Using the NAACCR Cancer Registry in i2b2 with HERON ETL

Dan Connolly
Biomedical Informatics Software Engineer
Division of Medical Informatics

KU Medical Center
The University of Kansas
KUMC Medical Informatics

- Russ **Waitman**, Director of Medical Informatics
- Software Engineers: Dan **Connolly**, Nathan **Graham**, Bhargav Adagarla, Matt **Hoag**, Mike Prittie, Lav Patel, Nazma Kotcherla
- Analysts, Honest Brokers: Tamara **McMahon**, Sravani Chandaka, Li Huang, Rachel Gyore, Maren Wennberg, Vince **Leonardo**
- Project Management: Steve Fennel, Brittany Zschoche, Hillary Sandoval
Kansas Cancer Registry (KCR)

- > 50,000 cancer records
- Population-based source of incidence info
- Survival, subpopulations
- Demographics
- Clinical Information
- Vital Status

The Incidence of Breast Cancer among Disabled Kansans with Medicare
Rogers, Austin R; Lai, Sue-Min; Keighley, John; Jungk, Jessica
URI: http://hdl.handle.net/2271/1364
Date: 2015-08-05
NAACCR Data Interchange Standard

Sites submit NAACCR format files to central (state) registries.

Great Plains Collaborative

1. The University of Kansas Medical Center (KUMC)
2. Children's Mercy Hospital CMH
3. Indiana / Regenstrief (IU)
4. University of Iowa Healthcare (UIOWA)
5. The University of Wisconsin-Madison (WISC)
6. The Medical College of Wisconsin (MCW)
7. Marshfield Clinic (Wisconsin) (MCRF)
8. The University of Minnesota Academic Health Center (UMN)
9. The University of Missouri (MU)
10. The University of Nebraska Medical Center (UNMC)
11. The University of Texas Health Sciences Center at San Antonio (UTHSCSA)
12. The University of Texas Southwestern Medical Center (UTSW)
GPC Breast Cancer Survey: 8 sites

1. The University of Kansas Medical Center (KUMC)
2. Children’s Mercy Hospital CMH
3. Indiana / Regenstrief (IU)
4. University of Iowa Healthcare (UIOWA)
5. The University of Wisconsin-Madison (WISC)
6. The Medical College of Wisconsin (MCW)
7. Marshfield Clinic (Wisconsin) (MCRF)
8. The University of Minnesota Academic Health Center (UMN)
9. The University of Missouri (MU)
10. The University of Nebraska Medical Center (UNMC)
11. The University of Texas Health Sciences Center at San Antonio (UTHSCSA)
12. The University of Texas Southwestern Medical Center (UTSW)
Share Thoughts on Breast Cancer
Study Principal Investigators

**Coordinating Site**
University of Iowa Holden Comprehensive Cancer Center:

❖ Elizabeth Chrischilles, PhD
❖ Ingrid Lizarraga, MBBS

**Participating Sites**
Marshfield Clinic:
- Robert Greenlee, PhD, MPH
- Adedayo A. Onitilo, MD, PhD, MSCR, FACP

Medical College of Wisconsin:
- Joan Neuner, MD, MPH

University of Kansas Medical Center:
- Jennifer Klemp, PhD, MPH
- Priyanka Sharma, MD

University of Minnesota:
- Anne Blaes, MD

University of Nebraska Medical Center:
- Ann Berger, PhD

University of Texas Southwestern Medical Center:
- Barbara Haley, MD

University of Texas San Antonio Medical Center:
- Amelie G. Ramirez, Ph.D., MPH

University of Wisconsin Carbone Cancer Center:
- Amy Trentham-Dietz, PhD
- Lee Gravatt Wilke, MD, FACS
Survey Prep

From a background cohort of 1000s per site, we* select a survey cohort of 100s.

*joined by Wendy He from KUMC Biostatistics
Survey Data Integration

For subjects that consent, we link registry and EHR data into a limited data set.
Breast Cancer Background Cohort

Inclusion Criteria

- Site: Breast
- Class of case: Analytic
- ...
# Background Cohort: Variables

<table>
<thead>
<tr>
<th>Path</th>
<th>Concept/Term</th>
<th>Other Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer Cases</td>
<td>101 Cancer Identification</td>
<td>0440 Grade</td>
</tr>
<tr>
<td>Cancer Cases</td>
<td>SEER Site Summary</td>
<td>Breast</td>
</tr>
<tr>
<td>Cancer Cases</td>
<td>101 Cancer Identification</td>
<td>0490 Diagnostic Confirmation</td>
</tr>
<tr>
<td>Cancer Cases</td>
<td>101 Cancer Identification</td>
<td>0521 Morph--Type&amp;Behav ICD-O-3</td>
</tr>
<tr>
<td>Cancer Cases</td>
<td>04 Follow-up/Recurrence/Death</td>
<td>1750 Date of Last Contact</td>
</tr>
<tr>
<td>Cancer Cases</td>
<td>04 Follow-up/Recurrence/Death</td>
<td>1760 Vital Status</td>
</tr>
<tr>
<td>Cancer Cases</td>
<td>04 Follow-up/Recurrence/Death</td>
<td>1860 Recurrence Date--1st</td>
</tr>
<tr>
<td>Cancer Cases</td>
<td>02 Demographic</td>
<td>0190 Spanish/Hispanic Origin</td>
</tr>
<tr>
<td>Cancer Cases</td>
<td>11 Stage/Prognostic Factors</td>
<td>3430 Derived AJCC-7 Stage Grp</td>
</tr>
<tr>
<td>Cancer Cases</td>
<td>02 Demographic</td>
<td>0160 Race 1</td>
</tr>
</tbody>
</table>

Additional variables include:
- Cancer Cases \02 Demographic: 0161 Race 2
- Cancer Cases \02 Demographic: 0162 Race 3
- Cancer Cases \02 Demographic: 0163 Race 4
- Cancer Cases \02 Demographic: 0170 Race Coding Sys--Current
- Cancer Cases \02 Demographic: 0180 Race Coding Sys--Original
- Cancer Cases \03 Cancer Identification: 0300 Sequence Number--Central
- Cancer Cases \03 Cancer Identification: 0410 Lateliness
- Cancer Cases \11 Stage/Prognostic Factors: 0820 Regional Nodes Positive
- Cancer Cases \11 Stage/Prognostic Factors: 0823 Regional Nodes Examined
- Cancer Cases \11 Stage/Prognostic Factors: 0868 CS Site-Specific Factor15
- Cancer Cases \11 Stage/Prognostic Factors: 0876 CS Site-Specific Factor2
- Cancer Cases \11 Stage/Prognostic Factors: 0877 CS Site-Specific Factor23
- Cancer Cases \11 Stage/Prognostic Factors: 0880 CS Site-Specific Factor 1
- Cancer Cases \11 Stage/Prognostic Factors: 0890 CS Site-Specific Factor 2
- Cancer Cases \11 Stage/Prognostic Factors: 0940 Derived AJCC-6 T
- Cancer Cases \11 Stage/Prognostic Factors: 0920 Derived SS2000
- Cancer Cases \11 Stage/Prognostic Factors: 1400 Derived AJCC-7 T
- Cancer Cases \04 Follow-up/Recurrence/Death: 1861 Recurrence Date--1st Flag
- Cancer Cases \03 Cancer Identification: 0390 Date of Diagnosis
- Cancer Cases \02 Demographic: 0220 Sex
- Cancer Cases \02 Demographic: 0184 Race 5
- Cancer Cases \03 Cancer Identification: 0490 Primary Site
- Cancer Cases \11 Stage/Prognostic Factors: 1200 Derived AJCC-6 Stage Grp
- Cancer Cases \11 Stage/Prognostic Factors: 1350 CS Metastases at DX
- Cancer Cases \11 Stage/Prognostic Factors: 0840 CS Met Eval
- Cancer Cases \06 Hospital-Specific: 0610 Class of Case
- Cancer Cases \06 Hospital-Specific: 0670 RX Hemp--Surg Prim Site
- Cancer Cases \02 Cancer Identification: 0400 Primary Site
- Cancer Cases \06 Hospital-Specific: 0610 Class of Case
- Cancer Cases \06 Hospital-Specific: 0540 Date of Birth
NAACCR Record: 100s of Items

Original approach:

- 2011 bid for NCI Designation
- Clues from:
  - Kimmel Cancer Center in Philadelphia
  - Group Health Cooperative in Seattle
# Fixed Width Fields

## Date of Birth

**Alternate Name**: Birth Date(SEER/CoC/CCCR)

<table>
<thead>
<tr>
<th>Item #</th>
<th>Length</th>
<th>Source of Standard</th>
<th>Year Implemented</th>
<th>Version Implemented</th>
<th>Year Retired</th>
<th>Version Retired</th>
<th>Column #</th>
</tr>
</thead>
<tbody>
<tr>
<td>240</td>
<td>8</td>
<td>SEER/CoC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>196 - 203</td>
</tr>
</tbody>
</table>

**Description**

Date of birth of the patient. See Chapter X for date format. If age at diagnosis and year of diagnosis are known, but year of birth is unknown, then year of birth should be calculated and so coded. Only the year should be entered, left-justified. Estimate date of birth when information is not available. It is better to estimate than to leave birth date unknown.
LOAD DATA
TRUNCATE
INTO TABLE "NAACR"."EXTRACT_INCR" ( "Record Type" position(1:1) CHAR,
"Registry Type" position(2:2) CHAR,
...
"Sex" position(192:192) CHAR,
"Age at Diagnosis" position(193:195) CHAR,
"Date of Birth" position(196:203) CHAR,
...
NAACCR Items Grouped by Section

### CHAPTER VII:

**RECORD LAYOUT TABLE (COLUMN # ORDER)**

The following table presents Version 15 of the NAACCR record layout. The table has column number, length, item number, item name, section, and note fields. Differences from Version 14 are marked “Revised” or “New” in the “Note” column of the table and highlighted in the body of the table. Text of previous versions is revealed by hovering over the highlighted cell. Revised and new items are summarized in Appendix F. Please note that “Retired” items are not reflected in this table.

<table>
<thead>
<tr>
<th>Column #</th>
<th>Length</th>
<th>Item #</th>
<th>Item Name</th>
<th>Section</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 1</td>
<td>1</td>
<td>16</td>
<td>Record Type</td>
<td>Record ID</td>
<td></td>
</tr>
<tr>
<td>2 - 2</td>
<td>1</td>
<td>30</td>
<td>Registry Type</td>
<td>Record ID</td>
<td></td>
</tr>
<tr>
<td>3 - 16</td>
<td>14</td>
<td>32</td>
<td>Reserved 00</td>
<td>Record ID</td>
<td></td>
</tr>
<tr>
<td>17 - 19</td>
<td>3</td>
<td>50</td>
<td>NAACCR Record Version</td>
<td>Record ID</td>
<td></td>
</tr>
<tr>
<td>20 - 29</td>
<td>10</td>
<td>45</td>
<td>NPI–Registry ID</td>
<td>Record ID</td>
<td></td>
</tr>
<tr>
<td>30 - 39</td>
<td>10</td>
<td>46</td>
<td>Registry ID</td>
<td>Record ID</td>
<td></td>
</tr>
<tr>
<td>40 - 41</td>
<td>2</td>
<td>98</td>
<td>Tumor Record Number</td>
<td>Record ID</td>
<td></td>
</tr>
<tr>
<td>42 - 49</td>
<td>8</td>
<td>26</td>
<td>Patient ID Number</td>
<td>Record ID</td>
<td></td>
</tr>
<tr>
<td>50 - 57</td>
<td>8</td>
<td>21</td>
<td>Patient System ID-Hosp</td>
<td>Record ID</td>
<td></td>
</tr>
<tr>
<td>58 - 94</td>
<td>37</td>
<td>370</td>
<td>Reserved 01</td>
<td>Record ID</td>
<td></td>
</tr>
<tr>
<td>95 - 144</td>
<td>50</td>
<td>70</td>
<td>Addr at DX–City</td>
<td>Demographic</td>
<td></td>
</tr>
<tr>
<td>145 - 146</td>
<td>2</td>
<td>98</td>
<td>Addr at DX–State</td>
<td>Demographic</td>
<td></td>
</tr>
<tr>
<td>147 - 155</td>
<td>9</td>
<td>169</td>
<td>Addr at DX–Postal Code</td>
<td>Demographic</td>
<td></td>
</tr>
</tbody>
</table>
Each Section is a Folder in i2b2
Items Inside Section Folders

- Cancer Cases (Lay-Person Hierarchy) [1,307,060 facts; 90,542 patients] - 90542
- Cancer Cases (NAACCR Hierarchy) [20,330,226 facts; 90,587 patients]
  - 01 Cancer Identification [1,838,230 facts; 90,542 patients] - 90542
  - 0380 Sequence Number-Central [97,348 facts; 90,542 patients] - 90542
  - 0390 Date of Diagnosis [97,348 facts; 90,542 patients] - 90542
  - 0391 Date of Diagnosis Flag
  - 0400 Primary Site [97,348 facts; 90,542 patients] - 90542
  - 0410 Laterality [97,348 facts; 90,542 patients] - 90542
  - 0419 Morph--Type&Behav ICD-O-2 [46,989 facts; 43,677 patients] - 43677
  - 0420 Histology (92-00) ICD-O-2 [47,153 facts; 43,790 patients] - 43790
  - 0430 Behavior (92-00) ICD-O-2 [47,153 facts; 43,790 patients] - 43790
  - 0439 Date of Multiple Tumors Flag [18,165 facts; 17,491 patients] - 17491
  - 0440 Grade [97,339 facts; 90,535 patients] - 90535
  - 0441 Grade Path Value [793 facts; 775 patients] - 775
  - 0442 Ambiguous Terminology DX [65,235 facts; 61,014 patients] - 61014
  - 0443 Date of Conclusive DX [32 facts; 32 patients] - 32
  - 0444 Multiple Rpt as One Prim [65,686 facts; 61,446 patients] - 61446
  - 0445 Date of Multiple Tumors [1,694 facts; 1,672 patients] - 1672
  - 0446 Multiplicity Counter [65,682 facts; 61,442 patients] - 61442
  - 0448 Date Conclusive DX Flag [19,741 facts; 18,917 patients] - 18917
  - 0449 Grade Path System [793 facts; 775 patients] - 775
  - 0450 Site Coding Sys--Current [97,348 facts; 90,542 patients] - 90542
### Coded Fields

#### GRADE

<table>
<thead>
<tr>
<th>Alternate Name</th>
<th>Item #</th>
<th>Length</th>
<th>Source of Standard</th>
<th>Year Implemented</th>
<th>Version Implemented</th>
<th>Year Retired</th>
<th>Version Retired</th>
<th>Column #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade, Differentiation, or Cell Lineage Indicator (SEER/CCCR)</td>
<td>440</td>
<td>1</td>
<td>SEER/CoC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>555-555</td>
</tr>
<tr>
<td>Grade/Differentiation (CoC)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Description**

Code for the grade or degree of differentiation of the reportable tumor. For lymphomas and leukemias, field also is used to indicate T-, B-, Null-, or NK-cell origin.

**Note:** Code 8 was adopted for use with lymphoma cases diagnosed in 1995 and later.

**Codes**

See the grade tables on page 67 of *ICD-O-3*. See also the most recent *CoC FORDS* manual and *SEER Program Code Manual*, for site specific coding rules and conversions.

1. Grade I
2. Grade II
3. Grade III
4. Grade IV
5. T-cell
6. B-cell
7. Null cell
8. NK (natural killer) cell
9. Grade/differentiation unknown, not stated, or not applicable

**Comment:** Use the most recent Hematopoietic and Lymphoid rules for assigning grades 5-8.
Code Values are Leaves in i2b2
De-identification in HERON

- No names, MRNs, ...identifiers
- No free-text
- Dates are shifted 0-364 days per patient
Not all Primary Site=C50 is Breast Cancer

Excludes lymphoma and leukemia M9590-9989 and Kaposi sarcoma M9140
# SEER Site Recode

**SEER Site Recode ICD-O-3 (1/27/2003) Definition**

The information provided in this table is also available in an ASCII text file (semicolons are used as the delimiters).

To see how this variable is used in SEER publications and to read about other site recode variables, please see the [SEER Incidence Site Recode](#).

<table>
<thead>
<tr>
<th>Site Group</th>
<th>ICD-O-3 Site</th>
<th>ICD-O-3 Histology (Type)</th>
<th>Recode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oral Cavity and Pharynx</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lip</td>
<td>C000-C009</td>
<td>excluding 9590-9989, and sometimes 9050-9055, 9140+</td>
<td>20010</td>
</tr>
<tr>
<td>Tongue</td>
<td>C019-C029</td>
<td></td>
<td>20020</td>
</tr>
<tr>
<td>Salivary Gland</td>
<td>C079-C089</td>
<td></td>
<td>20030</td>
</tr>
<tr>
<td>Floor of Mouth</td>
<td>C040-C049</td>
<td></td>
<td>20040</td>
</tr>
<tr>
<td>Gum and Other Mouth</td>
<td>C030-C039, C050-C059, C060-C069</td>
<td></td>
<td>20050</td>
</tr>
</tbody>
</table>
/* Lip */ when (site between 'C000' and 'C009')
and not (histology between '9590' and '9989'
or histology between '9050' and '9055'
or histology = '9140') then '20010'

...
SEER Site Recode
ER/PR hidden in CS Site-Specific Factors

| Cancer Cases \ 11 Stage/Prognostic Factors | 1920 Regional Nodes Examined |
| Cancer Cases \ 11 Stage/Prognostic Factors | 2869 CS Site-Specific Factor 15 |
| Cancer Cases \ 11 Stage/Prognostic Factors | 2876 CS Site-Specific Factor 22 |
| Cancer Cases \ 11 Stage/Prognostic Factors | 2877 CS Site-Specific Factor 23 |
| Cancer Cases \ 11 Stage/Prognostic Factors | 2880 CS Site-Specific Factor 1 |
| Cancer Cases \ 11 Stage/Prognostic Factors | 2890 CS Site-Specific Factor 2 |
| Cancer Cases \ 11 Stage/Prognostic Factors | 2940 Derived AJCC-6 T |
| Cancer Cases \ 11 Stage/Prognostic Factors | 3020 Derived SS2000 |
| Cancer Cases \ 11 Stage/Prognostic Factors | 3120 Derived AJCC-7 T |
CS Site-Specific Factors
# CS Site-Specific Factors

| Cancer Cases 111 Stage/Prognostic Factors | 0830 Regional Nodes Examined |
| Cancer Cases 111 Stage/Prognostic Factors | 2869 CS Site-Specific Factor 15 |
| Cancer Cases 111 Stage/Prognostic Factors | 2876 CS Site-Specific Factor 22 |
| Cancer Cases 111 Stage/Prognostic Factors | 2877 CS Site-Specific Factor 23 |
| Cancer Cases 111 Stage/Prognostic Factors | 2880 CS Site-Specific Factor 1 |
| Cancer Cases 111 Stage/Prognostic Factors | 2890 CS Site-Specific Factor 2 |
| Cancer Cases 111 Stage/Prognostic Factors | 2994 Derived AJCC:6 T |
| Cancer Cases 111 Stage/Prognostic Factors | 3020 Derived EJCC:6 T |

- **Breast**: 128,115 patients, 7,700 cases
  - **01**: Estrogen Receptor (ER) Assay [8,051 facts; 7,700 patients] - 7700
    - **000**: OBSOLETE DATA CONVERTED V0203 (<10 facts)
    - **100**: Positive elevated [6,160 facts; 5,927 patients] - 5927
    - **020**: Negative/norma/within normal limits [1,576 facts; 1,562 patients] - 1562
    - **030**: Borderline; undetermined whether positive or negative (<10 facts)
    - **080**: OBSOLETE DATA CONVERTED V0203
    - **099**: Not applicable; Information not collected for this case
    - **099**: Test ordered, results not in V0203 (<10 facts)
    - **0997**: Test ordered, results not in chart (<10 facts)
    - **0999**: Test not done (test not ordered and not performed) [130 facts; 129 patients] - 129
    - **999**: Unknown or no information [184 facts; 163 patients] - 163
  - **02**: Progesterone Receptor (PR) Assay [8,051 facts; 7,700 patients] - 7700
  - **03**: Number of Positive Isocentric Level III Axillary Lymph Nodes [6,080 facts; 5300 cases] - 5300
Exhaustive NAACCR Ontology
Demographics - Usable
Staging - Overwhelming
Class of Case: Tedious

Main distinction:
- Analytic
- Non-Analytic
Edited NAACCR Ontology
HERON NAACCR Approach

● At KUMC: >50K cases from the 1960’s
  ○ Updated monthly since 2011
● Exhaustive ETL, Ontology
● Supported 8-site GPC Breast Cancer Survey
● Refinements:
  ○ Edited Ontology
  ○ SEER Site Summary recode
  ○ Site-specific factors
Abstract

The NAACCR tumor registry is an important data source for cancer research. Data are aggregated at state levels from many academic medical centers using a standardized record format. KUMC integrated its NAACCR tumor registry into its enterprise clinical data repository, HERON, in 2011 as part of its bid for NCI designation. In 2015, our approach was adopted at 8 sites in our PCORNet CDRN, the Greater Plains Collaborative, and was used to support a survey of hundreds of breast cancer patients from several states. The initial approach exhaustively exposed the NAACCR records via the i2b2 query tool but overwhelmed most users with its ontology, limiting the utility to those with expert knowledge of both the NAACCR and i2b2 data representations. The current approach addresses a number of problems with the original approach by:

- promoting a subset of the ontology that is most often relevant
- synthesizing SEER site summary from primary site and morphology
- addressing limitations of the original approach for site-specific factors