1. **Background**

One of the primary sources of relatively high quality cancer data is Cancer Registries. NAACCR (<https://www.naaccr.org/data-standards-data-dictionary/>) is a data standard used to code data in the US Cancer Registries. To support ETL of Cancer Registry data into OMOP CDM, NAACCR data dictionary should be represented in the OMOP Vocabulary.

Moreover, in the absence of established data models and semantic annotations in cancer, NAACCR is arguably the best existing data dictionary that covers majority of cancer types and includes critical diagnostic features and high level treatment classification used in cancer epidemiology. We recommend using NAACCR data dictionary as a foundational OMOP vocabulary to support representation of diagnostic modifiers and high level treatment classifications.

As one of our future steps in the development of cancer specific vocabulary in OMOP, we will map NAACCR to the evolving or established proper ontology standards like SNOMED and LOINC.

1. **Overall Approach**
2. **Status of NAACCR concepts in the OMOP Vocabulary**

NAACCR data dictionary is not a formal ontology, nor is it currently mapped to any standard formal ontology. Long term, we are planning to map NAACCR data dictionary to a standard like LOINC and SNOMED. Before this is done, NAACCR concepts will be standard concepts.

1. **Ingestion Scope**

Data elements from the following NAACCR sections will be ingested:

* 'Cancer Identification'
* 'Stage/Prognostic Factors'
* 'Treatment-1st Course'

1. **Attribute-Value Pairs**

Since diagnostic modifiers will be stored as attribute-value pairs, the corresponding concept relationships will be persisted in the vocabulary.

1. **Diagnosis-specific and generic NAACCR concepts**

NAACCR includes both, diagnosis specific (e.g. Gleason Score) and generic concepts that can be used with multiple diagnoses (e.g. Laterality). Diagnoses (combinations of histology and site) are grouped into NAACCR schemas, and those schemas are linked to diagnosis-specific modifiers. These schemas will be preserved in the OMOP vocabulary to support accurate diagnosis-driven ETL processes.

1. **Binding of NAACCR variables with cancer diagnoses/diagnostic schemas**

I do not recommend using NAACCR bindings as a source (only source) of OMOP relationships between cancer diagnosis and their modifiers. They will not simplify the ETL of the NAACCR data beyond what the proposed below naming conventions accomplish. Nor will they represent a complete set of modifiers for each cancer diagnosis.

1. **Detailed Implementation**

In the text below NAACCR items are used interchangeably with NAACCR variables. Also, in the examples, for clarity, concept\_id often have concept\_name or concept\_code values instead of meaningless integer values.

1. **Scope**

A list of the NAACCR variables and their permissible values that should be ingested in the OMOP vocabulary is in this document: <https://docs.google.com/spreadsheets/d/1t1UO1Zs3XJhVW0xDeaC_ibCNpgiRP_D1B0hpCfGlPOU/edit?usp=sharing>

We excluded:

* ICD-O-2 diagnoses
* Secondary diagnoses

1. **Ingestion of NAACCR schemas into vocabulary**

There are 3 different bindings of diagnosis-dependent NAACCR variables (items):

* SEER EOD (Extent of Disease) Data : <https://staging.seer.cancer.gov/eod_public/list/1.4/>
* SEER TNM (Tumour Node Metastasis) Data: <https://staging.seer.cancer.gov/tnm/list/1.9/>
* NAACCR SSDI (Site Specific Data Items): <https://apps.naaccr.org/ssdi/list/>

SEER EOD includes NAACCR SSDI. These bindings are accomplished via schemas. Schemas are bundle combinations of ICD-O-3 site/histology diagnoses. The schemas for SEER EOD and SEER TNM overlap but are not equivalent. Some schema examples:

| **Binding** | **Schema** | **Primary Site** | **Histology** |
| --- | --- | --- | --- |
| EOD | Prostate | C619 | 8000-8700, 8720-8790, 9700-9701 |
| TNM | Prostate | C619 | 8000-8110, 8140-8576, 8940-8950, 8980-8981 |
| EOD | Adrenal Gland | C740-C741, C749 | 8000-8671, 8681-8683, 8691, 8720-8790, 9700-9701 |
| TNM | Adrenal Gland | C740 | 8010, 8140, 8370 |
|  |  | C749 | 8370 |

NAACCR schemas should be added to the OMOP Vocabulary as non-standard concepts to support diagnosis-driven ETL from CancerRegistry. Concept codes of NAACCR schemas are composed by concatenation of the binding type and NAACCR Schema name. For example:

| **concept\_name** | **domain\_id** | **vocabulary\_id** | **concept\_class\_id** | **standard\_concept** | **concept\_code** |
| --- | --- | --- | --- | --- | --- |
| NAACCR EOD Prostate Schema | Condition | NAACCR | Cancer Identification |  | EOD\_Prostate |
| NAACCR TNM Prostate Schema | Condition | NAACCR | Cancer Identification |  | TNM\_Prostate |

NAACCR schemas should be linked to pre-coordinated ICD-O-3 combinations of Histology and Topography in Concept\_Relationship table. For example:

| **concept\_id\_1** | **concept\_id\_2** | **relationship\_id** |
| --- | --- | --- |
| NAACCR EOD Prostate Schema | ‘8000/3- C619’ | NAACR Schema has Member |
| ‘8000/3- C619’ | NAACCR EOD Prostate Schema | Member of NAACR Schema |
| NAACCR TNM Prostate Schema | ‘8000/3- C619’ | NAACR Schema has Member |
| ‘8000/3- C619’ | NAACCR TNM Prostate Schema | Member of NAACR Schema |

1. **Ingestion of NAACCR variables and permissible values**

NAACCR variables (items) have unique identifiers. The majority of variables have the same semantic meaning in different diagnostic schemas. However, there is a small subset of variables that have completely different meaning in different schemas. These variables can be identified using their naming schema. In the data dictionary, their names start with ‘CS Site-Specific Factor’ followed by an integer (e.g. ‘CS Site-Specific Factor 7’). In binding schemas, they start with “SSF” prefix followed by an integer and then variable name. For example, NAACCR item #2880 is used by two completely different variables in Breast and Brain schemas:

Brain: [SSF1: World Health Organization (WHO) Grade Classification](https://staging.seer.cancer.gov/cs/input/02.05.50/brain/ssf1/?breadcrumbs=(~schema_list~),(~view_schema~,~brain~))

Breast: [SSF1: Estrogen Receptor (ER) Assay](https://staging.seer.cancer.gov/cs/input/02.05.50/breast/ssf1/?breadcrumbs=(~schema_list~),(~view_schema~,~breast~))

Permissible values have unique codes within the context of one NAACCR variable. However, even within the same variable, semantic meaning of a permissible value with the same code may vary depending on the diagnostic schema. Unlike variables with the same ID and different semantic meaning, permissible values with different semantic meaning are not easily identifiable. For example, NAACCR item #2810, CS Extension, has a permissible value code 050 that means “Benign or borderline brain tumor” in Brain schema and “Paget disease of nipple WITHOUT underlying tumor” in Breast schema.

Taking into consideration NAACCR coding ambiguity, we propose the following naming conventions for concept\_code of NAACCR variables and permissible values in the OMOP vocabulary. These conventions:

* ensure uniqueness of concept codes for NAACCR variables and their permissible values;
* preserve semantic value of each variable and permissible value within different diagnostic schemas;
* avoid duplication of variables/codes that carry the same meaning to an extent possible;

We also propose persisting explicit relationships between variables and their permissible values in the concept\_relationship table.

**A. If a NAACCR variable is not bound to any diagnostic schemas:**

Concept\_code for such variable should be the same as NAACCR item number

Concept\_code for permissible values of such variable should be composed by concatenating the respective NAACCR item number and permissible value code.

Concept\_name for such variable should be the same as NAACCR item name

Concept\_name for permissible values of such variable should be composed by concatenating the respective NAACCR item name and permissible value name.

Relationships between the variable and its permissible values should be stored in Concept\_Relationship table using ‘Has Answer’ – ‘Answer Of’ relationship pairs.

For example, the NAACCR item #410 'Laterality' and its permissible values should be ingested like this:

**CONCEPT**

| **concept\_id** | **concept\_name** | **domain\_id** | **vocabulary\_id** | **concept\_class\_id** | **standard\_concept** | **concept\_code** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | Laterality | Measurement | NAACCR | Cancer Identification | S | 410 |
| 2 | Laterality - Not a paired site | Meas Value | NAACCR | Cancer Identification | S | 410-0 |
| 3 | Laterality - Right: origin of primary | Meas Value | NAACCR | Cancer Identification | S | 410-1 |

**CONCEPT\_RELATIONSHIP**

| **concept\_id\_1** | **concept\_id\_2** | **relationship\_id** |
| --- | --- | --- |
| 1 | 2 | Has Answer |
| 2 | 1 | Answer of |
| 1 | 3 | Has Answer |
| 3 | 1 | Answer of |

**B. If a NAACCR variable is bound to one or more diagnostic schemas but its semantic meaning is the same:**

Concept\_code for such variable should be the same as NAACCR item number

Concept\_code for permissible values of such variable should be composed by concatenating the diagnostic schema name, respective NAACCR item number, and permissible value code.

Concept\_name for such variable should be the same as NAACCR item name

Concept\_name for permissible values of such variable should be composed by concatenating the respective diagnostic schema name, NAACCR item name, and permissible value name.

Relationships between the variable and its permissible values should be stored in Concept\_Relationship table using ‘Has Answer’ – ‘Answer Of’ relationship pairs as shown in A.

For example, the NAACCR item #2810, ‘CS Extension’, and its permissible values for Breast and Brain schemas should be ingested like follows:

**CONCEPT**

| **concept\_name** | **domain\_id** | **vocabulary\_id** | **concept\_class\_id** | **standard\_concept** | **concept\_code** |
| --- | --- | --- | --- | --- | --- |
| CS Extension | Measurement | NAACCR | Cancer Identification | S | 2810 |
| Breast-CS Extension - In situ: noninfiltrating | Meas Value | NAACCR | Cancer Identification | S | Breast-2810-000 |
| Breast-CS Extension - Paget disease of nipple WITHOUT underlying tumor | Meas Value | NAACCR | Cancer Identification | S | Breast-2810-050 |
| ……… |  |  |  |  | Breast-2810-... |
| Brain-CS Extension - Benign or borderline brain tumor | Meas Value | NAACCR | Cancer Identification | S | Brain-2810-050 |
| Brain-CS Extension -Supratentorial tumor confined to: Cerebral hemisphere or meninges of cerebral hemisphere | Meas Value | NAACCR | Cancer Identification | S | Brain-2810-100 |
| ……….. |  |  |  |  | Brain-2810-... |

**C. If a NAACCR variable is bound to one or more diagnostic schemas and its semantic varies between the schemas:**

As mentioned above, such variables can be identified using their naming schema that start with “SSF” prefix followed by an integer and then variable name.

Concept\_code for such variable should be composed by concatenating the diagnostic schema name and the NAACCR item number, thus generating two distinct OMOP concepts.

Concept\_code for permissible values of such variable should be composed by concatenating the diagnostic schema name, respective NAACCR item number, and permissible value code.

Concept\_name for such variable should be composed by concatenating the diagnostic schema name and NAACCR item name

Concept\_name for permissible values of such variable should be composed by concatenating the respective diagnostic schema name, NAACCR item name, and permissible value name.

Relationships between each of the two variables and their respective permissible values should be stored in Concept\_Relationship table using ‘Has Answer’ – ‘Answer Of’ relationship pairs as shown in A.

For example, the NAACCR variables under the item #2810 and their permissible values for Breast and Brain schemas should be ingested like follows:

**CONCEPT**

| **concept\_name** | **domain\_id** | **vocabulary\_id** | **concept\_class\_id** | **standard\_concept** | **concept\_code** |
| --- | --- | --- | --- | --- | --- |
| Breast-SSF1: Estrogen Receptor (ER) Assay | Measurement | NAACCR | Cancer Identification | S | Breast-2880 |
| Breast- Estrogen Receptor (ER) Assay - Positive/elevated | Meas Value | NAACCR | Cancer Identification | S | Breast-2880-010 |
| Breast- Estrogen Receptor (ER) Assay - Negative/normal; within normal limits | Meas Value | NAACCR | Cancer Identification | S | Breast-2880-020 |
| Breast-……….. | Meas Value | NAACCR | Cancer Identification | S | Breast-2880-... |
| Brain-SSF1: World Health Organization (WHO) Grade Classification | Measurement | NAACCR | Cancer Identification | S | Brain-2880 |
| Brain-WHO Classification - Grade I | Meas Value | NAACCR | Cancer Identification | S | Brain-2880-010 |
| Brain-WHO Classification - Grade II | Meas Value | NAACCR | Cancer Identification | S | Brain-2880-020 |
| Brain-……….. |  |  |  |  | Brain-2880-... |

1. **Ingestion of NAACCR permissible values representing numbers and numeric ranges**

Some NAACCR items include permissible values that represent numeric values or numeric ranges. For example, NAACCR item #2800, [CS Tumor Size](https://staging.seer.cancer.gov/cs/input/02.05.50/brain/size/?breadcrumbs=(~schema_list~),(~view_schema~,~brain~)), has four different kinds of permissible values (see below). One, 001-998, can be any numeric value in the range of 1-998. For example, the value of ‘015’ corresponds to 15 mm tumor size. The other three represent varieties of numeric ranges. Each of these permissible values is associated with a measurement unit.

| **Scenario** | **Code** | **Description** |
| --- | --- | --- |
| 1 | 001-988 | 001 - 988 millimeters (mm)  (Exact size to nearest mm) |
| 2 | 989 | 989 mm or larger |
| 3 | 991 | Described as "less than 1 centimeter (cm)" |
| 4 | 992 | Described as "less than 2 cm," or "greater than 1 cm," or "between 1 cm and 2 cm" |

According to the OMOP conventions, numeric attribute values are recorded in the value\_as\_number field in the Measurement table and units in unit\_concept\_id. Numeric ranges are represented by by two records each having the respective value of the upper and lower limit of the range and operator\_concept\_id. If a range is one-sided, only one record is created.

To automate ETL to an extent possible, we propose to persist these permissible values in the vocabulary.

The NAACCR permissible values that represent various numeric ranges (scenarios 2-4) will be stored in the vocabulary the same way as any other permissible value (see scenarios A, B, and C described in the previous section).

**CONCEPT**

| **concept\_name** | **domain\_id** | **vocabulary\_id** | **concept\_class\_id** | **standard\_concept** | **concept\_code** |
| --- | --- | --- | --- | --- | --- |
| 989 mm or larger | Meas Value | NAACCR | Cancer Identification | S | Brain-2800-989 |
| Described as "less than 1 centimeter (cm)" | Meas Value | NAACCR | Cancer Identification | S | Brain-2800-991 |
| Described as "less than 2 cm," or "greater than 1 cm," or "between 1 cm and 2 cm" | Meas Value | NAACCR | Cancer Identification | S | Brain-2800-992 |

In addition to this, we propose to add a new vocabulary table that will support explicit representation of the lower and upper limit numeric values, their units and math operators.

Add a new table, **CONCEPT\_NUMERIC**:

| **Field** | **Required** | **Type** | **Description** |
| --- | --- | --- | --- |
| concept\_id | Yes | integer | A foreign key that refers to a respective concept in the Standardized Vocabularies. |
| value\_as\_number | Yes | float | A value of the concept expressed as a numeric value. |
| unit\_concept\_id | No | integer | A foreign key to a Standard Concept ID of the concept units in the Standardized Vocabularies that belong to the 'Unit' domain. |
| operator\_concept\_id | yes | float | A foreign key identifier to the predefined Concept in the Standardized Vocabularies reflecting the mathematical operator that is applied to the value\_as\_number. Operators are <, <=, =, >=, > and these concepts belong to the 'Meas Value Operator' domain. |

Numeric ranges that have lower and upper limits will be represented by two records in this table. Numeric ranges containing only lower or upper limit will be represented by one record.

Below are examples of records in CONCEPT\_NUMERIC for scenarious 2-4:

**CONCEPT\_NUMERIC**

| **concept\_id** | **value\_as\_number** | **unit\_concept\_id** | **operator\_concept\_id** |
| --- | --- | --- | --- |
| XXXXXXX (‘989 mm or larger’) | 989 | YYYYYYY (‘mm’) | ZZZZZZZ(‘>=’) |
| XXXXXXX (‘Described as "less than 2 cm," or "greater than 1 cm," or "between 1 cm and 2 cm"’) | 1 | YYYYYYY (‘cm’) | ZZZZZZZ(‘>=’) |
| XXXXXXX (‘Described as "less than 2 cm," or "greater than 1 cm," or "between 1 cm and 2 cm"’) | 2 | YYYYYYY (‘cm’) | ZZZZZZZ(‘<=’) |

For scenario #1, permissible value ‘001 - 988 millimeters (mm) (Exact size to nearest mm)’, a concept will not be created and the value will be inserted as is in the measurement table. To constrain the value of the NAACCR item to numeric values only, we will associate it with the “Numeric” concept in the CONCEPT\_RELATIONSHIP table. To associate this value with units, the NAACCR item concept will be related with the respective unit concept in the CONCEPT\_RELATIONSHIP table. For example:

| **concept\_id\_1** | **relationship\_id** | **concept\_id\_2** |
| --- | --- | --- |
| XXXXXX(‘CS Tumor Size’) | has type | YYYYYYY(‘Numeric’) |
| XXXXXX(‘CS Tumor Size’) | has units | ZZZZZZZ(‘mm’) |

To constrain the NAACCR item to its permissible range (scenario #1), we create the following entries in the Concep Concept\_Numeric and Concept\_Relationship tables.

Concept table record:

| **concept\_name** | **domain\_id** | **vocabulary\_id** | **concept\_class\_id** | **standard\_concept** | **concept\_code** |
| --- | --- | --- | --- | --- | --- |
| 001 - 988 millimeters (mm)  (Exact size to nearest mm) | Metadata | NAACCR | Permissible range |  | Brain@2800@001-988 |

A record in Concept\_numeric is related to the concept above for “Brain@2800@001-988” code:

| **concept\_id** | **value\_as\_number** | **unit\_concept\_id** | **operator\_concept\_id** |
| --- | --- | --- | --- |
| XXXXXXX (‘001-988) | 1 | YYYYYYY (‘mm’) | ZZZZZZZ(‘>=’) |
| XXXXXXX (‘001-988) | 988 | YYYYYYY (‘mm’) | ZZZZZZZ(‘<=’) |

A record in Concept\_relationship along with a relationship ‘has permissible range’:

| **concept\_id\_1** | **relationship\_id** | **concept\_id\_2** |
| --- | --- | --- |
| XXXXXX(‘CS Tumor Size’) | has permissible range | YYYYYYY(‘001-988’) |

1. **Binding of NAACCR variables with cancer diagnoses/diagnostic schemas**

NAACCR bindings are an artificial workaround that supports their sub-optimal coding system. Although explicit bindings of a NAACCR item with a diagnostic schema represent true linkage between the cancer diagnosis and its modifiers, absence of an explicit binding does not indicate that the modifier is not bound to the diagnosis.

For example, laterality is not bound to any diagnostic schemas. However, it does not indicate that it binds to all diagnostic schemas. We know that it binds to Breast or Brain but does not bind to Pancreas. It only indicates that semantic meaning of this NAACCR item and its permissible value codes do not change when they are used within different diagnostic groups.

For this reason, I do not recommend using NAACCR bindings as a source of OMOP relationships between cancer diagnosis and their modifiers. They will not simplify the ETL of the NAACCR data (proposed ETL approach is described in the next section) variables nor will they represent a complete set of modifiers for each cancer diagnosis.

1. **Binding of NAACCR variables with cancer diagnoses**

In order to explicitly connect cancer diagnosis in the Condition table with cancer modifiers in the Measurement table, we connect each cancer diagnosis (schema?) to its respective NAACCR items (modifiers) in the vocabulary???

NAACCR bindings are an artificial workaround that supports their sub-optimal coding system. Although explicit bindings of a NAACCR item with a diagnostic schema represent true linkage between the cancer diagnosis and its modifiers, absence of an explicit binding does not indicate that the modifier is not bound to the diagnosis.

For example, laterality is not bound to any diagnostic schemas. However, it does not indicate that it binds to all diagnostic schemas. We know that it binds to Breast or Brain but does not bind to Pancreas. It only indicates that semantic meaning of this NAACCR item and its permissible value codes do not change when they are used within different diagnostic groups.

For this reason, I do not recommend using NAACCR bindings as a source of OMOP relationships between cancer diagnosis and their modifiers. They will not simplify the ETL of the NAACCR data (proposed ETL approach is described in the next section) variables nor will they represent a complete set of modifiers for each cancer diagnosis.

1. **Connecting NAACCR dates with respective variables**

For automation of ETL of dates into Measuremtn.measurement\_datetime, we propose to explicitly connect NAACCR variables with their designated date variables in Concept\_Relationship. The designated date variables should be marked as non-standard. They should not be ETLed into Measurement as separate records. An example of a connected variable and date is:

NAACCR Item #2800 (CS Tumor Size) ***‘has Date’*** NAACCR Item #443 (Date Conclusive Dx)

A list of connected variables and dates is here: <https://docs.google.com/spreadsheets/d/1t1UO1Zs3XJhVW0xDeaC_ibCNpgiRP_D1B0hpCfGlPOU/edit#gid=1863350377>

1. **ETL of NAACCR Data into OMOP CDM**

This section covers only mapping of NAACCR variable-value pairs to OMOP concepts to populate the following Measurement table fields: measurement\_concept\_id, value\_as\_concept\_id or value\_as number. It provides demonstration of how the proposed representation of the NAACCR data dicitionary will support the ETL process. The complete ETL instructions for populating cancer diagnosis and diagnostic modifiers will be described separately.

The following four parameters in the source data will determine how NAACCR variable and its value will be mapped to respective OMOP concepts:

1. Cancer histology coded in ICD-O-3, NAACCR item #521 ‘Morph--Type&Behav ICD-O-3’ (e.g. ‘9141/3’)
2. Cancer site in ICD-O-3, NAACCR item #400, ‘Primary site’ (e.g. ‘C50.1’)
3. NAACCR variable item # (e.g. item #2810, ‘CS Extension’)
4. NAACCR variable value code (e.g. ‘000’, ‘In situ: noninfiltrating’)

**ETL crosswalk** **steps**

For clarity, we have broken the ETL process into the following steps. The combined SQL is provided at the end of this section.

1. Identify NAACR diagnostic schema concept\_code

This concept\_code participates in the naming of NAACCR variables and permissible values and will be used to find proper mapping between the NAACCR variable and permissible value and respective OMOP concepts.

* 1. Concatenate histology and site into one concept\_code like this:

‘9141/3-C501’

* 1. Find concept\_code of the NAACCR diagnostic schema for this cancer diagnosis in the OMOP vocabulary like this:

SELECT s.concept\_code

FROM concept AS s

JOIN concept\_relationship on s.concept\_id = concept\_id\_1

AND relationship\_id = ‘NAACR Schema has Member’

JOIN concept d on d.concept\_id\_2 = d.concept\_id

AND d.vocabulary\_id = ‘ICD-O-3’

AND d.concept\_code = ‘9141/3-C501’

WHERE s.vocabulary\_id = ‘NAACCR’

The returned concept\_code should be ‘Breast’

1. Find OMOP concept corresponding to the NAACCR variable

Depending on the uniqueness of the NAACCR item ID, its concept\_code in OMOP vocabulary will be based either on the item ID alone or on concatenation of the diagnostic schema and item ID. Therefore, the query to find the concept should include both possibilities.

For the item ID #2810, ‘CS Extension’, the query will look like this:

SELECT concept\_id

FROM concept

WHERE vocabulary\_id = ‘NAACCR’

AND (concept\_code = ‘2810’ OR concept\_code = ‘Breast-2810’)

The resulting concept\_id will be recorded in Measurement.measurement\_concept\_id.

If a concept not found, it means that the respective NAACCR variable is not incorporated in the OMOP Vocabulary.

1. Find OMOP concept of the NAACCR permissible value

To uniquely map the source value to the corresponding OMOP concept, we will use an approach similar to the one we used for the variable, based on concept\_code naming conventions for NAACCR permissible values. Depending on the uniqueness of the NAACCR item ID, concept\_code of its permissible values in OMOP vocabulary will be based either on concatenation of the NAACCR item ID and permissible value code or on concatenation of the diagnostic schema, NAACCR item ID, and permissible value code. Therefore, the query to find the concept should include both possibilities.

For the permissible value ‘000’, ‘In situ: noninfiltrating’, of item ID #2810, ‘CS Extension’, the query will look like this:

SELECT concept\_id

FROM concept

WHERE vocabulary\_id = ‘NAACCR’

AND (concept\_code = ‘2810-000’ OR concept\_code = ‘Breast-2810-000’)

If a concept is not found, it implies that the value is numeric and should be recorded in Measurement.value\_as\_number as it appears in the source. The value should be verified to be truly numeric.

**ETL in one combined SQL statement**

A combined ETL statement for PostgreSQL is provided below. The *data\_source* table contains source data, including *site*, *histology*, NAACCR *variable\_id* and its *value*. The *measurement\_concept\_id* and *measurement\_attribute\_value* contain the respective values for populating measurement\_concept\_id and value\_as\_concept\_id\_value\_as\_number.

**SELECT** S.histology, s.site, s.variable\_id **as** item, s."value" **as** code

, d.concept\_code **as** condition\_concept

, d.concept\_name **as** **condition**

, c1.concept\_name **as** **schema**

, c2.concept\_id **as** measurement\_concept\_id

, c2.concept\_code **as** attribute\_code

, c2.concept\_name **as** **attribute**

, c3.concept\_code **as** value

, c3.concept\_id **as** value\_as\_concept\_id

**-- Numeric concepts**

, **coalesce**(cn.value\_as\_number,s.**value**) **as** measurement\_attribute\_value

, cn.unit\_concept\_id **as** unit\_concept\_id

, cn.operator\_concept\_id **as** operator\_concept\_id

**FROM** data\_source **AS** s

-- Getting schema

**LEFT** **JOIN** concept d

**ON** d.vocabulary\_id = 'ICD-O-3' **AND** d.concept\_code = s.histology || '-' || s.site

**LEFT** **JOIN** concept\_relationship cr1

**ON** d.concept\_id = cr1.concept\_id\_2 **AND** cr1.relationship\_id = 'Schema has Member'

**LEFT** **JOIN** concept **AS** c1

**ON** cr1.concept\_id\_1 = c1.concept\_id **AND** c1.vocabulary\_id = 'NAACCR'

-- Getting variable

**JOIN** concept **AS** c2

**ON** c2.vocabulary\_id = 'NAACCR' **AND** (c2.concept\_code = s.variable\_id **OR** c2.concept\_code = c1.concept\_code || '-' || s.variable\_id)

-- Getting permissible value

**LEFT** **JOIN** concept **AS** c3

**ON** c3.vocabulary\_id = 'NAACCR' **AND** (c3.concept\_code = s.variable\_id || '-' || s.**value** **OR** c3.concept\_code = c1.concept\_code || '-' || s.variable\_id || '-' || s.**value**)

**-- Numeric concepts**

**LEFT** **JOIN** concept\_numeric **AS** cn

**ON** c3.concept\_id = cn.concept\_id

The DDL for Data\_Source table and SQL to populate the tables used in testing the ETL are given in the Appendix.

1. **OUTSTANDING ISSUES**

Vocabulary, immediate needs

1. Permissible value names will require some additional editing. Some are too long, some include line breaks and other special characters.
2. Decide what to do with different flavors of ‘Unknown’: ingest or not ingest? If ingesting, they should be mapped to 0. If not ingesting, should we add them to the vocabulary so that they are recognized?
3. Coding system handling. Create hierarchy of cancer registry coding systems for staging, grading, etc.? Or, better, should coding system be attached to the concepts? Ex: “NAACCR-Stage1”?
4. Handling of provenance concepts. There are two types of provenance concepts: one describing diagnostic source (e.g. pathologist, clinician) and another describing record source (e.g. pathology report, EHR). Which one should we use?
5. What do we do with non-numeric NAACCR variables without permissible values (e.g. CS\*)? Do we need them?
6. Should some of the variables go to Treatment domain?
7. Double check how site-specific variable can be detected: ‘CS Site-Specific Factor…’ or ‘SSF…’?
8. Check variables that I marked as “Do we need it” or “What is it”and decide whether to ingest them or not .

Vocabulary, next steps

1. Map known lab values to LOINC

For ETL

1. Attach appropriate date to each modifier - Crystal?
2. Add provenance instructions
3. Complete

**APPENDIX A**

DDL for Data\_Source table and SQL to populate tables used in testing the ETL

1. Data\_Source

--DDL

**CREATE** **TABLE** public.data\_source (

site **varchar**(100) **NULL**,

histology **varchar**(100) **NULL**,

variable\_id **varchar**(100) **NULL**,

"value" **varchar**(100) **NULL**,

id **varchar**(100) **NOT** **NULL**,

**CONSTRAINT** data\_source\_pk **PRIMARY** **KEY** (id)

)

**WITH** (

**OIDS**=**FALSE**

);

--Insert Records

**INSERT** **INTO** public.data\_source

(site, histology, variable\_id, "value", id)

**VALUES**('C501', '8500/3', '410', '1', '2');

**INSERT** **INTO** public.data\_source

(site, histology, variable\_id, "value", id)

**VALUES**('C711', '9451/3', '2810', '050', '5');

**INSERT** **INTO** public.data\_source

(site, histology, variable\_id, "value", id)

**VALUES**('C501', '8500/3', '230', '51', '9');

**INSERT** **INTO** public.data\_source

(site, histology, variable\_id, "value", id)

**VALUES**('C501', '8500/3', '2880', '020', '7');

**INSERT** **INTO** public.data\_source

(site, histology, variable\_id, "value", id)

**VALUES**('C501', '8500/3', '2810', '050', '3');

**INSERT** **INTO** public.data\_source

(site, histology, variable\_id, "value", id)

**VALUES**('C619', '8000/3', '410', '0', '1');

**INSERT** **INTO** public.data\_source

(site, histology, variable\_id, "value", id)

**VALUES**('C711', '9451/3', '230', '39', '8');

**INSERT** **INTO** public.data\_source

(site, histology, variable\_id, "value", id)

**VALUES**('C711', '9451/3', '2880', '020', '6');

**INSERT** **INTO** public.data\_source

(site, histology, variable\_id, "value", id)

**VALUES**('C711', '9451/3', '410', '1', '4');

1. Concept

--Insert Records

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(16, 'Brain-SSF1: World Health Organization (WHO) Grade Classification', 'Measurement', 'NAACCR', 'NAACCR Item', 'S', 'Brain-2880', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(3, 'NAACCR EOD Breast Schema', 'NAACCR', 'NAACCR', 'NAACCR Schema', **NULL**, 'Breast', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(20, 'Breast- Estrogen Receptor (ER) Assay - Negative/normal; within normal limits', 'Meas Value', 'NAACCR', 'NAACCR Code', 'S', 'Breast-2880-020', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(9, 'Laterality - Right: origin of primary', 'Meas Value', 'NAACCR', 'NAACCR Code', 'S', '410-1', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(12, 'Breast-CS Extension - Paget disease of nipple WITHOUT underlying tumor', 'Meas Value', 'NAACCR', 'NAACCR Code', 'S', 'Breast-2810-050', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(17, 'Brain-WHO Classification - Grade I', 'Meas Value', 'NAACCR', 'NAACCR Code', 'S', 'Brain-2880-010', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(19, 'Breast- Estrogen Receptor (ER) Assay - Positive/elevated', 'Meas Value', 'NAACCR', 'NAACCR Code', 'S', 'Breast-2880-010', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(1, 'NAACCR EOD Prostate Schema', 'NAACCR', 'NAACCR', 'NAACCR Schema', **NULL**, 'Prostate', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(8, 'Laterality - Not a paired site', 'Meas Value', 'NAACCR', 'NAACCR Code', 'S', '410-0', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(13, 'Brain-CS Extension - Benign or borderline brain tumor', 'Meas Value', 'NAACCR', 'NAACCR Code', 'S', 'Brain-2810-050', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(18, 'Brain-WHO Classification - Grade II', 'Meas Value', 'NAACCR', 'NAACCR Code', 'S', 'Brain-2880-020', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(6, 'Anaplastic Oligodendroglioma of Frontal lobe', 'Condition', 'ICD-O-3', 'Cancer', **NULL**, '9451/3-C711', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(14, 'Brain-CS Extension -Supratentorial tumor confined to: Cerebral hemisphere or meninges of cerebral hemisphere', 'Meas Value', 'NAACCR', 'NAACCR Code', 'S', 'Brain-2810-100', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(10, 'CS Extension', 'Measurement', 'NAACCR', 'NAACCR Item', 'S', '2810', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(7, 'Laterality', 'Measurement', 'NAACCR', 'NAACCR Item', 'S', '410', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(5, 'Infiltrating Duct Carcinoma of Central Portion of Breast', 'Condition', 'ICD-O-3', 'Cancer', **NULL**, '8500/3-C501', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(21, 'Age at diagnosis', 'Measurement', 'NAACCR', 'NAACCR Item', 'S', '230', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(15, 'Breast-SSF1: Estrogen Receptor (ER) Assay', 'Measurement', 'NAACCR', 'NAACCR Item', 'S', 'Breast-2880', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(11, 'Breast-CS Extension - In situ: noninfiltrating', 'Meas Value', 'NAACCR', 'NAACCR Code', 'S', 'Breast-2810-000', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(4, 'NAACCR EOD Brain Schema', 'NAACCR', 'NAACCR', 'NAACCR Schema', **NULL**, 'Brain', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept

(concept\_id, concept\_name, domain\_id, vocabulary\_id, concept\_class\_id, standard\_concept, concept\_code, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(2, 'Tumor, malignant, NOS of Prostate Gland', 'Condition', 'ICD-O-3', 'Cancer', **NULL**, '8000/3-C619', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

1. Concept\_Relationship

--Insert Records

**INSERT** **INTO** public.concept\_relationship

(concept\_id\_1, concept\_id\_2, relationship\_id, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(5, 3, 'Member of Schema', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept\_relationship

(concept\_id\_1, concept\_id\_2, relationship\_id, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(4, 6, 'Schema has Member', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept\_relationship

(concept\_id\_1, concept\_id\_2, relationship\_id, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(4, 21, 'Schema has Member', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept\_relationship

(concept\_id\_1, concept\_id\_2, relationship\_id, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(21, 4, 'Member of Schema', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept\_relationship

(concept\_id\_1, concept\_id\_2, relationship\_id, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(3, 21, 'Schema has Member', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept\_relationship

(concept\_id\_1, concept\_id\_2, relationship\_id, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(2, 1, 'Member of Schema', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept\_relationship

(concept\_id\_1, concept\_id\_2, relationship\_id, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(3, 5, 'Schema has Member', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept\_relationship

(concept\_id\_1, concept\_id\_2, relationship\_id, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(21, 3, 'Member of Schema', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept\_relationship

(concept\_id\_1, concept\_id\_2, relationship\_id, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(6, 4, 'Member of Schema', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);

**INSERT** **INTO** public.concept\_relationship

(concept\_id\_1, concept\_id\_2, relationship\_id, valid\_start\_date, valid\_end\_date, invalid\_reason)

**VALUES**(1, 2, 'Schema has Member', TO\_DATE('2000-01-01','YYYY-MM-DD'), TO\_DATE('2090-01-01','YYYY-MM-DD'), **NULL**);