Building a Collaborative Informatics Platform for Translational Research: An IMI Project Experience

Prof. Yike Guo

Department of Computing

Imperial College London
Living in the Era of BIG

• Big Data: Massive amounts of information derived from dry/wet lab investigations, feasibility studies and clinical trials.

• Big Science: Research silos are evaporating with the merging of scientific methods. Traditional hypothesis-testing studies will couple with data-driven research.

• Big Collaboration: As evidence accumulates, personalized medicine will become a reality, and patient-specific cancer interventions will become available. Teams of disease specialists, researchers and bioinformaticians working in concert in a virtual frontier.
Collaboration Platform is the Key for Doing Big Science with Big Data
New Science Economy Model: Public-Private Partnership for Research
The Requirements for PPP-based Translational Research

• IT Infrastructure: hosted and shared by a community

• Data management: shared with configurable access control by various subgroups of a community

• Software: PaaS for development community

• Analysis: collaborative analysis

• Knowledge management: Dynamic integration of knowledge based on semantics
PPP Example: EU IMI

- New model for Public Private Partnership in Life Science
- Focus on Efficacy in 5 Disease Areas, Safety, Knowledge Management and Education & Training
- IMI projects address key pre-competitive bottlenecks identified by Pharma
U-BIOPRED Project: An IMI Project

Innovative Medicines Imitative (IMI)
• Worlds largest public private partnership
• 2 Billion Euro: 1 Billion from EU, 1 Billion from EFPIA

U-BIOPRED: Unbiased Biomarkers for predicting respiratory disease outcomes
• 40 Party Collaboration between 10 Pharma & 30 Academic Medical Centres

Knowledge Management Work Package
• Imperial Data Co-ordinator, JnJ TranSMART
• AZ, Roche, UCB, CNRS
UBIOPRED Knowledge Product: Handprints

1. Reaching international consensus on diagnostic criteria
2. Creating adult/pediatric cohorts and biobanks
3. Creating novel biology ‘handprints’ by combining molecular, histological, clinical and patient-reported data
4. Validating such ‘handprints’ in relation to exacerbations and disease progression
5. Refining the ‘handprints’ by using preclinical and human exacerbation models
6. Predicting efficacy of gold-standard and novel interventions
7. Refining the diagnostic criteria and phenotypes
8. Establishing a platform for exchange, education and dissemination
A Typical Querying Process

- **Query**
  - Sub-set of « X » matched patient
  - Specific « Y » criteria
  - Sub-set of patients
  - « Z » measurements originating from different partners

**Integrated analysis methods**
- 3 pathways
  - Search for connection to inflammation regulating drugs
  - Query for disease association
  - Create own pathway

**Omics Data export to analysis tool**
- Export Omicsdata-matrix and context to tab. files
  - Upload exported files into Application (cloud/systems)
  - Perform analysis
  - Export analysis results into KR
  - Select patients which provided sample associated the results and check their physiologic data for significant differences.

**Examples**
- e.g. age
- e.g. diagnostic
- e.g. proteomics
The Challenge in Building the KM for UBIOPRED

• The consortium formed as a loosely coupled virtual organisation
• Such a VO conducts highly collaborative and multidisciplinary scientific activities
• Data is generated by many people in the VO with a wide range of connected modalities
• Knowledge is build through an iterative knowledge production process where data need to be incrementally collected, flexibly integrated, systematically analysed and interactively reported
• Project has a period but knowledge need to be managed forever
• Very big task with very little money
U-BIOPRED KM System Design Goal and KM Key Features

- Cohort-based information integration to support clinical driven integrative biomarker study
- Cloud based infrastructure to facilitate collaborative translational research
- An integrated framework supports the management of a collaborative knowledge production process (curation, analysis, content enhancement, reporting, modeling and simulation)
- A sustainable environment with longevity and total ownership: workflow-based PaaS for integrative analytics:
  - Curation ETL workbenches
  - Analytics workbenches
- A scalable platform adaptable for future science/technology development (such as NGS technology; medical image et.al)
UBIOPRED Collaborative TR Platform

Analytical Register: Research Collaboration Management

TranSMART: Cohort-based Omics Data Management

IDBS InforSense Workflow: Analytical Engine + Development PaaS

30/04/2012 Deploying TranSmart to UBIOPRED
UBIOPRED TR Platform Components

- Content Building: Curation (ETL/Ontology)
- Collaborative Study Management (Collaboration Management)
- System Biology Study (Modelling and Simulation)
- Omics Experiment Management (ELN/LIMS)
- Study Data Management (tranSMART)
- Biomarker Data Analysis (Analytical Workflow)
- Collaborative Biomarker Data Analysis (Analytical Workflow)
Curation: Building Study Contents and Background Knowledge

The curation process begins by converting data from original sources into a common format. The data is then organized into a common structure and common ontology or vocabulary. Data is tagged for future referencing and searching, at the record level by Concepts (disease, tissue, platform). Quality-approved data is sent through the ETL Pipeline. Data is analyzed and compared against similarly tagged data, and any unusual features are noted.

tranSMART-based UBIOPRED Study Data Management System
Omics Experiment Management

Omics experiments (Proteomics, lipidomics experiments (WP7) and pre-clinical study (WP6) are managed with EWB of IDBS
Study Data Management

Study oriented curated data together with patient information are integrated and warehoused in tranSMART for analysis (WP8)

From John Shon et.al tranSMART, AMIA TBI 2012
Workflow-based Analytics: A Proteomics Example

- Q-TOF mass spectrometry experiment
  - Mass Spectra
  - MS data preprocessing
    - m/z peak data
      - Calibration
      - Baseline Correction
      - Noise Reduction and Smoothing
      - Peak Detection
      - Peak Alignment
  - Peak Selection
    - Selected peaks
      - Cross-validation
        - CV training set
        - Supervised learning (e.g., SVM, discriminant analysis)
        - Classification model
      - CV testing set
  - Label prediction
    - Use Monte Carlo method to randomly select other biomarker detection models and alternative parameter settings
  - Model performance
Workflows in InforSense
Workflow Deployed as Cloud Applications
A Development Cycle of Cloud Application Building via Workflow

Current development/deployment paradigm + Elasticity + Economic Model
Analytical Register: Enables Collaborative Knowledge Management
Mapping Research Collaboration in UBIOPRED
Analytical Register: A Study View

U-BIOPRED Analytical Register

Study: U-BIOPRED

About

The U-BIOPRED study will collect various data types to support the generation of the Handprint. Each data set is used to generate an individual fingerprint that in turn will be combined into a specific handprint.

Data available

<table>
<thead>
<tr>
<th>Data Class</th>
<th>Assay/Data Source</th>
<th>Cell Types</th>
<th>Number of Samples</th>
<th>Contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical</td>
<td>ECRF</td>
<td></td>
<td></td>
<td>Kenny Dyson</td>
</tr>
<tr>
<td>Clinical</td>
<td>Patient Reported Outcomes</td>
<td></td>
<td></td>
<td>Kenny Dyson</td>
</tr>
<tr>
<td>Clinical</td>
<td>Patient Diaries</td>
<td></td>
<td></td>
<td>Florian Singer</td>
</tr>
<tr>
<td>Transcriptomics</td>
<td>Human Genome U133 Plus 2.0 Array</td>
<td></td>
<td></td>
<td>WP7-LOIC</td>
</tr>
<tr>
<td>Proteomics</td>
<td>Unselective ESI</td>
<td></td>
<td></td>
<td>Paul Skipp</td>
</tr>
<tr>
<td>Proteomics</td>
<td>MSII</td>
<td></td>
<td></td>
<td>Paul Skipp</td>
</tr>
<tr>
<td>Lipidomics</td>
<td></td>
<td></td>
<td></td>
<td>Tony Postle</td>
</tr>
<tr>
<td>Lipidomics</td>
<td>Eicosanoids</td>
<td></td>
<td></td>
<td>Craig Wheelock</td>
</tr>
<tr>
<td>Lipidomics</td>
<td></td>
<td></td>
<td></td>
<td>WP7-Kallnow</td>
</tr>
<tr>
<td>Breathomics</td>
<td>eNose</td>
<td></td>
<td></td>
<td>Peter Benkman</td>
</tr>
<tr>
<td>Breathomics</td>
<td></td>
<td></td>
<td></td>
<td>Philips</td>
</tr>
<tr>
<td>Metagenomics</td>
<td></td>
<td></td>
<td></td>
<td>Peter Hewarth</td>
</tr>
<tr>
<td>CT</td>
<td>AiProm</td>
<td></td>
<td></td>
<td>Sumit Gupta</td>
</tr>
<tr>
<td>Histology</td>
<td>Waiting for Agreement</td>
<td>Bronchoscopy</td>
<td></td>
<td>Susan Wilson</td>
</tr>
</tbody>
</table>

Work Packages in this study:

- UBIO-WP10
- UBIO-WP2
- UBIO-WP3
- UBIO-WP4
- UBIO-WP5
- UBIO-WP6
- UBIO-WP7
- UBIO-WP8
- UBIO-WP9
Work Package: UBIO-WP5

Investigations

- Viral induced changes handprints
- The 8.334.140.1 axis in asthma
- Exome re-sequencing in asthma
- Investigation of Genes from GWAS of lung function in asthma
- Genome Wide Association study of Airway Remodelling Measures in asthma
- GWAS - Quantitative Trait Loci approaches in asthma
- The role of urokinase receptor in asthma
- Epithelial genetics in asthma

Researchers

- Norbert Krug
- Peter Sterk
- Scott Wagers
Analytical Register: A Cohort View

U-BIOPRED Analytical Register

Cohorts

Add new cohort

Filter by Study: All - AirProm - Debrecon RMA - ECLIPSE - Fan Macrophage - U-BIOPRED

Showing all cohorts.

<table>
<thead>
<tr>
<th>Name</th>
<th>Study</th>
<th>Researcher</th>
<th>Size (1)</th>
<th>Size (2)</th>
<th>Transmart</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sputum Microarray for Gold II and Gold III patients</td>
<td>ECLIPSE</td>
<td>Shu Lin Yan (LOIC)</td>
<td>69</td>
<td>58</td>
<td>178/586</td>
</tr>
<tr>
<td>Severe asthma with irreversible airflow limitation (as defined by Ton Binike at al)</td>
<td>U-BIOPRED</td>
<td>Ariane Wagner (AMST)</td>
<td>178/006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severe asthma with irreversible airflow limitation (as defined by TENOR)</td>
<td>U-BIOPRED</td>
<td>Ariane Wagner (AMST)</td>
<td>178/0076</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severe asthma with irreversible airflow limitation (as defined by ERS guidelines)</td>
<td>U-BIOPRED</td>
<td>Ariane Wagner (AMST)</td>
<td>178/0111</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identifying severe asthma with irreversible airflow limitation (as defined by Ton Binike et al)</td>
<td>ECLIPSE</td>
<td>Michelle Osmond (LOIC)</td>
<td>132</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identifying severe asthma with irreversible airflow limitation (as defined by TENOR)</td>
<td>ECLIPSE</td>
<td>Michelle Osmond (LOIC)</td>
<td>191</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PEV1 vs Exacerbation Rate</td>
<td>ECLIPSE</td>
<td>Anthony Rowe (JUL)</td>
<td>148</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COPD Patients for Clustering Analysis</td>
<td>ECLIPSE</td>
<td>Manohar Barwatage (LOIC)</td>
<td>148</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal BMI</td>
<td>U-BIOPRED</td>
<td>David Gibson (LOIC)</td>
<td>148</td>
<td></td>
<td></td>
</tr>
<tr>
<td>View Cohort 1</td>
<td></td>
<td></td>
<td>Yang Li (LOIC)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

11 found.

Add new cohort
Analytical Register: Interfacing to Cohort Selection
Analytical Register – Cohort View

Cohort: Sputum Microarray for Gold II and Gold III patients

Subset 1
Size: 69
Query:
Affymetrix GeneChip HG-U133 Plus 2. Sputum and GOLD: Gold II

Subset 2
Size: 58
Query:
Affymetrix GeneChip HG-U133 Plus 2. Sputum and GOLD: Gold III

Used in the following:
- Investigation: Validate ECLIPSE microarray data loaded into transSMART

Subject Totals

<table>
<thead>
<tr>
<th>Subset 1</th>
<th>Both</th>
<th>Subset 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>69</td>
<td>0</td>
<td>58</td>
</tr>
</tbody>
</table>

Histogram of Age

Comparison of Age

Subset 1
Mean: 65.29
Median: 60
IQR: 2
SD: 6.62
Data Points: 69

Subset 2
Mean: 64.14
Median: 64
IQR: 2
SD: 6.03
Data Points: 58
Analytical Register: An Investigation View

U-BIOPRED Analytical Register

Investigations

Filter by:
- Status
- Study
- Topic
  - AirProm
  - Diocere RNA
  - ECLIPSE
  - FanMacrophage
  - U-BIOPRED

Work Package
- Category

Showing all investigations from U-BIOPRED, all work packages, in any categories.

The role of urokinase receptor in asthma

Proposed Initiator: Dominick Shaw

- AirProm
- Diocere RNA
- ECLIPSE
- FanMacrophage
- U-BIOPRED

Working with U-BIOPRED Ubio-WP3

Epithelial genetics in asthma

Proposed Initiator: Dominick Shaw

16 Apr 2012 - 13:59: Background: GWAS data in asthma suggests that multiple genes of relevance to airway epithelial function are potentially altered. These findings require the characterisation of cellular responses in cancers and non-cancers of relevant risk alleles.

Aims: To identify gene sets and non-genes of relevant risk alleles and characterise primary bronchial epithelial cell functions relevant to these targets e.g. EMT-A3 SNPs and TGF-1 signaling,

- genotypes U-BIOPRED Ubio-WP3
- Ubio-WP4
- Ubio-WP5
- Ubio-WP6
- Ubio-WP7

GWAS - Quantitative Trait Loci approaches in asthma

Proposed Initiator: Dominick Shaw

16 Apr 2012 - 12:05: Background: Data relating genetic polymorphisms with mRNA expression in asthma subjects is sparse. These advances in genetic approaches it is now possible to generate a GWAS SNP dataset and global mRNA expression profile from any tissue or cell. These data will potentially aid the translation of new findings from asthma GWAS and potentially our understanding of asthma in new directions.

Aims: To explore GWAS and quantify gene expression in all subjects e.g. Atopy score, by bronchial epithelial cells, by airway fibroblasts, by airway smooth muscle cells and by pathological blood mononuclear cells and generate QTL datasets,

- ECLIPSE, U-BIOPRED Ubio-WP3
- Ubio-WP4
- Ubio-WP5
- Ubio-WP6

Genome Wide Association study of Airway Remodeling Measures in asthma

Proposed Initiator: Dominick Shaw

16 Apr 2012 - 10:54: Background: Structural alterations termed “airway remodeling” occur in asthma and the extent of remodeling is variable. Airway remodeling is a characterised by smooth muscle hypertrophy/hyperplasia, subepithelial fibrosis, basement membrane thickening, increased subepithelial matrix deposition, angiogenesis and more recently alterations in epithelial functions have been identified including metabolism and goblet cell hyperplasia. The mechanisms underlying airway remodeling remain to be determined.
Investigation: Is there a relationship between Exacerbation Rate and FEV-1 in COPD subjects?

Research Hypothesis

A core dogma of COPD research is that as FEV-1 decreases the number of exacerbations a COPD sufferer experiences will increase. Or to put simply it is expected that exacerbations will increase as lung function deteriorates.

This study investigates whether this correlation exists in the Eclipse study data.

Cohorts Used

- FEV1 vs Exacerbation Rate (TranSMART: 1787925)

Method and Results
Method and Results

Based on the ECLIPSE cohort of patients we fitted linear models between the % predicted normal post-dose FEV1 score and the exacerbation rate that patients had in Years 1, 2 & 3 of the ECLIPSE study. Selection of the patient data was conducted through TranSMART and the patient data extracted and analysed using the R platform. (The R Code for this analysis is attached here)

- Year 1 - FEV1 VS Exacerbation Rate (R2 = 0.07121):

![Graph showing FEV1 vs Exacerbation Rate (Year 1)]

- Year 2 - FEV1 VS Exacerbation Rate (R2 = 0.06442):

![Graph showing FEV1 vs Exacerbation Rate (Year 2)]
Analytical Register: Interfacing to Sample Tracking
Analytical Register: Fully Integrated with TranSMART

### Summary Statistics

**Query Summary for Subset 1**

(\(\text{BMI} \text{ Studies} \land \text{BMI} \text{ Subjects} \land \text{BMI} \text{ Data} \land \text{BMI} \text{ function tests} \land \text{Reversibility/FEV1 post salbutamol (L) \land Baseline Day 1} \land \text{AND} \land \text{BMI} \text{ Studies} \land \text{BMI} \text{ Subjects} \land \text{BMI} \text{ Data} \land \text{BMI} \text{ function tests} \land \text{Reversibility/FVC post salbutamol (L) \land Baseline Day 1} \land \text{AND} \land \text{BMI} \text{ Studies} \land \text{BMI} \text{ Subjects} \land \text{BMI} \text{ Data} \land \text{BMI} \text{ Subjects Demographic} \land \text{BMI} \text{ Gender}) \land \text{OR} \land \text{BMI} \text{ Studies} \land \text{BMI} \text{ Subjects} \land \text{BMI} \text{ Data} \land \text{BMI} \text{ Subjects Demographic} \land \text{BMI} \text{ Data} \land \text{BMI} \text{ Age at Consen}) \land \text{AND} \land \text{BMI} \text{ Studies} \land \text{BMI} \text{ Subjects} \land \text{BMI} \text{ Data} \land \text{BMI} \text{ Group} \land \text{BMI} \text{ Cohort A: Severe non-smoking asthma} \land \text{OR} \land \text{BMI} \text{ Studies} \land \text{BMI} \text{ Subjects} \land \text{BMI} \text{ Data} \land \text{BMI} \text{ Group} \land \text{BMI} \text{ Cohort B: Severe smoking asthma})

### Subject Totals

<table>
<thead>
<tr>
<th>Subset 1</th>
<th>Both</th>
<th>Subset 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### Comparison of Age

**Histogram of Age**

**Comparison of Age**

---

**Notes:**

1. The search results are filtered by various conditions related to BMI and demographic data, indicating a comprehensive analysis approach.
2. The subject totals show no matches for the specified criteria in Subset 1 and Subset 2.
3. The histogram suggests a distribution analysis, potentially for age-related data, but no specific age ranges are shown.

---

**Further Analysis:**

- Explore additional filters for more specific data insights.
- Use the histogram for age-specific comparisons in clinical studies.
UBIOPRED
System Biology Study

1. We have developed p38 pathway model in SBML and perform simulation throughput MATLAB.
2. We have proposed a novel glucocorticoid signalling pathway model in SBML.
3. We have proposed some hypotheses to suggest possible crosstalk between GC and p38 pathways.

4. We then construct the integrated pathway model in SBML.

5. Test the stability of this integrated model.
Extending UBIOPRED to Support Multiple Projects

Project 1

Virtual Container

DB Instance

Transmart Service

Data Centre

Project n
eTRIKS: European translational research and knowledge management services
Conclusion: the end of the beginning

- UBIOPRED knowledge management system supports cross-institutional large scale translational research
- UBIOPRED knowledge management system supports the entire knowledge production process
- For PPP-based TR projects, collaborative research support in a VO environment is circuitual
- Modern technology such as cloud and workflow provide the technical foundation for its implementation
- The development is moving forward to support IMI translational projects in general by eTRIKS project