High Performance Phenotyping

Proposal based on GELLO and UIMA-AS processing model
Outline

- Overview of GELLO / UIMA-AS solution
  - Problem Statement
  - Technology Overview
  - Data Capture of Phenotype Query
  - Stage 1 – Exclusion Processing
  - Stage 2 – Inclusion Processing
Problem Statement

• How do we create a generalized high performance phenotyping solution, that is potentially useful?
  • Characteristics
    • Usable by researchers and physician community
    • Supports the standardized representation of criteria
    • Capable of process millions of patient records
    • Supported across divergent practice systems
    • Interoperates with the other SHARP initiatives
Technology Overview

- Current Informatics Cloud technology is based on a system design, utilizing poly-core processors (2x12), +100Gb memory capacities, and +10 Tb storage.

- Performance characteristics
  - CPU ≈ 2.2 GHz or (2,200,000,000 instructions / second)
  - Disk ≈ 5 ms avg. seek or (200 reads / second)

1 disk seek ≈ 11,000,000 instruction execution
Data Capture of Phenotype Query

X-Form Entry of specifications

Stage 1
Exclusion Processing:
Remove bulk of patients from candidate list.

Standardized Exclusion Criteria

Stage 2
Inclusion Processing:
Determine candidate inclusion via review of Clinical Element Models.

Detailed Study Inclusion Criteria
Stage 1 - Exclusion Processing

- Each patient is assigned a 100 byte, pheno-index
  - Proposed content:
    - 30 bytes basic demographics
      - 10 Bytes for Patient Id, as used in CEM dataset.
      - 20 Bytes for gender code, birth date and Zip code
    - 70 bytes current health
      - 10 (4 Byte) selected physical quantity values
      - 240 (30 Byte x 8 bit flags) selected diagnostic codes

- Given 5 million patients, each with 100 byte pheno-index, we can create a read-only memory structure utilizing on 500 million bytes or (0.5 Gb memory).
Stage 1 - Exclusion Processing

Identify the set of Patients to move to Stage 2

Exclusion Record

Exclusion Processor

PhenoIndexer

NLP Processing / Data Normalization

UIMA-AS Processing

CEM Processing

Pheno Indexer

CEM Database

Standardized Exclusion Criteria

Detailed Study Inclusion Criteria GELLO Source
Stage 1 - Exclusion Processing

- Why propose an in-memory exclusion processing?
  - Given 5 million patients and 253 pheno-index features per patient. \textit{E.g.} 3 demographic + 10 PQ values + 240 diagnostic flags
  - If each feature is evaluated using approx. 20 machine instructions, then on a single 2.2 GHz core machine...

Estimate runtime = \((5,000,000 \times 253 \times 20)/2.2 \text{ billion.}\)

= 11.5 seconds

Utilizing 10 cores, we could potentially improve performance.
Stage 1 - Challenges

- Standard features to be utilized in the Pheno-Index need to be identified.
  - Could consider review of previous criteria utilized to determine likely candidate features.
- X-Form data entry section for exclusion needs design.
- Transform from X-Form to Exclusion record, needs to be informed by Exclusion processing input requirements.
- Output from Stage 1 to Stage 2, may benefit from a batching of candidate patient IDs in an exchange record.
  - This will need to be evaluated.
Stage 2 - Inclusion Processing

1. Standardized Exclusion Criteria
2. Detailed Study Inclusion Criteria
3. GELLO Source
4. X-Form Transform to GELLO
5. Inclusion Record
6. Patient identified from Stage 1
7. GELLO Inclusion Criteria from Stage 2

NLP Processing / Data Normalization

- UIMA-AS Processing
- CEM Processing
- Pheno Indexer

CEM Database
What is GELLO?

- GELLO is a class-based, object-oriented (OO) language that is built on existing standards. GELLO expression language is based on the Object Constraint Language (OCL), developed by the Object Management Group. Relevant components of OCL have been selected and integrated into the GELLO to provide a suitable framework for manipulation of clinical data for decision support in health care.
- The GELLO language can be used to:
  - Build up expressions to extract and manipulate data from medical records.
  - Construct decision criteria by building up expressions to reason about particular data features/values. These criteria can be used in decision-support knowledge bases such as those designed to provide alerts and reminders, guidelines, or other decision rules.
  - Create expressions, formulae, etc. for other applications.
GELLO syntax (Excerpt)

```plaintext
maintenance:
title: Screening for elevated calcium-phosphate product;
library:
purpose: provide an alert if the product of the blood calcium and phosphorus exceeds a certain threshold in the setting of renal failure;
explanation: An elevated Ca-P04 product suggests a tendency toward renal osteodystrophy and predisposes to soft-tissue calcification;

let lastCreatinine : Observation = Observation→ select(code="SNOMED-CT", "xxxxxx") .sortBy(effectiveTime.high).last()

let lastCalcium : Observation = Observation→ select(code="SNOMED-CT", "yyyyy") .sortBy(effectiveTime.high).last()

let lastPhosphate : Observation = Observation→ select(code="SNOMED-CT", "zzzzz") .sortBy(effectiveTime.high).last()

let renal_failure_threshold : PhysicalQuantity = Factory.PhysicalQuantity("2.0, mg/dl")

let threshold_for_osteodystrophy : int = 70

let renal_failure : Boolean = if lastCreatinine <> null and lastCreatine.value.greaterThan(renal_failure_threshold) then true else false Endif

let calcium_phosphate_product : real = if lastCalcium <> null and lastPhosphate <> null then lastCalcium.value * lastPhosphate.value else -1 endif
```
Stage 2 - Inclusion Processing

Patient identified from Stage 1

GELLO Inclusion Criteria from Stage 2

UIMA-AS based Phenotype Processor

- Identify Required Fields
- CEM Query Processing
- Execute Logical Evaluation
- Store Results

Query CEM for required fields.

CEM Database

Candidates
Stage 2 - Inclusion Processing

- Potential performance improvements in Stage 2:
  - Establish parallel processing nodes for accessing CEM data from CEM Repository.
  - Partition the CEM patient data into multiple databases, (E.g. 10 separate CEM databases with data from patients with the same last digit in their Patient-ID.)
    - Establish multiple UIMA-AS streams of processing, assigning patients with the same last digit to the same pipeline.
  - GELLO is proposed as a model of a language that we might adopt for CEM data access and processing. Need further investigation to determine adherence to the GELLO Std.
DISCUSSION