

# i2b2 and tranSMART 2017: first experiences of interoperability

Mauro Bucalo  
Matteo Gabetta  
BIOMERIS

Ward Weistra  
Jan Kanis  
Jarno Van Erp  
THE HYVE

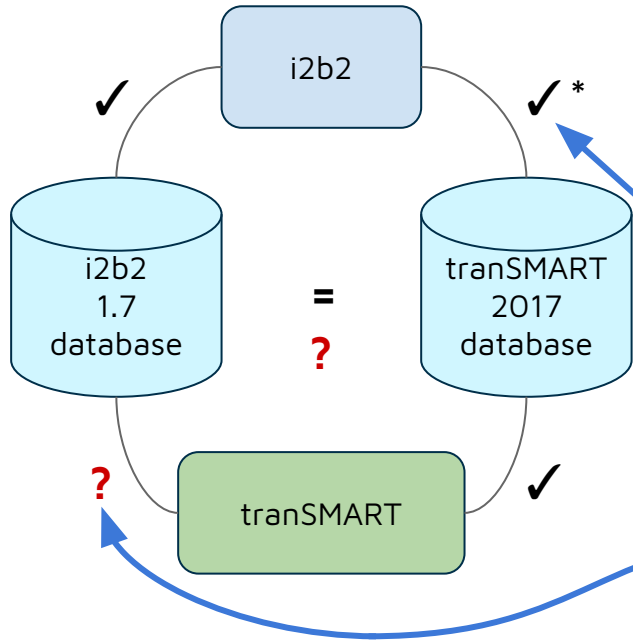
OCT 6, 2017 - PARIS



## Basic Idea

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

# The idea



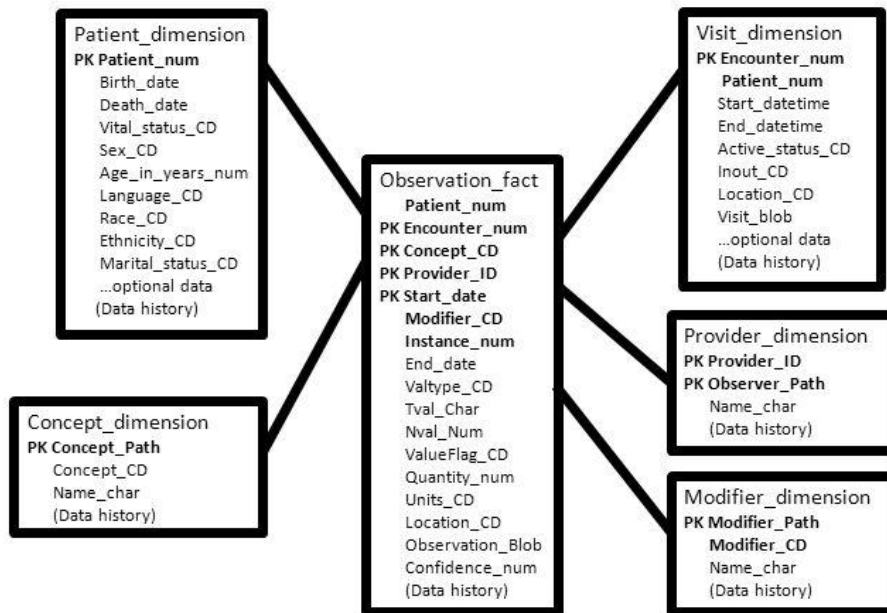
- History
  - **tranSMART** was built on top of **i2b2** with the same database, but diverted over time
  - **tranSMART 17.1 project** restored i2b2 star schema
- Let's test how similar they really are!
  - In Harvard (June) we showed that **i2b2** runs (with minor changes) on **tranSMART 2017 database**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

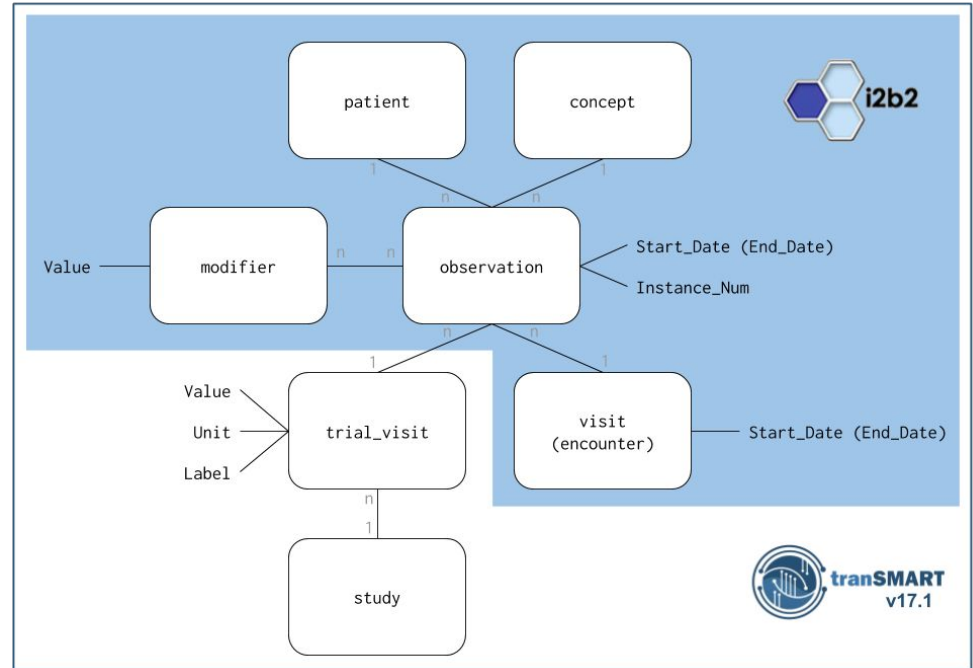


i2b2 Design Document; Data Repository (CRC) Cell; Partners Healthcare System, 1.7.1  
-image created by Jim Campbell, Dan Connolly, PCORI/GPC Standard Data Model

# 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

The screenshot displays the i2b2 BIOMERIS Query & Analysis Tool interface. The browser window title is "i2b2 Web Client". The address bar shows "Not Secure 34.250.185.67". The tool's header includes "Project: tranSMART", "User: i2b2 User", and navigation links for "Find Patients", "Analysis Tools", "Message Log", "Help", "Change Password", and "Logout".

The interface is divided into several sections:

- Navigate Terms:** A tree view on the left lists various study categories such as "Private Studies", "Public Studies", "CATEGORICAL\_VALUES", "CLINICAL\_TRIAL", "CLINICAL\_TRIAL\_HIGHDIM", "EHR", "EHR\_HIGHDIM", "GSE8581", "MIX\_HD", "MS\_Hecker\_GSE46293", "MS\_Hundeshaugen\_GSE33464", "MS\_Thamilarasan\_GSE42763", "Oracle\_1000\_Patient", "RNASEQ\_TRANSCRIPT", "SHARED\_CONCEPTS\_STUDY\_A", "SHARED\_CONCEPTS\_STUDY\_B", "SHARED\_HD\_CONCEPTS\_STUDY\_A", "SHARED\_HD\_CONCEPTS\_STUDY\_B", "Training", "TUMOR\_NORMAL\_SAMPLES", "Shared Clinical Data", "Shared Design Factors", "Shared Samples and Timepoints", and "Vital Signs".
- Query Tool:** The main workspace shows a query named "GSE8581@23:15:51". The temporal constraint is set to "Treat all groups independently". The query is structured into three groups, each with a "Dates" field and a "Treat Independently" dropdown. The "GSE8581" term is placed in the first group's "Dates" field. Below the query groups, there are instructions: "one or more of these" (pointing to the first group), "AND" (between groups), and "drop a term on here" (pointing to the second group).
- Run Query:** A button labeled "Run Query" is visible, along with a "Clear" button and a "New Group" button. The status "1 Group" is displayed.
- Results:** The "Graph Results" tab is active, showing a summary of the query results. It indicates "Number of patients" as **58** for the query "GSE8581@23:15:51".

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

The image displays the transmartApp UI interface for the i2b2 database. The main window features a top navigation bar with tabs: **Browse**, **Analyze**, **Sample Explorer**, **Gene Signature/Lists**, and **GWAS**. Below this is a secondary navigation bar with options: **Comparison**, **Summary Statistics**, **Grid View**, **Advanced Workflow**, **Data Export**, and **Export**.

On the left side, there is an **Active Filters** section with a dropdown menu set to **and** and buttons for **Filter** and **Clear**. Below it is the **Navigate Terms** panel, which shows a hierarchical tree view of the database structure. The tree is expanded to show the following structure:

- Across Trials
- Clinical Trials
  - Oasis
    - Clinical Measures
      - Clinical Dementia Rating (CDR)
        - mild dementia
        - moderate dementia
        - nondemented
        - very mild dementia
      - MMSE score
    - Demographics
    - Derived Anatomic Volumes
    - Images
  - Custom Metadata
  - Demographics
    - Age
    - Gender
      - Female
      - Male
      - Unknown

A green arrow points from the **Male** term in the **Navigate Terms** panel to the **i2b2 Query & Analysis Tool** overlay. This tool window, titled **Subset 1**, contains its own **Navigate Terms** panel and a **Find** input field. The tree view in the tool is identical to the one in the main panel, but with the **Male** term selected. The tool also includes **Include** and **Exclude** buttons for filtering terms.

# Glowing Bear cohort builder on i2b2 1.7 database

The screenshot displays the 'Data Selection' tab of the Glowing Bear cohort builder. The interface is divided into a left sidebar and a main content area. The sidebar, titled 'Tree', contains a search box for 'Filter tree nodes' and a list of folders: 'Vital Signs', 'Public Studies', 'Private Studies', and 'BIRN'. Below this are sections for 'Studies', 'Concepts', 'Saved Patient Sets', and 'Saved Observation Sets'. The main content area has a navigation bar with 'Dashboard', 'Data Selection' (active), 'Analysis', and 'Export'. It shows 'Select patients (133 patients selected)' and 'Inclusion criteria: 133 patients included.' with an 'add criterion' button. Below that, 'Exclusion criteria: 0 patients excluded.' also has an 'add criterion' button. A summary states '133 patients match your selection.' and includes a 'Patient set name (option)' input field and a 'Save patient set' button. At the bottom, it shows 'Select observations (0 observations, 0 concepts selected)'.



# The same query on all

1000 Genomes Demographics

- Clinical Trials
- Oasis
  - Clinical Measures
  - Demographics
    - Education
      - Beyond college
      - College grad.
      - High school grad.
      - Less than high school grad.
      - Some college
    - Hand Orientation
    - Socioeconomic Status
  - Derived Anatomic Volumes
    - Atlas scaling factor
    - Est. Total Intracranial Volume
    - Normalized Whole Brain Volume
  - Images
  - Custom Metadata
  - Demographics
  - Diagnoses
  - Diagnoses (ICD10)
  - Expression Profiles Data
  - Laboratory Tests
  - Medications
  - Procedures
  - Proteins

Query Name: Atlas s-Beyond @18:23:56

Temporal Constraint: Treat all groups independently

Group 1	Group 2	Group 3
Dates	Dates	Dates
Occurs > 0x	Occurs > 0x	Occurs > 0x
Exclude	Exclude	Exclude
Treat Independently	Treat Independently	Treat Independently
Atlas scaling factor	Beyond college	
one or more of these	AND	one or more of these
	AND	drop a term on here

Run Query Clear 2 Groups New Group

Show Query Status Graph Results Query Report

Number of patients

**16**

For Query "Atlas s-Beyond @18:23:56"

# The same query on all

The image displays two overlapping web browser windows. The left window is the 'i2b2 Query & Analysis Tool' showing a hierarchical tree of categories like '1000 Genomes Demographics', 'Clinical Trials', 'Oasis', 'Demographics', 'Education', etc. The right window is 'Dataset Explorer' showing the same query results. It includes a 'Query Summary for Subset 1' with a complex query string, a 'Subject Totals' table, a bar chart for 'Age', and a box plot for 'Subset 1'.

**Query Summary for Subset 1**

```
(@BIRN|BIRN|oasis|Derived Anatomic Volumes|Atlas scaling factor) AND (@BIRN|BIRN|oasis|Demographics|Education|Beyond college)
```

**Subject Totals**

Subset 1	Both	Subset 2
1	0	0

**Age**

Age	Frequency
10	2.0
15	2.0
20	1.0
25	5.0
30	1.0
35	1.0
40	1.0
45	1.0
50	1.0
55	1.0
60	1.0
65	1.0
70	1.0

**Subset 1**

Mean: 38.69
Median: 30.5
IQR: 40.0
SD: 20.75
Data Points: 16

**Box Plot**

# The same query on all

The image displays three overlapping browser windows illustrating the same query across different stages of a data analysis tool:

- Left Window (i2b2 Web Client):** Shows the 'i2b2 Query & Analysis Tool' interface. The 'Navigate Terms' panel is expanded to show a hierarchical tree structure. The 'Education' category is selected, with sub-items like 'Beyond college', 'College grad.', 'High school grad.', 'Less than high school grad.', and 'Some college' visible.
- Middle Window (Dataset Explorer):** Shows the 'Dataset Explorer' interface. The 'Active Filters' section is empty. The 'Navigate Terms' panel is expanded to show the same hierarchical tree structure as the left window, with 'Beyond college' selected.
- Right Window (Data Selection):** Shows the 'Data Selection' interface. The 'Tree' panel is expanded to show the same hierarchical tree structure as the left window, with 'Beyond college' selected. The 'Inclusion criteria' section shows two criteria: 'for Concept \BIRN\oasis\Derived Anatomic Volumes\Atlas scaling factor\' with a value range of 'between min:0.919 and max:1.551', and 'for Concept \BIRN\oasis\Demographics\Education\Beyond college\'.



## 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**.

The screenshot shows a web interface for selecting a study. A dropdown menu is open, displaying a list of study identifiers. The interface includes a search bar, a plus sign, and a dropdown arrow. Below the dropdown, there is a label 'Exclusion crite' and a text '0 patients exc'. At the bottom, there is another plus sign and a label 'add criterion'.

Study Identifier
CATEGORICAL_VALUES
CLINICAL_TRIAL
CLINICAL_TRIAL_HIGHDIM
ORACLE_1000_PATIENT
RNASEQ_TRANSCRIPT
SHARED_CONCEPTS_STUDY_





# 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 optimization)
  - 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