Electronic Health Record-based Phenotyping Algorithm for Familial Hypercholesterolemia

PseudoCode

Authors and contacts:

Principal Investigator:
Iftikhar Kullo, MD Kullo.Iftikhar@mayo.edu

Adelaide Arruda-Olson, MD, PhD ArrudaOlson.Adelaide@mayo.edu
Carin Smith Smith.Carin@mayo.edu
Hongfang Liu, PhD Liu.Hongfang@mayo.edu
Majid Rastegar Mojarad.Majid@mayo.edu
Maya Safarova, MD, PhD Safarova.Mayya@mayo.edu
Parvathi Balachandran, MBBS Balachandran.Parvathi@mayo.edu
Saeed Mehrabi Mehrabi.Saeed@mayo.edu
Sunghwan Sohn, PhD Sohn.Sunghwan@mayo.edu
Xiao Fan, PhD Fan.Xiao@mayo.edu
Yijing Cheng Cheng.Yijing@mayo.edu

Address for correspondence:
Cardiovascular Biomarkers Research Laboratory
Department of Cardiovascular Diseases
Mayo Clinic, Stabile Building, Office 4-50
200 First Street SW, Rochester, MN 55905

Modification of a version 1.0 from May 11, 2015

Familial Hypercholesterolemia

An ounce of prevention is worth a pound of cure. Henry de Bracton (c. 1210-1268, British jurist, priest).

INTRODUCTION

Familial hypercholesterolemia (FH) is the most common Mendelian genetic disorder which is associated with dramatically increased lifetime risk for premature atherosclerotic cardiovascular disease (ASCVD) due to elevated plasma low-density lipoprotein cholesterol (LDL-C) levels.[1-3] The risk of early-onset coronary heart disease (CHD) is higher among those with FH-causing mutation compared to those without, regardless of their LDL-C levels.[4] The risk of developing early-onset CHD is 22 times higher in people with LDL-C >=190 mg/dL and a FH mutation.[2] A 100-fold higher risk of mortality from CHD has been observed in patients aged 20-39 years with untreated FH when compared to the general population.[5] The life expectancy is significantly shortened in untreated FH patients, mainly due to myocardial infarction.[6, 7] FH can be diagnosed based on clinical presentation and/or genetic testing results,[8] with genetic testing used as the "gold standard". Three screening approaches have been proposed for identifying index cases of FH: targeted, opportunistic and universal, followed by cascade (family) screening.[13] Targeted identification of FH using electronic health records (EHRs) is one of the efficient approaches for screening of patients with family members who have elevated LDL-C levels among other FH-related clinical criteria.[8, 9]

Various clinical criteria have been proposed to identify suitable subjects for genetic testing. The Dutch Lipid Clinic Network (DLCN) criteria are deemed to be the most specific and sensitive.[10, 11] To estimate the probability of a FH-mutation carrier status, DLCN criteria weigh clinical stigmata of FH encompassing results of lipid panel testing, personal and family history of hypercholesterolemia or premature ASCVD, presence of xanthomas on extensor tendons or thickening of the Achilles tendon, and early corneal arcus. Timely initiation of lipid-lowering treatment almost equilibrates the risk of ASCVD in patients with clinical diagnosis of FH to that of general population.[12] However, most countries diagnose less than 20% of all estimated cases and, in essence, less than 1% of patients with FH are aware of their condition. A significant proportion of them are identified only after their first ischemic event has occurred.[13] There is a need to develop systematic approaches to identify patients with FH and to conduct cascade screening of their relatives.[8] This EHR-based algorithm is intended to optimize screening and identification of patients with FH among individuals with severe hypercholesterolemia and therefore increase awareness, detection and control of FH.

We provide a pseudocode to identify cases and controls for primary hypercholesterolemia and FH. We have utilized a standardized diagnostic algorithm based on the validated set of clinical criteria, i.e. modified DLCN criteria,[14] including 4 diagnostic groups, i.e. family history and personal history of hypercholesterolemia or premature ASCVD, hallmarks of FH on physical examination, and plasma LDL-C levels to identify patients at high risk of a FH-causing mutation carrier status. The highest applicable score is chosen per diagnostic group. A definite FH diagnosis can be made if the total score is >8 points, a probable FH is made when 6 to 8 points are scored, a possible FH is referred to a range of 3 to 5 points, and an unlikely FH is made if the subject scores 0 to 2 points. The prototype of this electronic phenotyping algorithm has been tested and validated for the purpose of targeted screening among individuals with severe hypercholesterolemia.[9]

ALGORITHM DESIGN AND AIMS

Figure 1 depicts notations for individual components of the FH eAlgirthm. Structured data are processed using preset codes and unstructured data are processed using natural language processing (NLP). This algorithm utilizes a two-stage approach to identify a case-control status.

Stage I identifies **cases and controls for primary hypercholesterolemia** using the structured data available in EHR (**Figure** 2). All patients above 18 years of age and with a lipid profile in EHR will be the parent sample set for identifying cases and controls.

Index date is defined as a date with the highest LDL-C levels in the EHR calculated using the Friedewald equation* or measured directly.

*[LDL-C] = [Total cholesterol] - [HDL-C] - ([TG]/5) for mg/dL
[LDL-C] = [Total cholesterol] - [HDL-C] - ([TG]/2.2) for mmol/L

HDL-C = high density lipoprotein cholesterol
LDL-C = low density lipoprotein cholesterol
TG = triglycerides

If the patient is on lipid lowering treatment within one year prior to the index date, a 30% reduction in LDL-C is expected as a therapy effect and adjusted LDL-C level is estimated by dividing index date LDL-C by 0.7.

Since LDL-C levels can be affected by a variety of medical and therapeutic conditions, a multi-step screening is implemented to detect and flag/exclude test results measured at the time of active underlying disorder causing an apparent increase in the LDL-C levels (secondary causes of hypercholesterolemia; **Table 2A**). All the patients identified to have secondary causes of hypercholesterolemia should be excluded. Patients with triglycerides levels more than 220 mg/dL and those who are pregnant (identified using ICD-9 coded, **Table 2B**) with LDL-C >=155 mg/dL need to be flagged. A "flagging" term has been introduced to increase flexibility of the system and fit a model to a set making reliable predictions on heterozygous FH accompanied by hypertriglyceridemia.

Stage II identifies **cases and controls for FH** by ascertaining **modified DLCN criteria for heterozygous FH** through both structured (Laboratory and Medication data, Clinical Diagnosis in Personal History) and unstructured (Family History of Hypercholesterolemia and Premature ASCVD, Personal History of premature ASCVD) data in clinical notes. Each patient gets scored based on four sets of parameters and the FH status is then ascertained as case, control or unknown, as described in **Figure 3**.

FINAL OUTPUT

Primary outcome: case/control/unknown status for FH
Secondary outcomes: (i) a case/control/unknown status for primary hypercholesterolemia, (ii) demographics of each individual (age at the time of qualifying LDL-C ascertainment, gender, race/ethnicity), (iii) lipid profile (total cholesterol, LDL-C, HDL-C, triglycerides), (iv) lipid-lowering treatment and difference in time between the index date and date of treatment ascertainment, (v) personal history of premature ASCVD and/or hypercholesterolemia, (vi) family history of premature ASCVD, (vii) xanthomas and/or early corneal arcus, (viii) DLCN score.

Proposed Genetic Analyses as of 2016
1. Discovery of new FH-causing genes through a sequencing approach
2. Identifying genetic risk variants for FH through a genome-wide association study approach

ACRONYMS
ASCVD = Atherosclerotic Cardiovascular Disease that includes:
- CHD - Coronary Heart Disease
- CVD - Cerebral Vascular Disease
- PAD - Peripheral Arterial disease

CPT = Current Procedural Terminology codes.
EHR=electronic health record
ICD-9-CM = International Classification of Diseases, Ninth Revision, Clinical Modification
LDL-C = Low-Density Lipoprotein Cholesterol
PPI = Patient Provided Information
Rx=Prescription
International System of Units (SI) conversion factor: To convert total, LDL-C, HDL-C to mmol/L, multiply by 0.0259, for triglycerides, multiply by 0.0113.

IMPLEMENTATION CONCEPT

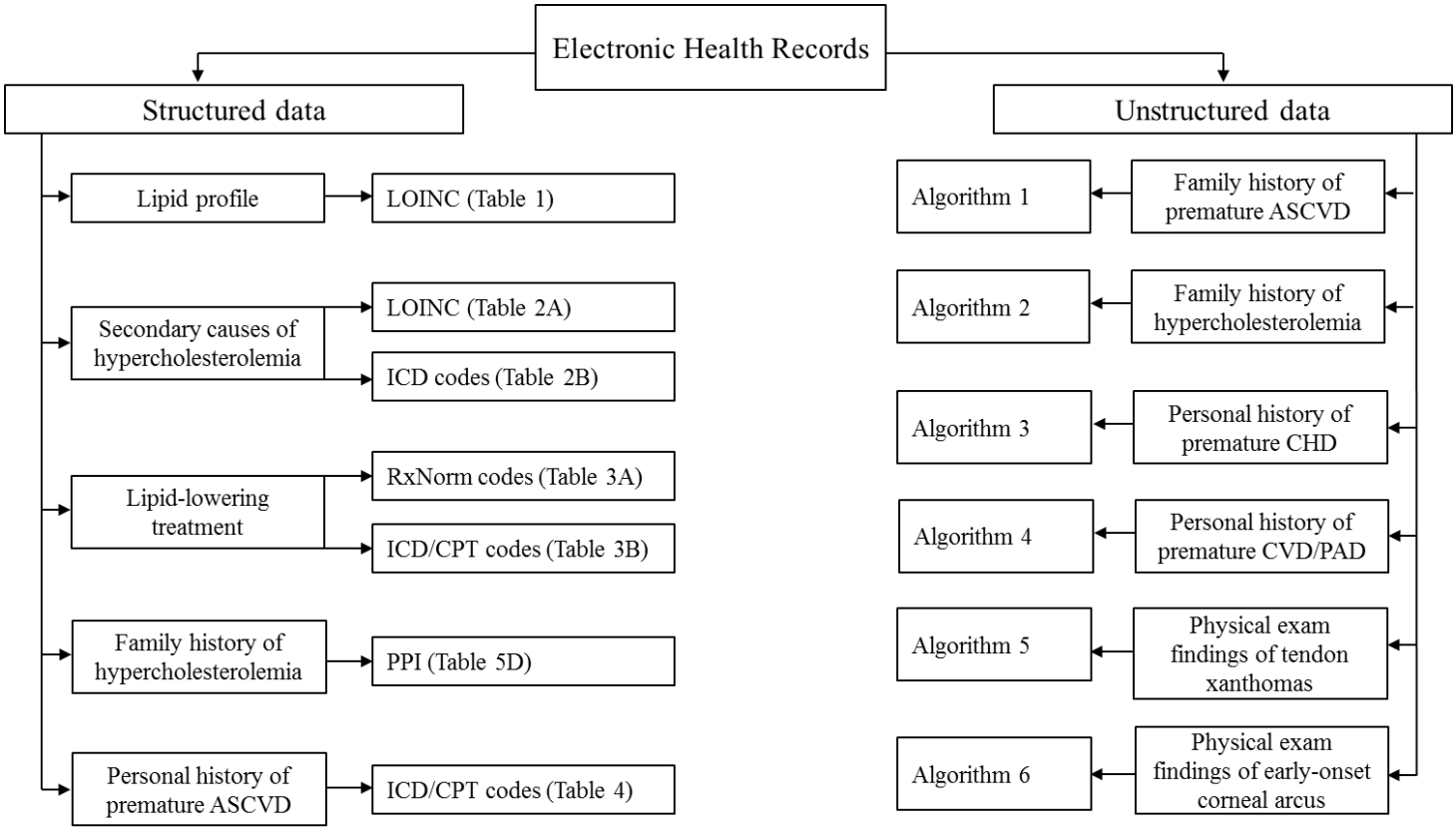

Figure 1. Input to the eAlgorithm for familial hypercholesterolemia.

Stage I: Assigning a case-control status for primary hypercholesterolemia in the general population (**Figure 2**).

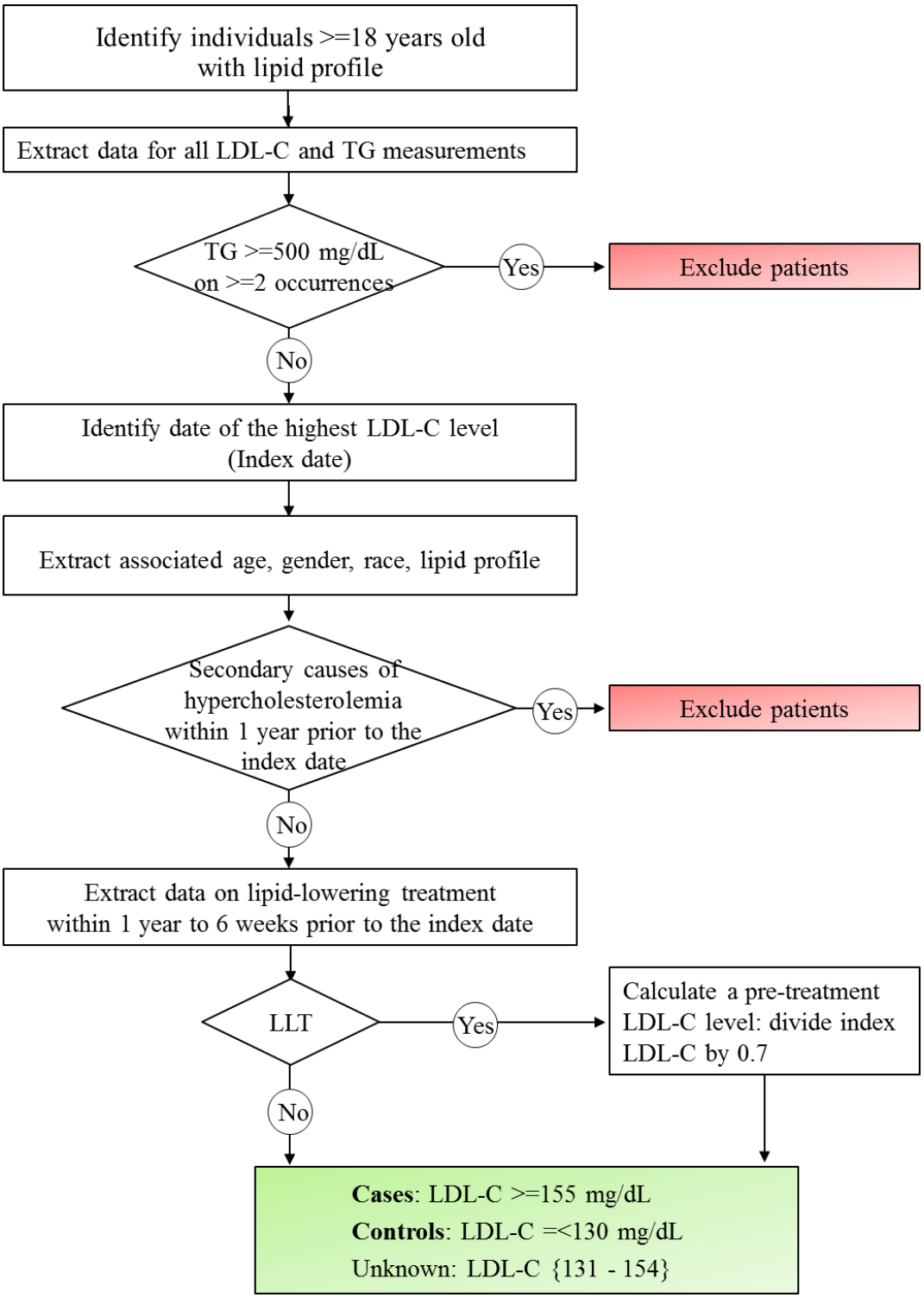

Figure 2. Logic used to ascertain primary hypercholesterolemia from EHR.

LLT=lipid-lowering treatment, LDL-C=low-density lipoprotein cholesterol, TG=triglycerides

Stage II: Use a scoring system (**Figure 3**) to weigh each component of the algorithm to ascertain FH case status in individuals with hypercholesterolemia. Per diagnostic group choose the highest applicable score. Provide a final score by summing up each individual score per each group.

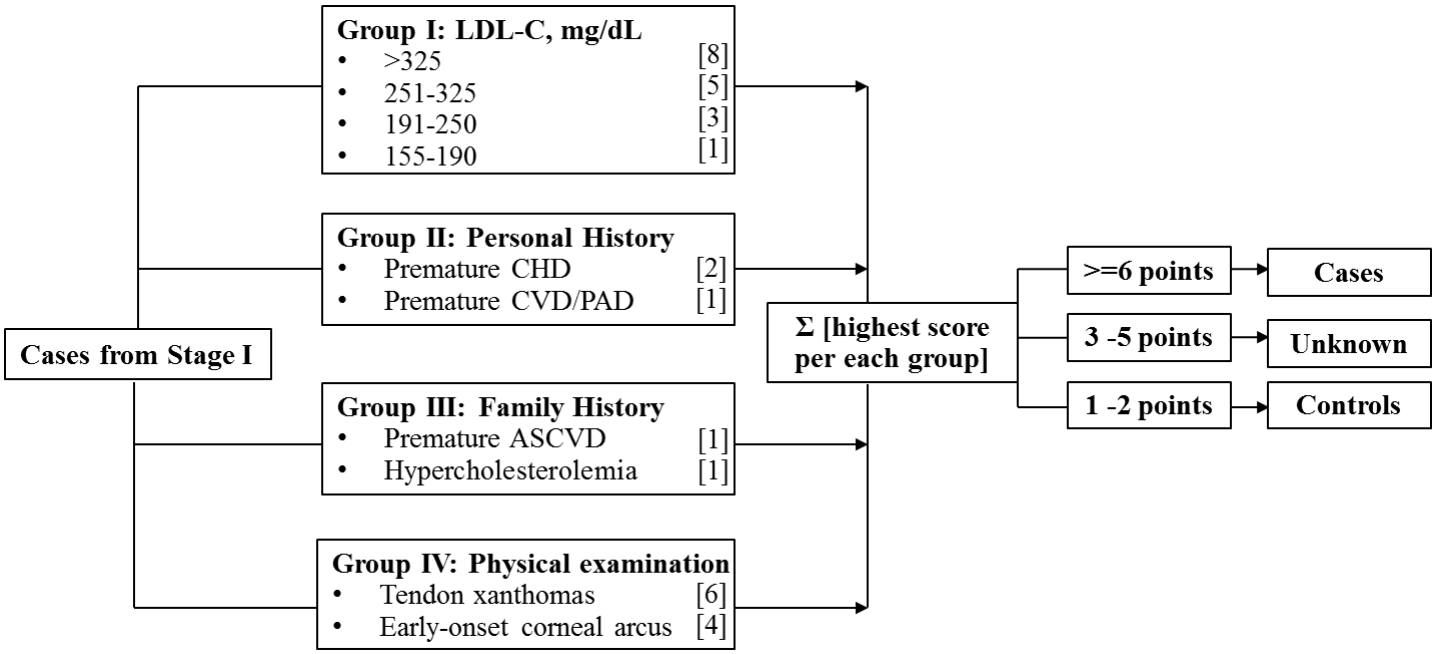

Figure 3. Logic and definitions used to ascertain familial hypercholesterolemia from EHR.

DATA ELEMENTS REQUIRED FOR THE EHR-BASED PHENOTYPING ALGORITHM FOR FH

Table 1. Laboratory test catalog with codes for lipid and lipoprotein test results.

Parameter	LOINC	DESCRIPTION
Total cholesterol (TC), mg/dL	2093-3	Mass/volume in Serum or Plasma
	48620-9	Mass/volume in Serum or Plasma ultracentrifugate
	35200-5	Mass or Molecules/volume in Serum or Plasma
	14647-2	Molecules/volume in Serum or Plasma
LDL-C, mg/dL	2089-1	Mass/volume in Serum or Plasma (beta-quantification)
	18262-6	Mass/volume in Serum or Plasma by Direct assay
	49132-4	Mass/volume in Serum or Plasma by Electrophoresis
	35198-1	Mass or Molecules/volume in Serum or Plasma
	39469-2	Molecules/volume in Serum or Plasma by Calculated
	12773-8	Units/volume in Serum or Plasma by Electrophoresis
	18261-8	Mass/volume in Serum or Plasma ultracentrifugate
	22748-8	Molecules/volume in Serum or Plasma
	13457-7	Mass/volume in Serum or Plasma by Calculated
	9346-8	Mass/volume in Serum or Plasma by Calculated
	2574-2	Mass/volume in Serum or Plasma
	14815-5	Molecules/volume in Serum or Plasma by Calculated
High-density lipoprotein (HDL), mg/dL	2085-9	Mass/volume in Serum or Plasma
	49130-8	Mass/volume in Serum or Plasma by Electrophoresis
	35197-3	Mass or Molecules/volume in Serum or Plasma
	12771-2	Presence in Serum or Plasma by Electrophoresis
	12772-0	Units/volume in Serum or Plasma by Electrophoresis
	18263-4	Mass/volume in Serum or Plasma ultracentrifugate
	27340-9	Presence in Serum or Plasma
	14646-4	Molecules/volume in Serum or Plasma
Triglycerides (TG), mg/dL	2571-8	Mass/volume in Serum or Plasma
	30524-3	Molecules/volume - 12h post cfst- in Serum or Plasma
	3048-6	Mass/volume -post CFst- in Serum or Plasma
	35217-9	Mass or Molecules/volume in Serum or Plasma
	28554-4	Presence in Serum or Plasma
	14927-8	Molecules/volume in Serum or Plasma
	47210-0	Molecules/volume -post CFst- in Serum or Plasma

LOINC=Logical Observation Identifiers Names and Codes

Secondary causes of hypercholesterolemia: a condition should be deemed active when abnormal laboratory values defined in Table 2A are present within 1 year prior to the index date; ICD-9 codes are used to ascertain pregnancy within 1 year prior to the index date.

Table 2A. Laboratory test catalog to ascertain secondary causes of hypercholesterolemia using codes for test results.

Condition	Parameter	Value	LOINC	
Hypothyroidism	Thyroid-stimulating hormone (TSH)	2 ULN (≥10 mIU/L)	11579-0	
			24348-5	Free T4 & TSH panel in Serum or Plasma
Biliary obstruction	Alkaline phosphatase	2 ULN (≥200 IU/L)	6768-6	Enzymatic activity/volume in Serum or Plasma
			12805-8	Units/volume in Serum or Plasma
Liver disease	Total bilirubin	2 ULN (>2.0 mg/dL)	35194-0	Mass and Molecules/volume in Serum or Plasma
			1975-2	Mass/volume in Serum or Plasma
			14631-6	Molecules/volume in Serum or Plasma
Nephrotic syndrome	Protein in a 24-h urine collection	>3g	21482-5	Mass/volume in 24 hour Urine
			2889-4	Mass/time in 24 hour Urine
			21028-6	Interpretation in 24 hour Urine
	Urine protein/creatinine ratio (=Albumin/Creat 24h Ur)	>3.0	13801-6	Protein/Creatinine [Mass Ratio] in 24 hour Urine
			2890-2	Protein/Creatinine [Mass Ratio] in Urine
Renal failure	Creatinine	>2.6 mg/dL	14682-9	Mass/volume in Serum or Plasma
			2160-0	Mass/volume in Serum or Plasma
			35203-9	Mass/volume in Serum or Plasma
			38483-4	Mass/volume in Blood
			59826-8	Moles/volume in Blood
			77140-2	Moles/volume in Serum, Plasma or Blood
			14682-9	Mass/volume in Serum or Plasma
	Estimated glomerular filtration rate (eGFR)	<15 mL/min/BSA	50261-7	Renal function panel and Glomerular filtration rate predicted in Serum or Plasma
			45066-8	Creatinine & Glomerular filtration rate predicted panel in Serum or Plasma
			48642-3	Glomerular filtration rate/1.73 m^2 predicted Female Flow in Serum or Plasma by Creatinine-based formula (MDRD)

			48642-3	Glomerular filtration rate/1.73 m^2 predicted Non-black [Flow] in Serum or Plasma by Creatinine-based formula (MDRD)
			48643-1	Glomerular filtration rate/1.73 m^2 predicted Black [Flow] in Serum or Plasma by Creatinine-based formula (MDRD)
			33914-3	Glomerular filtration rate/1.73 m^2 predicted Flow in Serum or Plasma by Creatinine-based formula (MDRD)
Diabetes	HbA1c	>9%	4549-2	Hemoglobin A1c (glycated HgB)/Hemoglobin Total in Blood by Electrophoresis
			17855-8	Hemoglobin A1c (glycated HgB)/Hemoglobin Total in Blood by Calculated
			17856-6	Hemoglobin A1c (glycated HgB)/Hemoglobin Total in Blood by High pressure liquid chromatography (HPLC)
			41995-2	Hemoglobin A1c (glycated HgB) Mass/volume in Blood
	Fasting Glucose	>200 mg/dL	1556-0	Fasting glucose [Mass/volume] in Capillary blood
		>220 mg/dL	1558-6	Fasting glucose [Mass/volume] in Serum or Plasma

ULN=upper limit of normal

Table 2B. ICD-9 codes to ascertain pregnancy.

Phenotype	Code number	Description
Pregnancy	V22	Normal pregnancy
	V23	Supervision of high-risk pregnancy
	645	Late pregnancy
	651	Multiple gestation
	652	Malposition and malpresentation of fetus

Table 3A. Compound, brand name and RxNorm codes for most commonly used lipid-lowering medications.

Compound	Brand name	RxNORM codes (Notation)
Simvastatin	Zocor; Lipex	36567
Fluvastatin	Lescol; Lescol xl	41127
Lovastatin	Mevacor; Altocor; Altoprev	6472
Pravastatin	Pravachol; Lipostat	42463
Pitavastatin	Livalo; Pitava	861634

Atorvastatin	Lipitor; Torvast; Atorlip; Tulip Torvacard	83367
Rosuvastatin	Crestor	301542
Cerivastatin	Lipobay; Baycol	221072 1152441
Niacin	Niaspan; Niacor	7393
Fenofibrate	Tricor; Triglide; Antara; Lipofen Fenoglide; Trilipix; Fibricor Lofibra	8703
Gemfibrozil	Lopid	4719
Ezetimibe	Ezetrol; Zetia	341248
Colesevelam	Welchol	141626
Cholestyramine	Questran Questran Light; Prevalite	2447
Colestipol	Colestid	2685
Mipomersen	Kynamro	1367839
Lomitapide	Juxtapid	1364479
Evolocumab	Repatha	1665895 1665900 1665904 1665906
Alirocumab	Praluent	1659156 1659161 1659165 1659167 1659177 1659179 1659182 1659183
Ezetimibe/simvastatin	Vytorin	495215
Niacin/simvastatin	Simcor	1372731
niacin/lovastatin	Advicor	327008
amlodipine/atorvastatin	Caduet	404914
sitagliptin/simvastatin	Juvisync	1372754

Table 3B. CPT codes of procedures and ICD-9-CM procedure codes for commonly used extracorporeal lipid-lowering treatment procedures.

Extracorporeal immunoadsorption and plasma reinfusion	36515 99.76
Extracorporeal selective adsorption or selective filtration and plasma reinfusion	36516

Premature Atherosclerotic Cardiovascular Disease (ASCVD) includes coronary heart disease (CHD), cerebral vascular disease (CVD) and peripheral arterial disease (PAD).
Premature ASCVD case status: presence of two or more pertinent diagnosis and/or procedural codes in EHR before age 56 in men and 66 in women.
Assigned codes should be evaluated at discharge from each encounter during the surveillance period.

Table 4. ICD-9-CM codes and CPT4 codes for events/procedures defining ASCVD cases.

Phenotype	Code number*	Description
CORONARY HEART DISEASE		
Angina; ICD-9-CM	413.0	Angina decubitus
	413.1	Prinzmetal angina
	413.9	Other and unspecified angina pectoris
Myocardial infarction; ICD-9-CM	410.00	Acute myocardial infarction of anterolateral wall episode of care unspecified
	410.01	Acute myocardial infarction of anterolateral wall initial episode of care
	410.02	Acute myocardial infarction of anterolatera l wall subsequent episode of care
	410.10	Acute myocardial infarction of other anterior wall episode of care unspecified
	410.11	Acute myocardial infarction of other anterior wall initial episode of care
	410.12	Acute myocardial infarction of other anterior wall subsequent episode of care
	410.20	Acute myocardial infarction of inferolateral wall episode of care unspecified
	410.21	Acute myocardial infarction of inferolateral wall initial episode of care
	410.22	Acute myocardial infarction of inferolateral wall subsequent episode of care
	410.30	Acute myocardial infarction of inferoposterior wall episode of care unspecified
	410.31	Acute myocardial infarction of inferoposterior wall initial episode of care
	410.32	Acute myocardial infarction of inferoposterior wall subsequent episode of care
	410.40	Acute myocardial infarction of other inferior wall episode of care unspecified
	410.41	Acute myocardial infarction of other inferior wall initial episode of care
	410.42	Acute myocardial infarction of other inferior wall subsequent episode of care
	410.50	Acute myocardial infarction of other lateral wall episode of care unspecified
	410.51	Acute myocardial infarction of other lateral wall initial episode of care

	410.52	Acute myocardial infarction of other lateral wall subsequent episode of care
	410.60	True posterior wall infarction episode of care unspecified
	410.61	True posterior wall infarction initial episode of care
	410.62	True posterior wall infarction subsequent episode of care
	410.70	Subendocardial infarction episode of care unspecified
	410.71	Subendocardial infarction initial episode of care
	410.72	Subendocardial infarction subsequent episode of care
	410.80	Acute myocardial infarction of other specified sites episode of care unspecified
	410.81	Acute myocardial infarction of other specified sites initial episode of care
	410.82	Acute myocardial infarction of other specified sites subsequent episode of care
	410.90	Acute myocardial infarction of unspecified site episode of care unspecified
	410.91	Acute myocardial infarction of unspecified site initial episode of care
	410.92	Acute myocardial infarction of unspecified site subsequent episode of care
	412	Old myocardial infarction
	429.71	Certain sequelae of myocardial infarction not elsewhere classified acquired cardiac septal defect
	429.79	Certain sequelae of myocardial infarction not elsewhere classified other
Coronary atherosclerosis/ chronic ischemic heart disease	414.00	Coronary atherosclerosis of unspecified type of vessel native or graft
	414.01	Coronary atherosclerosis of native coronary artery
	414.02	Coronary atherosclerosis of autologous vein bypass graft
	414.03	Coronary atherosclerosis of nonautologous biological bypass graft
	414.04	Coronary atherosclerosis of artery bypass graft
	414.05	Coronary atherosclerosis of unspecified bypass graft
	414.06	Coronary atherosclerosis of native coronary artery of transplanted heart
	414.07	Coronary atherosclerosis of bypass graft (artery) (vein) of transplanted heart
Percutaneous coronary revascularization CPT-4 Codes:	92920-92921; 92924-92925; 92928-92929; 92933-92934; 92937-92938; 92941; 92943-92944; 92980-82, 92984, 92995-6, 92973-92974	
Percutaneous coronary revascularization ICD-9-CM	36.01-36.07, 36.09 00.66	

Procedure Codes:		
Coronary bypass surgery CPT-4 Codes:	33510-33514 33516-33519 33521-33523 33533-33536	
Coronary bypass surgery ICD9P	36.10 - 36.19, 36.2	
CEREBROVASCULAR DISEASE		
ICD-9:	434	Occlusion of cerebral arteries
Stroke	434.00	Cerebral thrombosis without mention of cerebral infarction
	434.01	Cerebral thrombosis with cerebral infarction
	434.10	Cerebral embolism without mention of cerebral infarction
	434.11	Cerebral embolism with cerebral infarction
	434.90	Cerebral artery occlusion, unspecified without mention of cerebral infarction
	434.91	Cerebral artery occlusion, unspecified with cerebral infarction
	437.0	Cerebral atherosclerosis
	437.1	Other generalized ischemic cerebrovascular disease
Transient Ischemic Attack	435.0	Basilar artery syndrome
	435.1	Vertebral artery syndrome
	435.2	Subclavian steal syndrome
	435.3	Vertebrobasilar artery syndrome
	435.8	Other specified transient cerebral ischemias
	435.9	Unspecified transient cerebral ischemia
Carotid artery disease	433	Occlusion and stenosis of precerebral arteries
	433.00	Occlusion and stenosis of basilar artery without mention of cerebral infarction
	433.01	Occlusion and stenosis of basilar artery with cerebral infarction
	433.10	Occlusion and stenosis of carotid artery without mention of cerebral infarction
	433.11	Occlusion and stenosis of carotid artery with cerebral infarction
	433.20	Occlusion and stenosis of vertebral artery without mention of cerebral infarction
	433.21	Occlusion and stenosis of vertebral artery with cerebral infarction
	433.30	Occlusion and stenosis of multiple and bilateral precerebral arteries without mention of cerebral infarction
	433.31	Occlusion and stenosis of multiple and bilateral precerebral arteries with cerebral infarction
	433.80	Occlusion and stenosis of other specified precerebral artery without mention of cerebral infarction
	433.81	Occlusion and stenosis of other specified precerebral artery with cerebral infarction
	433.90	Occlusion and stenosis of unspecified precerebral artery without mention of cerebral infarction
	433.91	Occlusion and stenosis of unspecified precerebral artery with cerebral infarction
ICD 9-CM procedure codes:	38.12	Endarterectomy of other vessels of head / neck
	38.11	Endarterectomy of intracranial vessels

	00.61	Extracranial percutaneous angioplasty
	39.28	Extracranial-intracranial vascular bypass
	00.63	Carotid artery stent
CPT procedure codes:	35301	Thromboendarterectomy of carotid, vertebral, subclavian
	37215	Transcatheter placement of intravascular stent(s), cervical carotid artery, open or percutaneous, including angioplasty, when performed, and radiological supervision and interpretation; with distal embolic protection
	37216	Transcatheter placement of intravascular stent(s), cervical carotid artery, open or percutaneous, including angioplasty, when performed, and radiological supervision and interpretation; without distal embolic protection.
PERIPHERAL ARTERIAL DISEASE		
ICD-9 codes:	440.20	Atherosclerosis of native arteries of the extremities unspecified
	440.21	Atherosclerosis of native arteries of the extremities with intermittent claudication
	440.22	Atherosclerosis of native arteries of the extremities with rest pain
	440.23	Atherosclerosis of native arteries of the extremities with ulceration
	440.24	Atherosclerosis of native arteries of the extremities with gangrene
	440.29	Other atherosclerosis of native arteries of the extremities
Exclude if ICD-9 codes are with ≥2 occurrences of the following codes	237.70	Neurofibromatosis, unspecified
	237.71	Neurofibromatosis, type 1
	237.72	Neurofibromatosis, type 2 (acoustic neurofibromatosis)
	237.73	Schwannomatosis
	237.79	Other neurofibromatosis
	443.1	Thromboangiitis obliterans (Buerger's disease)
	446.0	Polyarteritis nodosa
	446.4	Wegener's granulomatosis
	446.5	Giant cell arteritis
	446.6	Thrombotic microangiopathy
	446.7	Takayasu's disease
	710.1	Systemic sclerosis
	747.10	Coarctation of aorta (preductal, postductal)
	747.11	Interruption of aortic arch
	747.22	Atresia and stenosis of aorta
	747.64	Lower limb vessel anomaly

RULE-BASED NLP PREDICTION

The NLP system for extracting results of a family history of ASCVD events and hypercholesterolemia, xanthomas and corneal arcus can be downloaded from PheKB.[15] The NLP system is an easily installed Java application with a simple graphical user interface. The user guide provides step-by-step user instructions. 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. The basis of the NLP system in the algorithm is to scan for pertinent keywords in unstructured data in clinical notes and returning a quantitative code that can be further used along with the return codes from other segments of the algorithm to diagnose FH.

The system is utilized for identification of (i) a family history of premature ASCVD, (ii) a family history of premature hypercholesterolemia, (iii) personal history of premature ASCVD, (iv) tendon xanthomas, (v) corneal arcus. Unstructured data sections of clinical notes scanned by the NLP system contain "Family History" and "Personal Medical History" sections. Structured data include PPI (patient provided information) in EHR.

NLP system for "Family History"
A set of keywords is searched to find positive history of ASCVD events. These events occurring in first-degree relatives within the predefined age brackets yield a positive family history of premature ASCVD as outlined in **Figure 4.** A similar system customized to detect a personal history of ASCVD can be used to complement the code-based logic (**Table 1**).

Figure 4. Conceptual outline of NLP logic for family history.

A "family history of ASCVD / hypercholesterolemia" concept identification from EHR: NLP system utilizes keywords to identify specific ASCVD events including coronary heart disease (CHD; **Table 5A**) and cerebral/peripheral vascular diseases (CVD/PAD; **Table 5B**) from the unstructured data in EHR. Terms specific for each category are listed in the column entitled '*Specific*' whereas the generic terms that usually accompany the specific terms in any combination are listed in the '*Generic*' column. For a positive search, one of the specific terms followed by any of the generic terms in the corresponding category should be present in the relevant section of the clinical notes. A definite positive result is also ascertained by searching for specific keywords in varying combinations listed in **Table 5C.**

Tables 5E-G provide keywords for negation, temporality, and a specific pedigree relationship. A definite negative result is also ascertained by searching for specific keywords in varying combinations as outlined in **Table 5H.** For example, if 'no history of early CHD' or 'negative history of premature heart disease', etc. is mentioned in clinical notes, a definite negative result is returned.

I. Keyword search by the NLP system

Table 5A. Keywords for a history of CHD.

CHD Category	Keywords	
	Specific	Generic
Conditions	heart cardiac myocardial infarction mi cad chd ascvd cardiovascular cardio_vascular	attack(s) disease(s) disorder(s)* problem(s) related_problem(s)
Procedures	bypass angioplasty(y|ies) stent(s) coronary artery coronary artery bypass cabg stenting aorto_iliac	surger(y|ies) procedure(s) placement repair(s) graft(s) bypass

Footnote: *For the EHR systems with a structured family history section

Table 5B. Keywords for a history of CVD/PAD.

CVD/PAD Category	Keywords	
	Specific	Generic
Conditions	transient ischemic tia cerebrovascular cerebral_vascular cva stroke(s) cvd critical limb ischemia claudication	attack(s) accident(s)
Other vascular	carotid_endarterectom(y|ies) carotid thrombo transcath femoro_popliteal	stenting stent placement(s) endarterectom(y|ies) placement(s) bypass

Table 5C. Proposed combinations of keywords to ascertain a *definite* positive history of *premature* CHD.

Category		Keywords
Specific words for positive history	early early_onset premature	cad chd heart_disease cardiac_disease cardiac_problem(s) cardiovascular_disease(s) coronary_heart_disease

Table 5D. Keywords for a family history of hypercholesterolemia.

- Hypercholesterolemia
- High_cholesterol
- Elevated_cholesterol
- Hyperlipidemia
- Dyslipidemia
- High_lipid|s
- Elevated_lipid|s

II. Temporal reasoning: keyword search is followed by a matching to determine events that occurred within gender-specific age brackets.

Table 5E. Terms to identify age brackets for premature ASCVD events.

Males	Females
less than 56 years|y|yrs	less than 66 years|y|yrs
=<55 years|y|yrs	=<65 years|y|yrs
early 50s	early 60s

III. Concept identification: the degree of relatedness is ascertained through keyword search listed in **Table 5F**, excluding events occurring in any relatives beyond this one degree of relation.

Table 5F. Terms to identify degree of relatedness.

First-degree relatives (inclusion criteria)	Other relatives (exclusion criteria)
father	uncle(s)
mother	aunt(s)

sister(s)	grandparent(s)
brother(s)	grandfather(s)
son(s)	grandmother(s)
daughter(s)	GF
sibling(s)	GM
kid(s)	PGF
children	PGM
child	MGF
	MGM
	cousin(s)
	nephew(s)
	niece(s)
	husband

IV. Exclusion criteria

A negation status for ASCVD events is identified through specific keyword search (**Table 5G**).

Table 5G. Keywords for negation identification

Category	**Keywords**
Excluded terms	murmur
	chronic_heart_failure
	heart_failure
	chf
	arrhythmia
	atrial_fibrillation
	af
	rheumatic
	valvular_heart_disease
	congenital_heart_disease
	late 60s IF in females || late 50s IF in males

Table 5H. Proposed combinations of keywords to ascertain a *definite* negative history of CHD.

Category	**Keywords**			
Specific words for negative history	no	(history) of/for	early	cad
	negative		early_onset	chd
	neg		premature	heart_disease
	no_positive			cardiac_disease
				cardiac_problem(s)
				cardiovascular_disease(s)

NLP ALGORITHM PSEUDOCODE

NLP system for a family history (Algorithm 1 for ASCVD events and Algorithms 2 for hypercholesterolemia):
A code of 1 is returned for a diagnosis of premature ASCVD (CHD/CVD/PAD) if the keyword search in the "Family History" of clinical notes detects any of the aforementioned ASCVD-specific terms in first-degree relatives within the predefined age brackets (Appendix 1**)**.

Algorithm 1 Function of retrieving family history of premature ASCVD from EHR

Retrieve patient's family history section in EHR using clinic identifier

Search a family history of CHD/CVD/PAD using keywords in **Table 5A; Table 5B.** Keywords for a history of CVD/PAD.**; Table 5G**

IF any ASCVD events is found

Identify the age of this family member when s/he was diagnosed with ASCVD

IF the age threshold is met according to **Table 5E.** Terms to identify age brackets for premature ASCVD events.

GOTO: Checkpoint

ELSE

The patient has negative family history of premature ASCVD

Return 0

ENDIF

ELSE

Search a family history of premature CHD using keywords in **Table 5C**and **Table 5H**

IF the premature CHD event is found

GOTO: Checkpoint

ELSE

The patient has negative family history of premature ASCVD

Return 0

ENDIF

ENDIF

Checkpoint:

Ascertain relatedness using keywords in **Table 5F**

IF the family member is a first-degree relative of the patient

The patient has positive family history of premature ASCVD

Return 1

ELSE

The patient has negative family history of premature ASCVD

Return 0

ENDIF

A code of 1 is returned for a family history of hypercholesterolemia if the keyword search in the "Family History" and/or PPI sections in EHR detects any of the aforementioned hypercholesterolemia-specific terms in first-degree relatives (Appendix 1, Figure 4B, NLP System, Logic for returning a positive/negative family history of hypercholesterolemia from EHR**)**.

Algorithm 2 Function of retrieving family history of hypercholesterolemia from EHR

Retrieve patient provided information section using clinic identifier

IF any ASCVD events is found

GOTO: Checkpoint

ELSE

Retrieve patient's family history section in EHR using clinic identifier

Search a family history of hypercholesterolemia using keywords in **Table 5D**.

IF hypercholesterolemia events is found

GOTO: checkpoint

ELSE

The patient has negative family history of hypercholesterolemia

Return 0

ENDIF

ENDIF

Checkpoint:

Ascertain relatedness using keywords in **Table 5F**

IF the family member is a first-degree relative of the patient

The patient has positive family history of hypercholesterolemia

Return 1

ELSE

The patient has negative family history of hypercholesterolemia

Return 0

ENDIF

NLP system for a personal history of premature ASCVD (Algorithm 3 for CHD and Algorithm 4 for CVD/PAD):
A "personal history of CHD and CVD/PAD" concept identification from EHR utilizes similar logic of target keyword search in the "Personal History" section of EHR. A code of 2 is returned for positive keyword search of CHD events within the standard age bracket. A code of 1 is returned for positive keyword search of CVD/PAD events within the standard age bracket.

When combined with the code-based logic for a personal history of ASCVD the earliest year of ASCVD event occurrence from codes or NLP system defines the status of "prematurity".
(Appendix 1, Figure 5A, NLP System, Logic for returning a positive/negative personal history of CHD from EHR; Figure 5B. NLP System, Logic for returning a positive/negative personal history of CVD/PAD.**)**.

Algorithm 3 Function of retrieving personal history of premature CHD from EHR

Retrieve patient's personal history section in EHR using clinic identifier

Search a personal history of CHD using keywords in **Table 5A; Table 5G**

IF CHD events is found

Identify the age of the patient when s/he was diagnosed with CHD

IF the age threshold is met according to **Table 5E.** Terms to identify age brackets for premature ASCVD events.

The patient has positive personal history of premature CHD

Return 2

ELSE

The patient has negative personal history of premature CHD

Return 0

ENDIF

ELSE

The patient has negative personal history of premature CHD

Return 0

ENDIF

Algorithm 4 Function of retrieving personal history of premature CVD/PAD from EHR

Retrieve patient's personal history section in EHR using clinic identifier

Search a personal history of CVD/PAD using keywords in **Table 5B.** Keywords for a history of CVD/PAD.

IF CVD/PAD events is found

Identify the age of the patient when s/he was diagnosed with CVD/PAD

IF the age threshold is met according to **Table 5E.** Terms to identify age brackets for premature ASCVD events.

The patient has positive personal history of premature CVD/PAD

Return 1

ELSE

The patient has negative personal history of premature CVD/PAD

Return 0

ENDIF

ELSE

The patient has negative personal history of premature CVD/PAD

Return 0

ENDIF

NLP system for xanthomas (Algorithm 5)
I. A code of 6 is returned for a positive target keyword search of "xanthomas" (**Table 6**) from the "Physical Examination" section in EHR.
II. If the keywords are identified in conjunction with 'No', return a code of 0.
(Appendix 1, Figure 6; NLP system; Logic for returning a positive/negative xanthoma variable from EHR).

Table 6. Keywords for xanthomas.

Acceptable terms
• Achilles_Tendinous_Xanthomas • Achilles_tendon_xanthoma • Cutaneous_Xanthomas • Nodular_Achilles_Tendons • Palmar_Xanthomas • Palpable_Xanthomas • Possible_Tendon_Xanthoma

- Tendinous_Xanthoma
- Tendinous_Xanthomas
- Tendinous_Xanthomata
- Tendon_Xanthoma
- Tendon_Xanthomas
- Tendon_Xanthomata
- Thickened_Achilles_Tendon
- Thickening_Of_Achilles_Tendon
- Tuberoeruptive_Xanthomas
- Tuberous_Xanthomas
- Xanthomas
- Xanthomas_Of_Achilles_Tendon
- Xanthomas_Of_The_Hands
- Xanthomas_On_The_Extensor_Surfaces

Excluded terms:

- Eruptive xanthomas
- Irruptive xanthomas
- Xanthomas around both eyes
- Xanthomas around the left eye
- Xanthomas around the right eye
- Eyelid xanthomas
- Eye lid xanthomas
- Xanthomas on the eye
- Xanthomas over eyelids
- Xanthomas over the eyes
- Xanthomas over both eyelids
- Xanthomas on the upper lids
- Xanthomas beneath both eyes
- Xanthomas beneath her eyes
- Xanthomas beneath his eyes
- Xanthomas beneath her eye
- Xanthomas beneath his eye
- Xanthomas beneath eyes
- Xanthomas beneath eye
- Xanthomas beneath eyelid
- Xanthomas beneath his eyelid
- Xanthomas beneath her eyelid
- Xanthomas on the right eyelid
- Xanthomas on the left eyelid
- Xanthomas on the face
- Xanthomas surrounding his eyes
- Xanthomas surrounding her eyes

- Xanthomas surrounding his eye
- Xanthomas surrounding her eye
- Scleral xanthomas
- Eruptive or papular xanthomas
- Xanthelasmas
- Xanthomas involving her face
- Xanthomas involving his face
- Sarcoid xanthomas
- Verruciform xanthomas; lesions on the scrotum
- Fibroxanthomas
- Verruciform xanthoma on the scrotum
- Verruciform xanthoma of the scrotum
- Verruciform xanthomas on the scrotum
- Verruciform xanthomas of the scrotum
- Verruciform xanthomas
- Achilles tendonitis

Algorithm 5 Function of retrieving xanthoma diagnosis from EHR

Retrieve patient's physical exam section in EHR using clinic identifier

Search a xanthoma diagnosis using keywords in **Table 6.**

IF xanthoma diagnosis is found

    The patient has positive diagnosis of xanthoma

    Return 6

ELSE

    The patient has negative diagnosis of xanthoma

    Return 0

ENDIF

NLP system for early corneal arcus (Algorithm 6)
I. A code of 4 is returned for a positive target keyword search of "arcus" from the "Physical Examination" section in EHR when occurring in an individual aged less than 45 years old.
II. Negation is identified if the keywords are in conjunction with 'No' in the "Physical Exam" section in EHR.
(Appendix 1, Figure 7, NLP system, Logic for returning a positive/negative arcus variable from EHR.

Table 7. Keywords for corneal arcus.

Corneal_arcus Arcus_cornealis Early_arcus_senilis

Algorithm 6 Function of retrieving early-onset arcus from EHR

Retrieve patient's physical exam section in EHR using clinic identifier
Search an arcus diagnosis using keywords in **Table 7**
IF arcus diagnosis is found
    Identify the age of the patient when s/he was diagnosed with arcus
    IF the age < 45
        The patient has positive diagnosis of early-onset arcus
        Return 4
    ELSE
        The patient has negative diagnosis of early-onset arcus
        Return 0
    ENDIF
ELSE
    The patient has negative diagnosis of early-onset arcus
    Return 0
ENDIF

A link to MedTagger installation and user guides:

http://ohnlp.org/index.php/MedTagger_Project_Page

REFERENCES

1. Goldberg AC, Hopkins PN, Toth PP, et al. Familial hypercholesterolemia: screening, diagnosis and management of pediatric and adult patients: clinical guidance from the National Lipid Association Expert Panel on Familial Hypercholesterolemia. *J Clin Lipidol.* 2011;5:S1-8.

2. Safarova MS, Kullo IJ. My approach to the patient with familial hypercholesterolemia. *Mayo Clin Proc.* 2016;91:770-786.

3. Perak AM, Ning H, de Ferranti SD, Gooding HC, Wilkins JT, Lloyd-Jones DM. Long-Term Risk of Atherosclerotic Cardiovascular Disease in US Adults With the Familial Hypercholesterolemia Phenotype. *Circulation.* 2016;134:9-19.

4. Khera AV, Won HH, Peloso GM, et al. Diagnostic Yield and Clinical Utility of Sequencing Familial Hypercholesterolemia Genes in Patients With Severe Hypercholesterolemia. *J Am Coll Cardiol.* 2016;67:2578-2589.

5. Mortality in treated heterozygous familial hypercholesterolaemia: implications for clinical management. Scientific Steering Committee on behalf of the Simon Broome Register Group. *Atherosclerosis.* 1999;142:105-112.

6. Alonso R, Mata P, Zambon D, Mata N, Fuentes-Jimenez F. Early diagnosis and treatment of familial hypercholesterolemia: improving patient outcomes. *Expert Rev Cardiovasc Ther.* 2013;11:327-342.

7. Thompson GR, Seed M, Naoumova RP, et al. Improved cardiovascular outcomes following temporal advances in lipid-lowering therapy in a genetically-characterised cohort of familial hypercholesterolaemia homozygotes. *Atherosclerosis.* 2015;243:328-333.

8. Gidding SS, Ann Champagne M, de Ferranti SD, et al. The agenda for familial hypercholesterolemia: A scientific statement from the American Heart Association. *Circulation.* 2015;132:2167-2192.

9. Safarova MS, Liu H, Kullo IJ. Rapid identification of familial hypercholesterolemia from electronic health records: The SEARCH study. *J Clin Lipidol.* 2016;10:1230-1239.

10. Benn M, Watts GF, Tybjaerg-Hansen A, Nordestgaard BG. Mutations causative of familial hypercholesterolaemia: screening of 98 098 individuals from the Copenhagen General Population Study estimated a prevalence of 1 in 217. *Eur Heart J.* 2016;37:1384-1394.

11. Damgaard D, Larsen ML, Nissen PH, et al. The relationship of molecular genetic to clinical diagnosis of familial hypercholesterolemia in a Danish population. *Atherosclerosis.* 2005;180:155-160.

12. Versmissen J, Oosterveer DM, Yazdanpanah M, et al. Efficacy of statins in familial hypercholesterolaemia: a long term cohort study. *BMJ.* 2008;337:a2423.

13. Nordestgaard BG, Chapman MJ, Humphries SE, et al. Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: guidance for clinicians to prevent coronary heart disease: consensus statement of the European Atherosclerosis Society. *Eur Heart J.* 2013;34:3478-3490a.

14. Benn M, Watts GF, Tybjaerg-Hansen A, Nordestgaard BG. Familial hypercholesterolemia in the danish general population: prevalence, coronary artery disease, and cholesterol-lowering medication. *J Clin Endocrinol Metab.* 2012;97:3956-3964.

15. Kirby JC, Speltz P, Rasmussen LV, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc.* 2016.