i2b2 and tranSMART 2017:

first experiences of interoperability

Mauro Bucalo
Matteo Gabetta
BIOMERIS

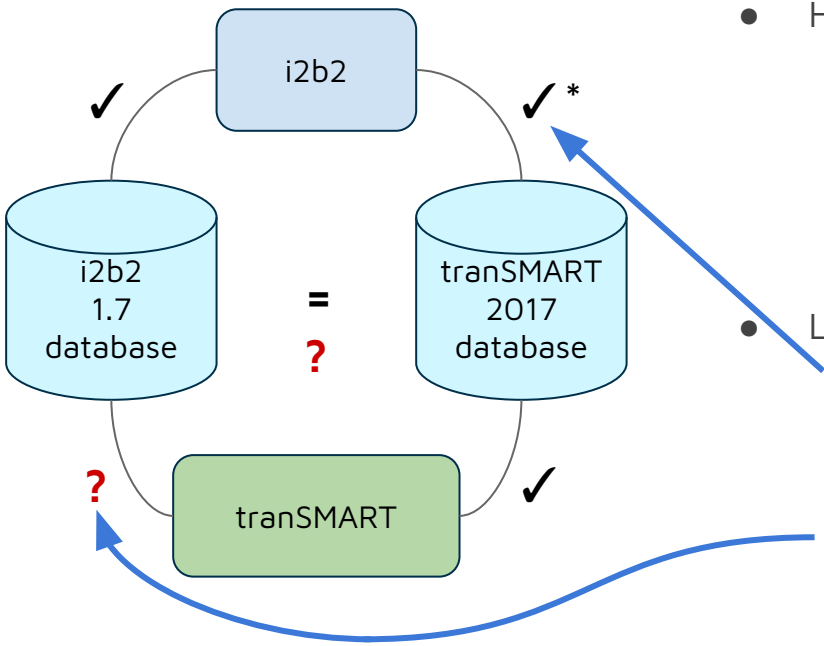
Ward Weistra
Jan Kanis
Jarno Van Erp
THE HYVE

OCT 6, 2017 - PARIS

Basic Idea

allow **i2b2** and **tranSMART**
to work on the same database
with opportune modifications

The idea

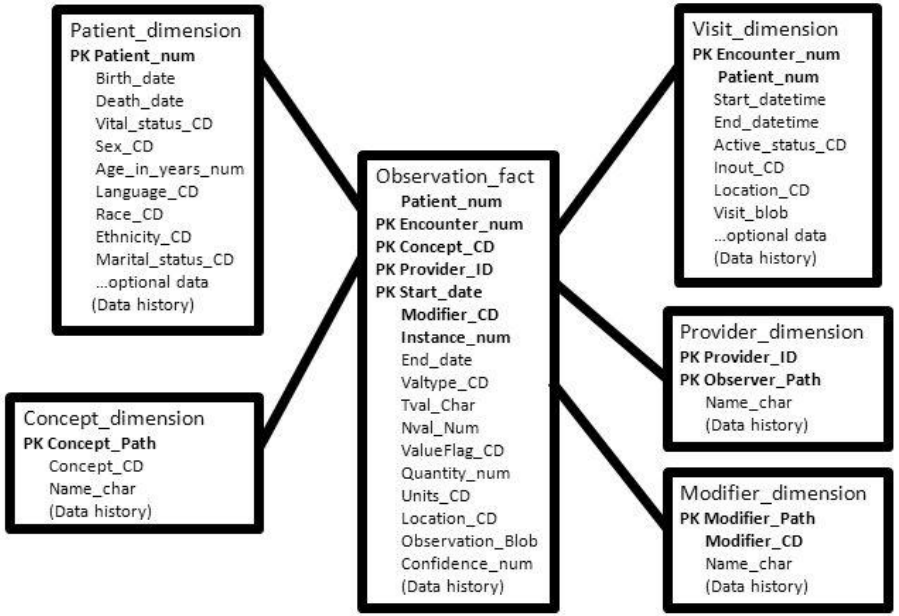


- History
  - **tranSMART** was built on top of **i2b2** with the same database, but diverted over time
  - **tranSMART 17.1 project** restored i2b2 star schema
- Let's test how similar they really are!
  - In Harvard (June) we showed that **i2b2** runs (with minor changes) on **tranSMART 2017 database**
  - Next up: Does **tranSMART 2017** run on an **i2b2 database**?

i2b2 1.7 database

The i2b2 clinical data model is pictured below.

- The ***visits***, ***dates*** on observations and ***modifiers*** allow for modelling time series and samples.
- The visit dimension has ***patient_num*** in its primary key. Hence these visits can only be linked to one patient.

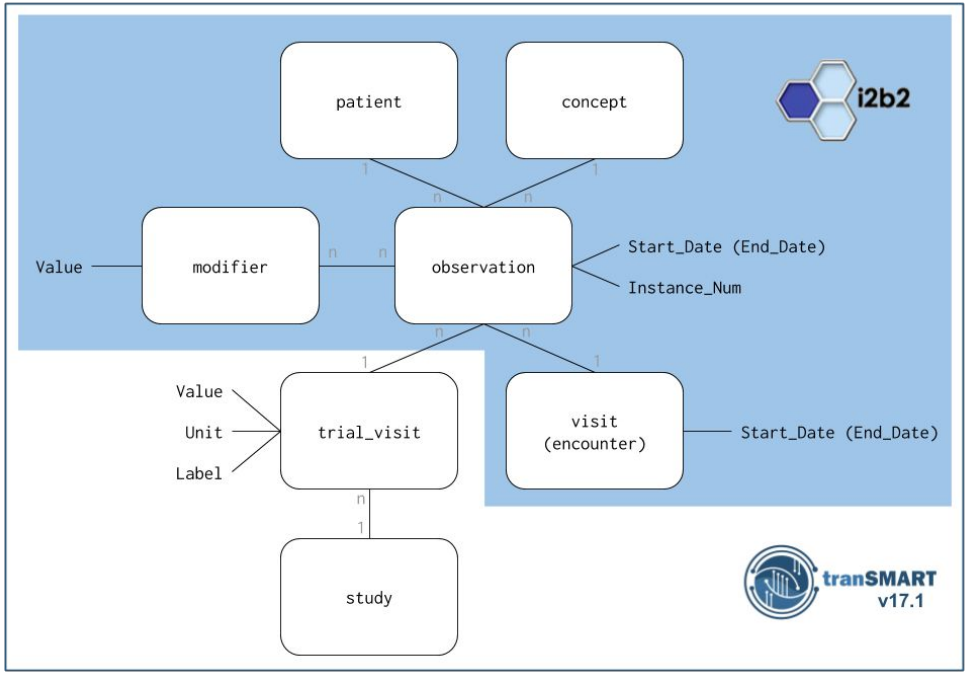


i2b2 Design Document; Data Repository (CRC) Cell; Partners Healthcare System, 1.7.1
-image created by Jim Campbell, Dan Connolly, PCORI/GPC Standard Data Model

tranSMART 17.1 database

The effective clinical data model in tranSMART 17.1 data model is pictured on the right.

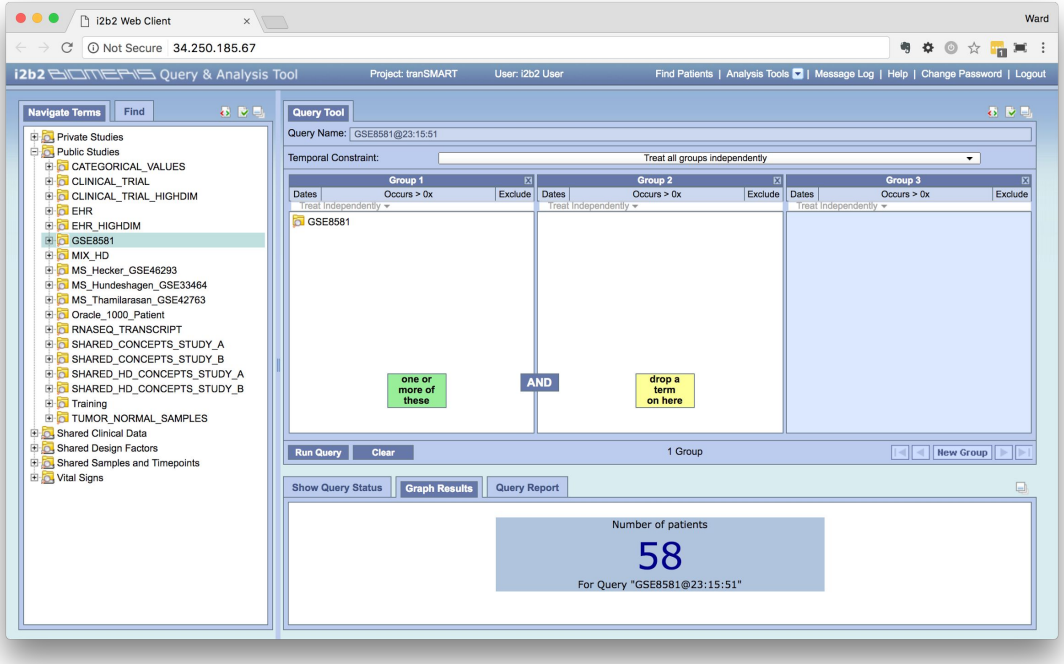
- Reinstatement of the **full i2b2 star schema dimensions** allows for modelling time series and samples.
- The link with **studies** is made explicit. It is made on observation level instead of concept level, to allow for cross-study concepts.
- **Trial visits** are introduced to model visits shared among patients (Baseline, Week 1, ...).



tranSMART 17.1 project: Removed differences in tranSMART compared to i2b2

- Dimensions
  - Restored the usage the i2b2 dimension columns and tables:
    start_date, **end_date** on observations, **visit/encounter**, **modifier_cd**, **instance_num**.
  - Restored **observation_fact** primary key and nullability by re-adding **encounter_num**, **start_date** and **instance_num**.
- High dimensional data linking
  - Deprecating **observation_fact.sample_cd** column. Link to high dimensional data is now made with a modifier instead of a tranSMART-specific column.
- Study linking
  - Deprecating **observation_fact.sourcesystem_cd** for storing the tranSMART study ID.

Flashback to June: i2b2 on tranSMART 2017 database

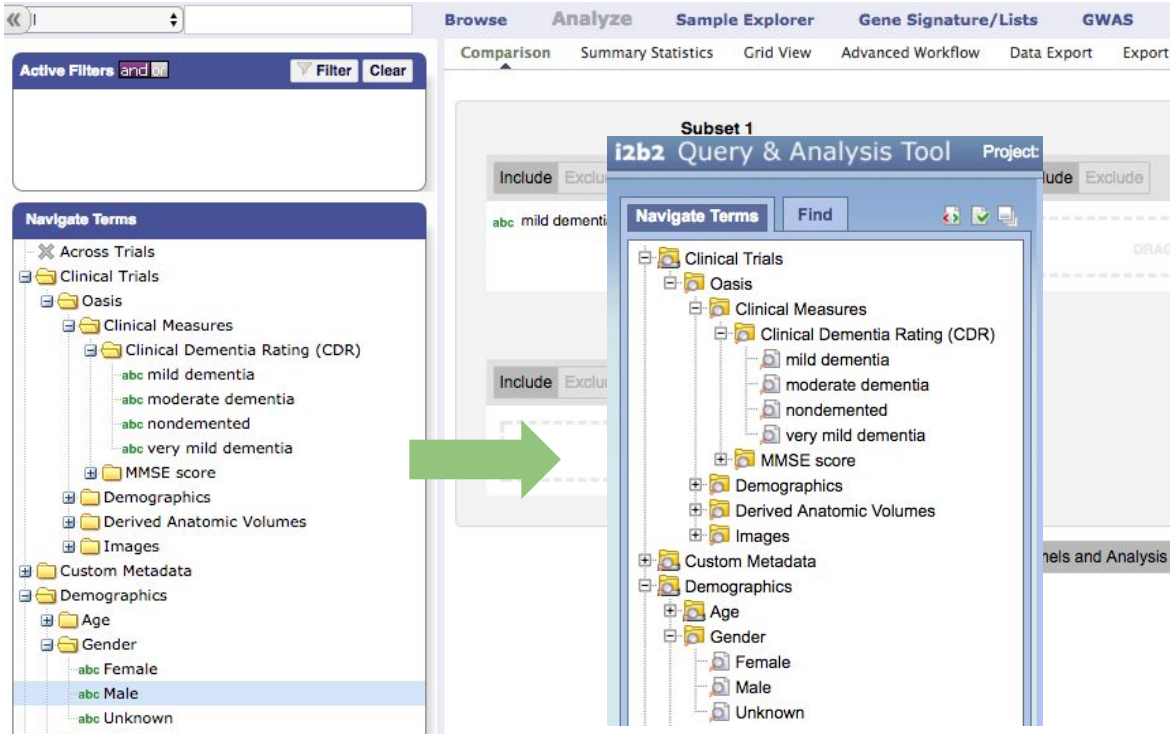


Works on a tranSMART 2017 database!
(sequence names need to be aligned)

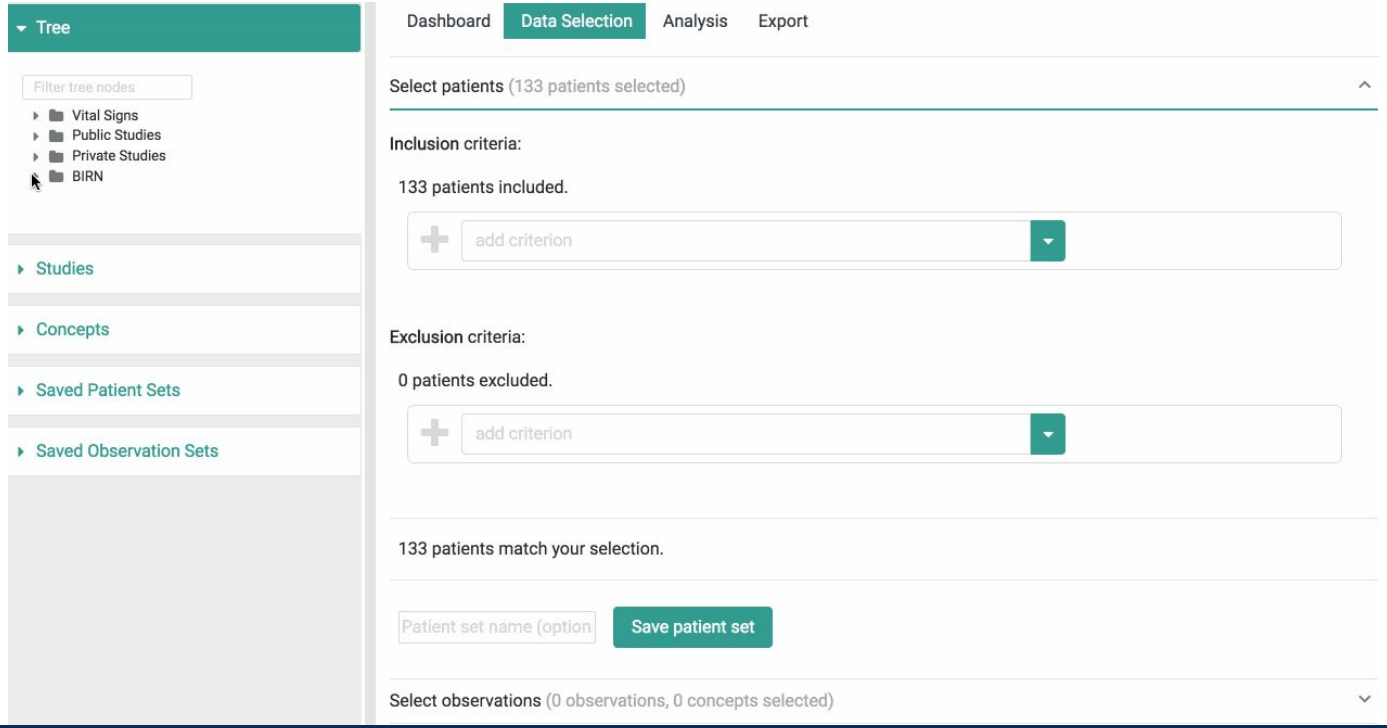
The steps make tranSMART run on i2b2 database

1. Take a clean tranSMART database and drop i2b2 tables
2. Add i2b2 **demodata** and **metadata** tables from i2b2 database:
   a. Change table format
      - Convert type **integer** to **numeric**
      - Add **trial_visit_num** to **observation_fact**
   b. Adjust metadata for all studies
      - Add **birn** entries to **i2b2_secure**
      - Add study OASIS to study table in **i2b2demodata** and a trial visit for this study
   c. Change visual attribute in **i2b2metadata**
      - tranSMART visually differentiates between categorical, numerical and high dimensional nodes (LA to LAN in numeric and LA to LAC in categorical)
   d. Satisfy tranSMART security model
      - Add one trial visit to each observation, and one study to the trial visit
   e. Adjust i2b2 data for tranSMART limitations
      - Convert date observations to text
      - Convert less than and greater than to exact number

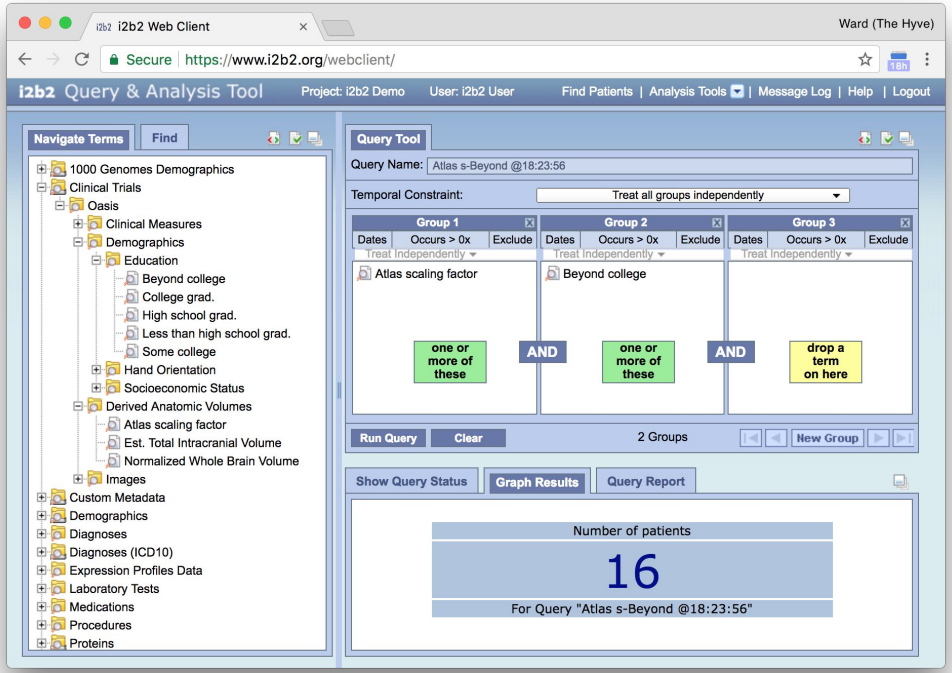
transmartApp UI on i2b2 database



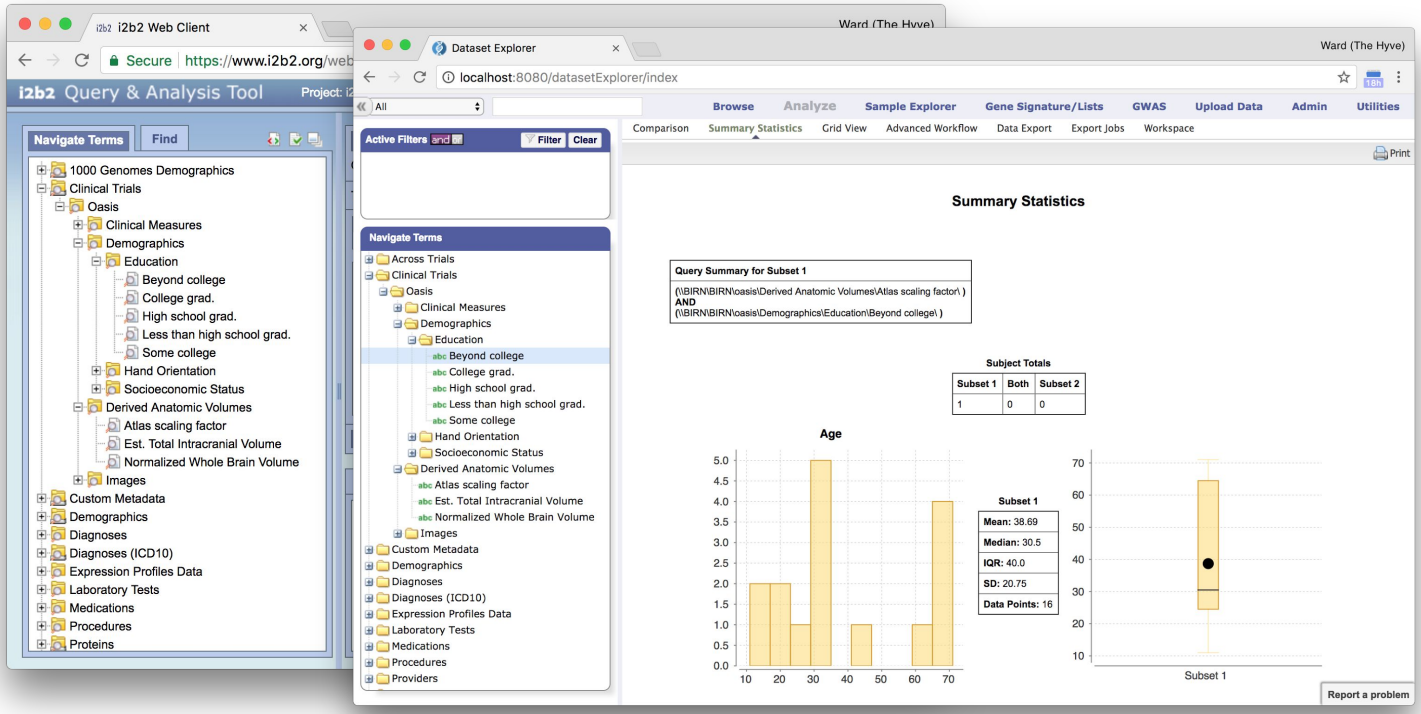
Glowing Bear cohort builder on i2b2 1.7 database



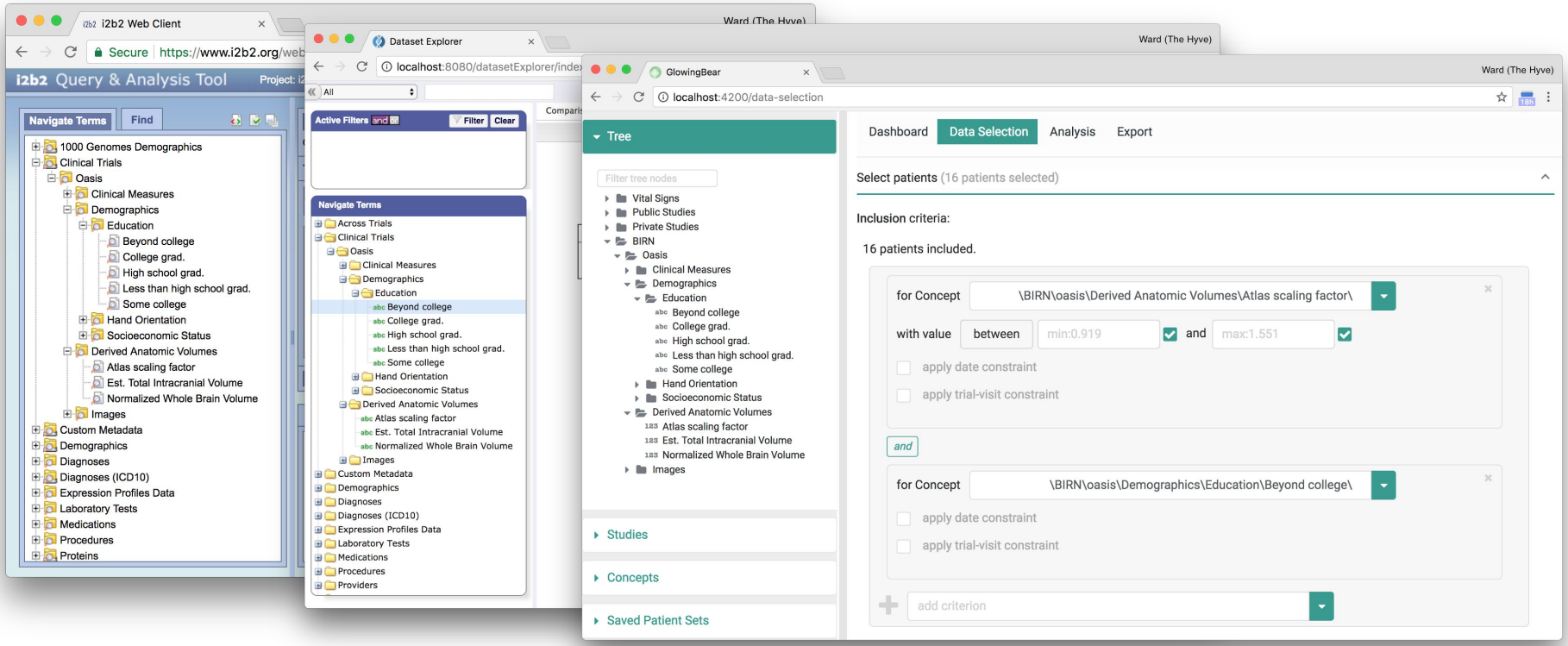
The same query on all



The same query on all



The same query on all



BIOMERIS

Other remaining differences

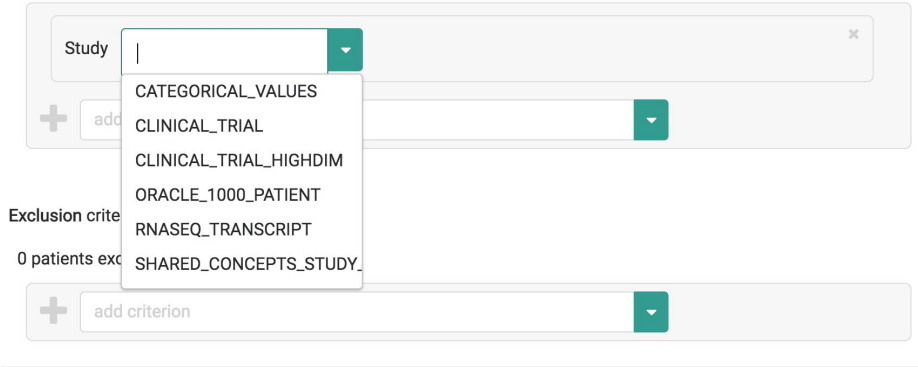
- Study tree nodes
    - tranSMART Ontology tree has study dimension tree nodes. Could work in i2b2 if corresponding tables are added.
- Sequence names
    - Sequence names in i2b2 different between Oracle and Postgres. tranSMART currently uses Oracle sequence names (as presented at Harvard meeting)

Projects in i2b2

- In i2b2 a user can access to different projects
- The data behind each project are stored in different schema or in different database (we use materialized views to simplify ETL procedures)
- User permission are granted using the PM cell

Studies and trial visits in tranSMART

- Both are essential for modelling clinical trials and tranSMART security
- The data behind each study are stored in the same database
- User permission are granted on a observation level using the added **trial_visit** column
    - Corresponding **study** and **trial_visit_dimensions** tables were added to **i2b2demodata**.
    - Studies are linked to observations via **trial visits**, such as to limit the number of additional columns in **observation_fact**.



Going forward to a shared database

- The **i2b2 PMC** is already considering:
  - Inclusion of the **trial_visit_num** column in the **i2b2demodata.observation_fact** table.
  - Inclusion of the study and **trial_visit_dimensions** tables to **i2b2demodata**.
- The **tranSMART PMC** is already considering:
  - Migration of the tables related to study dimensions and large scale file storage to non-i2b2 schema.
- The Foundation is planning a **database working group** to think further about alignment (and optimalization)
  - We hope this working group will consider our lessons learned and bring the community to **one shared i2b2 tranSMART database**!

Credits

- The Hyve's Team
  - Ward Weistra
  - Jan Kanis
  - Jarno Van Erp
- Biomeris Team
  - Mauro Bucalo
  - Matteo Gabetta