Meaningful query using SNOMED CT

How can semantic searches improve patient recruitment?

Brandon Ulrich & Orsolya Bali – snowowl@b2international.com

Background

Data that help the health industry improve patient care and reduce costs will be their most valuable asset in 5 years.

Two-thirds of health organizations expect their secondary data use to increase significantly.

The availability of semantically rich electronic health records (EHRs) utilizing SNOMED CT as a reference terminology continues to grow providing new opportunities for secondary use.

Problem

Current systems are unable to perform real-time, semantically rich searches on EHR data.
Real-time queries are:

- Limited to detailed information from a single patient
- Limited to aggregated information for all patients with a small number of predefined concepts
- Unable to access semantics

Which limits extraction of knowledge from the EHR data:

- Limiting the value of clinical decision support
- Preventing use of medical knowledge and semantic relationships between concepts to identify associations in existing data
- Limiting continuous identification of emerging associations as data accumulates

Furthermore:

- The volume of EHR data is growing exponentially
- Traditional data warehouses struggle to exploit the full meaning within the EHRs, as the data is built around a limited number of concepts

Example: Improving clinical trial patient recruitment – Arthritis study protocol

Background

A clinical trial is testing a medication that helps patients in certain risk groups avoid developing arthritis.

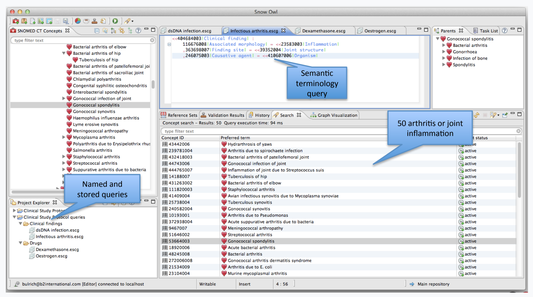
We semantically search 1,200,000 EHRs from the Observational Medical Outcomes Partnership (OMOP) to identify candidate patients for the trial based on study inclusion and exclusion criteria.

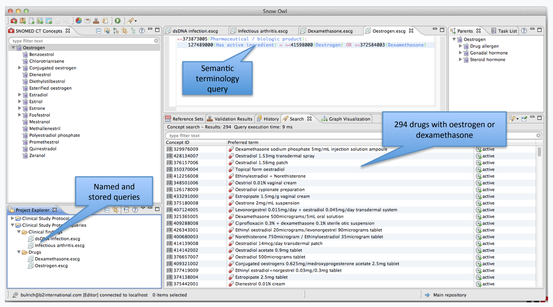
Inclusion criteria: High arthritis risk and suitable age

Inclusion 1: Exposure to certain DNA virus families increases the risk of arthritis. Instead of specifying all the possible relevant diseases (e.g. viral pneumonia, human papilloma virus), simply specify all diseases CAUSED BY a non-enveloped dsDNA virus.
Inclusion 2: Patient is less than 32 years old.

Exclusion criteria: Cause and Active Ingredient

Exclusion 1:
We need to exclude some patients that have had arthritis or a joint inflammation, but only where the condition was CAUSED BY an organism. (e.g. virus or bacteria).

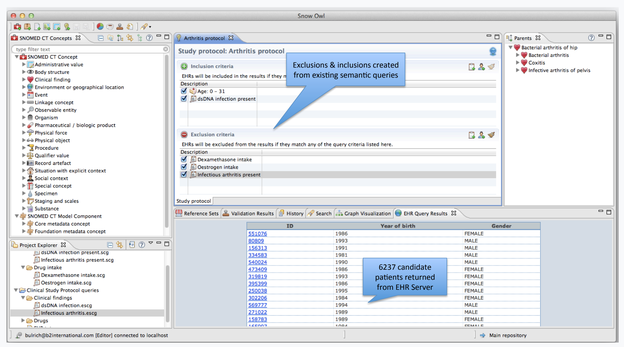




Exclusion 2:
Some substances have a positive effect on arthritis, so patients that have taken drugs with an ACTIVE INGREDIENT of dexamethasone or estrogen should be excluded. Don't identify all possible drugs and trade names, simply exclude the patient based on the active ingredients of the drugs they have taken.

Results

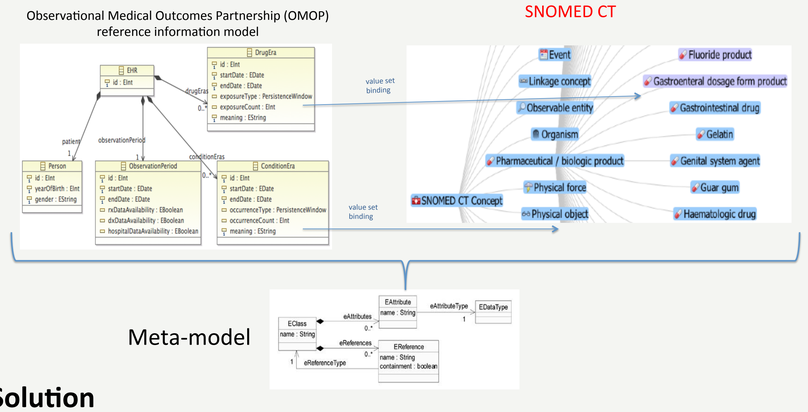
Saves time: Don't need to specify every infection virus by virus.
Improves accuracy: All drugs containing any dexamethasone or estrogen are automatically included.
Future-proof: When the criteria are reused in future trials, newly discovered viruses and released drugs are automatically included.



Ad hoc search result: 1,200,000 EHRs semantically searched in 4.4 seconds identifying 6,237 candidate patients.

Note: Hardware was a single 2007 vintage Xeon 2.0 GHz 8-core server

Questions? Drop by the B2i Healthcare booth in the exhibition area



Solution

- Information model combined with SNOMED CT that encapsulates semantic information
- Meaning not constrained to a fixed set of concepts
- Query language to leverage stored semantic meaning
- Real-time queries can be run on operational stores without requiring extraction and aggregation to analytical data stores

Scalable

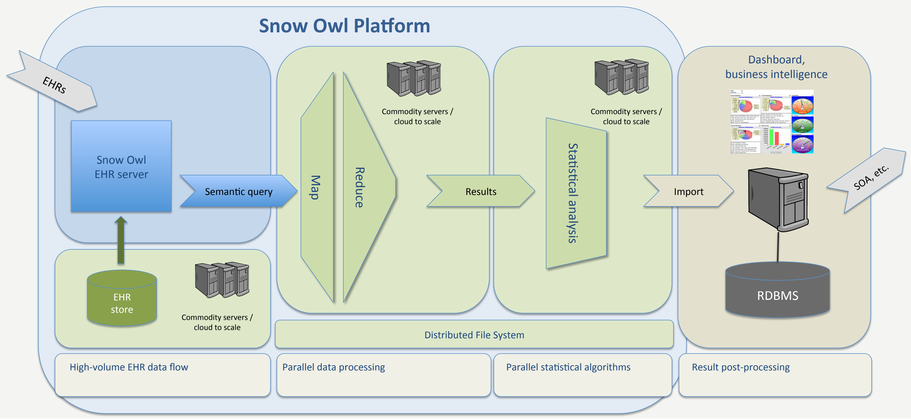
- Supports national to global numbers of electronic health records
- Platform supports massive parallel processing either on local networks or in the cloud

Flexible

- Works with different information model standards, including HL7, openEHR, EN 13606, LRA, LIM, UML, MOF as it is agnostic to the internal structure of the EHR data
- Configure instead of code; designed to support changes to information models and terminology

Tooling support

- Terminology authoring, mapping, and subset creation, information model authoring, EHR mapping
- Support for report design and analysis



Implications

Unlocks Semantics: Meaningful queries allow semantic aggregation instead of simple hierarchical grouping—avoiding combinatorial explosion later in the data analysis stage. Allows queries and analytics for meaningful aggregates such as OMOP's Health Outcomes of Interest (e.g. acute renal failure).

Model versus Code: Flexible, modeling paradigm allows reconfiguration as opposed to coding effort.

Scalability: Massively parallel platform allows real-time meaningful queries and analytics.

References

- "Big data: The next frontier for innovation, competition, and productivity," McKinsey, May 2011
- "Secondary uses of Electronic Health Record data in Life Sciences," Deloitte, June 2010
- "Transforming healthcare through secondary use of health data," PriceWaterhouseCoopers, October 2009