Assessing Data Quality of Rheumatoid and Psoriatic Arthritis Patients in the *All of Us* Research Program

Matthew Spotnitz, MD, MPH, John Giannini, PhD, Emily Clark, MPH, Yechiam Ostchega, PhD, RN, Tamara R Litwin, PhD, MPH, Lew Berman, PhD, MS

All of Us Data Collection Process

Authorize and share electronic health record data

Answer surveys

Provide physical measurements

Provide biosamples to be stored at biobanks

Share data from Fitbit devices

All of Us Data Types

The *All of Us* Research Program's Data and Research Center (DRC) curates a range of different data types as part of the data collection process. The numbers below reflect the number of participants with each data type available.

>633,000
with survey responses

>509,000
with physical measurements

>447,000
with genotyping arrays

>414,000
with short-read whole genome sequences (WGS)

>393,000
with electronic health record data

>97,000
with structural variant data

>59,000
with Fitbit records

>36,000
with Exploring the Mind task data

>2,700
with long-read WGS

Data as of February 2025 Curated Data Repository (CDR) v8 data release

Prior Studies

Topic	Title	Journal	Year
Ductal Carcinoma in Situ (DCIS)	Application of a Data Quality Framework to Ductal Carcinoma In Situ Using Electronic Health Record Data From the *All of Us* Research Program	JCO CCI	2024
Mastectomy	Assessing the Data Quality Dimensions of Partial and Complete Mastectomy Cohorts in the *All of Us* Research Program: Cross-Sectional Study	JMIR Cancer	2024
Surgical Oncology	Assessing the Data Quality Dimensions of Surgical Oncology Cohorts in the *All of Us* Research Program	JCO CCI	2025

Rheumatoid and Psoriatic Arthritis

- Rheumatoid and psoriatic arthritis (RA and PsA) are autoimmune diseases that have overlapping clinical symptoms and can cause debilitating joint pain.
- The American College of Rheumatology (ACR) and the European League Against Rheumatism (EULAR) recommend treating both conditions with Disease-modifying antirheumatic drugs (DMARDs).
- However, some practice patterns deviate from the guidelines.
- Therefore, generating real-world evidence may be valuable for research on the optimal treatment of RA and PsA.
- Our study aims to determine whether data from the *All of U*s program are fit for use for RA and PsA phenotypes.

Methods

- Study Design: Nested case-control
- Case Phenotypes: 1+ ICD/SNOMED diagnosis (dx) code
  - Rheumatoid Arthritis
  - Psoriatic Arthritis
- Control Phenotypes: No RA or PsA diagnosis
- Sensitivity analysis phenotypes: (i) 2 dx+, 30d+ (ii) Dx + DMARD (iii) 2dx+, 30d+, DMARD
- Manual selection of disease specific concept sets
- Data Sources: *All of Us* EHR and survey data

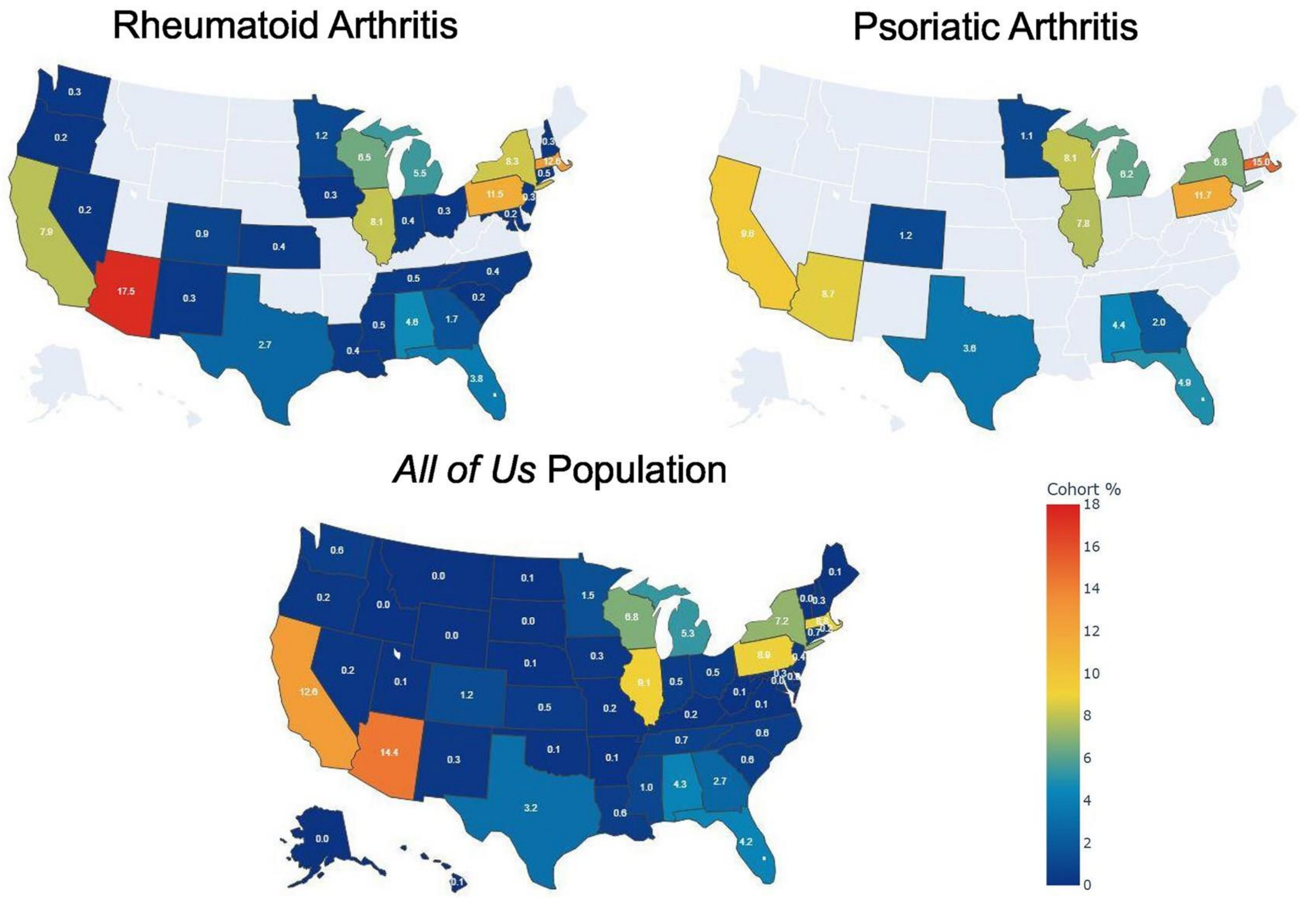
Data Quality Dimensions

- Conformance: Consistent Data Representation
- Completeness: Data Availability
- Concordance: Data Element Agreement
- Plausibility: Believable Data Elements
- Temporality: Expected Temporal Pattern

Berman L, Ostchega Y, Giannini J, et. al. Application of a Data Quality Framework to Ductal Carcinoma in Situ Using Electronic Health Record Data from the All of Us Research Program. JCO Clinical Cancer Informatics 2024 Aug:8:e2400052.

	RA Case n (%)	PsA Case n (%)	Controls n (%)
	10,753 (100)	1,836 (100)	380,389 (100)
Race / Ethnicity			
White	6375 (59.3)	1477 (80.4)	222,628 (58.5)
Black	2154 (20.0)	102 (5.6)	73,471 (19.3)
Hispanic	1936 (18.0)	210 (11.4)	69,578 (18.3)
Asian	208 (1.9)	49 (2.7)	14,509 (3.8)
MENA	104 (1.0)	≤20	4134 (1.1)
NHPI	≤20	≤20	1177 (0.3)
AIAN	587 (5.5)	56 (3.1)	15,789 (4.2)
Race Ethnicity None Of These	131 (1.2)	22 (1.2)	3882 (1.0)
Skip/Prefer Not To Answer	246 (2.3)	≤20	6745 (1.8)
Sex			
Female	8394 (78.1)	1146 (62.4)	229,585 (60.4)
Male	2219 (20.6)	674 (36.7)	146,802 (38.6)
None of the above or Skip	132 (1.3)	≤20	3788 (1.0)
Age at Diagnosis			
18-39	1050 (9.8)	262 (14.3)	10,7687 (28.3)
40-59	3987 (37.1)	753 (41.0)	131,449 (34.6)
60-79	5201 (48.4)	776 (42.3)	129,092 (33.9)
80+	515 (4.8)	44 (2.4)	12,017 (3.2)

Geospatial Distribution



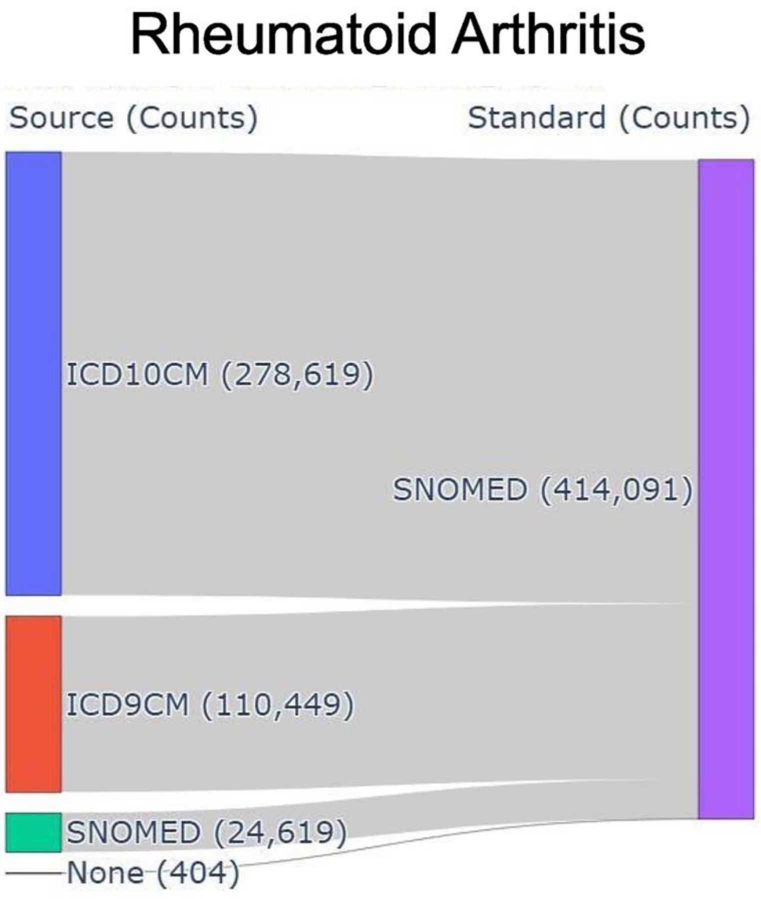
Rheumatoid Arthritis

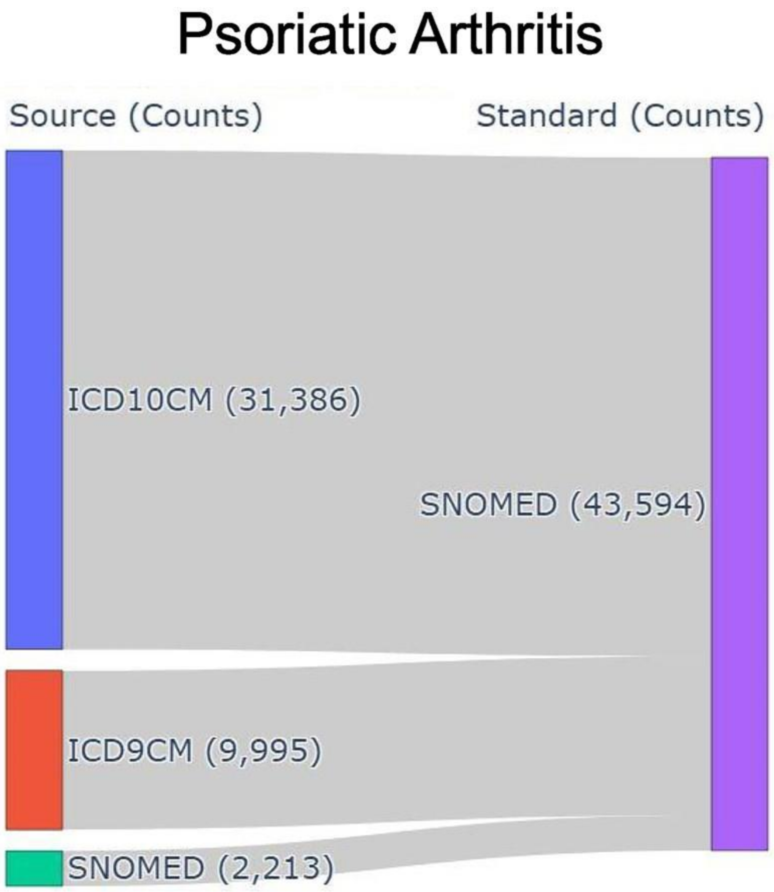
Psoriatic Arthritis

All of Us Population

States that contributed the most to the RA and PsA cohorts: MA, NY, PA, IL, WI, AZ, CA, FL

Conformance





Mostly ICD 10 source codes for both cohorts

Concept Completeness

Concept	RA Case (n, %)	PsA Case (n, %)	Control (n, %)
Anti-CCP	3292 (30.6)	392 (21.4)	12,394 (3.3)
Biologic DMARD	2573 (23.9)	832 (45.3)	3111 (8.2)
CRP	7008 (65.2)	1140 (62.1)	82,348 (21.6)
ESR	7457 (69.3)	1243 (67.7)	89,467 (23.5)
Foot X-ray	4541 (42.2)	657 (35.8)	57,908 (15.2)
Glucocorticoids	9147 (85.1)	1496 (81.5)	203,737 (53.6)
MRI Upper Extremity	501 (4.7)	209 (11.4)	17,579 (4.6)
NSAIDs	8741 (81.3)	1398 (76.1)	211,964 (55.7)
Rheumatoid Factor	4640 (43.2)	585 (31.9)	27,429 (7.2)
Wrist/Hand X-ray	5127 (47.7)	757 (41.2)	57,055 (15.0)
csDMARD	4887 (45.4)	707 (38.5)	8099 (2.1)

Multiple concepts with <50% completeness

Concordance

Variable	RA	PsA
No. Bivariate Pairs	66	136
No. $\rho>0.5$	3	3

Only 3 bivariate pairs with $\rho>0.5$

Plausibility

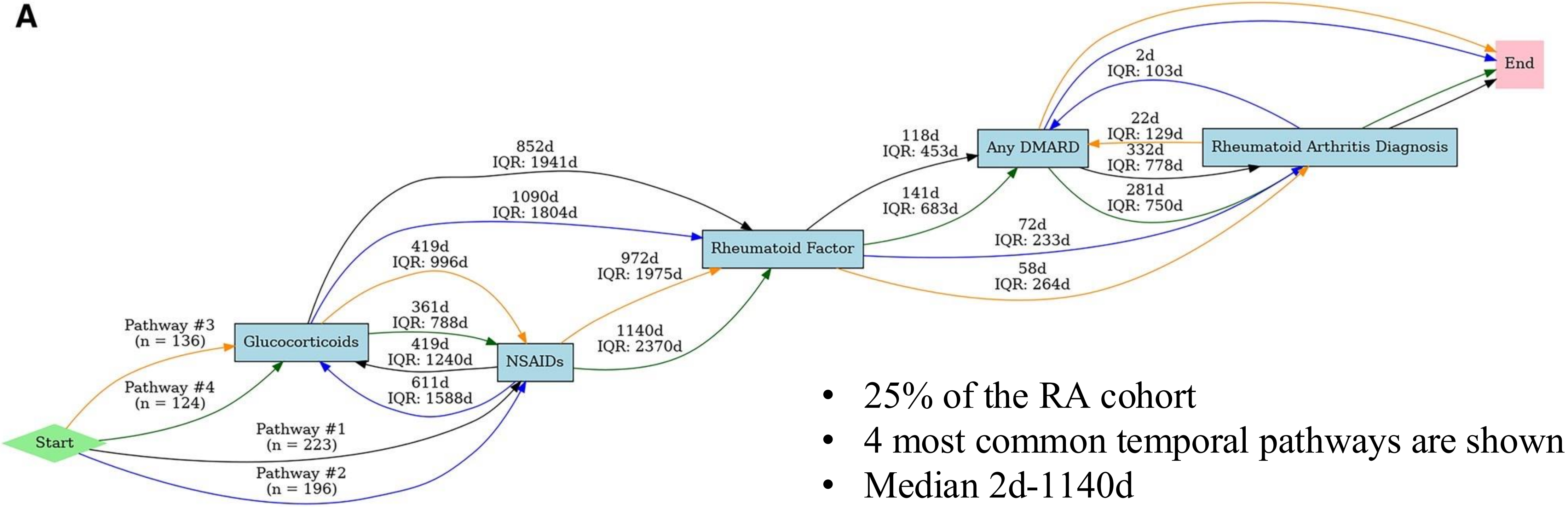
- Disease specific concepts were most prevalent for RA and PsA cases between 40-79 years of age
- RA and PsA cases self-reported higher percentages of poor or fair general health on a survey compared to controls ($p<0.001$)
- In the RA and PsA cohorts, methotrexate was the most prevalent conventional synthetic DMARD (27.5%, 30.0%)
- In the PsA cohort, TNF inhibitors were the most prevalent biologic DMARD (35.9%)

Temporality

Analysis	Rheumatoid Arthritis No. (%)	Median (IQR) (In Days)	Psoriatic Arthritis No. (%)	Median (IQR) (In Days)
RF to Diagnosis	4640 (43.2)	2.8 (111.8)	585 (31.9)	8.9 (132.8)
Diagnosis to csDMARD	4890 (45.5)	0 (74.7)	707 (38.5)	0 (79.9)
Diagnosis to NSAID	8742 (81.3)	-41.2 (300.0)	1398 (76.1)	-83.7 (345.7)
Diagnosis to Glucocorticoids	9147 (85.1)	-17.3 (263.7)	1496 (81.5)	-79.8 (313.3)
RF to csDMARD	2799 (26.0)	6.6 (82.8)	314 (17.1)	18.0 (118.2)
RF to Biologic DMARD	1428 (13.3)	64.9 (236.8)	298 (16.2)	43.5(180.5)
RF to NSAID	4051 (37.7)	-72.1 (306.5)	497 (27.1)	-97.9 (359.7)
RF to Glucocorticoids	4138 (38.5)	-32.1 (266.5)	510 (27.8)	-121.3 (338.8)

Variance in completeness, medians, IQRs

Temporality: RA



- 25% of the RA cohort
- 4 most common temporal pathways are shown
- Median 2d-1140d
- IQR 22-2370d
- Similar results in the PsA cohort

Completeness Sensitivity Analysis

- Subphenotypes reduced cohort size by up to 60%
- Concept set missingness differed by less than 12% compared to the main phenotypes

Discussion

- Application of a data quality framework to arthritis cohorts
- Data missingness for multiple concept sets
  - Non-digitized and out of network records
  - Concordance, plausibility, and temporality analyses were affected
  - Alternative phenotyping had a small effect on data missingness
- Observed variance in concept set sequences
  - Inconsistent coding and mapping practices, parallel treatment guidelines, differing insurance company preferences and practice patterns
- Conformance data were complete and recent (ICD-10)

Limitations

- Lack of external data
- Unmeasured RA/PsA misclassification and false positives
- Hard to differentiate missingness vs. variance practice patterns
- Uneven geospatial distribution
- Large variance in temporal data
- Manual concept selection
- Minimal unstructured data
- Adult population

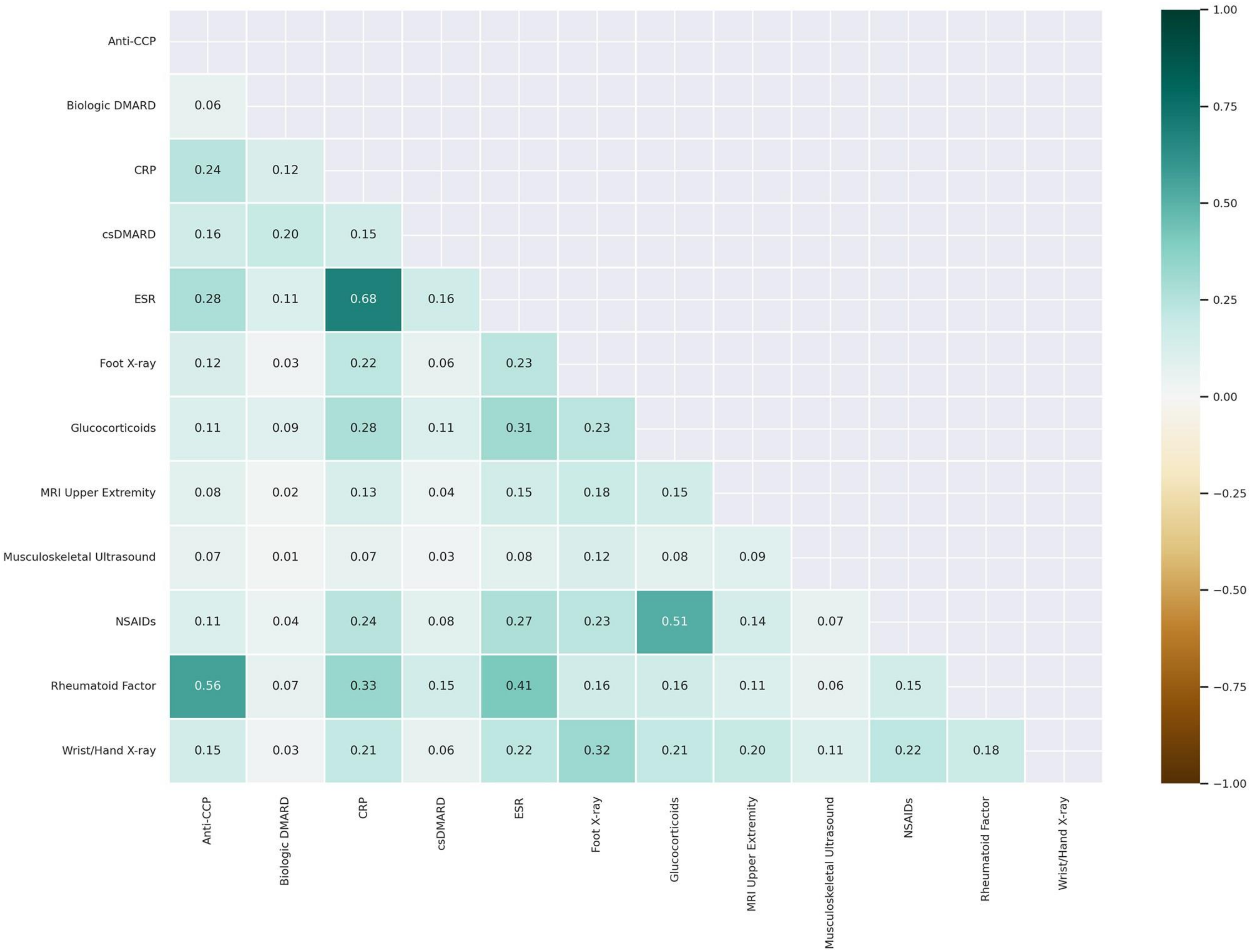
Conclusion

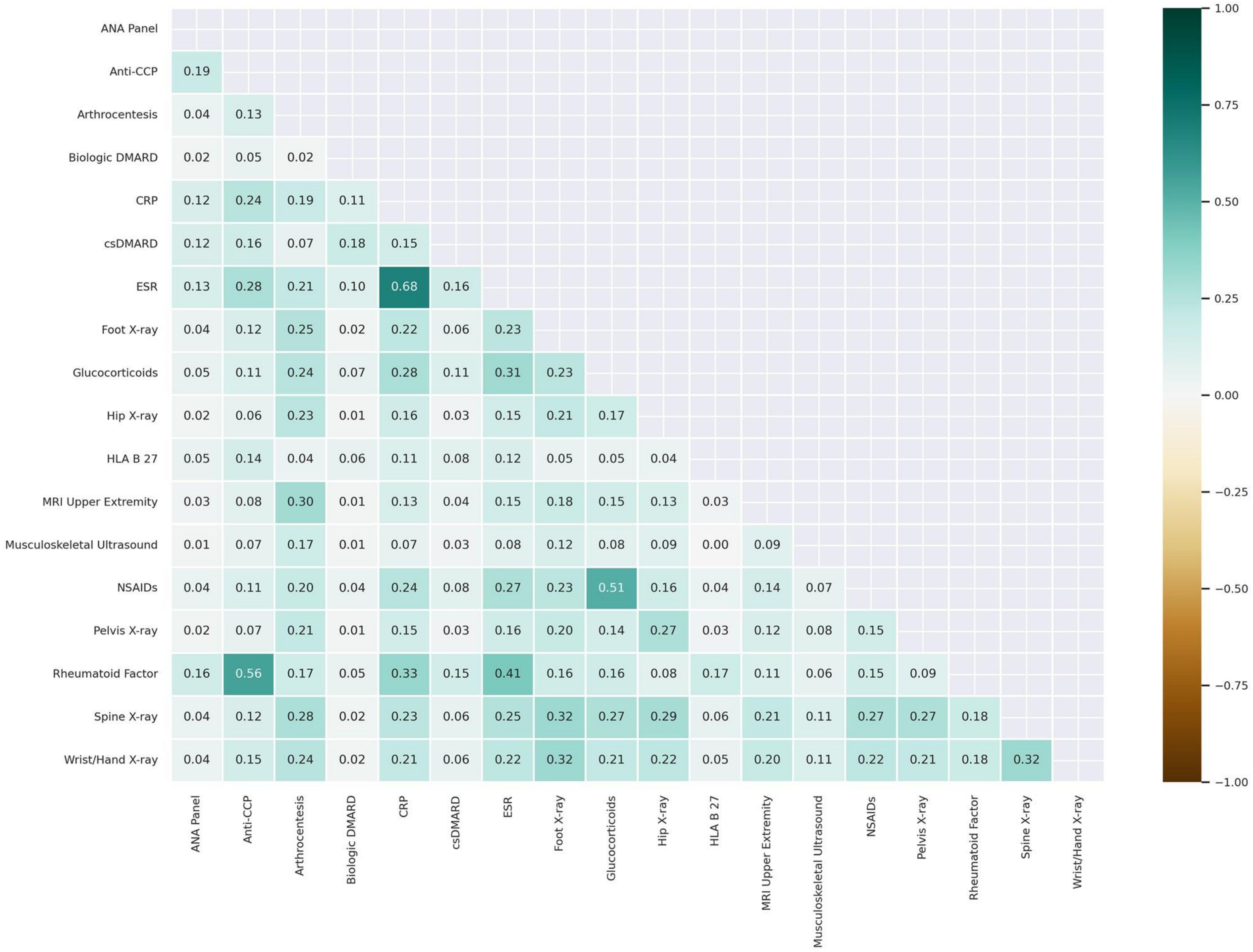
- Our data quality framework was implemented on rheumatoid and psoriatic arthritis cohorts to determine fitness for use. Further research is warranted to improve data quality for those conditions within the OMOP CDM.
- This approach can be generalized to other diseases and data types.
- Initiatives such as the Center for Linkage and Acquisition for Data (CLAD) may lead to improvements in completeness and data quality overall
- Future Direction: Inflammatory Bowel Disease data quality manuscript under review at JAMIA Open.

Acknowledgements

- Dr. Lew Berman
- Dr. John Giannini, Emily Clark, Dr. Tamara Litwin, Dr. Yechiam Ostchega
- Dr. Alison Lin
- Dr. Susan Gregurick
- ODSS
- NIH
- Dr. Patrick Ryan
- Craig Sachson
- OHDSI

Extra Slides


Anti-CCP
Biologic DMARD 0.06
CRP 0.24 0.12
csDMARD 0.16 0.20 0.15
ESR 0.28 0.11 0.68 0.16
Foot X-ray 0.12 0.03 0.22 0.06 0.23
Glucocorticoids 0.11 0.09 0.28 0.11 0.31 0.23
MRI Upper Extremity 0.08 0.02 0.13 0.04 0.15 0.18 0.15
Musculoskeletal Ultrasound 0.07 0.01 0.07 0.03 0.08 0.12 0.08 0.09
NSAIDs 0.11 0.04 0.24 0.08 0.27 0.23 0.51 0.14 0.07
Rheumatoid Factor 0.56 0.07 0.33 0.15 0.41 0.16 0.16 0.11 0.06 0.15
Wrist/Hand X-ray 0.15 0.03 0.21 0.06 0.22 0.32 0.21 0.20 0.11 0.22 0.18
Anti-CCP
Biologic DMARD
CRP
csDMARD
ESR
Foot X-ray
Glucocorticoids
MRI Upper Extremity
Musculoskeletal Ultrasound
NSAIDs
Rheumatoid Factor
Wrist/Hand X-ray
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


ANA Panel
Anti-CCP 0.19
Arthrocentesis 0.04 0.13
Biologic DMARD 0.02 0.05 0.02
CRP 0.12 0.24 0.19 0.11
csDMARD 0.12 0.16 0.07 0.18 0.15
ESR 0.13 0.28 0.21 0.10 0.68 0.16
Foot X-ray 0.04 0.12 0.25 0.02 0.22 0.06 0.23
Glucocorticoids 0.05 0.11 0.24 0.07 0.28 0.11 0.31 0.23
Hip X-ray 0.02 0.06 0.23 0.01 0.16 0.03 0.15 0.21 0.17
HLA B 27 0.05 0.14 0.04 0.06 0.11 0.08 0.12 0.05 0.05 0.04
MRI Upper Extremity 0.03 0.08 0.30 0.01 0.13 0.04 0.15 0.18 0.15 0.13 0.03
Musculoskeletal Ultrasound 0.01 0.07 0.17 0.01 0.07 0.03 0.08 0.12 0.08 0.09 0.00 0.09
NSAIDs 0.04 0.11 0.20 0.04 0.24 0.08 0.27 0.23 0.51 0.16 0.04 0.14 0.07
Pelvis X-ray 0.02 0.07 0.21 0.01 0.15 0.03 0.16 0.20 0.14 0.27 0.03 0.12 0.08 0.15
Rheumatoid Factor 0.16 0.56 0.17 0.05 0.33 0.15 0.41 0.16 0.16 0.08 0.17 0.11 0.06 0.15 0.09
Spine X-ray 0.04 0.12 0.28 0.02 0.23 0.06 0.25 0.32 0.27 0.29 0.06 0.21 0.11 0.27 0.27 0.18
Wrist/Hand X-ray 0.04 0.15 0.24 0.02 0.21 0.06 0.22 0.32 0.21 0.22 0.05 0.20 0.11 0.22 0.21 0.18 0.32
ANA Panel
Anti-CCP
Arthrocentesis
Biologic DMARD
CRP
csDMARD
ESR
Foot X-ray
Glucocorticoids
Hip X-ray
HLA B 27
MRI Upper Extremity
Musculoskeletal Ultrasound
NSAIDs
Pelvis X-ray
Rheumatoid Factor
Spine X-ray
Wrist/Hand X-ray
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Plausibility

csDMARD	**Rheumatoid arthritis (n, %)**	**Psoriatic arthritis (n, %)**
Methotrexate	2958 (27.5)	549 (30.0)
Sulfasalazine	1044 (9.8)	181 (9.9)
Leflunomide	1000 (9.3)	94 (5.1)
Hydroxychloroquine	3175 (29.5)	135 (7.4)
Any csDMARD	4887 (45.4)	707 (38.5)

Completeness Sensitivity Analysis: Less than 12% missingness difference

	2dx+, 30d+	**1dx+, DMARD**	**2dx+, 30d+, DMARD**
RA			
No., %	7164 (66.6)	5,379 (49.7)	4,372 (40.7)
PsA			
No., %	1282 (69.8)	1035 (56.4)	849 (46.2)

Concordance

Variable	RA	PsA
No. Bivariate Pairs	66	136
No. $\rho>0.5$	3	3
Concepts	1. ESR and CRP ($\rho=0.68$) 2. Rheumatoid Factor and anti-CCP ($\rho = 0.56$) 3. NSAIDs and glucocorticoids ($\rho = 0.51$)	1. ESR and CRP ($\rho = 0.68$) 2. Rheumatoid Factor and anti-CCP ($\rho = 0.56$) 3. NSAIDs and glucocorticoids ($\rho = 0.51$)

Concept Completeness

Concept	RA Case (n, %)	PsA Case (n, %)	Control (n, %)
Anti-CCP	3292 (30.6)	392 (21.4)	12,394 (3.3)
Biologic DMARD	2573 (23.9)	832 (45.3)	3111 (8.2)
CRP	7008 (65.2)	1140 (62.1)	82,348 (21.6)
ESR	7457 (69.3)	1243 (67.7)	89,467 (23.5)
Foot X-ray	4541 (42.2)	657 (35.8)	57,908 (15.2)
Glucocorticoids	9147 (85.1)	1496 (81.5)	203,737 (53.6)
MRI Upper Extremity	501 (4.7)	209 (11.4)	17,579 (4.6)
NSAIDs	8741 (81.3)	1398 (76.1)	211,964 (55.7)
Rheumatoid Factor	4640 (43.2)	585 (31.9)	27,429 (7.2)
Wrist/Hand X-ray	5127 (47.7)	757 (41.2)	57,055 (15.0)
csDMARD	4887 (45.4)	707 (38.5)	8099 (2.1)