UBC02 Transformation Documentation

UBC02 has two transformations, one for case centric data (enrollment) and one for sample data. The input files for these transformations can be found in Box at <https://nih.app.box.com/folder/128824338275>.

The transformations map the columns from the input files to node.properties in the database. Mappings can be one-to-one (i.e., one column in the input file maps to a single node.property), or many-to-one (i.e., two or more columns in the input file are concatenated to map to a single node.property). In addition some columns map the values in the input file to a set of accepted values in the database for the given node.property.

# Case Transformation

## Case Node Mappings

|  |  |  |  |
| --- | --- | --- | --- |
| **PROPERTY** | **MAPPING** | **VALUE MAPPING FROM** | **VALUE MAPPING TO** |
| case\_id | Study Code + "-" + Patient ID |  |  |
| patient\_id | Patient ID |  |  |

## Demographic Node Mappings

|  |  |  |  |
| --- | --- | --- | --- |
| **PROPERTY** | **MAPPING** | **VALUE MAPPING FROM** | **VALUE MAPPING TO** |
| breed | Breed |  |  |
| patient\_age\_at\_enrollment | Age at Registration |  |  |
| date\_of\_birth | Date of Birth |  |  |
| sex | Sex (M,F, or U) |  |  |
| weight | Body Weight (kg) |  |  |
| neutered\_indicator | Neutered Status Indicator (Y,N, U) |  |  |

## Diagnosis Node Mappings

|  |  |  |  |
| --- | --- | --- | --- |
| **PROPERTY** | **MAPPING** | **VALUE MAPPING FROM** | **VALUE MAPPING TO** |
| disease\_term | Diagnosis |  |  |
| primary\_disease\_site | Disease Site General Location |  |  |
| stage\_of\_disease | Stage of Disease |  |  |
| date\_of\_diagnosis | Date of Diagnosis |  |  |
| best\_response | Best Tumor Response |  |  |
| treatment\_data | Treatment Outcome Available |  |  |
| follow\_up\_data | Follow-up Available |  |  |
| pathology\_report | Pathology Evaluation Detail Available |  |  |
| concurrent\_disease | Concurrent Disease |  |  |
| concurrent\_disease\_type | Concurrent Disease Type |  |  |

## Enrollment Node Mappings

|  |  |  |  |
| --- | --- | --- | --- |
| **PROPERTY** | **MAPPING** | **VALUE MAPPING FROM** | **VALUE MAPPING TO** |
| date\_of\_registration | Date of Registration |  |  |
| site\_short\_name | Study site |  |  |
| cohort\_description | Study Arm (Breed High Risk) + “;” + Cohort Description |  |  |
|  |  | High Risk Breed;Luminal | High Risk Breed; Luminal |
|  |  | High Risk Breed;Basal | High Risk Breed; Basal |
|  |  | Average Risk Breed;Luminal | Average Risk Breed; Luminal |
|  |  | Average Risk Breed;Basal | Average Risk Breed; Basal |
|  |  | Healthy Controls;Healthy Controls | Healthy Controls |

## Registration Node Mappings

NOTE: There are two types of registration nodes: 1) what the transformation calls local registration nodes (these are simply the study patient id and the study code) or biobank registration nodes (these are the “biobank” ids and the “biobank” identifier).

### Local Registration Nodes

|  |  |  |  |
| --- | --- | --- | --- |
| **PROPERTY** | **MAPPING** | **VALUE MAPPING FROM** | **VALUE MAPPING TO** |
| registration\_id | Patient ID |  |  |
| registration\_origin | Study Code |  |  |

### 

### Biobank Registration Nodes

|  |  |  |  |
| --- | --- | --- | --- |
| **PROPERTY** | **MAPPING** | **VALUE MAPPING FROM** | **VALUE MAPPING TO** |
| registration\_id | Alt Reg ID |  |  |
| registration\_origin | Alt Reg Origin |  |  |

## Additional Mappings

In addition, we include some extra columns to the transformed files. These columns are used by the data loader.

|  |  |
| --- | --- |
| **NODE.PROPERTY** | **MAPPING** |
| study.clinical\_study\_designation | Study Code |
| cohort.cohort\_id | Study Code + “-“ + cohort\_description (refer to Enrollment Node Mappings for details of this derived value) |
| study\_arm.arm\_id | Study Code + “-“ + Study Arm (Breed High Risk) |
| enrollment\_id | Study Code + "-" + Patient ID |
| diagnosis\_id | Study Code + "-" + Patient ID |

## Input Columns Not Used

Submitters sometimes include columns with data that are not going to be included in our database. We leave these columns in the input file, but the transformation ignores them. For the case input file, these columns include:

* More Specific Tumor Location
* Tumor Size (cm3)
* Age at Diagnosis
* Supplementary File
* Alternate Registration ID
* Alternate Registration Origin

# Sample Transformation

## Sample Node Mappings

|  |  |  |  |
| --- | --- | --- | --- |
| **PROPERTY** | **MAPPING** | **VALUE MAPPING FROM** | **VALUE MAPPING TO** |
| sample\_id | Study Code + “-“ + Sample ID |  |  |
| sample\_site | Sample Site |  |  |
| physical\_sample\_type | Sample Type |  |  |
| general\_sample\_pathology | General Sample Pathology |  |  |
| tumor\_sample\_origin | Tumor Sample Origin |  |  |
| summarized\_sample\_type | Sample Type + “;” + General Sample Pathology + “;” + Tumor Sample Origin |  |  |
|  |  | Tissue;Normal;Not Applicable | Normal Tissue |
|  |  | Tissue;Malignant;Primary | Primary Malignant Tumor Tissue |
| specific\_sample\_pathology | Specific Sample Pathology |  |  |
| date\_of\_sample\_collection | Sample Collection Date |  |  |
| sample\_chronology | Sample Acquisition Timing |  |  |
| necropsy\_sample | Necropsy (Sample Collected at Necropsy) |  |  |
| percentage\_tumor | Percent Tumor |  |  |
| tumor\_grade | Grade (1-4 With 4 Being Highest Grade) |  |  |
| sample\_preservation | Sample Preservation Method |  |  |

## Additional Mappings

In addition, we include some extra columns to the transformed files. These columns are used by the data loader.

|  |  |
| --- | --- |
| **NODE.PROPERTY** | **MAPPING** |
| case.case\_id | Study Code + "-" + Patient ID |

## Input Columns Not Used

Submitters sometimes include columns with data that are not going to be included in our database. We leave these columns in the input file, but the transformation ignores them. For the sample input file, these columns include:

* Drugs At Sample Aquisition