Phenotyping Algorithm for Autoimmune Disease (AID) V4

Updates (Compared to V3):

- Updated serology tests LOINC codes

Updates (Compared to V2):

- Combined coding dictionary for cases and controls files
- Added data dictionary for co-variables extraction

I. Background and Significance

The goal of the AID phenotype definition is to identify an individual as case or control using structured electronic health record (EHR) elements. The autoimmune disease case phenotype spans 6 medical specialties, 9 groups of disease, and 51 diseases (46 non-overlapping diseases) (Table 1 and 2). Groups of diseases are loosely based on organ system. The 5 diseases causing the discrepancy between the 51 diseases and 46 non-overlapping diseases are due to the diseases belonging to parent diseases that are included in the definition (i.e., Giant Cell Arteritis and Cerebral Arteritis are children concepts to the parent concept Arteritis, Optic Papillitis is a child concept to Optic Neuritis, Ocular Cicatricial Pemphigoid is a child concept to parent Pemphigoid, Dermatomyositis is a child concept of Polymyositis). In addition to excluding autoimmune diseases, the autoimmune disease control phenotype excludes auto-inflammatory diseases and the presence of 10 sets of antibody tests (Table 3).

Each disease is defined by a collection of ICD9CM, ICD10CM, and/or SNOMED codes. A total of 1528 diagnosis codes were used to determine case status, and 49960 diagnosis codes and measurement codes were used for controls.

II. Development

- Autoimmune disease case and control phenotype was developed using structured EHR data using domain knowledge and standards.
- Validation was performed against expert opinion, including local experts from Rheumatology, Nephrology, Dermatology, Neurology, and Gastroenterology.

III. Definition

A flow diagram depicting how participants in a cohort are sorted into case and control cohort is in Figure 1. Pseudocode is described in Figure 2.

III.i. Autoimmune Disease Case

The presence of at least one autoimmune disease qualifies the patient to be a case.

A patient meeting the following condition (*Case Cohort: Condition A*) is defined as having at least one autoimmune disease:

> *Case Cohort: Condition A.* Any 3 diagnosis codes on distinct days within the same disease with the first and last diagnoses at least 7 days apart. Codes defining each disease are listed in AIDalgorithm_V1_coding_cases.csv and are grouped by disease name. For patients with diagnosis codes for type 1 diabetes mellitus (T1DM), an additional exclusion criterion of no type 2 diabetes mellitus (T2DM) codes is applied. Codes defining T2DM are listed in file AIDalgorithm_V1_coding_cases.csv under variable name, "Type 2 Diabetes Mellitus (T2DM)". Patients meeting the case eligibility criteria will be tagged as belonging to group(s) and disease(s) based on the presence of diagnosis code(s).

III.ii. Autoimmune Disease Control

The absence of any selected auto-inflammatory condition, autoimmune disease, and positive serology test qualifies a patient to be a control.

A patient not having any auto-inflammatory condition, autoimmune disease, and positive serology test must meet both of the following conditions (*Control Cohort: Condition A* and *Control Cohort: Condition B*):

> *Control Cohort: Condition A.* No autoimmune and auto-inflammatory diagnosis codes. Codes are listed in file AIDalgorithm_V1_coding_control.csv under variable name, "Auto-inflammatory + Autoimmune (SNOMED)".
>
> *Control Cohort: Condition B.* No instances of any positive serologies as defined by institutional and assay recommendations from a list of serologies (Table 3). Codes are listed in file AIDalgorithm_V1_coding_control.csv under variable names starting with "Serology".

Note: Opening *.csv in Excel strips the ".0" in ICD9 codes.

IV. Autoimmune Disease Implementation

The coding dictionary for the algorithm variables (i.e., autoimmune diseases, auto-inflammatory diseases, serologies) is index by condition in Table 4. Table 4 refers to 2 csv files: AIDalgorithm_V1_coding_cases.csv and AIDalgorithm_V1_coding_controls.csv.

The coding dictionary is also located on OHDSI ATLAS (http://www.ohdsi.org/web/atlas/#/cohortdefinitions) under Cohorts ID 102998 for cases and Cohorts ID 919472 for controls.

IV.I Coding dictionary CSV files

Each disease, or subphenotype, is defined by one or more codes belonging to one or more terminology vocabulary. For example, the codes for the subphenotype, "Vasculitis: Takayasu's disease", are located in AIDalgorithm_V1_coding_cases.csv and are the following:

- ICD10CM M31.4 Aortic arch syndrome [Takayasu]
- SNOMED 359789008 Takayasu's disease
- ICD9CM 446.7 Takayasu's disease

IV.II Using OHDSI ATLAS

1. Go to http://www.ohdsi.org/web/atlas/#/cohortdefinitions
2. Click on "Cohorts" on the left panel
3. Search for case or controls by typing 102998 or 919472, respectively, into "Filter:"
4. Click on the cohort in the search results
5. Select the "Concept Set" Tab located on the right panel under title
6. To export concept sets, click "Export All Concept Sets to CSV"

V. Covariates

Demographic information: eMERGE_ID, Age, Sex, Race, Ethnicity, Case_Control_Unknown_Status.
AID subphenotype: eMERGE_ID, subPhenotype.
AID algorithm relevant diagnosis: eMERGE_ID, Age_diagnosis, diagnosis_code, code_type.
Serology test (Table 3): eMERGE_ID, serology_test_name, age_at_serology_test, serology_test_result.
Complement (C3 and C4) test: eMERGE_ID, complement_test_name, age_at_complement_test, complement_test_result
Quantitative immunoglobulins (IgA, IgG, and IgM) test: eMERGE_ID, immunoglobulin_test_name, age_at_immunoglobulin_test, immunoglobulin_test_result

VI. Accompanying Documents

The documents for the Autoimmune Disease Phenotype are as follows:

1. AutoimmuneDiseaseAlgorithm_V4.pdf – Main document containing the AID algorithm definition.
2. Validation Instructions.docx – Validation instructions for secondary site.
3. AIDalgorithm_coding_V3.txt – Lists of codes used for Autoimmune Disease Case and Control Phenotype (corresponds to Table 4). This includes mapped OMOP vocabulary (conceptId).
4. AIDalgorithm_dd_V1.xls – AID co-variable data dictionary.

Figures and Tables

Table 1. The autoimmune disease case phenotype spans 9 groups of disease, and 51 diseases (46 non-overlapping diseases) or subphenotypes

D	Grouping Name	Disease Name
1	Arthritis	Ankylosing spondylitis
2	Arthritis	Bechet's disease
3	Arthritis	Palindromic rheumatism
4	Arthritis	Polymyalgia rheumatica
5	Arthritis	Psoriatic arthritis
6	Arthritis	Reiter's syndrome
7	Arthritis	Rheumatoid arthritis (RA)
8	Connective	Lupus erythematosus
9	Connective	Mixed Connective Tissue Disease (MCTD)
10	Connective	Sarcoidosis
11	Connective	Scleroderma
12	Connective	Sjogren's syndrome
13	Endocrine	Graves' Disease
14	Endocrine	Hashimoto's thyroiditis
15	Endocrine	T1D
16	GI	Autoimmune hepatitis
17	GI	Celiac Disease
18	GI	Crohn's disease
19	GI	Primary biliary cholangitis (PBC)
20	GI	Ulcerative colitis (UC)
21	Heme	Antiphospholipid syndrome (APS)
22	Heme	Autoimmune hemolytic anemia (AIHA)
23	Heme	Autoimmune neutropenia
24	Heme	Evans syndrome
25	Heme	Thrombocytopenic purpura (TTP)
26	Muscle	Dermatomyositis*[1]
27	Muscle	Polymyositis*[1]
28	Muscle	Inflammatory and immune myopathies
29	Neuro	Guillain-Barre Syndrome
30	Neuro	Lambert-Eaton syndrome
31	Neuro	Multiple sclerosis
32	Neuro	Myasthenia gravis
33	Neuro	Myelitis transversa
34	Neuro	Optic neuritis*[2]
35	Neuro	Optic Papillitis*[2]
36	Neuro	Schilder's disease
37	Skin	Alopecia areata
38	Skin	Dermatitis herpetiformis
39	Skin	Pemphigoid*[3]
40	Skin	Ocular cicatricial pemphigoid*[3]
41	Skin	Pemphigus
42	Skin	Psoriasis
43	Skin	Pyoderma
44	Skin	Raynaud
45	Skin	Vitiligo
46	Vasculitis	Arteritis*[4]
47	Vasculitis	Cerebral Arteritis*[4]
48	Vasculitis	Giant Cell Arteritis*[4]
49	Vasculitis	Goodpasture's syndrome
50	Vasculitis	Granulomatosis
51	Vasculitis	Takayasu's disease

* Denotes diseases that overlap with other diseases and are combined into one sub-phenotype.

Table 2. Medical Specialty and Groupings

Medical Specialty	Name of Group of Diseases
Rheumatology	Arthritis
	Connective Tissue
	Muscular (Muscle)
	Vasculitis
Endocrinology	Endocrine
Gastroenterology	Gastrointestinal (GI)
Hematology	Hematologic (Heme)
Neurology	Neurological (Neuro)
Dermatology	Skin

Table 3. Serology Antibody Tests

#	Antibody Tests:
1	Anti Nuclear Antibody (ANA)
2	Anti-Cytoplasmic Neutrophil Antibodies (ANCA)*
3	Anti-DNA Antibody (dsDNA)
4	Cyclic Citrullinated Peptide Antibody (CCP)
5	Rheumatoid Factor (RF)
6	Beta 2 Glycoprotein I Antibody (B2 Glycoprotein 1, B2Glyc)
7	RNA Polymerase 3 Antibody (RNA PIII, RNAP3)
8	Anti-Cardiolipin Antibodies (Cardiolipin)*
9	Centromere Antibody IgG (Centromere)
10	Extractable Nuclear Antibodies (ENA)*

Anti-Cytoplasmic Neutrophil Antibodies include:
Serum Anti Neutrophil Cytoplasmic Antibody C Test (cANCA)
Serum Anti Neutrophil Cytoplasmic Antibody P Test (pANCA)
Anti-Cardiolipin Antibodies include:
Anti-Cardiolipin IgG Antibody Measurement (Cardiolipin_IgG)
Anti-Cardiolipin IgM Antibody Measurement (Cardiolipin_IgM)
Extractable Nuclear Antibodies include:
Jo 1 Autoantibody (J01)
Ribonucleoprotein (U1RNP) Autoantibody (antiRNP)
Sclerdoma Autoantibody(Scl 70)
Smith Autoantibodies (Sm)
Sjogren's Syndrome A (SS-A) Antibodies IgG (SSA)
Sjogren's Syndrome B (SS-B) Antibodies IgG (SSB)

Table 4. An index to the code dictionary sorted by each condition and description of data

ID	*Condition*, Description of Data	**Filename**, Variable Names within file	
A1	*Case Cohort: Condition A,* Autoimmune diseases	**AIDalgorithm_V1_coding_cases.csv,**	
		Arthritis: Ankylosing spondylitis	Muscle: Dermatomyositis
		Arthritis: Behcet's disease	Muscle: Inflammatory and immune myopathies
		Arthritis: Palindromic rheumatism	
		Arthritis: Polymyalgia rheumatica	Muscle: Polymyositis
		Arthritis: Psoriatic arthritis	Neuro: Guillain-Barre Syndrome
		Arthritis: Reiter's syndrome	Neuro: Lambert-Eaton syndrome
		Arthritis: Rheumatoid arthritis (RA)	Neuro: Multiple sclerosis
		Connective: Lupus erythematosus	Neuro: Myasthenia gravis
		Connective: Mixed Connective Tissue Disease (MCTD)	Neuro: Myelitis transversa
			Neuro: Optic neuritis
		Connective: Sarcoidosis	Neuro: Optic Papillitis
		Connective: Scleroderma	Neuro: Schilder's disease
		Connective: Sjogren's syndrome	Skin: Alopecia areata
		Endocrine: Graves' Disease	Skin: Dermatitis herpetiformis
		Endocrine: Hashimoto's thyroiditis	Skin: Ocular cicatricial pemphigoid
		Endocrine: T1D	Skin: Pemphigoid
		GI: Autoimmune hepatitis	Skin: Pemphigus
		GI: Celiac Disease	Skin: Psoriasis
		GI: Crohn's disease	Skin: Pyoderma
		GI: Primary biliary cholangitis (PBC)	Skin: Raynaud
		GI: Ulcerative colitis (UC)	Skin: Vitiligo
		Heme: Antiphospholipid syndrome (APS)	Vasculitis: Arteritis
		Heme: Autoimmune hemolytic anemia (AIHA)	Vasculitis: Cerebral Arteritis
			Vasculitis: Giant Cell Arteritis
		Heme: Autoimmune neutropenia	Vasculitis: Goodpasture's syndrome
		Heme: Evans syndrome	Vasculitis: Granulomatosis
		Heme: Thrombocytopenic purpura (TTP)	Vasculitis: Takayasu's disease
A2	*Case Cohort: Condition A,* T2DM	**AIDalgorithm_V1_coding_cases.csv,**	
		Type 2 Diabetes Mellitus (T2DM)	
A3	*Control Cohort: Condition A,* Autoimmune diseases and Auto-inflammatory diseases	**AIDalgorithm_V1_coding_controls.csv,**	
		Auto-inflammatory + Autoimmune	
B1	*Control Cohort: Condition B,* Serologies	**AIDalgorithm_V1_coding_controls.csv,**	
		Serology: ANA	Serology: B2 Glycoprotein 1
		Serology: ANCA	Serology: RNA PIII
		Serology: dsDNA	Serology: Cardiolipin
		Serology: CCP	Serology: Centromere IgG
		Serology: RF	Serology: ENA

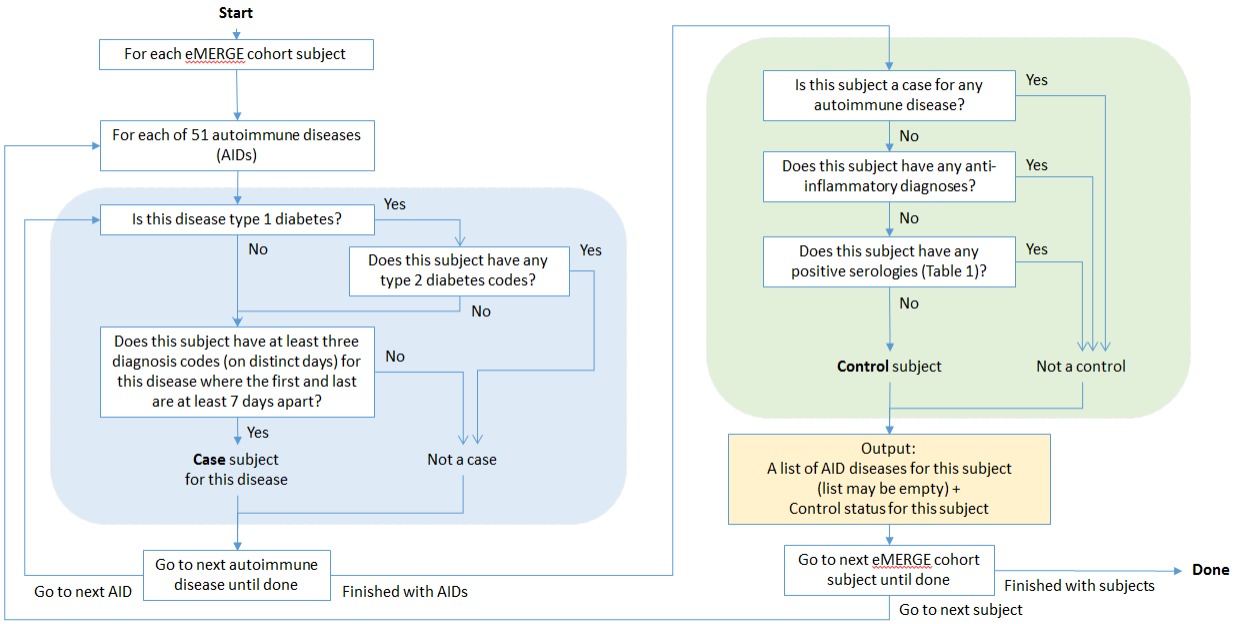

Figure 1. Flow diagram of phenotype logic for case and control definition

FOR each subject **IN** eMERGE cohort

FOR each disease **IN** the 51 autoimmune diseases from Table 1**:** D=1 Ankylosing Spondylitis **TO** D=51 Takayasu's disease

IF (this disease **IS** D=15 Type 1 diabetes (T1D))
AND (this subject has one or more type 2 diabetes codes from [1]) **THEN**

Subject is not a case for this disease D=15 Type 1 diabetes (T1D)

ELSEIF this subject has at least 3 diagnoses codes from [2] on distinct days for this disease where the first and last diagnoses are at least 7 days apart **THEN**

Subject is a case for this disease D

ELSE
Subject is not a case for this disease D

ENDIF

ENDFOR

IF this subject is a case for one or more autoimmune diseases **THEN**
EXIT FROM "FOR each eMERGE cohort subject"

//The following code is needed only if the subject is not a case
IF (this subject does NOT have one or more autoimmune diseases and one or more anti-inflammatory diagnoses from [3]) **AND** (this subject does NOT have one or more positive serologies using codes from [4] and using institution specific codes for serologies listed in Table 3) **THEN**

Subject is a control

ELSE
Subject is not a control

ENDIF

ENDFOR

[1] AIDalgorithm_V1_coding_cases.csv Variable Name: "Type 2 Diabetes Mellitus (T2DM)"
[2] AIDalgorithm_V1_coding_cases.csv Variable Name: D. For example, if D=51 Vasculitis: Takayasu's disease are listed under Variable Name: "Vasculitis: Takayasu's disease" and are the following:

- ICD10CM M31.4 Aortic arch syndrome [Takayasu]
- SNOMED 359789008 Takayasu's disease
- ICD9CM 446.7 Takayasu's disease

[3] AIDalgorithm_V1_coding_control.csv Variable Name: "Auto-inflammatory + Autoimmune (SNOMED)"
[4] AIDalgorithm_V1_coding_control.csv Variable Name: "Serology: ANA,Serology: ANCA", "Serology: dsDNA", "Serology: CCP", "Serology: RF", "Serology: B2 Glycoprotein 1", "Serology: RNA PIII", "Serology: Cardiolipin", "Serology: Centromere IgG", "Serology: ENA", "Serology: ICA + IA-2 + ICA 512 (Islet cell)", "Serology: GAD (Glutamate decarboxylase)", "Serology: Tissue Transglutaminase IgA", "Serology: Endomysium IgG + IgA", "Serology: Gliadin IgA + IgG", "Serology: ASCA IgG IgA (ASCA in MED)", or "Serology: OmpC+CBir1"

Figure 2. Pseudocode of phenotype logic for case and control definition