**CAAD**

**Carotid Artery Atherosclerosis Disease**

**Phenotype Algorithm Pseudo Code**

**Group Health/University of Washington**

**Version: December 29, 2014 (Final)**(Clarification edit made 5/22/2015)

**Contact Us**:

David S. Carrell ([carrell.d@ghc.org](mailto:carrell.d@ghc.org), 206-287-2705)

David Cronkite ([cronkite.d@ghc.org](mailto:cronkite.d@ghc.org), 206-287-2153)

Aaron Scrol ([scrol.a@ghc.org](mailto:scrol.a@ghc.org), 206-287-2150)

**Updates**:

* Final version posted on PheKB 12/29/2014
* Originally posted on PheKB 11/18/2014

Table of Contents

[1. Introduction: 2](#_Toc407711007)

[2. Development 2](#_Toc407711008)

[3. Data Requirements 2](#_Toc407711009)

[4. General Implementation Steps 3](#_Toc407711010)

[5. Using the NLP System 3](#_Toc407711011)

[6. General Inclusion Criteria 4](#_Toc407711012)

[7. CAAD Case Criteria 5](#_Toc407711013)

[8. CAAD Control Criteria 5](#_Toc407711014)

[9. GH/UW Counts of Cases, Controls, and Excluded Patients 6](#_Toc407711015)

[10. Data Dictionary and Covariates 7](#_Toc407711016)

[11. Flow diagram 8](#_Toc407711017)

# 1. Introduction:

The overall prevalence of asymptomatic carotid artery atherosclerosis disease (CAAD) is estimated at 50% in the general US population. Symptomatic disease is estimated at 2%–9%.[1] A higher prevalence of 5%–9% with symptomatic disease is anticipated in patients over 65 years of age, which is of public health relevance because it means that 1.3–2.4 million Americans in this age group are being affected by CAAD.[2,3,4]

Implementation of an electronic medical record (EMR) based phenotype for CAAD within the Electronic Medical Records and Genomics (eMERGE) Network[5] will employ a variety of clinical data types including patient encounter records, diagnoses codes, surgical procedure codes, natural language processing (NLP), laboratory measurements, and medication data.

Phenotype logic is described below beginning in section 6 (General Inclusion Criteria). A flow diagram representing the logic and definitions used to identify CAAD cases and CAAD controls in provided in section 10 as Figure 10.1.

# 2. Development

This phenotype was developed at Group Health/University of Washington (GH/UW). Structured data from the Group Health data warehouse and clinical imaging reports from the Group Health EMR were used to define cases and controls. The algorithm defines two categories of cases (severe CAAD and moderate CAAD) and two categories of controls (measured and unmeasured).

GH/UW developed a portable NLP system used to extract carotid stenosis exam results from imaging reports. This system was initially shared with developers at Marshfield, Northwestern and Mayo Clinic. Each receiving site processed a sample of reports and conducted error analyses, assisted by the portable NLP system’s built-in diagnostic reporting feature. GH/UW iteratively adapted the portable NLP system to address performance issues at each of these sites before releasing final versions of the system. The initial chart review validation was conducted at GH/UW. A secondary validation was conducted at [SITE NAME].

# 3. Data Requirements

To identify patients qualifying as cases or controls for this algorithm and to provide covariates needed for analyses the following types of data are needed:

* Patient demographics
* Encounter history
* Free text imaging reports (for natural language processing)
* Diagnosis codes
* Procedure codes
* Medication records

# 4. General Implementation Steps

For efficiency we recommend implementing the CAAD algorithm in the following general steps.

1. Identify patients meeting criteria described in section 6, General Inclusion Criteria (below).
2. Assemble clinical text reports of imaging studies that may include assessments of carotid artery stenosis (e.g., carotid ultrasound studies, CT studies of the neck, CS studies of the neck).
3. Process the assembled clinical text reports using the portable NLP system posted on [PheKB](https://phekb.org) with other CAAD phenotype files, then store the structured results produced by the system (in either CSV format or a database table).
4. Using your preferred programming language (e.g., SAS) process the NLP-extracted data in the CSV file to identify patients meeting the measured carotid stenosis levels described for cases and controls (sections 7 and 8 below). SAS code for processing the NLP system’s CSV file to produce these patient-level summaries is posted on [PheKB](https://phekb.org) (CAAD\_Process\_NLP\_Data.sas).
5. Assemble the structured data needed to determine case and control status as described in sections 7 and 8 (below).
6. Combine the NLP-extracted structured data (Step D above) with other patient data (Step E above) to identify CAAD cases and CAAD controls (sections 7 and 8 below).
7. Using the data dictionary posted on [PheKB](https://phekb.org), prepare the required datasets of observations and covariates for cases and controls.
8. As a final “verification” step, randomly select and review the data for two or three of each type of cases and two or three of each type of controls to assure case and control assignments were made using the correct data and logic.

# 5. Using the NLP System

The portable NLP system for extracting results of carotid artery imaging studies can be downloaded from [PheKB](https://phekb.org). A set of fictitious imaging reports useful for testing the NLP system after you install it locally are included in the NLP package. An NLP system User Guide is also available on [PheKB](https://phekb.org).

The NLP system is an easily installed Java application with a simple graphical user interface. The User Guide provides step-by-step user instructions. If you encounter any problems installing or using the NLP system please email or call the contacts listed on page 1.

The NLP system supports two modes for inputting text documents: a) reading the documents as individual files stored in a folder of the local or network file system, or b) reading the documents as text fields from a relational database table. Similarly, output from the NLP system may be stored in either a comma separated value (CSV) format file in the file system or a database table.

As described further in the User Guide, the NLP system generates output at several levels of granularity, each displayed on a separate tab in the user interface. These outputs are intended to simplify local review of results and to facilitate error analysis.

We designed this NLP system to be used as a tool in a remote collaboration between the implementing site (your site) and the developing site (GH/UW). GH/UW personnel are available to help you identify any errors the NLP system may make when processing your reports, incorporate revisions into the NLP system designed to address these errors, and make available via [PheKB](https://phekb.org) an updated version of the NLP system.

The aim of the NLP system is to create accurate structured data at the level of independent mentions within an imaging study report. For example, from the statement “50-69% stenosis of the left internal carotid artery” the system is designed to extract the minimum reported stenosis value (i.e., 50%) the maximum reported stenosis value (i.e., 69%), and associated document date, document identifier, and patient identifier. These structured data can then be processed by the local programmer along with other structured patient data to implement the phenotype algorithm. The NLP system does not carry out the final step of summarizing the extracted results to the patient level. As noted above, a SAS program (CAAD\_Process\_NLP\_Data.sas) is available on [PheKB](https://phekb.org) to illustrate the process of summarizing the NLP data to the patient level.

# 6. General Inclusion Criteria

All patients must satisfy the following criterion before being considered for inclusion in the study as a case or control.

6.1. In the period preceding the date of the last known contact with the patient (e.g., an outpatient or inpatient encounter) the patient must have evidence of contact that satisfies at least one of the following three criteria:

6.1.1. Contact on at least two (2) occasions ≥365 days apart in the preceding 5 years (e.g., 6/23/2011 and 9/30/2014), OR

6.1.2. Contact in at least three (3) different calendar quarters in the preceding 5 years (e.g., 1/1/2014 and 3/15/2014 and 6/30/2014), OR

6.1.3. Contact in at least four (4) different calendar months in the preceding 5 years (e.g., 9/1/2011 and 10/1/2011 and 11/1/2011 and 12/1/2011).[[1]](#footnote-1)

We require this minimum level of contact to assure that patients with CAAD have opportunities for the disease to be documented in their EMR chart. Because risk of CAAD increases with age we require the contact to be within the last five years of the patients most recent known contact with the health care system.

# 7. CAAD Case Criteria

In addition to satisfying the general inclusion criterion (6.1 above), patients may be considered as potential CAAD cases if they:

7.1. Do not have laboratory evidence (ever) of a maximum total cholesterol level >400.

*Note: In the Group Health laboratory system the relevant cholesterol result is labeled as “TOTAL CHOLESTEROL.” Other reported lab results (e.g., for HDL, LDL, and triglycerides) may be ignored at Group Health.*

*CAAD Case Definition #1*

Patients meeting the case eligibility criteria (7.1) will satisfy our CAAD case definition #1 if in addition they:

7.2. Have evidence from a carotid imaging study of >50% carotid artery stenosis (at least unilaterally),

*OR*

7.3. Have at least one procedure code in the medical record indicating the received an endarterectomy (evidenced by an ICD-9 or CPT procedure code). Qualifying procedure codes are:

00.63 (ICD: Percutaneous insertion of carotid artery stent(s))

38.12 (ICD: Endarterectomy of other vessels of head and neck including carotid)

35301 (CPT: Thromboendarterectomy, including patch graft, if performed, carotid …)

35390 (CPT: Reoperation, carotid, thromboendarterectomy, > 1 month after orig. …)

*CAAD Case Definition #2*

Patients meeting the case eligibility criteria (7.1) will satisfy our CAAD case definition #2 if in addition they:

7.4. Have evidence from a carotid imaging study of >-=16% and ≤49% carotid artery stenosis (at least unilaterally),

*OR*

7.5. Have CAAD diagnoses (ICD-9 code 433.1, 433.10, or 433.11) on at least two occasions ≥30 days apart at any time in the past.

# 8. CAAD Control Criteria

In addition to satisfying the general inclusion criterion (6.1 above), patients may be CAAD controls if they:

8.1. Have evidence from a carotid imaging study of ≤15% carotid artery stenosis *bilaterally* (i.e., there is no evidence of stenosis >15% in either carotid artery),

*OR*

8.2.1 Do not have evidence (ever) of having received an imaging study of the carotid arteries,

*AND*

8.2.2 Do not have any CAAD diagnoses (ICD-9 code 433.1, 433.10, or 433.11),

*AND*

8.2.3 Have never received a carotid repair procedure. Qualifying procedure codes are:

00.63 (ICD: Percutaneous insertion of carotid artery stent(s))

38.12 (ICD: Endarterectomy of other vessels of head and neck including carotid)

35301 (CPT: Thromboendarterectomy, including patch graft, if performed, carotid, …)

35390 (CPT: Reoperation, carotid, thromboendarterectomy, > 1 month after orig. …)

00.61 (ICD: Percutaneous angioplasty or atherectomy of precerebral … vessel(s))

38.02 (ICD: Incision of other vessels of head and neck)

38.62 (ICD: Other excision of other vessels of head and neck)

0075T (CPT: Transcatheter placement of extracranial, vertebral, or … )

0076T (CPT: Transcatheter placement of extracranial, vertebral, or … )

35501 (CPT: Bypass graft, with vein, common carotid-ipsilateral internal carotid …)

33891 (CPT: Bypass graft, with other than vein, transcervical retropharyngeal …)

*Note: The above list includes procedures that do not qualify as carotid repair procedures for purposes of case definition. This is intentional. The above 11 codes optimize sensitivity for detecting possible carotid repair procedures that exclude a patient from being a control; the four codes used to define cases optimize specificity in defining procedures that qualify a patient as a case.*

# 9. GH/UW Counts of Cases, Controls, and Excluded Patients

Table 1 shows frequency counts of Group Health/University of Washington patients qualifying as CAAD cases, CAAD controls, or excluded per this CAAD phenotype pseudo. This cohort tends to be older than many other eMERGE cohorts, but the distributions by age group may be informative. Also keep in mind that most GH/UW patients receive care through a managed care model, which may also influence rates of CAAD detection compared to other settings.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1. Frequency counts of Group Health/University of Washington patients qualifying as cases, controls, or excluded per the CAAD phenotype pseudo code by age group.** | | | | | | | | |
| **Age Group** | **Case Def. #1** | | **Case Def. #2** | | **Control** | | **Exclude** | **Total** |
| **Case by 7.2** | **Case by 7.3** | **Case by 7.4** | **Case by 7.5** | **Control by 8.1** | **Control by 8.2** |
| **50-59** | 0 | 0 | 3 | 1 | 8 | 523 | 64 | 599 |
| **60-69** | 5 | 4 | 9 | 6 | 33 | 1265 | 183 | 1505 |
| **70-79** | 6 | 6 | 21 | 9 | 24 | 672 | 130 | 868 |
| **80-89** | 29 | 30 | 43 | 40 | 70 | 1167 | 418 | 1797 |
| **90+** | 24 | 7 | 37 | 20 | 56 | 657 | 247 | 1048 |
| **All** | 64 | 47 | 113 | 76 | 191 | 4284 | 1042 | 5817 |

# 10. Data Dictionary and Covariates

The CAAD data dictionary with detailed definition of all data elements required for implementing the phenotype is available on [PheKB](https://phekb.org).

Covariates include measures of relevant comorbidities (e.g., cardiovascular disease, diabetes) and medications (e.g., lipid-lowering drugs), smoking status, body mass index, and demographics. Some covariates are measured as patient-level constants or flag variebles (e.g., race) and others are captured as repeated measures (e.g., lipid lowering medications). Major categories of covariates described in the data dictionary are:

Patient-level data represented as constants or computed flag variables:

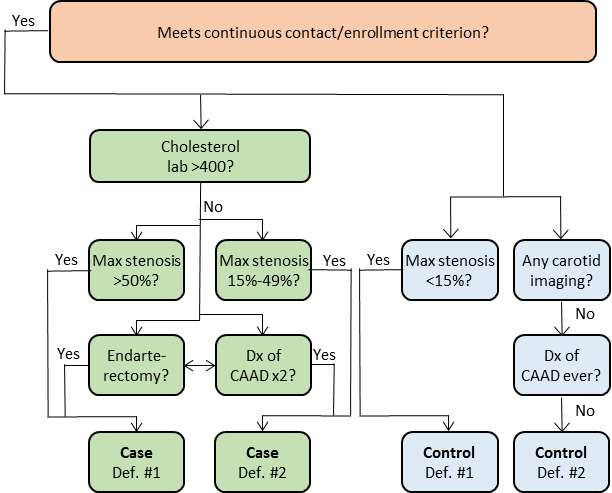
* Patient demographics
* Vital status/follow up status
* Smoking status
* Familial hypercholesterolemia diagnoses
* Type 1 diabetes mellitus (T1DM) diagnoses
* Type 2 diabetes mellitus (T2DM) diagnoses
* Coronary artery disease (CAD) diagnoses
* Peripheral Arterial Disease (PAD) diagnoses
* Endarterectomy (ENDART) procedures
* Cardiac revascularization (CRV) procedures

Repeated measure data:

* BMI
* Diabetes medications
* Lipid lowering medications
* Lipid laboratory results

# 11. Flow diagram

Figure 1 below represents the logic and definitions used to define two categories of CAAD cases and two categories of CAAD controls.



**Figure 10.1. Flow diagram of CAAD phenotype logic for definition of cases and controls.**

1. SAS code for operationalizing the enrollment/contact criteria using patient encounter data is available upon request from David Carrell, carrell.d@ghc.org. [↑](#footnote-ref-1)