Type 2 Diabetes (T2DM) Phenotype Algorithm

Implementation Document

Mass General Brigham eMERGE

|  |  |
| --- | --- |
| Document Version: | 1.0.00-001 |
| T2DM\_filterpositive\_2/12/2021 | 12Feb21 |

Table of Contents

Contents

[Section 1. Overview 3](#_Toc64967825)

[Section 2. Feature Dictionary 5](#_Toc64967826)

[2.1. Create a summary table 5](#_Toc64967827)

[2.1.1. Table with feature counts by subject 5](#_Toc64967828)

[2.1.2. Add flag the subjects who meet filter and data flow criteria 5](#_Toc64967829)

[2.3 References: 6](#_Toc64967830)

[2.4 Files for PheKB 6](#_Toc64967831)

[Section 3. Appendices 7](#_Toc64967832)

[2.5 Feature Definitions 7](#_Toc64967833)

[2.5.1 COD\_DX\_T2DM 7](#_Toc64967834)

[2.5.2 COD\_DX\_T1DM 7](#_Toc64967835)

[2.5.3 COD\_MED\_diabetes 7](#_Toc64967836)

[2.6 T2DM Feature Distribution Data Dictionary 9](#_Toc64967837)

# Overview

In 2014, this type 2 DM (T2DM) algorithm was created using a logistic regression classifier.  The overall data set of 5,000 Mass General Brigham (MGB) Biobank genotyped subjects was filtered to identify subjects with at least one ICD9/ICD-10 code for Type 2 diabetes mellitus (T2DM) (Section 2.5.1). Then this set was screened for a data floor defined as the presence of at least three ICD9/ICD10 codes (for any ICD code) and at least one medical note in the electronic health record (EHR) for a data floor threshold to ensure enough EHR data was available for algorithm development. Chart reviews were performed 284 charts selected from this set to establish a gold standard training set. Domain experts performed the chart reviews to classify confirmed T2DM (definite or probable vs. not enough evidence for a diagnosis of T2DM, or unknown).

The training set results were used to train the regression model. A set of EHR attributes (called features) associated with the disease were defined using Surrogate-Assisted Feature Extraction (SAFE) 1. Potential predictors were “crowd-sourced” from 5 publicly available knowledge sources: Wikipedia, Medscape, Merck Manuals Professional Edition, Mayo Clinic Diseases and Conditions, and MedlinePlus Medical Encyclopedia. These 5 sources yield UMLS concepts as candidate features. A penalized logistic regression model (LASSO) was trained to classify Type 2 DM in training set (2,3). The regression model identified relative weights (beta coefficients) of features significantly associated with T2DM. The final model included coded variables, and natural language processing (NLP) derived concepts. To facilitate portability of the algorithm, coded clinical variables were created to reflect the NLP concepts (eg. mentions of T2DM were classified as a count of ICD9/ICD10 codes for T2DM) selected by elastic net as predictive of T2DM status. A second model with coded variables only was trained where elastic net was used to select variables predictive of T2DM status in the training set. This algorithm produced relative weights (beta coefficients) of the features which were predictive of T2DM status. Threshold cutoffs were selected to classify T2DM status holding the specificity at 95%, where those above the cutoff were assigned case status, those below were assigned non-case status.

In 2021, the Mass General Brigham (MGB) phenotype team modified the 2014 algorithm to add ICD10 codes for T2DM and T1DM, and to add new diabetes medications (Section 2.5.3). The MGB 36,400+ Biobank genotyped subject dataset was filtered to identify subjects with at least one ICD9/ICD-10 code for T2DM (Section 2.5.1), who met the data floor, of at least three ICD9/ICD10 codes and at least one medical note in the EHR. Chart reviews were performed on a random set of 208 charts selected from this set to establish a gold standard training set. Domain experts (two endocrinology fellows) performed chart reviews to classify confirmed T2DM (definite or probable vs. not enough evidence for a diagnosis of T2DM) for the validation set. A second validation set from University of Alabama at Birmingham (UAB) was created using chart reviews of 198 charts that met the screening and data floor criteria. In addition, chart reviews were performed on 200 charts that did not meet the screening criteria (e.g. did not have any ICD9/10 codes for T2DM) at UAB.

This document describes how to implement the computed T2DM phenotype algorithm created at MGB. This document describes each feature used by the T2DM phenotype algorithm created at MGB. The model predicts current and past history of disease with no temporal constraints. As T2DM is not an uncommon disease, we prioritized optimizing specificity (i.e. reducing the false positive rate) over sensitivity (i.e. detecting all true positives). Once the model was created, a threshold value based on specificity>95% and positive predictive value (PPV) was selected to identify cases, with those above the cutoff identified as cases.

MGB assessed the PPV of the modified 2014 algorithm (ICD10 codes added, new DM medications added) as the rate of true positives in those classified as cases by the algorithm in Validation Set 1 (MGB), and Validation Set 2 (UAB). PPV was 91% in the MGB 2014 training set, 90% in the 2021 MGB validation set, and 94% in the UAB validation set. This document describes implementation of the filter, data floor, and selection of variables to provide a feature dictionary (Section 2.1) to the MGB phenotyping team for defining T2DM cases and controls in the eMERGE III genotyped set.

**Algorithm Training and Validation Results**

2014 Model Training

2014 Gold standard training set (N=284 charts reviewed): Yes 91, Probable 10, No 180, Unknown 3

Prevalence (Y+P) = 0.38

Holding the specificity at >95%, a cutoff was used to define case/non-case status. At a MGB cutoff of 0.593, sensitivity is 76.6%, specificity is 95.4%, positive predictive value is 90.8%, negative predictive value is 86.9%, AUC is 0.952.

|  |  |
| --- | --- |
| Accuracy Parameter  | T2DM  |
| Sensitivity  | 0.776 |
| Specificity | 0.954 |
| Positive Predictive Value (PPV) | 0.908 |
| Negative Predictive Value (NPV) | 0.869 |

2021 Updated Model Validation

MGB Gold standard training set (N=208 charts reviewed): Yes 138, No 59, Unknown 11

Prevalence (Y)= 0.66

Positive predictive value = 90%

|  |  |
| --- | --- |
| Accuracy Parameter  | T2DM  |
| Sensitivity  | 0.6087 |
| Specificity | 0.8714 |
| Positive Predictive Value (PPV) | 0.9032 |
| Negative Predictive Value (NPV) | 0.5304 |

2021 Updated Model Validation

UAB Gold standard training set (N=198 charts reviewed): Yes 184, No 14

Prevalence (Y)= 0.93

PPV = 94%

|  |  |
| --- | --- |
| Accuracy Parameter  | T2DM  |
| Sensitivity  | 0.809783 |
| Specificity | 0.357143 |
| Positive Predictive Value (PPV) | 0.943038 |
| Negative Predictive Value (NPV) | 0.125 |

# Feature Dictionary

The table below lists the features for the T2DM phenotype algorithm. The list of codes for each feature is listed in the Appendices in section 3.

#### List of T2DM Features

|  |  |  |
| --- | --- | --- |
| Feature\_ID  | Beta (weight) | Feature Description  |
| Intercept | 3.184 | Model intercept (beta 0) |
| Patient\_dxenct | -1.512 | Total number of encounters (visits), per subject, with any ICD-9/10 code (any dx code, not limited to T2DM) |
| COD\_DX\_T2DM | 2.475 | Count of coded diagnosis of type 2 DM ICD9/10 codes |
| COD\_DX\_T1DM |  -1.320 | Count of coded diagnosis of type 2 DM ICD9/10 codes |
| COD\_MED\_diabetes | 0.978 | Count of coded prescriptions for diabetes medications |

**\*\*** OMOP mapping for all codes can be found **T2DM\_Feature\_OMOP\_Mapping.csv** file by feature name. **\*\***

## Create a feature distribution table

### Table with feature counts by subject

For each coded feature, count the distinct dates in which a subject has a code from that feature.

**Example**:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SUBJID |  Patient\_dxenct | COD\_ DX\_ T2DM | COD\_DX\_T1 DM | COD\_MED\_diabetes | Meets ICD9/10 filter | Meets data floor |
|  |  |  |  |  |  |  |
| 100 | 256 | 240 | 49 | 84 | Y | Y |
| 200 | 38 | 1 | 1 | 34 | Y | Y |
| 300 | 343 | 319 | 38 | 77 | Y | Y |
| 400 | 279 | 5 | 0 | 0 | Y | Y |
| 500 | 34 | 0 | 0 | 0 | N | Y |
| 600 | 154 | 10 | 2 | 18 | Y | Y |
| 700 | 2 | 1 | 0 | 0 | Y | N |

### Add flag the subjects who meet filter and data flow criteria

Identify which subjects meet the following filter criteria in feature table:

ICD9/10 Filter

* Has at least 1 code from the COD\_DX\_T2DM feature defined in section **2.5.1. COD\_DX\_T2DM**

Data Floor

* Has at least 3 diagnosis codes for **any** ICD9/10 code
	+ With 30 days or more between first and last date of diagnosis code
	+ With at least one diagnosis code after ‘1/1/2005’
* At least one clinical note

## References:

1. Yu S, Liao KP, Shaw SY, Gainer VS, Churchill SE, Szolovits P, Murphy SN, Kohane IS, Cai T. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *J Am Med Inform Assoc.* Sep 2015;22(5):993-1000.]
2. Zou H. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association.* 2014/10/15 2006;101(476):1418-1429.
3. Zhang HH, Lu W. Adaptive Lasso for Cox's proportional hazards model. *Biometrika.* 2007;94(3):691-703.

## Files for PheKB

Create a feature distribution file using the feature distribution data dictionary on PheKB (T2DM\_DD\_FeatureDistribution.csv).

Create a demographics file of subjects using the demographics data dictionary on PheKB (T2DM\_DD\_demographics.csv).

A sample of both the demographics data dictionary and the feature distribution data dictionary can be found in section 2.6 of the appendices.

# Appendices

## Feature Definitions

### COD\_DX\_T2DM (used for filter criteria)

|  |  |  |
| --- | --- | --- |
| Feature Type | Code | Name |
| Diagnosis (ICD9) | 250.X0\* | Diabetes with X manifestations, type II, not stated uncontrolled |
| Diagnosis (ICD9) | 250.X2\* | Diabetes with X manifestations, type II, uncontrolled |
| Diagnosis (ICD10) | E11.X | Type 2 diabetes  |
| Diagnosis (ICD10) | E11.XX | Type 2 diabetes with complications |
| Diagnosis (ICD10) | E11.XXX | Type 2 diabetes with specific complications |
| Diagnosis (ICD10) | E11.XXXX\* | Type 2 diabetes mellitus with ocular complications, right eye, left eye |
| \* X stands for any number from 0 to 9.  |

### COD\_DX\_T1DM

|  |  |  |
| --- | --- | --- |
| Feature Type | Code | Name |
| Diagnosis (ICD9) | 250.X1\* | Diabetes with X manifestations, type 1 [juvenile type], not stated as uncontrolled |
| Diagnosis (ICD9) | 250.X3\* | Diabetes with X manifestations, type 1 [juvenile type], uncontrolled |
| Diagnosis (ICD10) | E10.X | Type 1 diabetes  |
| Diagnosis (ICD10) | E10.XX | Type 1 diabetes mellitus with complications |
| Diagnosis (ICD10) | E10.XXX | Type 1 diabetes mellitus with specific complications |
| Diagnosis (ICD10) | E10.XXXX\* | Type 1 diabetes mellitus with ocular complications, right eye, left eye |
| \*X stands for any number from 0 to 9.  |

### COD\_MED\_diabetes

The following is a listing of medications (name only) from Al’lona Furmanchuk, Northwestern.

For mappings of RXNorm to OMOP concept\_id see **T2DM\_Feature\_OMOP\_Mapping.csv.**

For the list of RXNorm CUIs, regular-expressions or NDC codes see the **DMmeds\_BMJpaper1.docx**.

|  |
| --- |
| Generic name (Brand Name) |
| *Alpha-glucosidase inhibitors:*acarbose (Precose, Glucobay)miglitol (Glyset)Voglibose (Basen) |
| *Dipeptidyl Peptidase IV Inhibitors:*alogliptin (Nesina)Anagliptin (Suiny)linagliptin (Tradjenta)saxagliptin (Onglyza)sitagliptin (Januvia)Teneligliptin (Tenelia)Vildagliptin (Galvus, Zomelis) |
| *Glucagon-like Peptide-1 Agonists:*Lixisenatide (Adlyxin, Lyxumia)Albiglutide (Tanzeum, Eperzan)Dulaglutide (Trulicity) |
| *Sodium glucose cotransporter (SGLT) 2 inhibitors:*dapagliflozin (Forxiga, Farxiga)canagliflozin (Invokana, Sulisent)empagliflozin (Jardiance) |
| *Sulfonylureas:*Acetohexamide (Dymelor) Dimelorglimepiride (Amaryl)gliclazide (Uni Diamicron)glipizide (Glucotrol, Minidiab, Minodiab, Glibenese, Glucotrol XL, Glipizide XL)glyburide or glibenclamide (DiaBeta, Glynase, Micronase, Glycron)chlorpropamide (Diabinese, Apo-Chlorpropamide, Glucamide, Novo-Propamide, Insulase)tolazamide (Tolinase, Glynase PresTab, Tolamide)tolbutamide (Orinase, Tol-Tab, Apo-Tolbutamide, Novo-Butamide)Glyclopyramide (Deamelin-S)Gliquidone (Glurenorm)Glibornuride (Glutril, Glibornurid, Glibornurida, Glibornuride, Glibornuridum)Glymidine sodium (glycodiazine, Gondafon, Glidiazine, Glymidine) |
| *Amylinomimetics:*Pramlintide (Symlin, SymlinPen 120, SymlinPen 60) |
| *Meglitinides:*nateglinide (Starlix)repaglinide (Prandin, NovoNorm) |
| *Insulins:*Insulin aspart (NovoLog)Insulin glulisine (Apidra)Insulin lispro (Humalog)Insulin inhaled (Afrezza)Regular insulin (Humulin R, Novolin R)Intermediate-Acting Insulins:Insulin NPH (Humulin N, Novolin N)Insulin detemir (Levemir)Insulin glargine (Lantus, Lantus SoloStar, Toujeo, Basaglar)Insulin degludec (Tresiba)Insulin aspart protamine/insulin aspart (NovoLog 50/50, NovoLog 70/30)Insulin lispro protamine/insulin lispro (Humalog 50/50, Humalog 75/25)ActrapidHypurinIletinInsulatardInsumanMixtardNovoMixNovoRapidOralin AbasaglarRyzodegV-go |
| *Combinations:*linagliptin-empagliflozin (Glyxambi)sitagliptin-simvastatin (Juvisync, Epistatin, Synvinolin, Zocor)metformin-alogliptin (Kazano)metformin-canagliflozin (Invokamet)metformin-dapagliflozin (Xigduo XR)metformin-empagliflozin (Synjardy)metformin-glipizide (Metaglip)metformin-glyburide (Glucovance)metformin-linagliptin (Jentadueto, Jentadueto XR)metformin-repaglinide (PrandiMet)metformin-saxagliptin (Kombiglyze XR)metformin-sitagliptin (Janumet, Janumet XR)metformin and vildagliptin (Eucreas)glimepiride-pioglitazone (Duetact)alogliptin-pioglitazone (Oseni)glimeperide-rosiglitazone (Avandaryl)Insulin/Lixisenatide |
| *Biguanides:*Metformin (Glucophage, Fortamet, Glumetza, Riomet, Riomet) |
| *Glucagon-like Peptide-1 Agonists:*Exenatide (Byetta, Bydureon)Liraglutide (Victoza, Saxenda) |
| *Thiazolidinediones:*rosiglitazone (Avandia)pioglitazone (Actos)Troglitazone (Noscal, Resulin, Rezulin, Romozin) |
| *Combinations:*metformin-pioglitazone (Actoplus, Actoplus Met, Actoplus Met XR, Competact)metformin-rosiglitazone (Avandamet)INSULIN- Liraglutideertugliflozin- metformin |

## T2DM Feature Distribution Data Dictionary

#### Diabetes Feature Distribution Data Dictionary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **VARNAME** | **VARDESC** | **TYPE** | **REQUIRED** | **VALUE** |
| SUBJID | The eMERGE unique ID of Subject | String | Yes |  |
| Sex | Sex of the participant | String | Yes | C46119=Male; C46110=Female; U=Unknown; NA=Not Assessed; .=Missing |
| Race | Race of the participant | String | Yes | C16352=Black or African American; C41259=American Indian or Alaska Native; C41260=Asian; C41261=White; C41219=Native Hawaiian or other Pacific Islander; C17998=Unknown; C43234=Not Reported |
| Ethnicity | Ethnicity of the participant | String | Yes | C17459=Hispanic or Latino; C41222=Not Hispanic or Latino; C41221=Unknown; .=Missing |
| Birth\_year  | Four-digit year of birth | integer | Yes |  |
| Meets\_ICD9\_10\_Filter | Flag to identify subjects that have at least one code from the **COD\_DX\_T2DM** feature  | String | Yes |  |
| Meets\_DataFloor | Flag to identify subjects that have meet the data floorHas at least 3 diagnosis codes for **any** ICD9/10 codeWith 30 days or more between first and last date of diagnosis codeWith at least one diagnosis code after ‘1/1/2005’At least one clinical note | String | Yes |  |
| Patient\_dxenct | Encounter CountTotal number of encounters (visits), per subject, with a coded ICD9/10 diagnosis (any diagnosis, not limited diagnosis used in this phenotype). | Integer | Yes |  |
| COD\_DX\_T2DM | The T2DM feature count of distinct dates in which a subject has a code from this feature. | Integer | Yes |  |
| COD\_DX\_T1DM | The T1DM feature count of distinct dates in which a subject has a code from this feature. | Integer | Yes |  |
| COD\_MED\_diabetes  | The diabetes medication feature count of distinct dates in which a subject has a code from this feature. | Integer | Yes |  |