Contents

[Overview 1](#_Toc132881731)

[Incidence data: 2](#_Toc132881732)

[SAS program name: 2](#_Toc132881733)

[Input 2](#_Toc132881734)

[Output variable list and variable names: 2](#_Toc132881735)

[Medicare: 5](#_Toc132881736)

[Enrollment: 5](#_Toc132881737)

[SAS program name 5](#_Toc132881738)

[Input 5](#_Toc132881739)

[Output 5](#_Toc132881740)

[Claims 5](#_Toc132881741)

[SAS program name 5](#_Toc132881742)

[Input 5](#_Toc132881743)

[Output 5](#_Toc132881744)

[Codes: 5](#_Toc132881745)

[Medicaid: 7](#_Toc132881746)

[Enrolls: 7](#_Toc132881747)

[SAS program name 7](#_Toc132881748)

[Input 7](#_Toc132881749)

[Output 7](#_Toc132881750)

[Health Claims: 7](#_Toc132881751)

[SAS program name 7](#_Toc132881752)

[Input 7](#_Toc132881753)

[Output 7](#_Toc132881754)

[Codes: 7](#_Toc132881755)

# Overview

This document showcases the process of creating preprocessed incidence, claims, and enrollment files which follow the formats to be used in the XXX package to generate prediction of breast cancer recurrence.

The included SAS programs show how the incidence, Medicare and Medicaid claims, and enrollment files are generated based on the data and their formats used to develop the package in Kentucky Cancer Registry. Users may have different formats or variable names in their incidence, claims or enrollment files. Users may need to use their programs to preprocessed incidence, claims and enrollments files to meet the input requirements for the package. The data dictionary provides the detailed information for these variables and requirements, which users should follow to generate these input files.

All SAS programs for processing and output files saved at Z:\UG3-UH3\UH3\Data Analysis\Data Processing Documents

# Incidence data:

### SAS program name:

*Read and Output Incidence File.sas*

### 

### Input File

The input file is a SAS data set created from the raw cancer registry incidence files. The attached SAS program is an example of the dataset we used. Users can use their own programs to generate the output dataset as long as it follows the output format below.

Selection criteria: Invasive, First or Only primary, Female Breast Cancer patients (SEERSite= 26000 and Behavior code=3, Sequence in (0,1)).

In the App, additional selection criteria are imposed, such as stage I-III, 12-month enrollment after cancer diagnosis, etc. Users can impose these additional selection criteria in the SAS program before generating output data.

### 

### Output variable list and variable names:

The crosswalk file, use this file to link enroll and claims for encrypted ids.

|  |  |  |
| --- | --- | --- |
| Variables | Variable names | Notes |
| Patient ID | patient\_id |  |
| Study ID | study\_id | Encrypted study ID |

Incidence file:

|  |  |  |
| --- | --- | --- |
| Variables | Variable names | Notes |
| Study ID | study\_id |  |
| Age at Diagnosis | DIAGAGE | Allowable Values:0-120, 999(Unknown) |
| Cancer Sequence | CentralSequenceNumber | |  |  | | --- | --- | | 00 | One primary only | | 01 | First of two or more primaries | | 02 | Second of two or more primaries | | -- | -- | | -- | (Actual number of this primary) | | -- |  | | 59 | Fifty-ninth or higher of fifty-nine or more primaries | | 99 | Unspecified or unknown sequence number of Federally required in situ or malignant tumors. Sequence number 99 can be used if there is a malignant tumor and its sequence number is unknown. (If there is known to be more than one malignant tumor, then the tumors must be sequenced.) | |
| Race | race | |  |  | | --- | --- | | 1 | White | | 2 | Black | | 3 | Other | |
| Primary Site | PrimarySite | Allowable Values: C followed by 3 digits, no special characters, no embedded blanks.  Using the topography codes listed in the International Classification of Diseases for Oncology, Third Edition (ICD-O-3). The topography code consists of a lead character (the letter ‘C’) followed by two numeric digits, a decimal point, and then one additional numeric digit. The decimal point is not entered as part of the code. |
| Stages | Ajccstage | Recoded based on Derived AJCC-7 Stage Grp   |  |  | | --- | --- | | 1 | Stage I (Derived AJCC-7 Stage Grp values :100-220) | | 2 | Stage II (Derived AJCC-7 Stage Grp values :300-430) | | 3 | Stage III (Derived AJCC-7 Stage Grp values :500-630) | | 4 | Stage IV(Derived AJCC-7 Stage Grp values :700 -740) | | 9 | Unknown (Derived AJCC-7 Stage Grp values:888,900, and 999) | |
| SEERSummStg | |  |  | | --- | --- | | 0 | In situ | | 1 | Localized | | 2 | Regional, direct extension only | | 3 | Regional, regional lymph nodes only | | 4 | Regional, direct extension and regional lymph nodes | | 7 | Distant | | 8 | Benign, borderline | | 9 | Unknown if extension or metastasis (unstaged, unknown, or unspecified) |   For years of diagnosis from 2010 to 2017, data were obtained from the variable “derivedSS2000”. For years of diagnosis starting in 2018, data were obtained from the variable “derivedSummaryStage2018”. |
| Tumor Grade | grade | |  |  | | --- | --- | | 1 | Grade I | | 2 | Grade II | | 3 | Grade III | | 4 | Grade IV | | 9 | Grade/differentiation unknown, not stated, or not applicable | |
| Date of Diagnosis | Date\_dx | Format: MM/DD/YYYY |
| Year of diagnosis | year\_diag | Format: YYYY |
| Laterality | Laterality | |  |  | | --- | --- | | 0 | Not a paired site | | 1 | Right: origin of primary | | 2 | Left: origin of primary | | 3 | Only one side involved, right or left origin unspecified | | 4 | Bilateral involvement at time of diagnosis, lateral origin unknown for a single primary; or both ovaries involved simultaneously, single histology; bilateral retinoblastoma; bilateral Wilms tumors | | 5 | Paired site: Midline tumor (effective with 01/01/2010 dx) | | 9 | Paired site, but no information concerning laterality | |
| Estrogen Receptor (ER) Assay | er\_stat | 1 positive, 0 negative, 9 unknown |
| Progesterone Receptor (PR) Assay | pr\_stat | 1 positive, 0 negative, 9 unknown |
| HER2 | her2\_stat | 1 positive, 0 negative, 9 unknown |
| Surgery of Primary Site | RXSummSurgPrimSite | |  |  | | --- | --- | | 00 | None | | 10-19 | Site-specific code; tumor destruction | | 20-80 | Site-specific codes; resection | | 90 | Surgery, NOS | | 98 | Site specific codes; special | | 99 | Unknown | |
| Scope of Regional Lymph Node Surgery | RXSummScopeRegLNSur | |  |  | | --- | --- | | 1 | Biopsy or aspiration of regional lymph node, NOS | | 2 | Sentinel lymph node biopsy [only] | | 3 | Number of regional lymph nodes removed unknown, not stated; regional lymph nodes removed, NOS | | 4 | 1 to 3 regional lymph nodes removed | | 5 | 4 or more regional lymph nodes removed | | 6 | Sentinel node biopsy and code 3, 4, or 5 at same time or timing not noted | | 7 | Sentinel node biopsy and code 3, 4, or 5 at different times | | 9 | Unknown or not applicable | |
| Surgical Procedure of Other Site | RXSummSurgOthRegDis | |  |  | | --- | --- | | 0 | None; diagnosed at autopsy | | 1 | Non-primary surgical procedure performed | | 2 | Non-primary surgical procedure to other regional sites | | 3 | Non-primary surgical procedure to distant lymph node(s) | | 4 | Non-primary surgical procedure to distant site | | 5 | Combination of codes 2, 3, or 4 | | 9 | Unknown | |
| First course of treatment included radiation therapy | RXSummRadiation | |  |  | | --- | --- | | 0 | None, diagnosed at autopsy | | 1 | Beam | | 2 | Implants | | 3 | Isotopes | | 4 | Combination of 1 with 2 or 3 | | 5 | Radiation, NOS | | 6 | Implants/Isotopes, NOS (Coding\_Proc = 03-11) | | 9 | Unknown | |
| Sequencing of radiation and surgery | RXSummSurgRadSeq | |  |  | | --- | --- | | 0 | No radiation and/or no cancer-directed surgery | | 2 | Radiation before surgery | | 3 | Radiation after surgery | | 4 | Radiation both before and after surgery | | 5 | Intraoperative radiation | | 6 | Intraoperative radiation with other radiation given before and/or after surgery | | 7 | Surgery both before and after radiation (for cases diagnosed 01/01/2012 and later) | | 9 | Both surgery and radiation given, but sequence unknown | |
| First course of treatment included chemotherapy | RXSummChemo | |  |  | | --- | --- | | 00 | None, chemotherapy was not part of the planned first course of therapy.  Diagnosed at autopsy. | | 01 | Chemotherapy administered as first course therapy, but the type and number of agents is not documented in the patient record. | | 02 | Single-agent chemotherapy administered as first course therapy. | | 03 | Multi-agent chemotherapy administered as first course therapy. | | 82 | Chemotherapy was not recommended/administered because it was contraindicated due to patient risk factors (i.e., comorbid conditions, advanced age, progression of tumor prior to administration, etc.) | | 85 | Chemotherapy was not administered because the patient died prior to planned or recommended therapy. | | 86 | Chemotherapy was not administered.  It was recommended by the patient's physician but was not administered as part of the first course of therapy.  No reason was stated in the patient record. | | 87 | Chemotherapy was not administered.  It was recommended by the patient's physician, but this treatment was refused by the patient, a patient's family member, or the patient's guardian.  The refusal was noted in the patient record. | | 88 | Chemotherapy was recommended, but it is unknown if it was administered. | | 99 | It is unknown whether chemotherapy was recommended or administered because it is not stated in the patient record. | |
| First course of treatment included chemotherapy | RXSummHormone | |  |  | | --- | --- | | **00** | None, hormone therapy was not part of the planned first course of therapy; not usually administered for this type and/or stage of cancer; diagnosed at autopsy only | | **01** | Hormone therapy was given as first course therapy. | | **82** | Hormone therapy was not recommended/administered because it was contraindicated due to patient risk factors (i.e., comorbid conditions, advanced age, progression of tumor prior to administration, etc.) | | **85** | Hormone therapy was not administered because the patient died prior to planned or recommended therapy | | **86** | Hormone therapy was not administered. It was recommended by the patient’s physician but was not administered as part of the first course of therapy. No reason was stated in the patient record. | | **87** | Hormone therapy was not administered. It was recommended by the patient’s physician, but this treatment was refused by the patient, a patient’s family member, or the patient’s guardian. The refusal was noted in the patient record. | | **88** | Hormone therapy was recommended, but it is unknown if it was administered. | | **99** | It is unknown whether a hormonal agent(s) was recommended or administered | |
| biologic response modifier (BRM) | RXSummBRM | |  |  | | --- | --- | | 00 | None, immunotherapy was not administered as part of first course treatment.  Diagnosed at autopsy. | | 01 | Immunotherapy was given as first course therapy. | | 82 | Immunotherapy was not recommended/administered because it was contraindicated due to patient risk factors. | | 85 | Immunotherapy was not administered because the patient died prior to planned or recommended therapy. | | 86 | Immunotherapy was not administered.  It was recommended by the patient's physician but was not administered as part of first course therapy.  No reason was stated in the patient record. | | 87 | Immunotherapy was not administered.  It was recommended by the patient's physician, but treatment was refused by the patient, a patient's family member, or the patient's guardian.  The refusal was noted in the patient record. | | 88 | Immunotherapy was recommended, but it is unknown if it was administered. | | 99 | It is unknown whether immunotherapy was recommended or administered because it is not stated in the patient record.  Death certificate only. | |
| Other Therapy | RXSummOther | |  |  | | --- | --- | | 00 | None, immunotherapy was not administered as part of first course treatment.  Diagnosed at autopsy. | | 01 | Immunotherapy was given as first course therapy. | | 82 | Immunotherapy was not recommended/administered because it was contraindicated due to patient risk factors. | | 85 | Immunotherapy was not administered because the patient died prior to planned or recommended therapy. | | 86 | Immunotherapy was not administered.  It was recommended by the patient's physician but was not administered as part of first course therapy.  No reason was stated in the patient record. | | 87 | Immunotherapy was not administered.  It was recommended by the patient's physician, but treatment was refused by the patient, a patient's family member, or the patient's guardian.  The refusal was noted in the patient record. | | 88 | Immunotherapy was recommended, but it is unknown if it was administered. | | 99 | It is unknown whether immunotherapy was recommended or administered because it is not stated in the patient record.  Death certificate only. | |
| sequencing of systemic therapy and surgical procedures | RXSummSystSurSeq | |  |  | | --- | --- | | 0 | No systemic therapy and/or no surgical procedure | | 2 | Systemic therapy before surgery | | 3 | Systemic therapy after surgery | | 4 | Systemic therapy both before and after surgery | | 5 | Intraoperative systemic therapy | | 6 | Intraoperative systemic therapy with other therapy given before or after surgery | | 9 | Both surgery and systemic therapy given, but sequence unknown | |
| TNM Clinical stages | TNMClinT, TNMClinN, TNMClinM | Clinical stage group as defined by the current AJCC 8th edition. |
| TNM Pathological stages | TNMPathT, TNMPathN, TNMPathM | Pathological stage group defined by the current AJCC 8th edition. |
| Regional Lymph nodes examined | RegNodesExamined | |  |  | | --- | --- | | 00 | No nodes were examined | | 01-89 | 1-89 nodes are examined (code exact number of nodes examined) | | 90 | 90 or more nodes were examined | | 95 | No regional nodes were removed, but aspiration OR core biopsy regional nodes was performed | | 96 | Regional lymph node removal was documented as a sampling, and the number of nodes is unknown/not stated | | 97 | Regional lymph node removal was documented as a dissection, and the number of nodes is unknown/not stated | | 98 | Regional lymph nodes were surgically removed, but the number of lymph nodes is unknown/not stated and not documented as a sampling or dissection; nodes were examined, but the number is unknown | | 99 | It is unknown whether nodes are examined; not stated in patient record | |
| Regional Lymph nodes positive | RegNodesPositive | |  |  | | --- | --- | | 00 | All nodes examined negative. | | 01-89 | 1 - 89 nodes positive (code exact number of nodes positive) | | 90 | 90 or more nodes are positive | | 95 | Positive aspiration OR core biopsy of lymph node(s) was performed | | 97 | Positive nodes are documented - number unspecified | | 98 | No nodes examined | | 99 | Unknown if nodes are positive; not applicable  Not documented in patient record | |
| Tumor size | CSTumorSize | Allowable Values: 000-999 (site-specific) |
| Other **Collaborative Staging (CS) variables** | CSLymphNodes | Allowable Values: 000-999 (site-specific) |
| CSLymphNodesEval | Allowable Values: 0-9 (site-specific) |
| CSExtension | Allowable Values: 000-999 (site-specific) |
| CSTumorSizeExtEval | Allowable Values: 0-9 (site-specific) |
| RUCA code | RUCA2000, RUCA2010 | |  |  | | --- | --- | | 1 | Urban commuting area-RUCA codes 1.0, 1.1, 2.0, 2.1, 3.0, 4.1, 5.1, 7.1, 8.1, and 10.1 | | 2 | Not an urban commuting area-All other RUCA codes except 99 | | 9 | Unknown or not applicable-census tract not available or RUCA code = 99 | |
| Rural-Urban Continuum Codes | RuralUrbanContin2013 | |  |  | | --- | --- | | 1 | Counties in metro areas of 1 million population or more | | 2 | Counties in metro areas of 250,000 to 1 million population | | 3 | Counties in metro areas of fewer than 250,000 population | | 4 | Urban population of 20,000 or more, adjacent to a metro area | | 5 | Urban population of 20,000 or more, not adjacent to a metro area | | 6 | Urban population of 2,500 to 19,999, adjacent to a metro area | | 7 | Urban population of 2,500 to 19,999, not adjacent to a metro area | | 8 | Completely rural or less than 2,500 urban population, adjacent to a metro area | | 9 | Completely rural or less than 2,500 urban population, not adjacent to a metro area | | 88 | Unknown-Alaska/Hawaii State/not official USDA Rural-Urban Continuum code | | 99 | Unknown/not official USDA Rural-Urban Continuum code | |
| Recurrence Date | Date\_1Recur | Format: MM/DD/YYYY |
| Vital Statues | VitalStat | |  |  | | --- | --- | | 1 | Alive | | 0 | Dead | |
| Histology code | ICDO3Histology | Allowable Values: 8000-9989. Refer to the International Classification of Diseases for Oncology, Third Edition, Second Revision Morphology (ICD-O-3) |
| Cause of Death | CauseOfDeath | Allowed Values: 4 digits (for ICD-9); for ICD-10, upper case letter followed by 3 digits or upper-case letter followed by 2 digits plus blank.  Use ICD-9 Cause of Death Code for deaths before Jan 1, 1999, and ICD-10 Code for deaths on or after Jan 1, 1999. |
| Date of birth | Date\_birth | Format: MM/DD/YYYY |
| *Other Cancer information (diagnosis date, sequence, and cancer site) was also linked for female breast cancer patients* | | |
| Cancer Sites other than female breast cancer | site\_O1-site\_O6 | Allowed Values: C followed by 3 digits, no special characters, no embedded blanks |
| Date of cancer case diagnosis (other than female breast cancer) | date\_O1-date\_O6 | Format: MM/DD/YYYY |
| Sequence of cancer cases (other than female breast cancer) | sequence\_O1-sequence\_O6 | |  |  | | --- | --- | | 00 | One primary only | | 01 | First of two or more primaries | | 02 | Second of two or more primaries | | -- | -- | | -- | (Actual number of this primary) | | -- |  | | 59 | Fifty-ninth or higher of fifty-nine or more primaries | | 99 | Unspecified or unknown sequence number of Federally required in situ or malignant tumors. Sequence number 99 can be used if there is a malignant tumor, and its sequence number is unknown. (If there is known to be more than one malignant tumor, then the tumors must be sequenced.) | |
| Date of last contact | Date\_lc | Format: MM/DD/YYYY |

# **Medicare:**

## **Enrollment file**:

### SAS program name

Read and Output Medicare Enrollment File.sas

### 

### Input

The input file is a SAS data file which is generated from the SAS programs provided by SEER\*Medicare. The data were based on an older version of the Medicare enrollment file, containing the patient ID and monthly enrollment index. The newer SEER\*Medicare data have different format. Users need to make modifications to make sure the output data follow the following format.

Data Dictionary ‘PEDSF2017.pdf’

<https://healthcaredelivery.cancer.gov/seermedicare/medicare/enroll.html>

Use ‘study\_id’ to link with incidence cross work file.

### Output

|  |  |  |
| --- | --- | --- |
| Variable | Variable Names | Notes |
| Study ID | study\_id | Study ID |
| Monthly Enroll Index | Mon1-Mon324 | One indicator for each month from 1/1991  to 12/2017.   |  |  | | --- | --- | | 0 | Not entitled | | 1 | Part A only | | 2 | Part B only | | 3 | Part A and B | |
| HMO enrollment indicator | GHO1-GHO324 | Indicates entitlement for each month from 1/1991  to 12/2017.  gho1 = January 1991  gho324 = December 2017   |  |  | | --- | --- | | 0 | Not Member of HMO | | 1 | Non-Lock-in, CMS to process Provider | | 2 | Non-Lock-in, GHO to process in-plan Part A& in-area Part B claims | | 4 | Fee-for-Service participant in case or disease  management demonstration projects  (effective 2005 forward) | | A | Lock-in, CMS to process provider claims | | B | Lock-in, GHO to process in-plan Part A and  in-area Part B claims | | C | Lock-in, GHO to process all provider claims | |

## 

## **Claims File**

### SAS program name

Read, Link and Output Medicare Claims File.sas

### 

### Input

The input data is a pre-processed SEER Medicare data with years of claims concatenated.

DME, MEDPAR, NCH, OUTPAT, PDE files between years 1999-2016.

The record (each row) in the output file is claim level.

Use ‘study\_id’ to link with incidence crosswalk file.

Only cases captured in the incidence data with Medicare enrollment are included in the claims output file.

### 

### Output

Outputs include ID to link across incidence file and claims files, claims dates, and codes.

|  |  |  |
| --- | --- | --- |
| Variable | Variable Names | Notes |
| Study ID | study\_id | Study ID-patient level |
| Claims ID | claim\_id | Claims level |
| Diagnose Codes | dgns\_cd1 -dgns\_cd25 | Allowable values: ICD-9 or ICD-10 codes. ICD-10 diagnosis codes were used starting October 2015. |
| Procedure Codes | PRCDR\_CD1 – PRCDR\_CD25 | Allowable values: ICD-9-CM code ICD-10-PCS code. 3-5 digits (for ICD-9 CM); for ICD-10-PCS, a 7-character alphanumeric codes. |
| HCPCs Code | hcpcs\_cd | Allowable values: Healthcare Common Procedure Coding System (HCPCS) code. Detailed coding refers to: https://www.cms.gov/medicare/regulations-guidance/physician-self-referral/list-cpt/hcpcs-codes |
| Revenue Center Codes | REV\_CNTR | Allowable values: 4-digitprovider-assigned revenue code. Detailed coding refers to: https://bluebutton.cms.gov/resources/variables/rev\_cntr/ |
| NDC codes | NDC\_CD (DME file), prod\_srvc\_id (PDE file) | Allowable values: National Drug Code (NDC) code. A 5-byte numeric labeler code, 4 byte numeric product code and a 2 byte numeric package code. |
| Dates | Claims\_date | Format: MM/DD/YYYY |
| File source | dfile | Indicator of input file: DME, NCH, MEDPAR, OUTPAT, or PDE |

# **Medicaid/Other Claims Sources:**

Medicare claims data can be considered as customized claims data based on individual claims sources, not necessarily limited to Medicaid data. The program here is only an example of what to be included in these type of claims data.

## 

## **Enrollment file:**

### SAS program name

Read, Link and Output Medicard Enrollment File.SAS

### 

### Input

Medicaid patient level enrollment status contains the patient ID and monthly enrollment index.

Use ‘study\_id’ to link with incidence cross work file.

Use ‘Id\_medicaid’ to link with Claims file.

### Output

|  |  |  |
| --- | --- | --- |
| Variable | Variable Names | Notes |
| Study ID | study\_id | Study ID |
| Medicaid ID | Id\_medicaid | Medicaid ID |
| Monthly Enroll Index | Mon1-Mon240, etc | One indicator for each month.  mon1 = 1/2000  mon2 = 2/2000  mon240 = 12/2019   |  |  | | --- | --- | | 0 | Not Enrolled | | 1 | Enrolled | |

## **Claims Files:**

### SAS program name

Read, Link and Output Medicaid Claims File.sas

### 

### Input

Pre-processed Medicaid data with years of claims concatenated

The record (each row) in the output file is claim level.

Use ‘id\_medicaid’ to link with enrollment file

### 

### Output

|  |  |  |
| --- | --- | --- |
| Variable | Variable Names | Notes |
| Study ID | study\_id | Study ID |
| Medicaid ID | Id\_medicaid | Medicaid ID |
| ICD Diagnosis Codes | CDE\_DIAG\_PRIM cde\_diag\_2-cde\_diag\_4 | Allowable values: ICD-9 or ICD-10 codes. ICD-10 diagnosis codes were used starting October 2015. |
| Procedure/ HCPCs codes | CDE\_PROC\_PRIM | Allowable values: ICD-9-CM code ICD-10-PCS code. 3-5 digits (for ICD-9 CM); for ICD-10-PCS, a 7-character alphanumeric codes. |
| Date of First service/ last service. | DTE\_FIRST\_SVC DTE\_LAST\_SVC | Format: YYYY-MM-DD |
| NDC codes | CDE\_NDC | Allowable values: National Drug Code(NDC) code. A 5 byte numeric labeler code, 4 byte numeric product code and a 2 byte numeric package code. |
| AHFS Therapeutic Class Code | CDE\_THERA\_CLS\_AHFS | Allowable values: AHFS Therapeutic Class Code. 5-8 digits.  Detailed coding refers to: https://www.oregon.gov/obnm/Documents/Formulary%20Information/AHFSClassificationwithDrugs2019.pdf |