HHS Public Access

Author manuscript

*Pharmacoepidemiol Drug Saf*. Author manuscript; available in PMC 2024 May 01.

Published in final edited form as:

*Pharmacoepidemiol Drug Saf*. 2023 May ; 32(5): 586–591. doi:10.1002/pds.5597.

# Validation of Serostatus of Rheumatoid Arthritis Using ICD-10 Codes in Administrative Claims Data

**Hemin Lee**[1], **Jeffrey A. Sparks**[2], **Su Been Lee**[1], **Kazuki Yoshida**[2], **Joan E. Landon**[1], **Seoyoung C. Kim**[1,2]

[1]Division of Pharmacoepidemiology and Pharmacoeconomics; Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

[2]Division of Rheumatology, Inflammation, and Immunity; Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

## Abstract

**Purpose:** To determine the accuracy of International Classification of Diseases- Tenth Revision (ICD-10) diagnosis codes for rheumatoid arthritis (RA) serostatus using a U.S. claims database (Optum Clinformatics Data Mart, Optum) and to compare the results to a previous validation study performed in IBM Marketscan Research Database (sensitivity 73%, positive predictive value, PPV, 84%).

**Methods:** In Optum (01/01/2016-03/31/2020) linked with laboratory results, we selected RA patients based on ≥2 ICD-10 diagnosis codes for RA (M05 or M06) and at least one dispensing of RA treatments. We included individuals with at least one laboratory result for rheumatoid factor (RF) or anti-cyclic citrullinated peptide (CCP) performed 365 days prior to and including the cohort entry date. An individual was "seropositive" if at least one of the 2 diagnosis codes used to define RA status was M05. "Seronegative" patients were required to have only M06. Secondary analyses were performed using subsets of M05 and M06 diagnosis codes. We calculated the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and kappa of M05 and M06 against the prespecified reference standard laboratory data.

**Results:** We identified 14,490 adult RA patients who had at least 1 RF or anti-CCP result. The number of patients identified for each reference standard definition ranged from 3,315 (reference standard definition: high + anti-CCP) to 13,636 (any +RF). PPV for seropositive RA, M05, was 77.1%. The PPV of M06 for seronegative RA was 61.6%. When we applied more restricted definitions of M05 and M06, the PPV for seropositive RA increased to 79.2%. The PPV for seronegative RA also notably increased to 89.5%.

**Corresponding author**: Seoyoung C. Kim, MD, ScD, MSCE, 1620 Tremont Street, Suite 3030, Boston MA 02120, USA, Phone: 1-617-278-0930, Fax: 1-617-232-8602, drsykim@gmail.com.

**Author Contributions:** H Lee, J Sparks, SB Lee, K Yoshida, JE Landon, SC Kim: Responsible for the work described in this paper. All authors were involved in the conception, design, and planning of the study. H Lee, SB Lee, JE Landon: Involved in the analysis of data. JA Sparks, K Yoshida, SC Kim: Involved in the interpretation of results, and substantially contributed to the drafting of the manuscript.

**Prior presentations**: Preliminary results from this work were presented at the virtual International Conference on Pharmacoepidemiology and Therapeutic Risk Management (ICPE) All Access meeting, September 2021.

**Conclusion:** ICD-10 codes (M05 and M06) can help identify RA serostatus in claims data, but their limitations should be acknowledged. The PPVs for seropositive and seronegative RA found in the Optum database were lower than those found in MarketScan, perhaps related to database variability or differing patient characteristics and clinical practice. When more restricted definitions of M05 and M06 were used, the PPVs for seropositive and seronegative RA improved to 79.2% and 89.5%, respectively.

## Plain language summary

Rheumatoid arthritis (RA) can be classified into seropositive and seronegative RA. International Classification of Diseases- Tenth Revision (ICD-10) diagnosis codes for RA distinguish these two serotypes, and a recent study performed in IBM MarketScan administrative claims database validated the performance of the diagnosis codes (sensitivity 73%, positive predictive value, PPV, 84%). In this study, we reevaluated the performance of the same codes using the Optum Clinformatics Data Mart database. After selecting a cohort of RA patients with laboratory results for rheumatoid factor (RF) or anti-cyclic citrullinated peptide (CCP), we calculated the sensitivity, specificity, PPV, and negative predictive value (NPV) of M05 and M06 against the prespecified reference standard laboratory data. The ICD-10 codes had good diagnostic accuracy for seropositive RA (sensitivity 67% and PPV 77% for the reference standard definition of any positive RF or anti-CCP). The performance of seronegative RA was similar. When more restrictive M05 and M06 codes were used, the PPV for seropositive and seronegative RA both increased (77.1 to 79.2% for seropositive RA, 61.6 to 89.5% for seronegative RA).

### Keywords

Rheumatoid arthritis; ICD-10; administrative claims; validation

## INTRODUCTION

As of October 1, 2015, the coding system of medical data in the United States has been updated from the International Classification of Diseases, 9$^{th}$ Revision, Clinical Modification (ICD-9-CM) to ICD-10-CM. Compared to 14,000 codes covered in the ICD-9 system, the ICD-10 system contains approximately 69,000 codes, encompassing detailed descriptions of disease locations, severity, and subtypes (1). In prior administrative claims data studies investigating rheumatoid arthritis (RA), cohorts were often defined based on a combination of at least two or more ICD diagnosis codes for RA in addition to a dispensing of disease-modifying anti-rheumatic drugs (DMARDs) (2). Under the ICD-9 system, 714.x was the only diagnosis code available to identify RA. However, after transitioning to ICD-10, it is now possible to make a distinction between RA with rheumatoid factor (M05) and without rheumatoid factor (M06). Seropositive RA, defined by the positive rheumatoid factor (RF) or anti-citrullinated peptide (anti-CCP) level, is associated with etiologies and genetic background different from seronegative RA and is often associated with worse disease course and prognosis (3) (4). If the RA serostatus can be determined through administrative claims data, one could provide a more comprehensive characterization of RA cohorts.

A recent work by Curtis and colleagues (5) examined the validity of ICD-10 diagnosis codes for serostatus in RA patients using the Rheumatology Informatics System for Effectiveness (RISE) electronic health record (EHR)-based registry and MarketScan database. In their work, the sensitivity for M05 as a proxy for seropositive RA was 73-76% and the positive predictive value (PPV) was 82-84%. However, these ICD-10 codes' performance has not been validated in other databases. Therefore, we aimed to conduct an external validation of the previous study (5) and examined the performance of ICD-10 diagnosis codes to classify RA serostatus in another U.S. commercial insurance claims database. In addition, we also performed secondary analyses using a more restrictive set of codes within M05 and M06.

## METHODS

### Study cohort:

Using Optum Clinformatics Data Mart database (01/01/2016-03/31/2020) linked with laboratory data, we created a cohort of RA patients based on 2 RA ICD-10 diagnosis codes of M05 or M06, separated by 7 to 365 days. The patients were additionally required to have at least one dispensing of methotrexate, other conventional disease-modifying antirheumatic drugs (DMARDs), targeted synthetic, or biologic DMARDs (2) (Supplementary table 1). The index date of cohort entry for an individual was the date when a patient fulfilled two RA diagnosis codes and a drug dispensing. We included patients at least 18 years old on the index date and who had continuous enrollment in the database for more than 365 days before the index date. Additionally, patients were required to have at least one laboratory test done for RF and/or anti-CCP with interpretable results during the 365-day baseline period prior to the index date.

### ICD-10 codes for serostatus:

RA serostatus was assessed based on the ICD-10 diagnosis codes for RA. M05 was a proxy for seropositive RA, whereas M06 was a proxy for seronegative RA. Because our cohort included patients with at least 2 RA codes, a patient was defined as seropositive if both diagnosis codes used for cohort entry were M05 or at least one of the two diagnosis codes was M05. A patient was classified as seronegative if both diagnosis codes used for cohort entry were M06. As a secondary analysis, we incorporated a more restrictive definition of M05 (Second RA diagnosis as M05.7x, M05.8x, M05.9x) and M06 (first and second RA diagnosis codes as M06.0x).

### Reference standard definitions:

We used laboratory test results (RF and/or anti-CCP) to validate the performance of ICD-10 codes for the serostatus. RF positivity was defined as an RF value greater than upper limit normal (ULN) or a dichotomous value of 'yes'. Additionally, among patients with numerical values, if a patient had an RF value greater than or equal to 3 times ULN at any time, the patient were flagged as a 'high positive RF'. Same criteria were used to define anti-CCP positive and high positive anti-CCP. In total, we assessed five reference standard definitions for seropositive RA: 1) any positive RF or anti-CCP, 2) any positive RF, 3) high positive RF, 4) any positive anti-CCP and 5) high positive anti-CCP. The reference standard definitions for M06 were the opposite of definitions for M05. Any positive RF or anti-CCP value was

the primary reference standard definition for seropositive RA. Similarly, not positive RF or anti-CCP was the primary definition for seronegative RA.

We calculated the sensitivity, specificity, PPV, negative predictive value (NPV), and kappa of M05 and M06 against the prespecified reference standard definitions. We analyzed descriptive statistics to characterize the RA cohorts based on serostatus ICD-10 diagnosis codes or laboratory results. Personal identifiers were removed from the data to protect patient confidentiality, and therefore, the requirement for patients' informed consent was waived. This study was approved by the Institutional Review Board of the Brigham and Women's Hospital, and data analysis was performed with SAS 9.4.

## RESULTS

During the study period, we identified a total of 132,343 RA patients with DMARD dispensing between 1/1/2016 – 3/31/2020 (Figure 1). After excluding patients aged <18 years and those with less than 365 days of continuous enrollment, 70,992 patients were included: the mean age was 62.9 and 75.9% of the population was female (Supplementary Table 2). Among these patients, 42% of patients (n=29,581) had at least one laboratory test for RF and/or anti-CCP, and 48% of the patients who performed RF and/or anti-CCP tests had corresponding laboratory results (n=14,490). For patients excluded due to no lab tests or no results, 40.3-42.9% were seropositive (i.e., had M05.x for their second diagnosis) and 57.1-59.7% were seronegative (i.e., had M06.x for their second diagnosis) (Supplementary Table 3).

For each reference standard definition for seropositive RA, the number of identified patients ranged from 3,315 (high positive anti-CCP) to 14,490 (any positive RF or anti-CCP) (Table 1). The sensitivity for seropositive RA was between 65.5% (any positive anti-CCP) and 68.5% (high positive RF) whereas the PPV was slightly higher ranging from 51.5% (high positive anti-CCP) to 77.1% (any positive RF or anti-CCP). The specificity was highest with the reference standard definition of any positive RF or anti-CCP (72.43%), and the NPV was highest under the definition of high positive anti-CCP (76.6%). The kappa values were between 0.28-0.41, comparable to the values in the previous validation study (0.24-0.45) (5). For seronegative RA, sensitivity ranged from 63.5% to 72.4%. PPVs for seronegative ranged from 61.6% to 76.6%. In a secondary analysis where we restricted M05, M06 codes, the sensitivity for seropositive RA improved (ranging from 92.6 to 97.4%) as did PPV (77.1 to 79.2%). A notable increase was seen in PPV for seronegative RA with subset of codes (from 61.6 to 89.5%) (Supplemental Table 4).

We assessed baseline characteristics of RA patients according to diagnosis codes or laboratory tests performed. There were no notable differences in the demographics, comorbidities, medications, and healthcare utilization patterns between the entire RA cohort and sub-cohorts of RA patients with M05 diagnosis code, M06 diagnosis code, RF and/or anti-CCP performed, and those with RF and/or anti-CCP results (Supplementary table 1).

## DISCUSSION

In this study, we validated ICD-10 codes M05 and M06 as a proxy for seropositive and seronegative RA in a U.S. administrative claims database, Optum. The ICD-10 codes had good diagnostic accuracy for seropositive RA (sensitivity 67.31% and PPV 77.12% for the reference standard definition of any positive RF or anti-CCP) and seronegative RA (sensitivity 72.43% and PPV 61.61% for the reference standard definition of not having any positive RF or anti-CCP). When subsets of codes within M05 and M06 were used in a secondary analysis, the PPVs for both seropositive and seronegative RA increased (77.1 to 79.2% for seropositive RA, 61.6 to 89.5% for seronegative RA).

Our primary findings were comparable but lower than the performance reported by Curtis *et al.* with the RISE registry and MarketScan data (PPV of 81 to 84% and sensitivity of 73 to 82% for seropositive RA) (5). Both Optum and MarketScan databases contain claims data from nationwide commercial health plans. The two databases have different payer systems, but the distribution of demographic factors is known to be similar (6). Both databases contain older patients including Medicare beneficiaries; however, the percentage of older patients in Optum is larger than that of MarketScan. Consequently, the mean age in the Optum cohort is greater (61.3) than the mean age (54.0) reported in the MarketScan validation study (5). Differences in the patient characteristics and the prevalence of RA in each database as well as other factors such as billing patterns or differences in clinical practice patterns may also affect the performances of M05 and M06.

Because our RA cohort allows for both incident and prevalent cases, we cannot exclude a possibility of “default” coding (i.e., assign M06.9, rheumatoid arthritis, unspecified for unknown laboratory status) when the RA serostatus of a patient is unknown in prevalent cases. A recent analysis of 87 RA cases in the Western Australian Rheumatic Disease Epidemiological Registry reflects our concern as the ICD-10 code M06.9 was the most frequently (65.5%) used RA diagnosis code, followed by M05.9 (seropositive RA, 13.8%) (7). Other possible scenarios are patients incorrectly reporting their laboratory results to their physicians as a part of their medical history (8) and variability across different laboratory assays.

We observed that seropositive RA classification using M05 performed better for RF-only classification compared to anti-CCP-only classification. The definition of M05 code is specific to positive RF and as of now, there are no ICD-10 codes reflecting the anti-CCP status. In clinical practice, however, the coding for two tests is often used interchangeably. Also, our study used lab results as the reference standard, rather than clinical diagnosis. Some patients may have had alternative diagnoses (e.g., polymyalgia rheumatica or spondyloarthritis) that were identified after initial coding for RA. Another notable finding of our study was increase in the PPVs for seropositive and seronegative RA when more restrictive sets of diagnoses codes were used. Future studies could consider using a subset of M05 and M06 codes in identifying serostatus of RA patients.

In conclusion, compared to the sensitivity and PPV for seropositive and seronegative RA found in MarketScan and the RISE registry, we found a slightly lower performance of

ICD-10 codes in Optum. However, using a more restrictive set of codes in M05 and M06 may further improve the performance characteristics.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Funding statement:

This study was funded by internal resources in the Division of Pharmacoepidemiology and Pharmacoeconomics at Brigham and Women's Hospital. Kim is in part supported by the NIH K24AR078959.

### Conflict of Interest disclosure:

Sparks has received research support from Bristol-Myers Squibb and performed consultancy for Bristol-Myers Squibb, Gilead, Inova Diagnostics, Optum, and Pfizer unrelated to this work. Yoshida is a consultant to OM1, Inc. and receives grant/research support NIAMS K23AR076453 unrelated to this work. Kim has received research grants to the Brigham and Women's Hospital from Roche, Pfizer, AbbVie, and Bristol-Myers Squibb for unrelated studies.

## Data Availability:

Data subject to third party restrictions. The data that support the findings of this study are available from Optum Clinformatics Data Mart. Restrictions apply to the availability of these data, which were used under license for this study. Data are available the authors with the permission of Optum Clinformatics Data Mart.

## REFERENCES

1. Caskey RN, Abutahoun A, Polick A, Barnes M, Srivastava P, Boyd AD. Transition to international classification of disease version 10, clinical modification: the impact on internal medicine and internal medicine subspecialties. BMC Health Serv Res. 2018;18(1):328. Epub 20180504. doi: 10.1186/s12913-018-3110-1. [PubMed: 29728145]
2. Kim SY, Servi A, Polinski JM, Mogun H, Weinblatt ME, Katz JN, et al. Validation of rheumatoid arthritis diagnoses in health care utilization data. Arthritis Res Ther. 2011;13(1):R32. Epub 20110223. doi: 10.1186/ar3260. [PubMed: 21345216]
3. Frisell T, Holmqvist M, Källberg H, Klareskog L, Alfredsson L, Askling J. Familial risks and heritability of rheumatoid arthritis: role of rheumatoid factor/anti-citrullinated protein antibody status, number and type of affected relatives, sex, and age. Arthritis Rheum. 2013;65(11):2773–82. doi: 10.1002/art.38097. [PubMed: 23897126]
4. Reilly PA, Cosh JA, Maddison PJ, Rasker JJ, Silman AJ. Mortality and survival in rheumatoid arthritis: a 25 year prospective study of 100 patients. Ann Rheum Dis. 1990;49(6):363–9. doi: 10.1136/ard.49.6.363. [PubMed: 2383059]
5. Curtis JR, Xie F, Zhou H, Salchert D, Yun H. Use of ICD-10 diagnosis codes to identify seropositive and seronegative rheumatoid arthritis when lab results are not available. Arthritis Res Ther. 2020;22(1):242. Epub 20201015. doi: 10.1186/s13075-020-02310-z. [PubMed: 33059732]
6. Voss EA, Ma Q, Ryan PB. The impact of standardizing the definition of visits on the consistency of multi-database observational health research. BMC Med Res Methodol. 2015;15:13-. doi: 10.1186/s12874-015-0001-6. [PubMed: 25887092]
7. Almutairi K, Inderjeeth C, Preen DB, Keen H, Rogers K, Nossent J. The accuracy of administrative health data for identifying patients with rheumatoid arthritis: a retrospective validation study using medical records in Western Australia. Rheumatol Int. 2021;41(4):741–50. Epub 20210223. doi: 10.1007/s00296-021-04811-9. [PubMed: 33620516]

8. Booth MJ, Clauw D, Janevic MR, Kobayashi LC, Piette JD. Validation of Self-Reported Rheumatoid Arthritis Using Medicare Claims: A Nationally Representative Longitudinal Study of Older Adults. ACR Open Rheumatol. 2021;3(4):239–49. Epub 2021/02/23. doi: 10.1002/acr2.11229. [PubMed: 33621434]

**Key Points**

1. ICD-10 diagnosis codes (M05, M06) can help classify the serostatus of rheumatoid arthritis patients in administrative claims data.
2. Using a subset of codes within M05 and M06 greatly improved the PPVs of both seropositive and seronegative RA when compared with the original definition (77.1 to 79.2% for seropositive RA, 61.6 to 89.5% for seronegative RA).
3. Performance may vary in other databases due to differences in the data structure, clinical or billing practices, or patient characteristics.

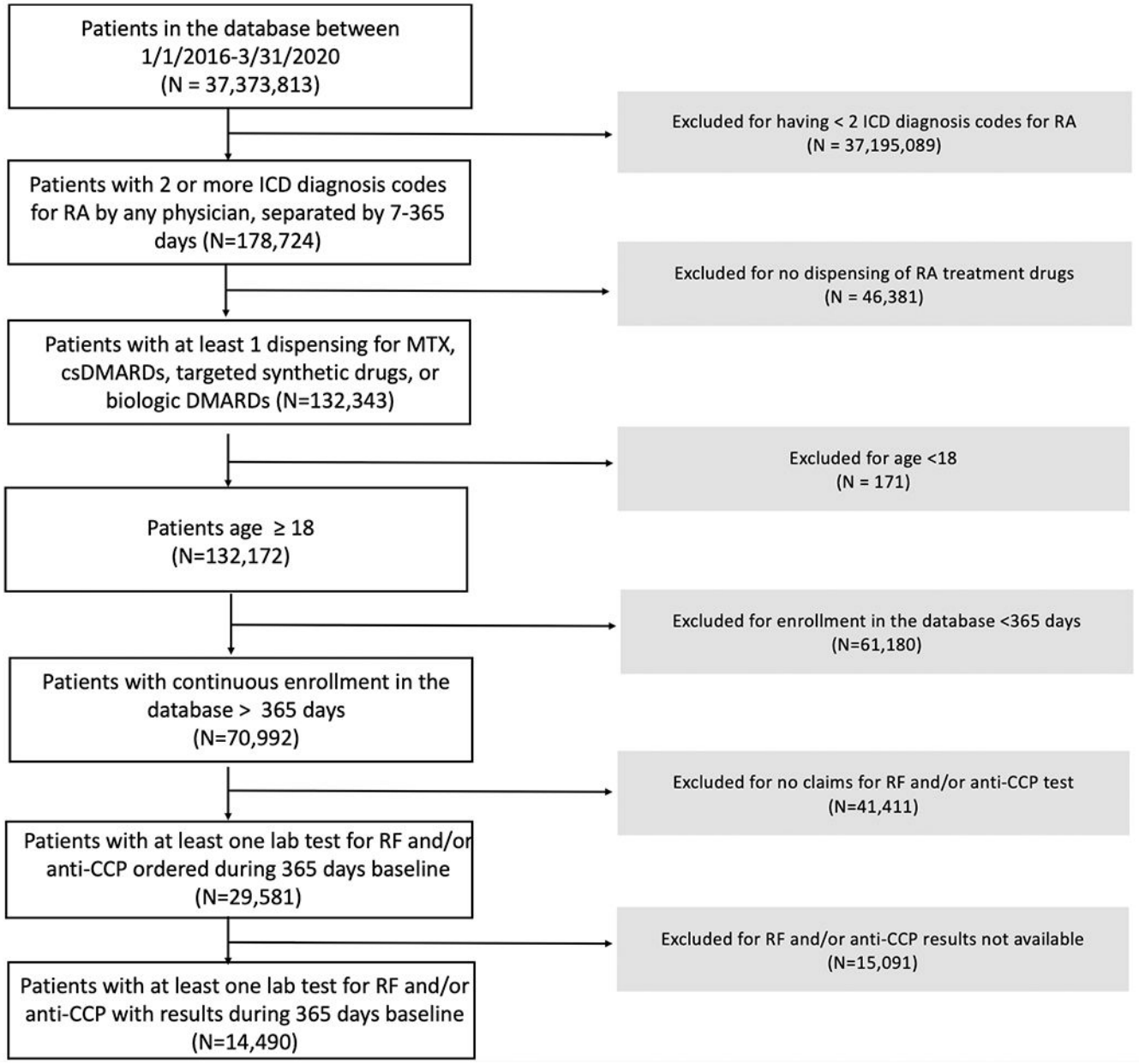


**Figure 1.**
Cohort flow diagram

**Table 1.**

Sensitivity, specificity, positive predictive value, and negative predictive values of M05 and M06 diagnosis codes in RA patients per each reference standard

| | Reference standard definition | Total number of identified patients | Sensitivity (95% CI) | Specificity (95% CI) | PPV (95% CI) | NPV (95% CI) | Kappa (95% CI) |
|---|---|---|---|---|---|---|---|
| Seropositive (M05.*) | Any positive RF or anti-CCP | 14,490 | 67.31 (66.31-68.32) | 72.43 (71.31-73.55) | 77.12 (76.16-78.08) | 61.61 (60.48-62.73) | 0.39 (0.37-0.40) |
| | Any positive RF | 13,636 | 67.61 (66.56-68.64) | 72.18 (71.04-73.33) | 76.00 (74.99-77.01) | 63.10 (61.95-64.25) | 0.37 (0.35-0.38) |
| | High positive RF | 10,494 | 68.49 (67.14-69.84) | 72.18 (71.04-73.33) | 65.54 (64.19-66.88) | 74.79 (73.66-75.91) | 0.41 (0.39-0.42) |
| | Any positive anti-CCP | 3,823 | 65.53 (63.29-67.77) | 63.47 (61.41-65.53) | 59.62 (57.41-61.83) | 69.11 (67.04-71.17) | 0.29 (0.26-0.32) |
| | High positive anti-CCP | 3,315 | 66.67 (64.02-69.31) | 63.47 (61.41-65.53) | 51.46 (48.99-53.92) | 76.63 (74.64-78.62) | 0.28 (0.25-0.32) |
| Seronegative (M06.*) | Not positive RF or anti-CCP | 14,490 | 72.43 (71.31-73.55) | 67.31 (66.31-68.32) | 61.61 (60.48-62.73) | 77.12 (76.16-78.08) | 0.39 (0.37-0.40) |
| | Not positive RF | 13,636 | 72.18 (71.04-73.33) | 67.61 (66.56-68.64) | 63.10 (61.95-64.25) | 76.00 (74.99-77.01) | 0.37 (0.35-0.38) |
| | Not positive high RF | 10,494 | 72.18 (71.04-73.33) | 68.49 (67.14-69.84) | 74.79 (73.66-75.91) | 65.54 (64.19-66.88) | 0.41 (0.39-0.42) |
| | Not positive anti-CCP | 3,823 | 63.47 (61.41-65.53) | 65.53 (63.29-67.77) | 69.11 (67.04-71.17) | 59.62 (57.41-61.83) | 0.29 (0.26-0.32) |
| | Not positive high anti-CCP | 3,315 | 63.47 (61.41-65.53) | 66.67 (64.02-69.31) | 76.63 (74.64-78.62) | 51.46 (48.99-53.92) | 0.28 (0.25-0.32) |