Weill Cornell Medical College

Architecture for Research Computing in Health (ARCH)

Using i2b2 to Find Patient Cohorts of Interest

To access i2b2, please visit https://i2b2.med.cornell.edu/webclient/ and login with your CWID and password. To request access, please contact i2b2-support@med.cornell.edu.

What data are available in i2b2?

As of April 2015, investigators can query de-identified data from Epic, the outpatient EHR of the Weill Cornell Physician Organization, for 2.6 million unique patients. Available data from Epic include:

- Demographics
  - Age (as of today; aka last date of data loaded into i2b2)
  - Age (at time of encounter)
  - Ethnicity
  - Language
  - Marital Status
  - Religion
  - Sex
  - State
  - Vitals Status
- Encounters: 36 million records
  - Recorded in Epic as individual clinic locations
  - Displayed in i2b2 using folders and subfolders based on academic departments and divisions
- Diagnoses: 22 million records
  - Recorded in Epic as encounter diagnoses using ICD-9 codes
  - Displayed in i2b2 using folders and subfolders based on the ICD-9 hierarchy
- Procedures: 96 millions records
  - Recorded in Epic as orders using Current Procedural Terminology (CPT)
  - Displayed in i2b2 using folders and subfolders based on the CPT hierarchy

What future data will be available to i2b2?

- Epic data
  - Medications – ***summer 2015***
  - Labs – ***summer 2015***
  - Social History

- Family History
- Other sources
  - Biospecimen inventory
  - Tumor Registry
  - Additional items from Allscripts (aka Eclipsys), Epic, and other systems by request; please contact i2b2-support@med.cornell.edu with requests

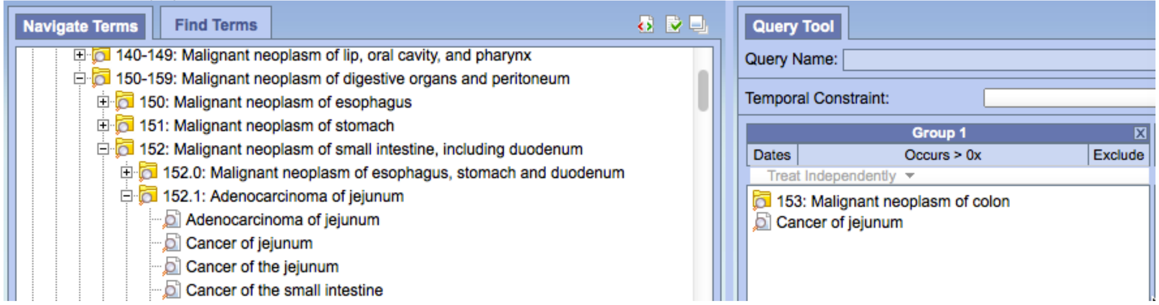
How do I create a query?

1. To optimize query run times, keep query elements as specific as possible by dragging **Folders** and **Concepts** from **Containers** into query groups.
   Containers: the highest level of hierarchy—Demographics, Diagnoses, Encounters, and Procedures—that contain folders & concepts and are expandable
   Folders: contains concepts (can contain other folders) and are expandable (e.g. Malignant neoplasms, Malignant neoplasms of digestive organs)
   Concept: the lowest level of hierarchy and is not expandable (e.g. Cancer of jejunum)
   - Drag **folder** and **concept** items into query groups

     folder (📁) **concept** (🗋)
   - Do NOT drag **container** items (denoted by a gray bar) into query groups

     container (📁)

GOOD QUERY – Do query folder "153: Malignant neoplasm of colon" and concept "Cancer of jejunum"



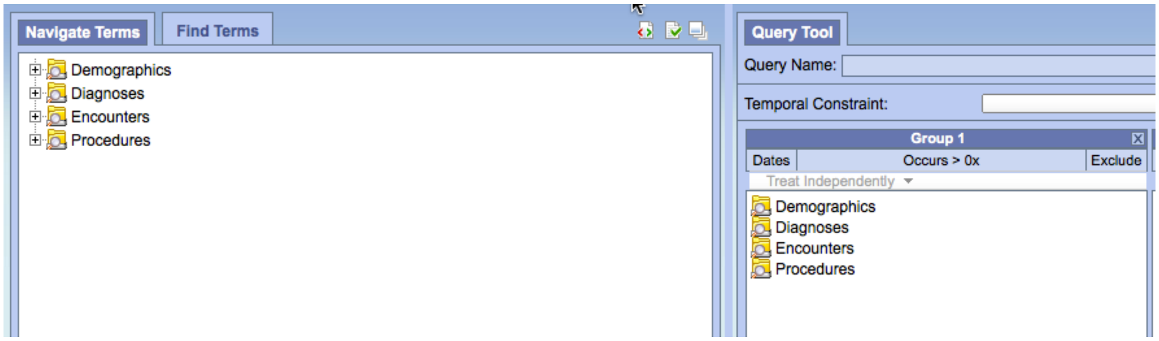
- 140-149: Malignant neoplasm of lip, oral cavity, and pharynx
- 150-159: Malignant neoplasm of digestive organs and peritoneum
  - 150: Malignant neoplasm of esophagus
  - 151: Malignant neoplasm of stomach
  - 152: Malignant neoplasm of small intestine, including duodenum
    - 152.0: Malignant neoplasm of esophagus, stomach and duodenum
    - 152.1: Adenocarcinoma of jejunum
      - Adenocarcinoma of jejunum
      - Cancer of jejunum
      - Cancer of the jejunum
      - Cancer of the small intestine

Query Name:

Temporal Constraint:

Group 1: 153: Malignant neoplasm of colon; Cancer of jejunum

BAD QUERY – Do NOT query containers "Demographics", "Diagnoses", "Encounters, "Procedures":



- Demographics
- Diagnoses
- Encounters
- Procedures

Query Name:

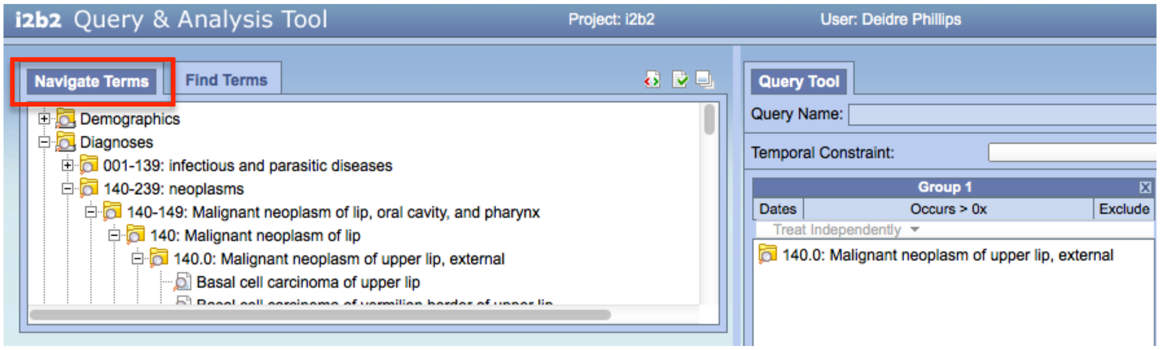
Temporal Constraint:

Group 1: Demographics; Diagnoses; Encounters; Procedures

2. There are two ways to search for terms:

   a. NAVIGATE TERMS (manual search)

      - For example, if you wish to query patient populations having lip cancer – start by expanding the **Diagnoses** container to view the *140-149: Malignant neoplasm of lip, oral cavity, and pharynx* folder, in order to narrow down the diagnosis selection to the *140: Malignant neoplasm of lip* folder:

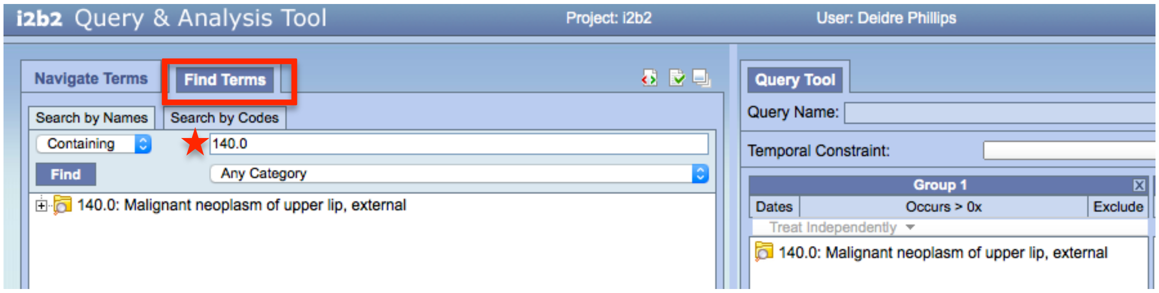


- Demographics
- Diagnoses
  - 001-139: infectious and parasitic diseases
  - 140-239: neoplasms
    - 140-149: Malignant neoplasm of lip, oral cavity, and pharynx
      - 140: Malignant neoplasm of lip
        - 140.0: Malignant neoplasm of upper lip, external
          - Basal cell carcinoma of upper lip
          - Basal cell carcinoma of vermilion border of upper lip

Group 1: 140.0: Malignant neoplasm of upper lip, external

   b. FIND TERMS (keyword search)

      - From the example above – search using the ICD-9 diagnosis code of interest:



Containing: 140.0

Any Category

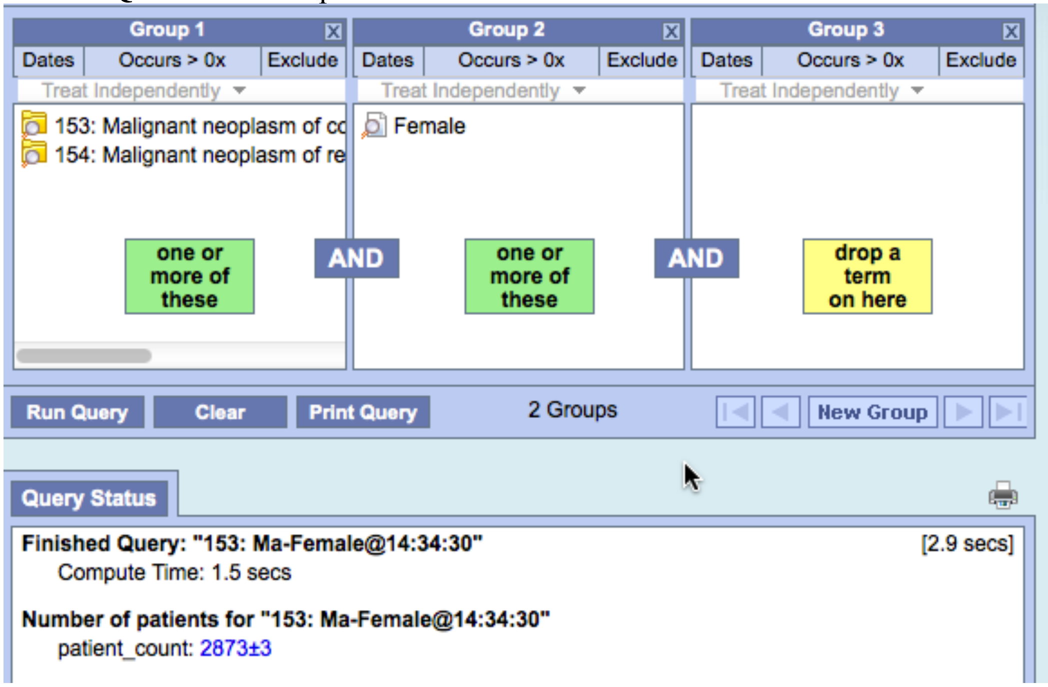
- 140.0: Malignant neoplasm of upper lip, external

Group 1: 140.0: Malignant neoplasm of upper lip, external

3. If a query times out when query elements are specified in a certain order, try reorganizing the groups and re-running the query.

4. To optimize query run times, place the query element that would logically return the **fewest** number of patients in Group 1.
    - Ex: “How many female patients had a diagnosis of colon cancer?”
        - Query elements:
            - Group 1→Fewest patient count = colon cancer
            - Group 2→Highest patient count = female

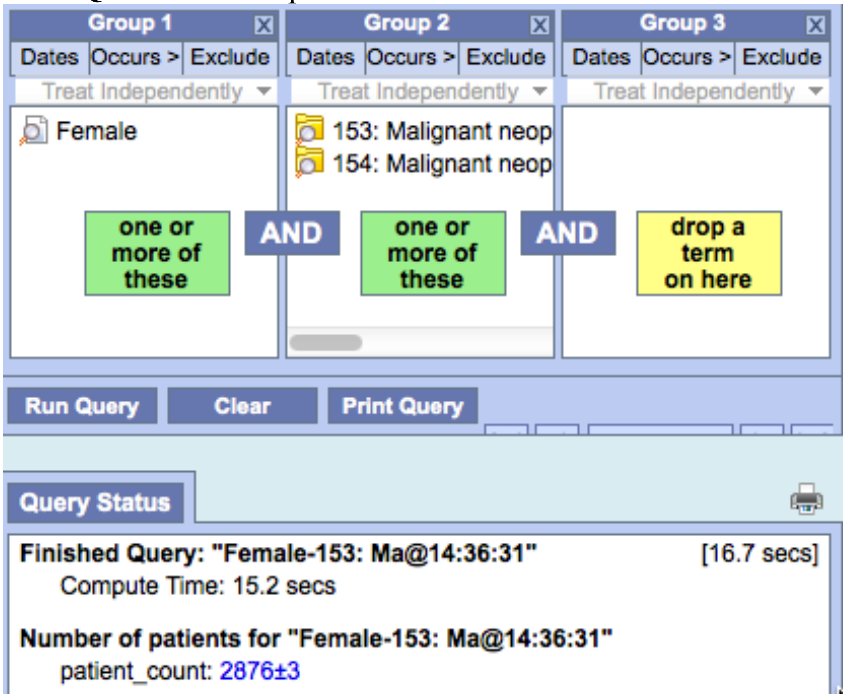
GOOD QUERY – Compute Time: 1.5 sec



Finished Query: "153: Ma-Female@14:34:30" [2.9 secs]
Compute Time: 1.5 secs

Number of patients for "153: Ma-Female@14:34:30"
patient_count: 2873±3

BAD QUERY – Compute time: 15.2 sec



Finished Query: "Female-153: Ma@14:36:31" [16.7 secs]
Compute Time: 15.2 secs

Number of patients for "Female-153: Ma@14:36:31"
patient_count: 2876±3

How does i2b2 store data and protect patient privacy?

i2b2 contains data that has been de-identified in accordance with the HIPAA Privacy Rule's Safe Harbor definition (http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html). The following eighteen patient identifiers are NOT available in i2b2:

- Names
- All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code and their equivalent geocodes, except for the initial three digits of a ZIP code if, according to the current publicly available data from the Bureau of the Census, (1) the geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people, and (2) the initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people are changed to 000
- All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date and date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of "age 90 or older"
- Telephone numbers
- Fax numbers
- E-mail addresses
- Social Security numbers
- Medical record numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers, including license plate numbers
- Device identifiers and serial numbers
- Web Universal Resource Locators (URLs)
- Internet Protocol (IP) addresses
- Biometric identifiers, including finger and voice prints
- Full-face photographic images and any comparable images
- Any other unique identifying number, characteristic or code

One of the various ways i2b2 upholds regulatory standards of PHI is through the use of **date shifting**. All patient records in i2b2 have been randomly date shifted backwards by 1 to 364 days. The exact number days shifted varies from patient to patient; however, the number of days shifted remains consistent per patient. As a result, date shifting in i2b2 preserves the sequence of care events (i.e. a procedure occurred following a diagnosis) but hinders the ability to identify seasonal trends (i.e. influenza diagnosis in November 2014).

Why do I receive a slightly different count as a result of running the same query?

A random number is intentionally added or subtracted to obfuscate the results, especially for diagnoses or procedures with dates. This obfuscation is necessary to ensure patient privacy.

How many years of data are included in the i2b2 data repository?

i2b2 contains all Epic data that has been collected and stored by WCMC clinical practices since 2000. The years of data will vary by clinical practice, depending on the start date of Epic implementation.

Currently i2b2 data is up-to-date with all available Epic data through 12/31/2014. By summer of 2015, Epic data will be refreshed monthly.

How reliable is the data in the i2b2 data repository?

The quality of data in i2b2 is dependent on the data quality of data entered into Epic by a clinical practice. Practices that frequently, consistently, and accurately enter data in Epic across time will have greater data quality available in i2b2.

After performing a de-identified query, can an investigator obtain identified data?

Yes, identified and provider-specific data is available to investigators who have IRB approval. Please share approved IRB documentation describing the items of interest from i2b2 with i2b2-support@med.cornell.edu to obtain identified data.

To learn more about ARCH, please visit http://arch.med.cornell.edu
If you have any questions, please contact i2b2-support@med.cornell.edu