Rheumatoid Arthritis Phenotype Algorithm

Testing and Training Document

Harvard eMERGE

Elizabeth Karlson

|  |  |
| --- | --- |
| Document Version: | 1.0.00-001 |
| Harvard\_eMERGE\_RA Algorithm Training and Testing | 11Oct2017 |

# Overview

This document describes the training and testing methods, including chart review guidelines, used to create and validate the rheumatoid arthritis phenotype. It can be used as a reference but is not required to implement the algorithm. To implement the algorithm, please see the Rheumatoid Arthritis Implementation document.

This algorithm was created using a machine-learning logistic regression model.  First, the Biobank dataset was screened for presence of ICD9/ICD-10 codes for rheumatoid arthritis. Chart reviews were performed among the screen positive set (at least one ICD9/ICD10 code) by a clinical expert to establish a gold standard Training Set of 60 cases and 142 non-cases based on 2010 American College of Rheumatology criteria for classification of rheumatoid arthritis1) (Appendix). The chart review results were used to train the regression model.  A set of NLP concept unique identifiers (CUIs) associated with the disease were defined using Automatic Feature Extraction (AFEP) 1. Potential predictors were “crowd-sourced” from freely available online medical knowledge sources such as Medline, Medscape and the Merck Manual by parsing articles for the phenotype using named entity recognition (NER).  Using NER allows identification of CUIs related to medical terms specific to the phenotype (NLP variables).  A domain expert selected codes from structured EHR data (COD variables) that mapped to the NLP concepts (e.g. medications). After applying frequency filters, EHR features, including NLP and COD variables were included in models. We tested models with NLP variables only, COD variables only, and NLP plus COD variables.

A penalized logistic regression model (LASSO) was trained to classify the RA in the gold-standard Training Set (2, 3). The regression model identified the relative weights (beta coefficients) of the features significantly associated with RA. Performance characteristics of the COD variables only model were determined to be satisfactory for testing the model in independent test sets. The overall AUC was 0.95.

At 95% specificity, the algorithm had a sensitivity of 90.9%, PPV of 92.3%, and NPV of 94%.

At 97% specificity, the algorithm at a sensitivity of 86.6%, PPV of 94.6%, and NPV of 92.3%.

**Performance Characteristics**

#### RA Algorithm Performance Characteristics among the Training Set

|  |  |  |
| --- | --- | --- |
| Metric  |  | RA, positive |
| PPV (precision)  | Positive Predictive Value (PPV) | 0.946 |
| TPR (sensitivity)  | Sensitivity or True Positive Rate (TPR)  | 0.866 |
| FPR (1-specificity) | False Positive Rate (FPR) | 0.03 |
| NPV | Negative Predictive Value (NPV) | 0.923 |

**Test Set 1 (Independent Validation at Harvard)**

An independent Test Set 1 was created by chart reviews of a randomly select set of 100 screen positive charts (34 Definite RA, 66 Not RA). A threshold value based on specificity = 97% for the Harvard algorithm in the training set was selected for which to identify cases and non-cases in the test set, with those above the cutoff identified as cases and those below, non-cases.

|  |
| --- |
| Test Set 1: Validation at Harvard |
|  |  | Chart Review |  |
|  |  | T | F |
| Algorithm | T | 30 | 2 |
|  | F | 4 | 64 |

PPV 94%, Sensitivity 88%, Specificity 98%, NPV 94%

The PPV (rate of true positives in those classified as cases by the algorithm) was 94% in Test Set 1.

**Test Set 2 (Secondary Validation at Vanderbilt)**

An independent Test Set 2 was selected at Vanderbilt for secondary validation from chart reviews performed for a prior RA algorithm portability testing project (5). In that project, 290 charts were reviewed for ACR criteria for rheumatoid arthritis (154 Definite RA, 136 Not RA). A threshold value based on specificity = 97% for the Harvard algorithm in the training set was selected for which to identify cases and non-cases in the test set, with those above the cutoff identified as cases and those below, non-cases.

|  |
| --- |
| Test Set 2: Validation at Vanderbilt |
|  |  | Chart Review |  |
|  |  | T | F |
| Algorithm | T | 128 | 11 |
|  | F | 26 | 125 |

PPV 92%, Sensitivity 83%, Specificity 92%, NPV 83%

The PPV (rate of true positives in those classified as cases by the algorithm) was 92% in Test Set 2.

## Reference:

**1.** Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO 3rd, Birnbaum

NS, Burmester GR, Bykerk VP, Cohen MD, Combe B, Costenbader KH, Dougados M, Emery

P, Ferraccioli G, Hazes JM, Hobbs K, Huizinga TW, Kavanaugh A, Kay J, Kvien TK,

Laing T, Mease P, Ménard HA, Moreland LW, Naden RL, Pincus T, Smolen JS,

Stanislawska-Biernat E, Symmons D, Tak PP, Upchurch KS, Vencovský J, Wolfe F,

Hawker G. 2010 Rheumatoid arthritis classification criteria: an American College

of Rheumatology/European League Against Rheumatism collaborative initiative.

Arthritis Rheum. 2010 Sep;62(9):2569-81. doi: 10.1002/art.27584. PubMed PMID:

20872595.

**2.** 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.

**3.** Zou H. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association.* 2014/10/15 2006;101(476):1418-1429.

**4.** Zhang HH, Lu W. Adaptive Lasso for Cox's proportional hazards model. *Biometrika.* 2007;94(3):691-703.

5. Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, Pacheco JA, Boomershine CS, Lasko TA, Xu H, Karlson EW, Perez RG, Gainer VS, Murphy SN, Ruderman EM, Pope RM, Plenge RM, Kho AN, Liao KP, Denny JC. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. J Am Med Inform Assoc. 2012 Jun;19(e1):e162-9. Epub 2012 Feb 28. PubMed PMID: 22374935.

# Appendix: Chart Review Guidelines

## Chart review Guidelines

The secondary validation site (Vanderbilt) applied the RA algorithm to a previously defined set of RA cases and non-cases based on prior chart reviews for the Arnett ACR Classification Criteria for RA. If implementation sites wish to verify the performance of the algorithm in their own dataset, ideally, charts for 50 cases defined by the algorithm should be reviewed using the ACR/EULAR 2010 Classification Criteria for rheumatoid arthritis (RA)1 and 50 controls should be reviewed for absence of RA classification criteria.

**Control chart review criteria:**

1. If no mention of rheumatoid arthritis is in clinic notes as a possible diagnosis - classify as No.
2. Then if no evidence of synovitis (swelling) of at least one joint – classify as No.

**Case chart review criteria: 2010 ACR/EULAR CLASSIFICATION CRITERIA FOR RA**

**Target Population** (Patients who);

1. Have at least 1 joint with definite clinical synovitis (swelling)\*
2. With the synovitis not better explained by another disease†

For patients who meet the target population criteria, assess for joint involvement in MCPs, PIPs, wrists, MTPs (small joints), knees, elbows, ankles, shoulders, hips (large joints). Note ACPA means anti-CCP antibodies.

**Classification criteria** for RA (score-based algorithm: **add score of categories A-D**;

A score of 6/10 is needed for classification of a patient as having definite RA) ‡

|  |  |
| --- | --- |
| 1. **Joint Involvement**§
 | **Points** |
| 1 large joint¶ | 0 |
| 2-10 large joints | 1 |
| 1-3 small joints (with or without involvement of large joints)# | 2 |
| 4-10 small joints (with or without involvement of large joints) | 3 |
| >10 joints (at least 1 small joint)\*\* | 5 |
| 1. **Serology (at least 1 test result is needed for classification) ††**
 | - |
| Negative RF and negative ACPA | 0 |
| Low-positive RF or low-positive ACPA | 2 |
| High-positive RF or high-positive ACPA | 3 |
| 1. **Acute-phase reactants (at least 1 test result is needed for classification) ‡‡**
 | - |
| Normal CRP and normal ESR | 0 |
| Abnormal CRP or abnormal ESR | 1 |
| 1. **Duration of symptoms**§§
 | - |
| < 6 weeks | 0 |
| >6 weeks | 1 |
| \* The criteria are aimed at classification of newly presenting patients. In addition, patients with erosive disease typical of rheumatoid arthritis (RA) with a history compatible with prior fulfillment of the 2010 criteria should be classified as having RA. Patients with longstanding disease, including those whose disease is inactive (with or without treatment) who, based on retrospectively available data, have previously fulfilled the 2010 criteria should be classified as having RA.† Differential diagnoses vary among patients with different presentations, but may include conditions such as systemic lupus erythematosus, psoriatic arthritis, and gout. If it is unclear about the relevant differential diagnoses to consider, an expert rheumatologist should be consulted.‡ Although patients with a score of <6/10 are not classifiable as having RA, their status can be reassessed and the criteria might be fulfilled cumulatively over time.§ Joint involvement refers to any *swollen* or *tender* joint on examination, which may be confirmed by imaging evidence of synovitis. Distal interphalangeal joints, first carpometacarpal joints, and first metatarsophalangeal joints are *excluded from assessment*. Categories of joint distribution are classified according to the location and number of involved joints, with placement into the highest category possible based on the pattern of joint involvement.¶ “Large joints” refers to shoulders, elbows, hips, knees, and ankles.# “Small joints” refers to the metacarpophalangeal joints, proximal interphalangeal joints, second through fifth metatarsophalangeal joints, thumb interphalangeal joints, and wrists.\*\* In this category, at least 1 of the involved joints must be a small joint; the other joints can include any combination of large and additional small joints, as well as other joints not specifically listed elsewhere (e.g., temporomandibular, acromioclavicular, sternoclavicular, etc.).†† Negative refers to IU values that are less than or equal to the upper limit of normal (ULN) for the laboratory and assay; low-positive refers to IU values that are higher than the ULN but >3 times the ULN for the laboratory and assay; high-positive refers to IU values that are >3 times the ULN for the laboratory and assay. Where rheumatoid factor (RF) information is only available as positive or negative, a positive result should be scored as low-positive for RF. ACPA = anti\_citrullinated protein antibody.‡‡ Normal/abnormal is determined by local laboratory standards. CRP = C-reactive protein; ESR = erythrocyte sedimentation rate.§§ Duration of symptoms refers to patient self-report of the duration of signs or symptoms of synovitis (e.g., pain, swelling, tenderness) of joints that are clinically involved at the time of assessment, regardless of treatment status. |