

---

# CTDPH INTERIM JULY 1, 2020 ESTIMATES: METHODOLOGY

---

August, 2024

Demographic Analytics Advisors and Connecticut Department of  
Public Health



Prepared by:

Demographic Analytics Advisors and the Surveillance Analysis and Research Unit, Connecticut  
Department of Public Health

Christopher Dick  
Founder and CEO  
Demographic Analytics Advisors

Karyn Backus  
Epidemiologist 4  
Connecticut Department of Public Health

The authors gratefully acknowledge the contributions of the following staff of the Connecticut  
Department of Public Health and CT Data Collaborative:

Laura E. Hayes, PhD  
Data Scientist  
Connecticut Department of Public Health

Cynthia Willner  
Senior Research Associate  
CT Data Collaborative

Michelle Riordan-Nold  
Executive Director  
CT Data Collaborative

This publication was supported by funds from the Department of Health and Human Services (HHS) Centers for Disease Control and Prevention (CDC) Notice of Funding Opportunity (NOFO) number CK19-1904, entitled Epidemiology and Laboratory Capacity for Prevention and Control of Emerging Infectious Diseases (ELC) (Grant No. # 6NU50CK000524) with funds made available under the Coronavirus Preparedness and Response Supplemental Appropriations Act, 2020 (P.L. 116-123); the Coronavirus Aid, Relief, and Economic Security Act, 2020 (the “CARES Act”) (P.L. 116-136); the Paycheck Protection Program and Health Care Enhancement Act (P.L. 116-139); and/or the Consolidated Appropriations Act, 2021, Division M – Coronavirus Response and Relief Supplemental Appropriations Act, 2021 (P.L. 116-260). The contents of the publication are solely the responsibility of the authors and do not necessarily represent the views of the HHS.

**Suggested Citation:**

Dick, C. and Backus, K. (2024). *CTDPH Interim July 1, 2020 Estimates: Methodology*. Connecticut Department of Public Health, Hartford, CT  
(<https://portal.ct.gov/dph/health-information-systems--reporting/population/town-population-with-demographics>)

## Table of contents

<b>Background</b>	<b>1</b>
<b>Process and Methodology</b>	<b>2</b>
Accessing Required Data Sources . . . . .	2
Reallocate “SOR” from the 2020 Census . . . . .	3
Rake the April 1, 2020 data to July 1, 2020 Totals . . . . .	5
<b>Quality Checks</b>	<b>5</b>
<b>Analysis</b>	<b>5</b>
Race / Ethnicity: Comparisons to the 2020 Census . . . . .	6
Age and Sex: Comparisons to the 2020 Census . . . . .	7
<b>Appendix</b>	<b>9</b>
Race by County . . . . .	9
Age / Sex Differences by County . . . . .	14

## Background

The 2020 Census was a decennial census unlike any other. Due to a confluence of factors, including the COVID-19 pandemic, changes in decennial procedures, and the implementation of a new privacy protection regime called differential privacy, data products from the 2020 Census have been delayed. One of the delayed data products is the modified race file (MRF) or modified age, race, and sex file (MARS) which contains data by county with “Some Other Race” (SOR) realigned to match with the race categories needed for the Population Estimates Program (PEP)<sup>1</sup>. It is likely this file will be out in late-2024, with estimates that incorporate the file released from the Census Bureau in Vintage 2024 (released in 2025). The Connecticut Department of Public Health seeks to publish and use estimates based on current racial and ethnic distributions of Connecticut’s communities<sup>2</sup>. Given the continued delays from the Census Bureau, Connecticut has developed an interim set of July 1, 2020 estimates for all towns in Connecticut by 5-year age groups, sex, race, and Hispanic origin. These estimates are based on the following:

---

<sup>1</sup>The decennial census and the American Community Survey are the only two data products that are both required and allowed to use “SOR” as a category. For all other data products the Census Bureau uses race categories without a “SOR” option.

<sup>2</sup>Currently, the Census Bureau’s population estimates are based on race and ethnic distributions from the 2010 Census with changes applied for births, deaths, and domestic and international migration.

- Data by age, race, sex, and Hispanic origin (ASRH) for the 2020 Census (both the Redistricting file as well as the Demographic and Housing Characteristics file).
- Modified Race proportions from the 2010 Census MRF.
- Unmodified Race data from 2010 Census Redistricting File.
- Town population total estimates for July 1, 2020 from the Vintage 2021 Estimates.
- County population estimates by age and sex for July 1, 2020 from the Vintage 2021 Estimates.

## Process and Methodology

At the most basic level, to create a town-level population estimate by ASRH requires us to project the April 1, 2020 population from the 2020 Census and control it to the town totals from the Census Bureau’s annual postcensal population estimates as well as the county age and sex counts from the same. The challenge with this approach is that the 2020 Census data do not come in the race groups needed for our estimates. In past censuses, the MRF was available to Connecticut in time to produce July 1 estimates for 2000 and 2010. Given the ongoing delays for the 2020 MRF we have to create our own MRF for the 2020 data. To build these estimates we use R statistical software for the following process:

1. Access required data sources
2. Reallocate “SOR” from the 2020 Census
3. Rake the April 1, 2020 data to July 1, 2020 Totals (town) and Age / Sex (county)

## Accessing Required Data Sources

We require 6 data sources for the development of these interim estimates, and each is accessed directly from the Census Bureau API or their FTP site. For data collected from the Census API we use the R package “tidycensus” to make our work much more efficient and easier to understand and read. For data pulled from the FTP site we use built in functions to read csv files and point them at the web location of the data. Data are accessed as follows:

**Table 1. Data Accessed**

Data File	Accessed From	Physical Location or Filename
2010 Census Redistricting File	Census API	Filename on API: “pl”

Data File	Accessed From	Physical Location or Filename
2010 Modified Race File	Census FTP	<a href="https://www2.census.gov/programs-surveys/popest/datasets/2010/modified-race-data-2010/stco-mr2010_al_mo.csv">https://www2.census.gov/programs-surveys/popest/datasets/2010/modified-race-data-2010/stco-mr2010_al_mo.csv</a>
2020 Census Redistricting File	Census API	Filename on API: "pl"
2020 Demographic and Housing Characteristics File	Census API	Filename on API: "dhc"
Vintage 2021 Town Population Estimates	Census FTP	<a href="https://www2.census.gov/programs-surveys/popest/datasets/2020-2021/cities/totals/sub-est2021_9.csv">https://www2.census.gov/programs-surveys/popest/datasets/2020-2021/cities/totals/sub-est2021_9.csv</a>
Vintage 2021 County Age and Sex Population Estimates	Census FTP	<a href="https://www2.census.gov/programs-surveys/popest/datasets/2020-2021/counties/asrh/cc-est2021-agesex-09.csv">https://www2.census.gov/programs-surveys/popest/datasets/2020-2021/counties/asrh/cc-est2021-agesex-09.csv</a>

## Reallocate "SOR" from the 2020 Census

Reallocating "SOR" is the most involved part of this process. This is a process that the Census Bureau generally does, and they plan to do it for 2020 as well, but due to schedule delays it is necessary for us to create our own pseudo-MRF.

To create our pseudo-MRF, we recode all of those who responded "SOR" to one of the 5 OMB accepted race categories or to a category for 2 or more of those 5 race categories. We use the following rules, which match the Census Bureau's approach in 2010 to the extent possible:

1. People who responded "Hispanic" will be unchanged because we only report "Hispanic, all races"
2. People who only responded as one of the 5 OMB races, or a combination thereof, will be left the same.
3. People who responded "SOR" plus one of the other 5 OMB races will be recoded into the OMB race alone category (e.g., if someone responded "White" and "SOR" they will become "White Alone"). This is exactly what the Census Bureau did in 2010 and plans to do in 2020.
4. People who responded "SOR" plus two or more of the other 5 OMB races will be recoded into the "Two or More Races" category (e.g., if someone responded "White", "Black", and "SOR" they will be put into the two or more races category. Note that in the DHC file, this is already how they are coded, but in the PL it is not). This is exactly what the Census Bureau did in 2010 and plans to do in 2020.
5. For those who responded "SOR" only we will use the county-level proportions from the 2010 Modified Race File to re-allocate these people into one of the 5 OMB race categories or into two or more races. The Census Bureau uses the household-level data and imputes race with

those in their household or nearby to create these re-codes. Since we do not have access to individual household level data, the best we can do is use the proportions at the lowest level of geography provided (county) from the 2010 MRF.

The most straightforward way to do this would be to apply these rules directly to the 2020 DHC. Unfortunately, the 2020 DHC only provides the following:

1. "SOR Alone"
2. "SOR Alone or in Combination"
3. "Two or More Races"

We need to understand how many people in the "SOR Alone or In Combination" group are combined with one other OMB race versus two or more other OMB races; we use the 2020 PL data to disaggregate the needed race detail. We use the PL to understand the number of people that responded "SOR" with 1 additional race, "SOR" with two or more additional races, or "SOR Alone". These data, by town and age group (over/under 18) will be used to parcel out the "SOR in Combination" in the DHC. Next, we prepare the 2020 DHC for reallocation of "SOR" and then merge in the 2020 PL data. We then proportioned out the "SOR in Combination" group into two groups: "SOR" plus one OMB race versus "SOR" plus two or more other OMB races.

This process provides us with fractional or non-whole number of people to recode from "SOR + 1" to OMB Alone categories. These data must be rounded so that we end up with the same number of "SOR + 1" for each row and the same number of recodes to each group. We use a 2-way controlled round to ensure both row totals and column totals stay the same as when we started the process. The process begins with all columns unrounded, starts with the smallest race groups and rounds them, recalculates the difference between the control totals and unrounded data, moves to the next smallest column and does the same, and then iterates through all columns until we have rounded totals that equal the control totals. After rounding the data we add these people into the correct OMB race group.

Next, we allocate those who responded "SOR Alone" to one of the 5 OMB race groups or any combination thereof. To recode the people who responded "SOR" only, we use the 2010 MRF and the 2010 Redistricting data to understand the proportion, by county, of people who were recoded to each group (including multiple races) in 2010. We then apply those proportions to the 2020 data.

We merge the 2010 Redistricting data with the 2010 MRF and evaluate how each race group changes, by county, between each file. Since the 2010 MRF does not provide the data that includes "SOR" this is the only way to see what proportion of people were recoded into each category. We calculate proportions for each OMB race group, as well as combinations thereof, that are recoded from "SOR". We append these proportions, by county, to the 2020 DHC data and multiply these proportions by the "SOR Alone" group. These data are then rounded to ensure that the total number of people in each age group and county remains the same.

## **Rake the April 1, 2020 data to July 1, 2020 Totals**

We are now ready to inflate or deflate each of these values based on town-level total growth and the county-level change in age and sex (AS), between April 1, 2020 and July 1, 2020. Essentially, we are keeping the race and Hispanic origin distribution constant, while changing their totals based on the aggregate change for the town over those three months. To do this we:

1. Calculate the change in population by town and county AS between April 1, 2020 and July 1, 2020 from the Vintage 2021 postcensal population estimates (which include the Census data from April 1). We use iterative proportional fitting to keep both the totals and characteristics constant.
2. Use two-way controlled rounding to round the cells while ensuring the town totals and county AS totals match the July 1, 2020 data from the postcensal population estimates.
3. Apply the race and Hispanic origin proportions from the 2020 Census (by modified race) by town and AS groupings
4. Apply a controlled round to the outcome to make sure the town AS distribution remains undisturbed.

## **Quality Checks**

Given the nature of the 3-month controlled inflation / deflation estimation procedure, only a couple of quality checks are necessary. We undertook the following quality checks, all with positive results:

1. During the creation of the pseudo-MRF, we ensured that totals by age, sex, and Hispanic Origin did not change, while race did change.
2. During the creation of the pseudo-MRF, we ensured that totals by town and county did not change.
3. During the raking process from April 1, 2020 to July 1, 2020 we ensured that the raked and rounded population by ASRH equaled the 2020 population from Vintage 2021 at the town total level.
4. We checked the differences in the ASRH distribution between April 1, 2020 (modified race) and July 1, 2020 (modified race). All differences were very small, with the largest change in percentage being smaller than 1/10th of 1%.

## **Analysis**

Given the fact that these estimates incorporate the use of the 2020 Census race data, while the estimates from the Census Bureau for the same time period currently incorporate the use of a

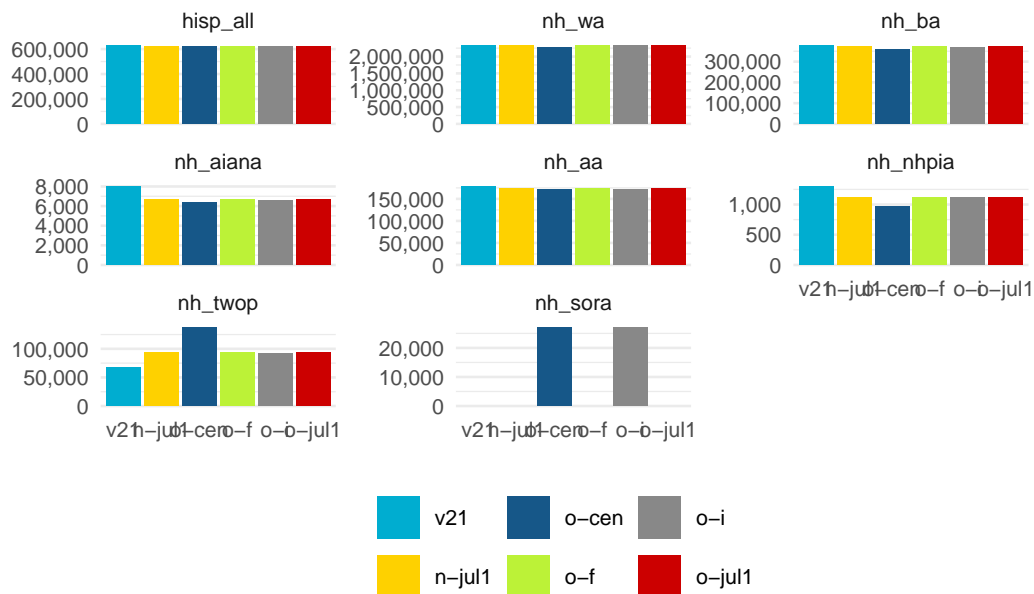
“Blended Base” with race characteristics from the 2010 Census-based postcensal population estimates, we analyze how race differs between the two data sets. Further, evaluate how race is different between the 2020 Census and our estimates based on the recoding of “SOR”. Finally, we assess how the age distribution differs between the data sources.

## Race / Ethnicity: Comparisons to the 2020 Census

First we assess how race changes between each of the steps of the process. In the chart below the following groups are included:

1. o-cen = Census 2020
2. o-i = recode of “SOR in Combination”
3. o-f = recode of all “SOR”
4. o-jul1 = f raked to July 1, 2020 town total populations from Vintage 2021 from the Census Bureau
5. n-jul1 = f raked to July 1, 2020 town total populations and county age / sex populations from Vintage 2021 from the Census Bureau
6. v21 = July 1, 2020 from the Vintage 2021 estimates for counties by ASRH from the Census Bureau

### Change in race coding by step





As we can see the population in “Two or more races” decreases as we recode the “SOR in Combination”. This is due to all of the people who reported “SOR” plus one other race who were then recoded to one of the other races alone. With the recoding of “SOR Alone” (the bar “o-f”) we see increases across all race groups, but they are smaller than the “SOR in Combination” recode step. Below we can see each of these comparisons in tabular format.

### Change in Race by Coding Step

source	hisp_all	nh_wa	nh_ba	nh_aiana	nh_aa	nh_nhpia	nh_twop	nh_sora
v21	630,417	2,324,950	378,756	8,015	178,541	1,301	67,545	NA
n-jul1	623,799	2,328,480	373,255	6,695	173,732	1,117	93,182	NA
o-cen	623,293	2,279,232	360,937	6,404	170,459	974	137,569	27,076
o-f	623,293	2,335,297	373,228	6,688	173,370	1,115	92,953	NA
o-i	623,293	2,317,357	366,529	6,591	171,380	1,113	92,605	27,076
o-jul1	622,451	2,331,136	372,638	6,688	173,307	1,115	92,925	NA

Looking at both the table and the chart, we can also make comparisons between Vintage 2021, which uses the Census Bureau’s “Blended Base” versus the interim DPH estimates. There are a few differences between each of these approaches, but the main one is that the interim DPH estimates use race from Census 2020, whereas the race data in the blended base is still coming mainly from Census 2010. The differences are stark for the two or more races group (though not as stark as they are between the 2020 Census and Vintage 2021). This tells us two things: (1) There are more people who define themselves as two or more races in 2020 than there were in 2010, and (2) the very large difference reported in 2020 is somewhat mitigated by recoding those who responded (or were coded as) “SOR in Combination”.

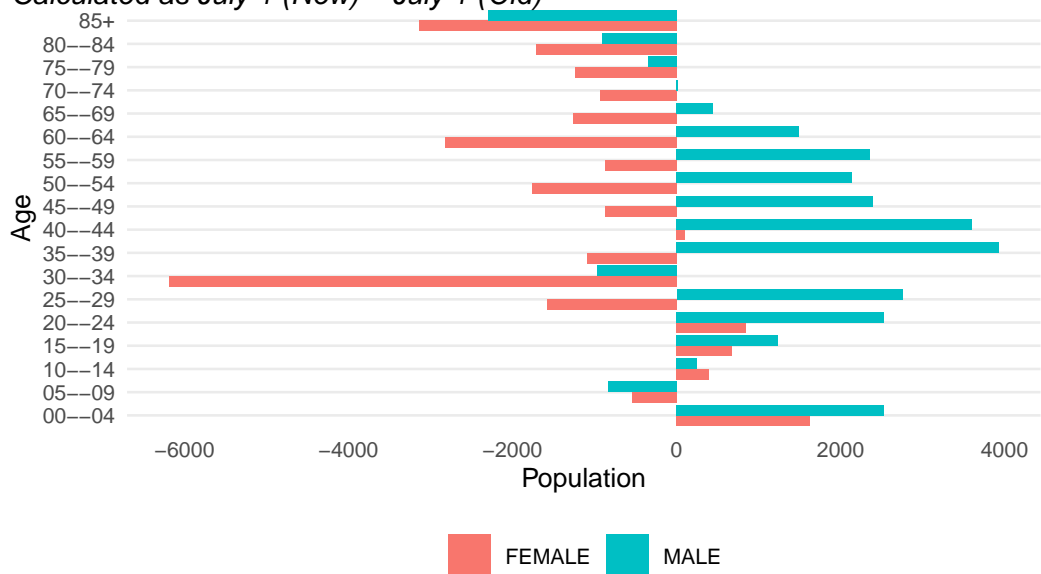
Results by county are similar and are included in the Appendix.

### Age and Sex: Comparisons to the 2020 Census

These July 1 estimates are controlled, by AS, to the Vintage 2021 county estimates, meaning data by AS match Vintage 2021 rather than the 2020 Census. Below we explore the differences between our new July 1 Estimates and a set of July 1 Estimates based on AS from the 2020 Census. This allows for an apples to apples comparison of how the age structure differs, on the same date, with an AS control to the Vintage 2021 county estimates, versus a control based on the 2020 Census.

## Population Differences: State by Age

*Calculated as July 1 (New) – July 1 (Old)*

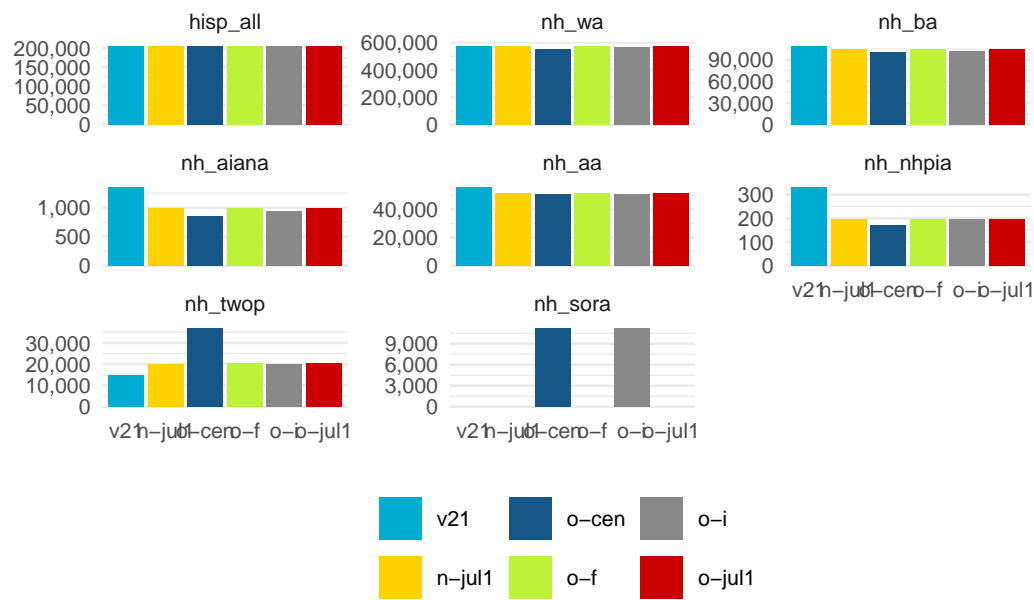


Results by county are included in the appendix.

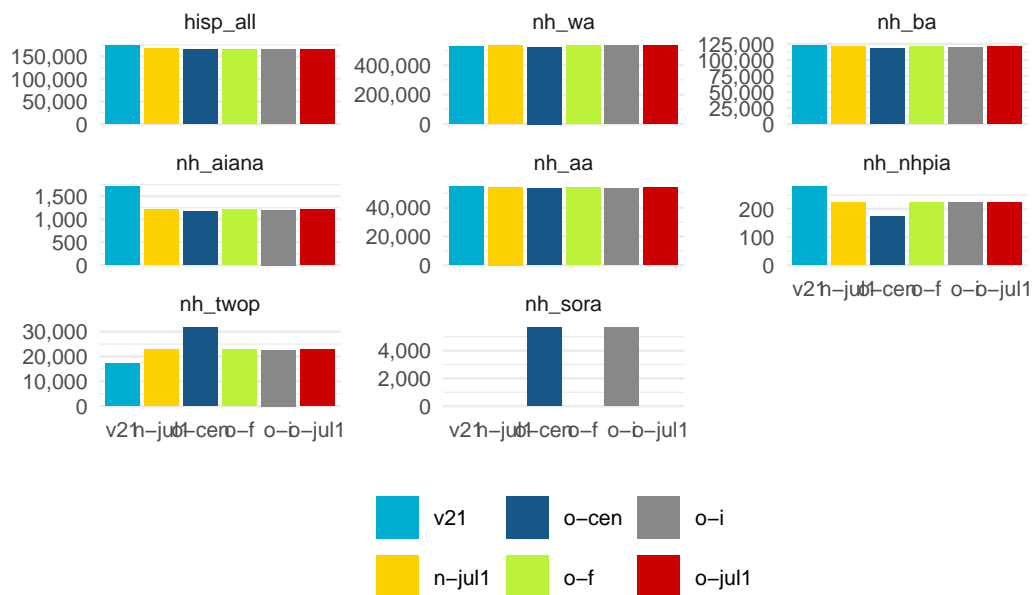
Appendix

Race by County

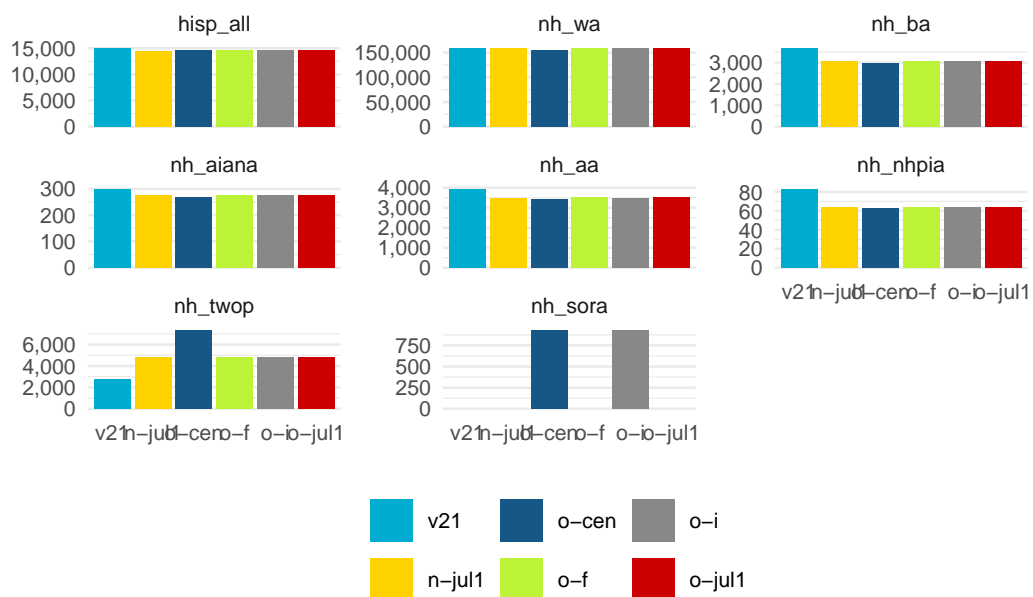
Change in race coding by step: Fairfield County



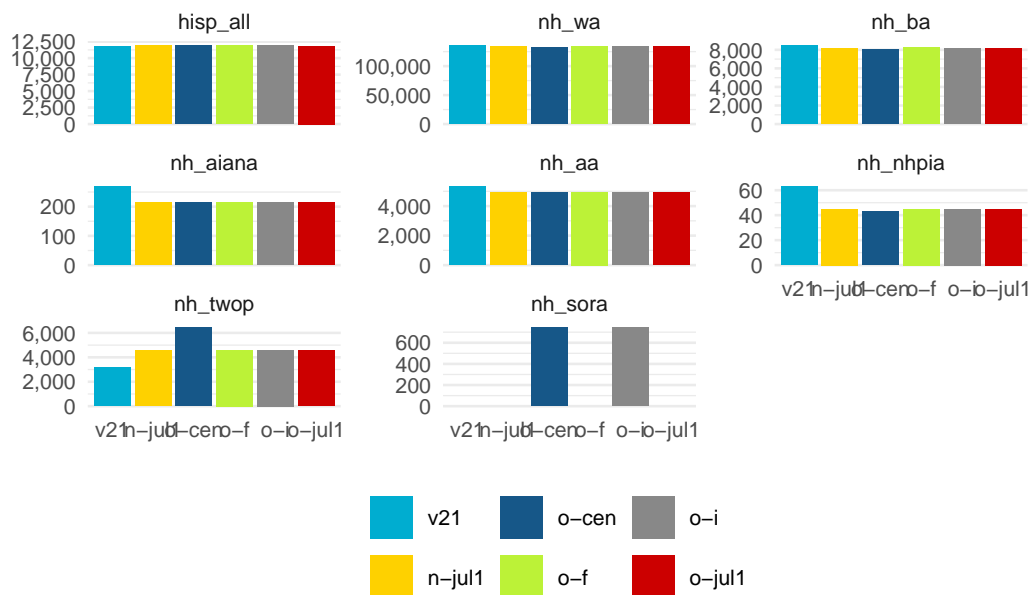
## Change in race coding by step: Hartford County



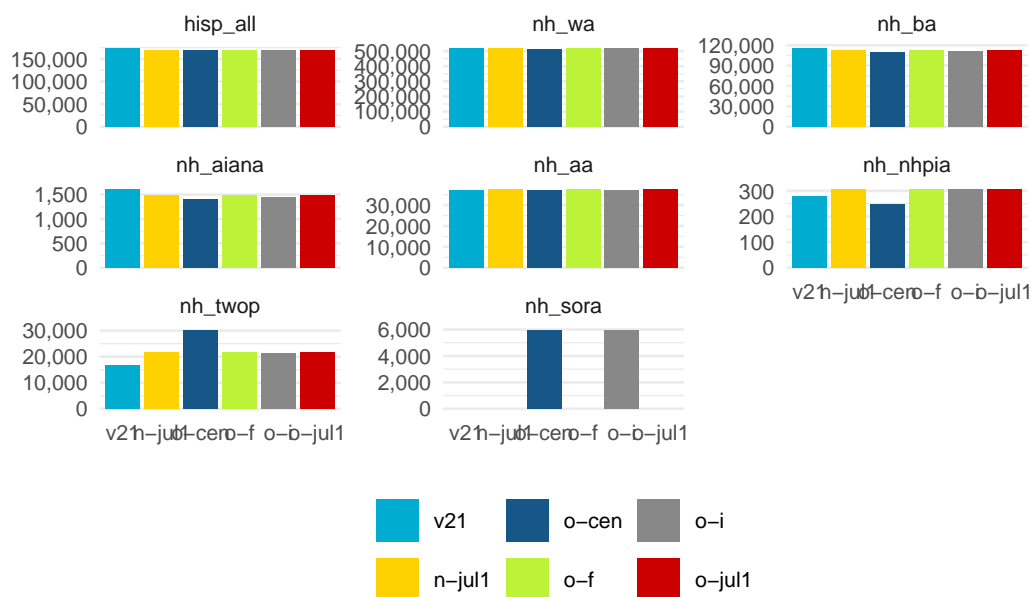
## Change in race coding by step: Litchfield County



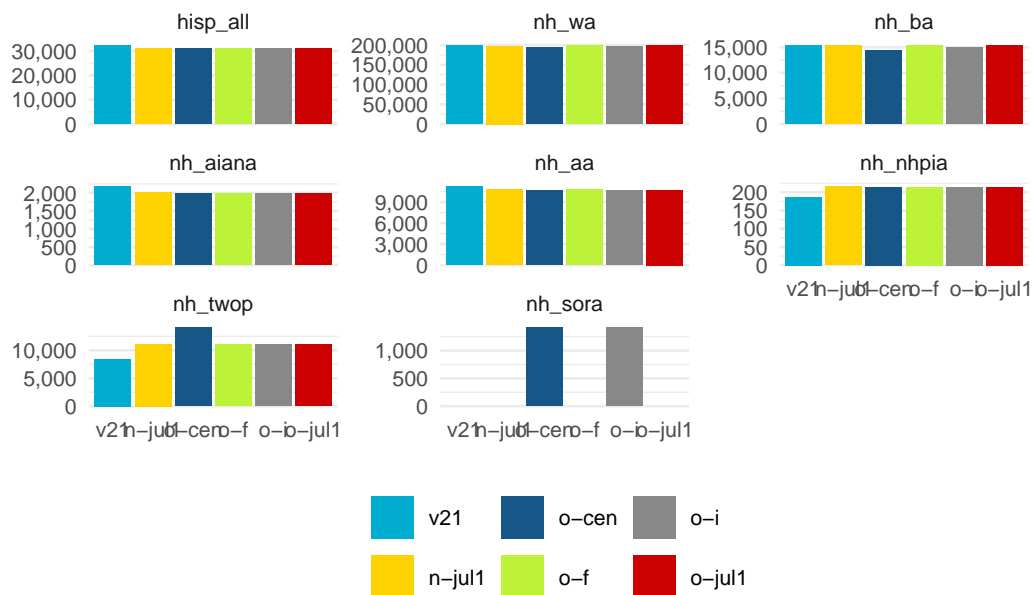
## Change in race coding by step: Middlesex County



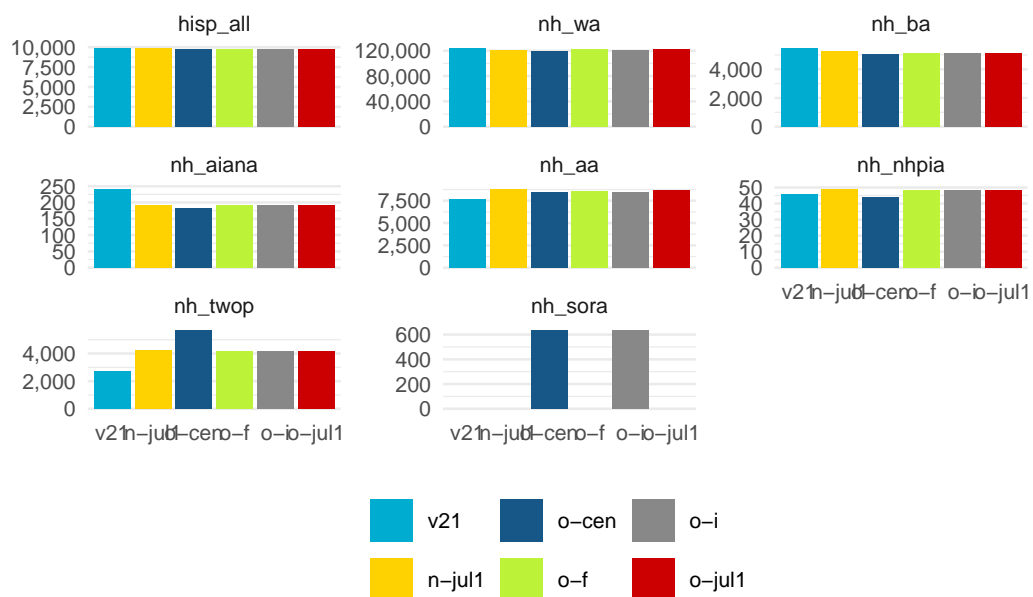
## Change in race coding by step: New Haven County



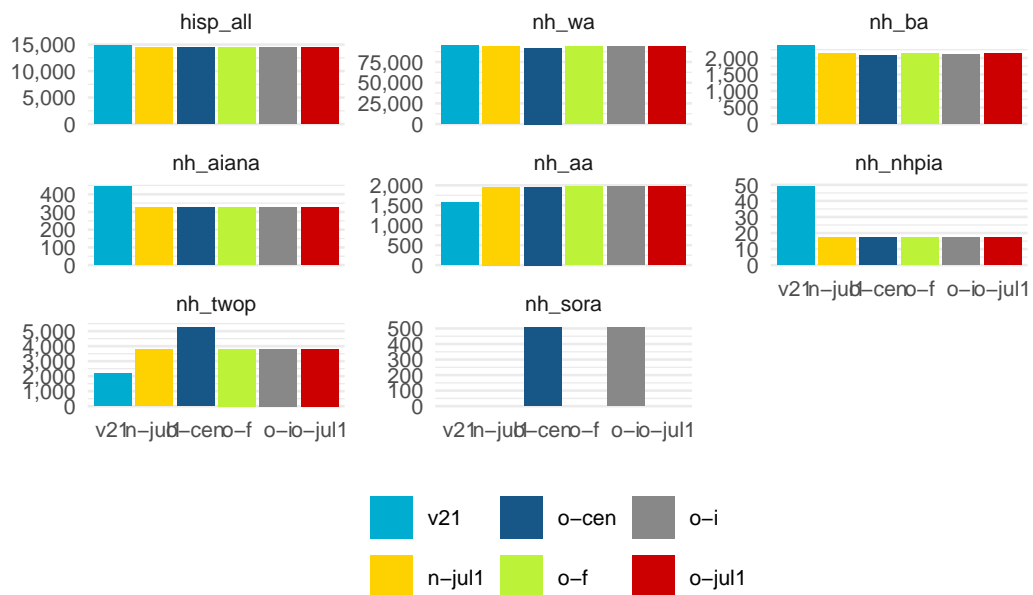
## Change in race coding by step: New London County



## Change in race coding by step: Tolland County



## Change in race coding by step: Windham County



## Age / Sex Differences by County

### Population Differences: Fairfield County by Age

Calculated as July 1 (New) – July 1 (Old)



### Population Differences: Hartford County by Age

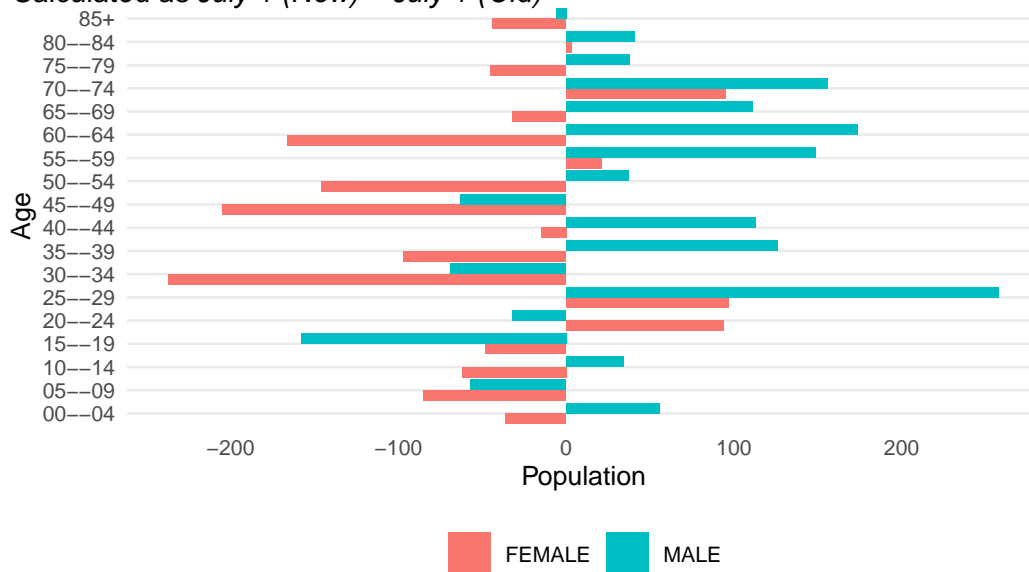
Calculated as July 1 (New) – July 1 (Old)





## Population Differences: Litchfield County by Age

Calculated as July 1 (New) – July 1 (Old)



## Population Differences: Middlesex County by Age

Calculated as July 1 (New) – July 1 (Old)



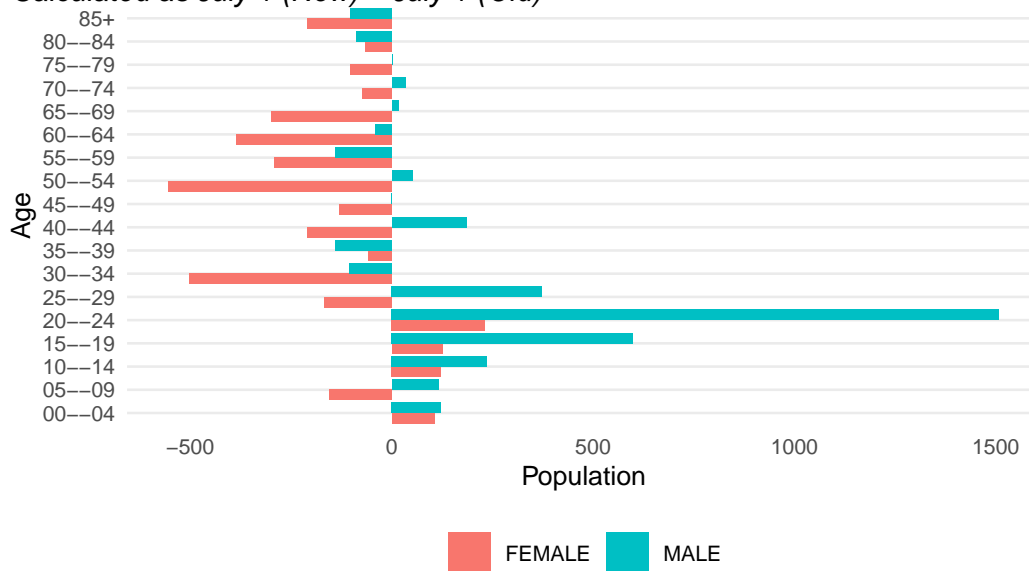
## Population Differences: New Haven County by Age

Calculated as July 1 (New) – July 1 (Old)



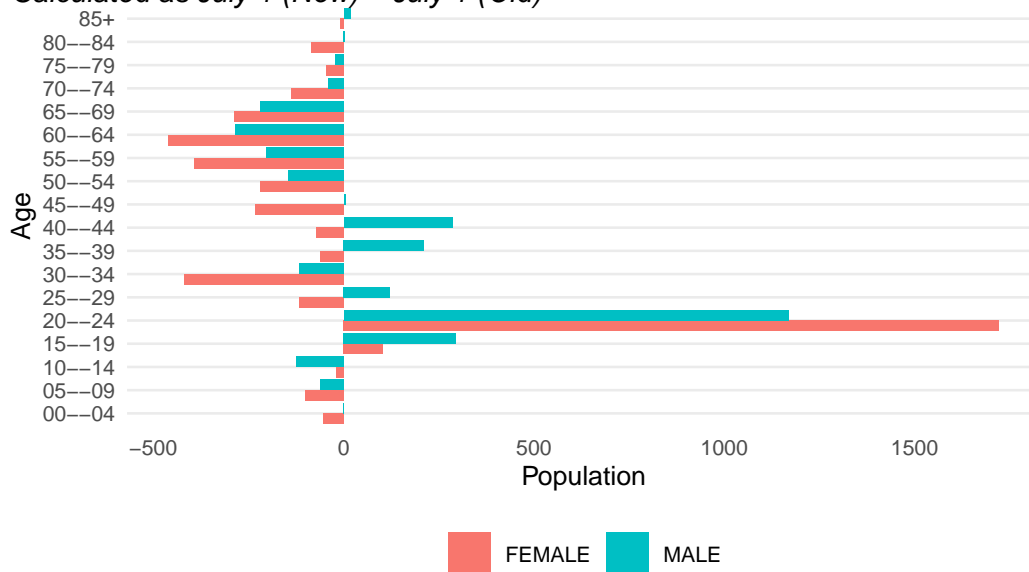
## Population Differences: New London County by Age

Calculated as July 1 (New) – July 1 (Old)



## Population Differences: Tolland County by Age

Calculated as July 1 (New) – July 1 (Old)



## Population Differences: Windham County by Age

Calculated as July 1 (New) – July 1 (Old)

