TECHNICAL PAPER

Cohort Creation and Characterization Using SAS® Health: Cohort Builder

Last update: September 2021

Contents

Relevant Products and Releases

- SAS® Health: Cohort Builder 2.1

Problem Definition

Researchers are continually searching for ways to increase productivity and efficiency. Using SAS® Health: Cohort Builder, you can analyze data from multiple data sources and vendors with unique analytic capabilities to address health care and life science use cases. Using the insight gained, users can make decisions about safety, efficacy, effectiveness, and costs.

Overview

Diabetes is a chronic health condition that we are familiar with. This health condition is fast-growing within populations globally as are the risk factors of diabetes-related complications. In the last decade, the prevalence of diabetes has increased drastically. According to the Centers for Disease Control and Prevention, 34.2 million U.S. adults have diabetes. Of these 34.2 million, one in five does not know that they currently have diabetes since the disease is often overlooked until serious health conditions arise ("What is Diabetes?", 2020). SAS Health: Cohort Builder can assist researchers define a cohort of diabetic patients quickly and with improved quality. You can use Cohort Builder to create, define, and analyze the cohort. The user can save definitions for later use with data, different cohorts, and comparisons. We want to build a cohort that consists of diabetic patients who have experienced a stroke, hypertension, and additional criteria. By creating this cohort with Cohort Builder, you can quickly identify the percentages of patients who fit the criteria. After building the cohort, the user can then use the Add-in Manager to create custom add-in templates for the cohort that is created and defined. Using the analytical add-ins, the user can specify information, code, user interface functionality, and execution controls and parameters to analyze the patient population.

Application and Data Description

A cohort is a patient population that satisfies the user-specified inclusion criteria over a period. It is important to define and design the cohort well enough to produce useful results. The data source that is used within the cohort contains sensitive patient data that is safeguarded with regulations to protect the privacy of medical records.

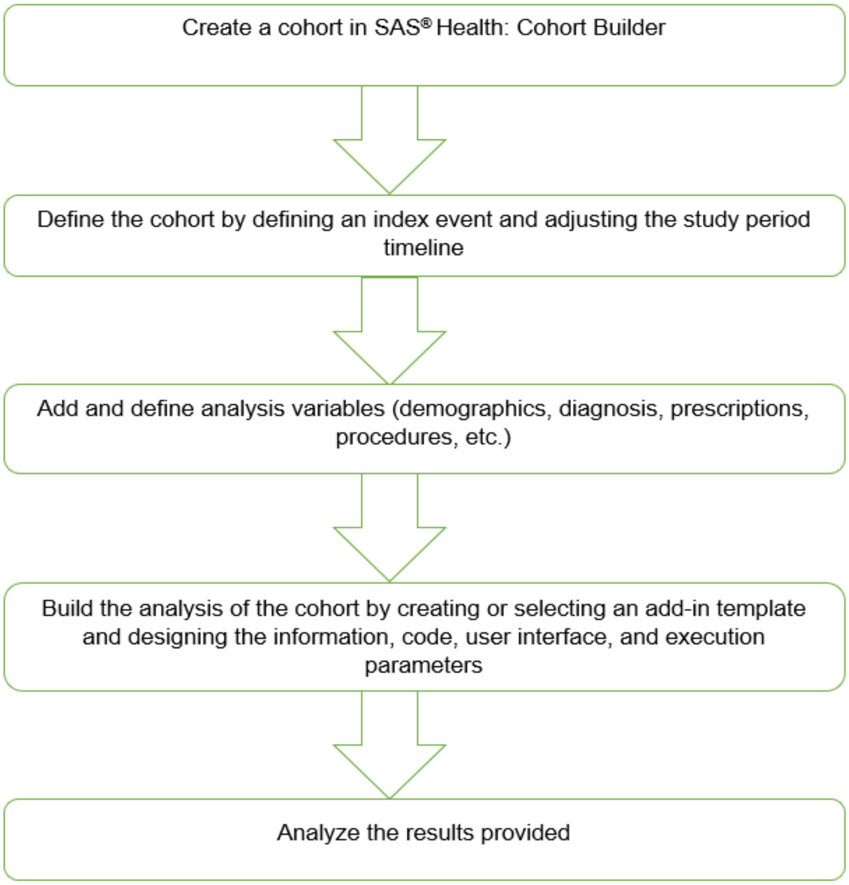
All cohorts that are created fall under two categories, which are defined in *SAS Health: Cohort Builder*.

- Population cohort – examines a population and variable relationships. This type of study is used to identify potential risk factors in the development of the disease as well as to inform the design of randomized clinical trials.

- Index event cohort – identifies populations that are based on the occurrence of an index event, which is a diagnosis, treatment, or procedure with associated criteria. The date on which the index event occurred is identified for each member of the cohort, allowing for other criteria to be conditional on their occurrence relative to the index event date. This type of study enables the analysis of certain clinical events that commonly occur within a subset of cohort members.

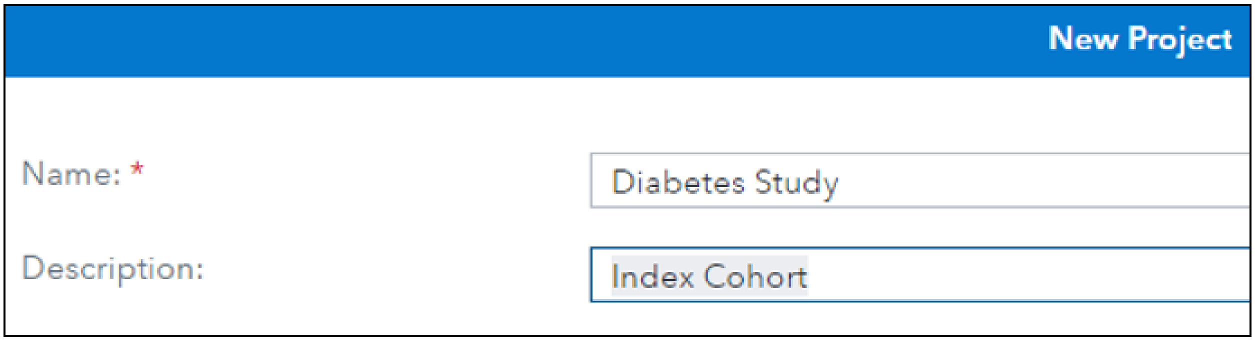
The process of defining an index event cohort and using Cohort Builder to analyze data is depicted as follows:

Figure 1. *Process to Define an Index Event*


Create a cohort in SAS® Health: Cohort Builder
Define the cohort by defining an index event and adjusting the study period timeline
Add and define analysis variables (demographics, diagnosis, prescriptions, procedures, etc.)
Build the analysis of the cohort by creating or selecting an add-in template and designing the information, code, user interface, and execution parameters
Analyze the results provided

The first step in using Cohort Builder is to create a project that contains all relevant cohorts. The user is then able to begin creating cohorts and editing existing cohorts. By selecting the tab to create a new cohort, the following pop-up appears. It enables the user to name, describe, and select a data source for the cohort.

Figure 2. *Specifying Information to Create a New Cohort*



New Project

Name: * Diabetes Study

Description: Index Cohort

After naming and selecting a data source for the cohort, the user is then prompted to define the cohort using concepts and index events. If the user does not define an index event, the cohort is considered a population cohort. In this example, the cohort is using a defined index event, **diabetes**, thus making it an index event cohort within the project.

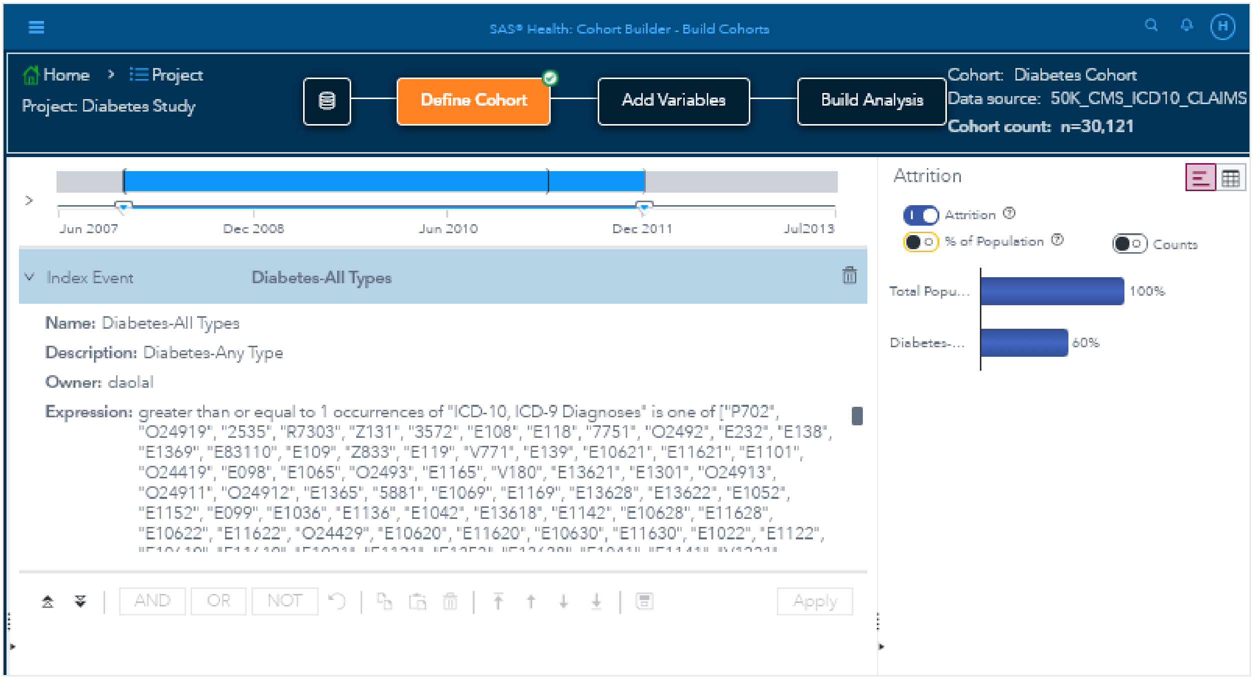
An index event contains properties of the defined expression:

- An index event expression is not based on a data source. Each is defined separately.
- An index event expression can include demographic fields.
- One data-based field must be included within the index event expression.
- An index event expression can include both simple and compound subexpressions.

The index event expression is used to determine the index event date that is associated with an index event and the index event period. The index event period is the index event start and end dates for a patient.

To define and create an index event, the user can search for descriptions and code values. The index event has been named **Diabetes**, which specifies that a patient has at least one occurrence of a Diabetes diagnosis (ICD-9 and ICD-10 codes) during the overall study dates.

Figure 3. *Defining the Cohort and Showing the Index Event Definition*



Project: Diabetes Study

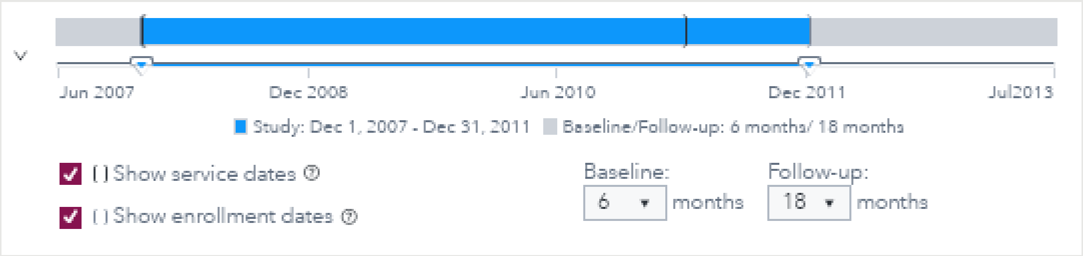
Cohort: Diabetes Cohort
Data source: 50K_CMS_ICD10_CLAIMS
Cohort count: n=30,121

Index Event Diabetes-All Types

Name: Diabetes-All Types
Description: Diabetes-Any Type
Owner: daolal
Expression: greater than or equal to 1 occurrences of "ICD-10, ICD-9 Diagnoses" is one of ["P702", "O24919", "2535", "R7303", "Z131", "3572", "E108", "E118", "7751", "O2492", "E232", "E138", "E1369", "E83110", "E109", "Z833", "E119", "V771", "E139", "E10621", "E11621", "E1101", "O24419", "E098", "E1065", "O2493", "E1165", "V180", "E13621", "E1301", "O24913", "O24911", "O24912", "E1365", "5881", "E1069", "E1169", "E13628", "E13622", "E1052", "E1152", "E099", "E1036", "E1136", "E1042", "E13618", "E1142", "E10628", "E11628", "E10622", "E11622", "O24429", "E10620", "E11620", "E10630", "E11630", "E1022", "E1122",

When applying the defined index event, Cohort Builder displays the cohort count, which is the number of patients within the given data source who identify as having diabetes within the study period. This study shows that out of the 50,000-count patient population, there are 29,416 patients who meet the criteria of being a diabetic patient between the study period of December 1, 2007 through December 31, 2011.

Figure 4. *Showing the Study-Period Time Line for the Specified Cohort*



■ Study: Dec 1, 2007 - Dec 31, 2011 ■ Baseline/Follow-up: 6 months/ 18 months

☑ ( ) Show service dates ⓘ
☑ ( ) Show enrollment dates ⓘ

Baseline: 6 ▾ months
Follow-up: 18 ▾ months

In the Define Cohort step (in Figure 1), the study period is depicted as a time line. The study period has defaulted to the earliest and latest for service and enrollment information that is found within the data source. The user can edit the time line to specify a certain study period, to allow service dates to be shown, and to allow enrollment dates to be shown on the time line. When creating an index event cohort, the user is also able to modify the baseline and follow-up periods that appear on the time line. The baseline and follow-up fields are required because the study start date for the patient is the cohort's index event date minus the number of months specified for the baseline (the default is six months) while the period end date for the patient is the index event date plus the number of months specified for the follow-up (the default is 18 months).

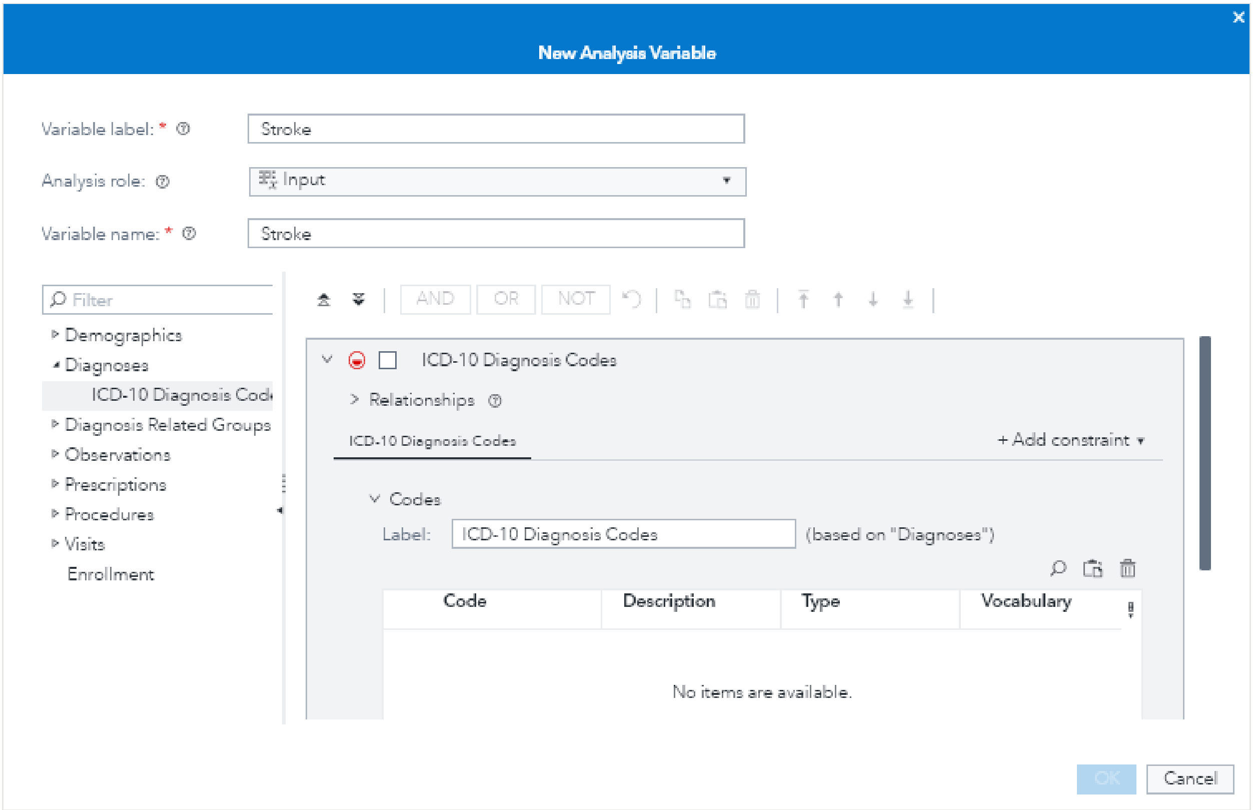
One important aspect of cohort studies is the ability to produce an adequate time line. Using Cohort Builder, the user can define a time line that is appropriate for the current study.

Our time shows that the cohort is being analyzed by looking into past medical events. A retrospective cohort study includes a patient population that is selected according to their past health records that indicate that they have been diagnosed with diabetes. Retrospective cohort studies allow for limited control over the data because it might be incomplete, inaccurate, or inconsistent. However, Cohort Builder is designed to assist clinicians in defining the inclusion and exclusion criteria, which enable the user to evaluate whether the cohort is biased. Also, this type of study costs less and is typically shorter than prospective cohort studies that investigate from present into the future (Song, J., and Chung, K).

After defining the cohort, your next step is to add variables. The user can include demographics and risk factors and can create new custom variables. These variables are applied to the cohort to create statistics and analytical results, which allow the user to identify potential patterns, trends, and relationships.

By creating new custom variables, Cohort Builder enables users to specify the variable label, the analysis role, and the variable name. The user can choose what type of variable to create from the list. For this study, the analysis variable, **Stroke**, is a diagnosis.

Figure 5. *Creating a Custom Analysis Variable to Search for Diagnostic Codes*



New Analysis Variable

Variable label: * Stroke

Analysis role: Input

Variable name: * Stroke

ICD-10 Diagnosis Codes

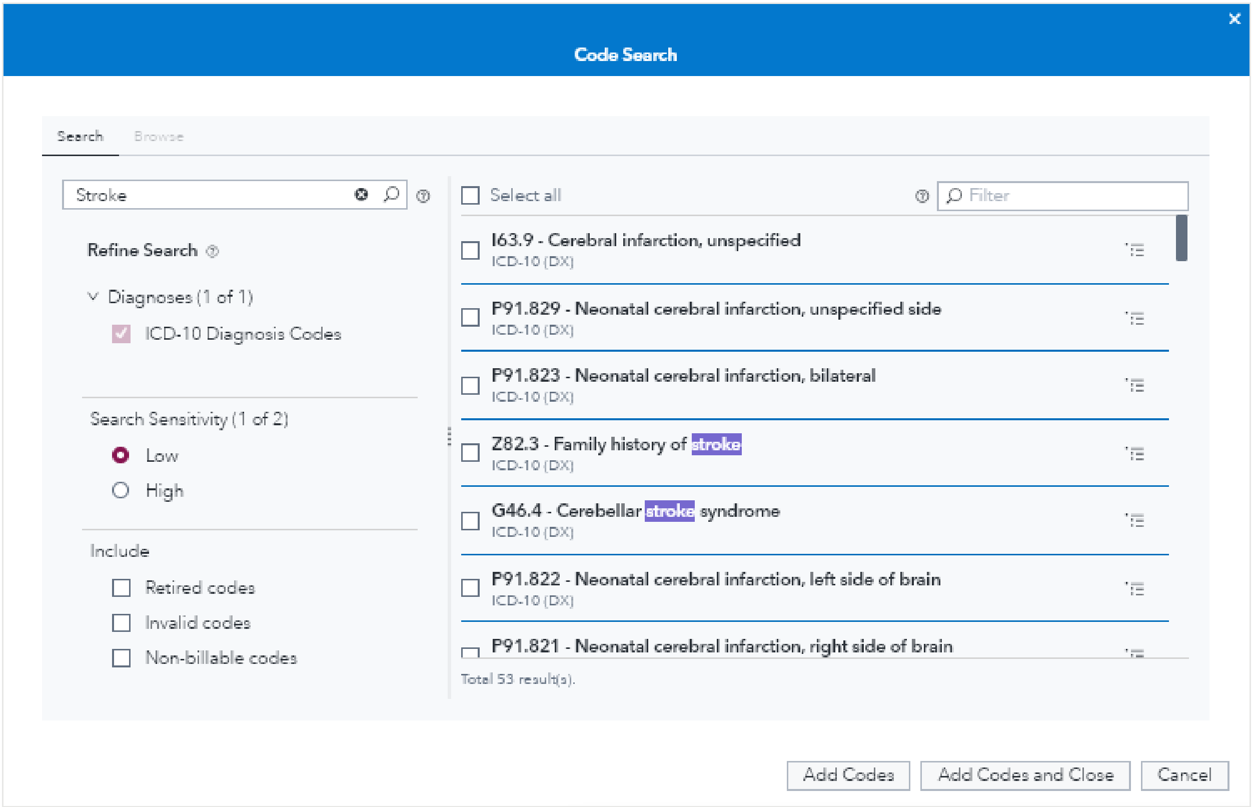
Label: ICD-10 Diagnosis Codes (based on "Diagnoses")

Code	Description	Type	Vocabulary

No items are available.

After the user enters the **Stroke** variable in the search field and clicks **Search**, Cohort Builder generates a list of ICD-10 diagnosis codes.

Figure 6. *Generating Code Search Results for the Custom Variable Stroke*



Stroke

Diagnoses (1 of 1)

☑ ICD-10 Diagnosis Codes

Search Sensitivity (1 of 2)

● Low

○ High

Include

□ Retired codes

□ Invalid codes

□ Non-billable codes

□ I63.9 - Cerebral infarction, unspecified
ICD-10 (DX)

□ P91.829 - Neonatal cerebral infarction, unspecified side
ICD-10 (DX)

□ P91.823 - Neonatal cerebral infarction, bilateral
ICD-10 (DX)

□ Z82.3 - Family history of stroke
ICD-10 (DX)

□ G46.4 - Cerebellar stroke syndrome
ICD-10 (DX)

□ P91.822 - Neonatal cerebral infarction, left side of brain
ICD-10 (DX)

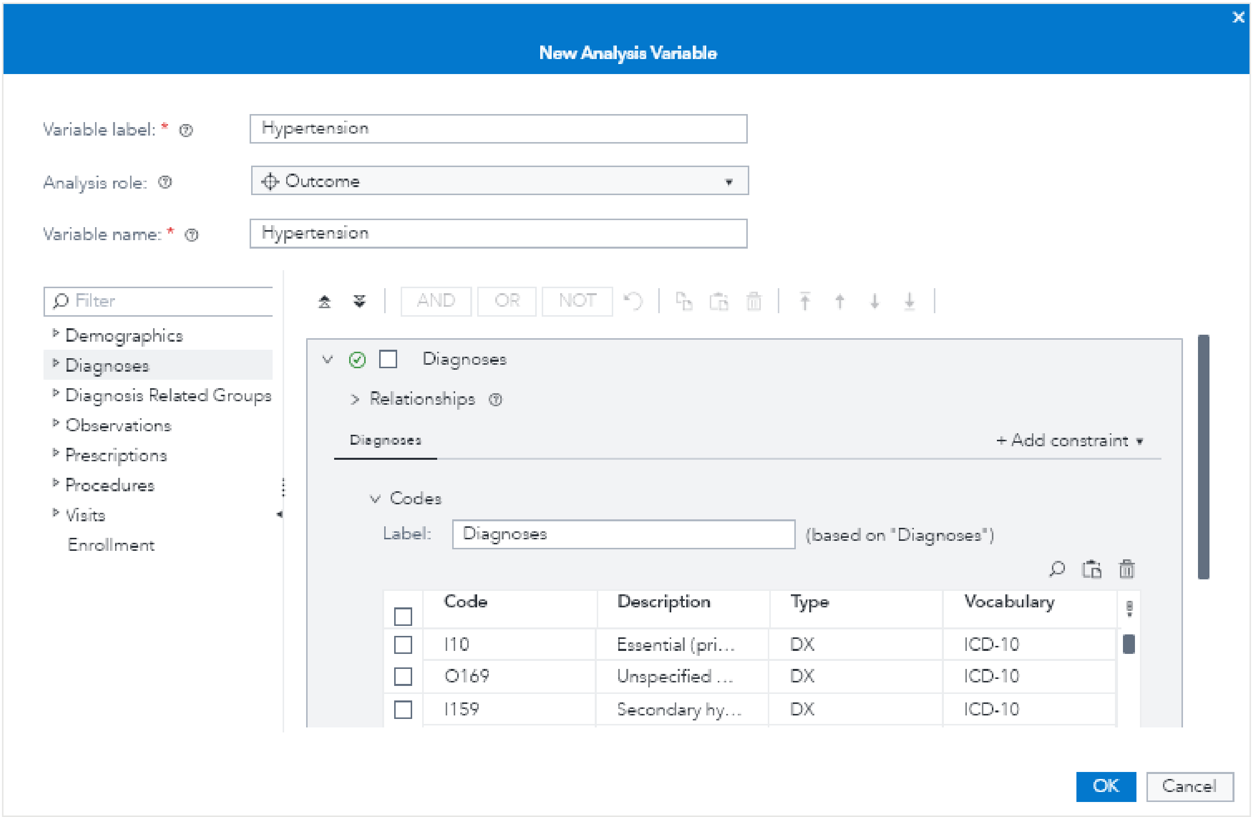
□ P91.821 - Neonatal cerebral infarction, right side of brain

Total 53 result(s).

After selecting all the ICD-10 diagnosis codes that are relevant to the analysis variable, the user can add the codes to generate the expression. The user can create as many analysis variables as necessary for analysis.

This example uses **Stroke**, **Hypertension**, and the demographic grouper **AGEGROUP_2** as variables for this study.

Figure 7. Displaying the Custom Variable Hypertension with the Diagnosis Codes



Variable label: * Hypertension

Analysis role: Outcome

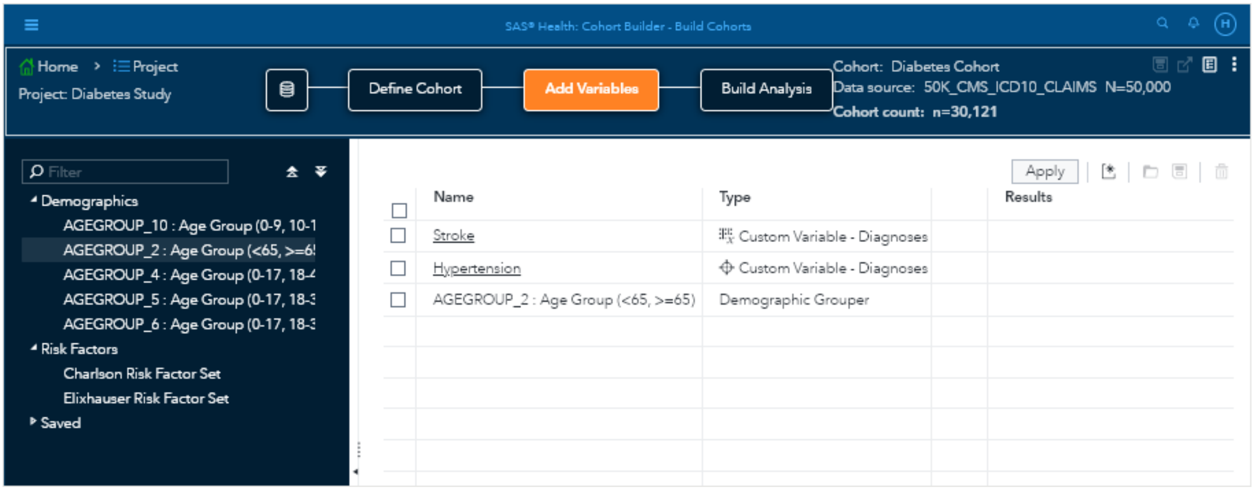
Variable name: * Hypertension

Label: Diagnoses (based on "Diagnoses")

Code	Description	Type	Vocabulary
I10	Essential (pri…	DX	ICD-10
O169	Unspecified …	DX	ICD-10
I159	Secondary hy…	DX	ICD-10

After the variables are added and applied to the defined cohort, the results appear in the **Add Variables** tab. This feature enables the user to view the result percentages and graphs.

Figure 8. *Displaying the Analysis Variables for this Study and the Calculated Results.*


SAS® Health: Cohort Builder - Build Cohorts
Home
Project
Define Cohort
Add Variables
Build Analysis
Filter
Demographics
AGEGROUP_10 : Age Group (0-9, 10-1
AGEGROUP_2 : Age Group (<65, >=6!
AGEGROUP_4 : Age Group (0-17, 18-4
AGEGROUP_5 : Age Group (0-17, 18-3
AGEGROUP_6 : Age Group (0-17, 18-3
Risk Factors
Charlson Risk Factor Set
Elixhauser Risk Factor Set
Saved
Apply

Project: Diabetes Study

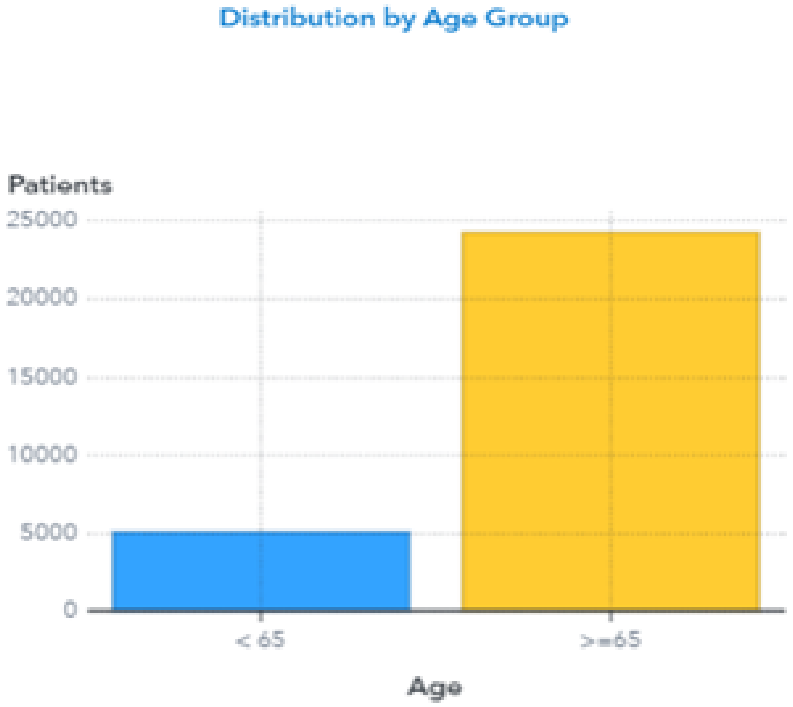
Cohort: Diabetes Cohort
Data source: 50K_CMS_ICD10_CLAIMS N=50,000
Cohort count: n=30,121

Name	Type	Results
Stroke	Custom Variable - Diagnoses	
Hypertension	Custom Variable - Diagnoses	
AGEGROUP_2 : Age Group (<65, >=65)	Demographic Grouper	

The number of patients who have experienced a stroke (indicated by the **Stroke** variable) within the cohort population is 1,675, which is 5.69%. The number of patients who have experienced hypertension (indicated by the

Hypertension variable) within the cohort population is 26,888, which is 91.41%. The demographic grouper **AGEGROUP_2** is defined as patients who are older than 65 or are 65 and younger. The following image depicts the age group distribution of the patient population within the cohort using the variable **AGEGROUP_2**. The following graph is generated by clicking the view graph icon.

Figure 9. *Displaying the Age Group Distribution of the Cohort Using the Variable AGEGROUP_2*



Cohort Builder includes a third tab, which contains add-ins that can help analyze the cohort further. The user can create add-in templates or use the Cohort Characterization add-in that is provided by SAS Health.

To create an add-in template, you can specify options in the information tab such as the template name, the description, the category, and whether the template is public or private. Within the code tab, the user can add SAS code for the add-in template. The user interface tab enables the user to set parameters in run times and add controls to the user interface work area. After adding code or user interface controls, the user can use the execution tab to test it. Running the code generates tabs to view the results provided in reports, output data sets, and logs. The output is downloadable.

This study uses the Cohort Characterization add-in template to generate a report that displays characterizations of patients within the cohort. The user must select the add-in template and then click **Edit**. By double clicking the add-in template, the user can create an add-in instance. Opening the Cohort Characterization add-in template in Edit mode enables the user to view and edit the information, the code, the user interface, and the execution tabs of the template.

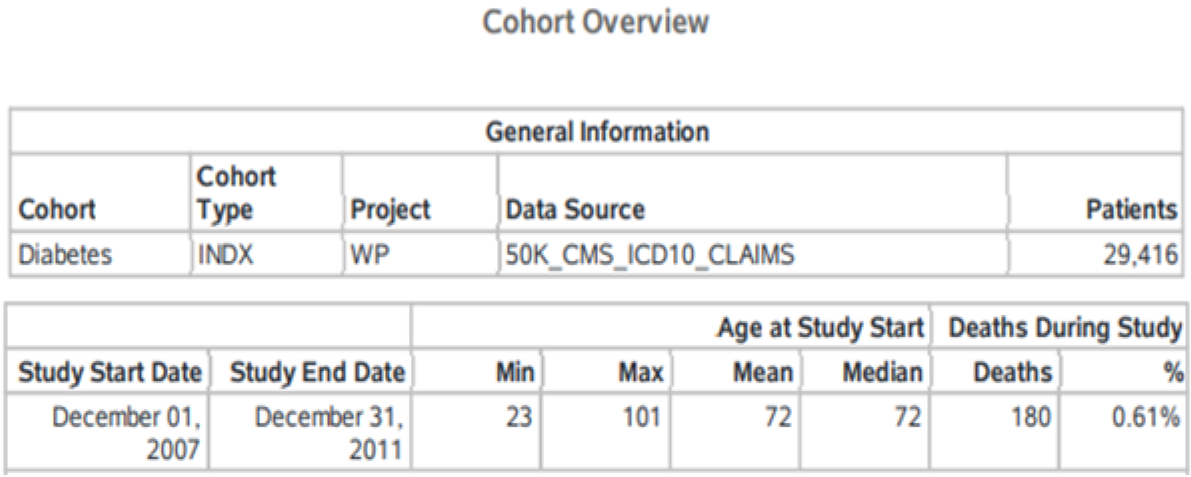
After writing and reviewing the code that is necessary for the analysis of the cohort, the user is then able to select whether they want to add a user interface and specify the settings of run-time parameters.

The execution tab is then available to run the user-specified add-in template. Running the Cohort Characterization add-in template produces a log, the results, and the output data tabs. The user can also select PDF or RTF report output format.

Results Analysis

The following images are the results of using the Cohort Characterization add-in template with the diabetic patient cohort that was created. The Cohort Overview displays general information about the cohort during its creation. It also generates statistics about patient ages and death rates.

Figure 10. *Cohort Overview*



Cohort Overview

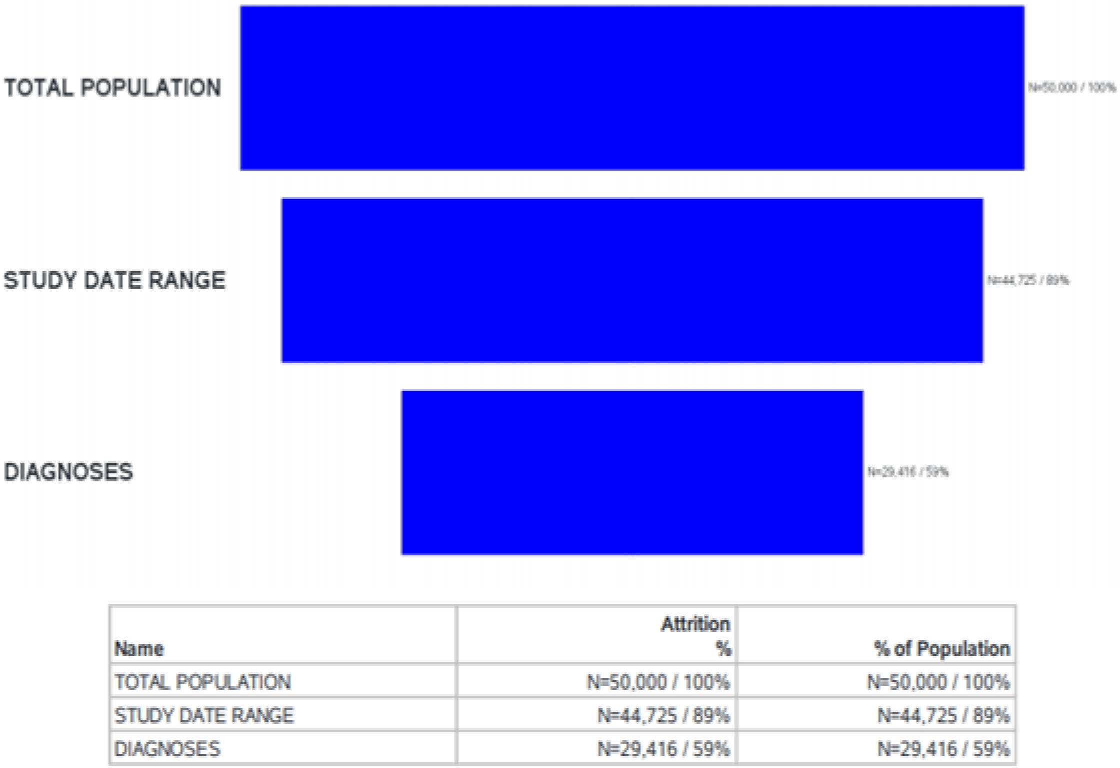
General Information				
Cohort	Cohort Type	Project	Data Source	Patients
Diabetes	INDX	WP	50K_CMS_ICD10_CLAIMS	29,416

		Age at Study Start				Deaths During Study	
Study Start Date	Study End Date	Min	Max	Mean	Median	Deaths	%
December 01, 2007	December 31, 2011	23	101	72	72	180	0.61%

The cohort overview enables the user to view the cohort type, the index, and the specified index event, diabetes. The user can also view the study period from December 01, 2007 through December 31, 2011 and see that over this four-year study, 180 deaths occurred during the study period. The youngest patient at the start of the study was 23 and the oldest was 101. The ages of the patients had a calculated mean and median of 72, which indicates that the overall age distribution of the patient population was more likely to be over the age of 65, as shown in Figure 10. After viewing the general information, the user can then proceed to analyze the cohort in more depth.

The following graph displays the number of patients in the data source, the patients who fit into the study period, and the patients who were diagnosed with the index event.

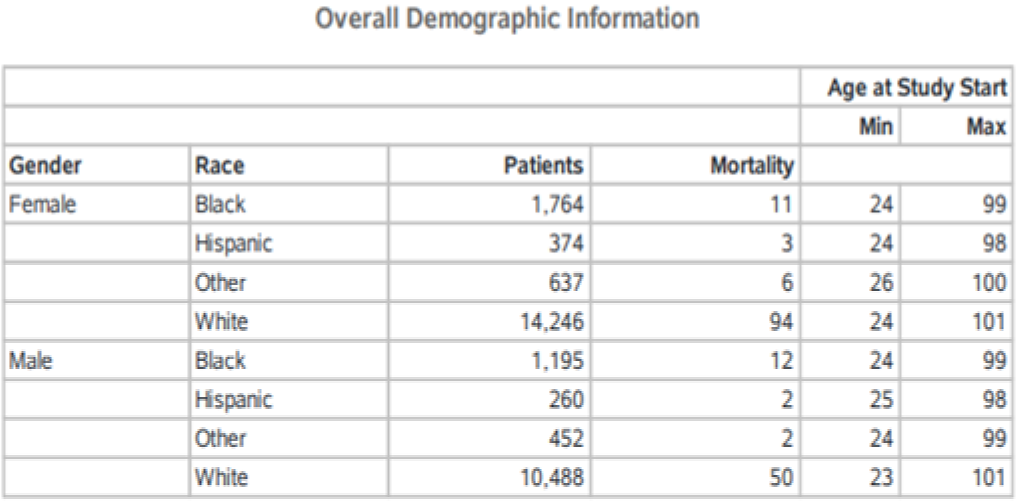
Figure 11. *Graphical Results from the Study*



Name	Attrition %	% of Population
TOTAL POPULATION	N=50,000 / 100%	N=50,000 / 100%
STUDY DATE RANGE	N=44,725 / 89%	N=44,725 / 89%
DIAGNOSES	N=29,416 / 59%	N=29,416 / 59%

The total population from the data source is 50,000 patients. Of these patients, 44,725 patients fall into the study date range. The number of patients who were diagnosed as diabetic during the study date range were 29,416. The user can also view the table to gather the percentages in relation to the total population to ensure that the data source is offering a representative patient population.

Figure 12. *Overall Demographic Information Table*

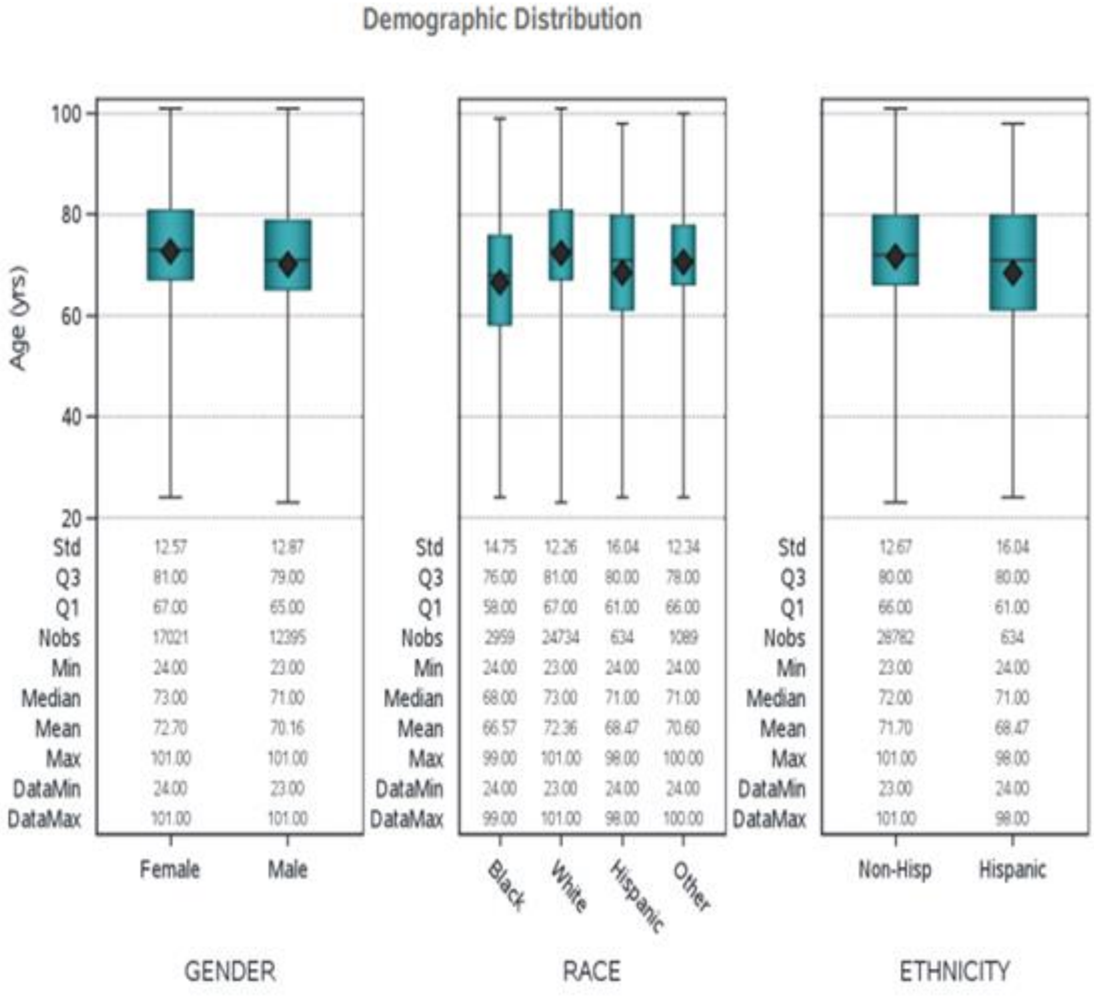


Overall Demographic Information

				Age at Study Start	
				Min	Max
Gender	Race	Patients	Mortality		
Female	Black	1,764	11	24	99
	Hispanic	374	3	24	98
	Other	637	6	26	100
	White	14,246	94	24	101
Male	Black	1,195	12	24	99
	Hispanic	260	2	25	98
	Other	452	2	24	99
	White	10,488	50	23	101

Personal characteristics are important for analyzing patient populations. Personal characteristics are age, race, gender, and so on, that are included in most data sets and are compared separately. First, the patients are grouped based on their gender, female or male. The patients are then categorized by their race, and the number of patients is calculated for each category, and finally, the mortality of patients is calculated. The youngest and oldest patients are also listed in relation to the patient categories. Analyzing this table enables the user to see that White female patients constitute the most patients as opposed to any other category of patient. During the study period, there are 14, 246 White female patients within the study with a calculated mortality, or number of deaths, of 94 patients.

Figure 13. *Demographic Distribution for Patient Gender, Race, and Ethnicity*



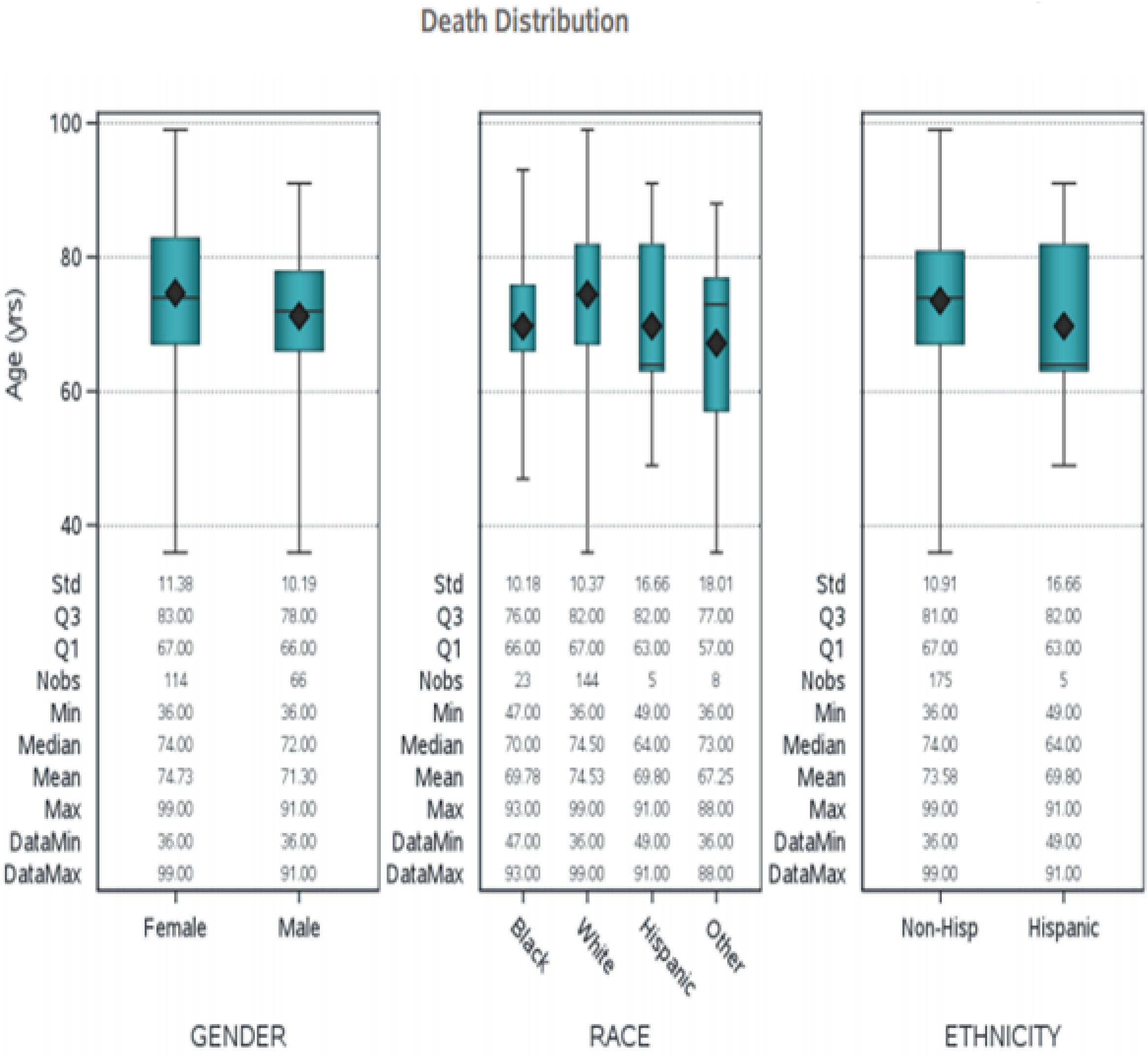
	Female	Male
Std	12.57	12.87
Q3	81.00	79.00
Q1	67.00	65.00
Nobs	17021	12395
Min	24.00	23.00
Median	73.00	71.00
Mean	72.70	70.16
Max	101.00	101.00
DataMin	24.00	23.00
DataMax	101.00	101.00

	Black	White	Hispanic	Other
Std	14.75	12.26	16.04	12.34
Q3	76.00	81.00	80.00	78.00
Q1	58.00	67.00	61.00	66.00
Nobs	2959	24734	634	1089
Min	24.00	23.00	24.00	24.00
Median	68.00	73.00	71.00	71.00
Mean	66.57	72.36	68.47	70.60
Max	99.00	101.00	98.00	100.00
DataMin	24.00	23.00	24.00	24.00
DataMax	99.00	101.00	98.00	100.00

	Non-Hisp	Hispanic
Std	12.67	16.04
Q3	80.00	80.00
Q1	66.00	61.00
Nobs	28782	634
Min	23.00	24.00
Median	72.00	71.00
Mean	71.70	68.47
Max	101.00	98.00
DataMin	23.00	24.00
DataMax	101.00	98.00

The demographic distribution is used as a comparative analysis to identify the characteristics and differences within the data. Using the demographic distribution to analyze data enables the user to identify the overall representation of the patient population. Epidemiologists are interested in ethnic and racial groups to identify the differences in susceptibility, biological, or cultural factors within these groups. This shows that the most represented group of patients are females within the gender category and White in the race category. This study does not contain a large representation for Hispanics.

Figure 14. *Death Distribution for Patient Gender, Race, and Ethnicity*



	Female	Male
Std	11.38	10.19
Q3	83.00	78.00
Q1	67.00	66.00
Nobs	114	66
Min	36.00	36.00
Median	74.00	72.00
Mean	74.73	71.30
Max	99.00	91.00
DataMin	36.00	36.00
DataMax	99.00	91.00

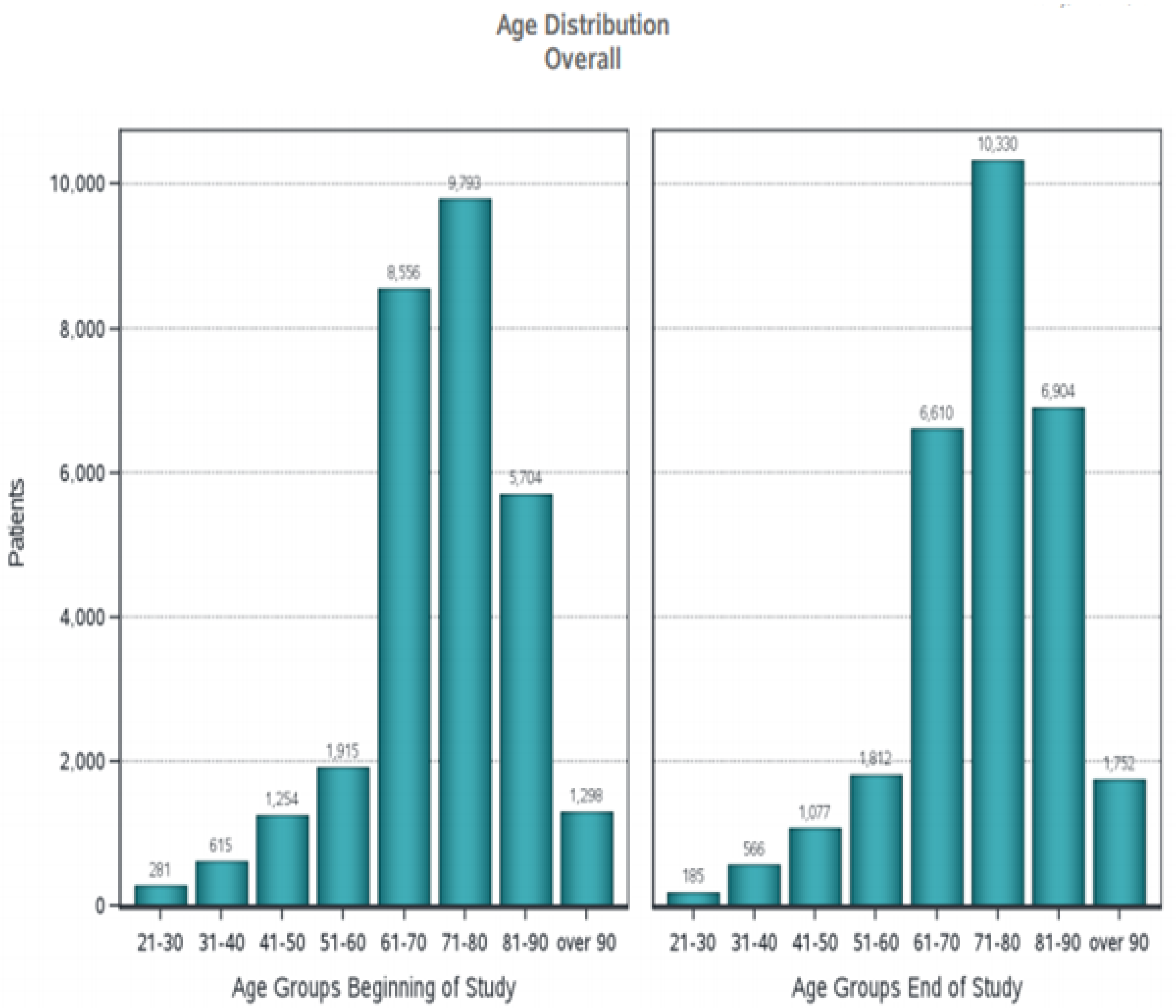
	Black	White	Hispanic	Other
Std	10.18	10.37	16.66	18.01
Q3	76.00	82.00	82.00	77.00
Q1	66.00	67.00	63.00	57.00
Nobs	23	144	5	8
Min	47.00	36.00	49.00	36.00
Median	70.00	74.50	64.00	73.00
Mean	69.78	74.53	69.80	67.25
Max	93.00	99.00	91.00	88.00
DataMin	47.00	36.00	49.00	36.00
DataMax	93.00	99.00	91.00	88.00

	Non-Hisp	Hispanic
Std	10.91	16.66
Q3	81.00	82.00
Q1	67.00	63.00
Nobs	175	5
Min	36.00	49.00
Median	74.00	64.00
Mean	73.58	69.80
Max	99.00	91.00
DataMin	36.00	49.00
DataMax	99.00	91.00

The number of deaths for this patient population is most represented by non-Hispanic, White patients. This type of distribution is an in-depth analysis of the 0.61% death rate that was previously noted in Figure 11. The least amount of death occurred within the Hispanic patient population. To further understand the discrepancies in the death rates, or any distribution of rate, the user can notice that the data source selected for this cohort contains a large representation of White patients. This can be viewed in Figure 13.

Epidemiologists consider age to be an important factor because with age comes different factors to consider such as susceptibility, latency, and physiological response to disease development.

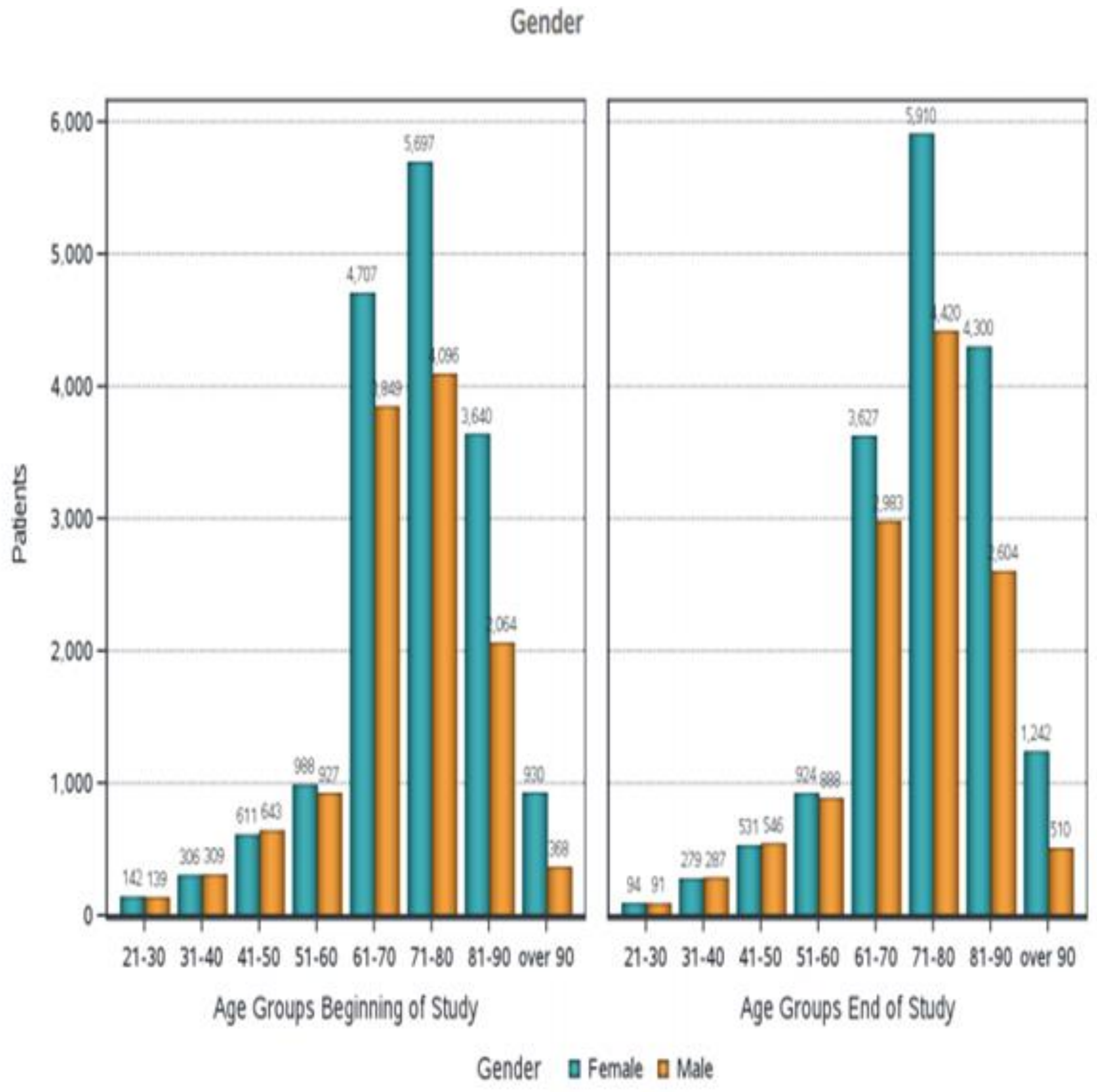
Figure 15. *Overall Age Distribution*



When analyzing age, epidemiologists typically use narrow age groups to identify age-related patterns. The overall age distribution that represents the patient population contains patients in every age group. The younger age groups are not represented as much as ages 61-90. This could be because the older population has more complications related to diabetes, and they visit health care facilities more often. Thus, it is more common, and noticeable, for older patients to be diagnosed.

Viewing the age distribution in terms of gender shows possible relationships and trends.

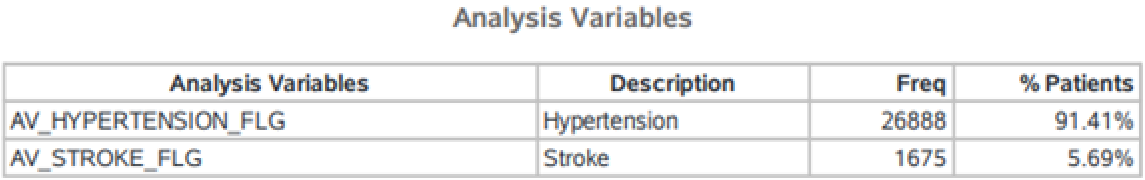
Figure 16. *Age Distribution by Gender*



The gender of a patient is important to consider when analyzing diseases due to genetic, hormonal, and anatomic differences. In most age groups, females have a higher patient count at the beginning and end of the study. Between the ages of 61 to over 90, females are notably more represented through this display of age distribution. This finding is not inconsistent with the previous displays because females have a larger population in this study.

The next table shows the analysis variables that were created during the cohort creation process in relation to the patients who have been diagnosed with hypertension or stroke.

Figure 17. *Analysis Variables and the Percentage of Patients with Hypertension or Stroke*



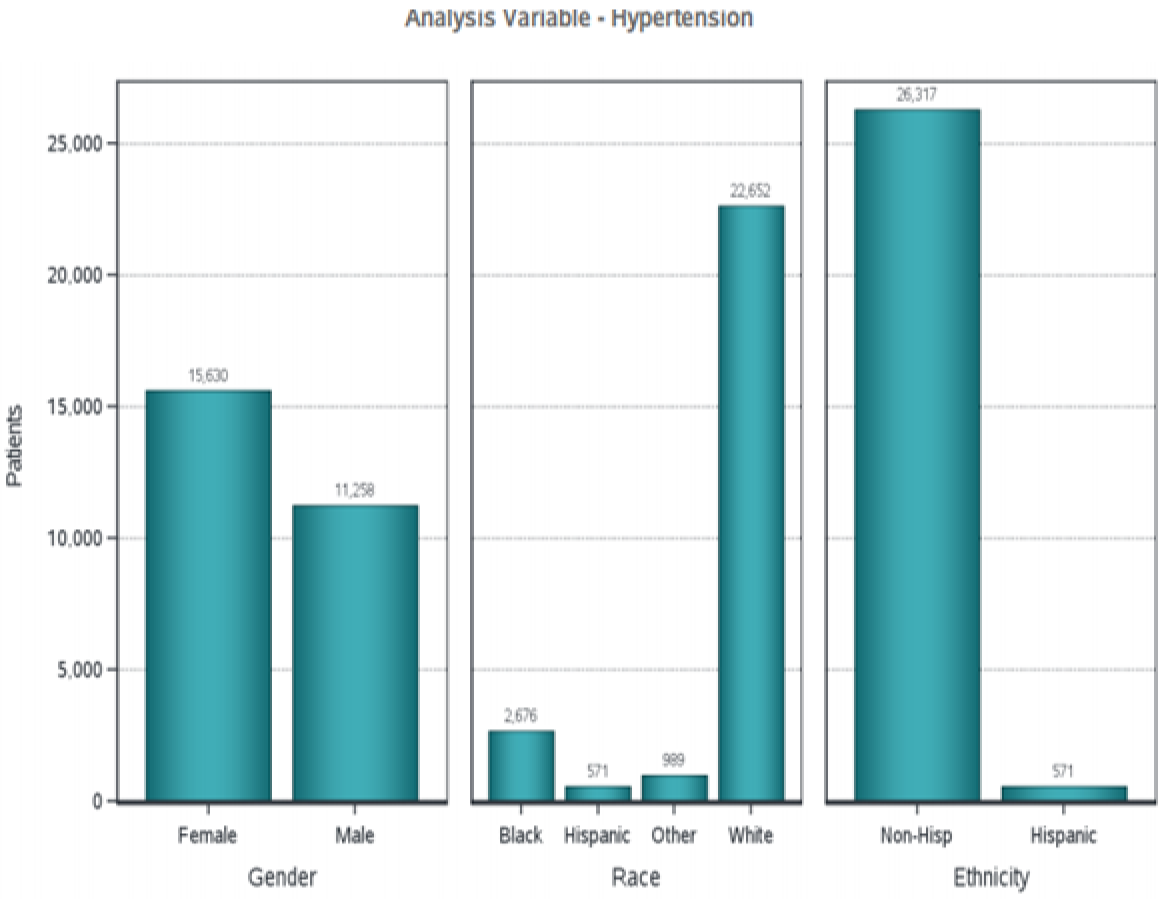
Analysis Variables

Analysis Variables	Description	Freq	% Patients
AV_HYPERTENSION_FLG	Hypertension	26888	91.41%
AV_STROKE_FLG	Stroke	1675	5.69%

The number of patients in this data source who have been diagnosed with diabetes and have experienced hypertension is over 90% of the patient population. Patients who have hypertension are experiencing high blood pressure and, when diagnosed with diabetes, a patient is more at risk for other diseases such as cardiovascular and kidney disease. Having both diabetes and hypertension leads to a greater probability of experiencing strokes.

Patients who experience high blood pressure while also being diagnosed as diabetic are more at risk for severe complications.

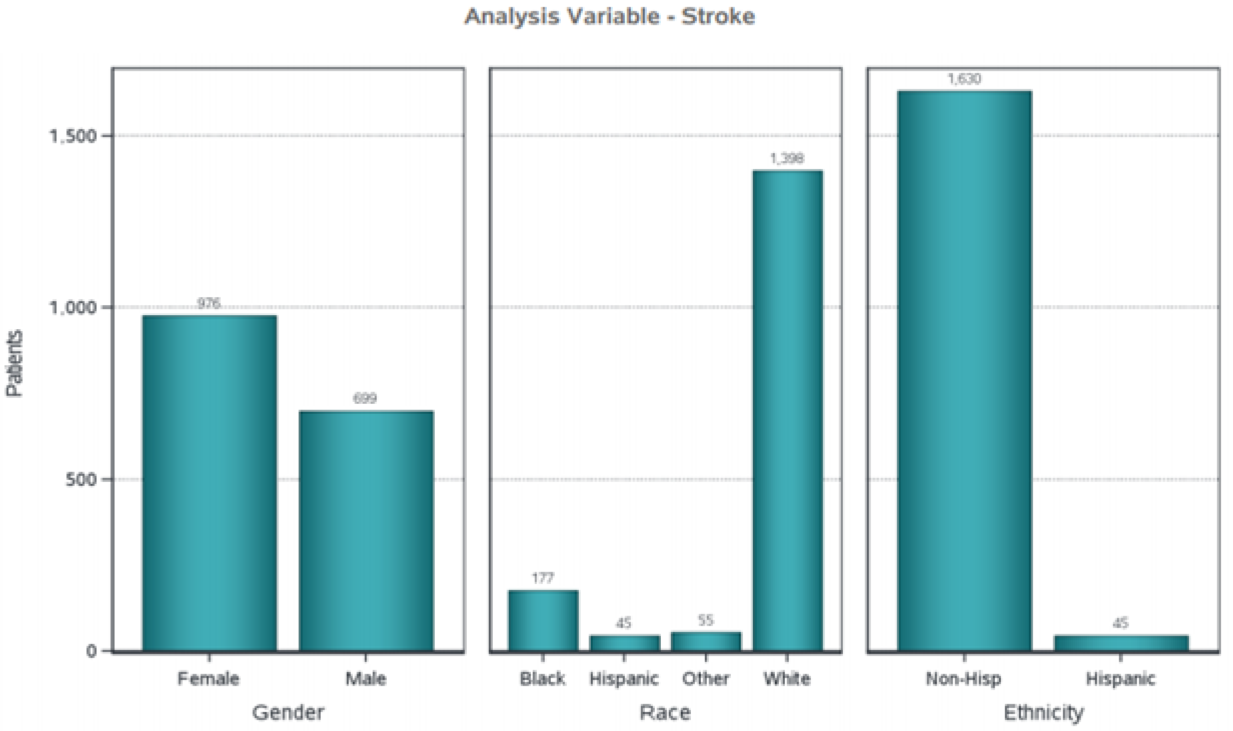
Figure 18. *Analysis Variable Hypertension in Relation to Patient Gender, Race, and Ethnicity*



Hypertension leads to heart attacks and strokes in many of these patients. Having high blood pressure puts a strain on the patient's heart and has a damaging effect on arteries. Knowing the existence of this relationship enables patients and medical providers to consider options that might help the patients who are diagnosed with diabetes and hypertension. The reality that this relationship exists can also provide medical experts with an initiative to identify ways to decrease the diabetic patient's likelihood of developing hypertension.

Strokes are a critical condition caused by blockages in the blood vessels in the brain or neck. Diabetic patients have a higher risk of having strokes if hypertension is also present.

Figure 19*. Analysis Variable Stroke in Relation to Patient Gender, Race, and Ethnicity*



The representation of diabetic patients who have had a stroke is 5.69%, as seen in Figure 20. Although this number is seemingly low, the study period of the cohort was over four years. Thus, the likelihood of patients being diagnosed with strokes would be dependent on the diabetic diagnosis and the hypertension condition that worsens over time.

Conclusion

The report generated for the cohort study shows an in-depth analysis of the characterizations of the cohort. This information is useful for understanding the relationships between diagnosed diabetic patients and stroke, hypertension, and age demographics. SAS Health: Cohort Builder provides all the tools, resources, and reports depicted throughout the study. With the help of Cohort Builder, the user can easily and quickly generate analytical studies of patient cohorts within minutes. Epidemiologists can use these tools to investigate and monitor patterns and relationships of disease in patient populations.

References

Song, J., and Chung, K. (2010, December). Observational Studies: Cohort and Case-Control Studies. Available at www.ncbi.nlm.nih.gov/pmc/articles/PMC2998589/#R2.

What is Diabetes? CDC (202, June 11). Available at www.cdc.gov/diabetes/basics/diabetes.html.

Release Information Content Version: 1.0 September, 2021.

Trademarks and Patents SAS Institute Inc. SAS Campus Drive, Cary, North Carolina 27513

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. R indicates USA registration. Other brand and product names are registered trademarks or trademarks of their respective companies.

To contact your local SAS office, please visit: sas.com/offices