PHR level. Additionally, we used both rs number and allele value as unique identifier for SNPs and we didn't use HGVS and HGNS standards.

4.3.2 Examples of DTC Genomic Data Formats

In DeCODEme file, SNP data is reported in a comma-separated text file (Figure 23).

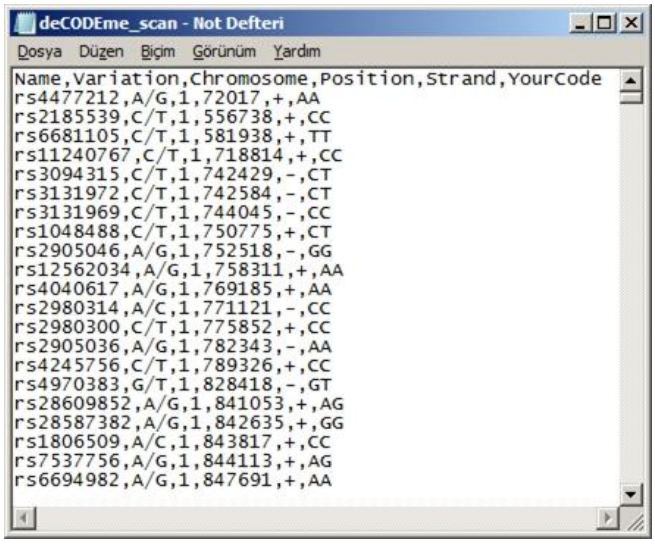
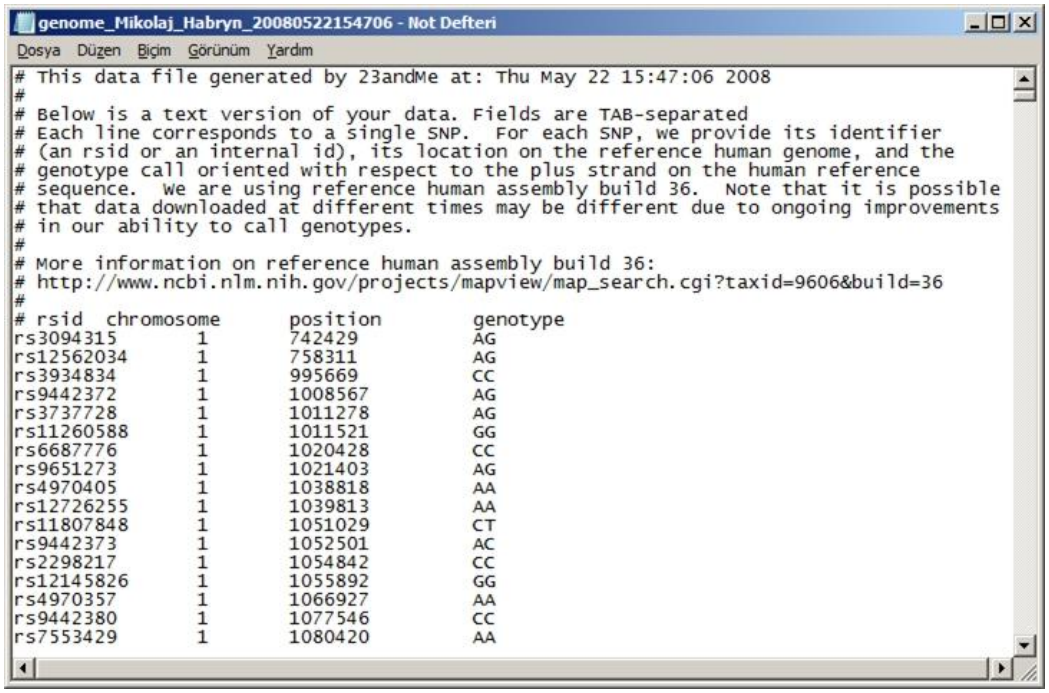


Figure 23: DeCODEme file format.

In this file format, "name" refer to rs number of SNP from dbSNP, "variation" contain are the possible allele nucleotides (A, C, G, T, or --), "chromosome" and "position" define the physical location of the SNPs. "Strand" refer to the orientation used to identify SNP e.g. plus or minus. "YourCode" provides the two genotypes for the sample.

There are several 23andMe file formats e.g. version 1, 2 and 3. SNP data is reported in a tab-separated text file (Figure 24).

In this file format, "rs_id" refer to rs number of SNP from dbSNP, "chromosome" and "position" identify the physical locations of the SNPs. "Genotype" ensures the two genotypes for the sample. In 23andMe file format, genotype values are oriented with respect to the plus strand on the human reference.


genome_Mikolaj_Habryn_20080522154706 - Not Defteri

Dosya Düzen Biçim Görünüm Yardım

```
# This data file generated by 23andMe at: Thu May 22 15:47:06 2008
#
# Below is a text version of your data. Fields are TAB-separated
# Each line corresponds to a single SNP.  For each SNP, we provide its identifier
# (an rsid or an internal id), its location on the reference human genome, and the
# genotype call oriented with respect to the plus strand on the human reference
# sequence.  We are using reference human assembly build 36.  Note that it is possible
# that data downloaded at different times may be different due to ongoing improvements
# in our ability to call genotypes.
#
# More information on reference human assembly build 36:
# http://www.ncbi.nlm.nih.gov/projects/mapview/map_search.cgi?taxid=9606&build=36
#
# rsid  chromosome      position        genotype
rs3094315       1       742429          AG
rs12562034      1       758311          AG
rs3934834       1       995669          CC
rs9442372       1       1008567         AG
rs3737728       1       1011278         AG
rs11260588      1       1011521         GG
rs6687776       1       1020428         CC
rs9651273       1       1021403         AG
rs4970405       1       1038818         AA
rs12726255      1       1039813         AA
rs11807848      1       1051029         CT
rs9442373       1       1052501         AC
rs2298217       1       1054842         CC
rs12145826      1       1055892         GG
rs4970357       1       1066927         AA
rs9442380       1       1077546         CC
rs7553429       1       1080420         AA
```

Figure 24: 23andMe file format.

4.3.3 Identifiers for SNP Data

In order to integrate structured genotype and phenotype data into any system, first requirement is to determine data components, terminology standards and identifiers of clinicogenomic information i.e. genotype data and its associated clinical interpretation.

Because of extensive usage and practicality we decided to use "rs number" to identify every single SNP. As explained before, SNPs are identified by rs number in many SNP resources, e.g. PharmGKB, AlzGene, PDGene, SzGene, SNPedia, GWAS Central Cancer GAMAdb and some sorts of personal genomic file formats (e.g. 23andMe, deCODEme, and Navigenics).

Because different alleles of SNPs may have different degrees and types of clinical impact, rs number is insufficient alone to identify the clinicogenomic significance of SNPs. For example, to have a heterozygote allele may not change the risk for the disease but homozygote allele of the same SNP variant may change the risk for a disease dramatically. To identify clinically relevant SNP, we need to use the combination of rs number and allele data as a minimum requirement (Attia, et al., 2009A).

Due to double stranded (plus and minus or forward and reverse respectively) nature of DNA, every SNP, can be identified using either of these strands. Sometimes, in various genomic databases, same SNPs alleles are defined with different alleles based on the orientation discrepancy (Attia, et al., 2009A). In our system, we have used plus strand as the standard.

4.3.4 SNP Data and NHIS-T CDA Transmission Schemas

In existing NHIS-T, medical and laboratory examination results are sent from hospitals to the central EHR databases as "Examination Result Transmission Data Set". The HL7 v3 CDA R2 conformant transmission schema of this data set contains several MHDS e.g. registration MHDS, result of tests MHDS, patient MHDS, etc. "Result of tests MHDS" involves data elements about examination features (order time, protocol number, result time, test result, reference value ranges, etc.). The data type of laboratory examination should be numeric or textual data regarding current schema standards (http://www.e-saglik.gov.tr/SaglikNet/SaglikNetDokumanlari.aspx).

Although WGS and other types of genotyping tests are acceptable as laboratory tests, they have different characteristics than other laboratory tests in routine practice. After clinical WGS test, a personal SNP data file which contains a huge amount of variant data is produced, in which all variant data need to be managed in an effective way. As explained in Chapter 2.3, to cope with the technical limitations, raw genomic data is stored the outside of the EMR and clinical interpretation of data is sent to the EMR database (Starren, et al., 2013), (Masys, et al., 2012), (Green, et al., 2013).

In our case, personal sequencing data is planned to store as raw data within a genomic laboratory information system. The clinically relevant part of individual SNP file and inferred clinical meaning of this data file is shared between corresponding databases via NHIS-T CDA schemas as an encapsulated text file.

HL7 v3 interoperability standards support encapsulated data type for text data (Benson, 2010). Encapsulated data can be used to transmit audio, video, images, genetic sequences, etc. In CDA, encapsulated data can be used in two ways i.e. directly incorporating into the CDA document or referencing by a URL. Mostly, encapsulated data types found in either <section> elements of the CDA document or in various clinical statement elements. Genomic sequence data may appeared in the <value> element of an <observation>. When the encapsulated data is provided through a reference the element containing the encapsulated data will contain a <reference> element (Boone, 2011).

Possible architectural alternatives for NHIS-T are discussed in detail in the following Chapter 5.

CHAPTER 5

DESIGN AND DEVELOPMENT

5.1 General Architecture of SNP Data Incorporated NHIS-T

In the proposed architecture, personal sequencing data is acquired and stored as raw data within a genomic laboratory information system (Figure 25). The clinically relevant personal SNP (CR-SNP) data extracted from personal SNP data file using the CR-SNP data list (genomic identifiers of clinicogenomic associations in clinicogenomic knowledge base). Then, personal CR-SNP data file is sent via NHIS-T infrastructure from genomic laboratory to central EHR databases as an encapsulated text file.

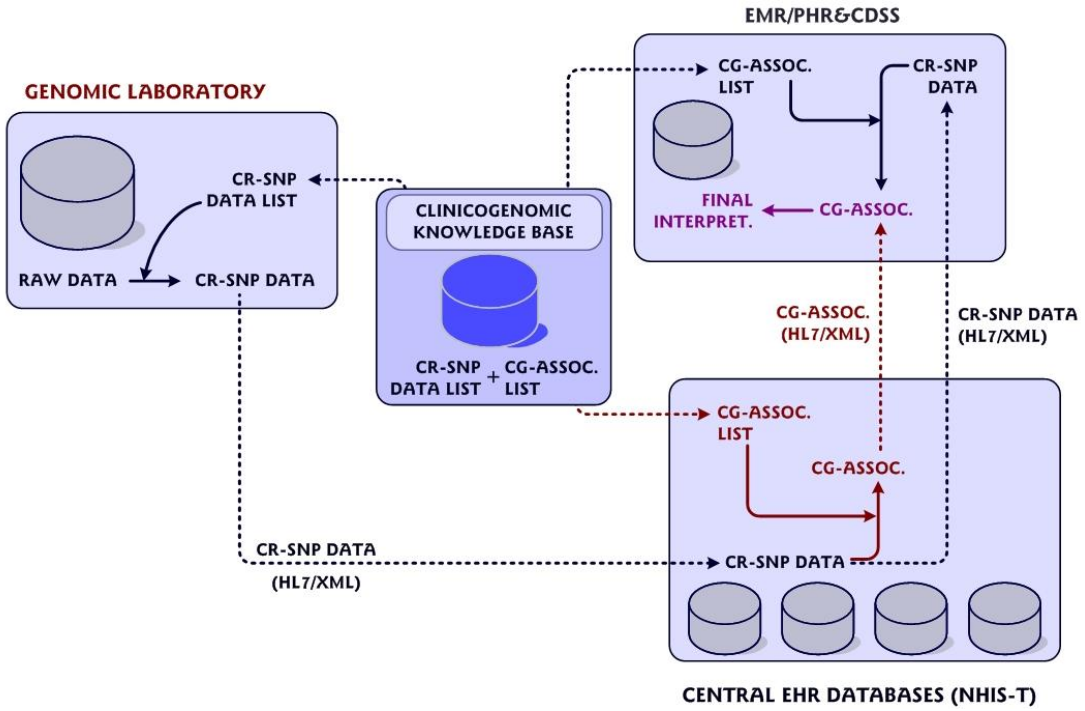


Figure 25: Extended architecture for genome enabled NHIS-T. Two possible sharing alternatives are drawn. Based on clinicogenomic knowledge base, extraction of clinicogenomic associations from personal CR-SNP data in EHR is the first choice (red lines), and extraction of these associations in EMR/PHR is the second (blue lines). In both alternatives, final clinicogenomic interpretation can be completed in end user. CR-SNP; Clinically Relevant SNP, CG-ASSOC.; Clinicogenomic Associations, CDSS: Clinical Decision Support System.

Personal CR-SNP file has to include SNP identifiers (e.g. rs number and allele data) in the HL7 CDA R2 schema. The received CR-SNP files are stored within the central EHR databases. Then CR-SNP files are processed to infer clinically relevant data by using the

clinicogenomic associations from knowledge base (Figure 26). Resulting personal clinicogenomic association files are sent to end users.

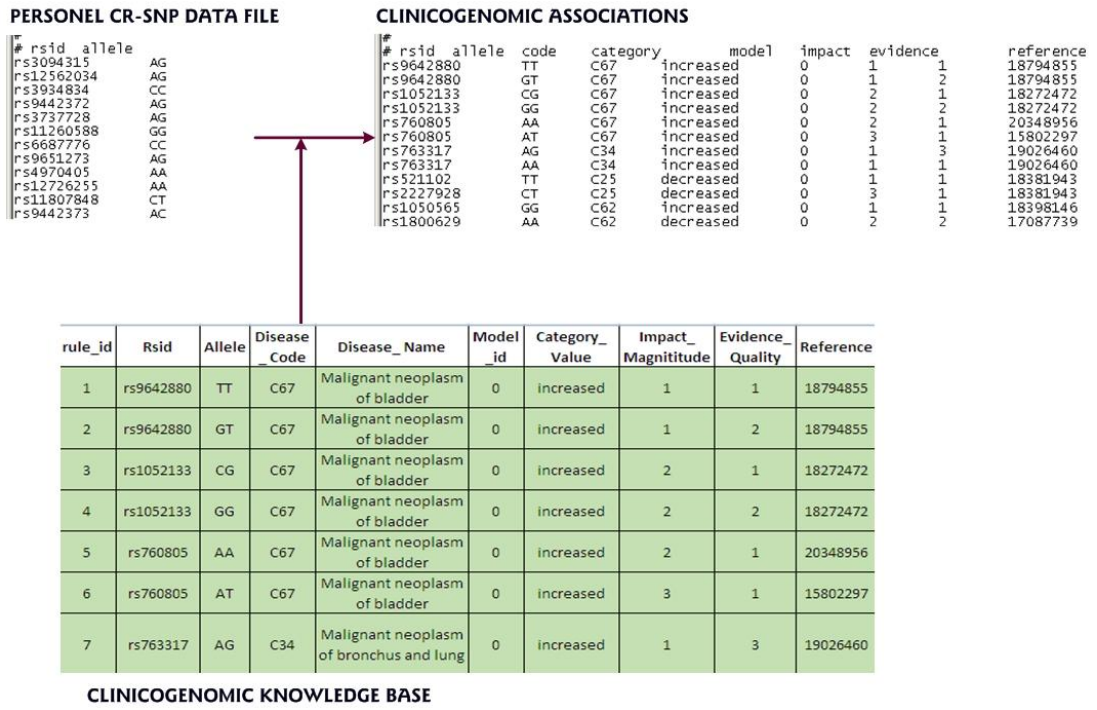


Figure 26: Converting CR-SNP to clinicogenomic associations based on clinicogenomic knowledge base.

Based on existing technical capabilities, to decrease the load of sharing clinicogenomic associations files, replicated knowledge bases can be used in client side web applications, and in central servers only CR-SNP data files can be stored. In this situation, clinicogenomic associations are inferred in client side replicated knowledge base. In this approach, the replication of reference central knowledge base to client side knowledge base must be synchronous.

When an authorized user, (patient, family practitioner, or a specialist in a hospital) need to reach personal CR-SNP or clinicogenomic association files, a request is sent to the central EHR and current data file is received via NHIS-T communication infrastructure.

In practice, to be able to share CR-SNP or personal clinicogenomic associations between central EHR databases and end-user systems, the capabilities of NHIS-T need to be extended (e.g. web services, client side inference and reporting capabilities, if demanded PHR).

When clinicogenomic knowledge base is updated by domain experts, according to type and level of change and preferred architecture, existing personal SNP data, CR-SNP data and/or clinicogenomic associations must be re-interpreted automatically in genomic laboratory system, national EHR repository and/or client side.

5.2 Complementary Components

Here, we have analyzed how we can use design and develop complementary capabilities using with the existing clinicogenomic information and assessment approaches. In general, our proposed architecture contains two complementary components i.e. knowledge base and decision support application (Figure 27). In our model, input data (genomic data e.g. SNP file) is processed using knowledge base and assessed by decision support applications. In scientific literature and daily practices, various assessment and reporting approaches are proposed e.g. collective reporting of associations, polygenic risk scoring, model based approaches etc. Detailed analyses of these methods are explained in "Ch.7: Assessment and Reporting Approaches".

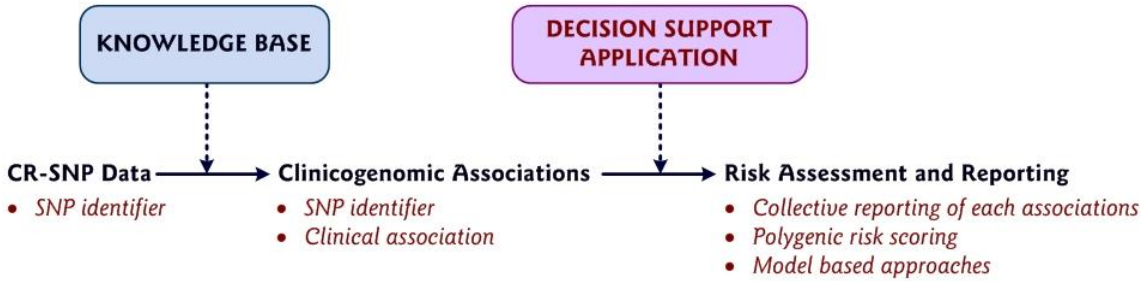


Figure 27: Complementary components and their functionalities.

5.2.1 Analysis of Clinicogenomic Associations and Models

The general analysis of clinicogenomic information specifically focuses on the SNP data and its clinical associations. In our study, we extracted more than 100 associations between SNPs and clinical information for cancer cases from SNPedia. SNPedia is a wiki source and contains categorized information summaries from scientific literature. Also, we profoundly studied various scientific publications to determine the assessment and reporting methods of clinicogenomic information for prostate cancer risk assessment.

In this step, we have determined that clinicogenomic associations have several characteristics in terms of clinical functionality, complexity of information, input variable types, and degree of clinical significance. Table 6 shows these features and examples as a whole.

5.2.2 Clinical Functionality

Various clinical functionality categories are determined for cancer cases including risk assessment (increased or decreased risk of disease), disease prognosis, treatment efficacy, and presence of adverse events.

"Clinical functionality" section of Table 6 ensures examples for each category. For risk assessment category, gathered associations indicated that T allele of rs798766 variation increases disease risk, whereas G allele of same variation has decreasing effect. Regarding disease prognosis, after diagnosis of certain disease, existence of a SNP have negative impact, such as rs713041(T) allele in breast cancer. Third type of association shows how a specific allele increase or decrease efficient of treatment. And last one, association (4),

presents an example of SNP associated adverse event for a specific drug. We have discovered that, most of (95%) of the extracted associations are SNPs associated with disease risk. We were able to collect less than ten SNPs about prognosis and only three SNPs about pharmacotherapy of cancer cases from SNPedia (Table 6, Examples 1-4).

Table 6: Categorization of different types of clinicogenomic associations from SNPedia.

Parameter	Examples of Information
Clinical Functionality	
Risk assessment	*(1)The T allele at rs798766 is associated with 1.24 higher odds of bladder cancer* (SNPedia, 2011), *but G allele of Rs3790844 appears to lower the risk of pancreatic cancer* (SNPedia, 2010).
Disease prognosis (poor)	*(2)The rs713041(T) allele is associated with an increased risk of death following breast cancer diagnosis, with a hazard ratio of 1.27 per rare allele* (SNPedia, 2007).
Efficacy of treatment	*(3)A testicular cancer patient with a rs1050565(G;G) genotype which is treated with bleomycin , has an odds ratio of 4.97 for testicular germ-cell cancer related death compared to (A;G) or (A;A) genotypes. The rs1050565(G;G) genotype also shows a higher prevalence of early relapses* (SNPedia, 2008).
Drug adverse events	*(4)rs12762549 and rs11045585 can be used to predict whether docetaxel will induce leukopenia/neutropenia* (SNPedia, 2008).
Information Complexity	
Simple	*(5)rs3218536(A) carriers also appear to be at lower risk for epithelial ovarian cancer. In a study of ~1,600 cases, the odds ratio for rs3218536(A;G) heterozygotes was 0.8 (CI: 0.7-1.0) and for the (quite rare) rs3218536(A;A) homozygotes 0.3 (0.1-0.9)).[PMID 15924337]* (SNPedia, 2005).
Combined	*(6)A report that attracted some media attention of a model for risk of prostate cancer was based on a combination of 5 SNPs plus family history, which the authors believe may account for 50% of the cancers. Although the model estimates risk, clinical parameters (such as age of onset and disease progression) are not predicted by this model. The 5 SNPs … are rs4430796, from ch 17q12, rs1859962, from ch 17q24.3, rs16901979, from ch 8q24 (region 2), rs6983267, from 8q24 (region 3), rs1447295, from 8q24 (region 1. Risk for prostate cancer (shown here as odds ratio, with CI) increases cumulatively based on the number of SNP risk genotypes for these 5 (or, second set of numbers, with family history counted as a 6th factor) as follows: 0: 1.00, 1: 1.50 (CI: 1.18-1.92); 1.62 (1.27-2.08), 2: 1.96 (1.54-2.49); 2.07 (1.62-2.64), 3: 2.21 (1.70-2.89); 2.71 (2.08-3.53), 4: 4.47 (2.93-6.80); 4.76 (3.31-6.84), 5: 4.47 (2.93-6.80); 9.46 (3.62-24.72), 6: 9.46 (3.62-24.72)* (SNPedia, 2008).
Input Variable Types	
Only SNP	*(7)To have rs1052133 (C,G) variant increases the risk of gallbladder cancer 1.9 times* (SNPedia, 2011).
SNP and sociodemographic	*(8)In African American males, CYP3A4 variation rs2740574 is associated with a ~10 fold higher risk of prostate cancer…* (SNPedia, 2008).
SNP and envirobehavioral (nutrition)	*(9)There is evidence that high intake of phytoestrogens substantially reduce prostate cancer risk among men with a variation of rs2987983 in the promoter region of the estrogen receptor-beta gene* (SNPedia, 2008).

Since there are insufficient information for the last three categories of clinical functionalities (i.e. prognosis, efficacy of treatment and adverse event of treatment), it would be misleading to design structured typologies based on these limited information. Therefore, we focused on clinicogenomic associations related with disease risk assessment.

5.2.3 Information Complexity

According to the complexity of components, clinicogenomic information can be categorized as simple or combined statements. "Information Complexity" section of Table 6 presents examples for both simple and combined statements.

Simple information is representable as independent associations including only one SNP variant and its associated clinical characteristics. Example (5) of Table 6 contains two simple clinicogenomic associations. In these statements, presence of one SNP data with certain allele has specified impact on a disease with odds ratios.

On the other hand, combined information can be acceptable as integrated statistical impact model of more than one SNP on a specific medical condition. Example (6) given in Table 6 is an example for combination of more than one simple association. This is an example of predictive model which are more sophisticated and valuable tools regarding clinical practice.

In SNPedia, we found three combined information (one for prostate cancer and two for breast cancer). All of these are focused to calculate and interpret cumulative impact of various SNPs on specific medical conditions. We can construct such types of combined information unifying several simple clinicogenomic associations within a risk assessment model. In some cumulative models, impact factors may be altered according to which alleles of SNPs are homozygotes or heterozygotes.

In models, existence of each association contributes to the total risk of the clinical condition for patients. In cumulative models, through analysis of patient's genotype, total count of associations are determined and calculated additively. Among risk assessment tools besides cumulative disease models, there are other ongoing efforts utilizing different data mining algorithms to interpret GWAS data for building various risk assessment models. For all types of these models, we can abstract a holistic model as combination of different number of simple independent associations.

Various models has been proposed to assess and report genomic data in scientific literature. These models, their components and calculation of their impact on clinical conditions are explained in "Ch.7: Assessment and Reporting Approaches".

5.2.4 Classification of Input Variables

Another characteristic of clinicogenomic information extracted from SNPedia is the diversity and complexity of input variables. It's known that, many of the complex diseases are aroused by virtue of interactions between environmental, behavioral, sociodemographic or/and biological (e.g. genomic, pathologic, etc.) parameters. Some information in SNPedia contain only one input variable, i.e. SNP variant, but other sources may include combination of SNP variant and additional variables about clinical findings, environmental, behavioral or sociodemographic parameters. In Table 6, 7th clinicogenomic association is an example of

one input with only SNP variable, whereas 8th and 9th associations include multiple types of variables.
Today, unfortunately most of the environmental and behavioral parameters are not recorded and stored in EMR/EHR in a structural way including NHIS-T. Therefore in this study, we limited our scope to share only SNP data and incorporate other types of data i.e. environmental, behavioral and sociodemographic parameters in the end user level to calculate total risk of disease.

5.2.5 Degree of Clinical Significance

For significance of clinicogenomic association, in SNPedia, magnitude is accepted as a subjective measure of interest for magnitude of impact and repute (good, bad) for quality of evidence, but these concepts are not well established.

Some of clinicogenomic knowledge sources contain different attributes to identify magnitude of clinical effects and strength of relationship between SNP variations and diseases. As a measurement of impact value for clinicogenomic associations, researchers usually prefer to use conventional approaches, e.g. odds ratios and relative risks for case control studies and cohort studies respectively (Attia, et al., 2009C). Most of clinicogenomic associations have small effect sizes and their credibility may largely depend on the success of control for errors and biases. To calculate and assign an evidence degree, several studies performed (Ioannidis, et al., 2008). Detailed information is given in "Ch. 6.4: Qualifiers for Clinicogenomic Associations."

5.3 Design of Standardized Association and Model Definitions

After analyzing the various sorts of clinicogenomic associations (independent and model based) from SNPedia and extraction of the scientific literature regarding assessment and reporting methods, we have compare the required data fields for knowledge base and decision support applications to determine which parameter will be included in our study (Table 7).

Table 7: Comparison of assessment and reporting types of associations and data fields.

Parameters	**Assessment and Reporting of Associations**			
	Collective Assessment of Independent Associations	**Polygenic Risk Scoring**	**Cumulative Models**	**Risk Parameters Sequences**
Association definition				
Association identifier	X		X	X
Rs number	X	X	X	X
Allele	X	X	X	X
Disease code	X	X	X	X
Disease name	X	X	X	X
Magnitude of impact (Odds ratio)	X			
Degree of evidence quality	X			

Table 7 (cont.): Comparison of assessment and reporting types of associations and data fields.

Parameters	Assessment and Reporting of Associations			
	Collective Assessment of Independent Associations	**Polygenic Risk Scoring**	**Cumulative Models**	**Risk Parameters Sequences**
Association definition				
Impact category		X		
Evidence category		X		
Impact value				
Branch_id				X
Model definition			**X**	**X**
Model type	X	X	X	X
Model name	X[a]	X[a]	X	X
Total impact			X	
Total count of SNPs				X
Branch_id				X
Narrative interpretation			X	X
External data definition				
Family history			X	X
Other type (BMI, etc.)				X
a= In this study, only increased risk covered				

5.3.1 Definition of Clinicogenomic Associations

After comparison of assessment and reporting types of associations and data fields, we have produced a standard representation of all types of clinicogenomic associations as Figure 28.

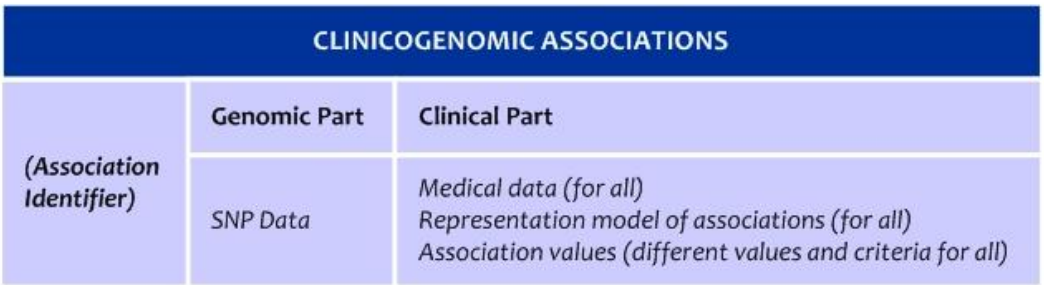


Figure 28: Standard representation of clinicogenomic associations.

Detailed data analysis of association typology is presented in Table 8. This table is used to develop the clinicogenomic knowledge base.

Table 8: Analysis of data fields for association parameters.

Data category	Parameters	Values
Association identifier	Assoc_id	Unique numeric value
SNP data	Rs number	Rs value
	Allele	Allele value (e.g. AA, AT, CG, etc.)
Medical data	Disease code	ICD-10 standard
	Disease name	
Representation model of associations	Model type	Independent associations, Cumulative model, Hybrid model based associations
	Model name	[CA];increased, decreased [CM]; number of SNP, name of author [RS]; name of author and date
Association Values	Parameter 1 [CA]; magnitude of impact, [CM]; impact value, [RS]; branch_id	[CA]; Numeric value, [CM]; numeric value [RS]; Numeric identifier
	Parameter 2 [CA]; degree of quality of evidence	[CA]; Numeric value (between 1 and 3)
	Parameter 3 [PS]; impact category	[PS]; 1,2,3 (corresponding Weak, Moderate, Strong respectively)
	Parameter 4 [PS]; evidence category	[PS]; 1,2,3 (corresponding Weak, Moderate, Strong respectively)
[CA] Collective Assessment of Independent Associations [PS] Polygenic Risk Scoring [CM] Cumulative Models [RS] Risk Parameter Sequences		

In this representation, clinicogenomic associations must have a unique identifier assigned automatically. In ClinGenKB, SNPs are identified by both rs number and allele value on the forward strand. If a SNP refers to more than one medical condition or model, then for every case, a new association was defined, and a different identifier is assigned for every different rule. Medical data category contains diagnosis code and name. Values of this data fields are selected from ICD-10, which is used as diagnostic terminology for diseases in current NHIS-T.

Model data has two components i.e. type and name. In ClinGenKB, we have two main clinicogenomic associations i.e. model based or model free (or independent) associations. Names for independent associations are categorized increased and decreased regarding potential risks or protective characteristics. In this study, we only focused on increased risk. Model based associations are used in predictive models.

Association values are tightly related to the type of model i.e. independent associations, cumulative model, and hybrid model based associations. For independent associations; odds ratios, degree of evidence quality, impact and evidence categories are appropriate and sufficient elements to evaluate clinical significance in an independent and complete way. In cumulative model based associations, it’s required to assign an impact value for every

association to calculate total personal risk value according to model definition table. For hybrid models, we calculate the total effects of variants using branch_id. In client side, all risk parameters involved in hybrid models grouping by "branch_id" and for all these groups, the total impact of risk parameters are determined. If one of these values is equal to the total value of the corresponding branch, it's accepted to have a risk of prostate cancer with accuracy, precision and recall values of the model.

5.3.2 Definition for Predictive Risk Models

We developed a model definition table to the process model based associations on the end user side (Table 9). In this table, the terms of "model type" identifies the category of models and "model name" entitles them. The fields of "total value" and "explanation" are map to the total impact of related SNPs and corresponding risk categories.

For cumulative models, these fields are about total impact and its explanation. For risk parameter sequences, the total value is referred to count of all SNPs for every branch, explanation is the interpretation of risk values regarding accuracy and precision. Risk parameter sequences used in our cases are hybrid model (SVM and decision tree) based associations and we need to determine additional data field to identify branches of decision tree.

These tables must be kept up to date and shared other stakeholders to use in their systems as a standard reference to interpret the model based associations in a proper manner. Also, model type and model name fields must be used as a standard reference for same fields in definition tables.

Table 9: Data field analysis of model definition table.

Data category	Parameters	Value (Domain)	Explanation
Model definition data	Model type	independent associations, cumulative model, hybrid model based associations	
	Model name	[CA];increased, decreased [CM]; number of SNP, name of author [RS]; Model number, name of author and date	
	Total value	Numeric value	[CM]Total impact [RS] Total count of SNPs
	Explanation 1	Text value	[CM]Explanation (Odds ratio) [RS]Explanation (Brief interpretation about risk assessment)
	Explanation 2	Text value	[RS]Explanation (Branch id)
[PS] Polygenic Risk Scoring , [CM] Cumulative Models, [RS] Risk Parameter Sequences			

5.4 Design and Development of Complementary Components

The complementary capabilities of the proposed prototype i.e. ClinGenKB and ClinGenWeb, which ensures clinical decision support capability is described in this section.

Requirements of a clinical decision support infrastructure are centrally managed repositories of machine readable medical knowledge, standardization of clinical decision support information provided for specific aspects of genomic medicine, standardized representation of patient data, standard approaches for leveraging machine readable medical knowledge, and standard approach for locating and retrieving patient data (Kawamoto, et al., 2009). The components of the proposed SNP incorporated NHIS-T system, and how it complies with the requirements is described in detail in Table 10.

Table 10: Analysis of the SNP incorporated NHIS-T regarding clinicogenomic decision support system requirements (Kawamoto, et al., 2009).

Requirement, Explanation and Examples	**SNP incorporated NHIS-T**
Centrally managed repositories of machine readable medical knowledge (e.g. PharmGKB, NCBI dbGaP, ClinVar)	ClinGenKB (specifically designed and developed for prostate cancer risk assessment)
Various formalisms for representing medical knowledge in a computer processable format (e.g., Arden Syntax, GELLO, GLIF, PROforma, SAGE, SEBASTIAN)	Vendor specific (BioXM Knowledge Management Environment)
Standardization of clinical decision support information provided for specific aspects of genomic medicine (e.g., HL7 refined message information models, openEHR Archetypes, HL7 Decision Support Service semantic profiles)	Model definition table (tabular representation of models, specifically designed and developed for prostate cancer risk assessment)
Standardized representation of patient data (e.g., HL7 data standards; openEHR Archetypes; SNOMED CT; LOINC; BSML; MAGE-ML; National Cancer Institute/caBIG Common Data Elements)	Rs number and allele value for SNP, ICD-10 for diagnostic terminology, encapsulated CR-SNP data in HL7 CDA R2 schema and CG-ASSOC. data for messaging.
Resources that enable mapping between different terminologies (e.g., Unified Medical Language System, National Cancer Institute Enterprise Vocabulary Services, HL7 Common Terminology Services standard)	Not needed yet for the coverage of system.
Standard approaches for leveraging machine readable medical knowledge (e.g., Arden Syntax, SAGE, PRODIGY, GLIF, SEBASTIAN, First DataBank Drug Information Framework, HL7 Decision Support Service)	ClinGenWeb, as a calculation and reporting capability (specifically designed and developed for prostate cancer risk assessment)

Table 10 (cont.): Analysis of the SNP incorporated NHIS-T regarding clinicogenomic decision support system requirements (Kawamoto, et al., 2009).

Requirement, Explanation and Examples	**SNP incorporated NHIS-T**
Regional and national initiatives for secure health data exchange (e.g., U.K. National Health Service Connecting for Health, U.S. Nationwide Health Information Network prototypes, caBIG, Indiana Health Information Exchange)	NHIS-T infrastructure

Complementary components of our system provide the following functions;

- Capturing, storing and real-time access to the individual data,
- Ensure communication channels between the end user, individual and domain expert,
- Assistance for the decision making process by assessing the individuals' risk status based on the entered data and generating reports,
- Recognition of patterns in the collected data set and assessment of risk degree.

5.4.1 General System Architecture

Knowledge base (i.e. specifically ClinGenKB for prostate cancer) contains the clinicogenomic associative rules. The rules in the knowledge base can be generated through a knowledge base editor or can be transferred from another application by the human medical experts. The general rules and common facts about the disease can be extracted from the medical books, websites, or human expert knowledge.

ClinGenWeb is a web-based system, including modules that cover the individual prostate cancer risks including personal SNPs, comorbidities, environmental and behavioral parameters. The system is composed of the data transfer, manual data entry, data storage, and data view and reporting functions.

Personal clinically relevant SNP (CR-SNP) data is stored in "raw data table" and personal clinicogenomic associations (CG-ASSOC) are stored in "processed data table". CR-SNP data converted to CG-ASSOC based on the knowledge base. Then, individual CG-ASSOC file is transferred to the ClinGenWeb CG-ASSOC table. If web services are built and integrated, it's possible to automatically transfer individual CR-SNP and CG-ASSOC files to the ClinGenWeb from central EHR servers. Additionally, comorbidities, environmental and behavioral parameters can be recorded by data entry module. Model and equation definition are recorded by model definition module. Reporting module provides structured outputs of the collected data set. The system has the capability to show different statistical reports and graphical representation of the collected data.

Figure 29 represents the proposed system architecture, main modules and interaction between modules.

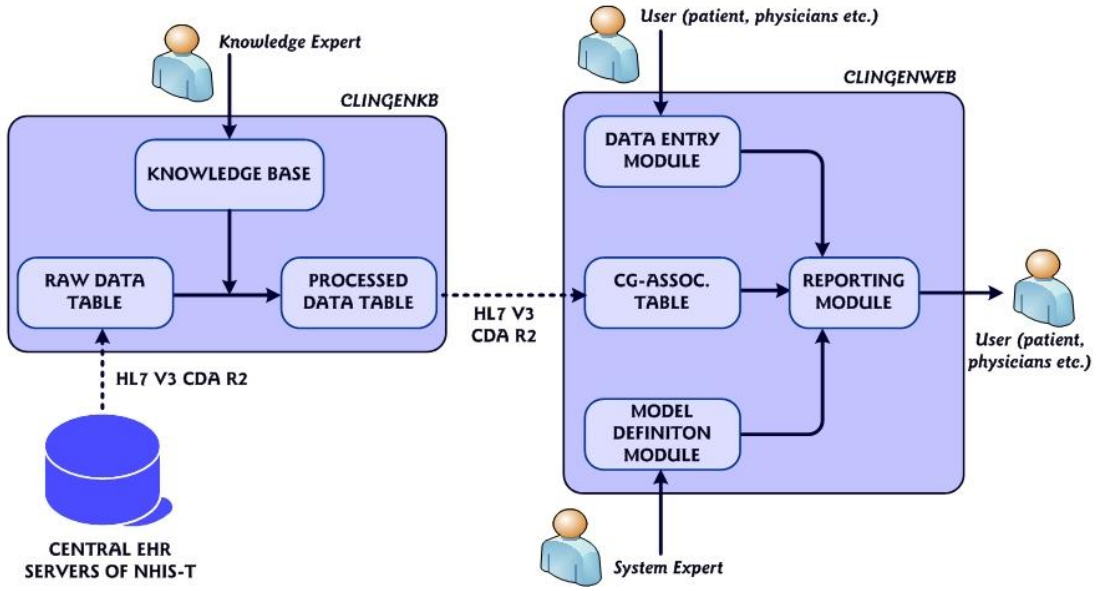


Figure 29: System architecture, main modules and the interactions between modules (focused on to complementary components).

5.4.2 Use Case Diagrams

Use-Case diagram of Unified Modeling Language (UML), is a kind of behavioral diagram which represents a graphical overview of the functionality provided by a system regarding actors, targets (i.e. use cases), and dependencies between use cases. An actor can be a human being or another system interacting with modeled system (Aleksovska - Stojkovska & Loskovska, 2011). The use case diagram of the complementary capabilities presented as a whole in Figure 30.

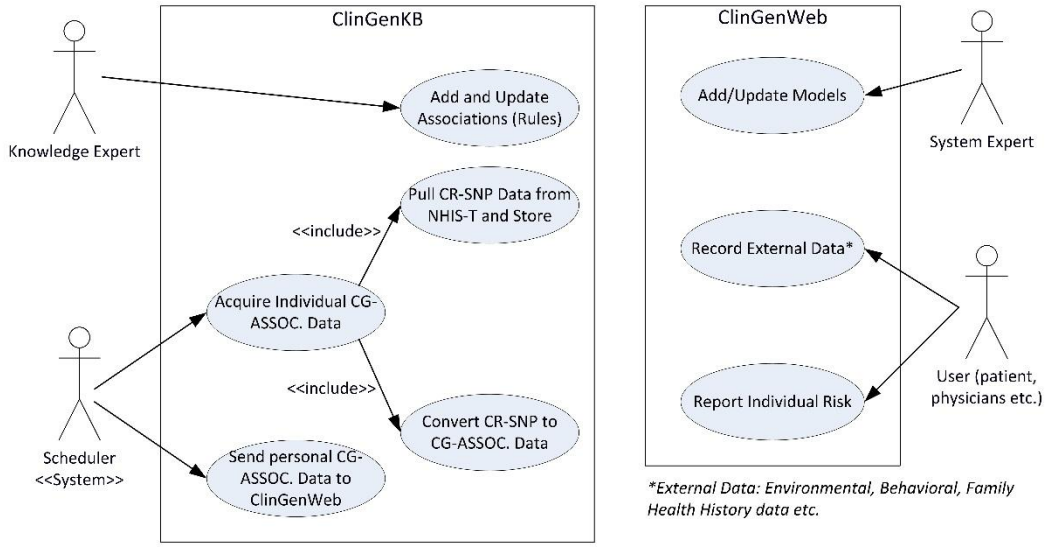


Figure 30: The use case diagram of our ClinGenKB and ClinGenWeb with NHIS-T infrastructure.

Major actors of our systems are NHIS-T, ClingenKB, ClinGenWeb Application, end user (physicians or patients), knowledge expert and system expert. Sending and storing functionalities will be performed automatically by NHIS-t infrastructure, ClinGenKB, and ClinGenWeb. Knowledge expert add and update clinicogenomic associations into the ClinGenKB. Conversion process of CR-SNP to the CG-ASSOC accomplished synchronously by ClinGenKB at the first uploading and then in each rule update session.

In ClinGenWeb application, system experts define reporting processes and models based on standard model definition table.

End users (i.e. patients, authorized physicians etc.) can record external data of patients (e.g. environmental, behavioral, family health history etc.) and report individual disease risk independently and using defined assessment models.

5.4.3 Data Model of Knowledge Base

We composed our domain model defining elements, annotations, relations and scopes of this components in BioXM™ based on association definition table (Figure 31). In this domain model, we have three types of elements; person, SNP variant and rule (clinical association). Every element has their specific annotations. Person element is related with the SNP variant by 'has' relation, referring that each patient might have a set of SNP variants. SNP identifiers are assigned to variants for ensuring uniqueness. Then each SNP variants are related with a rule (clinical association) element as input.

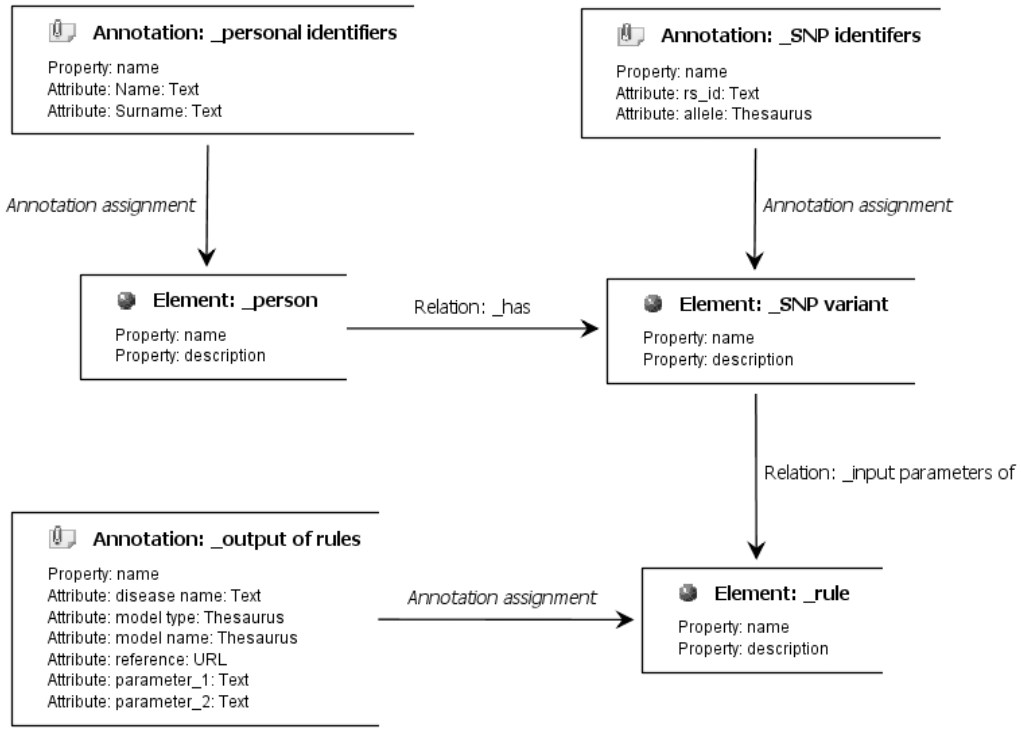


Figure 31: The graphical representation of data model of the proposed ClinGenKB implemented with BioXM™ Knowledge Management Environment.

5.4.4 Implementation of Knowledge Base

We have imported the content of knowledge base definition table as an external file with scripts. This content can be updated with subsequent importing operations. If a new association is generated or existing associations are changed or cancelled, authors can prepare all these changes in an external source according to association definition table and then easily can upload all of them via BioXM compatible files. After an importing process, clinicogenomic associations can be sorted and managed by administrator from table (Figure 32a).

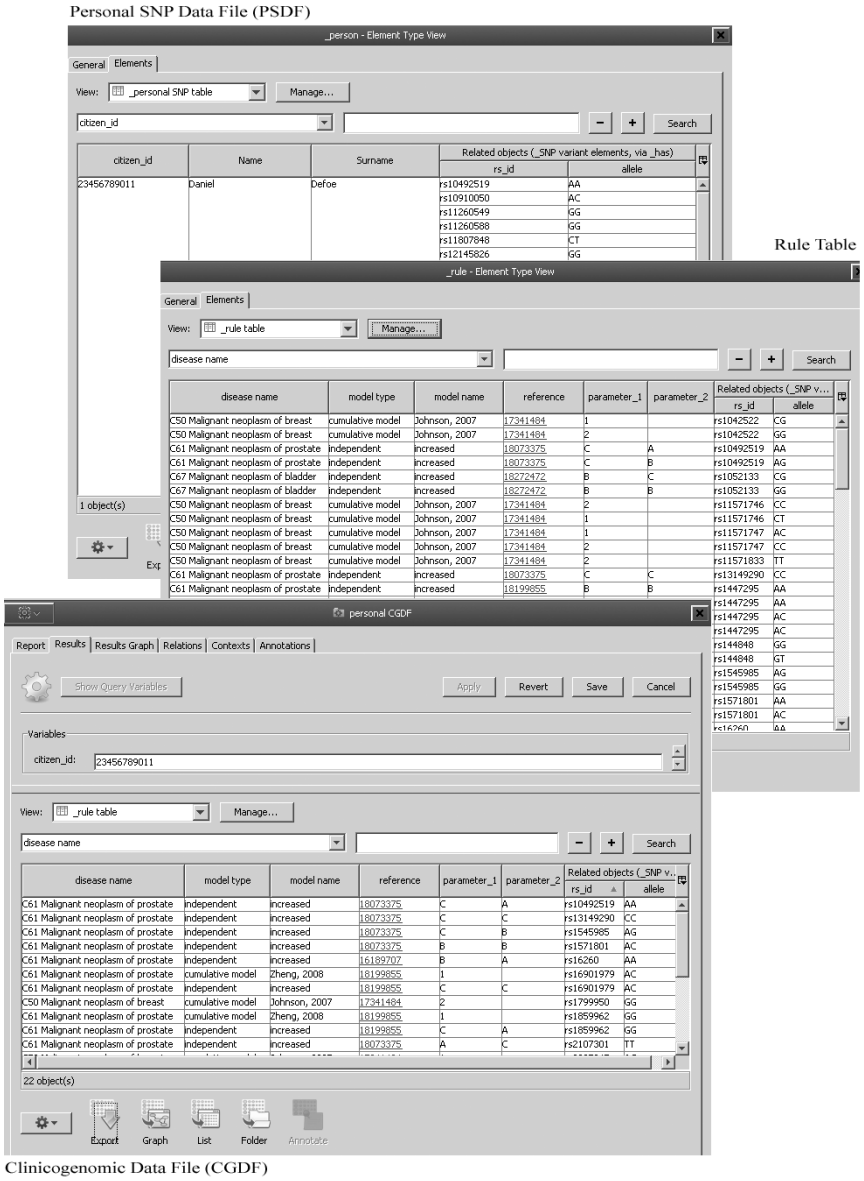


Figure 32: Some screens from ClinGenKB.

Clinicogenomic associations are prepared in an external source as suitable for ClinGenKB association definition table and uploaded with importing script. Personal CR-SNP data is stored in ClinGenKB as a separate file and accessed with a citizen identifier. Personal CR-SNP data can be easily converted to clinicogenomic associations data file based on the content of ClinGenKB and exported as a text file.

In addition, we can store personal CR-SNP data as a separate file on BioXM™ (Figure 32b). CR-SNP data can be easily converted to clinicogenomic associations based on the content of ClinGenKB, and this data file is exported as a text file. For all individuals, whom his/her CR-SNP data stored in BioXM™, whenever it's needed, it's possible to access personal CR-SNP data and to produce new clinicogenomic associations data files based on current ClinGen. All of these files can be accessible with a citizen identifier for our prototype and can be sorted according to data categories (Figure 32c).

5.4.5 Entity-Relationship Diagram of Decision Support

An entity–relationship (ER) model is a data model to draw the data aspects of a relational database. The main components of ER models are entities and the relationships. A possible ER diagram for decision support and reporting application is presented in Figure 33.

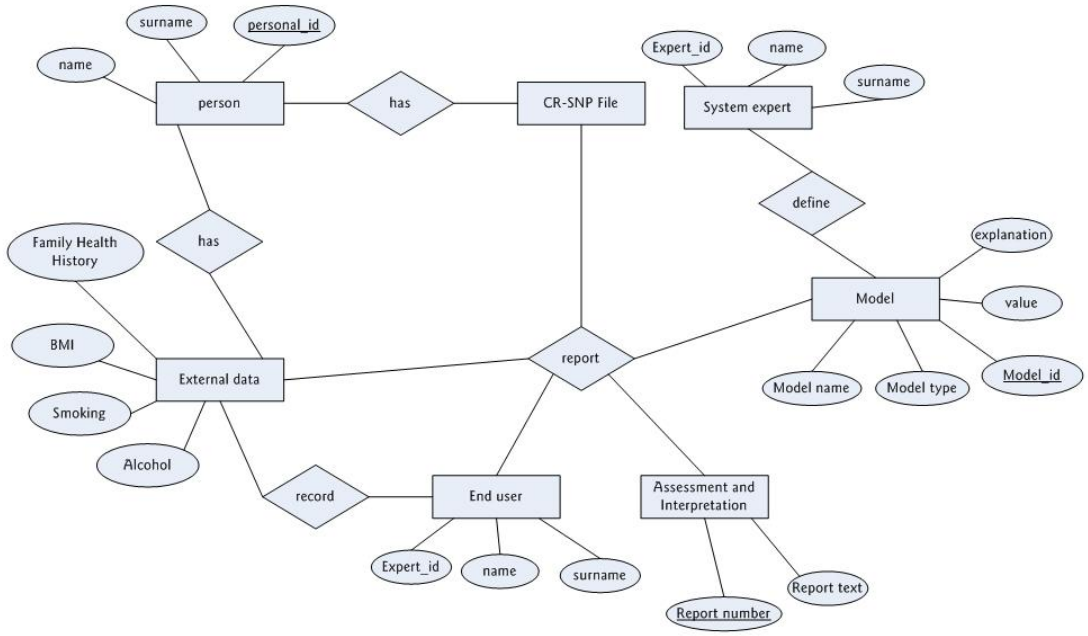


Figure 33: Entity-relationship diagram of decision support and reporting application.

5.4.6 Implementation of Decision Support and Reporting Application

During the implementation of our prototype, we have selected the Zoho Reports™ , and a simple prototype system has been developed by using the Zoho Reports™ on the client side. Our prototype (specifically named as ClinGenWeb for prostate cancer) is built as a web application processing genomic associations, clinical and environmental risk parameters. In this application, it's possible to report the relevant clinicogenomic SNPs or to assess the independent risk based on the models with the combination of conventional health data and clinicogenomic associations.

In ClinGenWeb, the personal predictive risk can be analyzed in three main categories i.e. collective reporting of independent associations, the complete assessment of clinically relevant SNPs (polygenic scoring), and model based interpretation of clinicogenomic associations. Some type of models are based on assessing only relevant SNPs. But a few models involve external data (family history, BMI, etc.). If collected, corresponding risk

factors for prostate cancer can be used to calculate the model based risk. Also, external personal data about clinical and some environmental risk factors for prostate cancer can be reported.

5.4.6.1. Collective Assessment and Reporting of Independent Associations

Reporting of all independent associations one by one will be very confusing, and the interpretation of these data by users will be time consuming. Instead of this approach, we considered using associated data in a category based graph which axes correspondingly to impact and evidence categories (Figure 34).

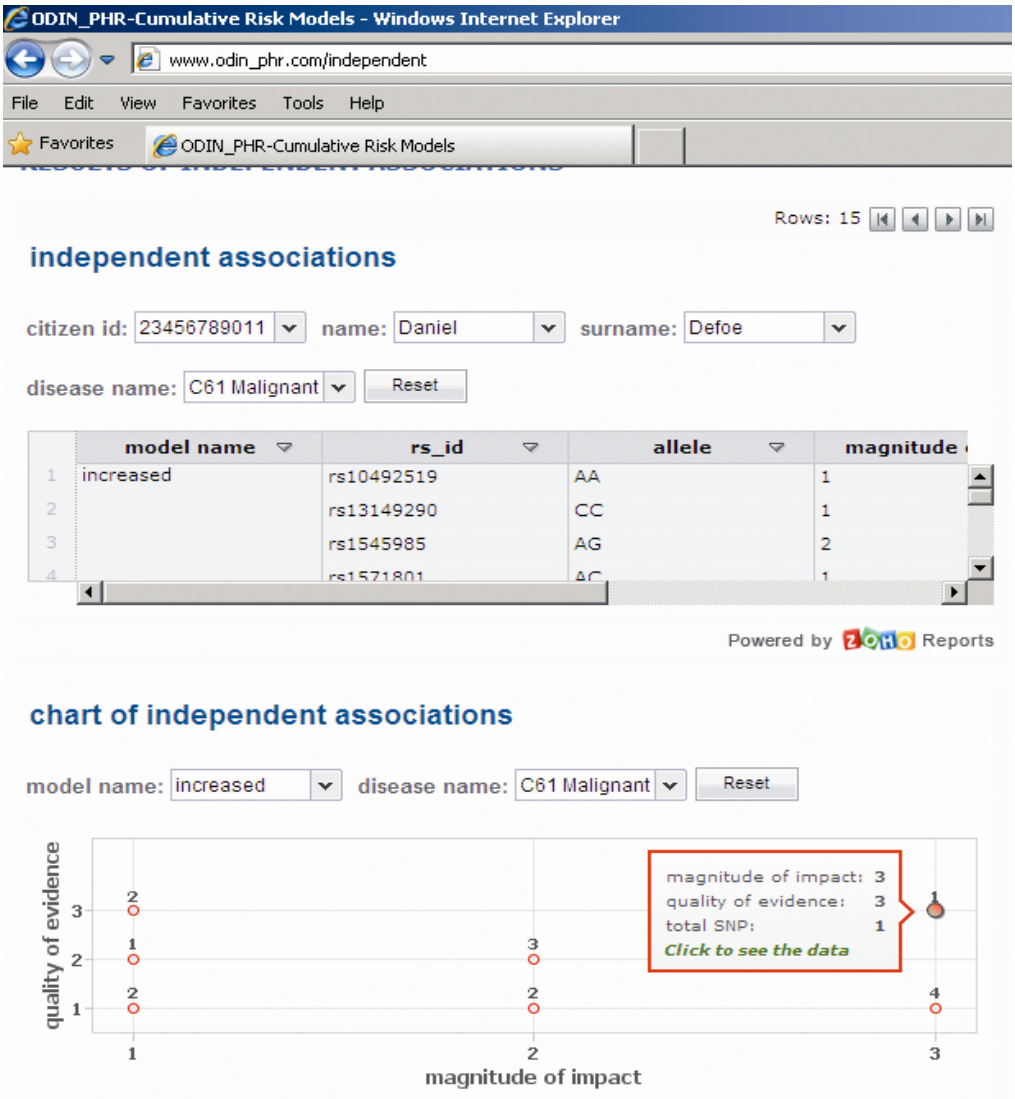


Figure 34: Visualization of independent associations in ClinGenWeb. Independent associations and their clinical significances (magnitude of impact and quality of evidence) are listed in the screen. Below, users can access the total graphical representation of clinically relevant variants for any disease. If users move cursor on a point, a brief explanation (count of variant according significance) is appeared and if clicking at this point, more detailed list of variants are sorted in the screen.

5.4.6.2. Reporting of Model Based Associations

Comparatively, results of model based interpretation give us more effective information to the end users for decision making. These are based on more accepted and proven integrated models.

In our study, we have two kinds of model i.e. cumulative and the hybrid model based association sets. In ClinGenWeb, results of these models and detailed explanation of reference values are presented to end users as a whole. If needed and end users can exploit detailed analysis of risk factors sorted below the total evaluation of the model (Figure 35).

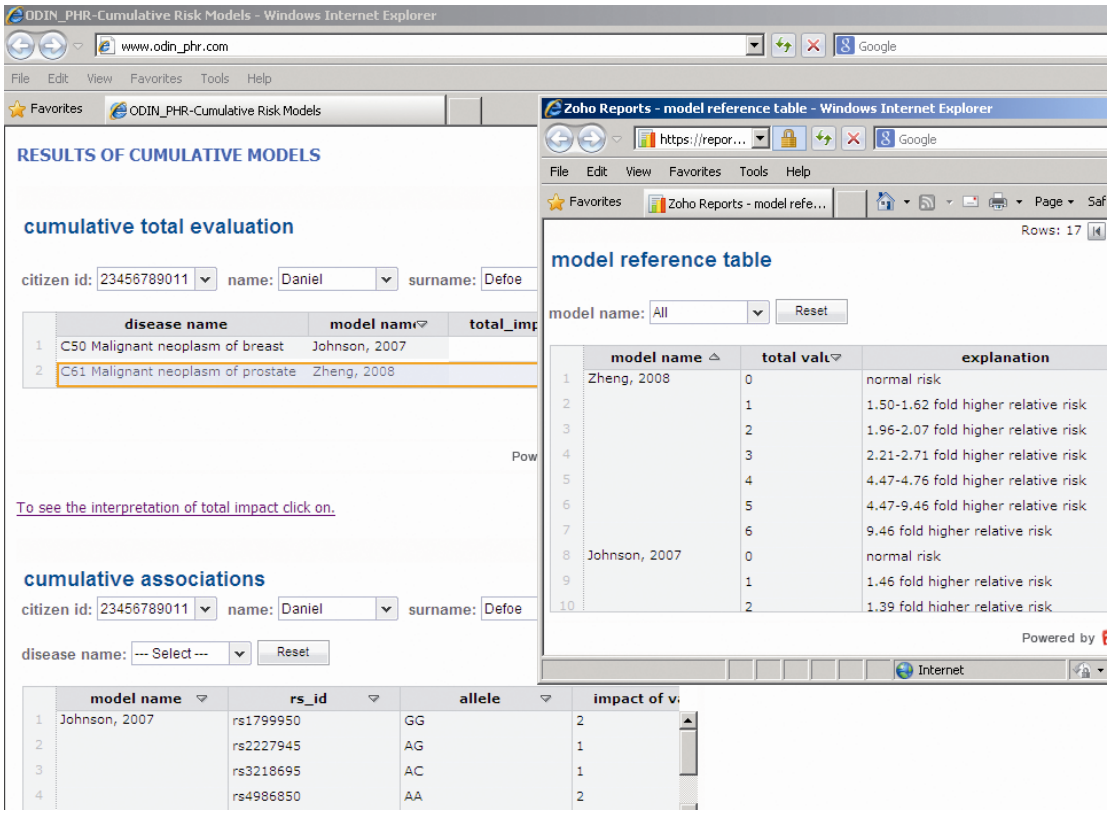


Figure 35: Reporting of model based rules in ClinGenWeb.
In client side, model based rules are processed using CGDF and model operators. Finally, total results of model based evaluation are presented for every disease. Users can access the whole impact (and meaning) of models clicking on interpretation link. List of clinically relevant variants is sorted following total evaluation, and if users want to access relevant variants for any disease, they can select proper disease(s) from "disease name" combobox.

5.4.6.3. Combining Clinicogenomic Associations and External Data

In ClinGenWeb, users can record and store other types of risk factors (family history, environmental, behavioral, and clinical data) to assess their prostate cancer risk. If this type of data is collected, these can be used as part of models or reported as a whole.

5.5 Interoperability Level of the Proposed Architecture

Interoperability is defined by Institute of Electrical and Electronics Engineers as "the ability of two or more systems or components to exchange information and to use the information that has been exchanged".

The HL7 EHR Interoperability Work Group has developed a framework, which covers three layers i.e. technical, semantic and process interoperability. *Technical interoperability* focus on to moves data from one system to another apart from its meaning. *Semantic interoperability* ensures that systems understand the data in the same way. This is specific to domain and context and usually involves the use of codes and identifiers. *Process interoperability* coordinates work processes, enabling the business processes at the organizations which have systems to work together.

Currently, the Health Level 7 (HL7) version 2 Messaging Standard is the most widely implemented message interface standard in the healthcare domain. However, being HL7 version 2 compliant does not support semantic interoperability between healthcare systems. To solve this problem, HL7 version 3 is developed, which is based on an object oriented data model i.e. RIM.

But some studies extended the interoperability definitions and models. The Levels of Conceptual Interoperability Model (LCIM) represents a hierarchy of capabilities for representing the meaning of information passed between systems, components, or services (Table 11) (Tolk, et al., 2009).

Table 11: Levels of Conceptual Interoperability Model (LCIM)

Level	**Type**	**Description**	**Requirements**
0	No interoperability	No build in capability to exchange data	None
I	Technical Interoperability	Systems are exchanging data with each other.	Ability to produce and consume data in exchange with systems external to self is required to be technically interoperable.
II	Syntactic Interoperability	Data exchange is taking place within an agreed to protocol, that all systems are able to produce to, and can consume them.	An agreed-to protocol that all can be supported by the technical level solution is required to achieve this level of interoperability.
III	Semantic Interoperability	Systems are exchanging a set of terms that they can semantically parse.	Agreement between all systems on a set of terms that grammatically satisfies the syntactic level solution requirements is required for this level.
IV	Pragmatic Interoperability	Systems will be aware of the context and meaning of information being exchanged.	A method for sharing meaning of terms is required, as well as a method for anticipating context. These both should be based on what exists at the semantic level.

Table 11 (cont.): Levels of Conceptual Interoperability Model (LCIM)

Level	Type	Description	Requirements
V	Dynamic Interoperability	Systems are able to reorient information production and consumption based on understood changes to meaning, due to changing context.	A method for defining meaning and context is required to achieve this level. The means of producing and consuming these definitions is required for dynamic interoperability.
VI	Conceptual Interoperability	Systems at this level are completely aware of each others information, processes, contexts, and modeling assumptions.	A shared understanding of the conceptual model of a system (exposing its information, processes, states and operations) must be possible in order to operate at this level.

Both HL7 Clinical Genomics model and NHIS-T are based on HL7 RIM and support semantic interoperability. For this reason, our model also is capable to support semantic interoperability.

Our architecture, also support pragmatic and dynamic interoperability. To ensure pragmatic interoperability, message sent by a system causes the effect intended by that system; i.e., the intended effect of the message is understood by the collaborating systems (Asuncion & van Sinderen, 2010). Standardizing and disseminating predictive models to use in decision support systems ensure pragmatic interoperability. In Ch. 5 we proposed a standardized definition table for predictive models. Pragmatic interoperability can only be achieved if systems are also syntactically and semantically interoperable.

Dynamic interoperability can be achieved if systems are able to reorient information production and consumption based on understood changes to meaning. We emphasized that it's an obligation to develop knowledge base system to re-interpret CR-SNPs using scientific changes and new information in our architecture. Re-interpretation of CR-SNP based on standard, accredited and curated knowledge base ensure us the dynamic interoperability.

CHAPTER 6

EXTRACTION OF CLINICOGENOMIC ASSOCIATIONS FOR PROSTATE CANCER CLINICOGENOMIC KNOWLEDGE BASE

6.1 General Approach

Clinicogenomic associations, namely associations between a specific clinical conditions and a specific genomic variant, can be discovered through GWAS studies. As explained in chapter 5.2, every clinicogenomic association identified through GWAS has three different data category i.e. variation identifier (e.g. rs number and allele), clinical condition (e.g. disease) and qualifiers of associations (magnitude of impact and quality of evidence).

To extract SNP-prostate cancer risk associations as the content of our knowledge base, we have developed a layered approach i.e. extraction, preprocessing, selection and assignment of evidence degree (Figure 36).

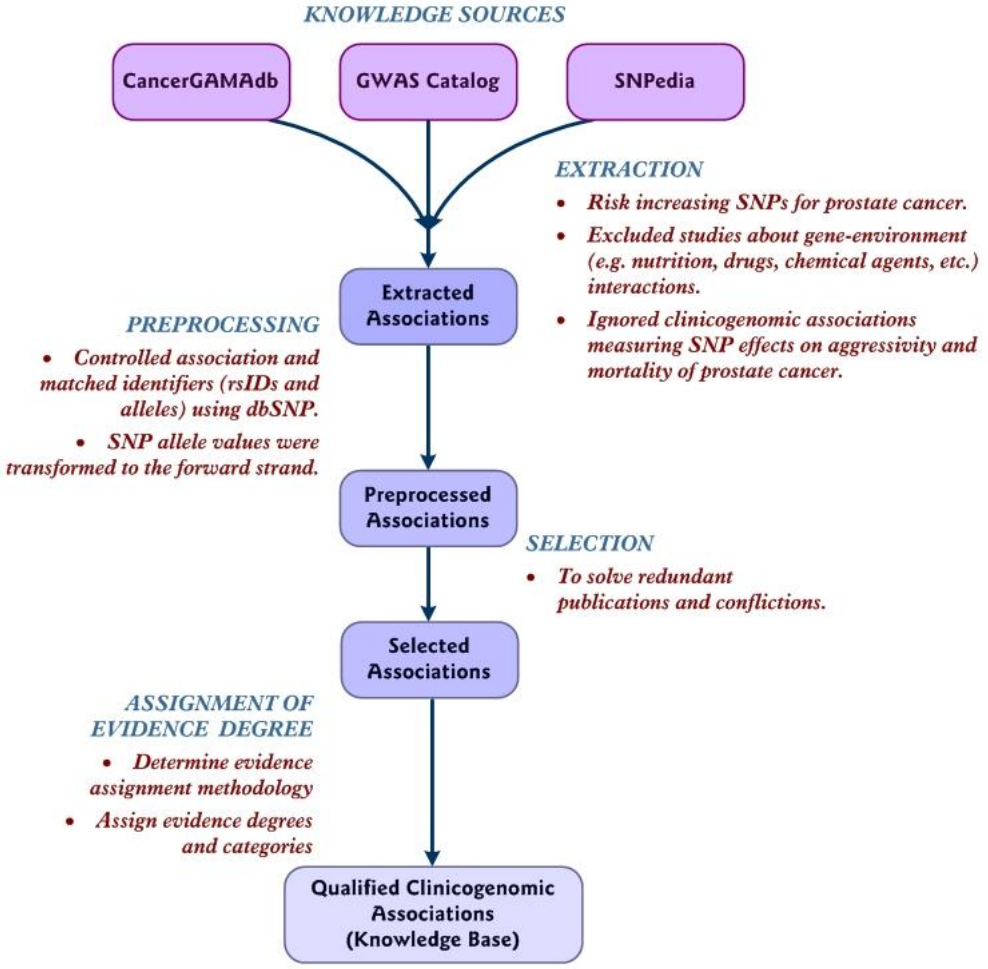


Figure 36: Associations extraction methodology for a clinicogenomic knowledge base.

6.2 Knowledge Sources

Since the completion of HGP, SNP-disease relationships have been extensively investigated and published in the medical literature. Results of these studies are mostly collected in various clinicogenomic knowledge sources in structured or narrative forms. To develop a structured clinicogenomic knowledge base for prostate cancer risk assessment, we need to determine reliable medical sources and collect clinicogenomic associations in a standardized form.

In our study, to extract these type of associations, we have preferred to utilize some publicly available knowledge sources i.e. The National Human Genome Research Institute (NHGRI) Catalog of Published Genome-Wide Association Studies (GWAS Catalog), Cancer Genome-wide Association and Meta Analyses Database (Cancer GAMAdb) and SNPedia. Cancer GAMAdb is a part of Cancer Genomic Evidence-based Medicine Knowledge Base (Cancer GEM KB) and provides GWAS researches and meta-analysis about clinicogenomic associations (http://www.hugenavigator.net/CancerGEMKB/caIntegratorStartPage.do) (Schully, et al., 2011). GWAS Catalog provides 1751 curated publications of 11,912 SNPs (Welter, et al., 2014). SNPedia (http://www.SNPedia.com) is a wiki resource of human genetic variation as published in peer-reviewed studies (Cariaso & Lennon, 2012).

6.3 Extraction and Preprocessing

We have collected clinicogenomic associations from the knowledge sources mentioned above, which are known to increase prostate cancer risk from, excluding studies about gene-environment (e.g. nutrition, drugs, chemical agents, etc.) interactions. We have also ignored clinicogenomic associations measuring SNP effects on degree and mortality of prostate cancer.

As SNP nomenclatures and notations are represented heterogeneously among different medical sources, proper unification and standardization of SNP identifiers was a critical step. We have checked all selected associations and corrected rs numbers and alleles by using dbSNP (http://www.ncbi.nlm.nih.gov/SNP) as our reference. For only one SNP rsID which had been merged to another SNP was updated, and allele values which had been identified based on reverse strand were transformed to the forward strand.

6.4 Qualifiers for Clinicogenomic Associations

For published clinicogenomic associations, the magnitude of impact and quality of evidence were critical to qualify the associations (Attia, et al., 2009B), (Van Allen, et al., 2013).

6.4.1 Magnitude of Impact

In clinical practices, absolute risk value of genetic variations are important but, in most of the disease-variation researches, absolute risk cannot be calculated due to lack of information about disease incidence (Janssens & van Duijn, 2009). As a measurement of impact value for clinicogenomic associations, researchers usually prefer to use conventional approaches, e.g. odds ratios and relative risks for case control studies and cohort studies respectively. These risks are presented with a confidence interval (Attia, et al., 2009C) (Figure 37).

	Disease (+)	Disease (-)
SNP- Allele (X)	Disease (+) AND SNP- Allele (X)	Disease (-) AND SNP- Allele (X)
SNP- Allele (Y)	Disease (+) AND SNP- Allele (Y)	Disease (-) AND SNP- Allele (Y)

$$\text{Odds Ratio} = \frac{[\text{Disease (+) AND SNP-Allele(X)}] * [\text{Disease (-) AND SNP-Allele (Y)}]}{[\text{Disease (+) AND SNP-Allele(Y)}] * [\text{Disease (-) AND SNP-Allele (X)}]}$$

Figure 37: Calculation of Odds ratio for disease associated SNPs.

In the literature, many SNPs with minor degree of association were published for the prediction of the prostate cancer risk. For a physician, it's impossible to interpret the every single SNP to determine appropriate clinical action.

Thus, to visualize and represent all of the personal risk SNPs in a graphical way, the magnitude of impact can be categorized. In our study, we have arbitrarily categorized the magnitude of impact (odds ratio) into three classes i.e. strong (>=2.50), moderate (>=1.50, <2.50) and weak (<1.50).

6.4.2 Quality of Evidence

In GWAS, various defects and biases about study design, genotyping or collected data quality affect the clinical value of results (Pearson & Manolio, 2008), (Attia, et al., 2009B), (Little, et al., 2009). The quality of evidence is scored based on the type of study and how well the study is conducted (Riegelman, 2010).

Most of clinicogenomic associations have small effect sizes and their credibility may largely depend on the success of control for errors and biases. Traditional epidemiological guidelines are not appropriate for the specifications for study of genetic epidemiology. To overcome this issue, a Human Genome Epidemiology Network (HuGENet) workshop organized in Venice, Italy in 2006 and the interim Venice guidelines were published to grade the cumulative evidence in genetic associations. This guideline is based on three criteria 1) the amount of evidence (sample size), 2) replication of studies (determining association in different studies), and 3) protection from bias (Ioannidis, et al., 2008) (Table 12).

Table 12: Venice interim guideline criteria for assessment of cumulative evidence on genetic associations (Ioannidis, et al., 2008)

Amount of evidence	**Category A:** Sample size >1000 **Category B:** Sample size >100 and <1000, **Category C:** Sample size <100 (total number in cases and controls assuming 1:1 ratio)
Extent of replication	**Category A:** Extensive replication including at least one well-conducted meta-analysis with little between-study inconsistency. **Category B:** Well-conducted meta-analysis with some methodological limitations or moderate between-study inconsistency. **Category C:** No association; no independent replication; failed replication; scattered studies; flawed meta-analysis or large inconsistency.
Protection from bias	**Category A:** Bias, if at all present, could affect the magnitude but probably not the presence of the association. **Category B:** No obvious bias that may affect the presence of the association but there is considerable missing information on the generation of evidence. **Category C:** Considerable potential for or demonstrable bias that can affect even the presence or absence of the association.

In genetic association studies, four kinds of biases were well recognized, which are phenotype misclassification, genotyping error, population stratification, selective reporting biases.

Large sample sizes tend to decrease the uncertainty for the impact values of the clinicogenomic associations. Also larger studies usually conducted by more experienced groups, so supposed to be affected less from selective reporting biases. The critical thresholds for sample size were determined as 100 and 1.000 regarding operational characteristics in Venice criteria.

Meta-analyses provide information on variation in effects across populations, and on using different methods. Lack of replication may be a marker for underlying biases and extensive replication may provide optimal evidence for clinicogenomic association. Also, the threshold of replication is a matter of considerable debate.

Regarding eliminating bias, investigators should assess studies regarding major biases e.g. phenotype, genotype, population stratification and selective reporting. These cover the two variables involved in the association, study specific confounding and field-wide bias.

After evaluation of a study, all considerations is categorized as A, B and C and finally, merged as a composite assessment using a semi quantitative index i.e. strong, moderate, and weak epidemiological credibility for genetic associations (Ioannidis, et al., 2008).

Today, Venice criteria are used to assess genomic association studies in several controlled and structured knowledge bases e.g. Alz-Gene, PD-Gene, SZ-Gene. For other various clinicogenomic association knowledge sources, different types of approaches are proposed to identify evidence quality. In ClinVar, evidence for clinical significance is categorized regarding study count and type such as in vitro studies, animal models, etc.

(http://www.ncbi.nlm.nih.gov/clinvar/intro). But for ClinVar, there is not a quantitative categorization of evidence quality.

In SNPedia, magnitude is constructed as a subjective measure of interest for magnitude of impact and repute (good, bad) for quality of evidence, but these concepts are not well established (http://snpedia.com/index.php/Genotype).

GET-Evidence (http://evidence.personalgenomes.org/about) categorizes clinicogenomic references according to their evidence degree (high, moderate, or low) and clinical significance (high, medium, or low) are used to produce impact score (Ball, et al., 2012). But evidence degree is not assigned in allele level. Also, this source is not sufficient for our study because only two SNP from our knowledge base is defined i.e. two associations as "Low clinical importance, Uncertain pathogenic", and 38 associations as "Insufficiently evaluated pathogenic").

6.5 Selection of Clinicogenomic Associations

In GWAS knowledge sources, generally there is multiple odds ratio for per clinicogenomic associations depending on the diversity of studies. To solve these value redundancies and confliction, there were two possible strategies, first merging all evidences and second selecting most compatible associations.

In our study, we preferred to follow the selection strategy because "merging" would need more time, domain expertise and cost. For our clinicogenomic association set, we have developed a four-phased selection approach to determine a reasonable value per SNP allele. Our selection approach is mostly inspired from Venice criteria.

In the first step, because all test data was gathered from Caucasians, we obtained have gathered the clinicogenomic association values from studies which were performed on this ethnic group. If there was not any study with Caucasians population, next we have preferred to use the mixed population results as second choice and results from other populations (Africans, Asians, etc.) as the last choice. This criterion was about avoiding population stratification bias. In the second step, we have assessed the study type and preferred meta-analysis results to research studies. Next, if we still had more than one association value, we evaluated the number of citation for the referenced article. Last two steps were adapted to be able to integrate credible studies and associations with replicated studies to reduce bias.
In the final step, we selected biggest odds ratio, when still required (Table 13).

Table 13: Selection criteria for extracted associations

Step	Category	Order of preference
1	Race and ethnicity	1. Caucasians 2. Mixed 3. Other races (Africans, Asians, etc.).
2	Study type	1. Meta-analysis 2. Research study
3	Credibility of journal	Highest citation number
4	Odds ratio	Highest value.

With this approach, we have extracted one odds ratio for every single clinicogenomic association in the knowledge sources.

6.6 Assignment of Evidence Degree

We have assigned a degree of evidence quality for all the association values, in order to indicate possible biases and faults in the results of genetic association studies. Ideally, it's preferred to evaluate association values regarding all bias sources (study design, genotyping problems, publication bias, etc.) of studies, but it is time consuming, costly and requires professional domain expertise. As the main objective of our study was to compare to different predictive model structures for prostate cancer regarding the clinical utility, we have generated a simpler approach using some indirect metrics to acquire a quality of evidence degree for every association.

In our approach, we have studied Venice criteria and extracted various indirect metrics to assign an evidence degree using PubMed publications and our knowledge sources. This method has a potential for automated evidence degree assignment. Matching of our parameters and Venice criteria is presented in Table 14.

Table 14: Comparison of selected criteria and Venice criteria for determining degree of evidence

Proposed Criteria	**Venice Criteria**		
	Amount of evidence (Sample size)	**Extent of replication**	**Protection from bias**
Citation number of article		✓	✓
Type of study and number of authors		✓ (Research article and Meta-analysis)	✓ (Number of Author)
Race and ethnicity of studied population			✓
Sample size (each of case and controls)	✓		
Number of article for SNP-prostate cancer relationship in PubMed		✓	✓
Number of cumulative models which involve SNP allele		✓	✓

In our approach, we used three main dimensions i.e. credibility of referenced article, reliability of the study and the scientific familiarity of SNP-disease relationship. To calculate the credibility of referenced article, "citation number of the article" and "type of study and number of authors" were used. Reliability of the studies were assessed by determination of race and ethnicity status and sample size (number of cases and controls). To evaluate the scientific familiarity of SNP-clinical condition relationship, we calculated the number of the scientific articles about SNP-prostate cancer relationship in PubMed and number of cumulative models which involve SNP allele. These criteria are also summarized in Table 15.

Table 15: Suggested model to assign the evidence degree for clinicogenomic associations.

Dimension	Order of preference	Value
Credibility of referenced article	Citation number of article	1-15=1 16-50=2 >50=3
	Type of study and number of authors	Research article-<10, author=1 Research article-<35, >=10 author=2 Research article->=35, author=3 Meta-analysis- <7, author=2 Meta-analysis->=7, author=3
Reliability of study	Race and ethnicity of studied population	Other races (Africans, Asians, etc.)=1 Mixed=2 Caucasians=3
	Sample size (each of case and controls)	Unknown or <100=1 100-1000=2 >1000=3
Scientific familiarity of SNP-disease relationship	Number of article for SNP-prostate cancer relationship in PubMed	<10=1 >=11, <39=2 >=80=3
	Number of cumulative models which involve SNP allele	None=1 1-2 model=2 >2 model=3
Degree of evidence quality	=Total value/6	1-<1.5= weak >=1.5-<2.3= moderate >=2.3= strong

We used two measurement parameters, namely "citation number of article" and "type of study and number of authors" to evaluate the "credibility of referenced article". Results for "citation number of article" are in Table 16. We have clustered citation numbers as three groups using k-means clustering algorithm and assigned values for these groups.

Table 16: Assigning value for parameter of "citation number"

Citation count	Number of cited articles in PubMed	"Citation number" Value
0	8	1
1	2	1
2	4	1
3	5	1
4	2	1
5	2	1
6	1	1
7	1	1
9	1	1
10	1	1

Table 16 (cont.): Assigning value for parameter of "citation number"

16	1	2
17	1	2
19	3	2
23	2	2
27	1	2
29	1	2
39	1	2
82	1	3
88	1	3
≥100	4	3

The "type of study and number of authors" is another critical factor. The meta-analysis publications are more credible from research articles, because it means that there are some different articles about same SNP-disease association values. For this reason, we categorized this parameter as two major category i.e. "research article" and "meta-analysis" and assigned arbitrary values regarding number of author (Table 17).

Table 17: Assigning value for parameter of "type of study and number of authors".

Type	**#of authors**	**# of articles**	**Assigned value**
Meta-analysis	3	28	2
Meta-analysis	4	16	2
Meta-analysis	5	8	2
Meta-analysis	6	4	2
Meta-analysis	8	2	3
Meta-analysis	9	46	3
Meta-analysis	10	10	3
Meta-analysis	11	4	3
Meta-analysis	12	2	3
Meta-analysis	13	2	3
Meta-analysis	15	2	3
Meta-analysis	17	4	3
Meta-analysis	25	4	3
Meta-analysis	100	4	3
Research article	6	2	1
Research article	8	2	1
Research article	9	2	1
Research article	13	12	2
Research article	20	1	2
Research article	31	12	2
Research article	39	4	3
Research article	49	4	3
Research article	61	17	3
Research article	70	4	3
Research article	≥100	26	3

For the reliability of the study, we have assessed “race and ethnicity status” and “sample size (number of cases and controls)”.In value assignment for “race and ethnicity of studied population”, as all of our test data were Caucasians, we assigned 3 point for studies performed on Caucasians, 2 point for studies about mixed populations including Caucasians and 1 point for other ethnicities (Table 18).

Table 18: Assigning value for parameter of “race and ethnicity of studied population”.

Ethnicity	Number of study	Category	Assigned value
African	1	Other	1
African-American	2	Other	1
Asian	7	Other	1
Japanese	3	Other	1
Japanese, Latin American	2	Other	1
General	22	Mixed	2
American White	2	Caucasian	3
Caucasian	38	Caucasian	3
European	24	Caucasian	3
European White	1	Caucasian	3
UK, Australian	1	Caucasian	3
White	3	Caucasian	3

For the parameter of “sample size”, we used same thresholds with Venice criteria, namely 1 point for studies where sample size was unknown or smaller than 100, 2 points for sample size between 100 and 1000 and 3 points for more 1000 individuals (Table 19).

Table 19: Assigning value for parameter of “sample size”.

Sample size	Number of cases and controls (both of)	Assigned value
Unknown or <100	22	1
100-1000	7	2
>1000	77	3

To evaluate the scientific familiarity of SNP-clinical condition relationship, we have calculated the “number of the scientific articles about SNP-prostate cancer relationship in PubMed” and “number of cumulative models which involve SNP allele”.

The “number of article for SNP-prostate cancer relationship in PubMed” directly point out, there are more than one studies about same SNP value. Results for this parameter are in

Table 20. We have clustered citation numbers as three groups using k-means clustering algorithm and assigned numeric values for these groups.

Table 20: Assigning value for parameter of "number of article for SNP-prostate cancer relationship in PubMed"

Number of articles for SNP-prostate cancer relationship in PubMed	Number of SNP-prostate cancer relationships	Assigned value
0	12	1
1	2	1
2	7	1
3	6	1
4	2	1
5	15	1
6	1	1
7	1	1
9	1	1
10	2	1
16	8	2
17	19	2
19	5	2
23	9	2
27	1	2
29	1	2
39	4	2
82	1	3
84	1	3
≥100	8	3

On the other hand, the "number of cumulative models which involve SNP allele" is an indirect parameter pointing more than one studies about same SNP value. The threshold values for this parameters are determined arbitrarily (Table 21).

Table 21: Assigning value for parameter of "number of cumulative models which involve SNP allele"

#of involved cumulative model for prostate cancer	#of SNP	Assigned value
0	89	1
1	9	2
2	3	2
3	1	3
4	2	3
≥5	2	3

Finally, the degree of evidence quality was calculated as the arithmetic average for each association. To visualize all of personal risk SNPs as a whole, we have categorized evidence values of associations in three classes like magnitude of impact. Evidence degree of clinicogenomic associations categorized as strong, moderate and weak (Table 15). It's clear that, as our approach is mostly based on literature survey, it is open to debate and further comparative studies are needed.

6.7 Overview of the Clinicogenomic Associations

Initially, we have determined 87 SNP alleles from the GWAS catalog, 32 SNP alleles from SNPedia and 236 SNP alleles from the Cancer GAMAdb, which are associated with increased prostate cancer risk. Through the extraction and selection processes of SNP-prostate cancer risk associations, we have excluded redundant, conflicted and incomplete associations. Finally, we have acquired 209 independent associations for increased risk of prostate cancer from the knowledge sources. Next, the evidence and impact categories to these associations are assigned (Appendix A).

To complete assessment of all different types of clinicogenomic associations, we have extracted counts of clinically relevant SNP alleles in terms of evidence and impact categories in Table 22.

Table 22: Distribution of clinicogenomic associations according to evidence degree

	Evidence Degree			
Impact Degree	**Strong**	**Moderate**	**Weak**	**Total**
Strong		5	2	**7**
Moderate		3	1	**4**
Weak	42	123	33	**198**
Total	**42**	**131**	**36**	**209**

CHAPTER 7

CLINICOGENOMIC ASSESSMENT AND REPORTING APPROACHES FOR PROSTATE CANCER

In literature, there are various types of risk assessment and reporting approaches e.g. collective reporting, model based risk assessment, polygenic risk scoring combining genomic and/or environmental data (Figure 38).

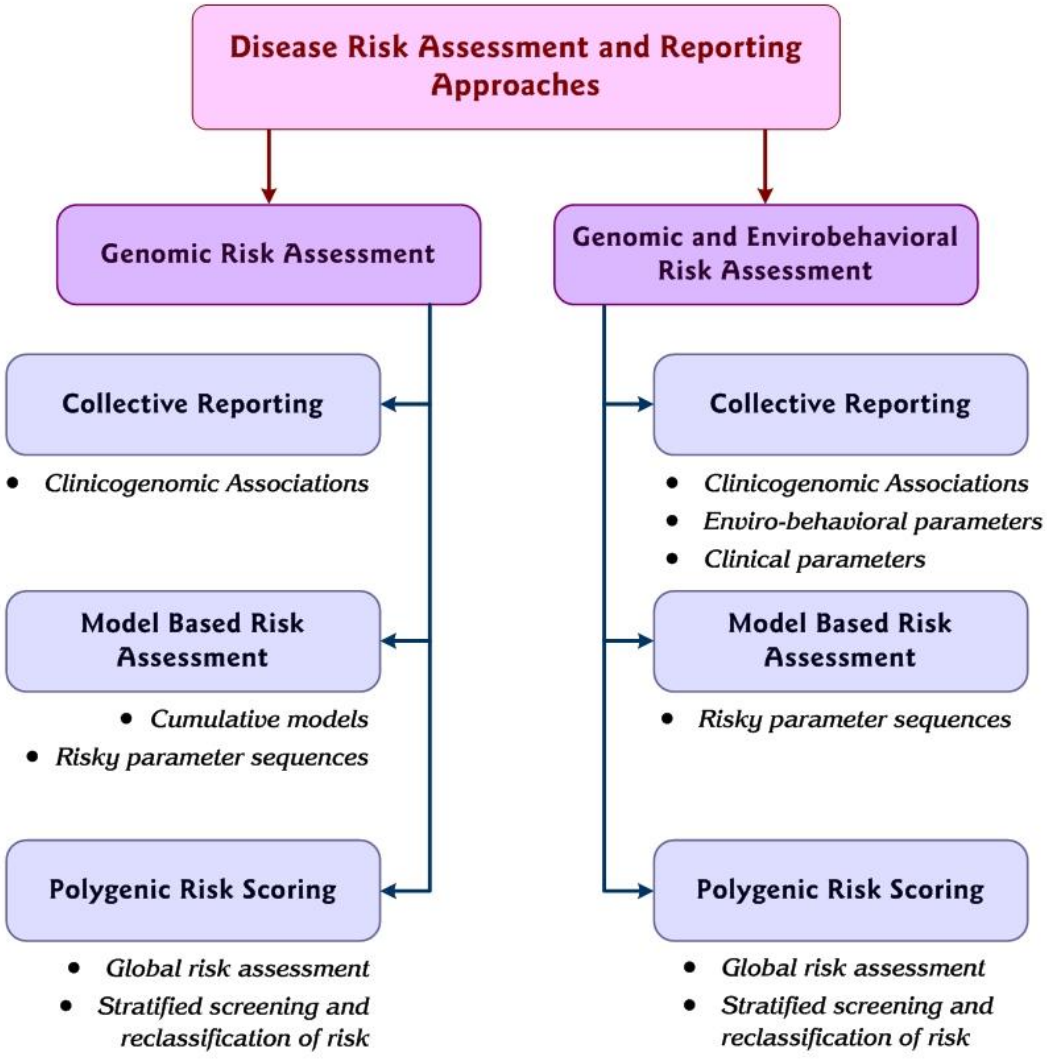


Figure 38: Different approaches to evaluate disease risk based on clinicogenomic associations.

7.1 Genomic Based Risk Assessment and Reporting Approaches

With the advent of NGS technologies, it's also possible to accomplish rapid and cheap whole genome sequencing (WGS). Researchers and clinicians expect that WGS will be one of the most important tools in personalized medicine age (Berg, et al., 2011), (Scheuner, et al., 2009), (Wright, et al., 2011). SNPs are about 90% of all the genomic variations. Although most are harmless, some of them have great values for disease risk assessment, and can be utilized in medical diagnostics and pharmaceutical products (Poo, et al., 2011), (Aronson, et al., 2012), (Drmanac, 2012).

With genome-wide association studies (GWAS), numerous SNPs can be recognized and examined for their associations regarding the pathogenesis of complex diseases (Pearson & Manolio, 2008). Following the GWAS and the filtering processes, hundreds to thousands of clinically relevant variations, which have the potential to be the underlying reason, could be extracted (Bamshad, et al., 2011), (Bick & Dimmock, 2011), (Biesecker, 2012), (Raffan & Semple, 2011).

In medical care processes, genomic data and its derivatives can be used on risk assessment, to predict disease susceptibility, targeted screening, clinical diagnosis, to predict the course of the disease, and to create a treatment plan and follow-up (Ginsburg & Willard, 2009), (Chan & Ginsburg, 2011).Today various methods and approaches are improving to utilize clinicogenomic associations e.g. collective reporting, model based approaches and polygenic risk scoring of personal genomic profile.

7.1.1 Collective Reporting of Genomic Risk Parameters

Independent associations and their effects as risk factors can be present one by one. If we report a limited number of independent associations, this approach may be useful. Especially, for diseased associated SNPs with strong impact and strong evidence can be shared by users one by one. At this point, using carefully chosen graphics and visualization techniques will be an efficient way.

In Partners Healthcare Center for Personalized Genetic Medicine, independent associations are reported one by one (Figure 39).

Also, various DTC companies report personal genomic risk using graphics. These graphics contain personal information representing relative risk and providing estimates of disease prevalence as a reference point (Figure 40).

LABORATORY FOR MOLECULAR MEDICINE
65 Landsdowne St., Cambridge, MA 02139
PHONE: (617) 768-8500 / FAX: (617) 768-8513
http://pcpgm.partners.org/lmm

PARTNERS HEALTHCARE

CENTER FOR PERSONALIZED GENETIC MEDICINE

A teaching affiliate of: HARVARD MEDICAL SCHOOL

Name: John Doe

DOB: 01/23/45
Sex: Male
Race: Caucasian

Accession ID: 0123456789
Specimen: Blood, Peripheral
Received: 01/23/45

Family #: F12345
Referring physician: John Smith, M.D.
Referring facility: Double Helix Hospital

SAMPLE GENERAL GENOME REPORT SAMPLE

Sequencing of this individual's genome was performed and covered 98.2% of all positions at 8X or higher, resulting in over 3.6 million variants compared to a reference genome. These data were analyzed to identify previously reported variants of potential clinical relevance as well as novel variants that could reasonably be assumed to cause disease (see methodology below). All results are summarized on page 1 with further details provided on subsequent pages.

RESULT SUMMARY

A. MONOGENIC DISEASE RISK: 1 VARIANT IDENTIFIED

This test identified 1 genetic variant that may be responsible for existing disease or the development of disease in this individual's lifetime.

Disease (Inheritance)	Phenotype	Gene Variant	Classification
A1. Episodic ataxia type II (Autosomal Dominant)	Poor coordination and balance	CACNA1A p.Arg2156GlyfsX32	Pathogenic

B. CARRIER RISK: 3 VARIANTS IDENTIFIED

This test identified carrier status for 3 autosomal recessive disorders.

Disease	Phenotype	Gene Variant	Classification	Carrier Phenotype*
B1. Cystic fibrosis	Chronic lung and digestive disease	CFTR c.1585-1G>A	Pathogenic	Infertility (moderate evidence)
B2. Myotonia congenita	Muscle disease	CLCN1 p.Arg894X	Pathogenic	Latent myotonia (case report only)
B3. Usher syndrome type II	Hearing loss and retinitis pigmentosa	USH2A p.Gly204ArgfsX12	Pathogenic	None reported

As a carrier for recessive genetic variants, this individual is at higher risk for having a child with one or more of these highly penetrant disorders. To determine the risk for this individual's children to be affected, the partner of this individual would also need to be tested for these variants. Other biologically related family members may also be carriers of these variants.*Carriers for some recessive disorders may be at risk for certain mild phenotypes. Please see variant descriptions for more information.

C. PHARMACOGENOMIC ASSOCIATIONS

This test identified the following variants associated with drug use and dosing. Additional pharmacogenomic results may be requested, but will require additional molecular confirmation prior to disclosure.

Drug	Risk and Dosing Information
C1. Warfarin	Decreased dose requirement.
C2. Clopidogrel	Typical risk of bleeding and cardiovascular events.
C3. Digoxin	Increased serum concentration of digoxin.
C4. Metformin	Typical glycemic response to metformin.
C5. Simvastatin	Lower risk of simvastatin-related myopathy.

D. BLOOD GROUPS

This test identified the ABO Rh blood type as O positive. Additional blood group information is available at the end of the report.

It should be noted that the disease risk section of this report is limited only to variants with evidence for causing highly penetrant disease, or contributing to highly penetrant disease in a recessive manner. Not all variants identified have been analyzed, and not all regions of the genome have been adequately sequenced. These results should be interpreted in the context of the patient's medical evaluation, family history, and racial/ethnic background. Please note that variant classification and/or interpretation may change over time if more information becomes available. For questions about this report, please contact the Genome Resource Center at GRC@partners.org.

Figure 39: A sample general genome report.

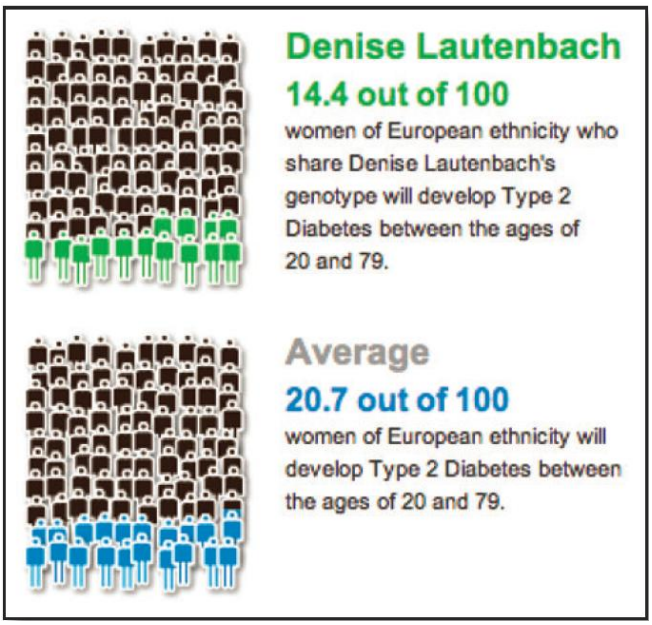


Figure 40: Sample pictographs from 23andMe results (Lautenbach, et al., 2013).

But, most of the clinically relevant SNPs have minor effect (Odds Ratio <1.50-2.00) and there are only limited number of different examples (Stranger, et al., 2011), (Kalf, et al., 2013). Also an individual can have more than 1,000,000 genetic variants (Starren, et al., 2012). Although the simplest reporting way of SNP variations is displaying these numerous variations in laboratory report, clinicians cannot interpret or evaluate these information stack. The volume of variation data integrated into clinical practice exceeds the boundaries of unsupported human cognition and interpretive capacity. Additionally the rapidly growing literature about clinicogenomic associations makes it more complicated to stay current for even the professionals (Masys, et al., 2012). For this reason, we need more sofisticated and improved approaches e.g. polygenic risk scoring and other types of model based approaches etc.

In our study, we have developed our system to report all disease associated SNPs individually with their magnitude of impact and quality of evidence. In evaluation step, we assess all cases and controls regarding amounts and qualifications of these clinicogenomic associations.

7.1.2 Genomic Risk Models

Despite the small impact degree of single clinicogenomic association, the combinations of various SNP alleles may be declarative in the pathogenesis of diseases. Some investigators attempt to improve models and multipanels assigning values for various SNP alleles and estimates entire risk of disease for more effective risk prediction (Manolio, 2010).

To combine these type of clinicogenomic associations, cumulative model is mainly used, but several different techniques also exist e.g. hybrid model based risky parameter sequences.

7.1.2.1 Cumulative Models

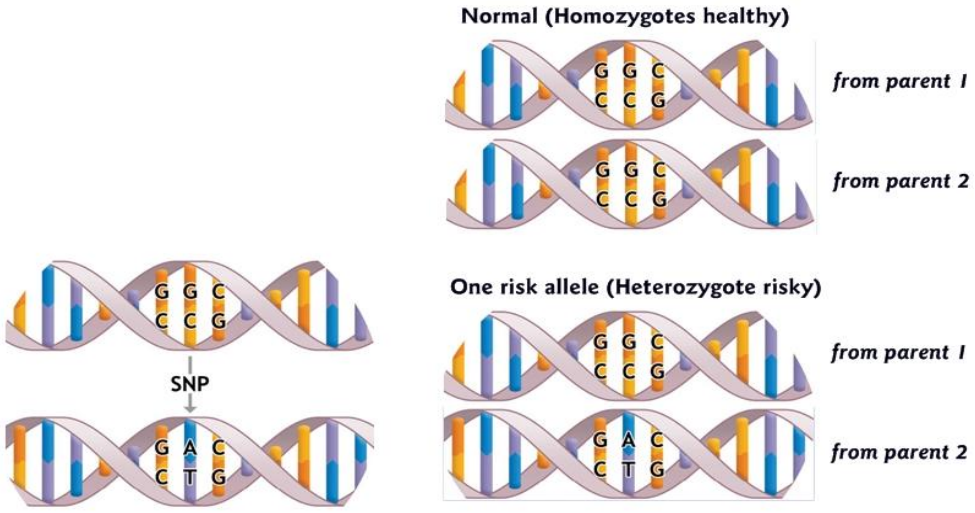
In the literature, several cumulative prediction models have been proposed but most of these are criticized regarding comprehensive evaluation especially for the lack of clinical utility (Janssens & van Duijn, 2009), (Little, et al., 2012).

We have extracted few of the cumulative predictive models from the scientific literature. For our study, we have checked rs number and allele values of climicogenomic associations and adapted to forward DNA strand using dbSNP. Models which involve additional external parameters e.g. family history were also determined. SNP alleles and genetic models of cumulative models are summarized in Table 23.

Table 23: Some cumulative risk prediction models for prostate cancer.

Chr	**rs number**	**Risk allele**	**17-SNP_Helfand** (Helfand, et al., 2011)	**9-SNP_Helfand** (Helfand, et al., 2010)	**5-SNP_Zheng** (Zheng, et al., 2008)	**5-SNP_Salinas** (Salinas, et al., 2009)	**4-SNP_Nam** (Nam, et al., 2009)	**3-SNP_Beuten** (Beuten, et al., 2009)
1	rs1819698	T						D
2	rs2710646	A		R				
2	rs721048	A	R					
3	rs10934853	A	D					
5	rs2736098	A	R					
5	rs401681	C	D					
6	rs1800629	A					D	
7	rs2348763	A					R	
8	rs1447295	A	D	D	D	D	D	
8	rs16901979	A	D	D	D			
8	rs16902094	G	D					
8	rs445114	T	D					
8	rs6983267	G	D	D	D	D		
8	rs6983561	C				D		
10	rs10993994	T	R	R				
11	rs10896450	G	D	D				
11	rs11228565	A	D					
15	rs10459592	T						
15	rs12439137	G						D
15	rs2470152	T						D
17	rs11649743	G	R					
17	rs1859962	G	R	R	R	R	R	
17	rs4430796	A	D	D	R	R		
19	rs8102476	C	D					
20	rs3787554	A						
23(X)	rs5945572	A	D	D				
External data		**Family History**	X		X	X		
Chr: Chromosome, D: Dominant, R: Recessive, X: Exist.								

Cumulative models can be considered as unification of impact of several clinicogenomic associations using arithmetic operators. For some SNPs, only homozygote alleles are involved in model (recessive model), but mostly heterozygote SNPs (dominant model) also are parts of models. Both in dominant and recessive models, the values of risk SNPs are accepted as one unit of impact. Alterations of SNPs' impact values regarding homozygote and heterozygote alleles are defined as the additive model (Figure 41). In our cases, dominant and recessive models are remarked as in Table 23.



Genetic models	Risk groups
Dominant	Homozygotes and heterozygotes individuals
Recessive	Only homozygotes individuals
Additive	Different risk degrees for homozygotes and heterozygotes individuals

Figure 41: Schematic representation of possible genetic models.

In cumulative models, existence of each association contributes to the total. For example, in 5-SNP_Zheng model, there are five different SNPs. The genetic model is dominant for three SNPs (rs1447295-A, rs16901979-A, rs6983267-G) and recessive for others (rs1859962-G, rs4430796-A). For dominant models, homozygote and heterozygote combination of alleles are identified as a risk factor in the same degree. For recessive models, only homozygote combinations are risk factors and heterozygote combinations are accepted as harmless. Through analysis of patient's genotype, total impact values of clinicogenomic associations are determined and calculated additively. Besides the associations for five different allelic SNPs, existence of prostate cancer in family history is added as an additional impact factor.

Table 24 presents how the increased risk of prostate cancer in terms of odds ratio is represented depending on total impact value. For example if patients without family history has only one impact factor, the risk of having prostate cancer increases by 1.5 compared to those who has none of the impact factors. If a patient has all of the five risk SNPs with specified alleles and a positive family history for prostate cancer, total impact is calculated as 6. According to the Table 24 this would correspond to an increased risk of 9.46 for having prostate cancer when compared to the general population.

Table 24: Reference table for 5-SNP_Zheng model (Zheng, et al., 2008).

Total impact	Odds ratio (without Family History)	Odds ratio (with Family History)
0	1.00 (by definition)	1.00 (by definition)
1	1.50 (CI: 1.18-1.92)	1.62 (CI: 1.27-2.08)
2	1.96 (CI: 1.54-2.49)	2.07 (CI: 1.62-2.64)
3	2.21 (CI: 1.70-2.89)	2.71 (CI: 2.08-3.53)
4	4.47 (CI: 2.93-6.80)	4.76 (CI: 3.31-6.84)
5	4.47 (CI: 2.93-6.80)	9.46 (CI: 3.62-24.72)
6		9.46 (CI: 3.62-24.72)

We prepared a reference tables for all models containing total impact of involved parameters and corresponding risk values. Full reference table for all cumulative models is in Appendix B.

In this study, we evaluated all our cases and controls regarding determined six predictive cumulative models and a hybrid based risky parameter sequences. Then the success of the results were interpreted and discussed.

7.1.2.2 Evidence based Probabilistic Models

Among risk assessment tools besides cumulative models, there are other ongoing efforts utilizing different data mining algorithms to interpret GWAS data for building various predictive models.

In order to present how these modeling approaches could be implemented in our prototype system, we also included such an examples into our study. This example is based on the works of "Yücebaş and Aydın Son" to assess prostate cancer risk and was developed through a hybrid approach combining Support Vector Machine (SVM) and ID3 decision tree based on "A Multiethnic Genome-wide Scan of Prostate Cancer" data set from dbGaP database (study accession no: phs000306 and version 2). The authors developed two kinds of model in thi study i.e. first (only SNP) model and second (combined SNP and envirobehavioral) model.

First hybrid model (only SNP model) includes 33 SNPs and their alleles and the accuracy, precision, and recall values of this model are %71.6, %72.69 and %68.96 respectively.

It’s possible to represent this kind of model using different approaches. In our study, we identified each branch of decision tree corresponding prostate cancer as an association set. When we transformed “Yücebaş-Aydın Son” first hybrid model (only SNP model), we captured 154 different association sets containing the combination of several different SNPs and alleles. Total list of decision tree and association sets is in Appendix C and D.

If an individual suitable for all parameters of one branch (i.e. association set), this individual has a prostate cancer risk with the accuracy and precision of complete model. Table 25 presents an example of the reference table for association sets of first hybrid model.

Table 25: Reference table for Yücebaş and Aydın Son model (Yücebaş & Aydın Son, 2014).

Branch_id	Total count of SNPs	Result
Branch_ 1	4	Prostate cancer risk (Accuracy: %71.6; Precision: %72.69; Recall: %68.96)
Branch_ 2	4	
Branch_ 3	7	
Branch_ 4	9	
….	….	
Branch_ 5	2	

The second (combined SNP and envirobehavioral) model was explained in 7.2.2.

7.1.3 Polygenic Risk Scoring

The dominant paradigm in human complex-trait genetics has been to map loci affecting disease risk and then to identify the causative mutations. With new technologies, SNPs could be used to produce a “genomic profile” for disease risk prediction testing hundreds of thousands of loci across the personal genome (Wray, et al., 2007), (Evans, et al., 2009).

Today, most of the SNP based risk assessment models have limited predictive utility and discriminative accuracy because most of the disease associated SNPs have small impacts (Evans, et al., 2009). It has been suggested that, genomic risk scores based on large numbers of SNPs could explain more heritability than models based on a small number and rigorously validated SNPs. But it’s required to process large data sets to build such a discriminative risk assessment models (Jostins & Barrett, 2011), (Wu, et al., 2013).

Genetic architecture of a disease refers to the number of genetic polymorphisms that affect risk of disease, the distribution of their allelic frequencies, the distribution of their effect sizes and their genetic mode of action (additive, dominant and/or epistatic). Prediction of genetic risk depends on the underlying genetic architecture. Indeed, the SNPs do not have to be the causative mutations. They just need to be in high linked disequilibrium with the causative mutations so that there is a consistent association between the SNP and disease risk (Wray, et al., 2008).

7.1.3.1. Global Risk Assessment

The number of possible genotype combinations exponentially increase with the number of contained variations (Janssens & van Duijn, 2009). A person has three different risk allele combinations for every locus i.e. homozygote healthy (zero risk allele), heterozygote risky (one risk allele), and homozygote risky (two risk alleles). For example, in a prostate cancer case with thirty-one risk loci, the number of possible risk alleles can be ranged from zero (all alleles healthy) to 62 (all alleles risky) and there are 3^{31} distinct possible combinations of these 31 alleles (Pashayan & Pharoah, 2012). For this reason, different type of polygenic prediction models were developed to combine the impact of disease associated SNP data e.g. count method, log-odds method, multiplicative model etc.

Count method is the calculation of total count of independent genomic risk alleles. In this method disease risk (N_{risk});

$$N\,(\text{risk}) = \sum x_i \qquad \textbf{EQUATION (1)}$$

where x_i=number of risk alleles (0, 1, 2) at SNP i. This method assumes that all risk alleles contribute equally to disease risk (Evans, et al., 2009).

Another method (log odds method), sums together the natural logarithm of the allelic odds ratio for each risk allele;

$$\log(\text{risk}) = \sum x_i \log\,(OR_i) \qquad \textbf{EQUATION (2)}$$

where OR_i is the allelic odds ratio (Evans, et al., 2009).

DTC testing companies typically employ a multiplicative model to calculate life time risk in the absence of an established method for combining SNP risk estimates i.e. multiplication of odds ratios of each genotype and average population risk. This model assumes that the independent SNPs occur and behave independently (Nusbaum, et al., 2013).

Scores can be transformed into binary outcomes by defining high risk to be individuals with a score greater than a threshold, and all others as low risk. The simplest measures of classification accuracy are the sensitivity and specificity of the test. These values vary with the choice of threshold, which represents the unavoidable trade-off between sensitivity and specificity (Jostins & Barrett, 2011). Related definitions presented in Figure 42.

A plot of the sensitivity against (1-specificity) for all possible choices of thresholds is known as a Receiver Operating Characteristic (ROC) curve. The area under the ROC curve (AUC) has the pleasing property of being equal to the probability that a randomly selected individuals with the disease has a higher score than a randomly selected healthy individuals (Jostins & Barrett, 2011).

In the evaluation phase, we assessed total impact of independent associations (polygenic risk score) based on five approaches i.e. number of SNP (based on both dominant and additive models), evidence-impact-SNP degree (based on both dominant and additive models) and the weighted version of number of SNP (dominant model).

	Disease (+)	Disease (-)
Test (+)	A	C
Test (-)	B	D

Definition	Formula	Explanation
Sensitivity	A / (A+B)	Probability that a test result will be positive when the disease is present (True Positive Rate-TPR, Recall).
Specificity	D / (C+D)	Probability that a test result will be negative when the disease is not present (True Negative Rate, TNR).
Positive Likelihood Ratio (+LR)	Sensitivity / (1-Specifity)	Ratio between the probability of a positive test result given the presence of the disease and the probability of a positive test result given the absence of the disease.
Negative Likelihood Ratio (-LR)	(1-Sensitivity) / Specificity	Ratio between the probability of a negative test result given the presence of the disease and the probability of a negative test result given the absence of the disease.
Positive Predictive Value (PPV)	A / (A+C)	Probability that the disease is present when the test is positive (Precision).
Negative Predictive Value (NPV)	D / (B+D)	Probability that the disease is not present when the test is negative.
Accuracy	(A+D) / (A+B+C+D)	Proportion of true results in the population.

Figure 42: Concepts and terms about binary classification of medical tests (Macaskill, et al., 2010), (Jenicek, 2013).

In the approach of "number of SNP", we calculated total count of existing relevant SNPs. In the dominant model, we only calculated the count of relevant SNPs, but in the additive model we considered the value of homozygote SNPs two times compared to heterozygote SNPs. In "evidence-impact-SNP" approach, for every existing SNP, we calculated an impact degree using evidence degrees (1, 2, and 3) and impact degrees (1, 2, and 3). Also, similar to the number of SNP calculation, for the additive model we assigned 1 and 2 to heterozygote and homozygote SNPs as coefficients respectively. In weighted method, un-analyzed SNPs were excluded.

Then, the sensitivity, specificity, PPV, NPV, ROC, and AUC, for every model was calculated with XLSTAT 2014.1.06. Finally, results for all of the methods were compared and interpreted to explain whether these tests were useful for diagnostic and/or screening purposes.

7.1.3.2. Stratified Screening and Reclassification of Disease Risk

Presently, predictive risk models based on the identified common susceptible variations have small values to assess disease risk. Recently, it's proposed that these susceptible common

variations can be used as screening tests for population level risk stratification (Pashayan & Pharoah, 2012).

Screening tests are presumptive diagnostic techniques whose purpose is not to establish a definitive diagnosis and prescribe treatment but to lead patients with positive results to a more complete diagnostic workup and evaluation and treatment if needed (Jenicek, 2013). Screening tests have high false-positive rates. Therefore, these tests are not ideal to predict given medical condition in a population and typically definitive diagnostic tests are used to precise diagnosis. The real advantage of population screening is to discover all possible cases of clinical conditions in the population (maximum sensitivity). Suspected individuals i.e. positive individuals regarding screening test, usually undergo subsequent procedures, interventions, and tests (Khoury, et al., 2013).

In genomic medicine, risk-stratified population screening can be applied as only polygenic risk profiling or combined with conventional risk factors (e.g. race, age, family history, etc.). By this approach, reclassifying risk, standard public health interventions could be applied more effective than conventional screening to each population stratum (Chowdhury, et al., 2013), (Dent T, et al., 2013), (Manolio, 2013), (Pashayan, et al., 2013).

In our study, ages of all samples were in the same range. Ethnicity of cases and controls were Caucasians. Family history data of samples were not exist. Therefore, we couldn't evaluate our data for stratified screening based on polygenic risk profiling and combined with conventional risk. We attempt to develop an example approach using genomic and envirobehavioral features in section 6.2.3.

7.2 Combined Genomic and Enviro-behavioral Risk Assessment

Common medical conditions e.g. heart disease, diabetes, schizophrenia, many types of cancers, and obesity are complex and multifactorial conditions which are caused by combination of multiple genes, lifestyle and environmental components (Janssens & van Duijn, 2008).

Genomic variations are very common in population. Each common variation may play a low role in the pathogenesis of complex diseases, but collectively all variations may be a strong reason of these. In the presence of specific variation patterns, with the involvement of environmental and behavioral causes clinical conditions may be triggered. In such cases, if people with risky genotype patterns avoid from some environmental agents, they can prevent themselves from possible manifest clinical conditions (National Cancer Institute, 2013A).

In disease risk assessment for common complex disease, it may be helpful to analysis the existence of traditional risk factors for specific types of medical conditions (e.g., family history, diet, physical activity etc.) by genetic rick score (Liu & Song, 2010).

Today, emerging technologies facilitated to collect different types of enviro-behavioral risk factors. Currently, smartphone-based mobile applications, pre-programmed questionnaires, wearable electronics and multi-sensor platforms (e.g. smart watches, wristband sensors, wearable sensor patches, artificial reality-augmented glasses, brain computer interfaces, wearable body metric textiles) can be used to collect personal physiological and psychological data (Paddock, 2013), (Swan, 2012). Today, most favorite mobile health applications are developed to track diet, exercises and weight management (Fox & Duggan, 2012).

Recently, there are growing the number of sensors and applications to monitor several physical, chemical and biological agents e.g. temperature, barometric pressure, humidity, air quality, carbon monoxide, radiation levels, airborne contaminants etc. Some types of these sensors are used to record and share data in real-time. Also, it's possible to track movements of individuals using mobile phones, and geolocation data can be used, for example, to help determine exposure e.g. air pollutants (Van Tongeren & Cherrie, 2012).

Eventually, various approaches and models are developed combining genomic and enviro-behavioral risk factors to assess possible risks for several common complex diseases.

In the literature, in addition to genetic factors, several diseases, sociodemographic characteristics, environmental and behavioral exposures are proposed as confounders of prostate cancer (Table 26). Therefore, we analyzed personal clinical and environmental characteristics using possible disease risk prediction methods which are meaningful for prostate cancer.

Table 26: List of various risk and protective factors for prostate cancer. (National Cancer Institute, 2013B), (Sartor, 2013)

Sociodemographic Data	
Age, Family health history, Ethnicity, Race	
Environmental Sources	
Nutrition and diet	Animal fat, fruits, legumes, yellow-orange and cruciferous vegetables, soy foods, dairy products, fatty fish, alcohol, coffee, green tea, modified citrus pectin, pomegranate.
Supplements	Multivitamins, supplement containing products (vitamin E -with or without selenium, folic acid, zinc, calcium, vitamin D, retinoid), zyflamend.
Drugs	5 alpha-reductase inhibitors, Non-steroidal anti-inflammatory drugs, statins, toremifene.
Medical procedures	Vasectomy, barium enema, hip or pelvis x-rays, and external beam radiation therapy for rectal cancer
Tobacco use	Tobacco products, smoking.
Personal Health Status (Internal Environment)	
Diagnosis	Prostatitis, prostatic intraepithelial neoplasia, syphilis, skin basal cell carcinoma, benign prostate hyperplasia (BPH), type 2 diabetes mellitus (T2DM).
Anatomic Measurements	High BMI.

7.2.1 Collective Reporting of Genomic and Envirobehavioral Parameters

The Coriell Personalized Medicine Collaborative has been used graphics to communicate genetic, environmental and lifestyle risks (Figure 43).

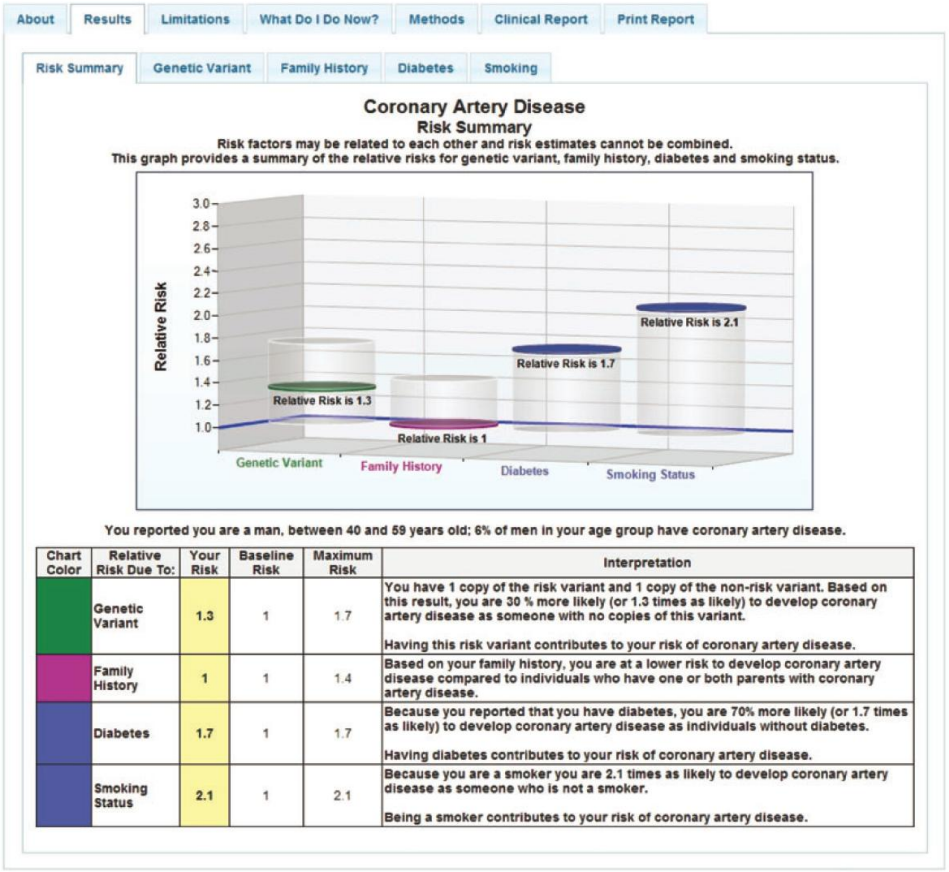


Chart Color	Relative Risk Due To:	Your Risk	Baseline Risk	Maximum Risk	Interpretation
	Genetic Variant	1.3	1	1.7	You have 1 copy of the risk variant and 1 copy of the non-risk variant. Based on this result, you are 30 % more likely (or 1.3 times as likely) to develop coronary artery disease as someone with no copies of this variant. Having this risk variant contributes to your risk of coronary artery disease.
	Family History	1	1	1.4	Based on your family history, you are at a lower risk to develop coronary artery disease compared to individuals who have one or both parents with coronary artery disease.
	Diabetes	1.7	1	1.7	Because you reported that you have diabetes, you are 70% more likely (or 1.7 times as likely) to develop coronary artery disease as individuals without diabetes. Having diabetes contributes to your risk of coronary artery disease.
	Smoking Status	2.1	1	2.1	Because you are a smoker you are 2.1 times as likely to develop coronary artery disease as someone who is not a smoker. Being a smoker contributes to your risk of coronary artery disease.

Figure 43: Sample risk summary from the Coriell Institute for Medical Research.

If we have not an approved and proven model integrating enviro-behavioral and genomic parameters as whole, it is impossible to accurately combine risks about genetic, familial, environmental and behavioral factors as one relative risk number. Because, it's not known how these risk factors interact to each other and determine overall disease risk. Therefore, the Coriell Institute for Medical Research preferred to provide risk values for each risk factor.

Before reporting risk factors, because occasionally there are conflicting scientific conclusions about effect type (risky, protective or normal), level of research evidence and impact of parameters in the literature, these scientific findings should be evaluated, curated and merged by domain experts.

During this process, the clinical impact value and quality of evidence degree should be assigned to each type of independent associations. This is a critical process because a risk factor may have a weak effect and strong evidence, but another factor may have a strong effect but weak evidence. When reporting these risk (or protective) factors as a whole, all aspects of the data need to be presented regarding the dimensions of clinical significance.

In our knowledge sources, there are several different approaches and techniques to categorize these parameters i.e. UpToDate Grading Guide (http://www.uptodate.com/home/grading-guide), NCI Levels of Evidence for Cancer Screening and Prevention Studies (http://www.cancer.gov/cancertopics/pdq/screening/levels-of-evidence/HealthProfessional). These grading methods should be compatible with each to be used as information sources in a single system.

To visualize all independent uniparametric associations as a whole, we can use scatter graphs where the axes corresponds to categories (low, medium, and high) of impact and quality (Figure 44).

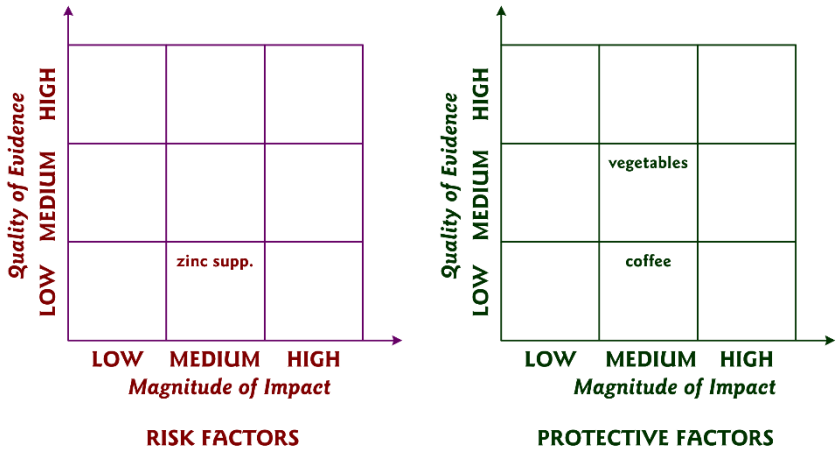


Figure 44: A graphical visualization of complete environmental parameters for an example case.

7.2.2 Envirogenomic Risk Models

Today, various researches carry out studies to develop envirogenomic risk models. A statistical approach, and software were developed using envirogenomic parameters and determining individual disease risk (Crouch, et al., 2013). Using this statistical risk model and software, it's suggested that, disease risk prediction of colorectal cancer could be possible tracking and managing envirogenomic profile (selected SNPs, alcohol intake, smoking, exercise levels, BMI, fibre intake and consumption of red and processed meat) and prevention of disease could be accomplished changing risky lifestyle factors (Yarnall, et al., 2013).

Another example of genome-environmental risk assessment model is based on a study of "Yücebaş and Aydın Son". In this study, to assess prostate cancer risk and was developed through a hybrid approach combining Support Vector Machine (SVM) and ID3 decision tree based on "A Multiethnic Genome-wide Scan of Prostate Cancer" data set from dbGaP database (study accession no: phs000306 and version 2). The authors developed two kinds of model in this study i.e. first (only SNP) model and second (combined SNP and envirobehavioral) model. In the second model of Yücebaş-Aydın Son hybrid based association sets (risky parameter sequences), SNPs and BMI data was used. In this model, to calculate the risk of some individuals we need smoking and alcohol consumption data.

The second hybrid model originally was developed for African Americans and contains 23 association sets containing 28 SNPs, BMI, alcohol, and cigarette usages. The accuracy, precision, and recall values of this model for African-Americans are %93.81, %96.55 and %90.92 respectively (Yücebaş & Aydın Son, 2014).

Similar to cumulative models, to prepare these hybrid models, in first, we checked rs numbers and adapted allele values of contained SNPs to forward DNA strand using dbSNP. After that, we converted the results of hybrid models as association sets. Total list of these association sets is in Appendix E and F. Finally, we prepared a reference tables for both models containing SNP parameters.

7.2.3 Polygenic Risk Scoring with Enviro-behavioral Parameters

7.2.3.1. Global Risk Assessment

Due to the complex genetic construction of many common diseases, it's hard to explain the associations and interactions between genetic and non-genetic risk factors. Thus, developing analytic models to integrate genetic and non-genetic factors for disease risk assessment is a still a critical problem (Salari, et al., 2012), (Khoury, et al., 2013).

In a study about cocaine dependence, a genome-environmental risk assessment model was developed using 330 SNPs and nine potentially cocaine related facets of environment. Such a genome-environmental risk assessment study simultaneously considers nearly one million predictors and their possible interactions. In this study, to handle such a large amount of data, a newly developed receiver operating characteristic approach i.e. tree-assembling ROC (TA-ROC), was used (Wei, et al., 2012).

In our study, we proposed a simple demonstrative model for evaluation phase. This model is needed to evaluate with large samples.

7.2.3.2. Stratified Screening and Reclassification of Disease Risk

Risk-stratified population screening approach can be applied as combined with conventional risk factors (e.g. race, age, family history, etc.). By this approach, reclassifying risk, standard public health interventions could be applied more effective than conventional screening to each population stratum (Chowdhury, et al., 2013), (Dent T, et al., 2013), (Manolio, 2013), (Pashayan, et al., 2013).

Disease associated SNP variations with enviro-behavioral or clinical parameters can be used to **reclassify** the subjects who are initially assigned to a low-risk category on the basis on different risk score. Reclassification can be of particular value in clinical decision making in people defined as intermediate risk by standard guidelines (Manolio, 2013).

In our study, we analyzed the effects of possible risk factors on polygenic risk scores and how we could use these factors to acquire more effective results.

CHAPTER 8

EVALUATION OF THE COMPLEMENTARY COMPONENTS FOR PROSTATE CANCER

8.1 Test Data

The proposed ClinGenKB and ClinGenWeb is evaluated by the real data (23andMe files) of personal genome project (https://my.personalgenomes.org/public_genetic_data) as use cases. This publicly available resource brings genomic, environmental and human trait data together.

Among the data in personal genome project, we have extracted four 23andMe files that belongs to men who have been diagnosed with prostate cancer. All of these patients were white and over 60 years of age. As controls white men greater than 60 years of age were selected. There were 15 healthy individual white men over the age of 60, whom 23andMe file was provided (Table 27).

Table 27: Characteristics of genomic data owners.

Participant id	Prostate cancer	Ancestral origin	Birth year
01-hu1213DA	Yes	Germany-Norway	1937
03-huD889CC	Yes	Ireland	1938
07-hu28F39C	Yes	United States	1943
13-hu6ED94A	Yes	United States-Austria	1950
02-hu59141C	No	United States-Canada	1937
04-huF7E042	No	United States-United Kingdom	1939
05-hu75BE2C	No	United States	1939
06-hu56B3B6	No	United States	1941
08-huB59C05	No	United States-Ireland	1943
10-hu7A2F1D	No	United States-Germany	1947
12-huD57BBF	No	United States	1949
14-huD7960A	No	Hungary-Ukraine-Russia	1951
15-hu2E413D	No	United States	1952
16-hu76CAA5	No	United States	1952
17-huA720D3	No	United States-United Kingdom	1953
18-hu63DA55	No	United States	1953
19-hu43860C	No	United Kingdom-Hungary	1954
20-huD00199	No	Germany-Poland	1954
21-huAC827A	No	United States-Sweden	1954

8.2 Preparation of Data

In the evaluation phase, first, we have generated personal CR-SNP data files for prostate cancer from the original 23andMe files based on clinicogenomic associations. Then, we have transferred clinicogenomic associations and the test data into ClinGenKB. Personal clinicogenomic associations were acquired processing personal CR-SNP data with a smart query based on ClinGenKB. After that, acquired clinicogenomic associations were transferred into ClinGenWeb. Also, the relevant personal health data were transferred from the personal genome project web site into the ClinGenWeb to be used in the determination of disease risk based on the models. Finally, validity of the implemented models and approaches were compared and discussed.

Prostate cancer assessment and reporting approaches and their specific examples which will be evaluated in this chapter are summarized in Figure 45.

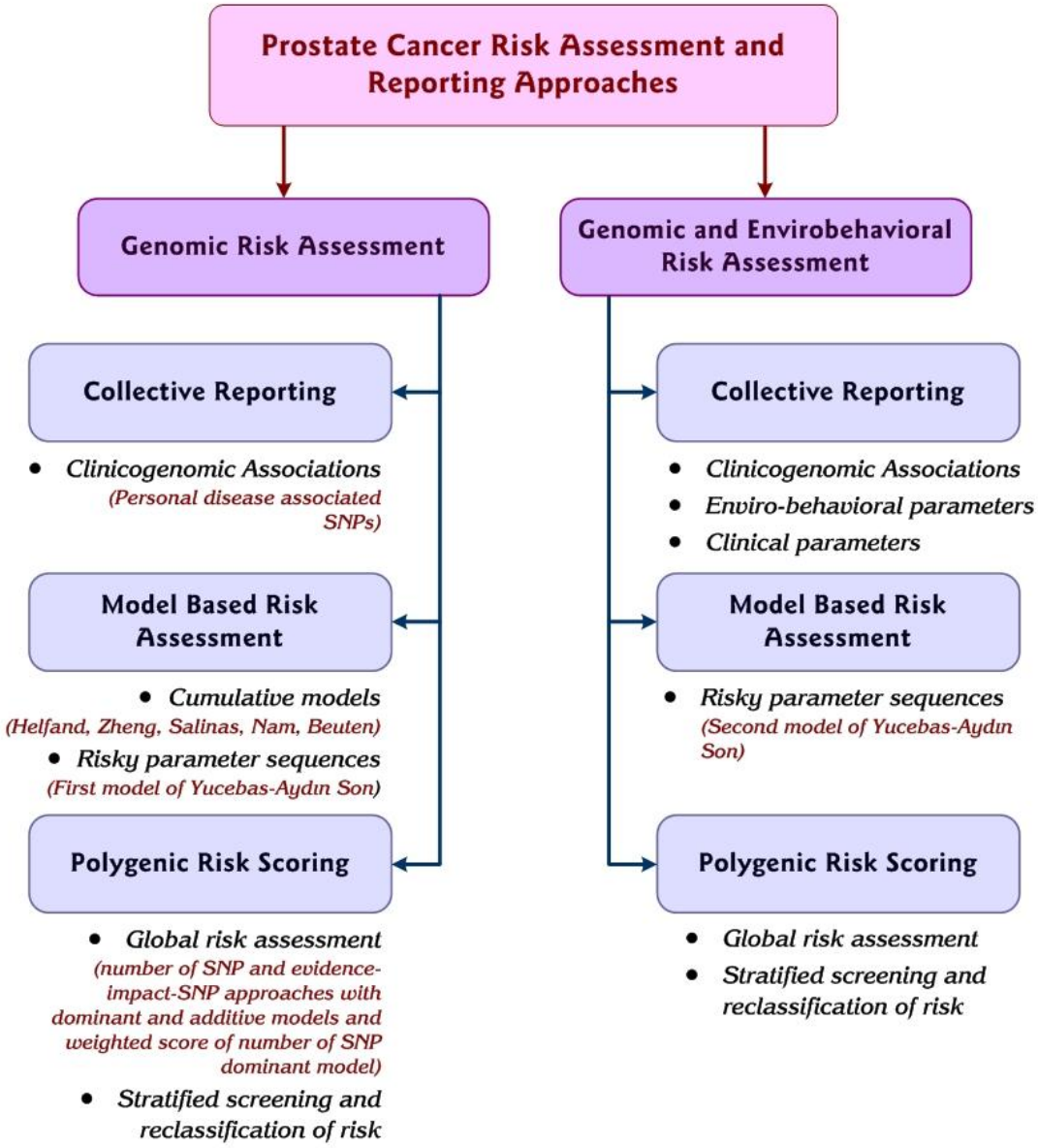


Figure 45: Prostate cancer assessment and reporting approaches.

8.3 Evaluation Results

Complete results (independent association assessment, model based evaluations, and clinical and environmental data) of test and evaluation processes were presented as a whole in Appendix G.

8.3.1 Collective Reporting of Independent Clinicogenomic Associations

In our test scenario, we have searched 106 SNPs in case and control groups. Results of these independent clinicogenomic associations are summarized in Table 28.

Table 28: Complete number of clinicogenomic associations.

	PERSONAL-ID	HOMO-ZYGOTE	HETERO-ZYGOTE	NORMAL	NOT-ANALYZED	TOTAL
CASES	01-hu1213DA	13	38	42	13	106
	03-huD889CC	13	32	41	20	106
	07-hu28F39C	21	28	48	9	106
	13-hu6ED94A	16	36	34	20	106
CONTROLS	17-huA720D3	18	36	43	9	106
	21-huAC827A	17	37	43	9	106
	10-hu7A2F1D	12	33	37	24	106
	08-huB59C05	21	32	45	8	106
	15-hu2E413D	19	34	45	8	106
	06-hu56B3B6	14	39	45	8	106
	04-huF7E042	20	26	43	17	106
	02-hu59141C	16	32	50	8	106
	12-huD57BBF	9	39	50	8	106
	14-huD7960A	15	31	51	9	106
	16-hu76CAA5	12	33	52	9	106
	05-hu75BE2C	9	36	52	9	106
	19-hu43860C	14	23	45	24	106
	20-huD00199	22	21	53	10	106
	18-hu63DA55	13	28	57	8	106

Due to version differences of 23andMe genomic test, some of the SNPs were not included in the analysis. Homozygote and heterozygote risk alleles are categorized as listed in table 28.

8.3.2 Genomic Risk Models Based Approaches

8.3.2.1. Cumulative Models

Overall results of cumulative models are in Appendix H. Results of these models are also summarized in Table 29.

Due to lack of family history data of individuals; we couldn't use this data to calculate cumulative risks. In our limited number of cases, cumulative models were not successful at predicting the outcome. But, like in the complete evaluation of independent associations, it must be considered that, non-analyzed SNPs might be distorting the results.

While developing Table 29, we accepted the individuals which have OR values more than 2.5 as under risk. In this table, unknown means that, if the unmeasured SNPs could be measured, there was a possibility to determine a risk.

Table 29: Summarized results for cumulative models

Model	**Case**			**Control**		
	OR>=2.5	**OR<2.5**	**Unknown**	**OR>=2.5**	**OR<2.5**	**Unknown**
17-SNP_ Helfand	1	-	3	2 (02-hu59141C, 12-huD57BBF)	10	3
9-SNP_ Helfand	1	3 (01-hu1213DA, 03-huD889CC, 07-hu28F39C)	-	1 (17-huA720D39)	12	2
5-SNP_ Zheng	-	4	-	-	15	-
5-SNP_ Salinas	-	4	-	-	15	-
4-SNP_ Nam	-	4	-	-	15	-
3-SNP_ Beuten	-	2	2	-	13	2

Excluding the unknown results, the total evaluation of all models are in the Table 30.

Table 30: Risk calculation using only SNP model.

GROUPS	**PERSONAL-ID**	**RESULTS**
Case	**01-hu1213DA**	**None**
Control	02-hu59141C	Yes (1)
Case	**03-huD889CC**	**None**
Control	04-huF7E042	None
Control	05-hu75BE2C	Yes
Control	06-hu56B3B6	None
Case	**07-hu28F39C**	**Yes (1)**
Control	08-huB59C05	None
Control	10-hu7A2F1D	None
Control	12-huD57BBF	Yes (1)

Table 30 (cont.): Risk calculation using only SNP model.

Case	**13-hu6ED94A**	**Yes (1)**
Control	14-huD7960A	None
Control	15-hu2E413D	None
Control	16-hu76CAA5	None
Control	17-huA720D3	Yes (1)
Control	18-hu63DA55	None
Control	19-hu43860C	None
Control	20-huD00199	None
Control	21-huAC827A	None

The sensitivity (recall) of cumulative models are %50 and the specificity of these models are %80. The accuracy rate is %73.68 and the precision rate is %40.

8.3.2.2. Probabilistic Models

The first model (only genotyping model) of Yücebaş-Aydın Son study was not able to predict the disease outcome for any of the cases, as summarized in Table 31.

Table 31: Risk calculation using only SNP model.

GROUPS	**PERSONAL-ID**	**RESULTS**
Case	**01-hu1213DA**	**Not exact**
Control	02-hu59141C	None
Case	**03-huD889CC**	**None**
Control	04-huF7E042	None
Control	05-hu75BE2C	Yes
Control	06-hu56B3B6	None
Case	**07-hu28F39C**	**None**
Control	08-huB59C05	None
Control	10-hu7A2F1D	None
Control	12-huD57BBF	None
Case	**13-hu6ED94A**	**None**
Control	14-huD7960A	None
Control	15-hu2E413D	None
Control	16-hu76CAA5	None
Control	17-huA720D3	None
Control	18-hu63DA55	None
Control	19-hu43860C	None
Control	20-huD00199	None
Control	21-huAC827A	None

One control was individual under prostate cancer risk and for one case risk there was an uncertain risk i.e. needed further analysis with additional SNPs. All of other individuals (cases and controls) did not have any risk regarding this model.

8.3.3 Polygenic Risk Scoring Based Approaches

In prostate cancer, known relevant SNPs have mostly modest odds ratio. Therefore, we assessed total number and values of relevant personal SNPs with four approaches, namely number of SNP dominant and additive models and evidence-impact-SNP dominant and additive models. The results of sensitivity, specificity, PPV, NPV, ROC and AUC for every model was presented and compared regarding the possible advantages of models as diagnostic and screening test below.

8.3.3.1. Number of SNP-Dominant Model

This method is the calculation of total count of individual genomic risk SNPs (regardless of allele characteristics). In this method disease risk (N_{risk});

$$N\,(\text{risk}) = \sum x_i \qquad \textbf{(EQUATION 3)}$$

where x_i=existing of risk SNPs (0 or 1) at SNP i. The values of cases and controls are presented in Table 32.

Table 32: Risk calculation using "Number of SNP-Dominant Model".

Group	Patient_id	Value
Control	19-hu43860C	37
Control	18-hu63DA55	41
Control	20-huD00199	43
Control	05-hu75BE2C	45
Control	10-hu7A2F1D	45
Control	16-hu76CAA5	45
Case	**03-huD889CC**	**45**
Control	04-huF7E042	46
Control	14-huD7960A	46
Control	02-hu59141C	48
Control	12-huD57BBF	48
Case	**07-hu28F39C**	**49**
Case	**01-hu1213DA**	**51**
Case	**13-hu6ED94A**	**52**
Control	08-huB59C05	53
Control	15-hu2E413D	53
Control	06-hu56B3B6	54
Control	17-huA720D3	54
Control	21-huAC827A	54

Minimum and maximum values of observations are 37 and 54 respectively. Mean value is 47, 842 and standard deviation is 4,845. Evaluation of "Number of SNP-Dominant Model" is in Table 33.

Table 33: Evaluation of "Number of SNP-Dominant Model".

Value	TP	TN	FP	FN	Sensitivity	Specificity	PPV	NPV	LR+	LR-	Accuracy
37,000	4	0	15	0	1,000	0,000	0,211		1,000		0,211
41,000	4	1	14	0	1,000	0,067	0,222	1,000	1,071	0,000	0,263
43,000	4	2	13	0	1,000	0,133	0,235	1,000	1,154	0,000	0,316
45,000	4	3	12	0	1,000	0,200	0,250	1,000	1,250	0,000	0,368
46,000	3	6	9	1	0,750	0,400	0,250	0,857	1,250	0,625	0,474
48,000	3	8	7	1	0,750	0,533	0,300	0,889	1,607	0,469	0,579
49,000	**3**	**10**	**5**	**1**	**0,750**	**0,667**	**0,375**	**0,909**	**2,250**	**0,375**	**0,684**
51,000	2	10	5	2	0,500	0,667	0,286	0,833	1,500	0,750	0,632
52,000	1	10	5	3	0,250	0,667	0,167	0,769	0,750	1,125	0,579
53,000	0	10	5	4	0,000	0,667	0,000	0,714	0,000	1,500	0,526
54,000	0	12	3	4	0,000	0,800	0,000	0,750	0,000	1,250	0,632

Test is positive if Value >= threshold value ***(Bold row)***
TP; True Positive, TN; True Negative, FP; False Positive, FN; False Negative, PPV; Positive Predictive Value, NPV; Negative Predictive Value, LR+; Positive Likelihood Ratio, LR-; Negative Likelihood Ratio.

Optimum threshold for diagnostic purposes is 49 for "Number of SNP-Dominant Model" (Figure 46).

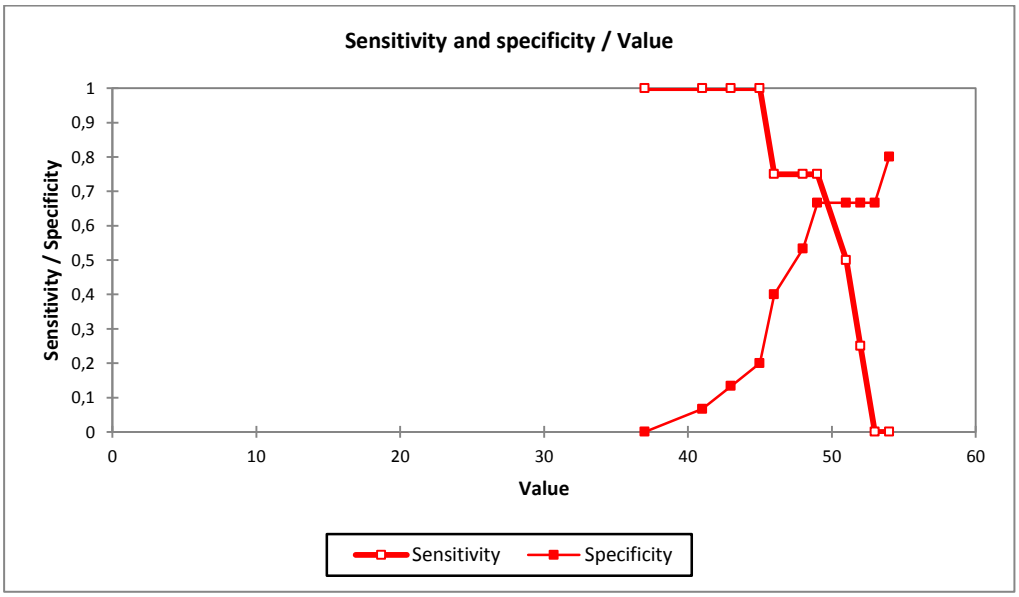


Figure 46: Graphical representation of optimum threshold for "Number of SNP-Dominant Model".

Receiver-Operating Characteristic (ROC) is drawn in Figure 47. Value of area under the curve (AUC) is 0,575.

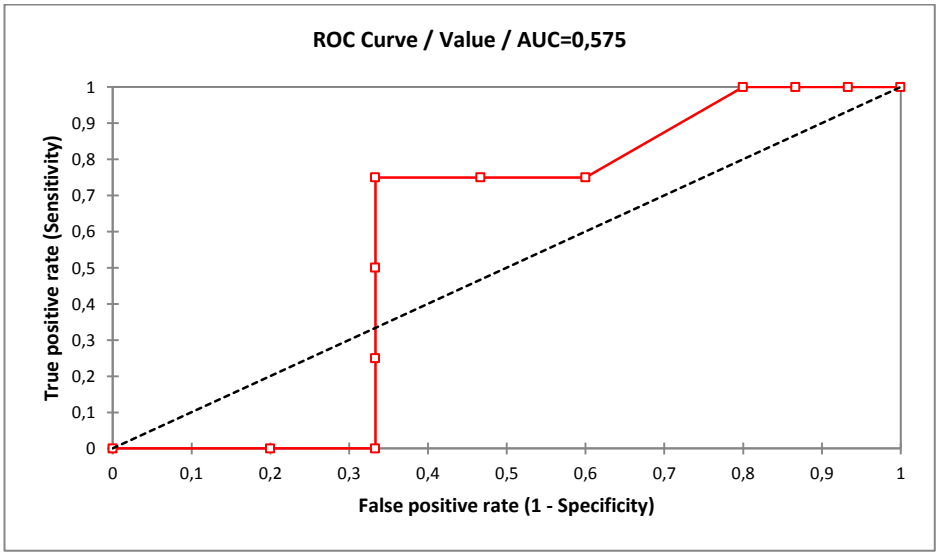


Figure 47: Receiver-Operating Characteristic (ROC) graph of "Number of SNP-Dominant Model".

8.3.3.2. Number of SNP-Dominant Model-Weighted Score

Because different version of analyzing, personal genome file may contain different type of SNPs. As explained in Table 28 (Complete number of clinicogenomic associations), some of clinicogenomic associations were not analyzed. Therefore, we transformed the results based on only analyzed SNPs and acquired weighted SNP risks by dividing total risky SNP by total analyzed risky SNP.

This method is the weighted mean of total count of individual genomic risk SNPs (regardless of allele characteristics). In this method weighted disease risk score (N_{risk});

$$N\,(\text{risk}) = \left(\sum x_i\right)/S \qquad \textbf{(EQUATION 4)}$$

where x_i=existing of risk SNPs (0 or 1) at SNP and S= number of analyzed SNP. The values of cases and controls are presented in Table 34.

Table 34: Weighted scores for Number of SNP-Dominant Model-Weighted Score

GROUP	PERSONAL-ID	HOMO-ZYGOTE	HETERO-ZYGOTE	NOT-ANALYZED	TOTAL	WEIGHTED SNP RISK (DM)
Case	**13-hu6ED94A**	**16**	**36**	**20**	**106**	**0,60**
Control	10-hu7A2F1D	12	33	24	106	0,55
Control	21-huAC827A	17	37	9	106	0,56
Control	17-huA720D3	18	36	9	106	0,56
Case	01-hu1213DA	13	38	13	106	0,55

Table 34 (cont.): Weighted scores for Number of SNP-Dominant Model-Weighted Score

GROUP	PERSONAL-ID	HOMO-ZYGOTE	HETERO-ZYGOTE	NOT-ANALYZED	TOTAL	WEIGHTED SNP RISK (DM)
Control	08-huB59C05	21	32	8	106	0,54
Control	06-hu56B3B6	14	39	8	106	0,54
Control	15-hu2E413D	19	34	8	106	0,54
Control	04-huF7E042	20	26	17	106	0,52
Case	**03-huD889CC**	**13**	**32**	**20**	**106**	**0,52**
Case	**07-hu28F39C**	**21**	**28**	**9**	**106**	**0,51**
Control	02-hu59141C	16	32	8	106	0,49
Control	12-huD57BBF	9	39	8	106	0,49
Control	19-hu43860C	14	23	24	106	0,45
Control	16-hu76CAA5	12	33	9	106	0,46
Control	05-hu75BE2C	9	36	9	106	0,46
Control	14-huD7960A	15	31	9	106	0,47
Control	20-huD00199	22	21	10	106	0,45
Control	18-hu63DA55	13	28	8	106	0,42

Minimum and maximum values of observations are 0,418 and 0,605 respectively. Mean value is 0,510 and standard deviation is 0,048. Evaluation of "Number of SNP-Dominant Model" is in Table 35.

Table 35: Evaluation of "Number of SNP-Dominant Model--Weighted Score"

Weighted SNP Risk	TP	TN	FP	FN	Sensitivity	Specificity	PPV	NPV	LR+	LR-	Accuracy
0,418	4	0	15	0	1,000	0,000	0,500		1,000		0,211
0,448	4	1	14	0	1,000	0,067	0,517	1,000	1,071	0,000	0,263
0,451	4	2	13	0	1,000	0,133	0,536	1,000	1,154	0,000	0,316
0,464	4	3	12	0	1,000	0,200	0,556	1,000	1,250	0,000	0,368
0,474	4	5	10	0	1,000	0,333	0,600	1,000	1,500	0,000	0,474
0,490	4	6	9	0	1,000	0,400	0,625	1,000	1,667	0,000	0,526
0,505	**4**	**8**	**7**	**0**	**1,000**	**0,533**	**0,682**	**1,000**	**2,143**	**0,000**	**0,632**
0,517	3	8	7	1	0,750	0,533	0,616	0,681	1,607	0,469	0,579
0,523	3	9	6	1	0,750	0,600	0,652	0,706	1,875	0,417	0,632
0,541	2	9	6	2	0,500	0,600	0,556	0,545	1,250	0,833	0,579
0,548	2	12	3	2	0,500	0,800	0,714	0,615	2,500	0,625	0,737
0,549	1	12	3	3	0,250	0,800	0,556	0,516	1,250	0,938	0,684
0,557	1	13	2	3	0,250	0,867	0,652	0,536	1,875	0,865	0,737
0,605	1	15	0	3	0,250	1,000	1,000	0,571	+Inf	0,750	0,842

Test is positive if Value >= threshold value ***(Bold row)*** TP; True Positive, TN; True Negative, FP; False Positive, FN; False Negative, PPV; Positive Predictive Value, NPV; Negative Predictive Value, LR+; Positive Likelihood Ratio, LR-; Negative Likelihood Ratio.

Optimum threshold for diagnostic purposes is 0,505 for "Number of SNP-Dominant Model-Weighted Score" (Figure 48).

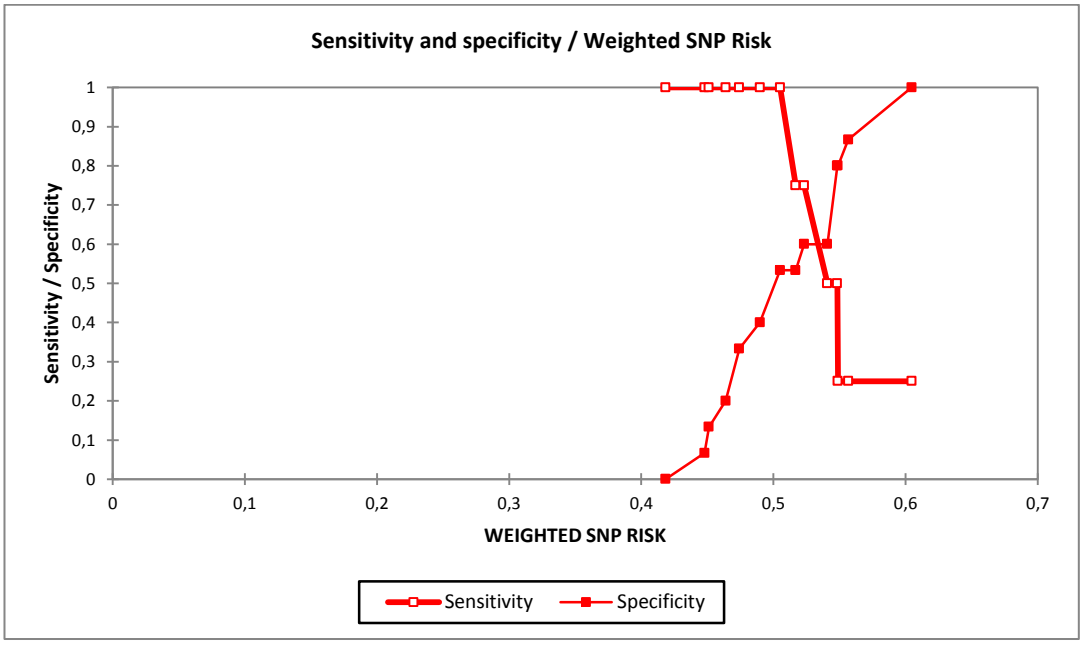


Figure 48: Graphical representation of optimum threshold for "Number of SNP-Dominant Model- Weighted Score".

Value of area under the curve (AUC) is 0,733. ROC graph of "Number of SNP-Dominant Model-Weighted Score" is presented in Figure 49.

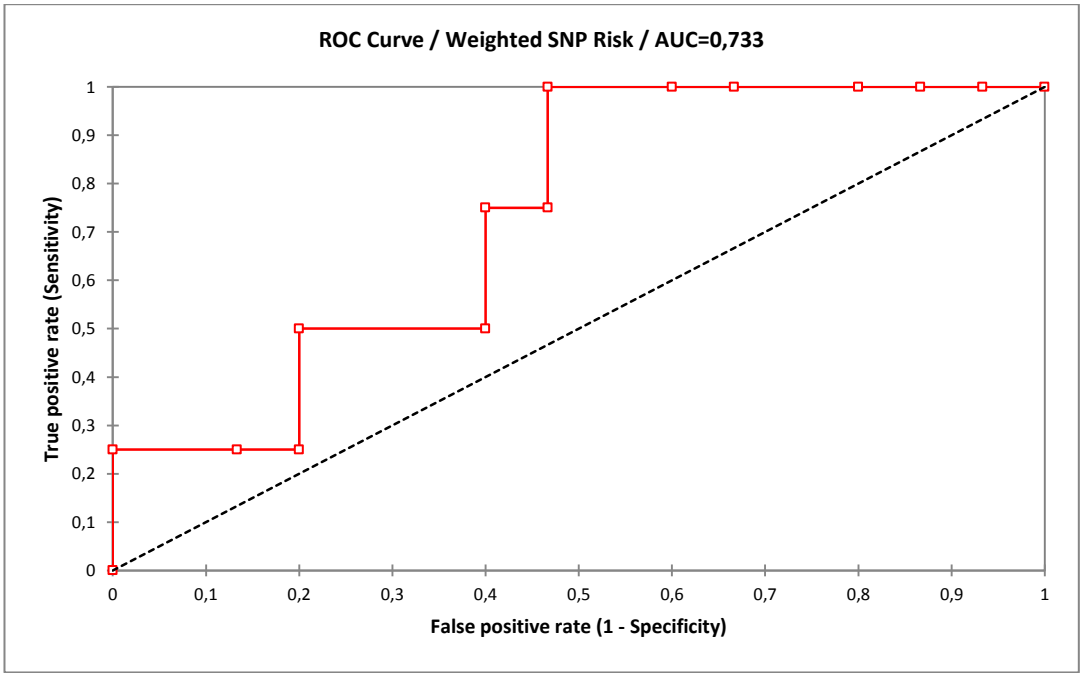


Figure 49: Receiver-Operating Characteristic (ROC) graph of "Number of SNP-Dominant Model-Weighted Score".

8.3.3.3. Number of SNP-Additive Model

This method is the calculation of total count of individual genomic risk alleles. In this method disease risk (N_{risk});

$$N\,(\text{risk}) = \sum x_i \qquad \textbf{(EQUATION 5)}$$

where x_i=number of risk alleles (0=homozygote healthy, 1=heterozygote risky, 2=homozygote risky) at SNP i. The values of cases and controls are presented in Table 36.

Table 36: Risk calculation using "Number of SNP-Additive Model".

Group	Patient_id	Value
Control	19-hu43860C	51
Control	05-hu75BE2C	54
Control	18-hu63DA55	54
Control	10-hu7A2F1D	57
Control	12-huD57BBF	57
Control	16-hu76CAA5	57
Case	**03-huD889CC**	**58**
Control	14-huD7960A	61
Control	02-hu59141C	64
Case	**01-hu1213DA**	**64**
Control	20-huD00199	65
Control	04-huF7E042	66
Control	06-hu56B3B6	68
Case	**13-hu6ED94A**	**68**
Case	**07-hu28F39C**	**70**
Control	21-huAC827A	71
Control	15-hu2E413D	72
Control	17-huA720D3	72
Control	08-huB59C05	74

Minimum and maximum values of observations are 51 and 74 respectively. Mean value is 63,316 and standard deviation is 7,079. Evaluation of "Number of SNP-Additive Model" is in Table 37.

Table 37: Evaluation of "Number of SNP-Additive Model".

Value	TP	TN	FP	FN	Sensitivity	Specificity	PPV	NPV	LR+	LR-	Accuracy
51,000	4	0	15	0	1,000	0,000	0,211		1,000		0,211
54,000	4	1	14	0	1,000	0,067	0,222	1,000	1,071	0,000	0,263
57,000	4	3	12	0	1,000	0,200	0,250	1,000	1,250	0,000	0,368
58,000	**4**	**6**	**9**	**0**	**1,000**	**0,400**	**0,308**	**1,000**	**1,667**	**0,000**	**0,526**

Table 37 (cont.): Evaluation of "Number of SNP-Additive Model".

Value	TP	TN	FP	FN	Sensitivity	Specificity	PPV	NPV	LR+	LR-	Accuracy
61,000	3	6	9	1	0,750	0,400	0,250	0,857	1,250	0,625	0,474
64,000	3	7	8	1	0,750	0,467	0,273	0,875	1,406	0,536	0,526
65,000	2	8	7	2	0,500	0,533	0,222	0,800	1,071	0,938	0,526
66,000	2	9	6	2	0,500	0,600	0,250	0,818	1,250	0,833	0,579
68,000	2	10	5	2	0,500	0,667	0,286	0,833	1,500	0,750	0,632
70,000	1	11	4	3	0,250	0,733	0,200	0,786	0,938	1,023	0,632
71,000	0	11	4	4	0,000	0,733	0,000	0,733	0,000	1,364	0,579
72,000	0	12	3	4	0,000	0,800	0,000	0,750	0,000	1,250	0,632
74,000	0	14	1	4	0,000	0,933	0,000	0,778	0,000	1,071	0,737

Test is positive if Value >= threshold value ***(Bold row)***
TP; True Positive, TN; True Negative, FP; False Positive, FN; False Negative, PPV; Positive Predictive Value, NPV; Negative Predictive Value, LR+; Positive Likelihood Ratio, LR-; Negative Likelihood Ratio.

Optimum threshold for diagnostic purposes is 49 for "Number of SNP-Additive Model" (Figure 50).

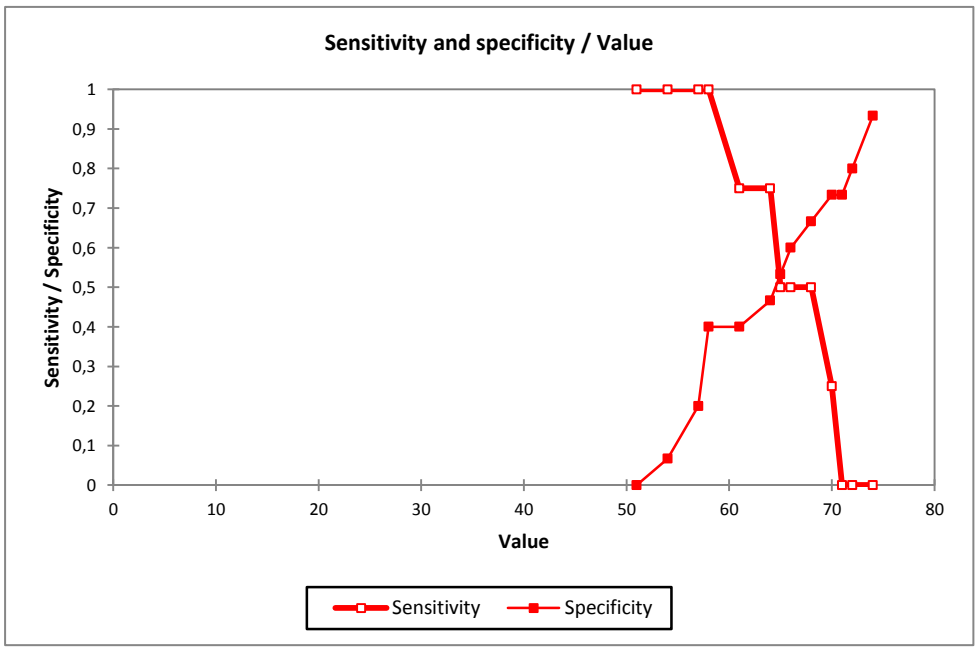


Figure 50: Graphical representation of optimum threshold for "Number of SNP-Additive Model".

Receiver-Operating Characteristic (ROC) is drawn in Figure 51. Value of area under the curve (AUC) is 0,583.

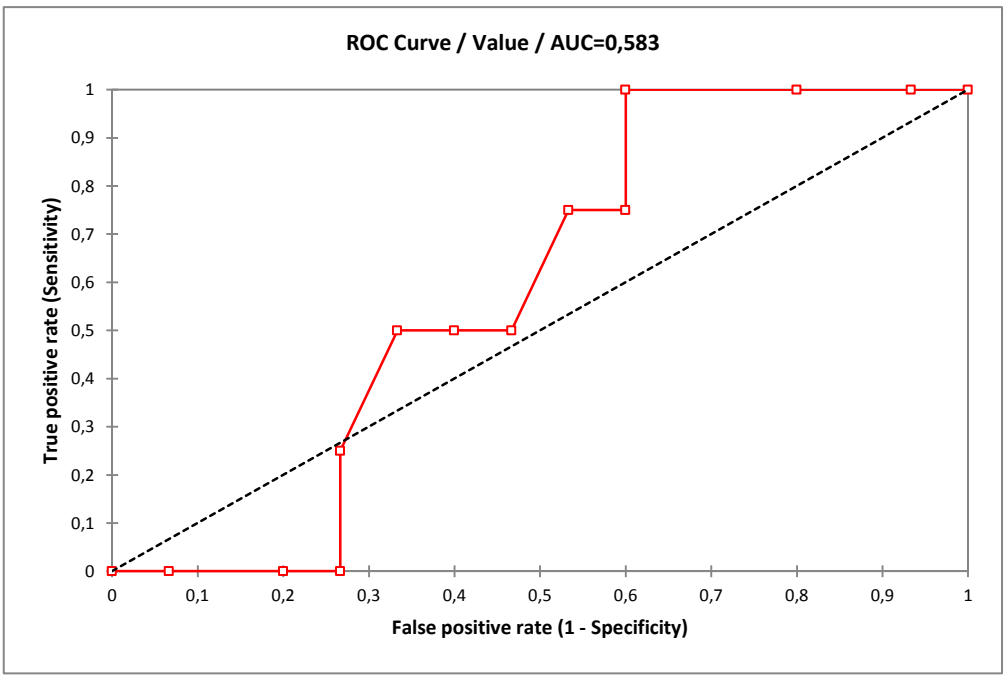


Figure 51: Receiver-Operating Characteristic (ROC) graph of "Number of SNP-Additive Model".

8.3.3.4. Evidence-Impact-SNP- Dominant Model

This method is the summation of the evidence and impact values of complete individual genomic risk SNPs regardless of allele characteristics. In this method disease risk (N_{risk});

$$N\,(\text{risk}) = \left(\sum x_i\right) I_i E_i \qquad \textbf{(EQUATION 6)}$$

where x_i=existing of risk SNPs (0 or 1), I_i=magnitude of impact degree, E_i=quality of evidence degree at SNP i. The values of cases and controls are presented in Table 38.

Table 38: Risk calculation using "Evidence-Impact-SNP- Dominant Model".

Group	Patient_id	Value
Control	19-hu43860C	122
Control	18-hu63DA55	132
Control	20-huD00199	133
Case	**03-huD889CC**	**137**
Control	14-huD7960A	142
Control	05-hu75BE2C	143
Control	16-hu76CAA5	143
Control	10-hu7A2F1D	145
Control	04-huF7E042	145
Control	12-huD57BBF	147

Table 38 (cont.): Risk calculation using "Evidence-Impact-SNP- Dominant Model".

Group	Patient_id	Value
Control	02-hu59141C	150
Case	**07-hu28F39C**	**152**
Case	**01-hu1213DA**	**159**
Control	15-hu2E413D	161
Case	**13-hu6ED94A**	**161**
Control	06-hu56B3B6	164
Control	17-huA720D3	166
Control	08-huB59C05	166
Control	21-huAC827A	170

Minimum and maximum values of observations are 122 and 170 respectively. Mean value is 149,368 and standard deviation is 13,363. Evaluation of "Evidence-Impact-SNP- Dominant Model" is in Table 39.

Table 39: Evaluation of "Evidence-Impact-SNP- Dominant Model".

Value	TP	TN	FP	FN	Sensitivity	Specificity	PPV	NPV	LR+	LR-	Accuracy
122,000	4	0	15	0	1,000	0,000	0,211		1,000		0,211
132,000	4	1	14	0	1,000	0,067	0,222	1,000	1,071	0,000	0,263
133,000	4	2	13	0	1,000	0,133	0,235	1,000	1,154	0,000	0,316
137,000	4	3	12	0	1,000	0,200	0,250	1,000	1,250	0,000	0,368
142,000	3	3	12	1	0,750	0,200	0,200	0,750	0,938	1,250	0,316
143,000	3	4	11	1	0,750	0,267	0,214	0,800	1,023	0,938	0,368
145,000	3	6	9	1	0,750	0,400	0,250	0,857	1,250	0,625	0,474
147,000	3	8	7	1	0,750	0,533	0,300	0,889	1,607	0,469	0,579
150,000	3	9	6	1	0,750	0,600	0,333	0,900	1,875	0,417	0,632
152,000	**3**	**10**	**5**	**1**	**0,750**	**0,667**	**0,375**	**0,909**	**2,250**	**0,375**	**0,684**
159,000	2	10	5	2	0,500	0,667	0,286	0,833	1,500	0,750	0,632
161,000	1	10	5	3	0,250	0,667	0,167	0,769	0,750	1,125	0,579
164,000	0	11	4	4	0,000	0,733	0,000	0,733	0,000	1,364	0,579
166,000	0	12	3	4	0,000	0,800	0,000	0,750	0,000	1,250	0,632
170,000	0	14	1	4	0,000	0,933	0,000	0,778	0,000	1,071	0,737

Test is positive if Value >= threshold value ***(Bold row)***
TP; True Positive, TN; True Negative, FP; False Positive, FN; False Negative, PPV; Positive Predictive Value, NPV; Negative Predictive Value, LR+; Positive Likelihood Ratio, LR-; Negative Likelihood Ratio.

Optimum threshold for diagnostic purposes is 49 for "Evidence-Impact-SNP- Dominant Model" (Figure 52).

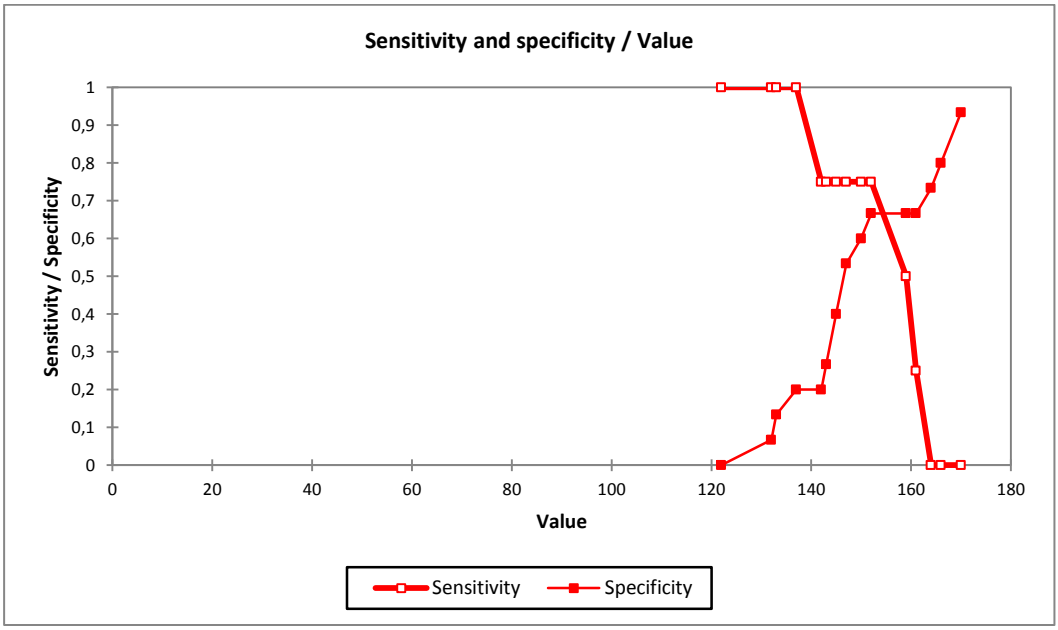


Figure 52: Graphical representation of optimum threshold for "Evidence-Impact-SNP-Dominant Model".

Value of area under the curve (AUC) is 0,558. ROC graph of "Evidence-Impact-SNP-Dominant Model" is presented in Figure 53.

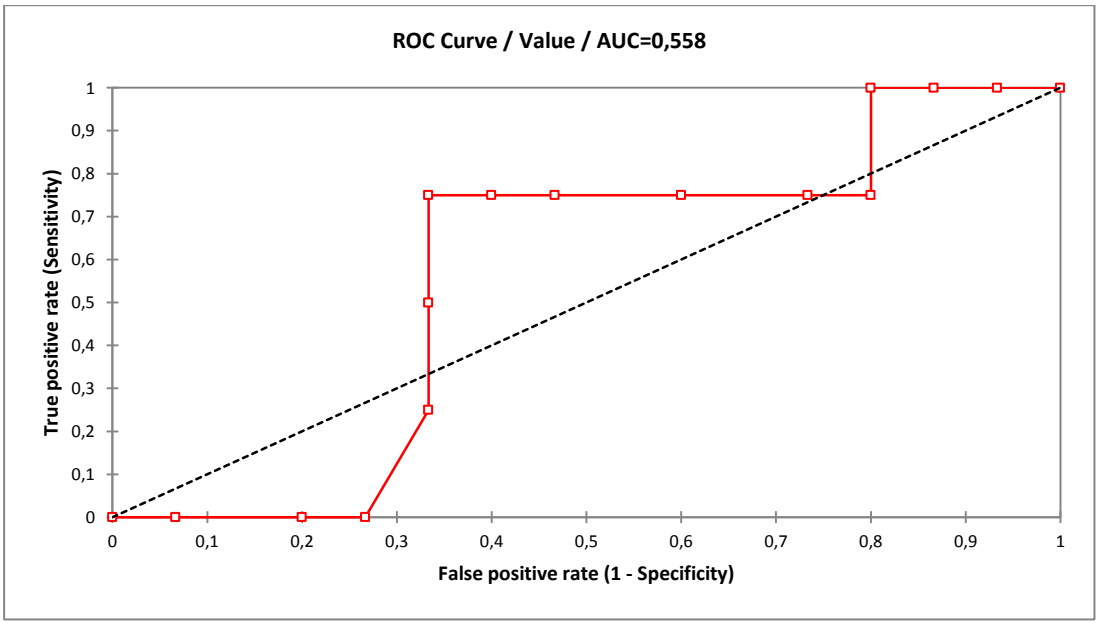


Figure 53: Receiver-Operating Characteristic (ROC) graph of Evidence-Impact-SNP-Dominant Model".

8.3.3.5. Evidence-Impact-SNP- Additive Model

This method is the sum of the evidence and impact values of complete individual genomic risk alleles. In this method disease risk (N_{risk});

$$N(\text{risk}) = \left(\sum x_i\right) I_i E_i \qquad \textbf{(EQUATION 7)}$$

where x_i=number of risk alleles (0=homozygote healthy, 1=heterozygote risky, 2=homozygote risky), I_i=magnitude of impact degree, E_i=quality of evidence degree at SNP i. The values of cases and controls are presented in Table 40.

Table 40: Risk calculation using "Evidence-Impact-SNP- Additive Model".

Group	Patient_id	Value
Control	08-huB59C05	236
Control	21-huAC827A	225
Control	17-huA720D3	222
Case	**07-hu28F39C**	**220**
Control	15-hu2E413D	219
Control	04-huF7E042	210
Case	**13-hu6ED94A**	**210**
Control	06-hu56B3B6	210
Control	20-huD00199	204
Control	02-hu59141C	202
Case	**01-hu1213DA**	**198**
Control	10-hu7A2F1D	190
Control	14-huD7960A	189
Control	16-hu76CAA5	182
Control	18-hu63DA55	179
Control	12-huD57BBF	177
Case	**03-huD889CC**	**177**
Control	05-hu75BE2C	172
Control	19-hu43860C	171

Minimum and maximum values of observations are 171 and 236 respectively. Mean value is 199,632 and standard deviation is 19,939. Evaluation of "Evidence-Impact-SNP- Additive Model" is in Table 41.

Table 41: Evaluation of "Evidence-Impact-SNP- Additive Model".

Value	TP	TN	FP	FN	Sensitivity	Specificity	PPV	NPV	LR+	LR-	Accuracy
171,000	4	0	15	0	1,000	0,000	0,211		1,000		0,211
172,000	4	1	14	0	1,000	0,067	0,222	1,000	1,071	0,000	0,263
177,000	4	2	13	0	1,000	0,133	0,235	1,000	1,154	0,000	0,316
179,000	3	3	12	1	0,750	0,200	0,200	0,750	0,938	1,250	0,316
182,000	3	4	11	1	0,750	0,267	0,214	0,800	1,023	0,938	0,368
189,000	3	5	10	1	0,750	0,333	0,231	0,833	1,125	0,750	0,421
190,000	3	6	9	1	0,750	0,400	0,250	0,857	1,250	0,625	0,474
198,000	**3**	**7**	**8**	**1**	**0,750**	**0,467**	**0,273**	**0,875**	**1,406**	**0,536**	**0,526**
202,000	2	7	8	2	0,500	0,467	0,200	0,778	0,938	1,071	0,474
204,000	2	8	7	2	0,500	0,533	0,222	0,800	1,071	0,938	0,526
210,000	2	9	6	2	0,500	0,600	0,250	0,818	1,250	0,833	0,579

Table 41 (cont.): Evaluation of "Evidence-Impact-SNP- Additive Model".

Value	TP	TN	FP	FN	Sensitivity	Specificity	PPV	NPV	LR+	LR-	Accuracy
219,000	1	11	4	3	0,250	0,733	0,200	0,786	0,938	1,023	0,632
220,000	1	12	3	3	0,250	0,800	0,250	0,800	1,250	0,938	0,684
222,000	0	12	3	4	0,000	0,800	0,000	0,750	0,000	1,250	0,632
225,000	0	13	2	4	0,000	0,867	0,000	0,765	0,000	1,154	0,684
236,000	0	14	1	4	0,000	0,933	0,000	0,778	0,000	1,071	0,737

Test is positive if Value >= threshold value ***(Bold row)***
TP; True Positive, TN; True Negative, FP; False Positive, FN; False Negative, PPV; Positive Predictive Value, NPV; Negative Predictive Value, LR+; Positive Likelihood Ratio, LR-; Negative Likelihood Ratio.

Optimum threshold for diagnostic purposes is 49 for "Evidence-Impact-SNP- Additive Model" (Figure 54).

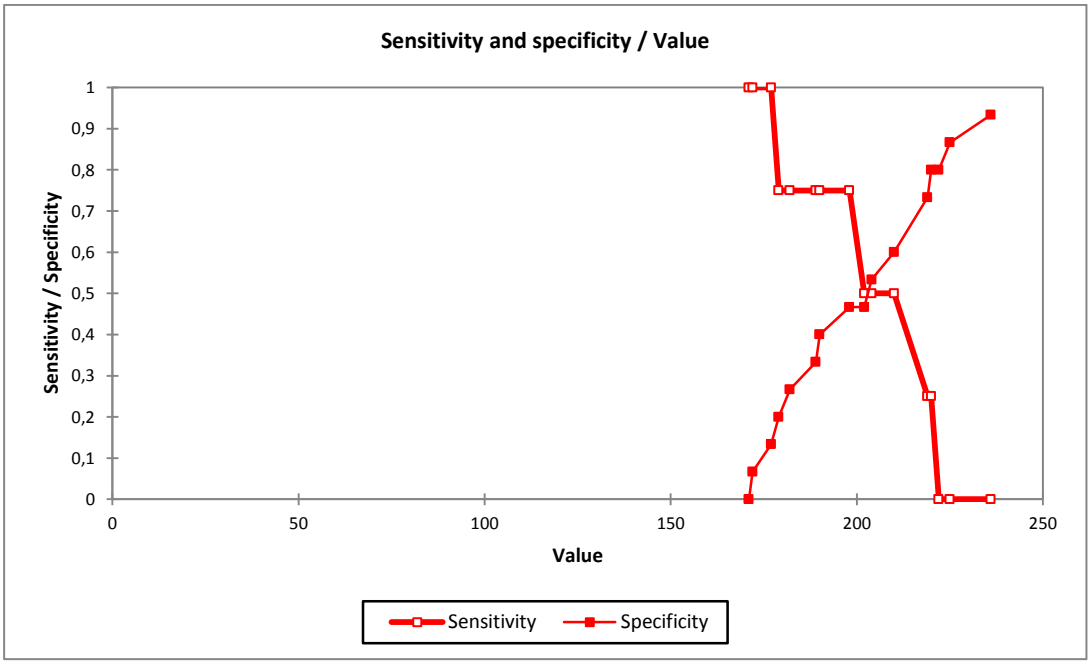


Figure 54: Graphical representation of optimum threshold for "Evidence-Impact-SNP-Additive Model".

Value of area under the curve (AUC) is 0,525. ROC graph of "Evidence-Impact-SNP-Additive Model" is presented in Figure 55.

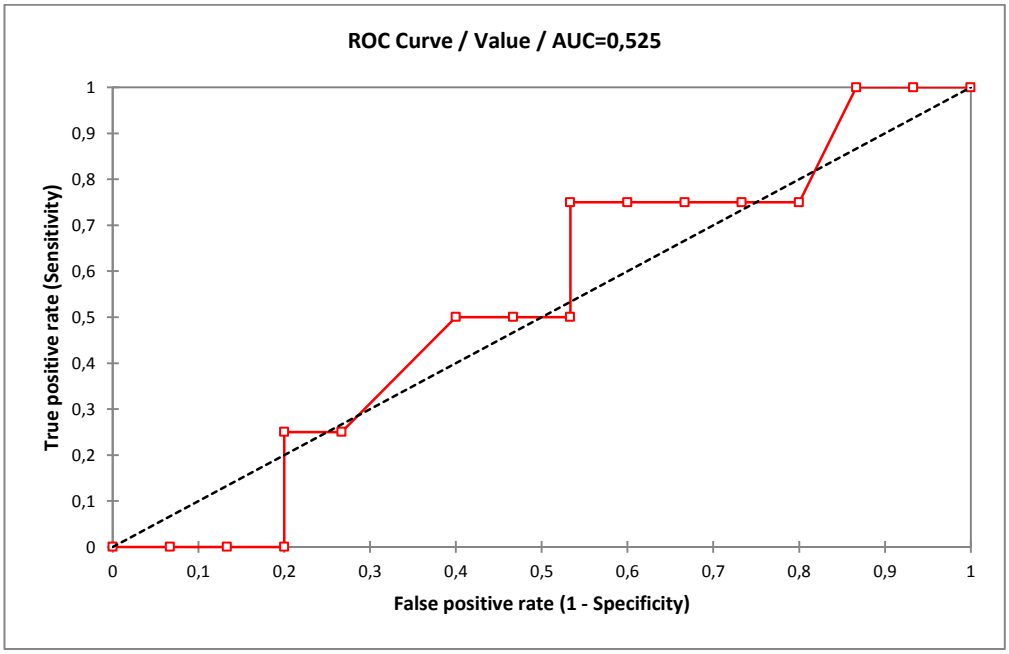


Figure 55: Receiver-Operating Characteristic (ROC) graph of "Evidence-Impact-SNP-Additive Model".

When we have evaluated and compared the performance of the models which explained these approaches and indicators, the "Number of SNP-Dominant Model-Weighted Score" become more prominent (Table 42). But, it must be remembered that, the sample size is too small and we need to evaluate these results in larger data samples.

Table 42: Comparison of different models regarding several performance indicators.

Models	Sensitivity	Specificity	PPV	NPV	LR+	LR-	Accuracy	AUC
Number of SNP-Dominant Model	0,750	0,667	0,375	0,909	2,250	0,375	0,684	0,575
Number of SNP-Dominant Model-Weighted Score	1,000	0,533	0,682	1,000	2,143	0,000	0,632	0,733
Number of SNP-Additive Model	1,000	0,400	0,308	1,000	1,667	0,000	0,526	0,583
Evidence-Impact-SNP-Dominant Model	0,750	0,667	0,375	0,909	2,250	0,375	0,684	0,558
Evidence-Impact-SNP-Additive Model	0,750	0,467	0,273	0,875	1,406	0,536	0,526	0,525

8.3.4 Collective Reporting of Genomic and Envirobehavioral Disease Risk Parameters

Prostate cancer is a polygenic multifactorial disease and both environmental, and genetic factors take important roles in its pathogenic mechanism. Therefore, if we analyze genomic risks with clinical and environmental characteristics, we can infer some useful results. But in literature, there are many controversial risk or protective factors about prostate cancer. Characteristics of cases and control regarding clinical and environmental risk factors for prostate cancer is summarized in Table 43.

Table 43: Clinical, environmental and behavioral risk factors of cases and control.

Risk factors	Individuals	Protective factors
Cases		
Hypercholesterolemia, BPH[1]	01-hu1213DA	
Syphilis	03-huD889CC	
Hypercholesterolemia, BPH Lipitor	07-hu28F39C	
Obesity, Hypercholesterolemia Simvastatin	13-hu6ED94A	
Controls		
Obesity, Multivitamins	02-hu59141C	T2DM Vegetable consumption, Regular physical activity (?)
BPH	04-huF7E042	TURP (2010)
	05-hu75BE2C	Regular physical activity
Obesity, Hypercholesterolemia, Chlamydia Infection, Alcoholism Ibuprofen, Multivitamin Folic Acid, Vitamin E, Selenium	06-hu56B3B6	Basal cell skin cancer, Lycopene, Pomegranate
Obesity	08-huB59C05	
Hypercholesterolemia Atorvastatin	10-hu7A2F1D	Non-melanoma skin cancer, Regular physical activity
Hypercholesterolemia, BPH Simvastatin, Aspirin Vasectomy	12-huD57BBF	Regular physical activity
Overweight, Hypercholesterolemia, BPH	14-huD7960A	T2DM
Overweight	15-hu2E413D	
Overweight Aspirin	16-hu76CAA5	Omega-3 Fish Oil
Hypercholesterolemia Aspirin, Multivitamin	17-huA720D3	Phytosterols, Omega-3 Fish Oil, Melatonin
	18-hu63DA55	Omega-3 Fish Oil
Overweight, Hypercholesterolemia Lovastatin	19-hu43860C	Non-melanoma skin cancer
Overweight, Hypercholesterolemia Atorvastatin	20-huD00199	
Overweight, Hypercholesterolemia Simvastatin	21-huAC827A	Hypogonadism
[1]Benign enlargement of prostate gland (Benign Prostate Hyperplasia, BPH) mostly is not accepted as a risk factor for prostate cancer, although the two frequently coexist.		

In enviro-behavioral and clinical evaluation, it was found that patient “03-huD889CC” had previously diagnosed as syphilis. In some publications, syphilis is accepted as a modest risk factor for prostate cancer (Sartor, 2013).

In high valued control individuals; “06-hu56B3B6” had basal cell carcinoma, 10-hu7A2F1D and 19-hu43860C had non-melanoma skin cancer and “21-huAC827A” had hypogonadism i.e. low level of testosterone. Both of these clinical conditions decreases the prostate cancer risk (National Cancer Institute, 2013B), (Sartor, 2013).

Also “06-hu56B3B6” and “17-huA720D3” uses several risky and protective drugs and supplements regarding prostate cancer risk.

In patient “08-huB59C05” and “15-hu2E413D”, we have not enough data to evaluate risk and protective factors. In health records of some cases and controls, there were several data about nutritional status, physical activity and usages of supplement data, etc. But, due to lack of precise measurement information (amount, period, duration, etc.) these data couldn’t use for evaluation.

8.3.5 Envirogenomic Model Based Approaches

In second hybrid model of Yücebaş-Aydın Son study, one patient was determined as under risk, two patients cannot be evaluated because of data incompleteness (smoking and alcohol consumption data) and one patient (03-huD889CC) was determined as risk-free. In controls, only one individual (04-huF7E042) was determined in risk group but six individuals was determined as risk-free. Eight individuals of this group cannot be evaluated due to data incompleteness. Although this model was produced for African Americans, and we had limited number of cases and controls for the evaluation process, this model was the most successful approach compared to other approaches. Interestingly, patient (03-huD889CC) who was determined as risk free is the same individual with the patient who was determined as in low risk group in the complete assessment approaches.

8.3.6 Polygenic Risk Scoring and Envirobehavioral Parameters Based Approaches

8.3.6.1. Global Risk Assessment

In scientific literature, various disease are indicated as important factors to increase or decrease the risk of prostate cancer e.g. syphilis, hypogonadism, non-melanoma skin cancers, diabetes mellitus etc.

There is a significant modest relationship between the risk of prostate cancer and the history of syphilis or gonorrhea (Sartor, 2013).

In some references, androgenic exposure of the prostate gland is studied as a risk factor. Ecological studies have ensured relationship between serum levels of testosterone and overall risk of prostate cancer (National Cancer Institute, 2013B). For this reason, hypogonadism may be a protective condition for prostate cancer. Additionally, testosterone supplementation for hypogonadism treatment does not increased risk of prostate cancer (Sartor, 2013).

Exposure to sunlight may ensure a protection for prostate cancer. The reason of this association is not clarified but the mechanism possibly related with vitamin D metabolism. Long term sunlight is also a major cause of non-melanoma skin cancers (Sartor, 2013).

And finally, patients with diabetes may have a lower risk of prostate cancer. The mechanism is not clear but the level of blood IGF-1 is proposed as a reason (Cancer Research UK, 2014).

We can produce a hypothetic risk assessment model based on personal genomic risk score and various risk factors to reclassify disease risk. In our model we used comorbidities, previous diseases and polygenic risk score based on number of SNP-Dominant Model. The values of cases and controls are presented in Table 44.

Table 44: Global risk assessment model containing polygenic risk score and co-morbidities.

GROUP	PERSONAL-ID	Total Risky SNP (DM)	Risky/ Protective comorbidities	Risk category of SNP	Polygenic risk value	Risky/ protective comorbidity value	Total Risk Value
Control	17-huA720D3	54		High	1,00	0,00	1,00
Control	21-huAC827A	54	hypogonadism	High	1,00	-1,00	0,00
Control	08-huB59C05	53		High	1,00	0,00	1,00
Control	15-hu2E413D	53		High	1,00	0,00	1,00
Control	06-hu56B3B6	53	basal cell carcinoma	High	1,00	-1,00	0,00
Case	**13-hu6ED94A**	**52**		**High**	**1,00**	**0,00**	**1,00**
Case	**01-hu1213DA**	**51**		**High**	**1,00**	**0,00**	**1,00**
Case	**07-hu28F39C**	**49**		**High**	**1,00**	**0,00**	**1,00**
Control	12-huD57BBF	48		Low	0,00	0,00	0,00
Control	02-hu59141C	48	T2DM	Low	0,00	-1,00	-1,00
Control	04-huF7E042	46		Low	0,00	0,00	0,00
Control	14-huD7960A	46	T2DM	Low	0,00	-1,00	-1,00
Case	**03-huD889CC**	**45**	**syphilis**	**Low**	**0,00**	**1,00**	**1,00**
Control	05-hu75BE2C	45		Low	0,00	0,00	0,00
Control	16-hu76CAA5	45		Low	0,00	0,00	0,00
Control	10-hu7A2F1D	45	Non-melanoma skin cancer	Low	0,00	-1,00	-1,00
Control	20-huD00199	43		Low	0,00	0,00	0,00
Control	18-hu63DA55	41		Low	0,00	0,00	0,00
Control	19-hu43860C	37	Non-melanoma skin cancer	Low	0,00	0,00	0,00

Risky comorbidites: Syphilis
Protective comorbidites: Hypogonadism, Basal Cell Carcinoma, Type 2 Diabetes Mellitus (T2DM), Non-Melanoma Skin Cancer.

In this model, for polygenic risk scores we determined high and low risk categories using 49 as a threshold and assigned 1 point for high risk category and 0 point for low risk category. Then, we determined risky and protective comorbidities and assigned 1 point for risky disease and -1 point for protective disease. Then we summed all these values and found total risk value for global risk assessment model containing polygenic risk score and comorbidities. Then we evaluated these values (Table 45).

Table 45: Evaluation of global risk assessment model.

Statistic	Value	Lower bound (95%)	Upper bound (95%)
Sensitivity	1,000	0,450	1,000
Specificity	0,786	0,515	0,929
False positive rate	0,214	0,025	0,404
False negative rate	0,000	0,000	0,000
PPV (Positive Predictive Value)	0,571	0,205	0,938
NPV (Negative Predictive Value)	1,000	1,000	1,000
LR+ (Positive likelihood ratio)	4,667	1,712	12,724
LR- (Negative likelihood ratio)	0,000		

8.3.6.2. Stratified Screening and Reclassification of Disease Risk

Using same parameters as in simple global risk assessment model, we developed a stratified screening model. In our model, firstly we evaluated comorbidities and excluded risky and protective comorbidities. We accepted risky comorbidities as “High Risk “category and protective comorbidities as “Low Risk” free from polygenic risk score.

After that, we have 3 cases and 9 controls and re-evaluate these individuals.

Total analysis of these approach is in Table 46.

Table 46: Evaluation of the stratified screening model.

Statistic	**Value**	**Lower bound (95%)**	**Upper bound (95%)**
Sensitivity	1,000	0,450	1,000
Specificity	0,800	0,539	0,935
False positive rate	0,200	0,020	0,380
False negative rate	0,000	0,000	0,000
PPV (Positive Predictive Value)	0,211	0,027	0,394
NPV (Negative Predictive Value)	0,571	0,205	0,938
LR+ (Positive likelihood ratio)	1,000	1,000	1,000
LR- (Negative likelihood ratio)	5,000	1,817	13,757

CHAPTER 9

DISCUSSION

9.1 Principal Results

In this study, we extended current architecture of a centralized national EHR i.e. NHIS-T and developed some complementary capabilities namely knowledge base (ClinGenKB) and reporting application (ClinGenWeb) to predict risk of diseases using SNP data.

With respect to interoperability, HL7 CG-SIG developed several standards and guidelines and try to cross the chasm between genomic laboratory and clinical practice. Comparing current and requiring infrastructure characteristics and determining some terminology standards for genome enabled messaging, NHIS-T can be adapted to HL7 CG standards because it's preferred to use HL7 v3 standards in NHIS-T.

Unique identification of SNP data is a critical issue in clinical genomics. In our system, due to simplicity and easiness, we proposed to use rs number and allele values for identification of SNPs. But, it's critical to remember that, some rs numbers have been merged in the course of time and to avoid inconsistencies. For this reason, SNP numbers must be checked out based on dbSNP and transformed into current values if required. Additionally, because different strand types are preferred among some clinicogenomic knowledge sources and publications, standardization of strand identification is another important point for SNP data incorporated clinical systems.

Regarding clinical terminology, we used existing NHIS-T standards e.g. ICD-10 for disease identification. For new data types (model name, model type, etc.), we produced our specific value categories.

To store and process of the huge amount of raw variant files, in our architecture, we accepted the idea defending of raw or processed genomic data was stored in genomic laboratory database, and clinically relevant variant data and/or clinicogenomic association information were shared between partners. To derive CR-SNP data from personal SNP data, we need to use a CR-SNP list. This list was designed as part of national level clinicogenomic knowledge base. This knowledge base is also required to transform CR-SNP data to clinicogenomic associations.

As it is emphasized in the literature, one of the most critical component of the genome enabled EHRs is to develop a national level knowledge base for clinicogenomic information. This capability must be kept up to date and manually curated by domain experts. For our study, we produced a prototype knowledge base (ClinGenKB) including clinicogenomic associations for prostate cancer.

Various different approaches are published to define clinical impact and evidence qualities of clinicogenomic associations in various knowledge sources. But there is still a lack of structured, objective and comprehensive methodologies for matching, selecting and merging different studies. In our prototype, we developed a simple methodology, but the best choice

will be to use available research standards (e.g. Venice criteria) to calculate and limit biases and faults for future clinicogenomic association studies.

ClinGenWeb is a prototype for end-user systems that present interpretations of clinicogenomic associations. To evaluate our system, we used real data from personal genome project. Collected data involved 23andMe data file, age, ethnicity, ancestral origin, clinical data, and some behavioral parameters. Age and ethnicity are extensively accepted as proven risk factors for prostate cancer. All of our cases and controls were selected from white men over 60 years old. Risk for prostate cancer is 1 in 15 for men aged 60 through 69 years, and 1 in 8 for men aged 70 years and older (National Cancer Institute, 2013C). There are several publications about some comorbidities and behavioral parameters.

ClinGenWeb uses both complete and model based interpretations for clinicogenomic associations. Independent associations may have very little importance for clinical processes alone but in complete interpretation, we tried to interpret all relevant data as a whole. After our experimental approaches, we considered that cases and controls could be divided into two or three different risk groups due to genetic heterogeneity. With the commissioning of WGS in clinical practice, similarity measurements of clinically relevant SNP patterns may be a new way to produce predictive models in genomic medicine, but this approach need to be supported with more phenotypic data and to be tested in large study samples.

Some authors proposed several cumulative models to predict prostate cancer, but we couldn't acquire meaningful results with these models in our subjects. Another original approach was to use hybrid (SVM+ID3 decision tree) model based associations. Only SNP model of this approach was not successful but another model i.e. combining genomic, clinical and behavioral factors was partly consistent. This model was produced for African-American, Latin and Japan individuals, and we used a template for African-Americans. Unfortunately, holistic enviro-genomic models are quite limited.

Another critical point is that clinical, environmental and behavioral data can be used to explain pathogenic and clinical heterogeneity and to clarify the complexity of results. With the support of clinical and behavioral data, we could interpret some contradictory results. Because, most of the environmental and behavioral data wasn't stored in EMR/EHRs in a structured manner, we added these type of data in end-user level.

Due to bipartite structure of our interpretation (i.e. conversion of CR-SNP into clinicogenomic associations and final clinical interpretation of associations) and final interpretation was accomplished in the end-user side, we combined both clinicogenomic associations and external parameters (such as BMI) which have been recorded or tracking by end users for decision making.

9.2 General Comparison of our Model with Prior Works

GeneInsight Suite is a comprehensive application environment to evaluate and share sequencing based test results between stakeholders. GeneInsight Clinic can be integrated with EMR or stand alone, GeneInsignt-Lab manages knowledge, and facilitates reporting. GeneInsight Network (VariantWire) provides the mechanism to connect laboratories and providers. Using this system, interpretations of sequencing based tests are shared with corresponding caregiver organizations. GeneInsight Suite allows clinicians to receive updates when new information on previously unknown variants is certified for clinical use.

There are critical differences between our system and GeneInsight (Table 47).

Table 47: Comparison with prior work in the field

Characteristics	**Our Work**	**GeneInsight**
Aim	As a part of national level EHR (NHIS-T)	Integrate peer stakeholders (laboratory and hospitals)
Architecture	Central	Federated
Medical scope		
Variant type	SNP	All type variation data
Scope of clinical process	Predictive risk assessment	General
Medical condition	Prostate cancer	General
Terminology and standards		
Identifier of variant	Rs_id, allele, plus strand	HGVS nomenclature
Messaging standard	HL7 v3 CDA R2, encapsulated data	HL7 v2, text data
Disease terminology	SNOMED-CT	ICD-10
Knowledge base/ variant database		
Owner	National level, as a part of central NHIS-T	Partners healthcare
Sources for associations	CancerGAMAdb, SNPedia, NHGRI GWAS Catalog	Literature
Content extraction	Specifically produced based on Venice criteria	Manual curation
Extracted content	Associations with impact degree evidence category	Associations with impact degree and evidence category
Interpretation of relevant SNP	Automated	Manual reporting (?)
Impact value	Yes (numeric and categorical values)	Categorical definition
Evidence value	Yes (numeric and categorical values)	Categorical definition
Evidence assignment method	Yes (offered)	Not determined
Decision support application for end users		
Reporting of independent clinicogenomic associations	Yes	Yes
Reporting of polygenic scores	Yes	None
Reporting of model based genomic interpretation	Yes	None
Integration of external data (environmental, behavioral, etc.). by end users	Yes	None

In first, our system is designed to aim as the part of a central national level EHR namely, NHIS-T. But GeneInsight interconnect corresponding caregiver organizations and genomic laboratory to share genomic interpretations.

While the architecture of NHIS-T is has a central service oriented nature, the architecture of EHR systems in USA is more federated. Both systems contain knowledge base for the interpretation and end user applications to collect and present clinicogenomic reports.

Variant database of GeneInsight is supported by domain experts of Partners Healthcare. Interpretations are reviewed by expert persons and presented as a collective manner. But these do not contain separate and structured impact degree or magnitude of evidence. Instead of these sorts of qualifiers, these systems use only a classification system containing several categories e.g. pathogenic and emphasize possible phenotype with evidence category.

In GeneInsight, interpretation and re-interpretation of critical variants are reported for clinical use as a collective way. These interpretations do not contain external data which is not in EMR, so not capable to process these types of data. Also, other types of assessment techniques e.g. model based assessment and polygenic scoring are not possible.

But in our system we presented all these types of assessment approaches. Additionally, clinical interpretation of SNP data is divided into two sequential processes, i.e. conversion of CR-SNP into clinicogenomic associations and clinical interpretation of them. Therefore, final interpretation can be completed in end-user application and so it's possible to use other kinds of data for risk prediction (environmental, behavioral, etc.).

9.3 Limitations

Complete implementation of SNP data incorporated NHIS-T in real systems was not possible due to regulative and technical issues at this stage. So, we restricted our focus to develop complementary capabilities as prototypes for NHIS-T i.e. ClinGenKB and ClinGenWeb which specifically targeted prostate cancer risk prediction.

In our study, we used SNP data, but recent studies show that the different type of variants (CNV, etc.) may be more responsible for clinical conditions.

In ClinGenKB, our critical focus is to generate a structured clinicogenomic representation for only risk prediction for prostate cancer. But in literature, there are several kinds of information related to different stages of clinical decision processes e.g. prognosis, pharmacogenomic, etc. In the real world project, this prototype has to be enhanced with other kinds of associations and diseases.

GWAS is based on "common disease, common variant" hypothesis. But, some authors proposed that, common variants can explain only a modest part of complex diseases and "common disease, rare variant" hypothesis was put forward (Lake, et al., 2012). Clinicogenomic associations using our knowledge base, based on GWAS researches and publication about them.

We obtained case and control data from personal genome project to evaluate our system and number of cases and controls were so limited. To determine the value of this system in clinical settings, we need comprehensive genomic, environmental, family health data and clinical conditions. Unfortunately, none of cases and controls had family history data, and we couldn't involve this critical parameter in our evaluation processes. Existing clinical data

about subjects didn't reflect the clinical and pathological heterogeneity of prostate cancer. Especially, we have not precise measurement information (amount, period, duration, etc.) about behavioral characteristics of subjects (diet, physical activity, supplements, etc.) and we couldn't interpret the possible effects of these parameters on prostate cancer risk.

Another limitation is to align clinical and bioinformatics domain terminologies in a consistent way. ICD classification is accepted as a standard for disease classification in many countries including Turkey. But, ICD-10 is not useful to manage all levels of clinical, pathologic and genetic heterogeneities. It is expected that, next version of ICD i.e. ICD-11 will be released in 2015 and this version can be integrated other medical terminologies such as SNOMED CT (Zafar & Ezat, 2012). Nevertheless, some authors proposed that, it's an unavoidable requirement to develop a new taxonomy of disease which will be based on information commons and knowledge network including a combination of molecular, social, environmental and clinical data and health outcomes (National Research Council, 2011).

We have collected clinicogenomic associations from literature. Due to ethnic characteristics of our subjects, we preferred primarily studies performed with Caucasians. But, the terms of ethnicity and race are beyond biological distinctions, might refer to sociocultural construct and affected both biological and environmental factors. For this reason, for a real world NHIS-T system, we will need data for Turkish population.

We do not have sufficient predictive models that can be used in clinical settings. Especially, we need approaches to assess complete analysis of clinically relevant SNPs. With the commissioning of WGS in clinical practice, similarity measurements of clinically relevant SNP patterns may be a new way to produce predictive models in genomic medicine, but this approach need to be enhanced with more phenotypic data and to be tested in large study samples.

On the other hand, in the present, holistic enviro-genomic models are quite limited. Due to most of the complex diseases progress as interaction of genomic and environmental factors, additionally we need more enviro-genomic models predictive such diseases.

CHAPTER 10

CONCLUSION AND FUTURE WORKS

Today, the healthcare systems are continuously evolving and transforming under the influence of developments in technology and globalization. A revolutionary paradigm shifting is changing the focus of medicine from traditional provider-centric approach to patient-centric personalized medicine. This paradigm shifting radically transforms clinical processes, medical education, and researches in theory and praxis. The commissioning of new health services based on emerging technologies (mobile health systems, pervasive applications, environmental sensors, body area sensor networks, etc.) also dramatically support these tendencies.

But in the light of literature on personalized medicine, we can argue that, the area of biomedical informatics did not begin to perform its essential mission on healthcare systems and the major shifting in healthcare practices increasingly have getting closer via genomic technologies. When we look at the big picture, we can see the emergence of evidence based managed healthcare systems with knowledge discovery capabilities driven by big data and knowledge infrastructure for sustainable, fair and effective care services.

In this respect, we consider that the next generation of health information systems will be constructed based on tracking and monitoring all aspects of individual health status in 24/7 and turn evidence based recommendations to empower individuals. Today, most of personal behavioral and environmental data is not a subject of EMR/EHR and even PHR contents. Characteristics of most environmental and behavioral data required frequent measurements and (nearly) continuous tracking. And, possibly if we extent PHR content (with genomic data) towards to involve environmental and behavioral factors, we can add value to disease risk assessment and prediction.

As we emphasized before, a national level manually curated and accredited knowledge base is the most important component of evidence based decision making. Based on this knowledge base, collected risk data will gain a predictive meaningful, and improvements in clinical sciences will be reflected individuals by reinterpretation of collected data. At this point, we need additional and improved analytic tools based on genomic and environmental parameters. To make easier to extract and manually curate existing references for domain experts, we aim to develop a knowledge repository integrating several knowledge bases with semantic technologies and adding some automatic evaluation techniques.

In healthcare systems, regarding public health and financial burden, most of the important diseases are in the complex nature. In the pathogenesis of complex diseases, interaction of genetic and environmental factors have critical importance, and ethnicity, race and geographic factors may play distinctive roles. Hence, it's necessary to have appropriate clinicogenomic information about subjected population and use this content for right peoples. Clinical data, environmental factors and family history are critical components, and it's needed to study of relationships between these parameters and genomic factors. Eventually, it will be effective and reasonable way to both enhance NHIS-T data fields to

record structured data which will be used in enviro-genetic studies, and conduct researches to acquire original data for population.

Omics area is not only represented by genomic data and in the near future different types of omic data is expected to be added to the routine clinical practices e.g. transcriptomics, proteomics, metabolomics, and epigenomics. Also, systems medicine is an impressive approach and possibly will increase the effectiveness of risk prediction strategies.

In addition, we aim to enhance our system, by integrating data warehouses for research. With this capability, integrated genomic and environmental data sets can also be used for clinical research. We will extract the meaningful relationship patterns via this system and, by using these patterns, we can calculate risks of groups who have similar characteristics e.g. family members or communities.

The major aim of our system is to provide true and actionable information for patients and their family practitioners. Our system will return evidence based recommendations to the individuals processing collected data, and make them more responsible about their preferences and consequences. Empowerment of individuals to participate their healthcare decisions is an emerging trend in personalized medicine. At this point, we need more understandable information sources and visual representation approaches intended for unprofessional individuals. Area of representation and reporting of clinicogenomic results should be the focus to develop new approaches, techniques and tools.

In last 10 years of Turkey, there has been a great effort to accomplish a transformation in national healthcare system based on information technologies. But yet, practical applications of personal genomics and integration into healthcare services are in its infancy and studies about personalized medicine are in academic level.

Our architecture and prototype which aim to incorporate personal SNP data into NHIS-T is in the preliminary level. Although, we need additional visions, works and tools extending our EHR capabilities for the future genome enabled healthcare systems, we believe that our work will enable a starting point for national healthcare system.

REFERENCES

Aleksovska - Stojkovska, L. & Loskovska, S., 2011. *Architectural and data model of clinical decision support system for managing asthma in school-aged children.* Mankato, MN, IEEE International Conference on Electro/Information Technology (EIT).

Alspach, J. G., 2011. The importance of family health history: your patients' and your own. *Crit Care Nurse,* 31(1), pp. 10-15.

American Medical Association, 2010. *Report of the Council on Science and Public Health , Annual Meeting, Genomic-based Personalized Medicine,* s.l.: s.n.

American Medical Association, 2012. *Report of the Council on Science and Public Health , Interim Meeting, Clinical Application of Next Generation Genomic Sequencing,* s.l.: s.n.

Arnold, G. L. & Vockley, J., 2011. Thoroughly modern medicine. *Mol Genet Metab,* Issue 104, pp. 1-2.

Aronson, S. J. et al., 2011. GeneInsight Suite: a platform to support laboratory and provider use of DNA-based genetic testing. *Hum Mutat,* Issue 32, pp. 532-536.

Aronson, S. J. et al., 2012. Communicating new knowledge on previously reported genetic variants. *Genet Med.*

Asuncion , C. H. & van Sinderen, M. J., 2010. Pragmatic Interoperability: A Systematic Review of Published Definitions. In: P. Bernus, G. Doumeingts & M. Fox, eds. *Enterprise Architecture, Integration and Interoperability.* Brisbane, Australia: Springer, pp. 164-175.

Attia, J. et al., 2009A. How to use an article about genetic association: A: Background concepts. *JAMA,* 301(1), pp. 74-81.

Attia, J. et al., 2009B. How to use an article about genetic association: B: Are the results of the study valid?. *JAMA,* 301(2), pp. 191-197.

Attia, J. et al., 2009C. How to use an article about genetic association: C: What are the results and will they help me in caring for my patients?. *JAMA,* 301(3), pp. 304-308.

Ball, M. P. et al., 2012. *A public resource facilitating clinical use of genomes.* USA, s.n., pp. 109, 11920-11927.

Balshaw, D. M. & Kwok, R. K., 2012. Innovative methods for improving measures of the personal environment. *Am J Prev Med,* Issue 42, pp. 558-559.

Bamshad, M. J. et al., 2011. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet,* Issue 12, pp. 745-55.

Barnes, M. R., 2010. Genetic Variation Analysis for Biomedical Researchers: A Primer. In: M. R. Barnes & G. Breen, eds. *Genetic Variation: Methods and Protocols.* s.l.:Humana Press, pp. 1-20.

Belmont, J. & McGuire, A. L., 2009. The futility of genomic counseling: essential role of electronic health records. *Genome Med,* Issue 1, p. 48.

Benson, T., 2010. *Principles of Health Interoperability HL7 and SNOMED.* London: Springer-Verlag.

Berg, J. S., Khoury, M. J. & Evans, J. P., 2011. Deploying whole genome sequencing in clinical practice and public health: meeting the challenge one bin at a time. *Genet Med,* 13(6), pp. 499-504.

Beuten, J. et al., 2009. Single and multigenic analysis of the association between variants in 12 steroid hormone metabolism genes and risk of prostate cancer. *Cancer Epidemiol Biomarkers Prev,* 18(6), pp. 1869-80.

Beyan, O. D. et al., 2013. *Querying Phenotype-Genotype Associations across Multiple Knowledge Bases using Semantic Web Technologies.* Chania, Greece, 13th IEEE International Conference on BioInformatics and BioEngineering.

Bick, D. & Dimmock, D., 2011. Whole exome and whole genome sequencing. *Curr Opin Pediatr,* Issue 23, pp. 594-600.

Biesecker, L. G., 2012. Opportunities and challenges for the integration of massively parallel genomic sequencing into clinical practice: lessons from the ClinSeq project. *Genet Med,* Issue 14, pp. 393-398..

Bloss, C. S., Schork, N. J. & Topol, E. J., 2011. Effect of direct-to-consumer genomewide profiling to assess disease risk. *N Engl J Med,* 364(6), pp. 524-34.

Boone, K. W., 2011. *The CDA™ Book.* s.l.:Springer.

Boyd, L. K., Mao, X. & Lu, Y. L., 2012. The complexity of prostate cancer: genomic alterations and heterogeneity. *Nat Rev Urol,* 9(11), pp. 652-64.

Brown, S. M., 2009. *Essentials of medical genomics.* 2nd ed. Hoboken, New Jersey, US: JohnWiley & Sons, Inc.

Cancer Research UK, 2014. *Prostate cancer risks and causes.* [Online]
Available at: http://www.cancerresearchuk.org/cancer-help/type/prostate-cancer/about/prostate-cancer-risks-and-causes
[Accessed 14 01 2014].

Cariaso, M. & Lennon, G., 2012. SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Res,* pp. D1308-12.

Chan, I. S. & Ginsburg, G. S., 2011. Personalized medicine: progress and promise. *Annu Rev Genomics Hum Genet,* Issue 12, pp. 217-44.

Chowdhury, S. et al., 2013. Incorporating genomics into breast and prostate cancer screening: assessing the implications. *Genet Med,* 15(6), pp. 423-32.

Chua, E. W. & Kennedy, M. A., 2012. Current state and future prospects of direct-to-consumer pharmacogenetics. *Front Pharmacol,* Issue 3, p. 152.

Cirulli, E. T. & Goldstein, D. B., 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet,* 11(6), pp. 415-25.

Cordero, P. & Ashley, E. A., 2012. Whole-Genome Sequencing in Personalized Therapeutics. *Clin Pharmacol Ther,* 91(6), pp. 1001-9.

Crouch, D. J., Goddard, G. H. & Lewis, C. M., 2013. REGENT: a risk assessment and classification algorithm for genetic and environmental factors. *Eur J Hum Genet,* 21(1), pp. 109-11.

Dent T, et al., 2013. *Stratified screening for cancer: Recommendations and analysis from COGS,* Cambridge, UK: PHG Foundation.

Dogac, A. et al., 2011. Electronic health record interoperability as realized in the Turkish health information system. *Methods Inf Med,* 50(2), pp. 140-9.

Downing, G. J., 2009. Key aspects of health system change on the path to personalized medicine. *Transl Res,* Issue 154, pp. 272-276.

Drmanac, R., 2012. Medicine. The ultimate genetic test. *Science,* Issue 336, pp. 1110-1112.

Dziuda, D. M., 2010. *Data mining for genomics and proteomics, analysis of gene and protein expression data.* s.l.:John Wiley & Sons, Inc..

Evans, D. M., Visscher, p. M. & Wray, N. R., 2009. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum Mol Genet,* 18(18), pp. 3525-31.

Fardy, J. M., 2009. Evaluation of Diagnostic Tests. In: P. Parfrey & B. Barrett, eds. *Methods of Molecular Biology, Clinical Epidemiology, vol. 473.* Totowa, NJ, US: Humana Press, pp. 127-136.

Feero, W. G., Bigley, M. B. & Brinner, K. M., 2008. New Standards and Enhanced Utility for Family Health History Information in the Electronic Health Record: An Update from the American Health Information Community's Family Health History Multi-Stakeholder Workgroup. *J Am Med Inform Assoc,* Issue 15, pp. 723-728.

Fox, S. & Duggan, M., 2012. *Mobile health 2012, pew internet & American life project.* [Online]
Available at: http://www.pewinternet.org/Reports/2012/Mobile-Health.aspx
[Accessed 01 03 2013].

Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J., 2009. Human genetic variation and its contribution to complex traits. *Nat Rev Genet,* 10(4), pp. 241-51.

Garets, D. & Davis, M., 2006. *Electronic Medical Records vs. Electronic Health Records: Yes, There Is a Difference,* s.l.: HIMSS Analytics.

Gerhard, G. S., Carey, D. J. & Steele, G. D., 2013. Electronic Health Records in Genomic Medicine. In: *Genomic and Personalized Medicine.* s.l.:Elsevier Inc., pp. 287-294.

Ginsburg, G. S. & Willard, H. F., 2009. Genomic and personalized medicine, foundations and applications. *Transl Res,* Issue 154, pp. 277-287.

Ginsburg, G. S. & Willard, H. F., 2013. *Essentials of genomic and personalized medicine.* 2nd ed. s.l.:Academic Press.

Glaser, J. et al., 2008. Advancing personalized health care through health information technology: an update from the American Health Information Community's Personalized Health Care Workgroup. *J Am Med Inform Assoc,* Issue 15, pp. 391-6.

Green, R. C., Rehm, H. L. & Kohane, I. S., 2013. Clinical Genome Sequencing. In: *Genomic and Personalized Medicine.* 2nd ed. s.l.:Academic Press.

Gubb, E. & Matthiesen, R., 2010. Introduction to Omics. In: *Bioinformatics Methods in Clinical Research, Methods in Molecular Biology.* s.l.:Humana Press.

Gullapalli, R. R. et al., 2012. Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. *J Pathol Inform,* Issue 3, p. 40.

Guttmacher, A. E., Collins, F. S. & Carmona, R. H., 2004. The family history--more important than ever.. *N Engl J Med,* Issue 351, pp. 2333-2336.

Häyrinen, K., Saranto, K. & Nykänen, P., 2008. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *Int J Med Inform,* 77(5), pp. 291-304.

Helfand, B. T. et al., 2010. Genetic prostate cancer risk assessment: common variants in 9 genomic regions are associated with cumulative risk. *J Urol,* 184(2), pp. 501-5.

Helfand, B. T., Kan, D., Modi, P. & Catalona, W. J., 2011. Prostate cancer risk alleles significantly improve disease detection and are associated with aggressive features in patients with a "normal" prostate specific antigen and digital rectal examination. *Prostate,* 71(4), pp. 394-402.

Helgason, A. & Stefánsson, K., 2010. The past, present, and future of direct-to-consumer genetic tests. *Dialogues Clin Neurosci,* 12(1), pp. 61-8.

HIMSS Global Enterprise Task Force, 2010. *Electronic Health Records: A Global Perspective, Part 1,* s.l.: Healthcare Information and Management Systems Society (HIMSS).

HL7 Clinical Genomic Work Group, 2013. *Clinical Genomics (CG).* [Online]
Available at: http://wiki.hl7.org/index.php?title=CG
[Accessed 31 10 2013].

HL7 Clinical Genomics SIG, 2005. *The Dynamic Model, V0.2,* San Diego: s.n.

HL7 International, 2012. *HL7 Reference Information Model.* [Online]
Available at: http://www.hl7.org/implement/standards/rim.cfm
[Accessed 15 01 2014].

Hoffman, M. A., 2007. The genome-enabled electronic medical record. *J Biomed Inform,* pp. 44-6.

Hoffman, M., Arnoldi, C. & Chuang, I., 2005. The clinical bioinformatics ontology: a curated semantic network utilizing RefSeq information. *Pac Symp Biocomput,* pp. 139-50.

Hoffman, M. A. & Williams, M. S., 2011. Electronic medical records and personalized medicine. *Hum Genet,* 130(1), pp. 33-9.

Hudson, K. L., 2011. Genomics, Health Care, and Society. *N Engl J Med,* 365(11), pp. 1033-41.

Ioannidis, J. P. et al., 2008. Assessment of cumulative evidence on genetic associations: interim guidelines. *Int J Epidemiol,* 37(1), pp. 120-32.

Jacob, H. J. et al., 2013. Genomics in clinical practice: lessons from the front lines. *Sci Transl Med,* 5(194), p. 194cm5.

Janssens, A. C. & van Duijn, C. M., 2008. Genome-based prediction of common diseases: advances and prospects. *Hum Mol Genet,* 17(R2)(R166-73), pp. R166-73.

Janssens, A. C. & van Duijn, C. M., 2009. Genome-based prediction of common diseases:methodological considerations for future research. *Genome Med,* 1(2), pp. 20.1-20.9.

Jenicek, M., 2013. *A Primer on Clinical Experience in Medicine, Reasoning, Decision Making, and Communication in Health Sciences.* 1 ed. Boca Raton, Florida: Taylor & Francis Group, LLC.

Jing, X. et al., 2012. Incorporating personalized gene sequence variants, molecular genetics knowledge, and health knowledge into an EHR prototype based on the Continuity of Care Record standard. *J Biomed,* 45(1), pp. 82-92.

Jostins, L. & Barrett, J. C., 2011. Genetic risk prediction in complex disease. *Hum Mol Genet,* Issue 20(R2), pp. R182-8.

Kabak, Y. et al., 2008. *The use of HL7 CDA in the National Health Information System (NHIS) of Turkey.* Greece, Proceedings of 9th Int. HL7 Interoperability Conference.

Kahn, S. D., 2011. On the future of genomic data. *Science,* Issue 331, pp. 728-729.

Kalf, R. R. et al., 2013. Variations in predicted risks in personal genome testing for common complex diseases. *Genet Med.*

Kannry, J. & Williams, M. S., 2013. The undiscovered country, the future of integrating genomic information into the EHR. *Genet Med,* 15(10), pp. 824-5.

Kawamoto, K., Lobach, D. F., Willard, H. F. & Ginsburg, G. S., 2009. A national clinical decision support infrastructure to enable the widespread and consistent practice of genomic and personalized medicine. *BMC Med Inform Decis Mak,* 9(17).

Khoury, M. J., Janssens, A. C. & Ransohoff, D. F., 2013. How can polygenic inheritance be used in population screening for common diseases?. *Genet Med,* 15(6), pp. 437-43.

Khoury, M. J. et al., 2009. The Scientific Foundation for personal genomics: recommendations from a National Institutes of Health-Centers for Disease Control and Prevention multidisciplinary workshop. *Genet Med,* 11(8), pp. 559-67.

Kose, I. et al., 2008. *Turkey's National Health Information System (NHIS).* Stockholm, s.n., pp. 170-177.

Krist, A. H. et al., 2011. Designing a patient-centered personal health record to promote preventive care. *BMC Med Inform Decis Mak,* Issue 11, p. 73.

Lake, N. J., Bozaoglu, K., Khan, A. W. & Jowett, J. B. M., 2012. Approaches for dissection of the genetic basis of complex disease development in humans. In: *Genetic Diversity n Microorganisms.* s.l.:InTech, pp. 309-39.

Lautenbach, D. M., Christensen, K. D., Sparks, J. A. & Green, R. C., 2013. Communicating genetic risk information for common disorders in the era of genomic medicine. *Annu Rev Genomics Hum Genet,* Issue 14, pp. 491-513.

Lioy, P. J. & Rappaport, S. M., 2011. Exposure science and the exposome: an opportunity for coherence in the environmental health sciences. *Environ Health Perspect,* Issue 119, pp. A466-467.

Little, J. et al., 2009. STrengthening the REporting of Genetic Association Studies (STREGA): an extension of the STROBE statement. *PLoS Med,* 6(2), p. e22.

Little, J. et al., 2012. *Multigene Panels in Prostate Cancer Risk Assessment. Evidence Report No. 209,* Rockville (MD): Agency for Healthcare Research and Quality (US).

Liu, S. & Song, Y., 2010. Building genetic scores to predict risk of complex diseases in humans: is it possible?. *Diabetes,* 59(11), pp. 2729-31.

Macaskill, P. et al., 2010. Chapter 10: Analysing and Presenting Results. In: *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.* s.l.:The Cochrane Collaboration.

Majewski, J. et al., 2011. What can exome sequencing do for you?. *J Med Genet,* Issue 48, pp. 580-589.

Manolio, T. A., 2010. Genomewide association studies and assessment of the risk of disease. *Engl J Med,* 363(2), pp. 166-76.

Manolio, T. A., 2013. Bringing genome-wide association findings into clinical use. *Nat Rev Genet,* 14(8), pp. 549-58.

Manolio, T. A. et al., 2013. Implementing genomic medicine in the clinic: the future is here. *Genet Med,* 15(4), pp. 258-267.

Marian, A. J., 2011. Medical DNA Sequencing. *Curr Opin Cardiol,* 26(3), pp. 175-180.

Masys, D. R. et al., 2012. Technical desiderata for the integration of genomic data into electronic health records. *J Biomed Inform,* 45(3), pp. 419-22.

Merritt, D., 2004. *Best Practices for Rule-Based Application Development.* [Online] Available at: http://msdn.microsoft.com/en-us/library/aa480020.aspx [Accessed 31 10 2013].

Miksad, R. A. et al., 2006. Prostate cancer in a transgender woman 41 years after initiation of feminization. *JAMA,* Issue 296, pp. 2316-2317.

Mirnezami, R., Nicholson, J. & Darzi, A., 2012. Preparing for precision medicine. *N Engl J,* 366(6), pp. 489-91.

Nam, R. K. et al., 2009. Utility of incorporating genetic variants for the early detection of prostate cancer. *Clin Cancer Res,* 15(5), pp. 1787-93.

National Cancer Institute, 2013A. *Cancer Genetics Risk Assessment and Counseling.* [Online]
Available at: http://www.cancer.gov/cancertopics/pdq/genetics/risk-assessment-and-counseling/HealthProfessional/page3
[Accessed 31 10 2013].

National Cancer Institute, 2013B. *Risk Factors for Prostate Cancer Development.* [Online]
Available at:
http://www.cancer.gov/cancertopics/pdq/prevention/prostate/healthprofessional/page3
[Accessed 31 10 2013].

National Cancer Institute, 2013C. *Genetics of Prostate Cancer.* [Online]
Available at:
http://www.cancer.gov/cancertopics/pdq/genetics/prostate/healthprofessional#Section_5
[Accessed 31 10 2013].

National Library of Medicine (US), 2013. *Inheriting Genetic Conditions, Inheritance patterns and understanding risk.* [Online]
Available at: http://ghr.nlm.nih.gov/handbook/inheritance?show=all
[Accessed 4 11 2013].

National Research Council, 2008. *Evidence-Based Medicine and the Changing Nature of Health Care: Meeting Summary (IOM Roundtable on Evidence-Based Medicine).* Washington, DC: The National Academies Press.

National Research Council, 2009. *Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement,* Washington, DC: The National Academies Press.

National Research Council, 2011. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease.* Washington, DC: The National Academies Press.

National Research Council, 2012. *Integrating Large-Scale Genomic Information into Clinical Practice: Workshop Summary.* Washington, DC: The National Academies Press.

Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S., 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet,* 12(6), pp. 443-51.

Nusbaum, R. et al., 2013. Translational Genomic Research: Protocol Development and Initial Outcomes following SNP Testing for Colon Cancer Risk. *Transl Behav Med,* 3(1), pp. 17-29.

OECD, 2007. *OECD e-Government Studies: Turkey.* s.l.:OECD Publishing.

Oetting, W. S., 2009. Clinical genetics & human genome variation: the 2008 Human Genome Variation Society scientific meeting. *Hum Mutat,* 30(5), pp. 852-6.

Office of the National Coordinator for HIT, 2008. *Personalized Healthcare Draft Detailed Use Case,* s.l.: U.S. Department of Health and Human Services ONC.

Okeh, U. M. & Ogbonna, L. N., 2013. Statistical evaluation of indicators of diagnostic test performance. *American Journal of BioScience* , 1(4), pp. 63-73.

Paddock, C., 2013. *Self-tracking rools help you stay healthy.* [Online]
Available at: http://www.medicalnewstoday.com/articles/254902.php
[Accessed 01 06 2013].

Pashayan, N. et al., 2013. Public health genomics and personalized prevention: lessons from the COGS project.. *J Intern Med,* 274(5), pp. 451-6.

Pashayan, N. & Pharoah, P., 2012. Population-based screening in the era of genomics. *Per Med,* 9(4), pp. 451-455.

Pearson, T. A. & Manolio, T. A., 2008. How to interpret a genome-wide association study. *JAMA,* 299(11), pp. 1335-44.

Pierce, B. A., 2010. *Genetics: A Conceptual Approach.* Fourth Edition ed. s.l.:W. H. Freeman;.

Poo, D. C., Cai, S. & Mah, J. T., 2011. UASIS: Universal Automatic SNP Identification System. *BMC Genomics,* Issue 12 (Suppl 3), p. S9.

Raffan, E. & Semple, R. K., 2011. Next generation sequencing – implications for clinical practice. *Br Med Bull.,* Issue 99, pp. 53-71.

Rakhmawati, N. A., Karnstedt, M. & Hausenblas, M., 2013. *A Comparison of Federation over SPARQL Endpoints Frameworks.* St. Petersburg, Russia, Knowledge Engineering and Semantic Web Conference (KESW 2013).

Rappaport, S. M. & Smith, M. T., 2010. Environment and disease risks. *Science,* Issue 330, pp. 460-461.

Republic of Turkey Ministry of Health, 2012. *HCRS System Codes, National Health Information System (NHIS).* [Online]
Available at: https://skrs3.sagliknet.saglik.gov.tr
[Accessed 31 10 2013].

Republic of Turkey Ministry of Health, 2013. *National Health Data Dictionary 2.0 (In Turkish).* [Online]
Available at: http://www.e-saglik.gov.tr/dosyalar/USVS2_30032012.pdf
[Accessed 31 10 2013].

Ribick, A., 2010. *Health Level Seven Clinical Genomics Version 2 Messaging Standard Implementation Guide Successfully Transmits Genomic Data Electronically.* [Online]
Available at: http://www.hl7.org/documentcenter/public_temp_AFE9DD91-1C23-BA17-0C7631EE4E63EBA4/pressreleases/hl7_press_20100119.pdf
[Accessed 1 11 2013].

Riegelman, R., 2010. *Public Health 101: Healthy People-Healthy Populations (Essential Public Health).* s.l.:Jones & Bartlett Learning.

Roden, D. M. & Tyndale, R. F., 2013. Genomic medicine, precision medicine, personalized medicine: what's in a name?. *Clin Pharmacol Ther,* 94(2), pp. 169-72.

Röhm, U. & Blakeley, J. A., 2009. *Data Management for High-Throughput Genomics.* USA, Association for Computing Machinery (ACM).

Salari, K., Watkins, H. & Ashley, E. A., 2012. Personalized medicine: hope or hype?. *Eur Heart J,* 33(13), pp. 1564-70.

Salinas, C. A. et al., 2009. Clinical utility of five genetic variants for predicting prostate cancer risk and mortality. *Prostate,* 69(4), pp. 363-72.

Sartor, A. O., 2013. *Risk factors for prostate cancer.* [Online]
Available at: http://www.uptodate.com/contents/risk-factors-for-prostate-cancer

Schaaf, C. F., Zschocke, J. & Potocki, L., 2012. *Human genetics : from molecules to medicine.* s.l.:Lippincott Williams & Wilkins, a Wolters Kluwer.

Scheuner, M. T. et al., 2009. Are electronic health records ready for genomic medicine. *Genet Med,* Issue 11, pp. 510-517.

Schneider, M. V. & Orchard, S., 2011. Omics Technologies, Data and Bioinformatics Principles. In: *Bioinformatics for Omics Data, Methods and Protocols.* s.l.:HumanaPress.

Schully, S. D. et al., 2011. Cancer GAMAdb: database of cancer genetic associations from meta-analyses and genome-wide association studies. *Eur J Hum Genet,* 19(8), pp. 928-30.

Sethi, P. & Theodos, K., 2009. Translational bioinformatics and healthcare informatics: computational and ethical challenges. *Perspect Health Inf Manag,* Issue 6, p. 1h.

Shabo, A., 2006. Clinical genomics data standards for pharmacogenetics and pharmacogenomics. *Pharmacogenomics,* Issue 7, pp. 247-253.

Shabo, A. et al., 2009. *HL7 Version 2 Implementation Guide: Clinical Genomics; Fully Loinc-Qualified Genetic Variation Model, Release 1 (1st Informative Ballot), HL7 Version 2.5.1.*. [Online]
Available at:
http://wiki.hl7.org/images/2/24/V2_CG_LOINCGENVAR_R1_I2_2009MAY.pdf

Shimokawa, K. et al., 2011. iCOD: an integrated clinical omics database based on the systems-pathology view. *BMC Genomics,* 11(Suppl 4), p. S19.

Shoenbill, K., Fost, N., Tachinardi, U. & Mendonca, E. A., 2013. Genetic data and electronic health records: a discussion of ethical, logistical and technological considerations. *J Am Med Inform Assoc,* p. [Epub ahead of print].

SNPedia, 2005. *Rs3218536.* [Çevrimiçi]
Available at: http://www.snpedia.com/index.php/Rs3218536
[Erişildi: 31 10 2013].

SNPedia, 2007. *Rs713041.* [Online]
Available at: http://www.snpedia.com/index.php/Rs713041
[Accessed 31 10 2013].

SNPedia, 2008. *Prostate Cancer.* [Online]
Available at: http://snpedia.com/index.php/Prostate_cancer
[Accessed 31 10 2013].

SNPedia, 2008. *Rs1050565.* [Online]
Available at: http://www.snpedia.com/index.php/Rs1050565
[Accessed 31 10 2013].

SNPedia, 2008. *Rs11045585.* [Online]
Available at: http://www.snpedia.com/index.php/Rs11045585
[Accessed 31 10 2013].

SNPedia, 2010. *Rs3790844.* [Online]
Available at: http://www.snpedia.com/index.php/Rs3790844
[Accessed 31 10 2013].

SNPedia, 2011. *Rs1052133.* [Online]
Available at: http://snpedia.com/index.php/Rs1052133
[Accessed 31 10 2013].

SNPedia, 2011. *Rs798766.* [Online]
Available at: http://www.snpedia.com/index.php/Rs798766
[Accessed 31 10 2013].

Starren, J. et al., 2012. *Crossing the Omic Chasm: Integrating Omic Data into the EHR.* San Francisco, AMIA.

Starren, J., Williams, M. S. & Bottinger, E. P., 2013. Crossing the omic chasm: a time for omic ancillary systems. *JAMA,* Issue 309, pp. 1237-1238.

Stepanov, V. A., 2010. Genomes, populations and diseases: ethnic genomics and personalized medicine. *Acta Naturae,* 2(4), pp. 15-30.

Stranger, B. E., Stahl, E. A. & Raj, T., 2011. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics,* 187(2), pp. 367-83.

Swan, M., 2012. Sensor mania! the internet of things, wearable computing, objective metrics, and the quantified self 2.0. *J Sens Actuator Netw,* 1(1), p. 217–253.

Tanaka, H., 2010. Omics-based medicine and systems pathology. A new perspective for personalized and predictive medicine. *Methods Inf Med,* 49(2), pp. 173-85.

Thomas, P. E. et al., 2011. Challenges in the association of human single nucleotide polymorphism mentions with unique database identifiers. *BMC Bioinformatics,* 12(Suppl 4), p. S4..

Thorn, C. F., Klein, T. E. & Altman, R. B., 2010. Pharmacogenomics and bioinformatics: PharmGKB. *Pharmacogenomics,* 11(4), pp. 501-5.

Tolk, A., Diallo, S. Y., King, R. D. & Turnitsa, C. D., 2009. A Layered Approach to Composition and Interoperation in Complex Systems. In: A. Tolk & L. C. Jain, eds. *Complex Systems in Knowledge-based Environments: Theory, Models and Applications.* Berlin Heidelberg: Springer-Verlag, pp. 41-74.

Tran, B. Q. & Gonzales, P., 2012. Standards and guidelines for personal health records in the United States: finding consensus in a rapidly evolving and divided environment. *J Health Med Informat.*

Ullman-Cullere, M. H. & Mathew, J. P., 2011. Emerging landscape of genomics in the electronic health record for personalized medicine. *Hum Mutat,* Issue 211, pp. 512-516.

Ury, A. G., 2013. Storing and interpreting genomic information in widely deployed electronic health record systems. *Genet Med,* 15(10), pp. 779-85.

Van Allen, E. M., Wagle, N. & Levy, M. A., 2013. Clinical analysis and interpretation of cancer genome data. *J Clin Oncol,* 31(15), pp. 1825-33.

Van Tongeren, M. & Cherrie, J. W., 2012. An integrated approach to the exposome. *Environ Health Perspect,* Issue A103–104, p. 120.

Verdonk, P. & Klinge, I., 2012. Mainstreaming sex and gender analysis in public health genomics. *Gend Med,* Issue 9, pp. 402-410.

Wei, C., Anthony, J. C. & Lu, Q., 2012. Genome-environmental risk assessment of cocaine dependence. *Front Genet,* 18(3), p. 83.

Weitzel, J. N. et al., 2011. Genetics, genomics, and cancer risk assessment, state of the art and future directions in the era of personalized medicine. *CA Cancer J Clin,* Issue 61, p. 327–359.

Welch, B. M. & Kawamoto, K., 2013. Clinical decision support for genetically guided personalized medicine: a systematic review. *J Am Med Inform Assoc,* 20(2), pp. 388-400.

Welter, D. et al., 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res,* 42(1), pp. D1001-6.

Whirl-Carrillo, M. et al., 2012. Pharmacogenomics Knowledge for Personalized Medicine. *Clin Pharmacol Ther,* 92(4), pp. 414-7.

Wild, C. P., 2012. The exposome: from concept to utility. *Int J Epidemiol,* Issue 41, pp. 24-32.

Winter, A. et al., 2011. *Health Information Systems, Architectures and Strategies.* 2nd ed. London: Springer-Verlag.

Wray, N. R., Goddard, M. E. & Visscher, P. M., 2007. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res,* 17(10), pp. 1520-8.

Wray, N. R., Goddard, M. E. & Visscher, P. M., 2008. Prediction of individual genetic risk of complex disease. *Curr Opin Genet Dev,* 18(3), pp. 257-63.

Wright, C. et al., 2011. *Next steps in the sequence, the implications of whole genome sequencing for health in the UK.* Cambridge: PHG Foundation.

Wu, J., Pfeiffer, R. M. & Gail, M. H., 2013. Strategies for developing prediction models from genome-wide association studies. *Genet Epidemiol,* 37(8), pp. 768-77.

Yarnall, J. M., Crouch, D. J. & Lewis, C. M., 2013. Incorporating non-genetic risk factors and behavioural modifications into risk prediction models for colorectal cancer. *Cancer Epidemiol,* 37(3), pp. 324-9.

Yücebaş, S. C. & Aydın Son, Y., 2014. A prostate cancer model build by a novel SVM-ID3 hybrid feature selection method using both genotyping and phenotype data from dbGaP. *PLOS One,* 9(3), p. e91404..

Zafar, A. & Ezat, W. P., 2012. Development of ICD 11: changes and challenges. *BMC Health Services Research,* 12(Suppl 1), p. I8.

Zheng, S. L. et al., 2008. Cumulative association of five genetic variants with prostate cancer. *N Engl J Med,* 358(9), pp. 910-9.

APPENDICES

APPENDIX A-) Complete List of Independent Associations

assoc_id	Chr	Gene	rs_id	Allele	variant_id	OR (CI)	cat(MI)	QE	cat(QE)
1	1	MTR	rs1805087	AG	rs1805087-AG	1.06 (1.00-1.12)	Weak	2,2	Moderate
2	1	MTR	rs1805087	GG	rs1805087-GG	1.06 (1.00-1.12)	Weak	2,2	Moderate
3	1	PTGS2	rs2745557	AG	rs2745557-AG	1.16 (1.00-1.33)	Weak	1,3	Weak
4	1	PTGS2	rs2745557	AA	rs2745557-AA	1.16 (1.00-1.33)	Weak	1,3	Weak
5	1	RNASEL	rs486907	CT	rs486907-CT	1.30 (1.1-1.5)	Weak	1,7	Moderate
6	1	RNASEL	rs486907	TT	rs486907-TT	1.30 (1.1-1.5)	Weak	1,7	Moderate
7	1	RNASEL	rs627928	AA	rs627928-AA	1.27 (1.13-1.44)	Weak	1,8	Moderate
8	1	RNASEL	rs627928	AC	rs627928-AC	1.27 (1.13-1.44)	Weak	1,8	Moderate
9	2	ITGA6	rs12621278	AG	rs12621278-AG	1.35 (1.27-1.44)	Weak	2,2	Moderate
10	2	ITGA6	rs12621278	AA	rs12621278-AA	1.35 (1.27-1.44)	Weak	2,2	Moderate
11	2	C2orf43	rs13385191	AG	rs13385191-AG	1.15 (1.10-1.21)	Weak	1,7	Moderate
12	2	C2orf43	rs13385191	GG	rs13385191-GG	1.15 (1.10-1.21)	Weak	1,7	Moderate
13	2	THADA	rs1465618	CT	rs1465618-CT	1.15 (1.04-1.26)	Weak	2,2	Moderate
14	2	THADA	rs1465618	TT	rs1465618-TT	1.15 (1.04-1.26)	Weak	2,2	Moderate
15	2	MLPH	rs2292884	AG	rs2292884-AG	1.14 (1.09-1.19)	Weak	2,2	Moderate
16	2	MLPH	rs2292884	GG	rs2292884-GG	1.14 (1.09-1.19)	Weak	2,2	Moderate
17	2	EHBP1	rs2710646	AC	rs2710646-AC	1.13 (1.04-1.22)	Weak	1,7	Moderate
18	2	EHBP1	rs2710646	AA	rs2710646-AA	1.13 (1.04-1.22)	Weak	1,7	Moderate
19	2	SRD5A2	rs523349	CC	rs523349-CC	1.11 (1.03-1.19)	Weak	1,8	Moderate
20	2	SRD5A2	rs523349	CG	rs523349-CG	1.11 (1.03-1.19)	Weak	1,8	Moderate

assoc_id	Chr	Gene	rs_id	Allele	variant_id	OR (CI)	cat(MI)	QE	cat(QE)
21	2	EHBP1	rs721048	AG	rs721048-AG	1.15 (1.10-1.21)	Weak	2,5	Strong
22	2	EHBP1	rs721048	AA	rs721048-AA	1.15 (1.10-1.21)	Weak	2,5	Strong
23	2		rs7584330	AG	rs7584330-AG	1.06 (1.02-1.09])	Weak	2,2	Moderate
24	2		rs7584330	GG	rs7584330-GG	1.06 (1.02-1.09])	Weak	2,2	Moderate
25	3	EEFSEC	rs10934853	AC	rs10934853-AC	1.12 (1.06-1.18)	Weak	2,3	Strong
26	3	EEFSEC	rs10934853	AA	rs10934853-AA	1.12 (1.06-1.18)	Weak	2,3	Strong
27	3	CLDN11	rs10936632	AC	rs10936632-AC	1.11(1.08-1.14)	Weak	2	Moderate
28	3	CLDN11	rs10936632	AA	rs10936632-AA	1.11(1.08-1.14)	Weak	2	Moderate
29	3		rs17023900	AG	rs17023900-AG	1.41 ()	Weak	1,2	Weak
30	3		rs17023900	GG	rs17023900-GG	1.41 ()	Weak	1,2	Weak
31	3	VGLL3	rs2660753	CT	rs2660753-CT	1.18 (1.06-1.31)	Weak	2,3	Strong
32	3	VGLL3	rs2660753	TT	rs2660753-TT	1.18 (1.06-1.31)	Weak	2,3	Strong
33	3	ZBTB38	rs6763931	AG	rs6763931-AG	1.04 (1.01-1.07)	Weak	2	Moderate
34	3	ZBTB38	rs6763931	AA	rs6763931-AA	1.04 (1.01-1.07)	Weak	2	Moderate
35	3		rs7629490	CT	rs7629490-CT	1.06 (1.04-1.09)	Weak	2,2	Moderate
36	3		rs7629490	TT	rs7629490-TT	1.06 (1.04-1.09)	Weak	2,2	Moderate
37	4	PDLIM5	rs12500426	AC	rs12500426-AC	1.08 (1.05-1.12)	Weak	2,3	Strong
38	4	PDLIM5	rs12500426	AA	rs12500426-AA	1.08 (1.05-1.12)	Weak	2,3	Strong
39	4	PDLIM5	rs17021918	CT	rs17021918-CT	1.14 (1.10-1.18)	Weak	2,2	Moderate
40	4	PDLIM5	rs17021918	TT	rs17021918-TT	1.14 (1.10-1.18)	Weak	2,2	Moderate

assoc_id	Chr	Gene	rs_id	Allele	variant_id	OR (CI)	cat(MI)	QE	cat(QE)
41	4	TET2	rs7679673	AC	rs7679673-AC	1.14 (1.09-1.20)	Weak	2,2	Moderate
42	4	TET2	rs7679673	AA	rs7679673-AA	1.14 (1.09-1.20)	Weak	2,2	Moderate
43	5		rs12653946	CT	rs12653946-CT	1.39 (1.22-1.57)	Weak	1,7	Moderate
44	5		rs12653946	TT	rs12653946-TT	1.39 (1.22-1.57)	Weak	1,7	Moderate
45	5	FGF10	rs2121875	AC	rs2121875-AC	1.05 (1.02-1.08)	Weak	2	Moderate
46	5	FGF10	rs2121875	CC	rs2121875-CC	1.05 (1.02-1.08)	Weak	2	Moderate
47	5	TERT	rs2242652	AG	rs2242652-AG	1.15 (1.11-1.19)	Weak	2	Moderate
48	5	TERT	rs2242652	AA	rs2242652-AA	1.15 (1.11-1.19)	Weak	2	Moderate
49	5	FGFR4	rs351855	AG	rs351855-AG	1.23 (1.08-1.40)	Weak	2	Moderate
50	5	FGFR4	rs351855	AA	rs351855-AA	1.23 (1.08-1.40)	Weak	2	Moderate
51	6	CCHCR1	rs130067	GT	rs130067-GT	1.05 (1.02-1.09)	Weak	2	Moderate
52	6	CCHCR1	rs130067	GG	rs130067-GG	1.05 (1.02-1.09)	Weak	2	Moderate
53	6	FOXP4	rs1983891	CT	rs1983891-CT	1.15 (1.09-1.21)	Weak	1,7	Moderate
54	6	FOXP4	rs1983891	TT	rs1983891-TT	1.15 (1.09-1.21)	Weak	1,7	Moderate
55	6	RFX6	rs339331	CT	rs339331-CT	1.22 (1.15-1.28])	Weak	1,7	Moderate
56	6	RFX6	rs339331	TT	rs339331-TT	1.22 (1.15-1.28])	Weak	1,7	Moderate
57	6	SOD2	rs4880	AG	rs4880-AG	1.12 (1.00–1.25)	Weak	2	Moderate
58	6	SOD2	rs4880	GG	rs4880-GG	1.18 (0.97–1.44)	Weak	2	Moderate
59	6	SLC22A1	rs651164	AG	rs651164-AG	1.15 (1.10-1.20)	Weak	2,2	Moderate
60	6	SLC22A1	rs651164	AA	rs651164-AA	1.15 (1.10-1.20)	Weak	2,2	Moderate

assoc_id	Chr	Gene	rs_id	Allele	variant_id	OR (CI)	cat(MI)	QE	cat(QE)
61	6	ESR1	rs9340799	AG	rs9340799-AG	1.31 (1.03-1.63)	Weak	1,8	Moderate
62	6	ESR1	rs9340799	GG	rs9340799-GG	1.31 (1.03-1.63)	Weak	1,8	Moderate
63	6	SLC22A3	rs9364554	CT	rs9364554-CT	1.17 (1.13-1.20)	Weak	1,7	Moderate
64	6	SLC22A3	rs9364554	TT	rs9364554-TT	1.17 (1.13-1.20)	Weak	1,7	Moderate
65	7	JAZF1	rs10486567	AG	rs10486567-AG	1.12 (1.02-1.25)	Weak	2,5	Strong
66	7	JAZF1	rs10486567	GG	rs10486567-GG	1.12 (1.02-1.25)	Weak	2,5	Strong
67	7	CYP3A4	rs2740574	CT	rs2740574-CT	9.30 (1.30-411)	Strong	1	Weak
68	7	CYP3A4	rs2740574	TT	rs2740574-TT	11.9 (1.6-533)	Strong	1	Weak
69	7	LMTK2	rs6465657	CT	rs6465657-CT	1.14 (1.05-1.23)	Weak	2,5	Strong
70	7	LMTK2	rs6465657	CC	rs6465657-CC	1.14 (1.05-1.23)	Weak	2,5	Strong
71	8	PCAT1	rs10086908	CC	rs10086908-CC	1.13 (1.08-1.19)	Weak	2,2	Moderate
72	8	PCAT1	rs10086908	CT	rs10086908-CT	1.13 (1.08-1.19)	Weak	2,2	Moderate
73	8		rs10090154	CT	rs10090154-CT	1.72 (1.58-1.86)	Weak	1,5	Weak
74	8		rs10090154	TT	rs10090154-TT	1.72 (1.58-1.86)	Weak	1,5	Weak
75	8	PCAT2	rs1016343	CT	rs1016343-CT	1.31 (1.20-1.42)	Weak	1,8	Moderate
76	8	PCAT2	rs1016343	TT	rs1016343-TT	1.31 (1.20-1.42)	Weak	1,8	Moderate
77	8	NKX3-1	rs10503733	GT	rs10503733-GT	1.29 ()	Weak	1,2	Weak
78	8	NKX3-1	rs10503733	TT	rs10503733-TT	1.29 ()	Weak	1,2	Weak
79	8		rs10505483	GT	rs10505483-GT	1.73 ()	Weak	1,3	Weak
80	8		rs10505483	TT	rs10505483-TT	1.73 ()	Weak	1,3	Weak

assoc_id	Chr	Gene	rs_id	Allele	variant_id	OR (CI)	cat(MI)	QE	cat(QE)
81	8	PCAT2	rs13252298	AG	rs13252298-AG	1.12 (1.05-1.18)	Weak	2,2	Moderate
82	8	PCAT2	rs13252298	AA	rs13252298-AA	1.12 (1.05-1.18)	Weak	2,2	Moderate
83	8		rs13254738	AG	rs13254738-AG	1.59 (1.38-1.84)	Weak	1,5	Weak
84	8		rs13254738	GG	rs13254738-GG	1.59 (1.38-1.84)	Weak	1,5	Weak
85	8	CASC8	rs1447295	AC	rs1447295-AC	1.47 (1.33-1.62)	Weak	2,7	Strong
86	8	CASC8	rs1447295	AA	rs1447295-AA	1.47 (1.33-1.62)	Weak	2,7	Strong
87	8	NKX3-1	rs1512268	CT	rs1512268-CT	1.17 (1.14–1.21)	Weak	2,2	Moderate
88	8	NKX3-1	rs1512268	TT	rs1512268-TT	1.17 (1.14–1.21)	Weak	2,2	Moderate
89	8		rs16901979	AA	rs16901979-AA	1.79 (1.53-2.11)	Weak	2,5	Strong
90	8		rs16901979	AC	rs16901979-AC	1.80 (1.55-2.09)	Weak	2,5	Strong
91	8	SLC25A37	rs2928679	AG	rs2928679-AG	1.13 (1.02-1.25)	Weak	2,2	Moderate
92	8	SLC25A37	rs2928679	AA	rs2928679-AA	1.13 (1.02-1.25)	Weak	2,2	Moderate
93	8		rs4242382	AG	rs4242382-AG	2.18 (1.57-3.02)	Moderate	1,7	Moderate
94	8		rs4242382	AA	rs4242382-AA	2.18 (1.57-3.02)	Moderate	1,7	Moderate
95	8		rs4242384	AC	rs4242384-AC	1.88 ()	Weak	2,3	Strong
96	8		rs4242384	CC	rs4242384-CC	1.88 ()	Weak	2,3	Strong
97	8	CASC8	rs445114	CT	rs445114-CT	1.22 (1.12-1.32)	Weak	2,3	Strong
98	8	CASC8	rs445114	TT	rs445114-TT	1.22 (1.12-1.32)	Weak	2,3	Strong
99	8	CASC8	rs620861	AG	rs620861-AG	1.16 (1.11-1.20)	Weak	2,2	Moderate
100	8	CASC8	rs620861	GG	rs620861-GG	1.16 (1.11-1.20)	Weak	2,2	Moderate

assoc_id	Chr	Gene	rs_id	Allele	variant_id	OR (CI)	cat(MI)	QE	cat(QE)
101	8	PCAT2	rs6470494	CT	rs6470494-CT	1.14 (1.09-1.19)	Weak	1,3	Weak
102	8	PCAT2	rs6470494	TT	rs6470494-TT	1.14 (1.09-1.19)	Weak	1,3	Weak
103	8	CASC8	rs6983267	GT	rs6983267-GT	1.31 (1.27-1.35)	Weak	2,7	Strong
104	8	CASC8	rs6983267	GG	rs6983267-GG	1.31 (1.27-1.35)	Weak	2,7	Strong
105	8		rs6983561	AC	rs6983561-AC	1.45 (1.30-1.61)	Weak	2,3	Strong
106	8		rs6983561	CC	rs6983561-CC	1.45 (1.30-1.61)	Weak	2,3	Strong
107	8	CASC8	rs7000448	CT	rs7000448-CT	1.4 (1.28-1.53)	Weak	1,3	Weak
108	8	CASC8	rs7000448	TT	rs7000448-TT	1.4 (1.28-1.53)	Weak	1,3	Weak
109	8		rs7837688	GT	rs7837688-GT	1.55 (1.39-1.76)	Weak	2	Moderate
110	8		rs7837688	TT	rs7837688-TT	1.55 (1.39-1.76)	Weak	2	Moderate
111	9	DAB2IP	rs1571801	GT	rs1571801-GT	1.27 (1.10-1.48)	Weak	2	Moderate
112	9	DAB2IP	rs1571801	TT	rs1571801-TT	1.27 (1.10-1.48)	Weak	2	Moderate
113	10	MSMB	rs10993994	CT	rs10993994-CT	1.18 (0.96-1.51)	Weak	2,3	Strong
114	10	MSMB	rs10993994	TT	rs10993994-TT	1.18 (0.96-1.51)	Weak	2,3	Strong
115	10	TCF7L2	rs12255372	GT	rs12255372-GT	1.05 (1.00-1.10)	Weak	2,3	Strong
116	10	TCF7L2	rs12255372	TT	rs12255372-TT	1.05 (1.00-1.10)	Weak	2,3	Strong
117	10	MGMT	rs2308321	GG	rs2308321-GG	2.02 (1.06-3.85)	Moderate	1,7	Moderate
118	10	MGMT	rs2308321	AG	rs2308321-AG	1.22 (1.01-1.47)	Weak	1,7	Moderate
119	10	CYP17A1	rs2486758	CT	rs2486758-CT	1.07 (1.00-1.14)	Weak	1,8	Moderate
120	10	CYP17A1	rs2486758	CC	rs2486758-CC	1.09 (0.95-1.26)	Weak	1,8	Moderate

assoc_id	Chr	Gene	rs_id	Allele	variant_id	OR (CI)	cat(MI)	QE	cat(QE)
121	10	CTBP2	rs4962416	CT	rs4962416-CT	1.28 (1.09-1.52)	Weak	1,5	Weak
122	10	CTBP2	rs4962416	CC	rs4962416-CC	1.28 (1.09-1.52)	Weak	1,5	Weak
123	10	WBP1L	rs6892	AG	rs6892-AG	1.08 (1.00-1.15)	Weak	1,8	Moderate
124	10	WBP1L	rs6892	GG	rs6892-GG	1.11 (0.95-1.30)	Weak	1,8	Moderate
125	10	CYP17A1	rs743572	AG	rs743572-AG	2.02 (1.14-3.58)	Moderate	1,5	Weak
126	10	CYP17A1	rs743572	GG	rs743572-GG	1.62 (1.10-2.37)	Weak	1,5	Weak
127	10	TIMM23B	rs7920517	AG	rs7920517-AG	1.23 (1.10-1.38)	Weak	1,5	Weak
128	10	TIMM23B	rs7920517	GG	rs7920517-GG	1.23 (1.10-1.38)	Weak	1,5	Weak
129	11	MYEOV	rs10896449	AG	rs10896449-AG	1.16 (1.11-1.22)	Weak	2,3	Strong
130	11	MYEOV	rs10896449	GG	rs10896449-GG	1.16 (1.11-1.22)	Weak	2,3	Strong
131	11	MYEOV	rs11228565	AG	rs11228565-AG	1.23 (1.16-1.31)	Weak	2,2	Moderate
132	11	MYEOV	rs11228565	AA	rs11228565-AA	1.23 (1.16-1.31)	Weak	2,2	Moderate
133	11	TH	rs7127900	AG	rs7127900-AG	1.25 (1.20-1.30)	Weak	2,2	Moderate
134	11	TH	rs7127900	AA	rs7127900-AA	1.25 (1.20-1.30)	Weak	2,2	Moderate
135	11		rs7130881	AG	rs7130881-AG	1.31 (1.20-1.44)	Weak	2,2	Moderate
136	11		rs7130881	GG	rs7130881-GG	1.31 (1.20-1.44)	Weak	2,2	Moderate
137	11	MYEOV	rs7931342	GT	rs7931342-GT	1.19 (1.11-1.27)	Weak	2,5	Strong
138	11	MYEOV	rs7931342	GG	rs7931342-GG	1.19 (1.11-1.27)	Weak	2,5	Strong
139	12	TUBA1C	rs10875943	CT	rs10875943-CT	1.07 (1.04-1.10)	Weak	2	Moderate
140	12	TUBA1C	rs10875943	CC	rs10875943-CC	1.07 (1.04-1.10)	Weak	2	Moderate

assoc_id	Chr	Gene	rs_id	Allele	variant_id	OR (CI)	cat(MI)	QE	cat(QE)
141	12	VDR	rs1544410	CT	rs1544410-CT	1.12 (1.00-1.25)	Weak	2	Moderate
142	12	VDR	rs1544410	TT	rs1544410-TT	1.12 (1.00-1.25)	Weak	2	Moderate
143	12	VDR	rs2107301	AA	rs2107301-AA	2.50 (1.52-4.00)	Strong	1,7	Moderate
144	12	VDR	rs2238135	GG	rs2238135-GG	1.95 (1.17-3.26)	Weak	1,7	Moderate
145	12	VDR	rs7975232	AC	rs7975232-AC	1.25 (1.02-1.53)	Weak	1,8	Moderate
146	12	VDR	rs7975232	CC	rs7975232-CC	1.25 (1.02-1.53)	Weak	1,8	Moderate
147	12	KRT8	rs902774	AG	rs902774-AG	1.17 (1.11-1.24)	Weak	2,2	Moderate
148	12	KRT8	rs902774	AA	rs902774-AA	1.17 (1.11-1.24)	Weak	2,2	Moderate
149	13		rs9600079	GT	rs9600079-GT	1.18 (1.12-1.24])	Weak	1,7	Moderate
150	13		rs9600079	TT	rs9600079-TT	1.18 (1.12-1.24])	Weak	1,7	Moderate
151	15		rs1042524	AG	rs1042524-AG	3.18 (1.02-9.90)	Strong	1,7	Moderate
152	15		rs1042524	AA	rs1042524-AA	3.18 (1.02-9.90)	Strong	1,7	Moderate
153	15	CYP1A1	rs1048943	GG	rs1048943-GG	1.31 (1.14-1.51)	Weak	1,7	Moderate
154	15	CYP1A1	rs1048943	AG	rs1048943-AG	1.31 (1.14-1.51)	Weak	1,7	Moderate
155	15	CYP1A1	rs4646903	AG	rs4646903-AG	1.33 (1.10-1.61)	Weak	1,5	Weak
156	15	CYP1A1	rs4646903	GG	rs4646903-GG	1.33 (1.10-1.61)	Weak	1,5	Weak
157	17	HNF1B	rs11649743	AG	rs11649743-AG	1.16 (1.11-1.22)	Weak	2,3	Strong
158	17	HNF1B	rs11649743	GG	rs11649743-GG	1.16 (1.11-1.22)	Weak	2,3	Strong
159	17	HOXB13	rs138213197	CT	rs138213197-CT	3.40 (2.2–5.4)	Strong	1,7	Moderate
160	17	HOXB13	rs138213197	TT	rs138213197-TT	3.40 (2.2–5.4)	Strong	1,7	Moderate

assoc_id	Chr	Gene	rs_id	Allele	variant_id	OR (CI)	cat(MI)	QE	cat(QE)
161	17		rs1859962	GT	rs1859962-GT	1.27 (1.18-1.37)	Weak	2,8	Strong
162	17		rs1859962	GG	rs1859962-GG	1.27 (1.18-1.37)	Weak	2,8	Strong
163	17	SHMT1	rs1979277	AG	rs1979277-AG	1.11 (1.00-1.22)	Weak	2	Moderate
164	17	SHMT1	rs1979277	AA	rs1979277-AA	1.11 (1.00-1.22)	Weak	2	Moderate
165	17	HNF1B	rs3760511	GT	rs3760511-GT	1.17 (1.11-1.23)	Weak	1,5	Weak
166	17	HNF1B	rs3760511	GG	rs3760511-GG	1.17 (1.11-1.23)	Weak	1,5	Weak
167	17		rs4054823	TT	rs4054823-TT	1.26 (1.16–1.36)	Weak	1,8	Moderate
168	17	HNF1B	rs4430796	AG	rs4430796-AG	1.22 (1.17-1.26)	Weak	2,8	Strong
169	17	HNF1B	rs4430796	AA	rs4430796-AA	1.22 (1.17-1.26)	Weak	2,8	Strong
170	17	ELAC2	rs4792311	AA	rs4792311-AA	1.12 (1.00-1.25)	Weak	1,8	Moderate
171	17	ELAC2	rs4792311	AG	rs4792311-AG	1.12 (1.00-1.25)	Weak	1,8	Moderate
172	17		rs6501455	AG	rs6501455-AG	1.1 (1.00-1.22)	Weak	1,5	Weak
173	17		rs6501455	AA	rs6501455-AA	1.1 (1.00-1.22)	Weak	1,5	Weak
174	17	ZNF652	rs7210100	AG	rs7210100-AG	1.51 (1.35-1.69)	Weak	1,7	Moderate
175	17	ZNF652	rs7210100	AA	rs7210100-AA	1.51 (1.35-1.69)	Weak	1,7	Moderate
176	17		rs7214479	CT	rs7214479-CT	1.16 (1.11-1.23)	Weak	1,3	Weak
177	17		rs7214479	TT	rs7214479-TT	1.16 (1.11-1.23)	Weak	1,3	Weak
178	17	HNF1B	rs7501939	CT	rs7501939-CT	1.19 (1.11-1.28)	Weak	2,2	Moderate
179	17	HNF1B	rs7501939	CC	rs7501939-CC	1.19 (1.11-1.28)	Weak	2,2	Moderate
180	19		rs103294	CT	rs103294-CT	1.28 (1.21-1.36)	Weak	1,7	Moderate

assoc_id	Chr	Gene	rs_id	Allele	variant_id	OR (CI)	cat(MI)	QE	cat(QE)
181	19		rs103294	CC	rs103294-CC	1.28 (1.21-1.36)	Weak	1,7	Moderate
182	19		rs11672691	AG	rs11672691-AG	1.11 (1.02-1.20)	Weak	2	Moderate
183	19		rs11672691	GG	rs11672691-GG	1.11 (1.02-1.20)	Weak	2	Moderate
184	19	TGFB1	rs1800470	AG	rs1800470-AG	1.28 (1.01-1.61)	Weak	1,8	Moderate
185	19	TGFB1	rs1800470	GG	rs1800470-GG	1.24 (1.02-1.52)	Weak	1,8	Moderate
186	19	XRCC1	rs25487	CT	rs25487-CT	1.48 (1.05-2.03)	Weak	1,3	Weak
187	19	XRCC1	rs25487	CC	rs25487-CC	1.48 (1.05-2.03)	Weak	1,3	Weak
188	19	KLK3	rs2735839	AG	rs2735839-AG	1.3 (1.11-1.51)	Weak	2,3	Strong
189	19	KLK3	rs2735839	GG	rs2735839-GG	1.3 (1.11-1.51)	Weak	2,3	Strong
190	19	PPP1R14A	rs8102476	CT	rs8102476-CT	1.12 (1.08-1.15)	Weak	2,3	Strong
191	19	PPP1R14A	rs8102476	CC	rs8102476-CC	1.12 (1.08-1.15)	Weak	2,3	Strong
192	19		rs887391	CT	rs887391-CT	1.15 (1.09-1.21)	Weak	2,2	Moderate
193	19		rs887391	TT	rs887391-TT	1.15 (1.09-1.21)	Weak	2,2	Moderate
194	21	RUNX1	rs928285	AG	rs928285-AG	1.24 (1.01-1.53)	Weak	2	Moderate
195	21	RUNX1	rs928285	AA	rs928285-AA	1.24 (1.01-1.53)	Weak	2	Moderate
196	22	TNRC6B	rs11704416	CG	rs11704416-CG	1.06 (1.02-1.10)	Weak	2	Moderate
197	22	TNRC6B	rs11704416	CC	rs11704416-CC	1.06 (1.02-1.10)	Weak	2	Moderate
198	22	COMT	rs4680	AG	rs4680 -AG	1.10 (0.9301.30)	Weak	1,8	Moderate
199	22	COMT	rs4680	AA	rs4680 -AA	1.10 (0.9301.30)	Weak	1,8	Moderate
200	22	BIK	rs5759167	GT	rs5759167-GT	1.18 (1.14-1.21)	Weak	2,2	Moderate

assoc_id	Chr	Gene	rs_id	Allele	variant_id	OR (CI)	cat(MI)	QE	cat(QE)
201	22	BIK	rs5759167	TT	rs5759167-TT	1.18 (1.14-1.21)	Weak	2,2	Moderate
202	22	TNRC6B	rs9623117	CT	rs9623117-CT	1.13 (1.05-1.22)	Weak	2,2	Moderate
203	22	TNRC6B	rs9623117	CC	rs9623117-CC	1.13 (1.05-1.22)	Weak	2,2	Moderate
204	23(X)	AR	rs5919432	CT	rs5919432-CT	1.06 (1.02-1.12)	Weak	2	Moderate
205	23(X)	AR	rs5919432	TT	rs5919432-TT	1.06 (1.02-1.12)	Weak	2	Moderate
206	23(X)	NUDT11	rs5945572	AG	rs5945572-AG	1.26 (1.19-1.34)	Weak	1,8	Moderate
207	23(X)	NUDT11	rs5945572	AA	rs5945572-AA	1.26 (1.19-1.34)	Weak	1,8	Moderate
208	23(X)	NUDT11	rs5945619	CT	rs5945619-CT	1.28 (1.11-1.47)	Weak	1,5	Weak
209	23(X)	NUDT11	rs5945619	CC	rs5945619-CC	1.28 (1.11-1.47)	Weak	1,5	Weak

Descriptions:

assoc_id: Association identifier

Chr: Chromosome

rs_id: rs number of SNP from dbSNP

variant_id: Variant identifier of SNP allele (include rs number and allele value)

OR (CI): Odds ratio (confidence interval)

Cat (MI): Category of magnitude of impact

QE: Quality of evidence

Cat (QE): Category of quality of evidence

APPENDIX B-) Reference Tables for Cumulative Models

Model Name	Count of risk factors	Odds Ratio (CI)	Reference (PubMed ID)
17-SNP_Helfand (only SNP, without F/H)	0-4	1.00 (by definition)	20860009
	5	1.6 (0.4-6.3)	
	6	1.0 (0.3-3.7)	
	7	1.4 (0.4-4.9)	
	8	1.4 (0.4-5.0)	
	9	2.5 (0.7-8.9)	
	10	3.1 (0.9-11.5)	
	>10	10.6 (2.7-42.0)	
17-SNP_Helfand (with F/H)	0-5	1.00 (by definition)	
	6	0.5 (0.2-1.3)	
	7	1.6 (0.7-3.5)	
	8	1.3 (0.6-2.8)	
	9	1.7 (0.8-3.8)	
	10	3.7 (1.6-8.4)	
	>10	11.2 (4.3-29.2)	
9-SNP_Helfand (only SNP)	0-1	1.00 (by definition)	20620408
	2	1.46 (0.74-2.86)	
	3	2.46 (1.29-4.66)	
	4	3.05 (1.60-5.79)	
	5	4.39 (2.24-8.61)	
	>6	5.75 (2.50-13.24)	
5-SNP_Zheng (only SNP, without F/H)	0	1.00 (by definition)	18199855
	1	1.50 (CI: 1.18-1.92)	
	2	1.96 (CI: 1.54-2.49)	
	3	2.21 (CI: 1.70-2.89)	
	4	4.47 (CI: 2.93-6.80)	
	5	4.47 (CI: 2.93-6.80)	
5-SNP_Zheng (with F/H)	0	1.00 (by definition)	
	1	1.62 (CI: 1.27-2.08)	
	2	2.07 (CI: 1.62-2.64)	
	3	2.71 (CI: 2.08-3.53)	
	4	4.76 (CI: 3.31-6.84)	
	5	9.46 (CI: 3.62-24.72)	
	6	9.46 (CI: 3.62-24.72)	
5-SNP_Salinas (only SNP, without F/H)	0	1.00 (by definition)	19058137
	1	1.48 (1.09-2.01)	
	2	1.88 (1.38-2.56)	
	3	2.97 (2.08-4.26)	
	>=4	3.36 (1.90-6.08)	
5-SNP_Salinas (with F/H)	0	1.00 (by definition)	
	1	1.41 (1.02-1.97)	
	2	2.25 (1.63-3.13)	
	3	3.43 (2.40-4.94)	
	4	3.65 (2.24-6.03)	
	>=5	4.92 (1.58-18.53)	

Model Name	Count of risk factors	Odds Ratio (CI)	Reference (PubMed ID)
4-SNP_Nam	0	1.00 (by definition)	19223501
	1	1.26 (1.0-1.6)	
	2	1.61 (1.3-2.1)	
	3	3.05 (2.0-4.6)	
	4	3.81 (1.2-12.3)	
3-SNP_Beuten	0	1.00 (by definition)	19505920
	1	1.39 (1.0-1.9)	
	2	1.56 (1.11-2.20)	
	3	2.87 (1.64-5.02)	

Descriptions:
CI: Confidence interval
PubMed ID: Identifier of scientific reference from PubMed

APPENDIX C-) Decision Tree Structure of First Hybrid Model Based Associations (Only SNP Model)

```
rs11720239 = AA
|  rs2999081 = AG
|  |  rs2811518 = AG
|  |  |  rs4793790 = AA: 1 {1=3, 2=0}
|  |  |  rs4793790 = AG: 2 {1=0, 2=1}
|  |  |  rs4793790 = GG: 1 {1=1, 2=0}
|  rs2999081 = CT: 2 {1=0, 2=16}
|  rs2999081 = CC: 2 {1=0, 2=89}
|  rs2999081 = GG
|  |  rs2811518 = AA
|  |  |  rs4793790 = AA
|  |  |  |  rs2811415 = AG
|  |  |  |  |  rs6798749 = GG
|  |  |  |  |  |  rs463967 = AG: 1 {1=3, 2=0}
|  |  |  |  |  |  rs463967 = GG
|  |  |  |  |  |  |  rs693913 = GG
|  |  |  |  |  |  |  |  rs585513 = AG: 1 {1=1, 2=0}
|  |  |  |  |  |  |  |  rs585513 = GG: 2 {1=0, 2=1}
|  |  |  |  rs2811415 = GG
|  |  |  |  |  rs6798749 = GG
|  |  |  |  |  |  rs463967 = AA
|  |  |  |  |  |  |  rs693913 = AA
|  |  |  |  |  |  |  |  rs585513 = AG: 2 {1=0, 2=2}
|  |  |  |  |  |  |  |  rs585513 = GG
|  |  |  |  |  |  |  |  |  rs2960482 = AA
|  |  |  |  |  |  |  |  |  |  rs2103869 = GG
|  |  |  |  |  |  |  |  |  |  |  rs3774796 = AA
|  |  |  |  |  |  |  |  |  |  |  |  rs1496306 = GG
|  |  |  |  |  |  |  |  |  |  |  |  |  rs7838995 = AA
|  |  |  |  |  |  |  |  |  |  |  |  |  |  rs7845891 = AA
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  rs6549458 = AA
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  rs9857492 = AA
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  rs2132528 = AA
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  rs17571004 = AG
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  rs2811388 = GG
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  rs569072 = GG
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  rs11079162 = AA
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  rs10420793 = AA
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  rs6992847 = AA
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  rs12797573 = AA
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  rs9643617 = AA
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  rs4921943 = AG: 1 {1=1, 2=0}
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  rs4921943 = GG: 2 {1=0, 2=1}
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  rs10420793 = AG: 1 {1=1, 2=0}
|  |  |  |  |  |  |  |  |  |  rs3774796 = AG: 1 {1=1, 2=0}
|  |  |  |  |  |  rs693913 = GG
|  |  |  |  |  |  |  rs585513 = AA: 1 {1=1, 2=0}
|  |  |  |  |  |  |  rs585513 = AG
```

| | | | | | | | | rs2960482 = AA
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | | rs17571004 = AG
| | | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | | rs569072 = AG
| | | | | | | | | | | | | | | | | | | | | rs11079162 = AA
| | | | | | | | | | | | | | | | | | | | | | rs10420793 = AA
| | | | | | | | | | | | | | | | | | | | | | | rs6992847 = AA
| | | | | | | | | | | | | | | | | | | | | | | | rs12797573 = AG
| | | | | | | | | | | | | | | | | | | | | | | | | rs9643617 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = AG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = GG: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | | | | | rs17571004 = GG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | rs6549458 = AC: 1 {1=2, 2=0}
| | | | | | | | | | | | | | | | | | rs7845891 = AC: 1 {1=1, 2=0}
| | | | | | | | | | rs2960482 = AG: 1 {1=3, 2=0}
| | | | | | | | | rs585513 = GG
| | | | | | | | | | rs2960482 = AA
| | | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | | | rs17571004 = AA: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | rs17571004 = AG
| | | | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | | | rs569072 = GG
| | | | | | | | | | | | | | | | | | | | | | rs11079162 = AA
| | | | | | | | | | | | | | | | | | | | | | | rs10420793 = AA
| | | | | | | | | | | | | | | | | | | | | | | | rs6992847 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | rs12797573 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | | rs9643617 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = GG
| | | | | | | | | | | | | | | | | | | | | | | | | | | | rs4384231 = AG: 1 {1=1,
2=0}
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | rs4384231 = GG: 2 {1=0,
2=1}
| | | | | | | | | | | | | | | | | | | | | rs17571004 = xx: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | rs6549458 = AC: 1 {1=1, 2=0}
| | | | | | | | | | | rs3774796 = AG: 1 {1=2, 2=0}
| | | | | | | | | rs2960482 = AG
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA

| | | | | | | | | | | | rs1496306 = AG
| | | | | | | | | | | | | rs7838995 = AG
| | | | | | | | | | | | | | rs7845891 = AC
| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | rs2132528 = AG
| | | | | | | | | | | | | | | | | rs17571004 = AA
| | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | rs569072 = GG
| | | | | | | | | | | | | | | | | | | | rs11079162 = AA
| | | | | | | | | | | | | | | | | | | | | rs10420793 = AA: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | | | rs10420793 = GG: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | | | | rs11079162 = AG: 1 {1=2, 2=0}
| | | | | | | | | | | | | | | | | | | rs17571004 = AG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | rs6549458 = CC: 1 {1=1, 2=0}
| | | | | | rs463967 = AG
| | | | | | | rs693913 = AA
| | | | | | | | rs585513 = AA: 2 {1=0, 2=1}
| | | | | | | | rs585513 = AG
| | | | | | | | | rs2960482 = AA
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | | rs7845891 = AA: 1 {1=4, 2=0}
| | | | | | | | | | | | | | | | | | rs7845891 = AC: 2 {1=0, 2=1}
| | | | | | | | | rs2960482 = AG: 1 {1=2, 2=0}
| | | | | | | | | rs2960482 = GG: 1 {1=2, 2=0}
| | | | | | | | rs585513 = GG
| | | | | | | | | rs2960482 = AA
| | | | | | | | | | rs2103869 = AG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | rs17571004 = AA: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | rs17571004 = AG: 2 {1=0, 2=1}
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | rs17571004 = AA
| | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | rs569072 = GG
| | | | | | | | | | | | | | | | | | | | rs11079162 = AA
| | | | | | | | | | | | | | | | | | | | | rs10420793 = AA: 1 {1=3, 2=0}

| | | | | | | | | | | | | | | | | | | | | | rs10420793 = AG: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | | rs17571004 = AG: 1 {1=4, 2=0}
| | | | | | | | | | | | | | | | | | rs17571004 = GG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | rs6549458 = AC: 1 {1=1, 2=0}
| | | | | | | | | rs2960482 = AG: 1 {1=2, 2=0}
| | | | | | | rs693913 = GG
| | | | | | | | rs585513 = AG
| | | | | | | | | rs2960482 = AA
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | | rs17571004 = AA: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | rs17571004 = AG
| | | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | | rs569072 = AG
| | | | | | | | | | | | | | | | | | | | | rs11079162 = AA
| | | | | | | | | | | | | | | | | | | | | | rs10420793 = AA
| | | | | | | | | | | | | | | | | | | | | | | rs6992847 = AA
| | | | | | | | | | | | | | | | | | | | | | | | rs12797573 = AA: 1 {1=2, 2=0}
| | | | | | | | | | | | | | | | | | | | | | | | rs12797573 = AG
| | | | | | | | | | | | | | | | | | | | | | | | | rs9643617 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = GG
| | | | | | | | | | | | | | | | | | | | | | | | | | | rs4384231 = AA: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | | | | | | | | | rs4384231 = AG: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | | | | | | | | | | | rs4384231 = GG
| | | | | | | | | | | | | | | | | | | | | | | | | | | | rs4147525 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | rs4680416 = AG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | rs4680416 = GG: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | | rs17571004 = GG: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | rs6549458 = AC
| | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | | rs17571004 = AA: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | rs17571004 = AG: 2 {1=0, 2=1}
| | | | | | | | | | | | | | rs7845891 = AC: 1 {1=1, 2=0}
| | | | | | | | | | | rs3774796 = AG: 1 {1=2, 2=0}
| | | | | | | | | rs2960482 = AG: 1 {1=3, 2=0}
| | | | | | | | | rs2960482 = GG: 1 {1=1, 2=0}
| | | | | | | | rs585513 = GG
| | | | | | | | | rs2960482 = AA
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | | rs7838995 = AA

| | | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | | rs17571004 = AG
| | | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | | rs569072 = AG: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | | | | rs569072 = GG
| | | | | | | | | | | | | | | | | | | | | rs11079162 = AA
| | | | | | | | | | | | | | | | | | | | | rs10420793 = AA
| | | | | | | | | | | | | | | | | | | | | | rs6992847 = AA
| | | | | | | | | | | | | | | | | | | | | | | rs12797573 = AA
| | | | | | | | | | | | | | | | | | | | | | | | rs9643617 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = AA: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = AG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = GG: 1 {1=3, 2=0}
| | | | | | | | | | | | | | | | | | | | | | | | rs10420793 = AG
| | | | | | | | | | | | | | | | | | | | | | | | rs6992847 = AA
| | | | | | | | | | | | | | | | | | | | | | | | rs12797573 = AA
| | | | | | | | | | | | | | | | | | | | | | | | rs9643617 = AA
| | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = GG
| | | | | | | | | | | | | | | | | | | | | | | | | rs4384231 = GG
| | | | | | | | | | | | | | | | | | | | | | | | | | rs4147525 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | | | rs4680416 = GG
| | | | | | | | | | | | | | | | | | | | | | | | | | | | rs6779266 = CC
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | rs9462806 = AA: 1
{1=2, 2=0}
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | rs9462806 = AG: 2
{1=0, 2=1}
| | | | | | | | | | | | | | | | | | | | | | | | | rs11079162 = AG: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | | | | | rs17571004 = GG
| | | | | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | | | | rs569072 = GG
| | | | | | | | | | | | | | | | | | | | | | | rs11079162 = AA
| | | | | | | | | | | | | | | | | | | | | | | rs10420793 = AG
| | | | | | | | | | | | | | | | | | | | | | | | rs6992847 = AA
| | | | | | | | | | | | | | | | | | | | | | | | rs12797573 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | rs9643617 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = AG
| | | | | | | | | | | | | | | | | | | | | | | | | | rs4384231 = AG: 2 {1=0,
2=1}
| | | | | | | | | | | | | | | | | | | | | | | | | | rs4384231 = GG: 1 {1=1,
2=0}
| | | | | | | | | | | | | | rs6549458 = AC: 1 {1=1, 2=0}
| | | | | | | | | | | | | rs7845891 = AC
| | | | | | | | | | | | | | rs6549458 = AA: 2 {1=0, 2=1}
| | | | | | | | | | | | | | rs6549458 = AC: 1 {1=1, 2=0}
| | | | | | | | | rs2960482 = AG
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = AG
| | | | | | | | | | | | | rs7838995 = AA: 1 {1=1, 2=0}
| | | | | | | | | | | | | rs7838995 = AG

| | | | | | | | | | | | | | rs7845891 = AC
| | | | | | | | | | | | | | | rs6549458 = AA: 1 {1=4, 2=0}
| | | | | | | | | | | | | | | rs6549458 = AC: 2 {1=0, 2=2}
| | | | | | | | | | | | | | rs7845891 = CC: 1 {1=1, 2=0}
| | | | | | | | | rs2960482 = GG: 1 {1=3, 2=0}
| | | | | | rs463967 = GG
| | | | | | | rs693913 = AA
| | | | | | | | rs585513 = AG: 1 {1=1, 2=0}
| | | | | | | | rs585513 = GG
| | | | | | | | | rs2960482 = AA
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | | rs17571004 = AA: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | rs17571004 = AG: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | rs6549458 = AC: 1 {1=2, 2=0}
| | | | | | | | | rs2960482 = AG
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = AG
| | | | | | | | | | | | | rs7838995 = AG
| | | | | | | | | | | | | | rs7845891 = AC
| | | | | | | | | | | | | | | rs6549458 = AA: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | rs6549458 = AC: 1 {1=4, 2=0}
| | | | | | rs693913 = GG
| | | | | | | rs585513 = AA
| | | | | | | | rs2960482 = AA: 1 {1=1, 2=0}
| | | | | | | | rs2960482 = AG: 2 {1=0, 2=2}
| | | | | | | rs585513 = AG
| | | | | | | | rs2960482 = AA
| | | | | | | | | rs2103869 = GG
| | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | rs7845891 = AA: 1 {1=7, 2=0}
| | | | | | | | | | | | | rs7845891 = AC: 2 {1=0, 2=1}
| | | | | | | | | rs2960482 = AG
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = AG
| | | | | | | | | | | | | rs7838995 = AG
| | | | | | | | | | | | | | rs7845891 = AC
| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | rs2132528 = AG
| | | | | | | | | | | | | | | | | | rs17571004 = AA
| | | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | | rs569072 = AG: 2 {1=0, 2=1}

| | | | | | | | | | | | | | | | | | | | | | | rs569072 = GG: 1 {1=3, 2=0}
| | | | | | | | | | | | | | | | | | | | | | rs17571004 = AG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | | | | rs17571004 = GG: 1 {1=1, 2=0}
| | | | | | | | | | | rs3774796 = AG: 1 {1=1, 2=0}
| | | | | | | | | rs585513 = GG
| | | | | | | | | rs2960482 = AA
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | | rs17571004 = AA
| | | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | | rs569072 = GG
| | | | | | | | | | | | | | | | | | | | | rs11079162 = AA
| | | | | | | | | | | | | | | | | | | | | | rs10420793 = AA: 1 {1=4, 2=0}
| | | | | | | | | | | | | | | | | | | | | | rs10420793 = AG
| | | | | | | | | | | | | | | | | | | | | | | rs6992847 = AA
| | | | | | | | | | | | | | | | | | | | | | | | rs12797573 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | | rs9643617 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = AA: 2 {1=0, 2=2}
| | | | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = AG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = GG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | | | | | | | | rs10420793 = GG: 1 {1=2, 2=0}
| | | | | | | | | | | | | | | | | | | | | | rs17571004 = AG: 1 {1=3, 2=0}
| | | | | | | | | | | | | | | | | | | | | | | rs17571004 = GG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | rs6549458 = AC
| | | | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | | | | rs17571004 = AG
| | | | | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | | | | rs569072 = GG
| | | | | | | | | | | | | | | | | | | | | | | rs11079162 = AA
| | | | | | | | | | | | | | | | | | | | | | | | rs10420793 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | rs6992847 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | | rs12797573 = AA: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | | | | | | | | | | rs12797573 = AG: 1 {1=2, 2=0}
| | | | | | | | | | | | | | | | | rs6549458 = CC
| | | | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | | | | rs17571004 = AA: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | | | | | rs17571004 = AG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | | rs7845891 = AC: 2 {1=0, 2=1}
| | | | | | | | | | | rs3774796 = AG: 1 {1=1, 2=0}
| | | | | | | | | rs2960482 = AG
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = AG
| | | | | | | | | | | | | rs7838995 = AG
| | | | | | | | | | | | | | rs7845891 = AC

| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | rs2132528 = AG
| | | | | | | | | | | | | | | | | | rs17571004 = AA
| | | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | | rs569072 = GG
| | | | | | | | | | | | | | | | | | | | | rs11079162 = AA
| | | | | | | | | | | | | | | | | | | | | | rs10420793 = AA
| | | | | | | | | | | | | | | | | | | | | | | rs6992847 = AG
| | | | | | | | | | | | | | | | | | | | | | | | rs12797573 = AA: 1 {1=2, 2=0}
| | | | | | | | | | | | | | | | | | | | | | | | | rs12797573 = xx: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | | | rs17571004 = AG
| | | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | | rs569072 = GG
| | | | | | | | | | | | | | | | | | | | | rs11079162 = AA
| | | | | | | | | | | | | | | | | | | | | | rs10420793 = AA: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | | | | | | rs10420793 = AG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | rs17571004 = GG: 2 {1=0, 2=3}
| | | | | | | | | | rs3774796 = AG: 2 {1=0, 2=1}
| | | | | | | | | rs2960482 = GG: 1 {1=2, 2=0}
| | | | rs2811415 = xx: 2 {1=0, 2=1}
| | | rs4793790 = AG
| | | | rs2811415 = AG
| | | | | rs6798749 = GG
| | | | | | rs463967 = AA: 1 {1=3, 2=0}
| | | | | | rs463967 = AG: 1 {1=1, 2=0}
| | | | | | rs463967 = GG
| | | | | | | rs693913 = GG
| | | | | | | | rs585513 = AG: 2 {1=0, 2=1}
| | | | | | | | rs585513 = GG: 1 {1=1, 2=0}
| | | | rs2811415 = GG
| | | | | rs6798749 = GG
| | | | | | rs463967 = AA
| | | | | | | rs693913 = AA
| | | | | | | | rs585513 = AA
| | | | | | | | | rs2960482 = AG
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = AG
| | | | | | | | | | | | | rs7838995 = AG
| | | | | | | | | | | | | | rs7845891 = AC
| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | rs2132528 = AG
| | | | | | | | | | | | | | | | | | rs17571004 = AA: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | rs17571004 = AG: 2 {1=0, 2=1}
| | | | | | | | rs585513 = AG
| | | | | | | | | rs2960482 = AG
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = AG
| | | | | | | | | | | | | rs7838995 = AG
| | | | | | | | | | | | | | rs7845891 = AC

| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | rs2132528 = AG
| | | | | | | | | | | | | | | | | | rs17571004 = AA: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | rs17571004 = GG: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | rs6549458 = AC: 1 {1=2, 2=0}
| | | | | | | | | rs585513 = GG
| | | | | | | | | | rs2960482 = AA: 1 {1=1, 2=0}
| | | | | | | | | | rs2960482 = AG: 2 {1=0, 2=1}
| | | | | | | | rs693913 = GG
| | | | | | | | | rs585513 = AA
| | | | | | | | | | rs2960482 = AA
| | | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | | | rs17571004 = AG
| | | | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | | | rs569072 = AA
| | | | | | | | | | | | | | | | | | | | | | rs11079162 = AG
| | | | | | | | | | | | | | | | | | | | | | | rs10420793 = AA: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | | | | | | | rs10420793 = AG: 1 {1=1, 2=0}
| | | | | | | | rs585513 = AG
| | | | | | | | | rs2960482 = AA
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | | rs17571004 = AA: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | rs17571004 = AG
| | | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | | rs569072 = AG
| | | | | | | | | | | | | | | | | | | | | rs11079162 = AG
| | | | | | | | | | | | | | | | | | | | | | rs10420793 = AA
| | | | | | | | | | | | | | | | | | | | | | | rs6992847 = AA
| | | | | | | | | | | | | | | | | | | | | | | | rs12797573 = AG
| | | | | | | | | | | | | | | | | | | | | | | | | rs9643617 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = AG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = GG: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | rs17571004 = GG: 1 {1=1, 2=0}
| | | | | | | | | | rs3774796 = AG: 1 {1=2, 2=0}
| | | | | | | | | rs2960482 = AG: 2 {1=0, 2=1}
| | | | | | | | rs585513 = GG
| | | | | | | | | rs2960482 = AA
| | | | | | | | | | rs2103869 = GG

| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | | rs17571004 = AA
| | | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | | rs569072 = GG
| | | | | | | | | | | | | | | | | | | | | rs11079162 = AG
| | | | | | | | | | | | | | | | | | | | | | rs10420793 = AA: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | | | | | | rs10420793 = AG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | | | rs17571004 = AG: 1 {1=5, 2=0}
| | | | | | | | | | | | | | | | | | | | | rs17571004 = GG
| | | | | | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | | | | rs569072 = GG
| | | | | | | | | | | | | | | | | | | | | | | rs11079162 = AG
| | | | | | | | | | | | | | | | | | | | | | | rs10420793 = AA
| | | | | | | | | | | | | | | | | | | | | | | | rs6992847 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | rs12797573 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | | rs9643617 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = GG
| | | | | | | | | | | | | | | | | | | | | | | | | | | | rs4384231 = AG: 1 {1=1,
2=0}
| | | | | | | | | | | | | | | | | | | | | | | | | | | | rs4384231 = GG: 2 {1=0,
2=1}
| | | | | | | | | | | | | | | rs6549458 = AC: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | rs7845891 = AC: 1 {1=1, 2=0}
| | | | | | | | | rs2960482 = AG
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = AG: 1 {1=1, 2=0}
| | | | | | | | | | | | rs1496306 = GG: 2 {1=0, 2=1}
| | | | | | rs463967 = AG
| | | | | | | rs693913 = AA
| | | | | | | | rs585513 = AA: 1 {1=1, 2=0}
| | | | | | | | rs585513 = AG
| | | | | | | | | rs2960482 = AA
| | | | | | | | | | rs2103869 = AG: 1 {1=1, 2=0}
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | | rs17571004 = AA: 2 {1=0, 2=2}
| | | | | | | | | | | | | | | | | | | rs17571004 = AG: 1 {1=1, 2=0}
| | | | | | | | | rs2960482 = AG: 1 {1=1, 2=0}
| | | | | | | | | rs585513 = GG: 1 {1=6, 2=0}
| | | | | | | rs693913 = GG

| | | | | | | | rs585513 = AA: 1 {1=4, 2=0}
| | | | | | | | rs585513 = AG
| | | | | | | | | rs2960482 = AA
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | | rs17571004 = AA: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | rs17571004 = AG
| | | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | | rs569072 = AG
| | | | | | | | | | | | | | | | | | | | | rs11079162 = AG
| | | | | | | | | | | | | | | | | | | | | | rs10420793 = AA
| | | | | | | | | | | | | | | | | | | | | | | rs6992847 = AA
| | | | | | | | | | | | | | | | | | | | | | | | rs12797573 = AG
| | | | | | | | | | | | | | | | | | | | | | | | | rs9643617 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = GG
| | | | | | | | | | | | | | | | | | | | | | | | | | | rs4384231 = AG
| | | | | | | | | | | | | | | | | | | | | | | | | | | | rs4147525 = AG: 1 {1=1,
2=0}
| | | | | | | | | | | | | | | | | | | | | | | | | | | | rs4147525 = GG: 2 {1=0,
2=2}
| | | | | | | | | | | | | | | | | | | | | | | | | | | | rs4384231 = GG: 1 {1=1,
2=0}
| | | | | | | | | | | | | | | | | rs17571004 = GG: 1 {1=1, 2=0}
| | | | | | | | | | rs3774796 = AG: 2 {1=0, 2=1}
| | | | | | | | | rs2960482 = AG: 1 {1=7, 2=0}
| | | | | | | | rs585513 = GG
| | | | | | | | | rs2960482 = AA
| | | | | | | | | | rs2103869 = AG: 1 {1=2, 2=0}
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | rs2132528 = AA: 1 {1=4, 2=0}
| | | | | | | | | | | | | | | | | rs2132528 = AG: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | rs6549458 = AC
| | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | | rs17571004 = AA: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | rs17571004 = AG: 2 {1=0, 2=2}
| | | | | | | | | rs2960482 = AG
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = AG
| | | | | | | | | | | | | rs7838995 = AG

| | | | | | | | | | | | | | rs7845891 = AC
| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | rs2132528 = AG
| | | | | | | | | | | | | | | | | | rs17571004 = AA
| | | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | | rs569072 = GG
| | | | | | | | | | | | | | | | | | | | | rs11079162 = AG
| | | | | | | | | | | | | | | | | | | | | | rs10420793 = AA
| | | | | | | | | | | | | | | | | | | | | | | rs6992847 = AG
| | | | | | | | | | | | | | | | | | | | | | | | rs12797573 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | rs9643617 = AG
| | | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = AG: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = GG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | | | | | | rs10420793 = AG
| | | | | | | | | | | | | | | | | | | | | | | | rs6992847 = AG
| | | | | | | | | | | | | | | | | | | | | | | | rs12797573 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | rs9643617 = AG
| | | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = AG: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = GG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | | | | | | | rs12797573 = AG: 1 {1=2, 2=0}
| | | | | | | | | | | | | | | | | | | | | | | | rs17571004 = AG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | | | | | rs6549458 = AC: 1 {1=1, 2=0}
| | | | | | | | | | rs693913 = xx: 2 {1=0, 2=1}
| | | | | | | | | rs463967 = GG
| | | | | | | | | rs693913 = AA
| | | | | | | | | | rs585513 = AG: 1 {1=2, 2=0}
| | | | | | | | | | rs585513 = GG
| | | | | | | | | | | rs2960482 = AA
| | | | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | | | | rs6549458 = AA: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | rs6549458 = AC
| | | | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | | | | rs17571004 = AA
| | | | | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | | | | rs569072 = GG
| | | | | | | | | | | | | | | | | | | | | | | rs11079162 = AG
| | | | | | | | | | | | | | | | | | | | | | | | rs10420793 = AG
| | | | | | | | | | | | | | | | | | | | | | | | | rs6992847 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | | rs12797573 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | | | rs9643617 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = AG: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = GG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | | | | | | | | | rs17571004 = AG: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | | | | | rs2960482 = AG: 1 {1=1, 2=0}
| | | | | | | | | rs693913 = GG
| | | | | | | | | | rs585513 = AA: 1 {1=2, 2=0}
| | | | | | | | | | rs585513 = AG

| | | | | | | | | rs2960482 = AA: 1 {1=3, 2=0}
| | | | | | | | | rs2960482 = AG
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = AG
| | | | | | | | | | | | | rs7838995 = AG
| | | | | | | | | | | | | | rs7845891 = AC
| | | | | | | | | | | | | | | rs6549458 = AA: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | rs6549458 = AC: 1 {1=2, 2=0}
| | | | | | | | | | | | | | | rs6549458 = CC: 1 {1=1, 2=0}
| | | | | | | | rs585513 = GG
| | | | | | | | | rs2960482 = AA
| | | | | | | | | | rs2103869 = AG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | | rs6549458 = AA: 1 {1=2, 2=0}
| | | | | | | | | | | | | | | rs6549458 = AC: 2 {1=0, 2=1}
| | | | | | | | | rs2103869 = GG
| | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | rs17571004 = AA: 1 {1=2, 2=0}
| | | | | | | | | | | | | | | | | rs17571004 = AG: 1 {1=4, 2=0}
| | | | | | | | | | | | | | | | | rs17571004 = GG: 2 {1=0, 2=2}
| | | | | | | | | | | | | | | rs6549458 = AC
| | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | rs17571004 = AA
| | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | rs569072 = GG
| | | | | | | | | | | | | | | | | | | | rs11079162 = AG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | | rs11079162 = GG: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | rs17571004 = AG: 2 {1=0, 2=1}
| | | | | | | | | | rs3774796 = AG: 1 {1=1, 2=0}
| | | | | | | | | rs2960482 = AG
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = AG
| | | | | | | | | | | | | rs7838995 = AG
| | | | | | | | | | | | | | rs7845891 = AC
| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | rs2132528 = AG
| | | | | | | | | | | | | | | | | rs17571004 = AA
| | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | rs569072 = GG
| | | | | | | | | | | | | | | | | | | | rs11079162 = AG

| | | | | | | | | | | | | | | | | | | | | | | | | | rs10420793 = AA: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | | | | | | | | | | | | rs10420793 = AG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | rs6549458 = AC: 2 {1=0, 2=1}
| | | | | | | | | | | rs3774796 = AG: 1 {1=1, 2=0}
| | | | | | | | | rs2960482 = GG: 2 {1=0, 2=1}
| | | rs4793790 = GG
| | | | rs2811415 = GG
| | | | | rs6798749 = GG
| | | | | | rs463967 = AA
| | | | | | | rs693913 = AA: 2 {1=0, 2=1}
| | | | | | | rs693913 = GG
| | | | | | | | rs585513 = AA: 2 {1=0, 2=1}
| | | | | | | | rs585513 = AG: 1 {1=1, 2=0}
| | | | | | | | rs585513 = GG
| | | | | | | | | rs2960482 = AA
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA: 2 {1=0, 2=1}
| | | | | | | | | | | rs3774796 = GG: 1 {1=2, 2=0}
| | | | | | | | | rs2960482 = AG
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = AG
| | | | | | | | | | | | | rs7838995 = AG
| | | | | | | | | | | | | | rs7845891 = AC
| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | rs2132528 = AG
| | | | | | | | | | | | | | | | | | rs17571004 = AG: 1 {1=2, 2=0}
| | | | | | | | | | | | | | | | | | rs17571004 = GG
| | | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | | rs569072 = GG
| | | | | | | | | | | | | | | | | | | | | rs11079162 = AG: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | | | rs11079162 = GG
| | | | | | | | | | | | | | | | | | | | | | rs10420793 = AA: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | | | | rs10420793 = AG: 2 {1=0, 2=1}
| | | | | | rs463967 = AG
| | | | | | | rs693913 = AA
| | | | | | | | rs585513 = AA: 2 {1=0, 2=1}
| | | | | | | | rs585513 = AG
| | | | | | | | | rs2960482 = AA
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | | rs6549458 = AA: 2 {1=0, 2=2}
| | | | | | | | | | | | | | | rs6549458 = AC: 1 {1=2, 2=0}
| | | | | | | | | rs2960482 = AG
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | | rs1496306 = AG
| | | | | | | | | | | | | rs7838995 = AG
| | | | | | | | | | | | | | rs7845891 = AC

| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | | rs2132528 = AG
| | | | | | | | | | | | | | | | | | rs17571004 = AA: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | rs17571004 = GG: 2 {1=0, 2=1}
| | | | | | | | rs585513 = GG
| | | | | | | | | rs2960482 = AA
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | rs17571004 = GG
| | | | | | | | | | | | | | | | | | rs2811388 = GG
| | | | | | | | | | | | | | | | | | | rs569072 = GG
| | | | | | | | | | | | | | | | | | | | rs11079162 = GG
| | | | | | | | | | | | | | | | | | | | | rs10420793 = AA
| | | | | | | | | | | | | | | | | | | | | | rs6992847 = AA
| | | | | | | | | | | | | | | | | | | | | | | rs12797573 = AA
| | | | | | | | | | | | | | | | | | | | | | | | rs9643617 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | rs4921943 = GG
| | | | | | | | | | | | | | | | | | | | | | | | | | rs4384231 = GG
| | | | | | | | | | | | | | | | | | | | | | | | | | | rs4147525 = GG
| | | | | | | | | | | | | | | | | | | | | | | | | | | | rs4680416 = GG
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | rs6779266 = CC
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | rs9462806 = AA
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | rs279878 = AA:
1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | rs279878 = AC:
2 {1=0, 2=1}
| | | | | | | rs693913 = GG
| | | | | | | | rs585513 = AA: 1 {1=1, 2=0}
| | | | | | | | rs585513 = AG
| | | | | | | | | rs2960482 = AA: 2 {1=0, 2=1}
| | | | | | | | | rs2960482 = AG: 1 {1=2, 2=0}
| | | | | | | | | rs2960482 = GG: 1 {1=1, 2=0}
| | | | | | | | rs585513 = GG
| | | | | | | | | rs2960482 = AA
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AA
| | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | rs9857492 = AA
| | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | rs17571004 = AG: 2 {1=0, 2=2}
| | | | | | | | | | | | | | | | | rs17571004 = GG: 1 {1=2, 2=0}
| | | | | | | | | | rs3774796 = AG: 1 {1=1, 2=0}
| | | | | | rs463967 = GG: 1 {1=7, 2=0}

rs11720239 = AG
| rs2999081 = AA: 2 {1=0, 2=2}
| rs2999081 = AG
| | rs2811518 = AA
| | | rs4793790 = AG
| | | | rs2811415 = AG
| | | | | rs6798749 = GG
| | | | | | rs463967 = GG
| | | | | | | rs693913 = AA: 1 {1=4, 2=0}
| | | | | | | rs693913 = GG: 2 {1=0, 2=2}
| | rs2811518 = AG
| | | rs4793790 = AA
| | | | rs2811415 = AA: 1 {1=1, 2=0}
| | | | rs2811415 = AG
| | | | | rs6798749 = AG
| | | | | | rs463967 = AA
| | | | | | | rs693913 = GG
| | | | | | | | rs585513 = AG
| | | | | | | | | rs2960482 = AA
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AG
| | | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | rs9857492 = AG
| | | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | | rs17571004 = AA: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | | | | rs17571004 = AG: 2 {1=0, 2=1}
| | | | | | rs463967 = AG
| | | | | | | rs693913 = AA: 2 {1=0, 2=1}
| | | | | | | rs693913 = GG
| | | | | | | | rs585513 = AG: 2 {1=0, 2=1}
| | | | | | | | rs585513 = GG
| | | | | | | | | rs2960482 = AA: 1 {1=1, 2=0}
| | | | | | | | | rs2960482 = AG: 2 {1=0, 2=1}
| | | | | | rs463967 = GG
| | | | | | | rs693913 = AA: 1 {1=5, 2=0}
| | | | | | | rs693913 = GG
| | | | | | | | rs585513 = GG
| | | | | | | | | rs2960482 = AA
| | | | | | | | | | rs2103869 = GG
| | | | | | | | | | | rs3774796 = AG
| | | | | | | | | | | | rs1496306 = GG
| | | | | | | | | | | | | rs7838995 = AA
| | | | | | | | | | | | | | rs7845891 = AA
| | | | | | | | | | | | | | | rs6549458 = AA
| | | | | | | | | | | | | | | | rs9857492 = AG
| | | | | | | | | | | | | | | | | rs2132528 = AA
| | | | | | | | | | | | | | | | | | rs17571004 = AA: 1 {1=3, 2=0}
| | | | | | | | | | | | | | | | | | rs17571004 = AG: 2 {1=0, 2=1}
| | | | | | | | | | | | | | | rs6549458 = AC: 1 {1=1, 2=0}
| | | | | | | | | | | | | | | rs7845891 = AC: 2 {1=0, 2=1}

```
|  |  |  |  |  rs6798749 = GG: 1 {1=1, 2=0}
|  |  |  |  |  rs6798749 = xx: 2 {1=0, 2=1}
|  |  |  rs4793790 = AG
|  |  |  |  rs2811415 = AG
|  |  |  |  |  rs6798749 = AG
|  |  |  |  |  |  rs463967 = AA: 1 {1=3, 2=0}
|  |  |  |  |  |  rs463967 = AG
|  |  |  |  |  |  |  rs693913 = AA
|  |  |  |  |  |  |  |  rs585513 = AG
|  |  |  |  |  |  |  |  |  rs2960482 = AG
|  |  |  |  |  |  |  |  |  |  rs2103869 = GG
|  |  |  |  |  |  |  |  |  |  |  rs3774796 = AG
|  |  |  |  |  |  |  |  |  |  |  |  rs1496306 = AG: 2 {1=0, 2=1}
|  |  |  |  |  |  |  |  |  |  |  |  rs1496306 = GG: 1 {1=1, 2=0}
|  |  |  |  |  |  |  |  rs585513 = GG: 1 {1=1, 2=0}
|  |  |  |  |  |  |  rs693913 = GG
|  |  |  |  |  |  |  |  rs585513 = AG: 1 {1=1, 2=0}
|  |  |  |  |  |  |  |  rs585513 = GG: 2 {1=0, 2=3}
|  |  |  |  |  |  rs463967 = GG
|  |  |  |  |  |  |  rs693913 = AA: 1 {1=2, 2=0}
|  |  |  |  |  |  |  rs693913 = GG
|  |  |  |  |  |  |  |  rs585513 = AG: 1 {1=2, 2=0}
|  |  |  |  |  |  |  |  rs585513 = GG: 2 {1=0, 2=4}
|  |  |  |  |  rs6798749 = GG: 1 {1=1, 2=0}
|  |  |  rs4793790 = GG
|  |  |  |  rs2811415 = AG
|  |  |  |  |  rs6798749 = AG
|  |  |  |  |  |  rs463967 = AG
|  |  |  |  |  |  |  rs693913 = AA: 2 {1=0, 2=1}
|  |  |  |  |  |  |  rs693913 = GG: 1 {1=2, 2=0}
|  |  |  |  |  |  rs463967 = GG: 1 {1=1, 2=0}
|  rs2999081 = TT: 2 {1=0, 2=7}
|  rs2999081 = CT: 2 {1=0, 2=31}
|  rs2999081 = CC: 2 {1=0, 2=3}
rs11720239 = GG
|  rs2999081 = AA: 1 {1=1, 2=0}
|  rs2999081 = TT: 2 {1=0, 2=8}
|  rs2999081 = CT: 2 {1=0, 2=6}
|  rs2999081 = CC: 2 {1=0, 2=2}
rs11720239 = zzz: 1 {1=1, 2=0}
```

APPENDIX D-) List of SVM-ID3 Hybrid Model Based Associations (Only SNP Model)

rs279878															
rs9462806															
rs6779266															
rs4680416															
rs4147525															
rs4384231															CT
rs4921943					CT				CT						CC
rs9643617					AA				AA						AA
rs12797573					TT				CT						TT
rs6992847					AA				AA						AA
rs10420793					AA	AG			AA						AA
rs11079162					AA	AA			AA						AA
rs569072					CC	CC			CT						CC
rs2811388					GG	GG			GG						GG
rs17571004					CT	CT			CT	CC				TT	CT
rs2132528					AA	AA			AA	AA				AA	AA
rs9857492					AA	AA			AA	AA				AA	AA
rs6549458					TT	TT			TT	TT	GT			TT	TT
rs7845891					AA	AA			AA	AA	AA	AC		AA	AA
rs7838995					AA	AA			AA	AA	AA	AA		AA	AA
rs1496306					GG	GG			GG	GG	GG	GG		GG	GG
rs3774796					TT	TT	CT		TT	TT	TT	TT		TT	TT
rs2103869					GG	GG	GG		GG	GG	GG	GG	CT	GG	GG
rs2960482					TT	TT	TT		TT	TT	TT	TT	CT	TT	TT
rs585513				CT	CC	CC	CC	TT	CT	CT	CT	CT	CT	CC	CC
rs693913				CC	TT	TT	TT	CC	CC	CC	CC	CC	CC	CC	CC
rs463967			AG	GG	AA	AA	AA	AA	AA	AA	AA	AA	AA	CC	AA
rs6798749			GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG
rs2811415			AG	AG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG
rs4793790	TT	CC	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT
rs2811518	CT	CT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT
rs2999081	CT	CT	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC
rs11720239	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA
Branch_id	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

rs279878																
rs9462806																
rs6779266																
rs4680416																
rs4147525																
rs4384231																
rs4921943																
rs9643617																
rs12797573																
rs6992847																
rs10420793			AA									AA				
rs11079162			AA	AG								AA				
rs569072			CC	CC								CC				
rs2811388			GG	GG								GG				
rs17571004			TT	TT	CT						TT	TT	CT	CC		
rs2132528			AG	AG	AG						AA	AA	AA	AA		
rs9857492			AA	AA	AA						AA	AA	AA	AA		
rs6549458	GT		TT	TT	AA	GG					TT	TT	TT	TT	GT	
rs7845891	AA		AC	AC	AC	AC	AA				AA	AA	AA	AA	AA	
rs7838995	AA		AG	AG	AG	AG	AA				AA	AA	AA	AA	AA	
rs1496306	GG		AG	AG	AG	AG	GG				GG	GG	GG	GG	GG	
rs3774796	TT	CT	TT	TT	TT	TT	TT				TT	TT	TT	TT	TT	
rs2103869	GG	GG	GG	GG	GG	GG	GG				AG	GG	GG	GG	GG	
rs2960482	TT	TT	CT	CT	CT	CT	TT	CT	CC		TT	TT	TT	TT	TT	CT
rs585513	CC	CC	CC	CC	CC	CC	CT	CT	CT		CC	CC	CC	CC	CC	CC
rs693913	CC	CC	CC	CC	CC	CC	TT	TT	TT		TT	TT	TT	TT	TT	TT
rs463967	AA	AA	AA	AA	AA	AA	AG	AG	AG		AG	AG	AG	AG	AG	AG
rs6798749	GG	GG	GG	GG	GG	GG	GG	GG	GG		GG	GG	GG	GG	GG	GG
rs2811415	GG	TT	GG	GG	GG	GG	GG	GG	GG		GG	GG	GG	GG	GG	GG
rs4793790	TT	TT	TT	TT	TT	TT	TT	TT	TT		TT	TT	TT	TT	TT	TT
rs2811518	TT	CC	TT	TT	TT	TT	TT	TT	TT		TT	TT	TT	TT	TT	TT
rs2999081	CC	CC	CC	CC	CC	CC	CC	CC	CC		CC	CC	CC	CC	CC	CC
rs11720239	AA	AA	AA	AA	AA	AA	AA	AA	AA		AA	AA	AA	AA	AA	AA
Branch_id	16	17	18	19	20	21	22	23	24		25	26	27	28	29	30

rs279878																
rs9462806													AA			
rs6779266													CC			
rs4680416				CT									CC			
rs4147525				AA									AA			
rs4384231			TT	CC									CC	CC		
rs4921943			CC	CC							CT	CC	CC	CT		
rs9643617			AA	AA							AA	AA	AA	AA		
rs12797573		TT	CT	CT							TT	TT	TT	TT		
rs6992847		AA	AA	AA							AA	AA	AA	AA		
rs10420793		AA	AA	AA							AA	AA	AG	AG		
rs11079162		AA	AA	AA							AA	AA	AA	AA		
rs569072		CT	CT	CT							CC	CC	CC	CC		
rs2811388		GG	GG	GG							GG	GG	GG	GG		
rs17571004	TT	CT	CT	CT	TT						CT	CT	CC	CC		
rs2132528	AA	AA	AA	AA	AA						AA	AA	AA	AA		
rs9857492	AA	AA	AA	AA	AA						AA	AA	AA	AA		
rs6549458	TT	TT	TT	TT	GT						TT	TT	TT	TT	GT	GT
rs7845891	AA	AA	AA	AA	AA	AC					AA	AA	AA	AA	AA	AC
rs7838995	AA	AA	AA	AA	AA	AA					AA	AA	AA	AA	AA	AA
rs1496306	GG	GG	GG	GG	GG	GG					GG	GG	GG	GG	GG	GG
rs3774796	TT	TT	TT	TT	TT	TT	CT				TT	TT	TT	TT	TT	TT
rs2103869	GG	GG	GG	GG	GG	GG	GG				GG	GG	GG	GG	GG	GG
rs2960482	TT	TT	TT	TT	TT	TT	TT	CT	CC		TT	TT	TT	TT	TT	TT
rs585513	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CC	CC	CC	CC	CC	CC
rs693913	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC
rs463967	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG
rs6798749	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG
rs2811415	GG	TT	GG	GG	TT	GG	GG	GG	GG	GG	GG	TT	GG	GG	GG	GG
rs4793790	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT
rs2811518	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT
rs2999081	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC
rs11720239	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA
Branch_id	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	

rs279878																
rs9462806																
rs6779266																
rs4680416																
rs4147525																
rs4384231																
rs4921943																
rs9643617																
rs12797573																
rs6992847																
rs10420793																AA
rs11079162																AA
rs569072												CC				CC
rs2811388												GG				GG
rs17571004							TT					TT	CT			TT
rs2132528							AA					AG	AG	AG		AA
rs9857492							AA					AA	AA	AA		AA
rs6549458			TT				TT	GT	GT			TT	TT	TT		TT
rs7845891			AC	CC			AA	AA	AC		AA	AC	AC	AC		AA
rs7838995	AA		AG	AG			AA	AA	AG		AA	AG	AG	AG		AA
rs1496306	AG		AG	AG			GG	GG	AG		GG	AG	AG	AG		GG
rs3774796	TT		TT	TT			TT	TT	TT		TT	TT	TT	TT	CT	TT
rs2103869	GG		GG	GG			GG	GG	GG		GG	GG	GG	GG	GG	GG
rs2960482	CT		CT	CT	CC		TT	TT	CT	TT	TT	CT	CT	CT	CT	TT
rs585513	CC		CC	CC	CC	CT	CC	CC	CC	TT	CT	CT	CT	CT	CT	CC
rs693913	CC		CC	CC	CC	TT	TT	TT	TT	CC	CC	CC	CC	CC	CC	CC
rs463967	AG		AG	AG	AG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG
rs6798749	GG		GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG
rs2811415	TT		TT	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG
rs4793790	TT		TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT
rs2811518	TT		TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT
rs2999081	CC		CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC
rs11720239	AA		AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA
Branch_id	46		47	48	49	50	51	52	53	54	55	56	57	58	59	60

rs279878																
rs9462806																
rs6779266																
rs4680416																
rs4147525																
rs4384231																
rs4921943	CT	CC														
rs9643617	AA	AA														
rs12797573	TT	TT				CT										
rs6992847	AA	AA				AA										
rs10420793	AG	AG	GG			AA				AA	AG					
rs11079162	AA	AA	AA			AA				AA	AG					
rs569072	CC	CC	CC			CC				CC	CC					
rs2811388	GG	GG	GG			GG				GG	GG					
rs17571004	TT	TT	TT	CT	CC	CT	CT			CT	CT					TT
rs2132528	AA	AA	AA	AA	AA	AA	AA			AG	AG					AG
rs9857492	AA	AA	AA	AA	AA	AA	AA			AA	AA					AA
rs6549458	TT	TT	TT	TT	TT	TT	GT			TT	TT					TT
rs7845891	AA	AA	AA	AA	AA	AA	AA			AC	AC					AC
rs7838995	AA	AA	AA	AA	AA	AA	AA			AG	AG					AG
rs1496306	GG	GG	GG	GG	GG	GG	GG			AG	AG					AG
rs3774796	TT	TT	TT	TT	TT	TT	TT	CT		TT	TT					TT
rs2103869	GG	GG	GG	GG	GG	GG	GG	GG		GG	GG					GG
rs2960482	TT	TT	TT	TT	TT	TT	TT	TT		CT	CT	CC				CT
rs585513	CC	CC	CC	CC	CC	CC	CC	CC		CC	CC	CC			CC	TT
rs693913	CC	CC	CC	CC	CC	CC	CC	CC		CC	CC	CC				TT
rs463967	GG	GG	GG	GG	GG	GG	GG	GG		GG	GG	GG	AA	AG	CC	AA
rs6798749	GG	GG	GG	GG	GG	GG	GG	GG		GG	GG	GG	GG	GG	GG	GG
rs2811415	TT	GG	TT	GG	GG	GG	GG	GG		GG	GG	GG	AG	AG	AG	GG
rs4793790	TT	TT	TT	TT	TT	TT	TT	TT		TT	TT	TT	CT	CT	CT	CT
rs2811518	TT	TT	TT	TT	TT	TT	TT	TT		TT	TT	TT	TT	TT	TT	TT
rs2999081	CC	CC	CC	CC	CC	CC	CC	CC		CC	CC	CC	CC	CC	CC	CC
rs11720239	AA	AA	AA	AA	AA	AA	AA	AA		AA	AA	AA	AA	AA	AA	AA
Branch_id	61	62	63	64	65	66	67	68		69	70	71	72	73	74	75

rs279878																
rs9462806																
rs6779266																
rs4680416																
rs4147525																
rs4384231												CT				
rs4921943						CT						CC				
rs9643617						AA						AA				
rs12797573						CT						TT				
rs6992847						AA						AA				
rs10420793				AG		AA				AG		AA				
rs11079162				AG		AG				AG		AG				
rs569072				TT		CT				CC		CC				
rs2811388				GG		GG				GG		GG				
rs17571004	TT			CT	TT	CT	CC			TT	CT	CC				
rs2132528	AG			AA	AA	AA	AA			AA	AA	AA				
rs9857492	AA			AA	AA	AA	AA			AA	AA	AA				
rs6549458	TT	GT		TT	TT	TT	TT			TT	TT	TT				
rs7845891	AC	AC		AA	AA	AA	AA			AA	AA	AA	AC			
rs7838995	AG	AG		AA	AA	AA	AA			AA	AA	AA	AA			
rs1496306	AG	AG		GG	GG	GG	GG			GG	GG	GG	GG	AG		
rs3774796	TT	TT		TT	TT	TT	TT	CT		TT	TT	TT	TT	TT		
rs2103869	GG	GG		GG	GG	GG	GG	GG		GG	GG	GG	GG	GG		AG
rs2960482	CT	CT	TT	TT	TT	TT	TT	TT		TT	TT	TT	TT	CT		TT
rs585513	CT	CT	CC	TT	CT	CT	CT	CT	CT	CC	CC	CC	CC	CC	TT	CT
rs693913	TT	TT	TT	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	TT	TT
rs463967	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AG	AG
rs6798749	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG
rs2811415	CT	GG	CT	GG	CT	GG	CT	GG	CT	GG	CT	GG	CT	GG	CT	GG
rs4793790	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT
rs2811518	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT
rs2999081	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC
rs11720239	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA
Branch_id	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	

rs279878																
rs9462806																
rs6779266																
rs4680416																
rs4147525						AG										
rs4384231						CT	CC									
rs4921943						CC	CC							CC	CC	
rs9643617						AA	AA							AG	AG	
rs12797573						CT	CT							TT	TT	CT
rs6992847						AA	AA							AG	AG	AG
rs10420793						AA	AA							AA	AG	AG
rs11079162						AG	AG							AG	AG	AG
rs569072						CT	CT							CC	CC	CC
rs2811388						GG	GG							GG	GG	GG
rs17571004	CT				TT	CT	CT		CC				TT	TT	TT	TT
rs2132528	AA				AA	AA	AA		AA			AA	AA	AG	AG	AG
rs9857492	AA				AA	AA	AA		AA			AA	AA	AA	AA	AA
rs6549458	TT				TT	TT	TT		TT			TT	GT	TT	TT	TT
rs7845891	AA				AA	AA	AA		AA			AA	AA	AC	AC	AC
rs7838995	AA				AA	AA	AA		AA			AA	AA	AG	AG	AG
rs1496306	GG				GG	GG	GG		GG			GG	GG	AG	AG	AG
rs3774796	TT				TT	TT	TT		TT			TT	TT	TT	TT	TT
rs2103869	GG				GG	GG	GG		GG		AG	GG	GG	GG	GG	GG
rs2960482	TT	CT			TT	TT	TT	CT	TT		TT	TT	TT	CT	CT	CT
rs585513	CT	CT	CC	TT	CT	CT	CT	CT	CT	CC	CC	CC	CC	CC	CC	CC
rs693913	TT	TT	TT	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC
rs463967	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG
rs6798749	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG
rs2811415	CT	GG	CT	GG	CT	GG	CT	GG	CT	GG	CT	GG	CT	GG	GG	GG
rs4793790	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT
rs2811518	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT
rs2999081	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC
rs11720239	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA
Branch_id	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	

rs279878																
rs9462806																
rs6779266																
rs4680416																
rs4147525																
rs4384231																
rs4921943				CC												
rs9643617				AA												
rs12797573				TT												
rs6992847				AA												
rs10420793				AG												AG
rs11079162				AG										AG		AG
rs569072				CC										CC		CC
rs2811388				GG										GG		GG
rs17571004	CT			TT								TT	CT	TT		TT
rs2132528	AG			AA								AA	AA	AA		AG
rs9857492	AA			AA								AA	AA	AA		AA
rs6549458	TT	GT		GT					GT	GG	TT	TT	TT	GT		TT
rs7845891	AC	AC		AA					AC	AC	AA	AA	AA	AA		AC
rs7838995	AG	AG		AA					AG	AG	AA	AA	AA	AA		AG
rs1496306	AG	AG		GG					AG	AG	GG	GG	GG	GG		AG
rs3774796	TT	TT		TT					TT	TT	TT	GG	TT	TT	CT	TT
rs2103869	GG	GG		GG					GG	GG	AG	GG	GG	GG	GG	GG
rs2960482	CT	CT		TT	CT			TT	CT	CT	TT	TT	TT	TT	TT	CT
rs585513	CC	CC	CT	CC	CC	TT		CT	CT	CT	CC	CC	CC	CC	CC	CC
rs693913	CC	CC	TT	TT	TT	CC		CC	CC	CC	CC	CC	CC	CC	CC	CC
rs463967	AG	AG	GG	GG	GG	GG		GG	GG	GG	GG	GG	GG	GG	GG	GG
rs6798749	GG	GG	GG	GG	GG	GG		GG	GG	GG	GG	GG	GG	GG	GG	GG
rs2811415	GG	CT	GG	GG	CT	GG		GG	GG	GG	GG	GG	GG	GG	GG	GG
rs4793790	CT	CT	CT	CT	CT	CT		CT	CT	CT	CT	CT	CT	CT	CT	CT
rs2811518	TT	TT	TT	TT	TT	TT		TT	TT	TT	TT	TT	TT	TT	TT	TT
rs2999081	CC	CC	CC	CC	CC	CC		CC	CC	CC	CC	CC	CC	CC	CC	CC
rs11720239	AA	AA	AA	AA	AA	AA		AA	AA	AA	AA	AA	AA	AA	AA	AA
Branch_id	106	107	108	109	110	111		112	113	114	115	116	117	118	119	120

rs279878									TT							
rs9462806									AA							
rs6779266									CC							
rs4680416									CC							
rs4147525									GG							
rs4384231									CC							
rs4921943									CC							
rs9643617									AA							
rs12797573									TT							
rs6992847									AA							
rs10420793						AA			AA							
rs11079162					AG	GG			GG							
rs569072					CC	CC			CC							
rs2811388					GG	GG			GG							
rs17571004				CT	CC	CC		TT	CC					CC		
rs2132528				AG	AG	AG		AG	AA					AA		
rs9857492				AA	AA	AA		AA	AA					AA		
rs6549458				TT	TT	TT	GT	TT	AA					TT		
rs7845891				AC	AC	AC	AA	AC	AA					AA		
rs7838995				AG	AG	AG	AA	AG	AA					AA		
rs1496306				AG	AG	AG	GG	AG	GG					GG		
rs3774796	CT		CC	TT	TT	TT	TT	TT	TT					TT	CT	
rs2103869	GG		GG	GG	GG	GG	GG	GG	GG					GG	GG	
rs2960482	CT		TT	CT	CT	CT	TT	CT	TT		CT	CC		TT		
rs585513	CC	CT	CC	CC	CC	CC	CT	CC	CC	TT	CT	CT	CC	CC		
rs693913	CC	CC	CC	AA	CC	CC	TT	CC	TT	CC	CC	CC	CC	CC		
rs463967	GG	GG	AA	CC	AA	AA	AG	AA	AG	AG	AG	AG	AG	AG	GG	
rs6798749	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	
rs2811415	GG	CT	CC	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	
rs4793790	CT	CT	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	
rs2811518	TT	TT	TT	CC	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	
rs2999081	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	
rs11720239	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	
Branch_id	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	

rs279878																
rs9462806																
rs6779266																
rs4680416																
rs4147525																
rs4384231																
rs4921943																
rs9643617																
rs12797573																
rs6992847																
rs10420793																
rs11079162																
rs569072																
rs2811388																
rs17571004			TT			TT										
rs2132528			AA			AA										
rs9857492			AG			AG										
rs6549458			TT			TT	GT									
rs7845891			AA			AA	AA									
rs7838995			AA			AA	AA									
rs1496306			GG			GG	GG				GG					
rs3774796			CT			CT	CT				CT					
rs2103869			GG			GG	GG				GG					
rs2960482			TT	TT		TT	TT				CT					
rs585513			CT	CC		CC	CC				CT	CC	CT		CT	
rs693913	TT		CC	CC	TT	CC	CC				TT	TT	CC	TT	CC	
rs463967	GG		AA	AG	GG	GG	GG			AA	AG	AG	AG	GG	GG	
rs6798749	GG		AG	AG	AG	AG	AG	GG	AG	AG	AG	AG	AG	AG	AG	GG
rs2811415	AG	AA	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG
rs4793790	CT	TT	TT	TT	TT	TT	TT	TT	TT	CT	CT	CT	CT	CT	CT	CT
rs2811518	TT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT
rs2999081	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT
rs11720239	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG	AG
Branch_id	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	

rs279878				
rs9462806				
rs6779266				
rs4680416				
rs4147525				
rs4384231				
rs4921943				
rs9643617				
rs12797573				
rs6992847				
rs10420793				
rs11079162				
rs569072				
rs2811388				
rs17571004				
rs2132528				
rs9857492				
rs6549458				
rs7845891				
rs7838995				
rs1496306				
rs3774796				
rs2103869				
rs2960482				
rs585513				
rs693913	CC			
rs463967	AG	GG		
rs6798749	AG	AG		
rs2811415	AG	AG		
rs4793790	CC	CC		
rs2811518	CT	CT		
rs2999081	CT	CT	CC	TT
rs11720239	AG	AG	AG	GG
Branch_id	151	152	153	154

Descriptions:

Branch_id: Identifier of hybrid model derived associations. Every single association has a different branch_id.

APPENDIX E-) Decision Tree Structure of Second Hybrid Model Based Associations (SNP-Environmental Combined, for African-Americans)

ethnicity = African American
| BMI≤22.5
| | rs11729739 = TT
| 22.5<BMI≤24.9
| | rs17701543 = AG
| | rs17701543 = GG
| | | rs9848588 = AG
| | | rs9848588 = GG
| | | | rs964130 = AA
| | | | | rs10195113 = AG
| | | | | rs10195113 = GG
| | | | | | rs1433369 = CT
| | | | | | rs1433369 = CC
| | | | | | | rs12733054 = CT
| | | | | | | rs12733054 = CC
| | | | | | | | rs17375010 = CT
| | | | | | | | rs17375010 = CC
| | | | | | | | | rs766045 = AG
| | | | | | | | | | Alcohol = NONE
| | | | rs964130 = AG
| 24.9<BMI≤29.9
| | smoking= none
| | | rs12201462 = CT
| | | rs12201462 = CC
| | | | rs4908656 = AA
| | | | | rs9462806 = AA
| | | | | | rs1974562 = AG
| | | | | rs9462806 = AG
| | | | rs4908656 = AC

| | smoking<5 years
| | | rs10954845 = AA
| | | | rs6997228 = AA
| | 5 years≤smoking<10 years
| | | rs10745253 = AG
| | 10 years≤smoking<20 years
| | | rs12980509 = CT
| | | rs12980509 = CC
| | | | rs2296370 = AA
| | | | rs2296370 = AG
| | 20 years≤smoking<30 years
| | | rs7843255 = AA
| | | rs7843255 = GG
| | 30 years≤smoking
| | | rs2194505 = CT
| 24.9<BMI
| | rs10517581 = AA
| | | rs2103869 = GG
| | | | rs10788555 = GG
| | | | | rs7067548 = AA
| | | | | | rs17001078 = TT
| | | | | | | rs918285 = TT
| | rs10517581 = AG
| | | rs9462806 = AG

APPENDIX F-) List of Second Hybrid Model Based Associations (SNP-Environmental Combined)

Branch id	Association parameters											
1	BMI≤22..5	rs11729739-TT										
2	22.5<BMI≤ 24.9	rs17701543-AG										
3	22.5<BMI≤ 24.9	rs17701543-GG	rs9848588-AG									
4	22.5<BMI≤ 24.9	rs17701543-GG	rs9848588-GG	rs964130-AA	rs10195113-AG							
5	22.5<BMI≤ 24.9	rs17701543-GG	rs9848588-GG	rs964130-AA	rs10195113-GG	rs1433369-CT						
6	22.5<BMI≤ 24.9	rs17701543-GG	rs9848588-GG	rs964130-AA	rs10195113-GG	rs1433369-CC	rs12733054-CT					
7	22.5<BMI≤ 24.9	rs17701543-GG	rs9848588-GG	rs964130-AA	rs10195113-GG	rs1433369-CC	rs12733054-CC	rs17375010-CT				
8	22.5<BMI≤ 24.9	rs17701543-GG	rs9848588-GG	rs964130-AA	rs10195113-GG	rs1433369-CC	rs12733054-CC	rs17375010-CC	rs766045-AG	Alcohol= none		
9	22.5<BMI≤ 24.9	rs17701543-GG	rs9848588-GG	rs964130-AG								
10	24.9<BMI≤ 29.9	smoking= none	rs12201462-CT									
11	24.9<BMI≤ 29.9	smoking= none	rs12201462-CC	rs4908656-AA	rs9462806-AA	rs1974562-AG						

Branch_id	Association parameters						
12	24.9<BMI≤29.9	smoking= none	rs12201462-CC	rs4908656-AA	rs9462806-AG		
13	24.9<BMI≤29.9	smoking= none	rs12201462-CC	rs4908656-AC			
14	24.9<BMI≤29.9	smoking <5 years	rs10954845-AA	rs6997228-AA			
15	24.9<BMI≤29.9	5 years ≤ smoking<10 years	rs10745253-AG				
16	24.9<BMI≤29.9	10 years ≤ smoking<20 years	rs12980509-CT				
17	24.9<BMI≤29.9	10 years ≤ smoking<20 years	rs12980509-CC	rs2296370-AA			
18	24.9<BMI≤29.9	10 years ≤ smoking<20 years	rs12980509-CC	rs2296370-AG			
19	24.9<BMI≤29.9	20 years ≤ smoking<30 years	rs7843255-AA				
20	24.9<BMI≤29.9	20 years ≤ smoking<30 years	rs7843255-GG				
21	24.9<BMI≤29.9	30 years ≤ smoking	rs2194505-CT				
22	24.9<BMI	rs10517581-AA	rs2103869-GG	rs10788555-GG	rs7067548-AA	rs17001078-TT	rs918285-TT
23	24.9<BMI	rs10517581-AG	rs9462806-AG				

Description:

Branch_id: Identifier of hybrid model derived associations. Every single association has a different branch_id.

APPENDIX G-) Complete results of test and evaluation processes

Evaluation	Approaches	Methods	01-hu1213DA
Genomic Evaluation	**Independent associations**	**Homozygote**	13
		Heterozygote	38
		Not-Analyzed	13
	Polygenic risk scores	**Number of SNP (DM)**	51
		Number of SNP (AM)	64
		Evidence-Impact-SNP (DM)	159
		Evidence-Impact-SNP (AM)	198
	Cumulative models	**17-SNP_Helfand**	8 (1)
		9-SNP_Helfand	2 (1)
		5-SNP_Zheng	1
		5-SNP_Salinas	1
		4-SNP_Nam	2
		3-SNP_Beuten	2 (1)
	Hybrid model	**Yücebaş-Aydın Son**	Not exact
EG Evaluation	**Hybrid model**	**Yücebaş-Aydın Son**	Not exact
SD Evaluation	**Family history**		Not available
	Age	birth year	1937 (76)
		PCa diagnosis year	2010 (73)
Clinical Evaluation	**Anatomic findings**	weight (kg)	86
		height (cm)	190
		BMI (kg/cm2)	23,82
		BMI (category)	Normal (BB)
	Medical conditions		Hypercholesterolemia
			BPH
Lifestyle (EB) Evaluation	**Drugs**		
	Surgical procedures		
	Supplements		
	Nutrition		
	Physical activity		
	Alcohol		

Evaluation	Approaches	Methods	03-huD889CC
Genomic Evaluation	**Independent associations**	**Homozygote**	13
		Heterozygote	32
		Not-Analyzed	20
	Polygenic risk scores	**Number of SNP (DM)**	45
		Number of SNP (AM)	58
		Evidence-Impact-SNP (DM)	137
		Evidence-Impact-SNP (AM)	177
	Cumulative models	**17-SNP_Helfand**	4 (5)
		9-SNP_Helfand	1
		5-SNP_Zheng	0
		5-SNP_Salinas	0
		4-SNP_Nam	1
		3-SNP_Beuten	1 (1)
	Hybrid model	**Yücebaş-Aydın Son**	None
EG Evaluation	**Hybrid model**	**Yücebaş-Aydın Son**	None
SD Evaluation	**Family history**		Not available
	Age	birth year	1938 (75)
		PCa diagnosis year	Not available
Clinical Evaluation	**Anatomic findings**	weight (kg)	64
		height (cm)	177
		BMI (kg/cm2)	20,43
		BMI (category)	Normal (AA)
	Medical conditions		Syphilis
Lifestyle (EB) Evaluation	**Drugs**		
	Surgical procedures		
	Supplements		
	Nutrition		
	Physical activity		
	Alcohol		

Evaluation	Approaches	Methods	07-hu28F39C
Genomic Evaluation	**Independent associations**	**Homozygote**	21
		Heterozygote	28
		Not-Analyzed	9
	Polygenic risk scores	**Number of SNP (DM)**	49
		Number of SNP (AM)	70
		Evidence-Impact-SNP (DM)	152
		Evidence-Impact-SNP (AM)	220
	Cumulative models	**17-SNP_Helfand**	9
		9-SNP_Helfand	3
		5-SNP_Zheng	1
		5-SNP_Salinas	1
		4-SNP_Nam	1
		3-SNP_Beuten	1 (1)
	Hybrid model	**Yücebaş-Aydın Son**	None
EG Evaluation	**Hybrid model**	**Yücebaş-Aydın Son**	Not exact
SD Evaluation	**Family history**		Not available
	Age	birth year	1943 (70)
		PCa diagnosis year	Not available
Clinical Evaluation	**Anatomic findings**	weight (kg)	70
		height (cm)	170
		BMI (kg/cm2)	24,22
		BMI (category)	Normal (BB)
	Medical conditions		Hypercholesterolemia
			BPH
Lifestyle (EB) Evaluation	**Drugs**		Lipitor
	Surgical procedures		
	Supplements		
	Nutrition		
	Physical activity		
	Alcohol		

Evaluation	Approaches	Methods	13-hu6ED94A
Genomic Evaluation	**Independent associations**	**Homozygote**	16
		Heterozygote	36
		Not-Analyzed	20
	Polygenic risk scores	**Number of SNP (DM)**	52
		Number of SNP (AM)	68
		Evidence-Impact-SNP (DM)	161
		Evidence-Impact-SNP (AM)	210
	Cumulative models	**17-SNP_Helfand**	8 (5)
		9-SNP_Helfand	4
		5-SNP_Zheng	1
		5-SNP_Salinas	1
		4-SNP_Nam	1
		3-SNP_Beuten	2 (1)
	Hybrid model	**Yücebaş-Aydın Son**	None
EG Evaluation	**Hybrid model**	**Yücebaş-Aydın Son**	Yes
SD Evaluation	**Family history**		Not available
	Age	birth year	1950 (63)
		PCa diagnosis year	2011 (61)
Clinical Evaluation	**Anatomic findings**	weight (kg)	58
		height (cm)	170
		BMI (kg/cm2)	20,07
		BMI (category)	Normal (AA)
	Medical conditions		Hypercholesterolemia
Lifestyle (EB) Evaluation	**Drugs**		Simvastatin
	Surgical procedures		
	Supplements		
	Nutrition		
	Physical activity		
	Alcohol		

Evaluation	Approaches	Methods	02-hu59141C
Genomic Evaluation	**Independent associations**	**Homozygote**	16
		Heterozygote	32
		Not-Analyzed	8
	Polygenic risk scores	**Number of SNP (DM)**	48
		Number of SNP (AM)	64
		Evidence-Impact-SNP (DM)	150
		Evidence-Impact-SNP (AM)	202
	Cumulative models	**17-SNP_Helfand**	9
		9-SNP_Helfand	3
		5-SNP_Zheng	2
		5-SNP_Salinas	2
		4-SNP_Nam	2
		3-SNP_Beuten	1 (1)
	Hybrid model	**Yücebaş-Aydın Son**	None
EG Evaluation	**Hybrid model**	**Yücebaş-Aydın Son**	None
SD Evaluation	**Family history**		Not available
	Age	birth year	1937 (76)
		PCa diagnosis year	
Clinical Evaluation	**Anatomic findings**	weight (kg)	108
		height (cm)	177
		BMI (kg/cm2)	34,47
		BMI (category)	Obese (DD)
	Medical conditions		Asbestosis
			T2DM
Lifestyle (EB) Evaluation	**Drugs**		
	Surgical procedures		
	Supplements		Multivitamin
	Nutrition		Vegetable servings (5 servings)
	Physical activity		regular physical activity ?
	Alcohol		

Evaluation	Approaches	Methods	04-huF7E042
Genomic Evaluation	**Independent associations**	**Homozygote**	20
		Heterozygote	26
		Not-Analyzed	17
	Polygenic risk scores	**Number of SNP (DM)**	46
		Number of SNP (AM)	66
		Evidence-Impact-SNP (DM)	145
		Evidence-Impact-SNP (AM)	210
	Cumulative models	**17-SNP_Helfand**	6 (4)
		9-SNP_Helfand	2
		5-SNP_Zheng	1
		5-SNP_Salinas	1
		4-SNP_Nam	2
		3-SNP_Beuten	0 (1)
	Hybrid model	**Yücebaş-Aydın Son**	None
EG Evaluation	**Hybrid model**	**Yücebaş-Aydın Son**	Yes
SD Evaluation	**Family history**		Not available
	Age	birth year	1939 (74)
		PCa diagnosis year	
Clinical Evaluation	**Anatomic findings**	weight (kg)	69
		height (cm)	177
		BMI (kg/cm2)	22,02
		BMI (category)	Normal (AA)
	Medical conditions		BPH
Lifestyle (EB) Evaluation	**Drugs**		
	Surgical procedures		TURP
	Supplements		
	Nutrition		
	Physical activity		
	Alcohol		

Evaluation	Approaches	Methods	05-hu75BE2C
Genomic Evaluation	**Independent associations**	**Homozygote**	9
		Heterozygote	36
		Not-Analyzed	9
	Polygenic risk scores	**Number of SNP (DM)**	45
		Number of SNP (AM)	54
		Evidence-Impact-SNP (DM)	143
		Evidence-Impact-SNP (AM)	172
	Cumulative models	**17-SNP_Helfand**	8
		9-SNP_Helfand	3
		5-SNP_Zheng	1
		5-SNP_Salinas	1
		4-SNP_Nam	1
		3-SNP_Beuten	1 (1)
	Hybrid model	**Yücebaş-Aydın Son**	Yes
EG Evaluation	**Hybrid model**	**Yücebaş-Aydın Son**	Not exact
SD Evaluation	**Family history**		Not available
	Age	birth year	1939 (74)
		PCa diagnosis year	
Clinical Evaluation	**Anatomic findings**	weight (kg)	75
		height (cm)	180
		BMI (kg/cm2)	23,15
		BMI (category)	Normal (BB)
	Medical conditions		
Lifestyle (EB) Evaluation	**Drugs**		
	Surgical procedures		
	Supplements		
	Nutrition		
	Physical activity		regular physical activity
	Alcohol		

Evaluation	Approaches	Methods	06-hu56B3B6
Genomic Evaluation	**Independent associations**	**Homozygote**	14
		Heterozygote	40
		Not-Analyzed	8
	Polygenic risk scores	**Number of SNP (DM)**	54
		Number of SNP (AM)	68
		Evidence-Impact-SNP (DM)	164
		Evidence-Impact-SNP (AM)	210
	Cumulative models	**17-SNP_Helfand**	7
		9-SNP_Helfand	3
		5-SNP_Zheng	2
		5-SNP_Salinas	2
		4-SNP_Nam	0
		3-SNP_Beuten	0 (1)
	Hybrid model	**Yücebaş-Aydın Son**	None
EG Evaluation	**Hybrid model**	**Yücebaş-Aydın Son**	None
SD Evaluation	**Family history**		Not available
	Age	birth year	1941 (72)
		PCa diagnosis year	
Clinical Evaluation	**Anatomic findings**	weight (kg)	99
		height (cm)	177
		BMI (kg/cm2)	31,6
		BMI (category)	Obese (DD)
	Medical conditions		Hypercholesterolemia
			Chlamydia Infection
			Basal cell skin cancer
Lifestyle (EB) Evaluation	**Drugs**		Ibuprofen
	Surgical procedures		
	Supplements		Multivitamin, Folic Acid, Vitamin E, Selenium, Lycopene, Pomegranate
	Nutrition		
	Physical activity		
	Alcohol		Alcoholism

Evaluation	Approaches	Methods	08-huB59C05
Genomic Evaluation	**Independent associations**	**Homozygote**	21
		Heterozygote	32
		Not-Analyzed	8
	Polygenic risk scores	**Number of SNP (DM)**	53
		Number of SNP (AM)	74
		Evidence-Impact-SNP (DM)	166
		Evidence-Impact-SNP (AM)	236
	Cumulative models	**17-SNP_Helfand**	8
		9-SNP_Helfand	3
		5-SNP_Zheng	1
		5-SNP_Salinas	1
		4-SNP_Nam	2
		3-SNP_Beuten	2 (1)
	Hybrid model	**Yücebaş-Aydın Son**	None
EG Evaluation	**Hybrid model**	**Yücebaş-Aydın Son**	None
SD Evaluation	**Family history**		Not available
	Age	birth year	1943 (70)
		PCa diagnosis year	
Clinical Evaluation	**Anatomic findings**	weight (kg)	99
		height (cm)	180
		BMI (kg/cm2)	30,56
		BMI (category)	Obese (DD)
	Medical conditions		
Lifestyle (EB) Evaluation	**Drugs**		
	Surgical procedures		
	Supplements		
	Nutrition		
	Physical activity		
	Alcohol		

Evaluation	Approaches	Methods	10-hu7A2F1D
Genomic Evaluation	**Independent associations**	**Homozygote**	12
		Heterozygote	33
		Not-Analyzed	24
	Polygenic risk scores	**Number of SNP (DM)**	45
		Number of SNP (AM)	57
		Evidence-Impact-SNP (DM)	145
		Evidence-Impact-SNP (AM)	190
	Cumulative models	**17-SNP_Helfand**	6 (4)
		9-SNP_Helfand	3 (1)
		5-SNP_Zheng	1 (1)
		5-SNP_Salinas	1 (1)
		4-SNP_Nam	1
		3-SNP_Beuten	1 (1)
	Hybrid model	**Yücebaş-Aydın Son**	None
EG Evaluation	**Hybrid model**	**Yücebaş-Aydın Son**	None
SD Evaluation	**Family history**		Not available
	Age	birth year	1947 (66)
		PCa diagnosis year	
Clinical Evaluation	**Anatomic findings**	weight (kg)	75
		height (cm)	177
		BMI (kg/cm2)	23,94
		BMI (category)	Normal (BB)
	Medical conditions		Hypercholesterolemia
			Non-melanoma skin cancer
Lifestyle (EB) Evaluation	**Drugs**		Atorvastatin
	Surgical procedures		
	Supplements		
	Nutrition		
	Physical activity		regular physical activity
	Alcohol		

Evaluation	Approaches	Methods	12-huD57BBF
Genomic Evaluation	**Independent associations**	**Homozygote**	9
		Heterozygote	39
		Not-Analyzed	8
	Polygenic risk scores	**Number of SNP (DM)**	48
		Number of SNP (AM)	57
		Evidence-Impact-SNP (DM)	147
		Evidence-Impact-SNP (AM)	177
	Cumulative models	**17-SNP_Helfand**	9
		9-SNP_Helfand	3
		5-SNP_Zheng	1
		5-SNP_Salinas	1
		4-SNP_Nam	1
		3-SNP_Beuten	1 (1)
	Hybrid model	**Yücebaş-Aydın Son**	None
EG Evaluation	**Hybrid model**	**Yücebaş-Aydın Son**	None
SD Evaluation	**Family history**		Not available
	Age	birth year	1949 (64)
		PCa diagnosis year	
Clinical Evaluation	**Anatomic findings**	weight (kg)	75
		height (cm)	177
		BMI (kg/cm2)	23,94
		BMI (category)	Normal (BB)
	Medical conditions		Hypercholesterolemia
			BPH
Lifestyle (EB) Evaluation	**Drugs**		Simvastatin, Aspirin
	Surgical procedures		Vasectomy
	Supplements		
	Nutrition		
	Physical activity		regular physical activity
	Alcohol		

Evaluation	Approaches	Methods	14-huD7960A
Genomic Evaluation	**Independent associations**	**Homozygote**	15
		Heterozygote	31
		Not-Analyzed	9
	Polygenic risk scores	**Number of SNP (DM)**	46
		Number of SNP (AM)	61
		Evidence-Impact-SNP (DM)	142
		Evidence-Impact-SNP (AM)	189
	Cumulative models	**17-SNP_Helfand**	6
		9-SNP_Helfand	3
		5-SNP_Zheng	1
		5-SNP_Salinas	1
		4-SNP_Nam	1
		3-SNP_Beuten	1 (1)
	Hybrid model	**Yücebaş-Aydın Son**	None
EG Evaluation	**Hybrid model**	**Yücebaş-Aydın Son**	Not exact
SD Evaluation	**Family history**		Not available
	Age	birth year	1951 (62)
		PCa diagnosis year	
Clinical Evaluation	**Anatomic findings**	weight (kg)	75
		height (cm)	172
		BMI (kg/cm2)	25,35
		BMI (category)	Overweight (CC)
	Medical conditions		Hypercholesterolemia
			BPH
			T2DM
Lifestyle (EB) Evaluation	**Drugs**		
	Surgical procedures		
	Supplements		
	Nutrition		
	Physical activity		
	Alcohol		

Evaluation	Approaches	Methods	15-hu2E413D
Genomic Evaluation	**Independent associations**	**Homozygote**	19
		Heterozygote	34
		Not-Analyzed	8
	Polygenic risk scores	**Number of SNP (DM)**	53
		Number of SNP (AM)	72
		Evidence-Impact-SNP (DM)	161
		Evidence-Impact-SNP (AM)	219
	Cumulative models	**17-SNP_Helfand**	6
		9-SNP_Helfand	1
		5-SNP_Zheng	0
		5-SNP_Salinas	0
		4-SNP_Nam	0
		3-SNP_Beuten	1 (1)
	Hybrid model	**Yücebaş-Aydın Son**	None
EG Evaluation	**Hybrid model**	**Yücebaş-Aydın Son**	Not exact
SD Evaluation	**Family history**		Not available
	Age	birth year	1952 (61)
		PCa diagnosis year	
Clinical Evaluation	**Anatomic findings**	weight (kg)	82
		height (cm)	180
		BMI (kg/cm2)	25,31
		BMI (category)	Overweight (CC)
	Medical conditions		
Lifestyle (EB) Evaluation	**Drugs**		
	Surgical procedures		
	Supplements		
	Nutrition		
	Physical activity		
	Alcohol		

Evaluation	Approaches	Methods	16-hu76CAA5
Genomic Evaluation	Independent associations	Homozygote	12
		Heterozygote	33
		Not-Analyzed	9
	Polygenic risk scores	Number of SNP (DM)	45
		Number of SNP (AM)	57
		Evidence-Impact-SNP (DM)	143
		Evidence-Impact-SNP (AM)	182
	Cumulative models	17-SNP_Helfand	8
		9-SNP_Helfand	2
		5-SNP_Zheng	1
		5-SNP_Salinas	1
		4-SNP_Nam	1
		3-SNP_Beuten	1 (1)
	Hybrid model	Yücebaş-Aydın Son	None
EG Evaluation	Hybrid model	Yücebaş-Aydın Son	Not exact
SD Evaluation	Family history		Not available
	Age	birth year	1952 (61)
		PCa diagnosis year	
Clinical Evaluation	Anatomic findings	weight (kg)	93
		height (cm)	177
		BMI (kg/cm2)	29,68
		BMI (category)	Overweight (CC)
	Medical conditions		
Lifestyle (EB) Evaluation	Drugs		Aspirin
	Surgical procedures		
	Supplements		Omega-3 Fish Oil
	Nutrition		
	Physical activity		
	Alcohol		

Evaluation	Approaches	Methods	17-huA720D3
Genomic Evaluation	**Independent associations**	**Homozygote**	18
		Heterozygote	36
		Not-Analyzed	9
	Polygenic risk scores	**Number of SNP (DM)**	54
		Number of SNP (AM)	72
		Evidence-Impact-SNP (DM)	166
		Evidence-Impact-SNP (AM)	222
	Cumulative models	**17-SNP_Helfand**	8
		9-SNP_Helfand	4
		5-SNP_Zheng	2
		5-SNP_Salinas	2
		4-SNP_Nam	2
		3-SNP_Beuten	1 (1)
	Hybrid model	**Yücebaş-Aydın Son**	None
EG Evaluation	**Hybrid model**	**Yücebaş-Aydın Son**	Not exact
SD Evaluation	**Family history**		Not available
	Age	birth year	1953 (60)
		PCa diagnosis year	
Clinical Evaluation	**Anatomic findings**	weight (kg)	79
		height (cm)	180
		BMI (kg/cm2)	24,38
		BMI (category)	Normal (BB)
	Medical conditions		Hypercholesterolemia
Lifestyle (EB) Evaluation	**Drugs**		Aspirin
	Surgical procedures		
	Supplements		Multivitamin, Phytosterols, Omega-3 Fish Oil, Melatonin
	Nutrition		
	Physical activity		
	Alcohol		

Evaluation	Approaches	Methods	18-hu63DA55
Genomic Evaluation	**Independent associations**	**Homozygote**	13
		Heterozygote	28
		Not-Analyzed	8
	Polygenic risk scores	**Number of SNP (DM)**	41
		Number of SNP (AM)	54
		Evidence-Impact-SNP (DM)	132
		Evidence-Impact-SNP (AM)	179
	Cumulative models	**17-SNP_Helfand**	6
		9-SNP_Helfand	2
		5-SNP_Zheng	1
		5-SNP_Salinas	1
		4-SNP_Nam	2
		3-SNP_Beuten	2 (1)
	Hybrid model	**Yücebaş-Aydın Son**	None
EG Evaluation	**Hybrid model**	**Yücebaş-Aydın Son**	None
SD Evaluation	**Family history**		Not available
	Age	birth year	1953 (60)
		PCa diagnosis year	
Clinical Evaluation	**Anatomic findings**	weight (kg)	91
		height (cm)	195
		BMI (kg/cm2)	23,93
		BMI (category)	Normal (BB)
	Medical conditions		
Lifestyle (EB) Evaluation	**Drugs**		
	Surgical procedures		
	Supplements		Omega-3 Fish Oil
	Nutrition		
	Physical activity		
	Alcohol		

Evaluation	Approaches	Methods	19-hu43860C
Genomic Evaluation	**Independent associations**	**Homozygote**	14
		Heterozygote	23
		Not-Analyzed	24
	Polygenic risk scores	**Number of SNP (DM)**	37
		Number of SNP (AM)	51
		Evidence-Impact-SNP (DM)	122
		Evidence-Impact-SNP (AM)	171
	Cumulative models	**17-SNP_Helfand**	6 (5)
		9-SNP_Helfand	3 (1)
		5-SNP_Zheng	1 (1)
		5-SNP_Salinas	1 (1)
		4-SNP_Nam	0
		3-SNP_Beuten	1 (1)
	Hybrid model	**Yücebaş-Aydın Son**	None
EG Evaluation	**Hybrid model**	**Yücebaş-Aydın Son**	Not exact
SD Evaluation	**Family history**		Not available
	Age	birth year	1954 (59)
		PCa diagnosis year	
Clinical Evaluation	**Anatomic findings**	weight (kg)	112
		height (cm)	195
		BMI (kg/cm2)	29,45
		BMI (category)	Overweight (CC)
	Medical conditions		Hypercholesterolemia
			Non-melanoma skin cancer
Lifestyle (EB) Evaluation	**Drugs**		Lovastatin
	Surgical procedures		
	Supplements		
	Nutrition		
	Physical activity		
	Alcohol		

Evaluation	Approaches	Methods	20-huD00199
Genomic Evaluation	Independent associations	Homozygote	22
		Heterozygote	21
		Not-Analyzed	10
	Polygenic risk scores	Number of SNP (DM)	43
		Number of SNP (AM)	65
		Evidence-Impact-SNP (DM)	133
		Evidence-Impact-SNP (AM)	204
	Cumulative models	17-SNP_Helfand	6
		9-SNP_Helfand	1
		5-SNP_Zheng	1
		5-SNP_Salinas	2
		4-SNP_Nam	0
		3-SNP_Beuten	0 (1)
	Hybrid model	Yücebaş-Aydın Son	None
EG Evaluation	Hybrid model	Yücebaş-Aydın Son	Not exact
SD Evaluation	Family history		Not available
	Age	birth year	1954 (59)
		PCa diagnosis year	
Clinical Evaluation	Anatomic findings	weight (kg)	75
		height (cm)	172
		BMI (kg/cm2)	25,35
		BMI (category)	Overweight (CC)
	Medical conditions		Hypercholesterolemia
Lifestyle (EB) Evaluation	Drugs		Atorvastatin
	Surgical procedures		
	Supplements		
	Nutrition		
	Physical activity		
	Alcohol		

Evaluation	Approaches	Methods	21-huAC827A
Genomic Evaluation	**Independent associations**	**Homozygote**	17
		Heterozygote	37
		Not-Analyzed	9
	Polygenic risk scores	**Number of SNP (DM)**	54
		Number of SNP (AM)	71
		Evidence-Impact-SNP (DM)	170
		Evidence-Impact-SNP (AM)	225
	Cumulative models	**17-SNP_Helfand**	7
		9-SNP_Helfand	2
		5-SNP_Zheng	1
		5-SNP_Salinas	1
		4-SNP_Nam	2
		3-SNP_Beuten	1 (1)
	Hybrid model	**Yücebaş-Aydın Son**	None
EG Evaluation	**Hybrid model**	**Yücebaş-Aydın Son**	Not exact
SD Evaluation	**Family history**		Not available
	Age	birth year	1954 (59)
		PCa diagnosis year	
Clinical Evaluation	**Anatomic findings**	weight (kg)	79
		height (cm)	170
		BMI (kg/cm2)	27,34
		BMI (category)	Overweight (CC)
	Medical conditions		Hypercholesterolemia
			Hypogonadism
Lifestyle (EB) Evaluation	**Drugs**		Simvastatin
	Surgical procedures		
	Supplements		
	Nutrition		
	Physical activity		
	Alcohol		

Descriptions: EG: Envirogenomic, SD: Sociodemographic, EB: Envirobehavioral, DM Dominant model, AM: Additive model.

APPENDIX H-) Personal disease risks for various cumulative models

Group	Person_id	17-SNP_ Helfand	9-SNP_ Helfand	5-SNP_ Zheng	5-SNP_ Salinas	4-SNP_ Nam	3-SNP_ Beuten
Case	01-hu1213DA	8 (1)	2 (1)	1	1	2	2 (1)
Case	03-huD889CC	4 (5)	1	0	0	1	1 (1)
Case	07-hu28F39C	9	3	1	1	1	1 (1)
Case	13-hu6ED94A	8 (5)	4	1	1	1	2 (1)
Control	02-hu59141C	9	3	2	2	2	1 (1)
Control	04-huF7E042	6 (4)	2	1	1	2	0 (1)
Control	05-hu75BE2C	8	3	1	1	1	1 (1)
Control	06-hu56B3B6	7	3	2	2	0	0 (1)
Control	08-huB59C05	8	3	1	1	2	2 (1)
Control	10-hu7A2F1D	6 (4)	3 (1)	1 (1)	1 (1)	1	1 (1)
Control	12-huD57BBF	9	3	1	1	1	1 (1)
Control	14-huD7960A	6	3	1	1	1	1 (1)
Control	15-hu2E413D	6	1	0	0	0	1 (1)
Control	16-hu76CAA5	8	2	1	1	1	1 (1)
Control	17-huA720D3	8	4	2	2	2	1 (1)
Control	18-hu63DA55	6	2	1	1	2	2 (1)
Control	19-hu43860C	6 (5)	3 (1)	1 (1)	1 (1)	0	1 (1)
Control	20-huD00199	6	1	1	2	0	0 (1)
Control	21-huAC827A	7	2	1	1	2	1 (1)
Values in the parenthesis are number of missing values (unanalyzed SNP alleles).							

CURRICULUM VITAE

Personal Information
Surname, Name: Beyan, Timur
Nationality: Turkish
Date and Place of Birth: 16.09.1968, Sakarya
Marital Status: Married
E-mail: tbeyan@yahoo.com

Education

Degree	Institution	Year of Graduation
PhD	Middle East Technical University, Informatics Institute, Health Informatics	2014
MS	Middle East Technical University, Informatics Institute, Information Systems	2005
BS	Gülhane Military Medical Academy, Medical Faculty	1993

Work Experience

Year	Place	Enrollment
2012-2013	Turkish Armed Forces, Health Command, Department of Health Information Systems and Technologies	Branch Manager of Planning and Management Office.
2008-2012	Turkish Armed Forces, Health Command, Department of Communication, Electronics and Information Systems	Health Information Systems Project Officer
2005-2008	Turkish Armed Forces, Health Command, Department of Communication, Electronics and Information Systems	Health Information Systems Analysis and Design Officer
1993-2005	Turkish Armed Forces (at various units and offices)	Several medical and administrative positions.

Publications

Books:

Baykal N., **Beyan T.** (2004*). Bulanık Mantık, İlke ve Temelleri (Fuzzy Logic, Principles and Foundations)* Bıçaklar Yayınevi.

Baykal N., **Beyan T.** (2004*). Bulanık Mantık, Uzman Sistemler ve Denetleyiciler (Fuzzy Logic, Expert Systems and Controllers)*. Bıçaklar Yayınevi.

Book Chapter:

Beyan T., Aydın Son Y. (2014). Emerging Technologies in Health Information Systems: Genomics Driven Wellness Tracking and Management System (GO-WELL) in N. Bessis

and C. Dobre *Big Data and Internet of Things: A Roadmap for Smart Environments.* Springer.

Research Papers:

Beyan T, Aydın Son Y (2014) *Incorporation of Personal Single Nucleotide Polymorphism Data into a National Level Electronic Health Record for Disease Risk Assessment, Part 1: Literature Review for Requirements,* Journal of Medical Internet Research (JMIR) Medical Informatics **(submitted)**.

Beyan T, Aydın Son Y (2014) *Incorporation of Personal Single Nucleotide Polymorphism Data into a National Level Electronic Health Record for Disease Risk Assessment, Part 2: SNP Data Integrated NHIS-T,* Journal of Medical Internet Research (JMIR) Medical Informatics **(submitted)**.

Beyan T, Aydın Son Y (2014) *Incorporation of Personal Single Nucleotide Polymorphism Data into a National Level Electronic Health Record for Disease Risk Assessment, Part 3: Content Extraction, Evaluation and Discussion,* Journal of Medical Internet Research (JMIR) Medical Informatics **(submitted)**.

Conference Papers:

Beyan T., Beyan OD. (2011). *Kaliteli Tıbbi Bakım İçin Sağlık Malzemeleri Tedarik Zincirinde Bilişim Teknolojilerine Dayalı Dönüşüm: Genel Bir Çerçeve Önerisi (Transformation of Health Materials Supply Chain Based on Information Technologies for Qualified Healthcare).* VIII. Ulusal Tıp Bilişimi Kongresi Bildirileri, TURKMIA'11 Proceedings.

Beyan OD., **Beyan T.**, Baykal N. (2010). *Sağlık Bakımı Performans Ölçümü Ontolojisi ve Bilgi Tabanı*. VII. Ulusal Tıp Bilişimi Kongresi Bildirileri, TURKMIA'10 Proceedings.

Koçgil OD., **Beyan T.**, Baykal N. (2009). *Sağlık Bakımı Performans Ölçümü İçin Bir Karşılaştırma Aracı Önerisi (A Framework Proposal for Health Care Performance Measurement and its Application to National Health Care System)*. The Ministry of Health of Turkey, the International Congress on Quality and Performance in Healthcare, Antalya, Turkey.

Koçgil OD., Baykal N., **Beyan T.** (2007). *Sağlık ve Sağlık Bakım Sistemlerinde Performans İzleme ve Değerlendirme (Performance Measurement and Assessment in Health Care Systems)*. The Ministry of Health of Turkey, the Second e-Health Congress, Antalya, Turkey.

Beyan T., Baykal N., Koçgil OD. (2007). *Kompleks Adaptif Sistem Olarak Sağlık Sistemleri ve Performans (Health Systems as Complex Adaptive Systems and Performance)*. The Ministry of Health of Turkey, the Second e-Health Congress, Antalya, Turkey.

Gulkesen KH., **Beyan T.**, Gul H., Bicakci K. (2006). *Reliability of Health Information on the Turkish Web Sites; Fever in Children at Home.* Medical Informatics 2006, Maastricht, Netherlands. European Notes in Medical Informatics, 2006; 2: 433-438.

Akman LE., **Beyan T.**, Kocgil OD., Mizani AM. (2005). *Yeni Bir Sağlık Standardı: Bakım Kaydının Sürekliliği (Continuity of Care Records)*. II. Ulusal Tıp Bilişimi Kongresi Bildirileri Medical Informatics '05 Turkey.

Thesis:

MS Thesis: Beyan T, A New Fuzzy-Chaotic Modelling Proposal for Medical Diagnostic Processes, 2005.

Foreign Language

English