ASSIST
Association Studies assisted by Inference and Semantic Technologies

Pericles A. Mitkas

Electrical and Computer Engineering
Aristotle University of Thessaloniki, Greece
&
Informatics and Telematics Institute
Center for Research and Technology-Hellas

September 18, 2008
Priority 2.4.11 --- Integrated Biomedical Information for Better Health

Development of methods and systems for:
- Medical knowledge discovery
- Support of prevention, diagnosis and treatment

through integration of:
- Clinical information related to tissues, organs
- Personal related information
- Information at the level of molecules and cells

Phenotypic data:

Genotypic data:

Any Project

New Products and/or Services

Barcelona 18/9/2008
Cervical Cancer
The second leading female cancer worldwide

Incident cases in 2000

<table>
<thead>
<tr>
<th>Region</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>64,895</td>
</tr>
<tr>
<td>Worldwide</td>
<td>470,600</td>
</tr>
</tbody>
</table>

Deaths in 2000

<table>
<thead>
<tr>
<th>Region</th>
<th>Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>28,548</td>
</tr>
<tr>
<td>Worldwide</td>
<td>230,000</td>
</tr>
</tbody>
</table>

In Europe, 80 women die every day of cervical cancer

*Age standardised incidence per 100,000 person years*
Human Papilloma Virus (HPV)
The necessary cause of cervical cancer
The causes of Cervical Cancer

- **HPV infection**
  - Transient Infections
  - Persistent Infection
  - CIN 1
  - CIN 2/3
  - Invasive cancer

- **HPV types and variants**
- **Host factors:**
  - Genetic susceptibility
  - Immunological factors
- **Hormonal factors**
  - Long-term OC use
  - High parity
- **Tobacco smoking**

Barcelona 18/9/2008
Association Studies

Epidemiological data
(lifestyle, marital status, sexual habits, smoking habits, etc)

Clinical data
(clinical tests, health status)

Genetic data
(viral type, variant, load, personal and family genotype alterations)

Medical Inferencing

Assembly of data

Hypothesis testing → Statistical evaluation → Conclusions

Goal:
Identify new markers of risk, diagnosis and prognosis
Current drawbacks in Association Studies

- Despite the increase in the number of studies describing phenotype-genotype associations there is limited progress
- Problems in unification and utilization of data from several similar “isolated” studies
  - Standardization in data collection
  - Standardization in methodology
  - Small sample sizes
  - Cost
  - Time consumption
  - Man power
  - Failure to attempt replication

“Research in isolation”
“Disposable patient study groups”
Medical data come in various forms

<table>
<thead>
<tr>
<th>Clinic 1</th>
<th>HPV Test</th>
<th>+/-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pap Test</td>
<td>Class 1/2/3/4</td>
<td></td>
</tr>
<tr>
<td>MTHFR</td>
<td>+/-</td>
<td></td>
</tr>
<tr>
<td>Colposcopy</td>
<td>Normal/LCIN/HCIN</td>
<td></td>
</tr>
<tr>
<td>Diagnosis</td>
<td>CIN 1/2/3, Ca</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clinic 2</th>
<th>HPV Test</th>
<th>+/-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pap Test</td>
<td>WNL, LCIN, HCIN, Ca</td>
<td></td>
</tr>
<tr>
<td>MTHFR</td>
<td>+/-</td>
<td></td>
</tr>
<tr>
<td>Colposcopy</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Diagnosis</td>
<td>CIN 1/2/3, Ca</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clinic 3</th>
<th>HPV Test</th>
<th>+/-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pap Test</td>
<td>Class 1/2/3/4</td>
<td></td>
</tr>
<tr>
<td>MTHFR</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Colposcopy</td>
<td>Normal/LCIN/HCIN</td>
<td></td>
</tr>
<tr>
<td>Diagnosis</td>
<td>CIN 1/2/3, Ca</td>
<td></td>
</tr>
</tbody>
</table>

Common Medical Data Repository

The case of Cervical Cancer
Data Inhomogeneity

- Differences in Language

**Clinic #1:**
- HPV_result = "16,18"
- Genetic_Info.P53 = A/A
- Genetic_Info.MTHFR = C/C

**Clinic #2:**
- HPV = "16,18"
- Polymorphism-P53 = A/A
- Polymorphism-MTHFR = C/C
Data Inhomogeneity

- Different Data Structures

Clinic #1:
HPV_result = “16,18”

Genetic_Info.P53 = A/A
Genetic_Info.MTHFR = C/C

Clinic #2:
HPV_16 = yes
HPV_18 = yes
HPV_31 = no, ...

Genetic_Info = “P53-A/A, MTHFR-C/C”
Data Inhomogeneity

- Differences in Information Encoding

Clinic #1:
- Cytology Results according to Papanikolaou Encoding
- HPV = “16,18”
- Smoking = “20 cigs / day, 15 years”

Clinic #2:
- Cytology Results according to Bethesda Encoding
- HPV = “high risk”
- Smoking = “heavy”
Data Inhomogeneity

- Different Methodologies

Clinic #1:
- HPV Detection with **PCR/RFLP** (Restriction Fragment Length Polymorphism)

Clinic #2:
- HPV Detection with **HCT** (Hybrid Capture Tube)

Different Reliability
ASSIS+

An Integrated Environment that will:

- Virtually unify multiple patient records
- Facilitate the construction of study groups “on demand”
- Enable association studies to combine phenotypic and genotypic data
- Automate the process of evaluating medical hypotheses

Focused on cervical cancer
Unification

- **Ontologies**
  - Semantic representation of data
    - ONE common terminology
    - ONE common taxonomy
  - Address the problems of differences in "language" and "structure"

- **Inference Rules**
  - Embody medical knowledge
  - Address the problems of differences in "encoding" and "methodology"
    - rely on ontologies
    - use tools from (fuzzy) logic
# ASSIST encoding: CYTOLOGY

(Pap-smear test results)

<table>
<thead>
<tr>
<th>Encoding</th>
<th>AUTH</th>
<th>Charite</th>
<th>Un. of Ghent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0: Normal Inflammation</td>
<td>PAPI</td>
<td>0: normal</td>
</tr>
<tr>
<td></td>
<td>Hyperplasia Metaplasia</td>
<td>PAPII</td>
<td>1: reactive changes</td>
</tr>
<tr>
<td></td>
<td>Carnification</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hyperkeratosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1: CIN1, ASCUS, LSIL, AGUS</td>
<td>PAPIII</td>
<td>2: ASCUS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PAPIIID</td>
<td>3: LSIL (CIN1 - koilocytosis -</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LSIL</td>
<td>mild dysplasia)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>7: AGC-NOS</td>
</tr>
<tr>
<td>2</td>
<td>2: ASC-H, CIN2, CIN3, Ca in situ, AIS</td>
<td>PAPIVa</td>
<td>5: HSIL (CIN2 - CIN3 - CIS -</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PAPIVb</td>
<td>moderate &amp; severe dysplasia)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HSIL</td>
<td>8: AGC-Favours neoplasia</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4: ASC-H</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9: AIS</td>
</tr>
<tr>
<td>3</td>
<td>3: Invasive cancer</td>
<td>PAPV</td>
<td>6: squamous cell carcinoma</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10: Adenocarcinoma</td>
</tr>
</tbody>
</table>
RULE 2:

IF  \(<\text{Histology}>\) definition present

THEN use \(<\text{Histology}>\) as \(<\text{severity index}>\) definition and the validity of the derived severity index is \text{HIGH}

ELSE IF \(<\text{Colposcopy}>\) definition present

THEN use \(<\text{Colposcopy}>\) definition as \(<\text{severity index}>\) definition and the validity of the derived severity index is \text{MEDIUM}

ELSE IF \(<\text{Cytology}>\) definition present

THEN use \(<\text{Cytology}>\) definition as \(<\text{severity index}>\) definition and the validity of the derived severity index is \text{LOW}

ELSE \(<\text{severity index}>\) not defined.
ASSIST Environment

- Ethics Advisory Board
- Maintenance Operator
- Helpdesk
- ASSIST Sites
  - Local Administrator
  - ASSIST Site Doctor
  - ASSIST Site Developer
- ASSIST platform
- Researchers
- Medical Expert

Barcelona 18/9/2008
System Architecture

IMA # 1
IMA # N

ASSIST Interfaces to medical archives

Medical Knowledge Base
Association Study
Inference Engine
Retrieval
Semantic Index and Query
Statistical Analysis
Data aggregation

Query from UI
Core Subsystem Response to UI

Query Module
Display Module
XML/RDF parser

ASSIST User Interface Subsystem

ASSIST Core Subsystem

Sesame
NCBI

Barcelona 18/9/2008
Interfaces to Medical Archives

IMA # 1
IMA # N

Medical Knowledge Base
Association Study
Inference Engine
Retrieval
Semantic Index and Query
Statistical Analysis
Data aggregation

Query Module
Display Module
XML/RDF parser

Query from UI
Core Subsystem Response to UI

ASSIST Interfaces to Medical Archives

ASSIST Core Subsystem

Syntactically Unified Patient Information from Medical Archive Interfaces (RDF)

IMA # N

ASSIST User Interface Subsystem

Barcelona 18/9/2008
ASSIST Core

IMA # 1

IMA # N

Syntactically Unified Patient Information from Medical Archive Interfaces (RDF)

ASSIST Interfaces to medical archives

Medical Knowledge Base

Association Study

Inference Engine

Retrieval

Data aggregation

Semantic Index and Query

Statistical Analysis

ASSIST Core Subsystem

Query Module

Display Module

XML/RDF parser

ASSIST User Interface Subsystem

IMA # N

Syntactically Unified Patient Information from Medical Archive Interfaces (RDF)

ASSIST Interfaces to medical archives

Sesame

NCBI
User Interface

- Medical Knowledge Base
- Association Study
- Inference Engine
- Retrieval
- Data aggregation
- Statistical Analysis
- Semantic Index and Query
- Query from UI
- Core Subsystem Response to UI
- Query Module
- Display Module
- XML/RDF parser

IMA # 1
IMA # N

Syntactically Unified Patient Information from Medical Archive Interfaces (RDF)
ASSIST Interfaces to medical archives

ASSIST Core Subsystem

Sesame

ASSIST User Interface Subsystem
ASSIST Prototype: Data Retrieval

Logically Connected Criteria
Data Retrieval: View Records

Anonymised!

Result Page(s) 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 Next

© 2006 ASSIST Project - A3sociation Studies assisted by Inference and Semantic Technologies - FP6-IST-2004-027510

Barcelona
Association Study: Data Validation

BARCELONA 18/9/2008

© 2006 ASSIST Project - Association Studies assisted by Inference and Semantic Technologies - FP6-187-2004-027510
UI - Data Display
UI

RDF Query

Get Data

Statistics

Valid

Input

Association Study

Get cached
data

Retrieval

RDF Data

Inference Engine

Medical Knowledge Base

Semantic Index

Sesame

RDF Query

UI

Get Data

Statistics

Valid

Input

Association Study

Get cached
data

Retrieval

RDF Data

Inference Engine

Medical Knowledge Base

Semantic Index

Sesame

Barcelona 18/9/2008
UI - Stat Results

Barcelona 18/9/2008
Three Pilot cases

Greek Pilot
Papageorgiou Hospital
1st Department of Obstetrics and Gynaecology

Belgian Pilot
University Hospital Ghent
Department of Genetics
Department of Obstetrics and Gynaecology
Dept of Anatomopathology
Centre of Molecular Diagnostics

German Pilot
Charite Hospital
Department of Gynaecology
Laboratory of Gynaecologic Tumor Immunology

- More than 2000 medical records including genetic data have been collected
- More are expected by the end of 2008

Barcelona 18/9/2008
The ASSIST Vision

ASSIS for Heart Diseases

ASSIS for Cervical Cancer

ASSIS for Breast Cancer

ASSIS for Colon Cancer
ASSIST Extension scenarios

- More clinics
  - The N+1 clinic problem
    - Cost?, Procedure?

- Association Studies in New domains
  - Other types of cancer
  - Other genetically related diseases

- Ontology changes
  - Change (addition/modifications) of stored medical knowledge
    - New info regarding HPV vaccination
      - This may influence the scope of association studies – the need is already obvious
    - New genetic variants (e.g., CNV vs. SNP)
More Clinics

IMA # 1
...
IMA # N
IMA # N+1

ASSIST interfaces to medical archives

Anonymization tool

IMA # 1
...
IMA # N
IMA # N+1

ASSIST interfaces to medical archives

MKB instantiation tool

IMA # N

ASSIST Core Subsystem

Statistical Analysis
Association Study
Inference Engine
Retrieval
Medical Knowledge Base
Semantic Index and Query

Data aggregation

Query Module
Display Module
XML/RDF parser

Query from UI
Core Subsystem Response to UI

UI - Ontology mapping tool

ASSIST User Interface Subsystem

Barcelona 18/9/2008
Ontology changes
http://assist.iti.gr

Thank you

Pericles A. Mitkas

Associate Director
Informatics & Telematics Institute
Center for Engineering Research and Technology – Hellas (CERTH)

Professor
Electrical and Computer Engineering
Aristotle Univ. of Thessaloniki
Email: mitkas@eng.auth.gr

Barcelona 18/9/2008