**Schaeffer Center Verification Algorithm for Identifying Dementia in Claims Data**Technical Documentation

Released March 2023

Patricia Ferido, Geoff Joyce, Bryan Tysinger, Sidra Haye, Johanna Thunell, Yi Chen, Julie Zissimopoulos

**Code on Github:** <https://github.com/USCSchaefferCenterDataCore/Schaeffer-Dementia-Algorithm/>

Funding support through National Institute on Aging P30AG066589, “Center for Advancing Sociodemographic and Economic Study of Alzheimer’s Disease and Related Dementias (CeASES‐ADRD),” National Institute on Aging P30AG043073, “Resource Center for Minority Aging Research.”

Table of Contents

Algorithm Overview………..…..………………………………..……………….……………………… 3

Dementia ICD Codes ……………………………………………………………………………4

Dementia Symptoms …………………………………………………………………………....4

Dementia Drugs …………………………………………………………………………………4

Package Description & Requirements ………..…..……………………..……………..………….….... 5

Output Codebook …….…………………………………………….………..…………………. 6

Step-by-Step Guide to Run Package …….………………..……...………...….………………………. 8

Input Data Requirements ……………………………………………………………………… 8

Input Macro Values …………………………………………………………………………….. 9

Common Pitfalls ……………………….….….……………………...…..………………………………. 10

Use Case: Calculating Incidence and Prevalence ……...………...…..………………………………. 10

Full Example ………………………………….……………………...…..………………………………. 12

References ……………………………….…….……………………...…..………………………………. 12

**Algorithm Overview**

This package identifies dementia in claims data as originally defined in “Measuring Alzheimer’s Disease and Other Dementias in Diverse Populations Using Medicare Claims Data” (Thunell, Ferido, & Zissimopoulos, 2019) and updated in “Estimates of Diagnosed Dementia Prevalence and Incidence in Traditional Medicare and Medicare Advantage” (Haye et al.). The algorithm uses diagnosis claims, dementia symptoms claims, and drug claims. A second claim or death within one year is required to verify an initial diagnosis and reduce measurement error from a ‘rule-out’ diagnosis.

Specifically, there are four methods for identifying dementia:

1. One dementia diagnosis code AND second verification dementia diagnosis within two years OR death within one year after diagnosis (diagnoses only)
2. One dementia diagnosis code AND one dementia symptom code within two years OR death within one year after diagnosis (diagnosis and symptom)
3. One dementia diagnosis code AND second verification diagnosis code OR drug claim in any order within two years OR death within one year after diagnosis (diagnosis and drug)
4. One dementia diagnosis code AND second verification diagnosis code OR one dementia symptom code OR drug code in any order within two years OR death within one year after the diagnosis (diagnosis, drug and symptom)

The following diagram draws out the components of each method:

All scenarios

1. Dx only

3. Dx + Rx

2. Dx + symp

in any order

Second dementia diagnosis claim in two years at least one day apart

OR

4. Dx + symp or Rx

OR

Dementia symptom claim in two years

OR

Dementia diagnosis claim

AND

Dementia drug claim in two years

OR

Death in one year

in any order

In cases where a drug or symptom claim verifies a diagnosis, the drug or symptom claim can precede the diagnosis and the date of the first event is considered the dementia date. The study, Thunell, Ferido, & Zissimopoulos, 2019, found that use of drugs and symptoms to identify dementia was important for identifying dementia, especially among racial/ethnic minorities. However, not all researchers may have access to drug claims so the package allows for identification without these data.

While the original paper used a two-year verification window for a second claim, other work has used a one-year verification window due to data limitations and found similar results. In 2016, incidence with one-year verification was 3.34 (3.33-3.35) and 3.35 (3.34-3.36) with two-year verification.Thus the package allows for the use of either a two-year or one-year verification window as the data allows. In both cases, verification by death always occurs within one year from a diagnosis.

**Dementia Diagnosis Codes**

The following tables list the ICD-9 and ICD-10 codes used for dementia. Descriptions for each code can be found in the included excel document found in /csv\_input/icdcodes.csv.

|  |  |
| --- | --- |
| **Condition** | **Code** |
| Dementia | ICD-9: 290.0, 290.10, 290.11, 290.12, 290.13, 290.20, 290.21, 290.3, 290.40, 290.41, 290.42, 290.43, 294.0, 294.10, 294.11, 294.20, 294.21, 294.8, 331.0, 331.11, 331.19, 331.2, 331.7, 331.82, 797  ICD-10: G30.0, G30.1, G30.8, G30.9, F01.50, F01.51, F02.80, F02.81, F03.90, F03.91, F05, G13.8, G31.01, G31.09, G31.1, G31.2, G31.83, G94, R41.81 |

**Dementia Symptoms**

We consider the following conditions as symptoms related to dementia: amnesia, aphasia, mild cognitive impairment and apraxia and agnosia. A full list of codes and descriptions can be found in the included excel document /csv\_input/icdcodes.csv.

|  |  |
| --- | --- |
| **Condition** | **Code** |
| Dementia Symptom | ICD-9: 780.93, 784.3, 331.83, 784.69  ICD-10: R41.1, R41.2, R41.3, R47.01, R48.2, R48.8, G318.4 |

**Dementia Drugs**

We consider the following drugs used to treat symptoms of dementia:

* donepezil
* galantamine
* rivastigmine
* memantine

We include an excel sheet with a list of relevant NDC’s at the time of this release (/csv\_input/ndc2019.csv); however, NDC’s are subject to change so we strongly recommend the use of generic name to identify drugs or a validation of relevant NDC’s with an external drug database.

**Package Description**  
This package allows a researcher to replicate the algorithm on any claims data or electronic health records with information on diagnoses, diagnosis date and death date by patient. This version of the package was written and tested on SAS Enterprise Guide 7.1, and should run with any compatible version of SAS. The package takes as input claims data, drug data, and user-defined values as outlined below. It outputs yearly datasets with variables for the dates of a verified dementia in that year and the first valid dates of dementia across all years.

**Censoring**Please note that in order to get valid results, you will need enough data for your specified verification window. If you select a two-year verification window, you will need at least two full years of data after your dates of interest and if you select a one-year verification window, you will need at least one full year of data after your dates of interest to avoid censoring. E.g., to verify a first dementia claim on December 31, 2021, you will need to include data through December 31, 2022 for one-year verification and through December 31, 2023 for two-year verification. All output data sets with potential censoring will be named with the suffix ‘\_censor’ and these data should be treated with caution.

**Output Codebook**

These are the variables for annual datasets at the patient-level.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| User Defined ID | Unique beneficiary identifier |
| First\_dx\_dt | Date of first verified dementia using dx + dx or dx + death. This variable uses all available years of data except for the years that potentially include censoring. |
| First\_dxsymp\_dt | Date of first verified dementia using dx + dx, dx + symp, or dx + death. This variable uses all available years of data except for the years that potentially include censoring. |
| First\_dxrx\_dt | Date of first verified dementia using dx + dx, dx + rx, or dx + death. This variable uses all available years of data except for the years that potentially include censoring. |
| First\_dxrxsymp\_dt | Date of first verified dementia using dx + dx, dx + symp, dx + rx, or dx + death. This variable uses all available years of data except for the years that potentially include censoring. |
| Scen\_dx\_dt[year] | Date of first dementia in year using dx + dx or dx + death |
| Scen\_dx\_vdt[year] | Date of verification using dx + dx or dx + death |
| Scen\_dx\_vtime[year] | Time (days) to verification using dx + dx or dx + death |
| Scen\_dx\_dttype[year] | Type of claim that occurred first using dx + dx or dx + death  1 - diagnosis |
| Scen\_dx\_vtype[year] | Type of claim that verified using dx + dx or dx + death  1 – diagnosis  4 - death |
| Dropdx[year] | Flags a potential data issue where date of verification by death occurs before the incident claim date |
| Scen\_dxsymp\_dt[year] | Date of first dementia in year using dx + dx, dx + symp, or dx + death |
| Scen\_dxsymp\_vdt[year] | Date of verification using dx + dx, dx + symp, or dx + death |
| Scen\_dxsymp\_vtime[year] | Time (days) to verification using dx + dx, dx + symp, or dx + death |
| Scen\_dxsymp\_dttype[year] | Type of claim that occurred first using dx + dx, dx + symp, or dx + death  1 – diagnosis  3 – symptom |
| Scen\_dxsymp\_vtype[year] | Type of claim that verified using dx + dx, dx + symp, or dx + death  1 – diagnosis  3 – symptom  4 – death |
| Dropdxsymp[year] | Flags a potential data issue where date of verification by death occurs before the incident claim date |
| Scen\_dxrx\_dt[year] | Date of first dementia in year using dx + dx, dx + rx, or dx + death method |
| Scen\_dxrx\_vdt[year] | Date of verification using dx + dx, dx + rx, or dx + death |
| Scen\_dxrx\_vtime[year] | Time (days) to verification using dx + dx, dx + rx, or dx + death |
| Scen\_dxrx\_dttype[year] | Type of claim that occurred first using dx + dx, dx + rx, or dx + death  1 – diagnosis  2 – drug |
| Scen\_dxrx\_vtype[year] | Type of claim that verified using dx + dx, dx + rx, or dx + death  1 – diagnosis  2 – drug  4 – death |
| Dropdxrx[year] | Flags a potential data issue where date of verification by death occurs before the incident claim date |
| Scen\_dxrxsymp\_dt[year] | Date of first dementia using dx + dx, dx + symp, dx + rx, or dx + death |
| Scen\_ dxrxsymp \_vdt[year] | Date of verification using dx + dx, dx + symp, dx + rx, or dx + death |
| Scen\_ dxrxsymp \_vtime[year] | Time (days) to verification using dx + dx, dx + symp, dx + rx, or dx + death |
| Scen\_ dxrxsymp \_dttype[year] | Type of claim that occurred first using dx + dx, dx + symp, dx + rx, or dx + death  1 – diagnosis  2 – drug  3 – symptom |
| Scen\_ dxrxsymp \_vtype[year] | Type of claim that verified using dx + dx, dx + symp, dx + rx, or dx + death  1 – diagnosis  2 – drug  3 – symptom  4 – death |
| Dropdxrxsymp[year] | Flags a potential data issue where date of verification by death occurs before the incident claim date |

The diagram below shows the general steps of the package:

**Drug Data** (Optional)

* Unique patient ID
* Service Date
* GNN or NDC

**Claims Data**

* Unique patient ID
* Death date
* Service date
* ICD diagnoses codes

**User-Defined Inputs** (/programs/input\_program.sas)

Variables defined in detail in Step-by-Step Guide – Step 11

* Project folder filepath
* ID
* Minimum year
* Maximum year
* Name of claims
* Whether to verify in 1 or 2 years
* Desired output data prefix
* Whether to use drug data
* Whether GNN or NDC included

**User Provided**

**Excel Sheet of valid ICD Codes** (/csv\_input/icd\_codes.csv)

**Wrapper program** (/programs/demv\_wrap.sas)

* pull\_dx\_clms.sas – Pulls the dementia and dementia symptom claims and makes unique at the patient-date level
* pull\_rx\_clms.sas – Pulls the dementia drug claims and makes unique at the patient-date level
* verify\_dementia.sas – Runs main algorithm

**Excel Sheet of valid NDCs** (/csv\_input/ndc2019.csv)

**Output** (desired location)Annual data sets with dementia date variables. Outlined in detail in Output Codebook pg. 6

**Step-by-Step Guide**

* 1. Copy the project folder into your personal directory and note the filepath. It will be an input macro variable later.

**Preparing the Input Claims Data**

The package will take multiple claims data as input. See pg. 4 for list of valid ICD codes. See pg. 5 for note about censoring and how many years to include.

* 2. Identify your claims data with diagnoses variables. When we use Medicare data, we use claims from the carrier, hospital outpatient, inpatient, skilled nursing facility, and home health agency files.

* 3. Identify your diagnosis variables. Rename them so that they have the prefix **icd\_dx** followed by a numeric suffix starting with **1**. The package will take any number of diagnosis variables. If you only have one diagnosis variable, name it **icd\_dx1.**

* 4. Identify your claim date variable**.** With Medicare claims, we use the clm\_thru\_dt. Name this variable **date.**
* 5. Identify your patient death date variable and make sure it is included in the file. In most cases, the death date will need to be merged onto your claims data set. Name the date variable **death\_date.**
* 6. Make sure your data include the following elements. For faster processing, drop any extraneous variables. Note the name and location of your data. These will be inputs to the package.

|  |  |  |
| --- | --- | --- |
| **Variable Type** | **Standardized Variable Names** | **Format** |
| Unique patient identifier\* | No standard name required | No standard format |
| ICD diagnosis codes\* | Icd\_dx1-icd\_dx[max]\* | Character\* |
| Claim date\* | Date\* | Date\* |
| Death date\* | Death\_date\* | Date\* |

If you are planning to use drug data then continue to Step #7, otherwise move on to Step #11.

**Preparing the Drug Claims Data (Optional)**

The package will take multiple drug data sets as input. See pg. 4 for list of included drugs. See pg. 5 for note on censoring and how many years of data to include.

* 7. Identify your drug claims data and your drug identifier variable. This data requires either generic name (GNN) or National Drug Code (National Drug Code). We include an excel sheet with a list of NDC’s for the relevant drugs at time of release (/csv\_input/ndc2019.csv); however, as NDCs are subject to change over time, we strongly recommend that you use GNN or do a validation of NDCs with an external drug database.
* 8a. If using GNN, then name the variable **gnn.**
* 8b. If using NDC, then name the variable **ndc.**
* 9. Identify your drug claim date variable. Name it **date**.
* 10. Make sure your drug data include the following elements. For faster processing, drop any extraneous variables. Note the name and location of your data. These will be inputs to the package.

|  |  |  |
| --- | --- | --- |
| **Variable Type** | **Standardized Variable Names** | **Format** |
| Unique patient identifier\* | No standard name required | No standard format |
| GNN or NDC \* | GNN\* or NDC\* | Character\* |
| Claim date\* | Date\* | Date\* |

**Preparing the Macro Input Variables**

* 11. Fill out your macro variables with the following values, paying special attention to the notes. **Do not enclose any of the values in quotes**.

|  |  |  |
| --- | --- | --- |
| **Input Macro Variable** | **Definition** | **Notes** |
| Projhome | Full filepath of main project folder |  |
| ID | Name of unique patient identifier |  |
| Minyear | First year that you want to output dementia dates | See pg. 5 for note on censoring. There should be enough time beyond the maxyear to allow for verification. The package will take into account all the years in the input datasets, even those outside of the minyear and maxyear range. If not enough data are provided to validate all the dates in the desired year range, then the output dataset will be labeled with ‘\_censor’ and dates in the dataset should be treated with caution. Dates from censored years will not be included in the ‘first’ variables. |
| Maxyear | Last year that you want output dementia dates |
| Dxclaims | List of processed input claims | This macro can take more than one claims data set as input separated by a space. A libref must be included with the data set name. If a permanent library is used, then the libname must be set in ‘input\_program.sas’. |
| Vyear | Desired years for verification window   1. 1-year verification 2. 2-year verification | Only accepts 1 or 2 as valid values |
| Outprefix | Desired prefix for annual output data sets. The format will be [outprefix][year] | A libref must be included with the data set name. If using a permanent library, then the libname must be set in ‘input\_program.sas’. |
| Userx | Whether or not drug data should be used  Y – Yes N – No | Only accepts Y or N as valid values |
| Rxclaims | If using drug data, then list of processed input drug claims | This macro can take more than one data set as input separated by a space. A libref must be included with the data set name. If a permanent library is used, then the libname must be set in ‘input\_program.sas’. |
| Hasgnn | If using drug data, then whether or not GNN or NDC is included  Y – GNN included  N – NDC included | Only accepts Y or N as valid values |

* 12. Run “input\_program.sas”. Check the log for errors.

With a successful run, you should have annual data sets for every year between your specified minyear and maxyear inclusive. See pg. 5 for the output codebook.

**Common Pitfalls**

The following are common issues you may run into when using this package:

* **Not including enough years of data** – See pg. 5 for note about censoring.
* **Including a range of years with different drug data availability** – You may have a range of years in which some have available drug data and some do not. The full range of years can be used and specified in ‘minyear’ and ‘maxyear’ macro variables and userx should be set to ‘Y’, but just note that you should disregard any variables for scenarios that use drug data to verify.
* **Not including librefs in your input data set names** **and output data set names**. **Not including corresponding libname statements.**

**Use Case: Calculating Incidence and Prevalence**

**Prevalence**

The calculation of prevalence requires merging the output with an enrollment file in order to ensure full availability of data during the study period. We recommend requiring continuous enrollment at least throughout the whole verification window until death, but this may depend on the study. After imposing enrollment restrictions, any of the following variables can be used to calculate a prevalence rate depending on which method you would like to use: **first\_dx\_dt, first\_dxsymp\_dt, first\_dxrx\_dt, first\_dxrxsymp\_dt, scen\_dx\_dt[year], scen\_dxsymp\_dt[year], scen\_dxrx\_dt[year], scen\_dxrxsymp\_dt[year]**. If you want to calculate prevalence based on ever having a valid dementia date, then you would want to use the **first** variables. If you want to calculate prevalence based on a valid dementia date in a calendar year or at a point in time, then it may be better to use the **scen** variables. The following examples are the steps one might take to calculate prevalence using diagnosis and symptoms in 2015:

**Desired outcome:** Prevalence in 2015 with 2-year verification window and looking only in 2015

1. Run algorithm using data from 2015 to 2017
   1. Data from 2016-2017 required for verification window
2. Merge and restrict output to a sample continuously enrolled from 2015 to 2017 or death
3. Calculate prevalence rate. Denominator is everyone in sample and numerator is everyone with a **scen\_dxsymp\_dt2015** value.

**Desired outcome**: Prevalence in 2015 with 2-year verification window and looking back to 2013.

1. Run algorithm using data from 2013 to 2017.
   1. Data from 2016-2017 required for verification window
2. Merge and restrict output to a sample continuously enrolled from 2013 to 2017 of death
3. Calculate prevalence rate. Denominator is everyone in sample and numerator is everyone with a **first\_dxsymp\_dt** value that occurs in 2015 or earlier.

**Incidence**

The calculation of incidence requires merging the output with an enrollment file in order to apply a “washout” period – a period in which a beneficiary is clear of any prior diagnoses. We commonly use 2 years as a washout period for dementia, which means for incidence in year *t,* we require continuous enrollment from year *t-2*, year ­*t-1*, and year *t* and no prior diagnosis of dementia prior to year *t*. To get a proper incidence rate, data from all of these years should be input into the model and data from the verification period as well. After imposing enrollment restrictions, any of the following variables can be used to identify incidence date depending on the method you want to use: **first\_dx\_dt, first\_dxsymp\_dt, first\_dxrx\_dt, first\_dxrxsymp\_dt**. This following example outlines the steps one might take to calculate incidence using diagnosis and symptoms.

**Desired outcome:** Incidence in 2015 with 2-year verification window and 2-year washout period

1. Run algorithm using data from 2013 to 2017
   * Data from 2013-2014 required for washout period
   * Data from 2016-2017 required for verification window
2. Merge and restrict output to everyone who is continuously enrolled from 2013 to 2017 or death
3. Use **first\_dxsymp\_dt** to drop everyone who had a diagnosis of dementia prior to 2015. You may find people who have a **first\_dxsymp\_dt** in 2013-2014. These people should excluded and the sample should now be an enrolled sample clear of dementia.
4. Calculate incidence rate in 2015. Denominator is everyone in above sample, numerator are those with a **first\_dxsymp\_dt** that occurs in 2015.

**Example**

A full example of the package run on synthetic claims data can be found in the /example/ folder.

**References**

Thunell, J., Ferido, P., & Zissimopoulos, J. (2019). Measuring Alzheimer's Disease and Other Dementias in Diverse Populations Using Medicare Claims Data. Journal of Alzheimer's Disease.

Haye S, Thunell J, Joyce G, Ferido P, Tysinger B, Jacobson M, Zissimopoulos J. (2023). Estimates of Diagnosed Dementia Prevalence and Incidence in Traditional Medicare and Medicare Advantage. Unpublished Manuscript.