mCode to OMOP Standard Library

Contents

Specimen . . . 2
Map from mCode to OMOP . . . 2
Cancer Staging . . . 5
Genomics . . . 6
HGVS . . . 6
HGNC . . . 7
ClinVar . . . 8
Final Genomics Library . . . 9
Units of Measure . . . 9
SNOMED . . . 10
LOINC . . . 15
ICD-10 CM . . . 15

Last updated on: 2021-01-02

```
library(mOMOP)
library(chariot)
library(tidyverse)
```

```
conn <- chariot::connectAthena()
```

This vignette takes a look at the Value Sets in the mCode Data Dictionary to create the standard library of Oncology concepts, available for loading in this package.

The mCode valuesets are retrieved from the publicly available data dictionary.

```
value_sets <- get_value_sets()
value_sets
```

```
## # A tibble: 6,603 x 5
##    `Value Set Name`   `Code System` `Logical Definitio~ Code  `Code Description`
##    <chr>              <chr>         <chr>               <chr> <chr>
##  1 CancerBodyLocatio~ SNOMED CT     includes codes des~ <NA>  <NA>
##  2 CancerDiseaseStat~ SNOMED CT     <NA>                3636~ Imaging (procedur~
##  3 CancerDiseaseStat~ SNOMED CT     <NA>                2524~ Histopathology te~
##  4 CancerDiseaseStat~ SNOMED CT     <NA>                7110~ Assessment of sym~
##  5 CancerDiseaseStat~ SNOMED CT     <NA>                5880~ Physical examinat~
##  6 CancerDiseaseStat~ SNOMED CT     <NA>                2507~ Tumor marker meas~
##  7 CancerDiseaseStat~ SNOMED CT     <NA>                3863~ Laboratory data i~
##  8 CancerDisorderVS   ICD-10-CM     <NA>                C000  Malignant neoplas~
##  9 CancerDisorderVS   ICD-10-CM     <NA>                C001  Malignant neoplas~
## 10 CancerDisorderVS   ICD-10-CM     <NA>                C002  Malignant neoplas~
## # ... with 6,593 more rows
```

To map the valuesets to OMOP concepts, it is split by `Code System`, the equivalent to the `vocabulary_id` field in OMOP's Concept table, for joining. As of 2021-01-02, there are 9 `Code Systems` used in mCode.

```
value_sets_by_vocab <- value_sets %>% rubix::split_by(col = `Code System`)
names(value_sets_by_vocab)
```

```
## [1] "ClinVar"
## [2] "http://cancerstaging.org"
## [3] "http://terminology.hl7.org/CodeSystem/v2-0487"
## [4] "http://unitsofmeasure.org"
## [5] "http://varnomen.hgvs.org"
## [6] "http://www.genenames.org/geneId"
## [7] "ICD-10-CM"
## [8] "LOINC"
## [9] "SNOMED CT"
```

Specimen

In mCode, Specimen representation is based on HL7's Code System while in OMOP, the Specimen domain has its own subset of concepts. Furthermore, mCode's
Specimen representation maps to other OMOP Concept Ids. The Specimen representation in this package therefore has 2 parts:

1. A map between mCode and OMOP Concepts outside OMOP's Specimen domain

2. OMOP's Specimen domain concepts

Map from mCode to OMOP

Specimen representation is derived from the valuesets.

```
specimen_library <- value_sets_by_vocab$`http://terminology.hl7.org/CodeSystem/v2-0487` %>%
    rubix::format_colnames()
head(specimen_library)
```

```
## # A tibble: 6 x 5
##   value_set_name   code_system           logical_definit~ code  code_description
##   <chr>            <chr>                 <chr>            <chr> <chr>
## 1 GeneticSpecimen~ http://terminology.h~ <NA>             AMN   Amniotic fluid
## 2 GeneticSpecimen~ http://terminology.h~ <NA>             BIFL  Bile Fluid
## 3 GeneticSpecimen~ http://terminology.h~ <NA>             BLD   Whole blood
## 4 GeneticSpecimen~ http://terminology.h~ <NA>             BLDA  Blood arterial
## 5 GeneticSpecimen~ http://terminology.h~ <NA>             BLDCO Cord blood
## 6 GeneticSpecimen~ http://terminology.h~ <NA>             BLDV  Blood venous
```

The specimen description is joined on the Concept Synonym Names in the Concept Synonym table to retrieve the Concept Id the mCode Specimen maps to.

```
specimen_omop_library1 <- join_on_concept_synonym_name(data = specimen_library,
    column = "code_description", case_insensitive = TRUE, conn = conn) %>%
    select(-concept_synonym_name, -language_concept_id) %>% rename(match_concept_id = concept_id)
```

```
head(specimen_omop_library1)
```

```
##          value_set_name                                    code_system
## 1 GeneticSpecimenTypeVS http://terminology.hl7.org/CodeSystem/v2-0487
## 2 GeneticSpecimenTypeVS http://terminology.hl7.org/CodeSystem/v2-0487
## 3 GeneticSpecimenTypeVS http://terminology.hl7.org/CodeSystem/v2-0487
```

```
## 4 GeneticSpecimenTypeVS http://terminology.hl7.org/CodeSystem/v2-0487
## 5 GeneticSpecimenTypeVS http://terminology.hl7.org/CodeSystem/v2-0487
## 6 GeneticSpecimenTypeVS http://terminology.hl7.org/CodeSystem/v2-0487
##   logical_definition code code_description match_concept_id
## 1               <NA>  SKN             Skin          1027716
## 2               <NA>  PLC         Placenta          1032268
## 3               <NA>  SAL           Saliva          1033195
## 4               <NA>  SPT           Sputum          1033465
## 5               <NA>  WND            Wound          1033739
## 6               <NA>  BON             Bone          1585831
```

The complete Concept representation is taken from the Concept Id.

```
specimen_omop_library2 <- join_on_concept_id(data = specimen_omop_library1,
    column = "match_concept_id", conn = conn)
```

```
head(specimen_omop_library2)
```

```
##          value_set_name                                    code_system
## 1 GeneticSpecimenTypeVS http://terminology.hl7.org/CodeSystem/v2-0487
## 2 GeneticSpecimenTypeVS http://terminology.hl7.org/CodeSystem/v2-0487
## 3 GeneticSpecimenTypeVS http://terminology.hl7.org/CodeSystem/v2-0487
## 4 GeneticSpecimenTypeVS http://terminology.hl7.org/CodeSystem/v2-0487
## 5 GeneticSpecimenTypeVS http://terminology.hl7.org/CodeSystem/v2-0487
## 6 GeneticSpecimenTypeVS http://terminology.hl7.org/CodeSystem/v2-0487
##   logical_definition code code_description match_concept_id concept_id
## 1               <NA>  SKN             Skin          1027716    1027716
## 2               <NA>  PLC         Placenta          1032268    1032268
## 3               <NA>  SAL           Saliva          1033195    1033195
## 4               <NA>  SPT           Sputum          1033465    1033465
## 5               <NA>  WND            Wound          1033739    1033739
## 6               <NA>  BON             Bone          1585831    1585831
##                         concept_name   domain_id vocabulary_id concept_class_id
## 1                               Skin Observation         LOINC     LOINC System
## 2                           Placenta Observation         LOINC     LOINC System
## 3                             Saliva Observation         LOINC     LOINC System
## 4                             Sputum Observation         LOINC     LOINC System
## 5                              Wound Observation         LOINC     LOINC System
## 6 Organ Transplant Description: Bone Observation           PPI           Answer
##   standard_concept                    concept_code valid_start_date
## 1             <NA>                       LP36760-4       1970-01-01
## 2             <NA>                        LP7477-5       1970-01-01
## 3             <NA>                        LP7565-7       1970-01-01
## 4             <NA>                        LP7600-2       1970-01-01
## 5             <NA>                        LP7726-5       1970-01-01
## 6             <NA> OrganTransplantDescription_Bone       2017-05-17
##   valid_end_date invalid_reason
## 1     2099-12-31           <NA>
## 2     2099-12-31           <NA>
## 3     2099-12-31           <NA>
## 4     2099-12-31           <NA>
## 5     2099-12-31           <NA>
## 6     2099-12-31           <NA>
```

To maintain a one-to-one representation of the original grain of information, the OMOP mappings are pivoted on the OMOP Domain to see how each mCode Specimen maps to by domain.

```
specimen_omop_library3 <- specimen_omop_library2 %>% merge_strip(into = "concept",
    domain_id) %>% pivot_wider(id_cols = !concept, names_from = domain_id,
    values_from = concept) %>% select(-match_concept_id) %>%
    distinct()
head(specimen_omop_library3)
```

```
## # A tibble: 6 x 14
##   concept_id value_set_name code_system logical_definit~ code  code_description
##        <dbl> <chr>          <chr>       <chr>            <chr> <chr>
## 1    1027716 GeneticSpecim~ http://ter~ <NA>             SKN   Skin
## 2    1032268 GeneticSpecim~ http://ter~ <NA>             PLC   Placenta
## 3    1033195 GeneticSpecim~ http://ter~ <NA>             SAL   Saliva
## 4    1033465 GeneticSpecim~ http://ter~ <NA>             SPT   Sputum
## 5    1033739 GeneticSpecim~ http://ter~ <NA>             WND   Wound
## 6    1585831 GeneticSpecim~ http://ter~ <NA>             BON   Bone
## # ... with 8 more variables: Observation <chr>, Condition <chr>, `Spec Anatomic
## #   Site` <chr>, Drug <chr>, `Meas Value` <chr>, Measurement <chr>,
## #   Specimen <chr>, `NA` <chr>
```

To gather a complete representation of Specimens, any missing Specimen concepts in OMOP are added to the library.

```
omop_specimen <- chariot::queryAthena("SELECT *
                      FROM omop_vocabulary.concept
                      WHERE domain_id = 'Specimen' AND concept_class_id = 'Specimen';",
    conn = conn) %>% merge_strip(into = "Specimen") %>% select(-Specimen_id)
```

```
head(omop_specimen)
```

```
## # A tibble: 6 x 1
##   Specimen
##   <chr>
## 1 [V] [S] 759813 Physical object specimen [SNOMED 1021000124109] [Specimen] [Sp~
## 2 [V] [S] 759820 Paper specimen [SNOMED 1031000124107] [Specimen] [Specimen]
## 3 [V] [S] 759825 Writing paper specimen [SNOMED 1041000124102] [Specimen] [Spec~
## 4 [V] [S] 759826 Envelope specimen [SNOMED 1051000124100] [Specimen] [Specimen]
## 5 [V] [S] 759829 Package specimen [SNOMED 1061000124103] [Specimen] [Specimen]
## 6 [V] [S] 759869 Clothing specimen [SNOMED 1071000124105] [Specimen] [Specimen]
```

```
specimen_omop_library <- specimen_omop_library3 %>% full_join(omop_specimen,
    by = "Specimen") %>% distinct()
head(specimen_omop_library)
```

```
## # A tibble: 6 x 14
##   concept_id value_set_name code_system logical_definit~ code  code_description
##        <dbl> <chr>          <chr>       <chr>            <chr> <chr>
## 1    1027716 GeneticSpecim~ http://ter~ <NA>             SKN   Skin
## 2    1032268 GeneticSpecim~ http://ter~ <NA>             PLC   Placenta
## 3    1033195 GeneticSpecim~ http://ter~ <NA>             SAL   Saliva
## 4    1033465 GeneticSpecim~ http://ter~ <NA>             SPT   Sputum
## 5    1033739 GeneticSpecim~ http://ter~ <NA>             WND   Wound
## 6    1585831 GeneticSpecim~ http://ter~ <NA>             BON   Bone
## # ... with 8 more variables: Observation <chr>, Condition <chr>, `Spec Anatomic
## #   Site` <chr>, Drug <chr>, `Meas Value` <chr>, Measurement <chr>,
## #   Specimen <chr>, `NA` <chr>
```

This file is written to the `data-raw/` folder for distribution if it does not already exist.

```
file <- file.path(getwd(), "data-raw", "specimen.csv")
if (!file.exists(file)) {
    write_csv(x = specimen_omop_library, file = file)
}
```

Cancer Staging

mCode uses the AJCC TNM Staging system, which correlates with NCIt concepts in the OMOP vocabulary.

```
cancer_staging_library <- value_sets_by_vocab$`http://cancerstaging.org` %>%
    rubix::format_colnames()
head(cancer_staging_library)
```

```
## # A tibble: 4 x 5
##   value_set_name    code_system    logical_definition     code  code_description
##   <chr>             <chr>          <chr>                  <chr> <chr>
## 1 TNMDistantMetast~ http://cancer~ includes codes from c~ <NA>  <NA>
## 2 TNMPrimaryTumorC~ http://cancer~ includes codes from c~ <NA>  <NA>
## 3 TNMRegionalNodes~ http://cancer~ includes codes from c~ <NA>  <NA>
## 4 TNMStageGroupVS   http://cancer~ includes codes from c~ <NA>  <NA>
```

As it is represented in the OMOP Vocabulary, NCIt is not separated by Tumor, Node, and Metastasis like it is in mCode.

```
ncit_omop_library <- chariot::queryAthena("SELECT *
                      FROM omop_vocabulary.concept
                      WHERE vocabulary_id = 'NCIt' AND concept_class_id = 'AJCC Category';",
    conn = conn)
```

```
head(ncit_omop_library)
```

```
## # A tibble: 6 x 10
##   concept_id concept_name domain_id vocabulary_id concept_class_id
##        <dbl> <chr>        <chr>     <chr>         <chr>
## 1    1537692 Vulvar Canc~ Measurem~ NCIt          AJCC Category
## 2    1537693 Retinoblast~ Measurem~ NCIt          AJCC Category
## 3    1537694 Vulvar Canc~ Measurem~ NCIt          AJCC Category
## 4    1537695 Nasopharyng~ Measurem~ NCIt          AJCC Category
## 5    1537700 Adrenal Cor~ Measurem~ NCIt          AJCC Category
## 6    1537780 Retinoblast~ Measurem~ NCIt          AJCC Category
## # ... with 5 more variables: standard_concept <chr>, concept_code <chr>,
## #   valid_start_date <date>, valid_end_date <date>, invalid_reason <chr>
```

NCIt `AJCC Category` concepts are therefore grouped based on pattern matching with the Concept Code.

```
ncit_omop_library2 <- ncit_omop_library
ncit_omop_library2$value_set_name <- ""
ncit_omop_library2 <- ncit_omop_library2 %>% mutate(value_set_name = case_when(grepl("^[cp]{1}M",
    concept_code) ~ "TNMDistantMetastasesCategoryVS", grepl("^[cp]{1}N",
    concept_code) ~ "TNMRegionalNodesCategoryVS", grepl("^[cp]{1}T",
    concept_code) ~ "TNMPrimaryTumorCategoryVS"))
head(ncit_omop_library2)
```

```
## # A tibble: 6 x 11
##   concept_id concept_name domain_id vocabulary_id concept_class_id
```

```
##        <dbl> <chr>        <chr>     <chr>         <chr>
## 1    1537692 Vulvar Canc~ Measurem~ NCIt          AJCC Category
## 2    1537693 Retinoblast~ Measurem~ NCIt          AJCC Category
## 3    1537694 Vulvar Canc~ Measurem~ NCIt          AJCC Category
## 4    1537695 Nasopharyng~ Measurem~ NCIt          AJCC Category
## 5    1537700 Adrenal Cor~ Measurem~ NCIt          AJCC Category
## 6    1537780 Retinoblast~ Measurem~ NCIt          AJCC Category
## # ... with 6 more variables: standard_concept <chr>, concept_code <chr>,
## #   valid_start_date <date>, valid_end_date <date>, invalid_reason <chr>,
## #   value_set_name <chr>
```

A correlate to mCode's `TNMStageGroupVS` value set is not present in the OMOP Vocabularies, but since it can be derived from the TNM categories, it is skipped.

```
cancer_staging_omop_library <- cancer_staging_library %>% left_join(ncit_omop_library2,
    by = "value_set_name")
head(cancer_staging_omop_library)
```

```
## # A tibble: 6 x 15
##   value_set_name code_system logical_definit~ code  code_description concept_id
##   <chr>          <chr>       <chr>            <chr> <chr>                 <dbl>
## 1 TNMDistantMet~ http://can~ includes codes ~ <NA>  <NA>                1537692
## 2 TNMDistantMet~ http://can~ includes codes ~ <NA>  <NA>                1537780
## 3 TNMDistantMet~ http://can~ includes codes ~ <NA>  <NA>                1537798
## 4 TNMDistantMet~ http://can~ includes codes ~ <NA>  <NA>                1537804
## 5 TNMDistantMet~ http://can~ includes codes ~ <NA>  <NA>                1537805
## 6 TNMDistantMet~ http://can~ includes codes ~ <NA>  <NA>                1537808
## # ... with 9 more variables: concept_name <chr>, domain_id <chr>,
## #   vocabulary_id <chr>, concept_class_id <chr>, standard_concept <chr>,
## #   concept_code <chr>, valid_start_date <date>, valid_end_date <date>,
## #   invalid_reason <chr>
```

This data is written to a `cancer_staging.csv` in the `data-raw/` folder for distribution if it does not already exist.

```
file <- file.path(getwd(), "data-raw", "cancer_staging.csv")
if (!file.exists(file)) {
    write_csv(x = cancer_staging_omop_library, file = file)
}
```

Genomics

HGVS, HGNC, and ClinVar are collapsed into a single `Genomics` category.

HGVS

mCode uses HGVS (http://varnomen.hgvs.org), but these codes are not in the OMOP Vocabulary and cannot be derived from the website. Therefore, it is skipped for now.

```
hgvs_library <- value_sets_by_vocab$`http://varnomen.hgvs.org` %>%
    rubix::format_colnames()
head(hgvs_library)
```

```
## # A tibble: 1 x 5
##   value_set_name code_system      logical_definition      code  code_description
```

```
##   <chr>          <chr>            <chr>                    <chr> <chr>
## 1 HGVSVS         http://varnomen~ All codes in http://va~ <NA>  <NA>
```

HGNC

Gene Names in mCode are presumed to be derived from HGNC and the entire HGNC subset of the OMOP Vocabularies are included in the library.

```
hgnc_library <- value_sets_by_vocab$`http://www.genenames.org/geneId` %>%
    rubix::format_colnames()
head(hgnc_library)
```

```
## # A tibble: 1 x 5
##   value_set_name code_system      logical_definition      code  code_description
##   <chr>          <chr>            <chr>                   <chr> <chr>
## 1 HGNCVS         http://www.gene~ All codes in http://ww~ <NA>  <NA>
```

```
hgnc_omop_library1 <- chariot::queryAthena("SELECT *
                     FROM omop_vocabulary.concept
                     WHERE vocabulary_id = 'HGNC';",
    conn = conn)
```

```
head(hgnc_omop_library1)
```

```
## # A tibble: 6 x 10
##   concept_id concept_name domain_id vocabulary_id concept_class_id
##        <dbl> <chr>        <chr>     <chr>         <chr>
## 1   35944910 CCDC77 (coi~ Measurem~ HGNC          Gene
## 2   35944911 INMT (indol~ Measurem~ HGNC          Gene
## 3   35944912 ZNF117 (zin~ Measurem~ HGNC          Gene
## 4   35944913 CKAP2 (cyto~ Measurem~ HGNC          Gene
## 5   35944914 ITPRIPL2 (I~ Measurem~ HGNC          Gene
## 6   35944916 PEX19 (pero~ Measurem~ HGNC          Gene
## # ... with 5 more variables: standard_concept <chr>, concept_code <chr>,
## #   valid_start_date <date>, valid_end_date <date>, invalid_reason <chr>
```

The OMOP HGNC concepts are joined to the mCode data dictionary set.

```
hgnc_omop_library2 <- hgnc_omop_library1 %>% mutate(value_set_name = "HGNCVS")
head(hgnc_omop_library2)
```

```
## # A tibble: 6 x 11
##   concept_id concept_name domain_id vocabulary_id concept_class_id
##        <dbl> <chr>        <chr>     <chr>         <chr>
## 1   35944910 CCDC77 (coi~ Measurem~ HGNC          Gene
## 2   35944911 INMT (indol~ Measurem~ HGNC          Gene
## 3   35944912 ZNF117 (zin~ Measurem~ HGNC          Gene
## 4   35944913 CKAP2 (cyto~ Measurem~ HGNC          Gene
## 5   35944914 ITPRIPL2 (I~ Measurem~ HGNC          Gene
## 6   35944916 PEX19 (pero~ Measurem~ HGNC          Gene
## # ... with 6 more variables: standard_concept <chr>, concept_code <chr>,
## #   valid_start_date <date>, valid_end_date <date>, invalid_reason <chr>,
## #   value_set_name <chr>
```

```
hgnc_omop_library <- hgnc_library %>% left_join(hgnc_omop_library2,
    by = "value_set_name")
head(hgnc_omop_library)
```

```
## # A tibble: 6 x 15
##   value_set_name code_system logical_definit~ code  code_description concept_id
##   <chr>          <chr>       <chr>            <chr> <chr>                 <dbl>
## 1 HGNCVS         http://www~ All codes in ht~ <NA>  <NA>               35944910
## 2 HGNCVS         http://www~ All codes in ht~ <NA>  <NA>               35944911
## 3 HGNCVS         http://www~ All codes in ht~ <NA>  <NA>               35944912
## 4 HGNCVS         http://www~ All codes in ht~ <NA>  <NA>               35944913
## 5 HGNCVS         http://www~ All codes in ht~ <NA>  <NA>               35944914
## 6 HGNCVS         http://www~ All codes in ht~ <NA>  <NA>               35944916
## # ... with 9 more variables: concept_name <chr>, domain_id <chr>,
## #   vocabulary_id <chr>, concept_class_id <chr>, standard_concept <chr>,
## #   concept_code <chr>, valid_start_date <date>, valid_end_date <date>,
## #   invalid_reason <chr>
```

ClinVar

Like HGNC, the entire ClinVar subset of the OMOP Vocabularies are included in the library.

```
clinvar_library <- value_sets_by_vocab$ClinVar %>% rubix::format_colnames()
head(clinvar_library)
```

```
## # A tibble: 1 x 5
##   value_set_name code_system logical_definition          code  code_description
##   <chr>          <chr>       <chr>                      <chr> <chr>
## 1 ClinVarVS      ClinVar     Includes codes from ClinVar <NA>  <NA>
```

```
clinvar_omop_library1 <- chariot::queryAthena("SELECT *
                      FROM omop_vocabulary.concept
                      WHERE vocabulary_id = 'ClinVar';",
    conn = conn)
```

```
head(clinvar_omop_library1)
```

```
## # A tibble: 6 x 10
##   concept_id concept_name domain_id vocabulary_id concept_class_id
##        <dbl> <chr>        <chr>     <chr>         <chr>
## 1   35968119 NC_000002.1~ Measurem~ ClinVar       Variant
## 2   35968121 NM_000038.6~ Measurem~ ClinVar       Variant
## 3   35968122 NM_000038.6~ Measurem~ ClinVar       Variant
## 4   35968123 NM_000038.6~ Measurem~ ClinVar       Variant
## 5   35968124 NM_000038.6~ Measurem~ ClinVar       Variant
## 6   35968125 NM_000038.6~ Measurem~ ClinVar       Variant
## # ... with 5 more variables: standard_concept <chr>, concept_code <chr>,
## #   valid_start_date <date>, valid_end_date <date>, invalid_reason <chr>
```

```
clinvar_omop_library2 <- clinvar_omop_library1 %>% mutate(value_set_name = "ClinVarVS")
head(clinvar_omop_library2)
```

```
## # A tibble: 6 x 11
##   concept_id concept_name domain_id vocabulary_id concept_class_id
##        <dbl> <chr>        <chr>     <chr>         <chr>
## 1   35968119 NC_000002.1~ Measurem~ ClinVar       Variant
## 2   35968121 NM_000038.6~ Measurem~ ClinVar       Variant
## 3   35968122 NM_000038.6~ Measurem~ ClinVar       Variant
## 4   35968123 NM_000038.6~ Measurem~ ClinVar       Variant
## 5   35968124 NM_000038.6~ Measurem~ ClinVar       Variant
```

```
## 6   35968125 NM_000038.6~ Measurem~ ClinVar       Variant
## # ... with 6 more variables: standard_concept <chr>, concept_code <chr>,
## #   valid_start_date <date>, valid_end_date <date>, invalid_reason <chr>,
## #   value_set_name <chr>
```

```
clinvar_omop_library <- clinvar_library %>% left_join(clinvar_omop_library2,
    by = "value_set_name")
head(clinvar_omop_library)
```

```
## # A tibble: 6 x 15
##   value_set_name code_system logical_definit~ code  code_description concept_id
##   <chr>          <chr>       <chr>            <chr> <chr>                 <dbl>
## 1 ClinVarVS      ClinVar     Includes codes ~ <NA>  <NA>                35968119
## 2 ClinVarVS      ClinVar     Includes codes ~ <NA>  <NA>                35968121
## 3 ClinVarVS      ClinVar     Includes codes ~ <NA>  <NA>                35968122
## 4 ClinVarVS      ClinVar     Includes codes ~ <NA>  <NA>                35968123
## 5 ClinVarVS      ClinVar     Includes codes ~ <NA>  <NA>                35968124
## 6 ClinVarVS      ClinVar     Includes codes ~ <NA>  <NA>                35968125
## # ... with 9 more variables: concept_name <chr>, domain_id <chr>,
## #   vocabulary_id <chr>, concept_class_id <chr>, standard_concept <chr>,
## #   concept_code <chr>, valid_start_date <date>, valid_end_date <date>,
## #   invalid_reason <chr>
```

Final Genomics Library

The final genomics library file is written if it does not already exist.

```
genomics_omop_library <- bind_rows(hgnc_omop_library, clinvar_omop_library)
file <- file.path(getwd(), "data-raw", "genomics.csv")
if (!file.exists(file)) {
    write_csv(x = genomics_omop_library, file = file)
}
```

Units of Measure

Like Specimen, the Units of Measure representation is limited and requires incorporation of the OMOP `UCUM` vocabulary valueset.

```
uom_library <- value_sets_by_vocab$`http://unitsofmeasure.org` %>%
    rubix::format_colnames()
head(uom_library)
```

```
## # A tibble: 6 x 5
##   value_set_name  code_system           logical_definiti~ code  code_description
##   <chr>           <chr>                 <chr>             <chr> <chr>
## 1 UnitsOfLengthVS http://unitsofmeasur~ <NA>              pm    Picometer
## 2 UnitsOfLengthVS http://unitsofmeasur~ <NA>              nm    Nanometer
## 3 UnitsOfLengthVS http://unitsofmeasur~ <NA>              mm    Millimeter
## 4 UnitsOfLengthVS http://unitsofmeasur~ <NA>              cm    Centimeter
## 5 UnitsOfLengthVS http://unitsofmeasur~ <NA>              m     Meter
## 6 UnitsOfLengthVS http://unitsofmeasur~ <NA>              ft-us Foot
```

```
ucum_omop_library1 <- chariot::queryAthena("SELECT *
                    FROM omop_vocabulary.concept
```

```
                    WHERE vocabulary_id = 'UCUM';",
    conn = conn)
```

```
head(ucum_omop_library1)
```

```
## # A tibble: 6 x 10
##   concept_id concept_name domain_id vocabulary_id concept_class_id
##        <dbl> <chr>        <chr>     <chr>         <chr>
## 1       8478 avidity ind~ Unit      UCUM          Unit
## 2       8479 centipoise   Unit      UCUM          Unit
## 3       8480 Ehrlich unit Unit      UCUM          Unit
## 4       8481 EV           Unit      UCUM          Unit
## 5       8482 pH           Unit      UCUM          Unit
## 6       8483 counts per ~ Unit      UCUM          Unit
## # ... with 5 more variables: standard_concept <chr>, concept_code <chr>,
## #   valid_start_date <date>, valid_end_date <date>, invalid_reason <chr>
```

```
uom_omop_library <- uom_library %>% bind_rows(ucum_omop_library1)
head(uom_omop_library)
```

```
## # A tibble: 6 x 15
##   value_set_name code_system logical_definit~ code  code_description concept_id
##   <chr>          <chr>       <chr>            <chr> <chr>                <dbl>
## 1 UnitsOfLength~ http://uni~ <NA>             pm    Picometer               NA
## 2 UnitsOfLength~ http://uni~ <NA>             nm    Nanometer               NA
## 3 UnitsOfLength~ http://uni~ <NA>             mm    Millimeter              NA
## 4 UnitsOfLength~ http://uni~ <NA>             cm    Centimeter              NA
## 5 UnitsOfLength~ http://uni~ <NA>             m     Meter                   NA
## 6 UnitsOfLength~ http://uni~ <NA>             ft-us Foot                    NA
## # ... with 9 more variables: concept_name <chr>, domain_id <chr>,
## #   vocabulary_id <chr>, concept_class_id <chr>, standard_concept <chr>,
## #   concept_code <chr>, valid_start_date <date>, valid_end_date <date>,
## #   invalid_reason <chr>
```

Write the data to a file if it does not already exist.

```
file <- file.path(getwd(), "data-raw", "unitsofmeasurement.csv")
if (!file.exists(file)) {
    write_csv(x = uom_omop_library, file = file)
}
```

SNOMED

The SNOMED subset of the library contains:

- Explicitly stated concepts
- Concepts that are descendants of a stated code
- Concepts that are ancestors of a stated code

To derive all concepts, the code in the `logical_definition` field is extracted based on the listed scenario above.

```
snomed_library <- value_sets_by_vocab$`SNOMED CT` %>% rubix::format_colnames() %>%
    distinct() %>% mutate_all(as.character) %>% extract(col = logical_definition,
```

```
    into = "descendants_of", regex = "includes codes descending from ([0-9]{1,})[^0-9]{1}.*$",
    remove = FALSE) %>% extract(col = logical_definition, into = "ancestors_of",
    regex = "excludes codes descending from ([0-9]{1,})[^0-9]{1}.*$",
    remove = FALSE) %>% mutate(all_codes = coalesce(code, ancestors_of,
    descendants_of)) %>% mutate_all(trimws)
head(snomed_library)
```

```
## # A tibble: 6 x 8
##   value_set_name code_system logical_definit~ ancestors_of descendants_of code
##   <chr>          <chr>       <chr>            <chr>        <chr>          <chr>
## 1 CancerBodyLoc~ SNOMED CT   includes codes ~ <NA>         123037004      <NA>
## 2 CancerDisease~ SNOMED CT   <NA>             <NA>         <NA>           3636~
## 3 CancerDisease~ SNOMED CT   <NA>             <NA>         <NA>           2524~
## 4 CancerDisease~ SNOMED CT   <NA>             <NA>         <NA>           7110~
## 5 CancerDisease~ SNOMED CT   <NA>             <NA>         <NA>           5880~
## 6 CancerDisease~ SNOMED CT   <NA>             <NA>         <NA>           2507~
## # ... with 2 more variables: code_description <chr>, all_codes <chr>
```

OMOP Concept table is joined to the mCode's SNOMED library by code.

```
snomed_omop_library1 <- chariot::join_on_concept_code(kind = "LEFT",
    data = snomed_library, column = "all_codes", where_in_concept_field = "vocabulary_id",
    where_in_concept_field_value = "SNOMED")
snomed_omop_library1 <- snomed_library %>% left_join(snomed_omop_library1,
    by = c("value_set_name", "ancestors_of", "descendants_of",
        "all_codes", "code_system", "logical_definition", "code",
        "code_description")) %>% distinct()
```

```
head(snomed_omop_library1)
```

```
## # A tibble: 6 x 18
##   value_set_name code_system logical_definit~ ancestors_of descendants_of code
##   <chr>          <chr>       <chr>            <chr>        <chr>          <chr>
## 1 CancerBodyLoc~ SNOMED CT   includes codes ~ <NA>         123037004      <NA>
## 2 CancerDisease~ SNOMED CT   <NA>             <NA>         <NA>           3636~
## 3 CancerDisease~ SNOMED CT   <NA>             <NA>         <NA>           2524~
## 4 CancerDisease~ SNOMED CT   <NA>             <NA>         <NA>           7110~
## 5 CancerDisease~ SNOMED CT   <NA>             <NA>         <NA>           5880~
## 6 CancerDisease~ SNOMED CT   <NA>             <NA>         <NA>           2507~
## # ... with 12 more variables: code_description <chr>, all_codes <chr>,
## #   concept_id <dbl>, concept_name <chr>, domain_id <chr>, vocabulary_id <chr>,
## #   concept_class_id <chr>, standard_concept <chr>, concept_code <chr>,
## #   valid_start_date <date>, valid_end_date <date>, invalid_reason <chr>
```

The resultset is then split into 3 based on whether the descendants (a) or ancestors (b) need to be derived, or if the concept is explicitly stated by code (c).

The descendants are derived for those mCode concepts that included descendants.

```
snomed_omop_library2a <- snomed_omop_library1 %>% filter(!is.na(descendants_of))
snomed_omop_library2a2 <- join_for_descendants(kind = "LEFT",
    data = snomed_omop_library2a, ancestor_id_column = "concept_id") %>%
    select(all_of(colnames(snomed_library)), starts_with("descendant_")) %>%
    rename_all(~str_remove_all(., pattern = "^descendant_"))
```

```
head(snomed_omop_library2a2)
```

```
##         value_set_name code_system
## 1 CancerBodyLocationVS   SNOMED CT
## 2 CancerBodyLocationVS   SNOMED CT
## 3 CancerBodyLocationVS   SNOMED CT
## 4 CancerBodyLocationVS   SNOMED CT
## 5 CancerBodyLocationVS   SNOMED CT
## 6 CancerBodyLocationVS   SNOMED CT
##                                       logical_definition ancestors_of
## 1 includes codes descending from 123037004 | Body Structure         <NA>
## 2 includes codes descending from 123037004 | Body Structure         <NA>
## 3 includes codes descending from 123037004 | Body Structure         <NA>
## 4 includes codes descending from 123037004 | Body Structure         <NA>
## 5 includes codes descending from 123037004 | Body Structure         <NA>
## 6 includes codes descending from 123037004 | Body Structure         <NA>
##   descendants_of code code_description all_codes concept_id
## 1      123037004 <NA>             <NA> 123037004    4048384
## 2      123037004 <NA>             <NA> 123037004    4002852
## 3      123037004 <NA>             <NA> 123037004   36717763
## 4      123037004 <NA>             <NA> 123037004    4230944
## 5      123037004 <NA>             <NA> 123037004   42605189
## 6      123037004 <NA>             <NA> 123037004    4097829
##                                         concept_name          domain_id
## 1                                    Body structure Spec Anatomic Site
## 2                                   Buccal embrasure Spec Anatomic Site
## 3                     Skin structure of left lower eyelid Spec Anatomic Site
## 4                                      Adenofibrosis        Observation
## 5 Intervertebral foramen of eighteenth thoracic vertebra Spec Anatomic Site
## 6                            Transitional cell carcinoma        Observation
##       vocabulary_id  concept_class_id standard_concept    concept_code
## 1            SNOMED    Body Structure                S       123037004
## 2            SNOMED    Body Structure                S       110326006
## 3            SNOMED    Body Structure                S       719884003
## 4            SNOMED Morph Abnormality                S        89115006
## 5 SNOMED Veterinary    Body Structure                S 336621000009107
## 6            SNOMED Morph Abnormality                S        27090000
##   valid_start_date valid_end_date invalid_reason
## 1       1970-01-01     2099-12-31           <NA>
## 2       1970-01-01     2099-12-31           <NA>
## 3       2017-01-31     2099-12-31           <NA>
## 4       1970-01-01     2099-12-31           <NA>
## 5       2014-01-31     2099-12-31           <NA>
## 6       1970-01-01     2099-12-31           <NA>
```

The ancestors are derived for those mCode concepts that excluded descendants.

```
snomed_omop_library2b <- snomed_omop_library1 %>% filter(!is.na(ancestors_of))
snomed_omop_library2b2 <- join_for_ancestors(kind = "LEFT", data = snomed_omop_library2b,
    descendant_id_column = "concept_id") %>% filter(min_levels_of_separation !=
    0) %>% select(all_of(colnames(snomed_library)), starts_with("ancestor_")) %>%
    rename_all(~str_remove_all(., pattern = "^ancestor_"))
```

```
head(snomed_omop_library2b2)
```

```
##                  value_set_name code_system
## 1 HistologyMorphologyBehaviorVS   SNOMED CT
```

```
## 2 HistologyMorphologyBehaviorVS   SNOMED CT
## 3 HistologyMorphologyBehaviorVS   SNOMED CT
## 4 HistologyMorphologyBehaviorVS   SNOMED CT
## 5 HistologyMorphologyBehaviorVS   SNOMED CT
## 6 HistologyMorphologyBehaviorVS   SNOMED CT
##
## 1 excludes codes descending from 399983006 | Papillary neoplasm, pancreatobiliary-type, with high gra
## 2 excludes codes descending from 399983006 | Papillary neoplasm, pancreatobiliary-type, with high gra
## 3 excludes codes descending from 399983006 | Papillary neoplasm, pancreatobiliary-type, with high gra
## 4 excludes codes descending from 399983006 | Papillary neoplasm, pancreatobiliary-type, with high gra
## 5 excludes codes descending from 399983006 | Papillary neoplasm, pancreatobiliary-type, with high gra
## 6 excludes codes descending from 399983006 | Papillary neoplasm, pancreatobiliary-type, with high gra
##   ancestors_of descendants_of code code_description all_codes concept_id
## 1    399983006           <NA> <NA>             <NA> 399983006    4019483
## 2    399983006           <NA> <NA>             <NA> 399983006    4030314
## 3    399983006           <NA> <NA>             <NA> 399983006    4042149
## 4    399983006           <NA> <NA>             <NA> 399983006    4043350
## 5    399983006           <NA> <NA>             <NA> 399983006    4048384
## 6    399983006           <NA> <NA>             <NA> 399983006    4133294
##                       concept_name          domain_id vocabulary_id
## 1     Adenoma AND/OR adenocarcinoma        Observation        SNOMED
## 2                          Neoplasm        Observation        SNOMED
## 3               Epithelial neoplasm        Observation        SNOMED
## 4 Morphologically altered structure        Observation        SNOMED
## 5                    Body structure Spec Anatomic Site        SNOMED
## 6                  In situ neoplasm        Observation        SNOMED
##    concept_class_id standard_concept concept_code valid_start_date
## 1 Morph Abnormality                S    115215004       1970-01-01
## 2 Morph Abnormality                S    108369006       1970-01-01
## 3 Morph Abnormality                S    118285006       1970-01-01
## 4 Morph Abnormality                S    118956008       1970-01-01
## 5    Body Structure                S    123037004       1970-01-01
## 6 Morph Abnormality                S    127569003       1970-01-01
##   valid_end_date invalid_reason
## 1     2099-12-31           <NA>
## 2     2099-12-31           <NA>
## 3     2099-12-31           <NA>
## 4     2099-12-31           <NA>
## 5     2099-12-31           <NA>
## 6     2099-12-31           <NA>
```

The concepts that are explicitly stated do not require any additional derivation.

```
snomed_omop_library2c <- snomed_omop_library1 %>% filter(is.na(descendants_of),
    is.na(ancestors_of))
head(snomed_omop_library2c)
```

```
## # A tibble: 6 x 18
##   value_set_name code_system logical_definit~ ancestors_of descendants_of code
##   <chr>          <chr>       <chr>            <chr>        <chr>          <chr>
## 1 CancerDisease~ SNOMED CT   <NA>             <NA>         <NA>           3636~
## 2 CancerDisease~ SNOMED CT   <NA>             <NA>         <NA>           2524~
## 3 CancerDisease~ SNOMED CT   <NA>             <NA>         <NA>           7110~
## 4 CancerDisease~ SNOMED CT   <NA>             <NA>         <NA>           5880~
## 5 CancerDisease~ SNOMED CT   <NA>             <NA>         <NA>           2507~
```

```
## 6 CancerDisease~ SNOMED CT   <NA>             <NA>         <NA>           3863~
## # ... with 12 more variables: code_description <chr>, all_codes <chr>,
## #   concept_id <dbl>, concept_name <chr>, domain_id <chr>, vocabulary_id <chr>,
## #   concept_class_id <chr>, standard_concept <chr>, concept_code <chr>,
## #   valid_start_date <date>, valid_end_date <date>, invalid_reason <chr>
```

All 3 subsets are recombined before writing to file.

```
snomed_omop_library <- bind_rows(snomed_omop_library2a2, snomed_omop_library2b2,
    snomed_omop_library2c)
head(snomed_omop_library)
```

```
##         value_set_name code_system
## 1 CancerBodyLocationVS   SNOMED CT
## 2 CancerBodyLocationVS   SNOMED CT
## 3 CancerBodyLocationVS   SNOMED CT
## 4 CancerBodyLocationVS   SNOMED CT
## 5 CancerBodyLocationVS   SNOMED CT
## 6 CancerBodyLocationVS   SNOMED CT
##                                       logical_definition ancestors_of
## 1 includes codes descending from 123037004 | Body Structure         <NA>
## 2 includes codes descending from 123037004 | Body Structure         <NA>
## 3 includes codes descending from 123037004 | Body Structure         <NA>
## 4 includes codes descending from 123037004 | Body Structure         <NA>
## 5 includes codes descending from 123037004 | Body Structure         <NA>
## 6 includes codes descending from 123037004 | Body Structure         <NA>
##   descendants_of code code_description all_codes concept_id
## 1      123037004 <NA>             <NA> 123037004    4048384
## 2      123037004 <NA>             <NA> 123037004    4002852
## 3      123037004 <NA>             <NA> 123037004   36717763
## 4      123037004 <NA>             <NA> 123037004    4230944
## 5      123037004 <NA>             <NA> 123037004   42605189
## 6      123037004 <NA>             <NA> 123037004    4097829
##                                          concept_name          domain_id
## 1                                     Body structure Spec Anatomic Site
## 2                                   Buccal embrasure Spec Anatomic Site
## 3                  Skin structure of left lower eyelid Spec Anatomic Site
## 4                                     Adenofibrosis        Observation
## 5 Intervertebral foramen of eighteenth thoracic vertebra Spec Anatomic Site
## 6                          Transitional cell carcinoma        Observation
##       vocabulary_id  concept_class_id standard_concept    concept_code
## 1            SNOMED    Body Structure                S       123037004
## 2            SNOMED    Body Structure                S       110326006
## 3            SNOMED    Body Structure                S       719884003
## 4            SNOMED Morph Abnormality                S        89115006
## 5 SNOMED Veterinary    Body Structure                S 336621000009107
## 6            SNOMED Morph Abnormality                S        27090000
##   valid_start_date valid_end_date invalid_reason
## 1       1970-01-01     2099-12-31           <NA>
## 2       1970-01-01     2099-12-31           <NA>
## 3       2017-01-31     2099-12-31           <NA>
## 4       1970-01-01     2099-12-31           <NA>
## 5       2014-01-31     2099-12-31           <NA>
## 6       1970-01-01     2099-12-31           <NA>
```

```
file <- file.path(getwd(), "data-raw", "snomed.csv")
if (!file.exists(file)) {
    write_csv(x = snomed_omop_library, file = file)
}
```

LOINC

```
loinc_library <- value_sets_by_vocab$LOINC %>% rubix::format_colnames() %>%
    mutate_all(as.character) %>% mutate_all(trimws) %>% distinct()
head(loinc_library)
```

```
## # A tibble: 6 x 5
##   value_set_name  code_system logical_definiti~ code  code_description
##   <chr>           <chr>       <chr>             <chr> <chr>
## 1 TumorMarkerTes~ LOINC       <NA>              1695~ 5-Hydroxyindoleacetate [M~
## 2 TumorMarkerTes~ LOINC       <NA>              3120~ 5-Hydroxyindoleacetate [M~
## 3 TumorMarkerTes~ LOINC       <NA>              1692~ 5-Hydroxyindoleacetate [M~
## 4 TumorMarkerTes~ LOINC       <NA>              1693~ 5-Hydroxyindoleacetate [M~
## 5 TumorMarkerTes~ LOINC       <NA>              1694~ 5-Hydroxyindoleacetate [M~
## 6 TumorMarkerTes~ LOINC       <NA>              7282~ 5-Hydroxyindoleacetate [M~
```

```
loinc_omop_library <- chariot::join_on_concept_code(kind = "LEFT",
    data = loinc_library, column = "code", where_in_concept_field = "vocabulary_id",
    where_in_concept_field_value = "LOINC")
loinc_omop_library <- loinc_library %>% left_join(loinc_omop_library,
    by = c("value_set_name", "code_system", "logical_definition",
        "code", "code_description")) %>% distinct()
```

```
head(loinc_omop_library)
```

```
## # A tibble: 6 x 15
##   value_set_name code_system logical_definit~ code  code_description concept_id
##   <chr>          <chr>       <chr>            <chr> <chr>                 <dbl>
## 1 TumorMarkerTe~ LOINC       <NA>             1695~ 5-Hydroxyindole~    3005148
## 2 TumorMarkerTe~ LOINC       <NA>             3120~ 5-Hydroxyindole~    3023028
## 3 TumorMarkerTe~ LOINC       <NA>             1692~ 5-Hydroxyindole~    3014670
## 4 TumorMarkerTe~ LOINC       <NA>             1693~ 5-Hydroxyindole~    3021106
## 5 TumorMarkerTe~ LOINC       <NA>             1694~ 5-Hydroxyindole~    3021385
## 6 TumorMarkerTe~ LOINC       <NA>             7282~ 5-Hydroxyindole~   43055680
## # ... with 9 more variables: concept_name <chr>, domain_id <chr>,
## #   vocabulary_id <chr>, concept_class_id <chr>, standard_concept <chr>,
## #   concept_code <chr>, valid_start_date <date>, valid_end_date <date>,
## #   invalid_reason <chr>
```

```
file <- file.path(getwd(), "data-raw", "loinc.csv")
if (!file.exists(file)) {
    write_csv(x = loinc_omop_library, file = file)
}
```

ICD-10 CM

```
icd10cm_library <- value_sets_by_vocab$`ICD-10-CM` %>% rubix::format_colnames() %>%
    mutate_all(as.character) %>% mutate_all(trimws) %>% mutate(code = str_replace_all(string = code,
    pattern = "(^[A-Z]{1}[0-9A-Z]{2})([0-9A-Z]{1}.*$)", replacement = "\\1.\\2")) %>%
    distinct()
head(icd10cm_library)
```

```
## # A tibble: 6 x 5
##   value_set_name  code_system logical_definiti~ code  code_description
##   <chr>           <chr>       <chr>             <chr> <chr>
## 1 CancerDisorder~ ICD-10-CM   <NA>              C00.0 Malignant neoplasm of ext~
## 2 CancerDisorder~ ICD-10-CM   <NA>              C00.1 Malignant neoplasm of ext~
## 3 CancerDisorder~ ICD-10-CM   <NA>              C00.2 Malignant neoplasm of ext~
## 4 CancerDisorder~ ICD-10-CM   <NA>              C00.3 Malignant neoplasm of upp~
## 5 CancerDisorder~ ICD-10-CM   <NA>              C00.4 Malignant neoplasm of low~
## 6 CancerDisorder~ ICD-10-CM   <NA>              C00.5 Malignant neoplasm of lip~
```

```
icd10cm_omop_library <- chariot::join_on_concept_code(kind = "LEFT",
    data = icd10cm_library, column = "code", where_in_concept_field = "vocabulary_id",
    where_in_concept_field_value = "ICD10CM")
icd10cm_omop_library <- icd10cm_library %>% left_join(icd10cm_omop_library,
    by = c("value_set_name", "code_system", "logical_definition",
        "code", "code_description")) %>% distinct()
```

```
head(icd10cm_omop_library)
```

```
## # A tibble: 6 x 15
##   value_set_name code_system logical_definit~ code  code_description concept_id
##   <chr>          <chr>       <chr>            <chr> <chr>                 <dbl>
## 1 CancerDisorde~ ICD-10-CM   <NA>             C00.0 Malignant neopl~   35206047
## 2 CancerDisorde~ ICD-10-CM   <NA>             C00.1 Malignant neopl~   35206048
## 3 CancerDisorde~ ICD-10-CM   <NA>             C00.2 Malignant neopl~   35206049
## 4 CancerDisorde~ ICD-10-CM   <NA>             C00.3 Malignant neopl~   35206050
## 5 CancerDisorde~ ICD-10-CM   <NA>             C00.4 Malignant neopl~   35206051
## 6 CancerDisorde~ ICD-10-CM   <NA>             C00.5 Malignant neopl~   35206052
## # ... with 9 more variables: concept_name <chr>, domain_id <chr>,
## #   vocabulary_id <chr>, concept_class_id <chr>, standard_concept <chr>,
## #   concept_code <chr>, valid_start_date <date>, valid_end_date <date>,
## #   invalid_reason <chr>
```

```
file <- file.path(getwd(), "data-raw", "icd10cm.csv")
if (!file.exists(file)) {
    write_csv(x = icd10cm_omop_library, file = file)
}
```

```
chariot::dcAthena(conn = conn)
```