Pathology: the Key to Translational Research

J. Felix Liu, Ph.D.  
Director  
Felix.liu@ufl.edu

Christopher P. Barnes  
Assistant Director  
cpb@ufl.edu

Clinical and Translational Science IT  
Clinical and Translational Science Institute  
University of Florida
Outline

+ The Key
+ Examples: Pathology Data in Action
+ Translational Research Cycle
+ Data to Information to Knowledge
+ Collaboration
+ Growth in Data
+ The Needs for Ontology
70% of the clinical data in the electronic medical record are from pathology.

70% of requests for data from the electronic medical record are for pathology data.
Pathology Data Usage in Research

- Clinical
- Translational
- Personalized Medicine
- Quality
- Outcomes
- Hospital Economics
The Key to Translational Research
Research Use Case

Prediction of Perioperative Acute Kidney Injury using Machine Learning

**Outcome:** Perioperative AKI - Change in sCr day 1-7
Research Use Case

Genetic Data Integration

The UF Customized Gene Chip

- 256 SNPs
  - Covers 120 genes
  - 251 pharmacogenetics SNPs
    - 8 for CYP2C19 clinical testing
    - 3 for sex markers
    - 2 for QA
  - $30 per test or 12¢ per SNP (1/10 of commercial costs)
**Genetic Data Integration**

Clopidogrel is used to prevent platelets, a type of cell found in the blood, from clumping together. When platelets clump together they can form clots which block the flow of blood.

| Reduced CYP2C19 activity | **Poor Metabolizer (3%)**  
|                          | Not likely to receive full benefit of clopidogrel. Increased risk for heart attack and stroke. |
|                         | **Intermediate Metabolizer (20%)**  
|                          | May not receive full benefit of clopidogrel. Possible increased risk for heart attack and stroke. |
| Typical CYP2C19 activity | **Extensive Metabolizer (38%)**  
|                          | Expected to benefit from standard clopidogrel dose. |
| Increased CYP2C19 activity | **Ultra-Rapid Metabolizer (32%)**  
|                          | Expected to process medication more quickly. Possible increased benefit. Possible increased risk for bleeding. |
| Uncertain CYP2C19 activity | **Metabolizer Status Unknown (7%)**  
|                          | Not enough data to determine clopidogrel response. |
Research Use Case

Genetic Data Integration

Personalized Medicine for Pharmacogenetics
Biorepository

- Patient consent for spare specimen collection for research use
- Informatics support for using these specimen in IRB approved research – including usage in de-identified way
- Linkage to other clinical data
- Collective efforts among 3 IT units
Electrophoresis - SPE
Electrophoresis - IFE
Breast Cancer Study
Breast Cancer Study

Breast Cancer Dashboard

Welcome, LIU43!

- Overview - SFBCAS
- Comprehensive - SFBCAS
- Mets Biochemistry - SFBCAS
- Report - TNBC

<table>
<thead>
<tr>
<th>Triple Neg</th>
<th>ER</th>
<th>PR</th>
<th>HER2: IHC</th>
<th>FISH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race</th>
<th>DX Age</th>
<th>Menopause Status</th>
<th>Grade</th>
<th>High Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between</td>
<td>40 and 60</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tumor Type</th>
<th>Basal Like Class</th>
<th>Gen Class</th>
<th>Chemo Text Keyword</th>
<th>Hormone Text Keyword</th>
<th>Recurrence Type</th>
<th>Post Recurrence Survival</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(All Choices)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Survival</th>
<th>Cause of Death</th>
<th>DFS</th>
<th>Between</th>
<th>and</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| TAB No. | Race | DX Age | Menopause Status | Last Contact Date | Block Name | ER | PR | HER2 IHC | HER2 FISH | Triple Negative | High Grade | Tumor Type | Gen_Class | Histology | Nodes Examined | Positive Nodes | Best Stage | Clinical Stage | Pathological Stage | Chemo Text | Hormone Treat Date | Hormone Text | Radiation Treat Date | Other Treat Date | First Recurrence Date | First Recurrence Site | First Recurrence Cause of Death | First Recurrence Survived | Post Recurrence Survival | Post Recurrence Site 3 | Post Recurrence Site 5 | Post Recurrence Site 10 | Post Recurrence Site 20 | Post Recurrence Site 50 | Post Recurrence Site 100 |
|---------|------|--------|------------------|-------------------|------------|----|----|----------|----------|-----------------|------------|------------|-----------|-----------|-----------------|----------------|-----------|----------------|-------------------|------------|-------------------|---------------|-------------------|-----------------|----------------------|----------------|---------------------|----------------|--------------------|---------------------|------------------------|----------------------|---------------------|------------------|
| 01-001  | White| 54 No  |                  | 01/01/2001        | Tumor1     | Negative | Negative | Negative | Yes         | 3               | high grade | invasive ductal | Basal       | 80003       | 01/01/2001 | AC           | 01/01/2000 | EBRM NEGATIVE | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 |
| 01-002  | White| 54 No  |                  | 01/01/2001        | Tumor1     | Negative | Negative | Negative | Yes         | 3               | high grade | invasive ductal | Basal       | 80003       | 01/01/2001 | AC           | 01/01/2000 | EBRM NEGATIVE | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 |
| 01-003  | Black| 45 No  |                  | 01/01/2001        | Tumor1     | Negative | Negative | Negative | Yes         | 3               | high grade | invasive ductal | Basal       | 80003       | 01/01/2001 | AC           | 01/01/2000 | EBRM NEGATIVE | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 |
| 01-004  | Black| 45 No  |                  | 01/01/2001        | Tumor1     | Negative | Negative | Negative | Yes         | 3               | high grade | invasive ductal | Basal       | 80003       | 01/01/2001 | AC           | 01/01/2000 | EBRM NEGATIVE | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 |
| 01-005  | Black| 45 No  |                  | 01/01/2001        | Tumor1     | Negative | Negative | Negative | Yes         | 3               | high grade | invasive ductal | Basal       | 80003       | 01/01/2001 | AC           | 01/01/2000 | EBRM NEGATIVE | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 |
| 01-006  | Black| 45 No  |                  | 01/01/2001        | Tumor1     | Negative | Negative | Negative | Yes         | 3               | high grade | invasive ductal | Basal       | 80003       | 01/01/2001 | AC           | 01/01/2000 | EBRM NEGATIVE | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 |
| 01-007  | Asian | 45 No  |                  | 01/01/2001        | Tumor1     | Negative | Negative | Negative | Yes         | 3               | high grade | invasive ductal | Basal       | 80003       | 01/01/2001 | AC           | 01/01/2000 | EBRM NEGATIVE | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 | 01/01/2000 |
Aberrant expression of DNA damage response proteins is associated with breast cancer subtype and clinical features

Data → Information → Knowledge

+ **Data**
  + symbols

+ **Information**
  + data that are processed to be useful
  + provides answers to "who", "what", "where", and "when"

+ **Knowledge**
  + application of data and information
  + Answers to "how"

+ **Understanding**
  + appreciation of "why"

+ **Wisdom**
  + evaluated understanding
The Dollar Sign

- For all systems to understand each other
- So computers can be used for reasoning
- Standard based, ontology enabled systems ($$
- Subject matter experts ($$$$
- More and better Research
- Collaboration and information exchange
Without Standard Terminology

TheBeta-Blocker Query

Find patients who are taking
Alprenolol, Carteolol, Levobunolol, Mepindolol,
Metipranolol, Nadolol, Oxprenolol, Penbutolol, Pindolol,
Propranolol, Sotalol, Timolol, Acebutolol, Atenolol,
Betaxolol, Bisoprolol, Esmolol, Metoprolol, Nebivolol,
Carvedilol, Celiprolol, Labetalol, Butaxamine
And the data is in X, Y & ....Z Tables
Hmm.... did I miss anything?
Find all patients who take one of the beta blockers in the class that is highlighted.
Example: I want patient’s Glucose Test Results

<table>
<thead>
<tr>
<th>Test Code</th>
<th>Test Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLUBFP</td>
<td>BODY FLUID GLUCOSE</td>
</tr>
<tr>
<td>GLU</td>
<td>GLUCOSE</td>
</tr>
<tr>
<td>GTT2</td>
<td>GLUCOSE TOLERANCE, 2 HOUR</td>
</tr>
<tr>
<td>GTT3</td>
<td>GLUCOSE TOLERANCE, 3 HOUR</td>
</tr>
<tr>
<td>GTT4</td>
<td>GLUCOSE TOLERANCE, 4 HOUR</td>
</tr>
<tr>
<td>GTT5</td>
<td>GLUCOSE TOLERANCE, 5 HOUR</td>
</tr>
<tr>
<td>GTT6</td>
<td>GLUCOSE TOLERANCE, 6 HOUR</td>
</tr>
<tr>
<td>GT1</td>
<td>GLUCOSE, 1 HOUR POST</td>
</tr>
<tr>
<td>GT2</td>
<td>GLUCOSE, 2 HOUR POST</td>
</tr>
<tr>
<td>GT3</td>
<td>GLUCOSE, 3 HOUR POST</td>
</tr>
<tr>
<td>GT4</td>
<td>GLUCOSE, 4 HOUR POST</td>
</tr>
<tr>
<td>GT5</td>
<td>GLUCOSE, 5 HOUR POST</td>
</tr>
<tr>
<td>GT0</td>
<td>GLUCOSE, FASTING</td>
</tr>
<tr>
<td>GLU1HR</td>
<td>GLUCOSE, 1 HOUR</td>
</tr>
<tr>
<td>GLU2HR</td>
<td>GLUCOSE, 2 HOUR</td>
</tr>
<tr>
<td>GSFLGLU</td>
<td>SPINAL FLUID GLUCOSE</td>
</tr>
<tr>
<td>UGL</td>
<td>URINE GLUCOSE, QUANT</td>
</tr>
<tr>
<td>UGLR</td>
<td>URINE GLUCOSE, RANDOM</td>
</tr>
</tbody>
</table>
To share and collaborate:

We need the data to be structured in a way that helps us interpret it.

+ IE: From the example before, which Glucose am I looking for?

We need to find ways to control the Tsunami of data that is coming at us every day.

We need to be able to quickly and easily collaborate and LINK data with other peoples data WITHOUT their help

+ Too often data in silos are difficult to use and share EVEN IF the data is in the same Standard format because, of issues like technology mismatch, policy mismatch or geographic mismatch

+ This is where **ontology** comes in.
Why should I care about ontology?

Because - Its all about the data

“The amount of data in the world today is equal to:

- Every person in the US tweeting three tweets per minute for 26,976 years
- Every person in the world having more than 215m high-resolution MRI scans a day
- More than 200bn HD movies – which would take a person 47m years to watch
- Fill 57.5bn 32GB iPads”

Unstructured Data is Growing faster than we can store it.

http://www.economist.com/node/15557443
Auntie M, the data’s coming
Where can we start bringing order to chaos?

- **Controlled Vocabularies** (less duplicate data)
  - isA controlled vocabulary is a set of terms with very well defined meaning.

- **Ontologies** (even less duplicate data because it can be LINKED instead of copied)
  - isA controlled vocabulary that DEFINES the RELATIONSHIPS between the concepts and terms in its domain.
We could always make another list?

**Logical Observation Identifiers Names and Codes (LOINC)**

“A universal code system will enable facilities and departments across the world to receive and send results from their areas for comparison and consultation and may contribute toward a larger public health initiative of improving clinical outcomes and quality of care. LOINC is one of the standards for use in U.S. Federal Government systems for the electronic exchange of clinical health information.”([http://en.wikipedia.org/wiki/LOINC](http://en.wikipedia.org/wiki/LOINC))

Does the list of data tell you anything about what hemoglobin relates to?

<table>
<thead>
<tr>
<th>LOINC</th>
<th>Unit code name</th>
<th>Unit code</th>
<th>Performed/reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>4548-4</td>
<td>Hemoglobin A1C, B</td>
<td>82080</td>
<td>Central Clinical Lab/in MICS</td>
</tr>
<tr>
<td>4548-4</td>
<td>Hemoglobin A1C, B</td>
<td>80947</td>
<td>Clinical Trials/not in MICS</td>
</tr>
<tr>
<td>4548-4</td>
<td>Hemoglobin A1C, B</td>
<td>124055</td>
<td>New England/not in MICS</td>
</tr>
<tr>
<td>4548-4</td>
<td>Hemoglobin A1C, B</td>
<td>82990</td>
<td>Kasson/in MICS</td>
</tr>
</tbody>
</table>
Isn’t that enough, I have my list of terms?

+ NO, we want a computer algorithm to KNOW that hemoglobin is a “binding protein” and a “protein” without a human having to make any real-time mappings or interpretation.

SNOMED Clinical terms AND ONTOLOGICAL relationships (http://bioportal.bioontology.org/ontologies/46896?p=terms&conceptid=38082009#visualization)
So what can I do with all these new “relationships”?

- Have a computer chew on it for you. Find and correlate related information via a computerized reasoning process, BEFORE its handed to poor overworked you.
You could find new collaborators based on genes known to be linked to breast cancer.

Computer:
*Find all faculty members whose genetic work is implicated in breast cancer*
Or you could create a computer Jeopardy Overlord!...

Thank you.

Questions?

?snoitseuQ