**TOPOGRAPHY**

1. All 330 parsed SEER codes are present
2. They have to be recoded to include dot. E.g. ‘C80.9’ instead of ‘C809’
3. Group SEER codes are not present.
4. 79 NCI codes are not in Mike’s parsed file:
	1. 2-digit codes that represent upper concept like ‘LIP’ or ‘C03’ (‘GUM’) . They are not explicitly listed in the original SEER file. However, they are in the text:

SITE\_CODE ‘C030-C031,C039-C041,C048-C052,C058-C062,C068-C069’; SITE\_TEXT ‘GUM, FLOOR OF MOUTH, & OTHER MOUTH’.

They need to be added to the parsed file.

* 1. The remaining codes are groups of codes (e.g. C69-C72), total 9.



1. 3 of the NCI codes are not unique

****

**HISTOLOGY**

1. Total ICDO Histology codes in NCI Metathesaurus = 1,135
2. 17 SEER ICDO Histology/Behavior codes are not in NCI Metathesaurus

These 17 missing SEER codes account for 388 SEER Histology/Topology pairs

1. There are 431 ICDO codes in NCI Metathesaurus that are not present in SEER. ****

**COMBINATIONS**

1. There are 833 unique ICDO histology/topography combinations in NCI Metathesaurus that are linked directly.

|  |  |  |  |
| --- | --- | --- | --- |
| 8014/3 | Large cell carcinoma with rhabdoid phenotype | C34.9 | Lung, NOS |
| 8045/3 | Combined small cell-squamous cell carcinoma | C34.9 | Lung, NOS |
| 8077/0 | Esophageal squamous intraepithelial neoplasia (dysplasia), low grade | C15 | ESOPHAGUS |
| 8077/0 | Esophageal squamous intraepithelial neoplasia (dysplasia), low grade | C15.9 | Esophagus, NOS |

1. There are 883 unique ICDO combinations that are linked via NCIt (precoordinated) codes.
2. There are 579 unique precoordinated NCIt codes:

269 NCIt codes correspond to unique combinations of ICDO codes

310 to multiple (2-4).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| C6876 | Large Cell Lung Carcinoma with Rhabdoid Phenotype | 8014/3 | Large cell carcinoma with rhabdoid phenotype | C34.9 | Lung, NOS |
| C9137 | Combined Small and Large Cell Lung Cancer | 8045/3 | Mixed small cell carcinoma | C34.9 | Lung, NOS |
| C27427 | Low Grade Esophageal Squamous Intraepithelial Neoplasia | 8077/0 | Cervical intraepithelial neoplasia, low grade | C15 | ESOPHAGUS |
| C27427 | Low Grade Esophageal Squamous Intraepithelial Neoplasia | 8077/0 | Cervical intraepithelial neoplasia, low grade | C15.9 | Esophagus, NOS |
| C4630 | Low-Grade Cervix Intraepithelial Neoplasia | 8077/0 | Cervical intraepithelial neoplasia, low grade | C53 | CERVIX UTERI |
| C4630 | Low-Grade Cervix Intraepithelial Neoplasia | 8077/0 | Cervical intraepithelial neoplasia, low grade | C53.9 | Cervix, NOS |
| C4630 | Low-Grade Cervix Intraepithelial Neoplasia | 8077/0 | Cervical intraepithelial neoplasia, low grade | C55 | UTERUS, NOS |
| C4630 | Low-Grade Cervix Intraepithelial Neoplasia | 8077/0 | Cervical intraepithelial neoplasia, low grade | C55.9 | Uterus, NOS |

1. Only 268 out of 49,831 SEER pairs that are covered by NCIt pre-coordinated codes (this is w/o cleaning)

Examples of missing codes:

|  |  |  |  |
| --- | --- | --- | --- |
| C70.0 | 8000/0 | Cerebral meninges | Neoplasm, benign |
| C70.1 | 8000/0 | Spinal meninges | Neoplasm, benign |
| C70.9 | 8000/0 | Dura mater, NOS | Neoplasm, benign |
| C71.0 | 8000/0 | Basal ganglia | Neoplasm, benign |
| C71.1 | 8000/0 | Frontal lobe | Neoplasm, benign |

APPENDIX. Queries

**--TOPOGRAPHY**

**-- Total Topography codes in NCI = 409**

**select** **count**(**distinct** code)

**from** mrconso m1

**where** m1.sab='ICDO'

**AND** **substr**(code,1,1) = 'C';

**--All parsed SEER codes are there**

**SELECT** \*

**FROM** ICDO\_HISTOLOGY\_TOPOGRAPHY

**WHERE** TOPOGRAPHY **NOT** **IN**

(**select** CODE

**from** mrconso m1

**where** m1.sab='ICDO'

**AND** **substr**(code,1,1) = 'C'

);

**-- No unparsed SEER codes are there**

**SELECT** **DISTINCT** SITE\_CODE, SITE\_TEXT

**FROM** ICDO\_HISTOLOGY\_TOPOGRAPHY\_ORIGINAL

**WHERE** SITE\_CODE **IN**

(**select** CODE

**from** mrconso m1

**where** m1.sab='ICDO'

**AND** **substr**(code,1,1) = 'C'

)

**ORDER** **BY** SITE\_CODE;

**-- 79 NCI codes are not in Mike’s parsed file**

**select** **count**(**distinct** code)

**from** mrconso m1

**where** m1.sab='ICDO'

**AND** **substr**(code,1,1) = 'C'

**AND** CODE **NOT** **IN**

(**SELECT** TOPOGRAPHY

**FROM** ICDO\_HISTOLOGY\_TOPOGRAPHY

);

**-- Most of them are 2-digit codes that represent upper concept like ‘LIP’ or ‘C03’ (‘GUM’) . They can be added by using substrings (subquery in the query below) or better regular expressions from site\_code. The remaining codes extracted by the query below are groups of codes (e.g. C69-C72), total 9.**

**select** **distinct** code, str

**from** mrconso m1

**where** m1.sab='ICDO'

**AND** **substr**(code,1,1) = 'C'

**AND** CODE **NOT** **IN**

(**SELECT** TOPOGRAPHY

**FROM** ICDO\_HISTOLOGY\_TOPOGRAPHY

**UNION** **ALL**

**SELECT** **DISTINCT** **substr**(site\_code, 1, 3)

**FROM** ICDO\_HISTOLOGY\_TOPOGRAPHY\_ORIGINAL

**UNION** **ALL**

**SELECT** **DISTINCT** **substr**(**replace**(site\_code,' ',''), **Instr**(**replace**(site\_code,' ',''),',',1,1)+1, 3)

**FROM** ICDO\_HISTOLOGY\_TOPOGRAPHY\_ORIGINAL

**UNION** **ALL**

**SELECT** **DISTINCT** **substr**(**replace**(site\_code,' ',''), **Instr**(**replace**(site\_code,' ',''),',',1,2)+1, 3)

**FROM** ICDO\_HISTOLOGY\_TOPOGRAPHY\_ORIGINAL

**UNION** **ALL**

**SELECT** **DISTINCT** **substr**(**replace**(site\_code,' ',''), **Instr**(**replace**(site\_code,' ',''),',',1,3)+1, 3)

**FROM** ICDO\_HISTOLOGY\_TOPOGRAPHY\_ORIGINAL

**UNION** **ALL**

**SELECT** **DISTINCT** **substr**(**replace**(site\_code,' ',''), **Instr**(**replace**(site\_code,' ',''),',',1,4)+1, 3)

**FROM** ICDO\_HISTOLOGY\_TOPOGRAPHY\_ORIGINAL

)

;

**-- 3 of the NCI codes are not unique**

**select** code, str, **count**(\*)

**from** mrconso m1

**where** m1.sab='ICDO'

**AND** **substr**(code,1,1) = 'C'

**GROUP** **BY** code, str

**HAVING** **count**(\*) > 1;

**Select** \* **from** mrconso

**Where** (code, str) **in** (

**select** code, str

**from** mrconso m1

**where** m1.sab='ICDO'

**AND** **substr**(code,1,1) = 'C'

**GROUP** **BY** code, str

**HAVING** **count**(\*) > 1);

**HISTOLOGY**

**-- Total Histology codes in NCI = 1,135**

**select** **count**(**distinct** code)

**from** mrconso m1

**where** m1.sab='ICDO'

**AND** **substr**(code,1,1) **NOT** **IN** ('C','T','M'));

**-- 17 SEER Histology/Behavior codes are not in NCI**

**SELECT** **DISTINCT** HISTOLOGY\_BEHAVIOR\_CODE, HISTOLOGY\_BEHAVIOR\_TEXT

**FROM** ICDO\_HISTOLOGY\_TOPOGRAPHY\_ORIGINAL

**WHERE** HISTOLOGY\_BEHAVIOR\_CODE **NOT** **IN**

(**SELECT** **distinct** code

**from** mrconso m1

**where** m1.sab='ICDO'

**AND** **substr**(code,1,1) **NOT** **IN** ('C','T','M'));

**-- Those 17 missing SEER codes account for 388 SEER Histology/Topology pairs**

**SELECT** **count**(\*)

**FROM** ICDO\_HISTOLOGY\_TOPOGRAPHY

**WHERE** HISTOLOGY **NOT** **IN**

(**SELECT** **distinct** code

**from** mrconso m1

**where** m1.sab='ICDO'

**AND** **substr**(code,1,1) **NOT** **IN** ('C','T','M'));

**-- There are 431 NCI codes that are not present in SEER**

**SELECT** **count** (**distinct** code)

**from** mrconso m1

**where** m1.sab='ICDO'

**AND** **substr**(code,1,1) **NOT** **IN** ('C','T','M')

**AND** code **NOT** **IN**

(**SELECT** **DISTINCT** HISTOLOGY

**FROM** ICDO\_HISTOLOGY\_TOPOGRAPHY

)

**ORDER** **BY** code;

**SELECT** **DISTINCT** code, **first\_value**(str) **OVER** (**partition** **by** code)

**from** mrconso m1

**where** m1.sab='ICDO'

**AND** **substr**(code,1,1) **NOT** **IN** ('C','T','M')

**AND** code **NOT** **IN**

(**SELECT** **DISTINCT** HISTOLOGY

**FROM** ICDO\_HISTOLOGY\_TOPOGRAPHY

)

;

**COMBINATIONS**

**-- There are 833 unique ICDO histology/topography combinations in NCI Metathesaurus that are linked directly.**

**SELECT** **count**(\*) **FROM** (

**select** **distinct**

 m1.code **AS** morphology\_code, **first\_value**(m1.str) **over** (**partition** **by** m1.code) **as** morphology,

 m3.code **AS** anatomy\_code, **first\_value**(m3.str) **over** (**partition** **by** m3.code) **as** anatomy

**from** mrconso m1

**join** mrrel **on** m1.cui=cui1 **and** m1.sab='ICDO'

**join** mrconso m3 **on** m3.cui=cui2 **and** m3.sab='ICDO' **AND** rela='Is\_Primary\_Anatomic\_Site\_Of\_Disease'

**order** **by** 1)

;

**-- There are 883 unique ICDO combinations that are linked via NCIt (precoordinated) codes.**

**SELECT** **count**(\*) **FROM** (

**select** **distinct**

 m1.code **AS** morphology\_code, **first\_value**(m1.str) **over** (**partition** **by** m1.code) **as** morphology,

 m3.code **AS** anatomy\_code, **first\_value**(m3.str) **over** (**partition** **by** m3.code) **as** anatomy

**from** mrconso m1

-- connection between NCI and ICDO through common CUI

**join** mrconso m2 **on** m2.cui=m1.cui **and** m1.sab='ICDO' **and** m2.sab='NCI'

-- connection between NCI and anatomical site through special relationship

**join** mrrel **on** m2.cui=cui1

**join** mrconso m3 **on** m3.cui=cui2 **and** m3.sab='ICDO' **AND** rela='Is\_Primary\_Anatomic\_Site\_Of\_Disease'

**order** **by** 1);

**-- There are 579 precoordinated NCIt codes corresponding to those 883 unique ICDO combinations**

**SELECT** **count**(\*)

**FROM**

(**select** **distinct**

 m2.code **AS** NCI\_code, **first\_value**(m2.str) **over** (**partition** **by** m2.code) **as** NCI

**from** mrconso m1

-- connection between NCI and ICDO through common CUI

**join** mrconso m2 **on** m2.cui=m1.cui **and** m1.sab='ICDO' **and** m2.sab='NCI'

-- connection between NCI and anatomical site through special relationship

**join** mrrel **on** m2.cui=cui1

**join** mrconso m3 **on** m3.cui=cui2 **and** m3.sab='ICDO' **AND** rela='Is\_Primary\_Anatomic\_Site\_Of\_Disease'

**order** **by** 1);

**-- 269 NCIt codes correspond to unique combinations of ICDO codes**

**310 to multiple (2-4)**

**SELECT** **count**(\*) **FROM** (

**SELECT** NCI\_code, NCI, **count**(\*) **FROM** (

**select** **distinct**

 m1.code **AS** morphology\_code, **first\_value**(m1.str) **over** (**partition** **by** m1.code) **as** morphology,

 m2.code **AS** NCI\_code, **first\_value**(m2.str) **over** (**partition** **by** m2.code) **as** NCI,

 m3.code **AS** anatomy\_code, **first\_value**(m3.str) **over** (**partition** **by** m3.code) **as** anatomy

**from** mrconso m1

-- connection between NCI and ICDO through common CUI

**join** mrconso m2 **on** m2.cui=m1.cui **and** m1.sab='ICDO' **and** m2.sab='NCI'

-- connection between NCI and anatomical site through special relationship

**join** mrrel **on** m2.cui=cui1

**join** mrconso m3 **on** m3.cui=cui2 **and** m3.sab='ICDO' **AND** rela='Is\_Primary\_Anatomic\_Site\_Of\_Disease'

**order** **by** 1)

**GROUP** **BY** NCI\_code, NCI

**HAVING** **count**(\*) > 1);

**-- There are only 268 out of 49,831 SEER pairs that are covered by NCIt pre-coordinated codes (this is w/o cleaning)**

**SELECT** **count**(\*) **FROM**

(

**SELECT** **distinct** topography, histology

**FROM** icdo\_histology\_topography

**WHERE** (histology, topography) **IN**

(

**select** **distinct**

 m1.code **AS** morphology\_code,

 m3.code **AS** anatomy\_code

**from** mrconso m1

-- connection between NCI and ICDO through common CUI

**join** mrconso m2 **on** m2.cui=m1.cui **and** m1.sab='ICDO' **and** m2.sab='NCI'

-- connection between NCI and anatomical site through special relationship

**join** mrrel **on** m2.cui=cui1

**join** mrconso m3 **on** m3.cui=cui2 **and** m3.sab='ICDO' **AND** rela='Is\_Primary\_Anatomic\_Site\_Of\_Disease'

)

);

**-- Example of missing SEER combinations**

**select** **DISTINCT**

m1.cui, m1.code **AS** morphology\_code, **first\_value**(m1.str) **over** (**partition** **by** m1.code) **as** morphology,

 m2.code **AS** NCI\_code, **first\_value**(m2.str) **over** (**partition** **by** m2.code) **as** NCI,

 rela **AS** LINK

**from** mrconso m1

**join** mrconso m2 **on** m2.cui=m1.cui **and** m1.sab='ICDO' **and** m2.sab='NCI'

**LEFT** **join** mrrel **on** m2.cui=cui1 **AND** rela='Is\_Primary\_Anatomic\_Site\_Of\_Disease'

**WHERE** m1.code = '8000/0';