

## National Dementia Workforce Study: Nursing Home Waves 1, 2 and 3 Sample Frames, At-A-Glance

The National Dementia Workforce Study aims to interview a nationally representative sample of frontline staff caring for residents with dementia in nursing homes. The sample frames for Waves 1, 2 and 3 of the survey included all federally certified nursing homes in the United States operating from 2021 to 2024 (Wave 1), from 2022 to 2024 (Wave 2), and from 2023 to 2025 (Wave 3).

This document describes the data and methods used to construct the three sample frames, and the SAS code used to construct the frames are available in the accompanying files (NDWS\_NH\_Sample\_Frame\_W1\_20250611.zip, NDWS\_NH\_Sample\_Frame\_W2\_20250611.zip, and NDWS\_NH\_Sample\_Frame\_W3\_20260512.zip)<sup>1</sup>. For those interested in running the code, we recommend using the Waves 2 or 3 versions because we updated the sections of the code from Wave 1 that processed the Minimum Data Set to make the code easier to follow.

### Data

To develop the nationwide sample frames of nursing homes that care for people living with dementia, we used data from the Centers for Medicare & Medicaid Services, including the Minimum Data Set and Nursing Home Compare. We also used data from Brown University's LTCFocus to obtain additional information about the nursing homes. The sample frame further used other publicly available sources, primarily for geocoding (see [Table 1](#)).

### Methods

Our overall approach to identifying the Wave 1 sample frame is available in the [steps to sample frame construction](#) section. The process involved two key steps, and within each step, there are multiple sub-steps (and multiple SAS programs) to implement the required logic and identify the sample frame. The following are the two key steps, and each has a link to the relevant section that provides additional details:

1. [Reading in, processing, and merging publicly available LTCFocus and Nursing Home Compare data](#)
2. [Processing the Minimum Data Set data, merging them with the LTCFocus and Nursing Home Compare data, and constructing the final sample frame](#)

For Wave 2, we streamlined and simplified several steps in the Wave 1 code. In addition, we fixed a trivial error in the Minimum Data Set code. We describe the updates to the code for Wave 2 in [Table 2](#). For Wave 3, we added new code to track in which waves we found each nursing home, as described in [Table 3](#).

<sup>1</sup> For readers without SAS, the programs can be read in any text editor. **Mac users can open the files with TextEdit but may need to go to the "Privacy & Security" menu in their System Settings to do so.**

## A. Background

The National Dementia Workforce Study (NDWS) aims to interview a nationally representative sample of frontline staff caring for residents with dementia in nursing homes. The sample frame for Wave 1 of the survey included all federally certified nursing homes in the United States operating between 2021 and 2024. Key facility characteristics required for sampling or nonresponse analyses included the number of federally certified beds, the average daily census of dementia patients, whether the facility had a dedicated Alzheimer’s disease special care unit, the percentage of residents with Medicaid as primary payer, and whether the facility was in a rural census tract. No one data source contained all the variables needed for the sample frame. For this reason, we used multiple data sources.

In this paper and accompanying documentation (SAS code used to develop the sample frame), we describe the data used to develop the sample frame (Section B) and the programming steps to construct the sample frame (Section C). We conclude with updates that we made to the data and code for Wave 2 and Wave 3 of the nursing home sample frame (Section D). *For users interested in running the code, we recommend using either the Wave 2 or Wave 3 version because we updated sections of the code that processes the Minimum Data Set (MDS) to make the code easier to follow.*

## B. Data

Table 1 describes the data sources used to construct the nursing home sample frame for Wave 1. For each source, we used the most recent available file at the time the work was conducted.

**Table 1. Data sources to construct the NDWS Wave 1 nursing home sample frame**

Data (years)	Key variables for sample frame construction	Access to data
Minimum Data Set (2020 and 2021)	<ul style="list-style-type: none"> <li>Average daily census of all residents in 2021</li> <li>Average daily census of residents with dementia in 2021</li> </ul>	Accessed via the Chronic Conditions Warehouse (CCW) Virtual Research Data Center (VRDC) under approved data use agreement (DUA)
LTCFocus facility-level file (2021)	<ul style="list-style-type: none"> <li>Whether the facility has an Alzheimer’s disease special care unit in 2021<sup>a</sup></li> <li>Percentage of facility residents whose primary support is Medicaid in 2021<sup>a</sup></li> </ul>	Publicly available with registration from Brown University: <a href="https://ltcfocus.org/">https://ltcfocus.org/</a>
Nursing Home Compare (January 2024 [most recent] and all months of 2021)	<ul style="list-style-type: none"> <li>Number of federally certified beds (using January 2024 data)</li> <li>Whether facilities found in 2021 MDS were also found in Nursing Home Compare January 2024 (that is, still in operation)</li> <li>Address data used to geocode facilities into urban or rural locations (using January 2024 data)</li> </ul>	Publicly available from the Centers for Medicare & Medicaid Services website: <a href="https://data.cms.gov/provider-data/data-set/4pq5-n9py">https://data.cms.gov/provider-data/data-set/4pq5-n9py</a>

Data (years)	Key variables for sample frame construction	Access to data
	<ul style="list-style-type: none"> <li>Average daily census of all residents (based on 2021 data; used to benchmark our average daily census calculations from the MDS to ensure accuracy)</li> </ul>	
U.S. Census Bureau’s 2010 TIGER/line shapefile	<ul style="list-style-type: none"> <li>Census tract polygon coordinates</li> </ul>	Publicly available from the IPUMS website: <a href="#">NHGIS Data Finder</a> <sup>b</sup>
Rural–Urban Commuting Area Codes (RUCA) 2010 file	<ul style="list-style-type: none"> <li>2010 RUCA code for each census tract</li> </ul>	Publicly available from the U.S. Department of Agriculture website: <a href="https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/">https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/</a> <sup>b</sup>

<sup>a</sup> This information was not available for facilities in Alaska or the District of Columbia because the LTCFocus facility-level file excludes facilities in Alaska and the District of Columbia.

<sup>b</sup> For wave 3, we also used shapefiles and RUCA codes based on the 2020 census tract definitions. This link provides access to the relevant files needed for both 2010 and 2020 census tract definitions.

We identified facilities in MDS, LTCFocus, and Nursing Home Compare based on their unique Medicare provider number, the Centers for Medicare & Medicaid Services Certification Number (CCN).

### C. Steps to sample frame construction

The sample frame resulted from two broad steps: (1) reading in, processing, and merging the publicly available sources (LTCFocus and Nursing Home compare) and (2) processing the MDS and merging to the publicly available sources. Within each step, several pieces of code implement the required data processing steps. The steps are described below. The code is available in the accompanying zip file, NDWS\_NH\_Sample\_Frame\_W1\_20250611.zip. For readers without SAS, the programs can be read in any text editor. **Mac users can open the files with TextEdit but may need to go to the "Privacy & Security" menu in their System Settings to do so.**

#### 1. Reading in, processing, and merging publicly available LTCFocus and Nursing Home Compare data

- [110 create ltcf file.sas](#). This program reads in the 2021 LTCFocus facility-level data set, limiting to variables needed for sample frame construction or that might be useful for data checks or nonresponse analyses. It also constructs numeric variables for whether each facility has an Alzheimer’s disease special care unit and whether the facility is for-profit or not.
- [120 create nhc21 file.sas](#). This program reads in all monthly Nursing Home Compare files for 2021 (this includes January through November 2021 and January 2022 because

there is no December 2021 file), limiting to variables needed to compare our calculation of facilities' average daily census using the MDS (in a downstream program) or that might be useful for other data checks. The code retains the most recent monthly record for each facility (it does not combine records for facilities that shared a zip code and name or a zip code and street address).

- [130 create\\_nhc24\\_file.sas](#). This program reads in the most recent publicly available month of Nursing Home Compare, including all variables available in the provider information file. The code constructs a numeric variable related to ownership type.
- [220 readin\\_ruca\\_xwalk.sas](#). This program reads in the most recent publicly available file from the U.S. Department of Agriculture that maps census tracts to Rural–Urban Commuting Area codes (RUCAs; available from [Rural–Urban Commuting Area Codes](#)). The program reads in the file from Excel, renames the variables in the file, and outputs the data in a SAS data set.
- [235 readin\\_shp\\_file.sas](#). This program uses proc mapimport to read in a shapefile<sup>1</sup> containing census tract boundaries for all U.S. census tracts based on the 2010 census (we use the 2010 census because the most current RUCA codes are still based on the 2010 census) and outputs the data in a SAS data set.
- [236 xform\\_shp\\_file.sas](#). This program uses proc gproject to transform the census tract boundaries in the shape file from census tract polygon coordinates to latitude and longitude coordinates and outputs the data in a SAS data set.
- [237 id\\_nhc24\\_tracts.sas](#). This program uses the most recent publicly available month of Nursing Home Compare (read in from the 130 program) to create geographic variables. Specifically, the code uses the provider address information to create a variable for state Federal Information Processing Series (FIPS) code and uses proc ginside to geocode each facility to the census tract where it is located.
- [240 create\\_nhc24\\_rural.sas](#). This program uses the geocoded, most recent month of publicly available Nursing Home Compare data (the output file from the 237 program) to assign facilities to RUCA codes and to construct a binary rural variable based on the RUCA values. It uses the publicly available 2010 census tract to RUCA code crosswalk available from the U.S. Department of Agriculture ([USDA ERS - Documentation](#)). In this program, the code also crosswalks the Social Security Administration (SSA) county codes from Nursing Home Compare to FIPS county codes using the publicly available crosswalk from the National Bureau of Economic Research ([SSA to FIPS State and County Crosswalk](#)).

---

<sup>1</sup> See the ArcMap website for a definition of shapefiles: <https://desktop.arcgis.com/en/arcmap/latest/manage-data/shapefiles/what-is-a-shapefile.htm>. Last accessed November 12, 2024.

- 310 merge ltcf nhc21.sas. This program merges the LTCFocus file that was read in and processed in the 110 program to the 2021 Nursing Home Compare file that was read in and processed in the 120 program. The code retains all facilities post-merge, even if they did not have a record in both input files.
- 320 merge final.sas. This program merges the previously merged LTCFocus and 2021 Nursing Home Compare data (from the 310 program) with the geocoded, most recent month of Nursing Home Compare data (from the 240 program). It retains all facilities post-merge, even if they did not have a record in both input files, and creates an indicator for which file(s) the facility was found in.

## **2. Processing the MDS data, merging to the LTCFocus and Nursing Home Compare data, and constructing final sample frame**

The MDS code for the Wave 1 sample frame development work contains five programs—400, 410, 420, 430, and 440—described in detail below.

- 400 process mds.sas. This program reads in the most recent two years of MDS data (2020 and 2021), limiting to key variables needed for sample frame construction or that might be useful for data checks or nonresponse analyses. The code then constructs flags for whether the MDS assessment contains evidence that the resident (1) had dementia (using items I4200\_ALZHMR\_CD = 1 or I4800\_DMT\_CD = 1), (2) died (using A0310F\_ENTRY\_DSCHRG\_CD = 12), and (3) was in a swing bed unit (using A0200\_PRVRDR\_TYPE\_CD=2). We did not find any records for swing bed units. This portion of the code also drops a small number of MDS assessments with missing CCNs.
- 410 avg daily census.sas. This program reads in the MDS files created in the 400 program as well as the combined LTCFocus and Nursing Home Compare file from the 320 program. The program implements the following steps:
  - Using the 2020 and 2021 MDS, the program calculates each facility's average daily census of residents in 2021, both overall and limited to those persons living with dementia (PLWD) as follows: it assigns a unique stay identification number to all MDS records for the same resident at the same facility (defined by CCN) that were part of the same stay in 2021 (that is, the stay either began in 2020 or earlier or it began with an admission record in 2021 and ended in a discharge or did not end in 2021). The code then collapses the file to one record per resident per stay per facility (a resident might have multiple stays at the same facility, in which case they would have multiple records for that CCN). The collapsed file includes variables for the length of stay in days, whether the stay was for a PLWD, and the race and ethnicity of the resident. The code then collapses the file once more to the facility level, summing the number of days across all stays as well as the number of days among PLWD and by race and ethnicity across all stays and

across all stays for PLWD. For each facility, the code divides the number of days by 365 to calculate the average daily census of all residents and for PLWD and for both groups by race and ethnicity.

- The program merges the processed MDS data to the file from the 320 program with LTCFocus and Nursing Home Compare data. The code then identifies CCNs from the 2021 MDS that did not merge to a CCN in the 2024 Nursing Home Compare data but for which the unmerged records from the 2021 MDS and 2024 Nursing Home Compare share the same street address and zip code, suggesting that the facility changed ownership between 2021 and 2024. The code also flags CCNs that share the same street address and zip code or zip code and facility name and assigns them a group identification number to easily identify facilities that might be part of the same campus. The code then limits the file to facilities that were found in both MDS 2021 and Nursing Home Compare 2024 and to the MDS-derived variables. We limit the file to MDS-derived variables to expedite the VRDC review process. (As described in the 430 program below, we re-merge the variables from LTCFocus and Nursing Home Compare onto this file by CCN locally.)
- 420 avg daily census suppress.sas. This program formats values of MDS-derived variables created in the prior program with "< 11" if the value is less than 11 or with an asterisk ("\*") if the value needs to be suppressed to prevent analysts from having the ability to back out the value of another variable with a cell size less than 11. The program masking\_formats.sas is included in the program to apply the small cell size formats. The Excel file created at the end of this program is downloaded from the VRDC to prepare for combining it with the file from the 320 program with LTCFocus and Nursing Home Compare data.
- 430 merge w vrdc.sas. This program reads in the Excel file containing the MDS-derived, facility-level variables downloaded from the VRDC and merges it with the file created in the 320 program by CCN to add the LTCFocus and Nursing Home Compare variables.
- 440 sample frame.sas. In this program, the code moves variables that will be most important for sampling to be the first variables listed in the file. It also removes variables that are not needed (primarily address-related variables from Nursing Home Compare 2021 and LTCFocus). The code also drops duplicate records for two facilities (all variables contained the same values across the two records for the facility, including name and address, except for a few staffing-related variables). The code then identifies facilities located in U.S. territories and removes them from the sample frame file. The code also creates the final set of missingness indicators. These include .M (for "missing") if a facility is missing information for a particular variable; .E (for "eleven") if a cell count is smaller than 11 (and had < 11, per program above); and .S (for "suppressed") if a cell was suppressed (formatted with an asterisk in the prior program) because otherwise one

could back out the value of a variable with a cell count smaller than 11. Finally, the program outputs an Excel version of the sample frame.

### D. Updates made to the code for subsequent years of the nursing home survey

In Table 2, we describe updates made to the Wave 1 nursing home sample frame code for Wave 2. For Wave 2, we continued to use 2021 LTCFocus (still the most recent file available), but we updated to use 2022 (full year) and September 2024 (most recent) Nursing Home Compare data and 2022 MDS data. The code is available in the accompanying zip file, NDWS\_NH\_Sample\_Frame\_W2\_20250611.zip.

**Table 2. Changes to the Wave 1 code for the Wave 2 sample frame**

Program	Change log
<b>Programs to read in and process LTCFocus and Nursing Home Compare data</b>	
110_create_ltcf_file.sas	<ul style="list-style-type: none"> <li>Updated to use a year macro so we can update to use the most recent LTCFocus file.</li> </ul>
120_create_nhc22_file.sas	<ul style="list-style-type: none"> <li>Updated to read in Nursing Home Compare 2022 data.</li> </ul>
130_create_nhc24_file.sas	<ul style="list-style-type: none"> <li>Updated to read in Nursing Home Compare data for September 2024 (the most recent available at the time we developed the Wave 2 sample frame). We also updated the set of variables read in from the raw data to account for three new variables added to the Nursing Home Compare data since the Wave 1 sample frame.</li> </ul>
237_id_nhc24_tracts.sas	<ul style="list-style-type: none"> <li>Updated to use the most recent Nursing Home Compare data (September 2024).</li> </ul>
240_create_nhc24_rural.sas	<ul style="list-style-type: none"> <li>Updated to use the most recent publicly available SSA to FIPS county crosswalk from the National Bureau of Economic Research.</li> </ul>
310_merge_ltcf_nhc22.sas	<ul style="list-style-type: none"> <li>Updated to use the most recent data (from programs above).</li> </ul>
320_merge_final.sas	<ul style="list-style-type: none"> <li>Updated to use the most recent data (from programs above).</li> </ul>
<b>Programs to read in and process MDS data, merge to the LTCFocus and Nursing Home Compare data, and construct final sample frame</b>	
000_nhsf_driver.sas (new)	<ul style="list-style-type: none"> <li>Created a new program that sets all the parameters needed in this section of code—for example, the years of MDS data to be used—as well as code to run the programs described below. It facilitates annual updates to the code by a user only having to update the input data and years once in this program. The driver also includes a program (<code>_masking_macro.sas</code>) with formats and macros for masking output if there are fewer than 11 units in key variables or output that otherwise must be suppressed.</li> </ul>
400_process_mds.sas	<ul style="list-style-type: none"> <li>Updated to read in the relevant years of MDS based on the settings in the driver program. The code now uses a do loop to read in MDS from the earliest year to the latest year (2021 to 2022 for Wave 2).</li> </ul>
410_avg_daily_census.sas	<ul style="list-style-type: none"> <li>Updated to use the relevant years of MDS based on the settings in the driver program (2021 and 2022 for Wave 2).</li> <li>Updated the line of code that erroneously set the end date of some stays to the end of the measurement period instead of the assessment date plus 150 days (this affected less than 1 percent of MDS assessments).</li> <li>Moved the code that deduplicated records in the merged LTCFocus and Nursing Home Compare file into this program (in Wave 1, this step was in the <code>440_sample_frame.sas</code> code).</li> <li>Updated the merge between the MDS and the file containing data from LTCFocus and Nursing Home Compare to simplify the subsequent steps to identify facilities that likely changed ownership. The Wave 2 code now outputs the merged records into different data sets rather than keeping all facilities in the merged file and using flags</li> </ul>

Program	Change log
	<p>to limit the file to the relevant records for identifying facilities that changed ownership. Now the code cross-checks the data set with MDS records that didn't merge to the most recent Nursing Home Compare data to the data set with the most recent Nursing Home Compare data that didn't merge to the MDS to identify records with the same street address and zip code that likely changed ownership. It then merges these records together and stacks them into the file of CCNs that merged between MDS and Nursing Home Compare 2024.</p> <ul style="list-style-type: none"> <li>Fixed an error from Wave 1—namely, although the Wave 1 code accurately identified CCNs that changed ownership (n = 22), it did not retain these facilities in the sample frame output. The Wave 2 code has been updated to append these facilities that changed ownership to the set of facilities in the sample frame.</li> <li>Moved the code that identified and dropped CCNs located in U.S. territories into this program (in Wave 1, this step was in the 440_sample_frame.sas program).</li> <li>Moved the code that suppressed small cell sizes and added meaningful missingness indicators (.M, .E, .S) to this program (in Wave 1, the small cell size suppression step was in the 420_avg_daily_census_suppress.sas program and the meaningful missingness indicators were added in program 440_sample_frame.sas).</li> <li>Removed the step to drop the LTCFocus and Nursing Home Compare variables from the MDS sample frame file before downloading from the VRDC and to now output a SAS file for downloading instead of Excel. This eliminated the need to re-merge the LTCFocus and Nursing Home Compare variables onto the file after downloading the MDS derived facility census and average daily count variables from the VRDC. (Thus, we did not need to use the 430 program from Wave 1 for Wave 2.)</li> </ul>
<p>420_nh_sample_frame.sas (formerly 440_nh_sample_frame.sas)</p>	<ul style="list-style-type: none"> <li>Identified facilities that were in both the Wave 1 sample frame and the Wave 2 sample frame but changed CCNs in the interim. The code flags these facilities and puts the Wave 1 CCN onto the Wave 2 record so that the sampling team can track facilities over time.</li> </ul>

In Table 3, we describe updates made to the Wave 2 nursing home sample frame code for Wave 3. Briefly, for Wave 3, we updated the input files to use 2022 LTCFocus, and 2023 (full year) and July 2025 (most recent) Nursing Home Compare data, and 2023 MDS data. We also updated the geocoding programs to assign facilities to census tracts based on both 2010 census tracts, consistent with waves 1 and 2, as well as newly available 2020 census tract definitions. Similarly, we continued to assign facilities to 2010 RUCA codes based on 2010 census tracts, also consistent with the earlier waves, in addition to assigning facilities to 2020 RUCA codes based on 2020 census tracts. The code is available in the accompanying zip file, NDWS\_NH\_Sample\_Frame\_W3\_20260512.zip.

**Table 3. Changes to the Wave 2 code for the Wave 3 sample frame**

Program	Change log
<b>Programs to read in and process LTCFocus and Nursing Home Compare data</b>	
110_create_ltcf_file.sas	<ul style="list-style-type: none"> <li>Set the year macro to 2022 instead of 2021 to use the most recent LTCFocus data</li> </ul>
120_create_nhc23_file.sas	<ul style="list-style-type: none"> <li>Updated to read in Nursing Home Compare 2023 data (instead of 2022).</li> </ul>
130_create_nhc25_file.sas	<ul style="list-style-type: none"> <li>Updated to read in Nursing Home Compare data for July 2025 (the most recent available at the time we developed the Wave 3 sample frame). We also updated the</li> </ul>

Program	Change log
	set of variables read in from the raw data to account for three variables that were dropped from Nursing Home Compare data since the Wave 2 sample frame.
220_readin_ruca_xwalk.sas	<ul style="list-style-type: none"> <li>No changes to this program since Wave 2. It reads in the 2010 RUCA data</li> </ul>
221_readin_ruca_xwalk_2020.sas	<ul style="list-style-type: none"> <li>This program is new. In Wave 2, we read in the 2010 RUCA-based definitions in the 220 program. This program mirrors the 220 program, but it reads in the 2020-based RUCA definitions</li> </ul>
235_readin_shp_file.sas	<ul style="list-style-type: none"> <li>Updated to read in 2010 and 2020 shape files of US census tracts</li> </ul>
236_xform_shp_file.sas	<ul style="list-style-type: none"> <li>Updated to transform 2010 and 2020 census tract boundaries to latitude and longitude coordinates</li> </ul>
237_id_nhc25_tracts.sas	<ul style="list-style-type: none"> <li>Updated this program to run it twice, once mapping facilities' addresses to 2010 Census-based census tract variables and again to map to 2020 Census-based census tract definitions. Also updated to run on the most recent Nursing Home Compare data (July 2025).</li> </ul>
240_create_nhc25_rural.sas	<ul style="list-style-type: none"> <li>Updated to use the most recent Nursing Home Compare data (July 2025) and to assign facilities to RUCA codes based on both 2010 and 2020 census tract definitions (In Waves 1 and 2, we used 2010 definitions only)</li> </ul>
310_merge_ltcf_nhc23.sas	<ul style="list-style-type: none"> <li>Updated to use the most recent data (from programs above).</li> </ul>
320_merge_final.sas	<ul style="list-style-type: none"> <li>Updated to use the most recent data (from programs above).</li> </ul>
<b>Programs to read in and process MDS data, merge to the LTCFocus and Nursing Home Compare data, and construct final sample frame</b>	
000_nhsf_driver.sas	<ul style="list-style-type: none"> <li>Updated to use the most recent data (2022 and 2023 MDS data and 2023 and 2025 Nursing Home Compare). The includes.sas program contains macros for producing the SAS log and lst files.</li> </ul>
400_process_mds.sas	<ul style="list-style-type: none"> <li>Updated to keep the race_wht_cd variable, which was populated in the MDS starting in October 2023 (<a href="#">Race and Ethnicity - Changes in Section A   MDS Consultants</a>) and replaced the original a1000f_wht_cd variable.<sup>1</sup></li> </ul>
410_avg_daily_census.sas	<ul style="list-style-type: none"> <li>Updated to use race_wht_cd if a1000f_wht_cd contained missing data.</li> </ul>
420_nh_sample_frame.sas	<ul style="list-style-type: none"> <li>Updated code to track which facilities were found in which waves, and whether their CCN ever changed between Wave 1 and Wave 3. Created variables to retain prior CCN(s) so that the survey team can track facilities that have changed CCNs over time.</li> <li>Created 4 flags for whether the facility was found in (1) Wave 3 only; (2) Waves 1 and 3 only; (3) Waves 2 and 3 only; or (4) Waves 1, 2, and 3</li> </ul>

<sup>1</sup> In addition to the variable for white race, the MDS replaced all other race and ethnicity variables starting in October 2023. Our code only used the white variable to flag white versus non-white residents. Users who plan to modify the code to measure average daily census of dementia residents with other races or ethnicities would need to update the code to retain the relevant, new race and ethnicity variables.