AUTHOR INFORMATION
Tjardo D. Maarseveen[1], Marc P. Maurits [1], Nils Steinz [1], Sytske Anne Bergstra [1], Bianca Boxma-de Klerk [1], Annette H. M. van der Helm-van Mil [1 2], Cornelia F. Allaart [1], Marcel J. T. Reinders [3 4], Thomas W J Huizinga [1], Erik B van den Akker [3 4], Rachel Knevel [1 5]

[1] Department of Rheumatology, Leiden University Medical Center, Leiden, Zuid-Holland, NL
[2] Department of Rheumatology, Erasmus Medical Center, Rotterdam, NL
[3] The Delft Bioinformatics Lab, Delft University of Technology, Delft, Zuid-Holland, NL
[4] Leiden Computational Biology Centre, Leiden University Medical Center, Leiden, Zuid-Holland, NL
[5] Rheumatology, Newcastle University Translational and Clinical Research Institute, Newcastle upon Tyne, UK

TITLE
Location of joint involvement differentiates Rheumatoid arthritis into different clinical subsets.

ABSTRACT

Objectives To aid etiology and treatment research of the very heterogeneous rheumatoid arthritis (RA) population, we aimed to identify phenotypically distinct RA subsets using baseline clinical data.
Method
We collected baseline numerical- (hematology work-up & age) and categorical variables (serology, joint location & sex) from the Electronic Health records (EHR) repository of the Leiden University Medical Center, comprising 1,387 unique first visits to the outpatient clinic. We used deep learning and graph clustering to identify phenotypically distinct RA subsets. To ensure the robustness of our findings, we tested a) cluster stability (1000 fold) b) physician confounding, c) association with remission and methotrexate failure, d) generalizability to a second different data set (Leiden Early Arthritis clinic; n=769).
Results We identified four subsets (C1-C4) of patients with rheumatoid arthritis that were delineated on the following characteristics: C1) arthritis in feet, C2) seropositive oligo-articular disease, C3) seronegative hand arthritis, C4) polyarthritis. Our validity analyses showed high stability (mean 78%-91%), no physician confounding, and a significant difference in methotrexate failure(*P-value*=6.1e-4) and occurrence of remission*(P-value*=7.4e-3), and generalizability to a second dataset. The hand-cluster (III) had the most favorable outcomes ($HR_{remission}$=1.65 (95%CI:1.20-2.29), $HR_{methotrexate}$=0.48 (95%CI:0.35-0.77)), particularly the ACPA-positive patients in this cluster, while in the other clusters the ACPA-negative patients did best. The clusters outperformed standard clinical variables, which were attributed to the hand and feet differentiation.
Conclusions
We discovered four phenotypically distinct subgroups of rheumatoid arthritis at baseline that associate with clinical outcomes. Furthermore, our study provides evidence for the presence of separate hand and foot subgroups in RA

Key messages

What is already known about the subject?

- Rheumatoid arthritis is a heterogeneous disease and clinicians have not identified the disease discerning patterns in clinical practice.
- Rheumatoid arthritis features differ for patients with and without ACPA and by age of onset.
- Electronic Health Records features an unprecedented wealth of information, but studying this multimodal high dimensional data source is a methodological challenge.
- Data-driven unsupervised techniques are able to identify hidden structures in big data.

What does this study add?

- We identified four novel RA clusters at baseline: feet involvement, oligo-articular disease, hand involvement and polyarthritis. In particular, the hand and feet clusters show a marked difference in outcomes, despite having a similar number of involved joints.
- The association of both ACPA and age of onset with long-term outcomes differed between the clusters, suggesting that the association between these markers might be different depending on the RA subset.

How might this impact on clinical practice or future developments?

- RA-patients with hand or feet involvement at baseline show differences in long term outcomes, which can be indicative of being distinct phenotypes, possibly requesting different treatment solutions.
- Grouping patients into clusters could unveil nuances in the disease entity that might be overlooked when categorizing them based solely on attributes like ACPA, age, and joint inflammation, as the association of these attributes with clinical outcomes differs across the four RA subsets.

Introduction

Rheumatoid arthritis (RA) is a heterogeneous disease. The current classification criteria for RA were developed to approximate the decision to start early treatment and the exclusion of other diseases. At clinical presentation, patients vary in the number- and pattern of joints involved, presence of extra articular manifestations and abnormalities in blood and synovial fluid [1, 2]. The heterogeneity of RA also manifests in long-term outcomes, namely prognosis, treatment response and comorbidities. This evident diversity may, among other things, impact the interpretation of the genetic component and downplay its importance altogether [3]. If phenotypic subsetting into more homogeneous groups is possible, there is a potential for better research into the etiology and enhancing the treatment of RA.

For centuries, pattern recognition on clinical variables by doctors has been the driving force of disease identification and examination of the underlying etiologic mechanisms. Thus far clinicians have not identified the relevant (sub)patterns in RA. The presence of ACPA [4, 5, 6, 7] and the age of onset [8, 9, 10] are commonly raised as possible dichotomous disease subsetting features. However, neither of these markers in isolation adequately addresses the heterogeneity and complexity of the disease. Because classical statistical approaches such as regression analysis fall short with respect to capturing the nuances within the data, we turn to clustering instead.

Cluster analysis has demonstrated its effectiveness in categorizing complex diseases (such as diabetes type II, asthma, osteoarthritis) into subtypes that differ in clinical outcomes or biological background [11, 12, 13]. In the context of RA, there is quite some focus on molecular phenotyping such as done by Lewis et al [14], who discovered patterns in synovial tissue at baseline, with the lymphoid-myeloid pathotype being a predictor for a poor outcome at disease onset [15]. Moreover, Jung et al [16] used clinical and comorbidity information for clustering and identified four subsets, including one that exhibited a higher likelihood for biological DMARD initiation. Likewise, Curtis et al [17] used clinical variables, but not perse at baseline, and identified five clusters that differed in disease activity, RA-duration and type of comorbidities.

In the current study, we chose to particularly focus on the clinical variables collected in routine clinical care, similarly to the study of Jung et al. Our focus lies on elucidating the clinical heterogeneity within RA and thus we do not take the comorbidities along with clustering, nor do we look at differences in prescription behavior (i.e. DMARD initiation). Rather, we look at the differences in long-term clinical outcomes to evaluate the clinical relevance of the RA clusters. Furthermore, we hypothesize that the location of the involved joints and the inflammatory patterns observed in the blood play a crucial role in subsetting RA, similar to their significance in distinguishing PsA from RA. [18]

In this context, the Electronic Health Records (EHR) data provides a powerful asset as it encompasses a wide variety of data types (laboratory values, clinical examination, demographics) that each offer a unique perspective on the patient's condition. The EHRs are collected as routine clinical care, and thus resemble the true patient population more closely than a study population collected with a particular hypothesis in mind. . The diversity of data types does however pose a methodological

challenge due to structural differences. Nonetheless, the recent surge of deep learning tools [19], offered the possibility to combine different EHR-layers into a representative patient embedding. The literature reports that clustering on top of these embeddings typically outperforms conventional techniques in the case of high dimensional or complicated data [20, 21, 22].
This sparked us to develop our current study in which we use state-of-the-art data-driven techniques to identify the disease-discerning signatures of RA using baseline clinical variables.

Methods

Patients

Our study comprises two different phases (Fig. S1): a developmental phase to identify the subtypes at baseline in training set (set A) and a validation phase that uses an independent historic cohort set (set B).

In set A we included patients from the rheumatology clinic of the Leiden University Medical Center, based on the physician's diagnosis within 1 year since the first visit. This was accomplished with our previously validated AI-model, which extracts the diagnosis from the format-free conclusion section of the physician notes [23, 24]. Hereby, we only considered patients after the initiation of the digital EHR-system on August 29th 2011 till the most recent data freeze on December 1st 2022.

For set B we selected RA-patients from the Leiden Early Arthritis Clinic (EAC) [25] between January 1st, 2004 till August 29th, 2011 to ensure there was no overlap with set A. Herein, the diagnosis of RA was defined as fulfilling 1987 RA criteria, measured cumulatively during the initial year. In both sets we required at least 1 year of follow-up. Prior to conducting the study, we acquired approval from the ethics committee of the LUMC. Patients and public were not involved during the development, execution, and dissemination of the study.

Selection of phenotypic variables

To construct patient phenotypic profiles we extracted information on serology (RF and ACPA), location of joint involvement (Tender- and Swollen joints), demographics, hematology workup and ESR at baseline (Table S1). In set B we had the same variables except for some hematology ones (Table S1).
Baseline was defined as the first visit to the clinic (set A) or the moment of inclusion in the cohort (set B) which generally happens at the first visit. Patients with missing lab or joint location variables were dropped (Fig S2).

Construction of the clusters

We combined the different data types from the EHR to construct a mathematical representation of the patients (i.e. a patient embedding) with a Multi-modal autoencoder (MMAE) [19]. We used PhenoGraph[26], a technique from the single cell field, to further organize the patients into subcommunities based on their similarities in clinical parameters. PhenoGraph was preferred over

traditional methods like K-means, since it is better at handling sparse data with many variables (see the supplemental material).

Cluster interpretation

For each cluster, we examined the characteristics and visualized the phenotype on a pictorial mannequin with an integrated heatmap to demonstrate the number of involved joints. We used a surrogate ML-technique to model the cluster assignment and subjected this model to a SHAP (SHapley Additive exPlanations) [27] analysis to retrospectively identify the most important variables per cluster. The SHAP plots show the strength and direction of impact of that variable for each patient (also those who are not assigned to that cluster).

Cluster validation

To confirm that our identified clusters comprised a stable and explainable partitioning, we performed a number of validation checks. We inferred the stability of the clusters across 1000 random subsets of the data, checked for possible bias/confounders (physician, disease duration, parvovirus B19 infection) and evaluated the long-term outcomes, namely: time to MTX-failure (defined by replacement of or adding an additional DMARD to MTX) and remission (DAS44 < 1.6) within one year. Finally, we evaluated the replicability on an external dataset (set B).

Statistical tests

We inferred statistical significance by employing Chi-squared for categorical and Mann Whitney U for continuous variables. When comparing more than two groups, we used Kruskal Wallis and post-hoc Dunn's test. For survival analysis we utilized the log-rank test to examine the overall trend and a univariate Cox-regression [28] to quantify the cluster differences. We applied the last-observation-carried forward (LOCF) approach to handle missing data. The proportional hazards assumption was verified by examining the Schoenfeld residuals [29]. Finally, to compare the fitness of the generalized models for MTX-response we used ANOVA. In all tests, we considered $P<0.05$ significant. All of our scripts are publicly available at Github [30].

Results

Patients

We retrieved 2,691 RA patients for training set A and 868 for test set B of whom, 1,387 and 769 were included in our study based on the availability of lab values and joint counts (Fig. S2, Table 1). Both datasets featured a typical early RA population [31, 32]. Patients in set B were slightly more often seropositive, somewhat younger and had a lower number of tender joints.

Table 1: Baseline table for both the development set (set A) and the replication set (set B)

	SET A	**SET B**
N	1387	769
Sex, female [1] (N(%))	894 (64.5)	523 (68.0)
Age [1] (SD, yr)	60.0 (15.0)	56.6 (15.7)
RF [1] (N(%))	722 (52.1)	443 (57.6)

ACPA [2] ***(N(%))***	676 (48.7)	406 (52.8)
ESR [2] ***(IQR, mm/hr)***	25 (11-45)	32 (17-54)
DAS44(3) (IQR)	3.5 (2.7-4.3)	3.4 (2.6-4.2)
SJC (IQR)	7 (4-12)	8 (4-14)
TJC (IQR)	10 (5-17)	7 (3-15)
DAS28(3) (IQR)	5.2 (4.3-6.2)	5.3 (4.3-6.1)
MTX (N(%))	1707 (858-2575)	1972 (1476-4033)
Symptom duration (IQR, days)	155 (62-512)	135 (70-256)

Where SD= standard deviation; RF= Rheumatoid factor; ACPA= Anti-cyclic citrullinated peptide (ccp) antibodies; ESR=erythrocyte sedimentation rate; IQR=interquartile range; DAS=Three component disease activity score (either 44 or 28 joint scheme); SJC=swollen joint count; TJC=tender joint count; MTX= prevalence of patients receiving methotrexate at baseline;

[Place figure 1 here]

Figure 1: Two-dimensional Uniform Manifold Approximation and Projection (UMAP)-representation of the patient embedding*. Here each patient is represented by a dot that is colored by the four clusters in the first plot and a gradient from high (blue) to low (red) in the subsequent plots. From left to right: dots are colored on corresponding cluster, Rheumatoid factor (RF) status, anti-cyclic citrullinated protein (ACPA) status, Age, disease activity score 28 (DAS28), erythrocyte sedimentation rate (ESR), Sex (1= Female, 0=Male).*

[Place figure 2 here]

Figure 2: Cluster summary plots showing the average joint involvement and the top 15 driving variables. *The mannequin is formatted as a heatmap, showing the prevalence (red=100%, yellow=0%) of joint involvement (tender or swollen). In the SHAP plots, the most informative features for each cluster are listed in descending order. Here the x-axis shows the strength and direction of impact of that variable for each patient (represented by a dot). The colour of the dot shows the initial value of the clinical variable (pink=high, blue=low). Where ACPA=anti-cyclic citrullinated peptide antibodies; ESR=erythrocyte sedimentation rate; IP= interphalangeal; L=left, MCH=mean corpuscular hemoglobin; MCHC=mean corpuscular hemoglobin concentration; MCP=metacarpophalangeal; MTP=metatarsophalangeal; PIP=proximal interphalangeal; R=right;*

<u>Four clusters separated by joint location, serology and blood values</u>

We detected four clusters in our patient embedding (Fig. 1). Notably, these clusters were not driven by any single clinical variable, as indicated by the wide dispersion of values. The clusters exhibited variations in joint location, amount of inflammation, age and differences in seropositivity (Figure 2, Table 2):

- <u>Cluster 1 (n=415) feet</u>
  moderate number of involved joints, particularly feet joints, younger patients, low leukocyte and thrombocyte levels.
- <u>Cluster 2 (n=380) oligo-articular</u>
  limited joint involvement and mostly seropositive patients.
- <u>Cluster 3 (n=323) hand</u>
  elderly patients symmetrical polyarthritis of hands seronegative.
- <u>Cluster 4 (n=269) polyarthritis</u>

majority seronegative polyarthritis in hand and feet though with lower ESR.

Notably, the identified patient clusters do not seem to represent different disease stages as the cluster with the longest symptom duration has the lowest joint count and vice versa. On average >80% of patients group together in the same cluster over all iterations in the stability analysis (Fig. S3&S4). In fact, the stability was better in our combined multi-modal approach than if we take each data type (numeric/categorical) separately (Fig. S5). Moreover, the clusters are not confounded by treating physicians (Fig. S6) and are recurring in the external validation set B (Fig. 3, S7).

Table 2. Baseline characteristics of the different patient clusters. *The clinical variables used for clustering are marked with the gamma symbol (ᵞ).*

	C1-Feet	C2-Oligo	C3-Hand	C4-Poly
N	415	380	323	269
Sex, female ᵞ [N(%)]	262 (63.1)	262 (68.9)	198 (61.3)	172 (63.9)
Age ᵞ (SD, yr)	56.8 (14.4)	59.7 (14.9)	68.5 (12.6)	55.1 (14.7)
RF ᵞ [N(%)]	245 (59.0)	236 (62.1)	121 (37.5)	120 (44.6)
ACPA ᵞ [N(%)]	241 (58.1)	224 (58.9)	97 (30.0)	114 (42.4)
ESR ᵞ (IQR, mm/hr)	22 (9-36)	29 (14-46)	31 (17-51)	22 (9-41)
DAS44(3) (IQR)	3.5 (3.0-4.0)	2.5 (2.0-2.8)	3.8 (3.2-4.4)	4.7 (4.0-5.5)
SJC (IQR)	8 (5-11)	3 (1-4)	10 (7-14)	14 (9-21)
TJC (IQR)	11 (8-15)	3 (2-5)	11 (8-15)	24 (18-31)
DAS28(3) (IQR)	5.3 (4.4-6.0)	4.2 (3.4-4.7)	5.6 (5.0-6.4)	6.6 (5.5-7.4)
MTX [N(%)]	354 (85.3)	275 (72.4)	240 (74.3)	218 (81.0)
Follow up (IQR, days)	1749 (860-2663)	1822 (1024-2587)	1267 (706-2179)	2032 (1141-2909)
Symptom duration (IQR, days)	154 (56-365)	217 (63-740)	122 (42-365)	155 (55-365)

SD= standard deviation; RF= Rheumatoid factor; ACPA= Anti-cyclic citrullinated peptide (ccp) antibodies; ESR=erythrocyte sedimentation rate; IQR=interquartile range; DAS=Three component disease activity score (either 44 or 28 joint scheme); SJC=swollen joint count; TJC=tender joint count; MTX= prevalence of patients receiving methotrexate at baseline;

[Place figure 3 here]

Fig. 3: Pictorial mannequins for replication set B and their original counterpart (set A) to show the affected joints for each cluster with color and size to depict prevalence. *Frequency is colored on a gradient from red (=100%) to yellow (=0%). When there is no colored dot, it signifies the absence of both swelling and pain at baseline for these patients.*

Validation on long-term outcomes

In Set A, approximately 80% of patients received MTX as an initial drug across all clusters. The Kaplan Meier curves show a difference in MTX failure between the clusters: 27%, 23%, 16%, 30% (for cluster 1-4, *P*=6.1e-4, Fig. 4a). The feet cluster (C1) was twice as likely to fail on MTX compared to the hand

cluster (C3), despite showing similar disease activity at baseline (HR:2.39 (1.34-4.27), *P*=3.0e-3). Similarly, C3 patients are more likely to stay on MTX than the most severe disease subtype (C4) (HR 0.48 (95% CI 0.35-0.77), *P*=9.5e-4).
The association between MTX-response and our clusters was not a random finding: the p-value with our clusters was lower compared to random cluster assignment in 99.9% of the instances (Fig. S8). Consistent with MTX response, we observed differences in remission rates: 44.3%, 47.4%, 55.7%, 38.5% (for cluster 1-4, *P*=7.4e-3, Fig. 4b) with the biggest difference between the hand C3 and polyarthritis group C4 (HR 1.65 (95% CI 1.2-2.29), *P*=2.4e-3). C3 also had the highest remission rate in the replication set, especially when compared to the feet cluster of C1 (*P=3*.0e-2) (Fig. S9).
The good response in cluster 3 raised the question whether this group overrepresented patients with parvovirus induced arthritis instead of RA, but none of our clusters were enriched for parvovirus positive patients (Fig s10).

[Place figure 4 here]
Figure 4: Kaplan Meier survival curves depicting MTX switch rates and remission (DAS44<1.6) as a function of the cluster membership at a follow-up period of 1 year. *Follow-up data only shown for MTX-starters (n=1,084; A) or patients with remission information (n=676; B). Global trend was inferred with log rank test, where p<0.05 is marked with **

Since the literature reports that ACPA is indicative of persistent disease [33], we examined whether the ACPA status was the main factor driving the difference in MTX failure. In the overall data we observed a higher treatment failure in ACPA positive than negative patients (27.6% versus 22.0%, Fig. 5a) though it was not significant. Intriguingly, the association of ACPA with MTX-failure differed within the clusters (Fig. 5b); within the good responder C3, the ACPA positive patients responded better to treatment than the ACPA negative patients and in C4 there was no difference in MTX failure based on ACPA.

[Place figure 5 here]
Figure 5: Kaplan Meier depicting global effect of ACPA on MTX switch rates (A), and local effect within clusters (B). *Here the dashed lines indicate the seropositive patients. Global trend was inferred with the log rank test, where p<0.05 is marked with an asterix.*

Finally, we checked whether the clusters add any additional information to a regression model of baseline variables known to associate with MTX failure (model 1). Here, the addition of the clusters (model 2) significantly improved the fit of the model based on known risk factors (*P*=1.8e-2) (Fig. S11A). This improvement was not mediated by differences in disease or symptom duration, delay in treatment or the number of affected joints (Table S5).
When we encoded the joint location into two binary variables for a) feet involvement and b) hand involvement and added this to the initial model (model 3), the clusters (model 4) no longer improved the model on MTX-failure. Thus, the hand and feet count provided a proxy for cluster membership (Fig. S11B-C).

Discussion

Through the application of deep learning and clustering, we identified four phenotypically distinct subtypes of RA at the baseline visit. These were characterized by arthritis in feet (C1), seropositive oligoarticular disease (C2), seronegative arthritis in hand in elderly patients (C3) and polyarthritis (C4). We demonstrated that our clusters are stable and not confounded by physicians or disease stage. Finally, we validated the findings with long-term outcomes MTX response and occurrence of remission and replicated the clusters in a separate dataset.

Our cluster analysis captured several interesting findings. The first notable finding is the observation of an elderly onset RA cluster with higher inflammation (ESR and SJC), lower prevalence of women and lower prevalence of ACPA and RF compared to the total group of RA patients. This is consistent with previous research on elderly onset RA (EORA).[9, 10] Then, our study uncovers more granular subtypes than EORA versus young onset RA (YORA). First of all there are younger people with a clinical pattern similar to the EORA in C3 and older people that cluster with younger seropositive patients. Also, our youngest cluster (C4) is defined by high joint involvement and low prevalence of ACPA and RF. The distinction between the 'old'(C3) and the 'young' (C4) cluster is the involvement of feet in C4. Our results contained a second cluster characterized by feet involvement, C1. The two feet clusters (C1,C4) showed worse long-term outcomes than the hand cluster (C3). This difference was not explained by differences in number of involved joints (i.e. C1 and C3 had similar joint counts), ACPA status or disease duration at baseline/initiation of treatment.

Important to note is the prevalence of ACPA positivity in the clusters. It is often hypothesized that ACPA divides RA into two sub entities, but studies of the phenotype of the disease at baseline do not find clear differences (only in established disease).[31,34] Our data suggests that other factors, such as the location of joint involvement are important for disease subsetting. Within these subsets the association of ACPA differs and deviates from current knowledge about ACPA and clinical features. ACPA's prevalence was lowest in the more RA typical clusters with polyarthritis (C3 and C4) and highest in the cluster with the lowest joint involvement (C2).
In set B (the replication cohort) we observed a similar pattern of ACPA prevalence between the clusters though less prominent. The (changing) RA classification criteria might influence misclassification making it more likely for patients to get an RA diagnosis when they have polyarthritis or are ACPA positive.[35] In our data, this cannot completely explain this effect. For set A, we did not use the classification criteria and for both sets we checked the RA diagnosis over the course of a year, which overcomes most of the initial clinical misclassification.

There was no notable difference in MTX continuation between a positive and negative ACPA-status. However, we did notice that within the clusters ACPA was differently associated with the continuation of MTX. Follow-up studies are required to study the underlying etiology in relation to ACPA.

The phenotype of C3 closely resembles the sign and symptoms as seen in patients infected by the parvo virus: moreover, these patients have the shortest symptom duration and are primarily seronegative. However, we did not observe an enrichment for parvovirus [36, 37], nor do we find an

enrichment of elicited tests which would be indicative of a suspicion by the clinician. Nonetheless, C3 may depict a form of self-remitting disease.

Our findings of 4 different clusters are based on the analysis of joint swelling at onset of disease. We have previously reported [38] that joint inflammation during the disease course tends to occur in the same joints, so it may be possible that the clusters we currently identified are stable over the disease course, but this has not been analysed yet.

One might wonder whether misdiagnosis of patients in our study influenced the results. For set A, we selected patients with a natural language algorithm, balancing Positive Predictive Value (PPV) against inclusiveness (both > 0.85).[23, 24] We reran the pipeline on the patients with a PPV for having RA (according to the physician) of 0.95 and higher. Except for an overall increase in ACPA prevalence and methotrexate initiation, we did not find differences in patient characteristics. Most notably, the differences between the clusters remained the same, as did the MTX failure and remission rate (*data not shown*).

Important to note is that we consider the physician's diagnosis as ground truth in our selection of patients. We moved away from the classification criteria as they were designed to create more homogeneous groups for the purpose of clinical trials, whereas we wanted to capture the full width of patients. Reassuringly, our patient selection did not differ too much from a cohort (set B) where all RA cases fulfilled the RA classification criteria and resembled many previous collections of early RA.[31, 32] This discovery was in line with our previous study, where we do not find large differences between different selection criteria [24].

Despite the robustness of our clusters, patients lie on a gradient (Fig. 2), rather than being segregated in clearly separable modules. The underlying structure that we identified could also be summarized into more or fewer clusters and the clusters might become clearer when more layers of information are added as well as the layer of time. We know from other studies that use molecular profiles from blood and synovial biopsies [14, 39], that adding gene expression patterns is also informative for subsetting RA. Unfortunately, with the limited data overlap, it is difficult to map these subsets to ours. For future research we aim to integrate pathology and expression patterns of cytokines with clinical data types. This might render increased insight into the taxonomy of RA.

Moreover, when such biological data are available, we can also identify the biological pathways driving these subsets. Another relevant step will be to explore how the clusters relate to the efficacy of alternative treatment strategies.

Our study design of using real-world data and unsupervised analyses put certain restraints on the conclusion about our data. Due to the real-world data, we cannot completely rule out that our clusters capture some noise, though we ruled out the factors which we expect to cause most of the noise. Secondly, we had to define MTX success based on adding or switching to a different drug. Probably some of the switches were due to side effects unrelated to efficacy. However, this would have resulted in a higher failure rate and particularly in the elderly cluster (C3) we observed less treatment switches. Finally, though we postulate that our clusters could reflect different aetiologic

subsets, we do not prove that. Our future studies will focus on collecting more biological material to substantiate this claim.

Within these limitations, we demonstrated the power of unsupervised- and data-driven tools to unlock hidden structures in the data. The integration of different EHR-modalities resulted in more stable clusters than clustering on categorical or numerical variables alone. In particular, the inclusion of joint involvement patterns appears to be a major axis of variation.

In conclusion, by clustering RA patients on their first presenting symptoms, we uncovered four phenotypes characterized by hand and feet involvement that associate with long-term treatment outcomes which were not part of the initial clustering. These cluster differences in outcome and clinical presentation may be indicative of a distinct etiology, though future research is required to affirm this potential biological link. Nonetheless, our data-driven approach offers a more granular picture of RA in the clinic, than the purely dichotomous division by age or ACPA.

References

1. Grassi W, De Angelis R, Lamanna G, Cervini C. The clinical features of rheumatoid arthritis. Eur J Radiol. 1998 May;27 Suppl 1:S18-24. doi: 10.1016/s0720-048x(98)00038-2. PMID: 9652497.
2. Heidari B. Rheumatoid Arthritis: Early diagnosis and treatment outcomes. Caspian J Intern Med. 2011 Winter;2(1):161-70. doi: 10.1016/s0720-048x(98)00038-2. PMID: 24024009; PMCID: PMC3766928.
3. Padyukov L. Genetics of rheumatoid arthritis. Semin Immunopathol. 2022 Jan;44(1):47-62. doi: 10.1007/s00281-022-00912-0. Epub 2022 Jan 27. PMID: 35088123; PMCID: PMC8837504
4. Syversen SW, Goll GL, van der Heijde D, et al. Prediction of radiographic progression in rheumatoid arthritis and the role of antibodies against mutated citrullinated vimentin: results from a 10-year prospective study. Annals of the Rheumatic Diseases 2010;69:345-351.5. doi: 10.1136/ard.2009.113092. Epub 2009 Jul 30. PMID: 19648126.
5. Matthijssen XME, Niemantsverdriet E, Huizinga TWJ, van der Helm–van Mil AHM Enhanced treatment strategies and distinct disease outcomes among autoantibody-positive and -negative rheumatoid arthritis patients over 25 years: A longitudinal cohort study in the Netherlands. PLOS Medicine 2020;17(9): e1003296. doi: 10.1371/journal.pmed.1003296
6. Pedersen M, Jacobsen S, Klarlund M, Pedersen BV, Wiik A, Wohlfahrt J, Frisch M. Environmental risk factors differ between rheumatoid arthritis with and without auto-antibodies against cyclic citrullinated peptides. Arthritis Res Ther. 2006;8(4):R133. doi: 10.1186/ar2022. PMID: 16872514; PMCID: PMC1779386.
7. Nordberg LB, Lillegraven S, Lie E and the ARCTIC working group, et al. Patients with seronegative RA have more inflammatory activity compared with patients with seropositive RA in an inception cohort of DMARD-naïve patients classified according to the 2010 ACR/EULAR criteria. Annals of the Rheumatic Diseases 2017;76:341-345. doi: 10.1136/annrheumdis-2015-208873
8. Sugihara T, Harigai M. Targeting Low Disease Activity in Elderly-Onset Rheumatoid Arthritis: Current and Future Roles of Biological Disease-Modifying Antirheumatic Drugs. Drugs Aging. 2016 Feb;33(2):97-107. doi: 10.1007/s40266-015-0341-2. PMID: 26833350; PMCID: PMC4756046.
9. Deal CL, Meenan RF, Goldenberg DL , et al. The clinical features of elderly-onset rheumatoid arthritis. A comparison with younger-onset disease of similar duration. Arthritis Rheum. 1985 Sep;28(9):987-94. doi: 10.1002/art.1780280905. PMID: 4038365.
10. Sugihara T, Harigai M. Targeting Low Disease Activity in Elderly-Onset Rheumatoid Arthritis: Current and Future Roles of Biological Disease-Modifying Antirheumatic Drugs. Drugs Aging. 2016 Feb;33(2):97-107. doi: 10.1007/s40266-015-0341-2. PMID: 26833350; PMCID: PMC4756046.
11. Ahlqvist E, Storm P, Käräjämäki A et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol. 2018 May;6(5):361-369. doi: 10.1016/S2213-8587(18)30051-2. Epub 2018 Mar 5. PMID: 29503172.
12. Howrylak JA, Fuhlbrigge AL, Strunk RC, Zeiger RS, Weiss ST, Raby BA; Childhood Asthma Management Program Research Group. Classification of childhood asthma phenotypes and long-term clinical responses to inhaled anti-inflammatory medications. J Allergy Clin Immunol. 2014 May;133(5):1289-300, 1300.e1-12. doi: 10.1016/j.jaci.2014.02.006. PMID: 24892144; PMCID: PMC4047642.

13. Angelini F, Widera P, Mobasheri A, et al. Osteoarthritis endotype discovery via clustering of biochemical marker data. Ann Rheum Dis. 2022 May;81(5):666-675. doi: 10.1136/annrheumdis-2021-221763. Epub 2022 Mar 4. PMID: 35246457.
14. Lewis MJ, Barnes MR, Blighe K, et al. Molecular Portraits of Early Rheumatoid Arthritis Identify Clinical and Treatment Response Phenotypes. Cell Rep. 2019 Aug 27;28(9):2455-2470.e5. doi: 10.1016/j.celrep.2019.07.091. PMID: 31461658; PMCID: PMC6718830
15. Lliso-Ribera G, Humby F, Lewis M, et al. Synovial tissue signatures enhance clinical classification and prognostic/treatment response algorithms in early inflammatory arthritis and predict requirement for subsequent biological therapy: results from the pathobiology of early arthritis cohort (PEAC). Ann Rheum Dis. 2019 Dec;78(12):1642-1652. doi: 10.1136/annrheumdis-2019-215751. Epub 2019 Oct 2. PMID: 31582377; PMCID: PMC6900253.
16. Jung SM, Park KS, Kim KJ. Clinical phenotype with high risk for initiation of biologic therapy in rheumatoid arthritis: a data-driven cluster analysis. Clin Exp Rheumatol. 2021 Nov-Dec;39(6):1282-1290. doi: 10.55563/clinexprheumatol/24zas6. Epub 2021 Feb 9. PMID: 33635214.
17. Curtis JR, Weinblatt M, Saag K, et al. Data-Driven Patient Clustering and Differential Clinical Outcomes in the Brigham and Women's Rheumatoid Arthritis Sequential Study Registry. Arthritis Care Res (Hoboken). 2021 Apr;73(4):471-480. doi: 10.1002/acr.24471. Epub 2021 Mar 13. PMID: 33002337; PMCID: PMC8048846.
18. Merola JF, Espinoza LR, Fleischmann R. Distinguishing rheumatoid arthritis from psoriatic arthritis. RMD Open. 2018 Aug 13;4(2):e000656. doi: 10.1136/rmdopen-2018-000656. PMID: 30167326; PMCID: PMC6109814.
19. Onken A, Yague JG, Sakata S. Deep multimodal autoencoders for identifying latent representations of spike counts and local field potentials. Bernstein Conference 2018. doi: 10.12751/nncn.bc2018.0223
20. Steinbach M, Ertöz L, Kumar V. The Challenges of Clustering High Dimensional Data. In: Wille LT, editor. New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2004. p. 273–309. Available from: https://doi.org/10.1007/978-3-662-08968-2_16
21. Cai, J., Fan, J., Guo, W., et al. Efficient deep embedded subspace clustering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022 (pp. 1-10).
22. Castela Forte J, Yeshmagambetova G, van der Grinten ML, et al. Identifying and characterizing high-risk clusters in a heterogeneous ICU population with deep embedded clustering. Sci Rep. 2021 Jun 8;11(1):12109. doi: 10.1038/s41598-021-91297-x. PMID: 34103544; PMCID: PMC8187398.
23. Maarseveen TD, Meinderink T, Reinders MJT, et al. Machine Learning Electronic Health Record Identification of Patients with Rheumatoid Arthritis: Algorithm Pipeline Development and Validation Study. JMIR Med Inform. 2020 Nov 30;8(11):e23930. doi: 10.2196/23930. PMID: 33252349; PMCID: PMC7735897.
24. Maarseveen TD, Maurits MP, Niemantsverdriet E, van der Helm-van Mil AHM, Huizinga TWJ, Knevel R. Handwork vs machine: a comparison of rheumatoid arthritis patient populations as identified from EHR free-text by diagnosis extraction through machine-learning or traditional criteria-based chart review. Arthritis Res Ther. 2021 Jun 22;23(1):174. doi: 10.1186/s13075-021-02553-4. PMID: 34158089; PMCID: PMC8218515.
25. de Rooy DP, van der Linden MP, Knevel R, Huizinga TW, van der Helm-van Mil AH. Predicting arthritis outcomes--what can be learned from the Leiden Early Arthritis Clinic? Rheumatology

(Oxford). 2011 Jan;50(1):93-100. doi: 10.1093/rheumatology/keq230. Epub 2010 Jul 16. PMID: 20639266.
26. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir el-AD, Tadmor MD, Litvin O, Fienberg HG, Jager A, Zunder ER, Finck R, Gedman AL, Radtke I, Downing JR, Pe'er D, Nolan GP. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell. 2015 Jul 2;162(1):184-97. doi: 10.1016/j.cell.2015.05.047. Epub 2015 Jun 18. PMID: 26095251; PMCID: PMC4508757.
27. Roth, A. E (Ed.). The Shapley Value: Essays in Honor of Lloyd S. Shapley. Cambridge: Cambridge University Press. 1988. doi:10.1017/CBO9780511528446
28. Cox, D. R. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 1972; 34(2), 187-202.
29. Schoenfeld, D. Chi-squared goodness-of-fit tests for the proportional hazards regression model. Biometrika, 1980; 67(1), 145-153.
30. Maarseveen TD. EHR Clustering RA. GitHub. Published: August 8, 2022. Link: https://github.com/levrex/EHR-Clustering-RA
31. Nishimura K, Sugiyama D, Kogata Y, et al. Meta-analysis: Diagnostic Accuracy of Anti–Cyclic Citrullinated Peptide Antibody and Rheumatoid Factor for Rheumatoid Arthritis. Ann Intern Med.2007;146:797-808. [Epub 5 June 2007]. doi:10.7326/0003-4819-146-11-200706050-00008
32. van Nies JAB, Tsonaka R, Gaujoux-Viala C, et al. Evaluating relationships between symptom duration and persistence of rheumatoid arthritis: does a window of opportunity exist? Results on the Leiden Early Arthritis Clinic and ESPOIR cohorts. Annals of the Rheumatic Diseases 2015;74:806-812.
33. Visser H, le Cessie S, Vos K, Breedveld FC, Hazes JM. How to diagnose rheumatoid arthritis early: a prediction model for persistent (erosive) arthritis. Arthritis Rheum. 2002 Feb;46(2):357-65. doi: 10.1002/art.10117. PMID: 11840437.
34. van der Helm-van Mil AH, Verpoort KN, Breedveld FC, Toes RE, Huizinga TW. Antibodies to citrullinated proteins and differences in clinical progression of rheumatoid arthritis. Arthritis Res Ther. 2005;7(5):R949-58. doi: 10.1186/ar1767. Epub 2005 Jun 14. PMID: 16207336; PMCID: PMC1257421.
35. Aletaha D, Neogi T, Silman AJ, et al. 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. PMID: 20872595.
36. Takahashi Y, Murai C, Shibata S, et al. Human parvovirus B19 as a causative agent for rheumatoid arthritis. Proc Natl Acad Sci U S A. 1998;95(14):8227-8232. doi:10.1073/pnas.95.14.8227
37. Naides SJ, Scharosch LL, Foto F, Howard EJ. Rheumatologic manifestations of human parvovirus B19 infection in adults. Initial two-year clinical experience. Arthritis Rheum. 1990 Sep;33(9):1297-309. doi: 10.1002/art.1780330901. PMID: 2169746.
38. Heckert SL, Bergstra SA, Matthijssen XME, et al. Joint inflammation tends to recur in the same joints during the rheumatoid arthritis disease course. Ann Rheum Dis. 2022 Feb;81(2):169-174. doi: 10.1136/annrheumdis-2021-220882. Epub 2021 Aug 30. PMID: 34462262.
39. Zhang F, Jonsson AH, Nathan A, et al. Cellular deconstruction of inflamed synovium defines diverse inflammatory phenotypes in rheumatoid arthritis. bioRxiv. 2022 Jan 1;2022.02.25.481990.

Footnotes

Contributors

RK developed the study design together with EB, TM and MP. TM ran the cluster analysis. BBdK replicated the survival analysis in the validation set. All authors contributed to the interpretation of the results, and provided ideas for further downstream analysis. AHM annotated the classification criteria. RK and TM drafted the first version of the manuscript. All authors reviewed and approved the final draft of the manuscript to be submitted.

Funding

This project has received funding from two large European Union's Horizon grants for Europe research and innovation. First for SQUEEZE (activity No. 101095052) and secondly for SPIDERR (activity No. 101080711). There was additional financial support from the ZonMW Klinische Fellow No. 40-00703-97-19069, as well as the Zonmw Open Competitie, No. 09120012110075.

Competing interest

The authors declare no competing interests.

Patient and public involvement

Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Ethics approval

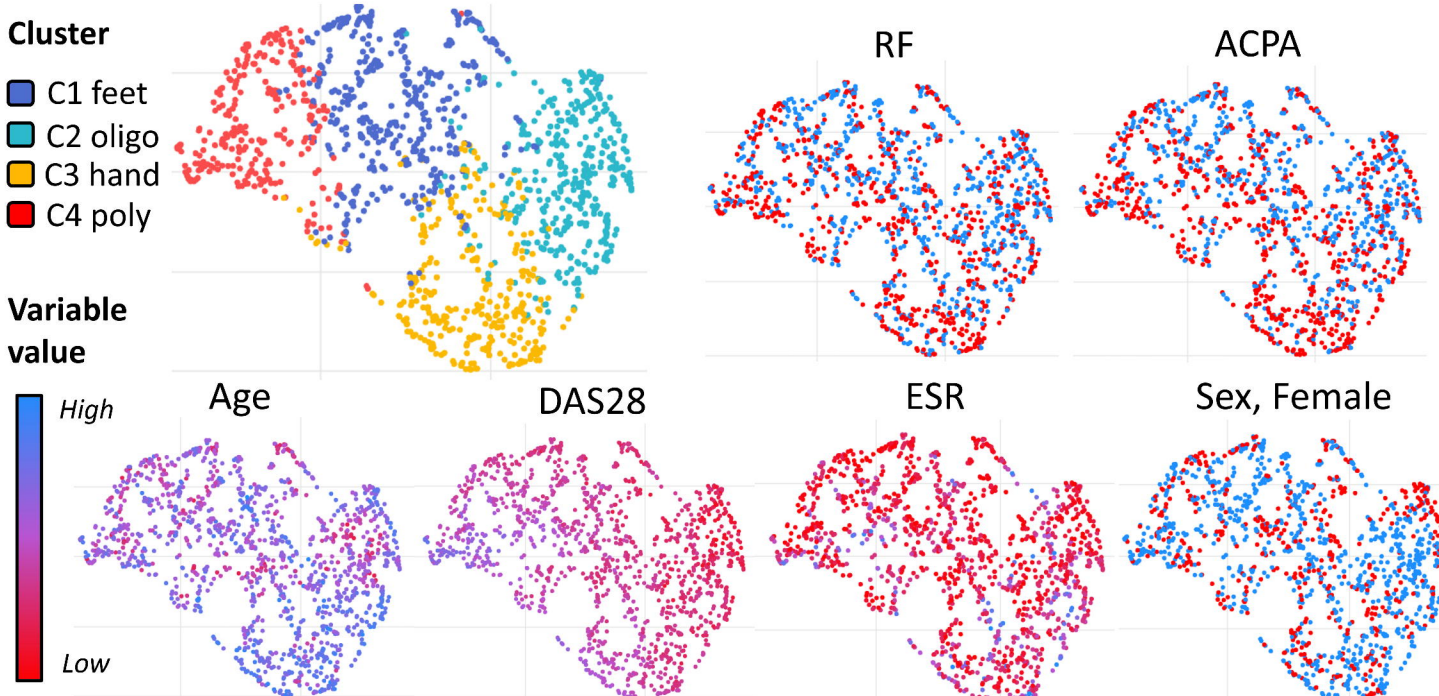
We received ethical approval from the Medical Ethics Committee (METC) at Leiden University Medical Center according to study protocol B18.057.

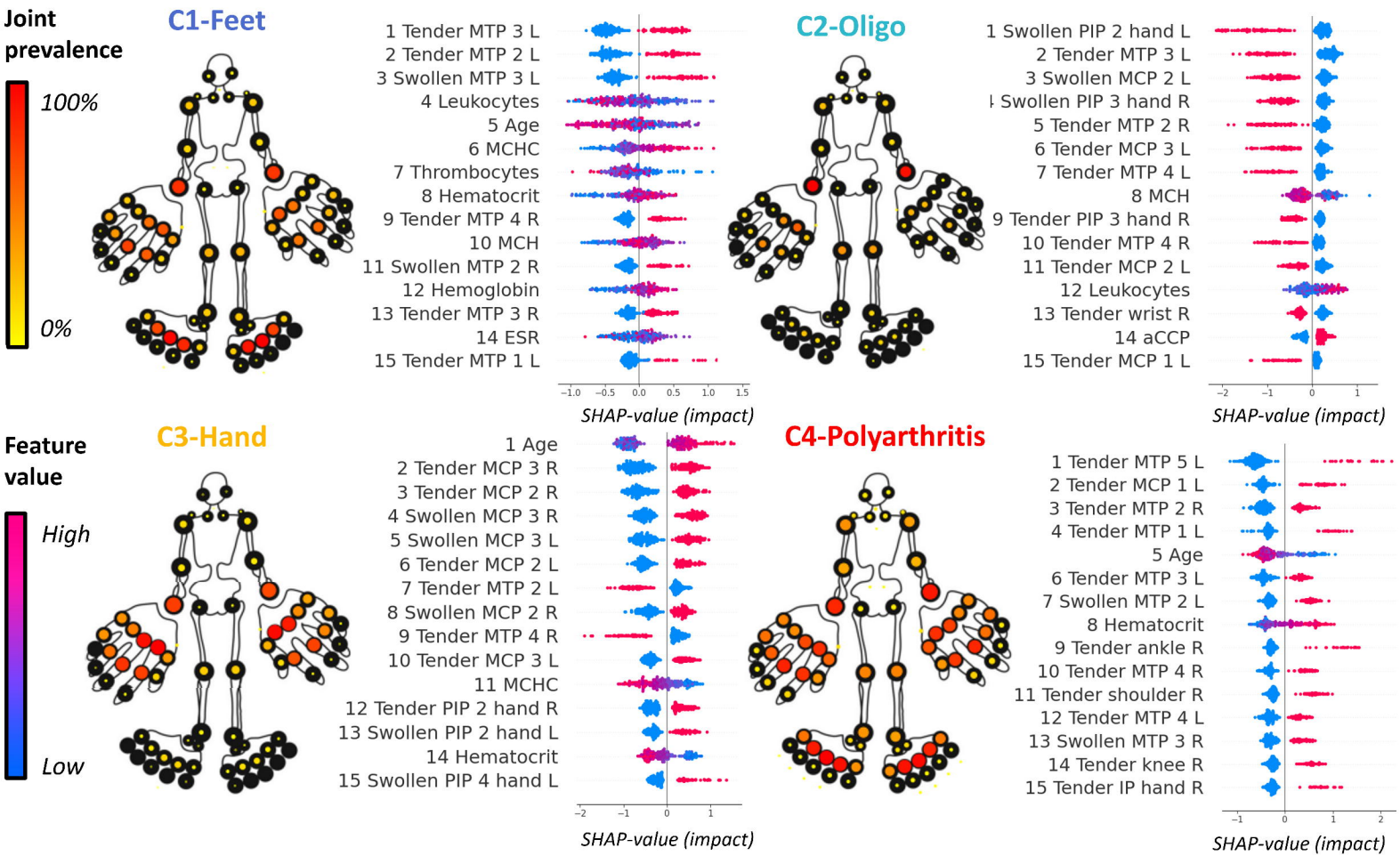
Data availability statement

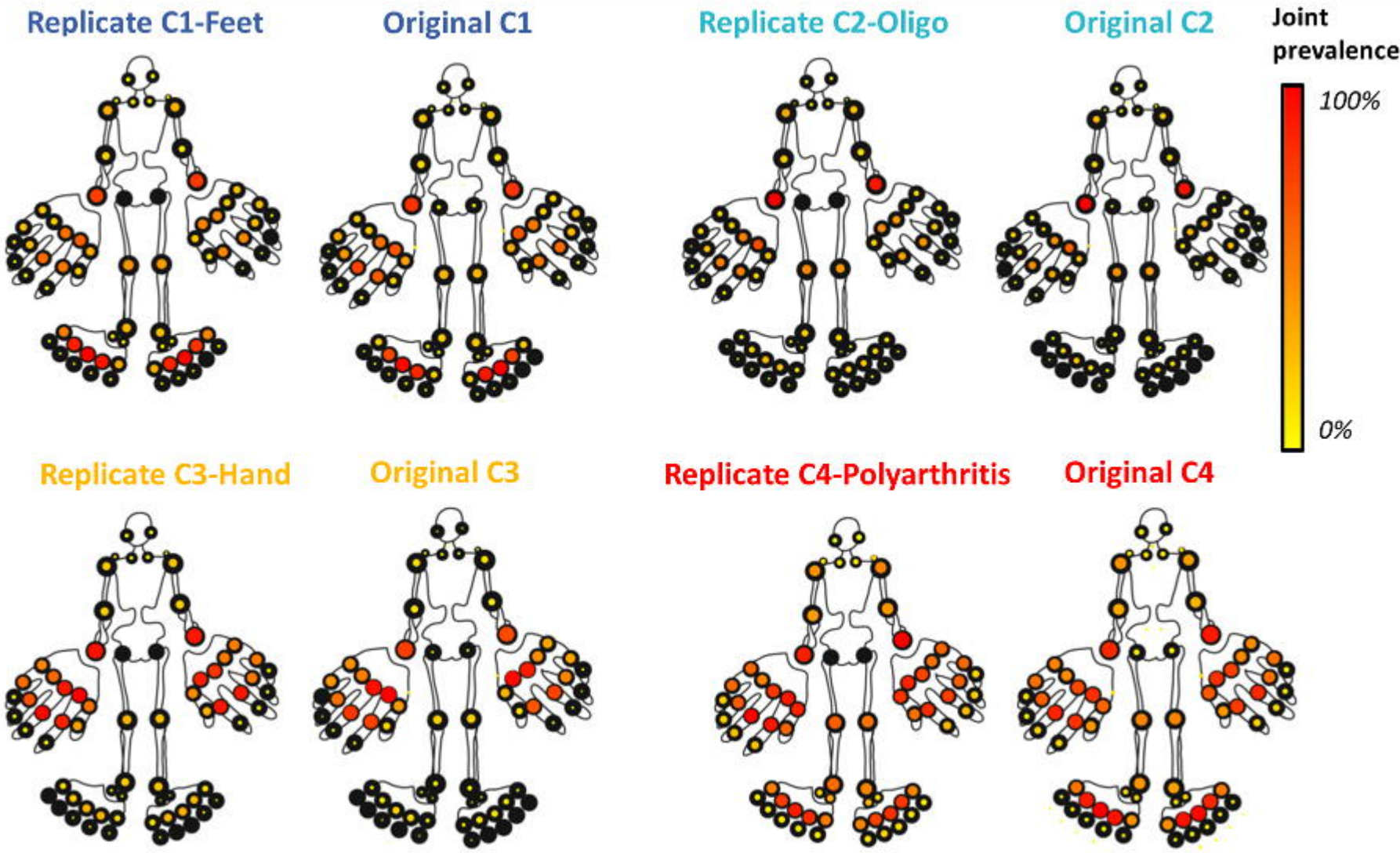
We have made our scripts available in a public repository at GitHub [30]. Study data is available upon reasonable request

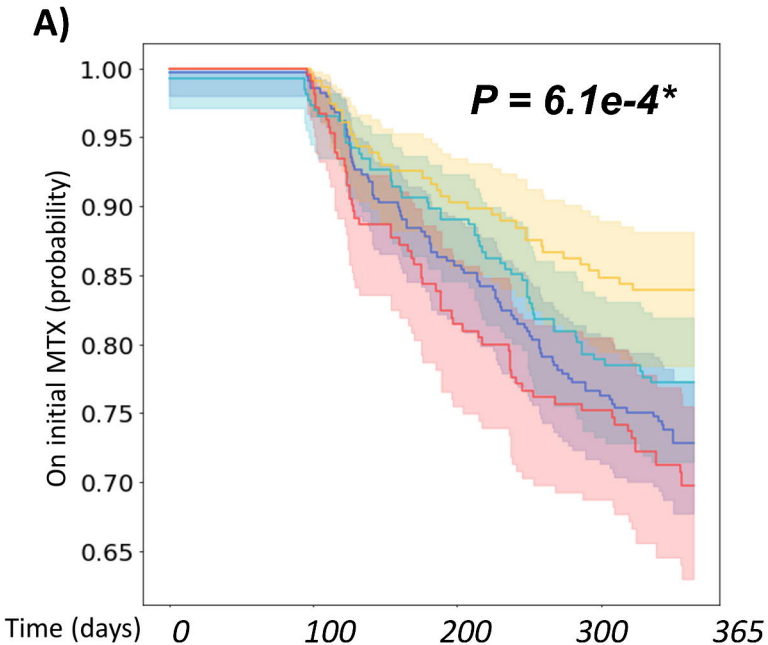
Acknowledgements

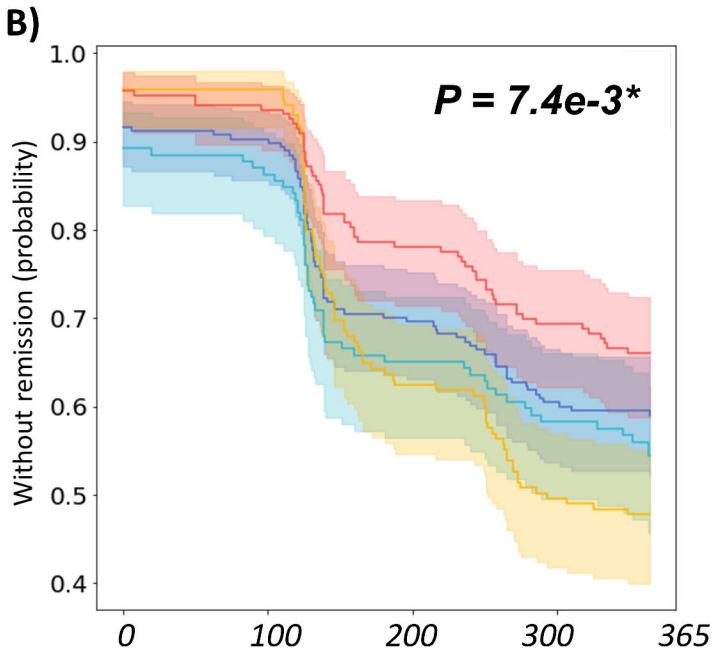
We thank Samantha Jurado-Zapata for her help with respect to the extraction and processing of the Electronic Health Record data from the Leiden University Medical Center


Cluster
C1 feet
C2 oligo
C3 hand
C4 poly
RF
ACPA
Variable value
High
Low
Age
DAS28
ESR
Sex, Female


Joint prevalence
100%
0%
C1-Feet
1 Tender MTP 3 L
2 Tender MTP 2 L
3 Swollen MTP 3 L
4 Leukocytes
5 Age
6 MCHC
7 Thrombocytes
8 Hematocrit
9 Tender MTP 4 R
10 MCH
11 Swollen MTP 2 R
12 Hemoglobin
13 Tender MTP 3 R
14 ESR
15 Tender MTP 1 L
−1.0 −0.5 0.0 0.5 1.0 1.5
SHAP-value (impact)
C2-Oligo
1 Swollen PIP 2 hand L
2 Tender MTP 3 L
3 Swollen MCP 2 L
4 Swollen PIP 3 hand R
5 Tender MTP 2 R
6 Tender MCP 3 L
7 Tender MTP 4 L
8 MCH
9 Tender PIP 3 hand R
10 Tender MTP 4 R
11 Tender MCP 2 L
12 Leukocytes
13 Tender wrist R
14 aCCP
15 Tender MCP 1 L
−2 −1 0 1
SHAP-value (impact)
Feature value
High
Low
C3-Hand
1 Age
2 Tender MCP 3 R
3 Tender MCP 2 R
4 Swollen MCP 3 R
5 Swollen MCP 3 L
6 Tender MCP 2 L
7 Tender MTP 2 L
8 Swollen MCP 2 R
9 Tender MTP 4 R
10 Tender MCP 3 L
11 MCHC
12 Tender PIP 2 hand R
13 Swollen PIP 2 hand L
14 Hematocrit
15 Swollen PIP 4 hand L
−2 −1 0 1
SHAP-value (impact)
C4-Polyarthritis
1 Tender MTP 5 L
2 Tender MCP 1 L
3 Tender MTP 2 R
4 Tender MTP 1 L
5 Age
6 Tender MTP 3 L
7 Swollen MTP 2 L
8 Hematocrit
9 Tender ankle R
10 Tender MTP 4 R
11 Tender shoulder R
12 Tender MTP 4 L
13 Swollen MTP 3 R
14 Tender knee R
15 Tender IP hand R
−1 0 1 2
SHAP-value (impact)

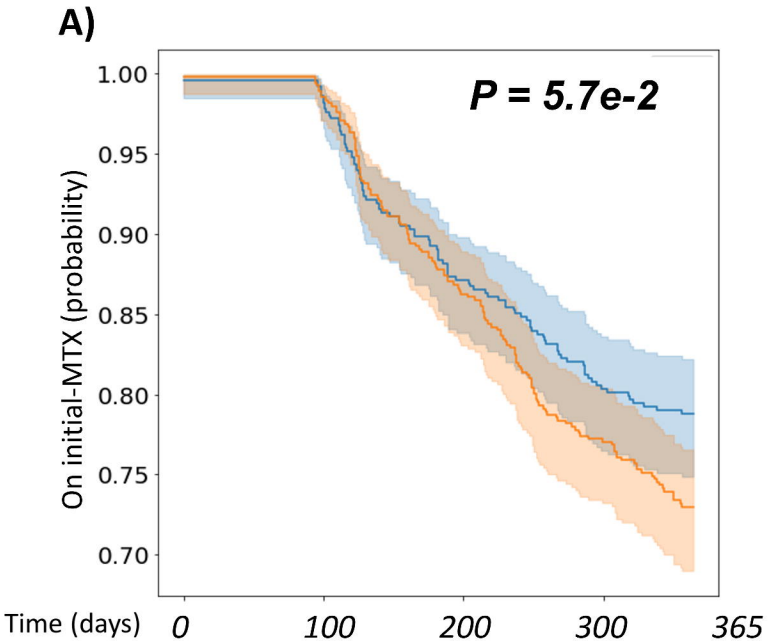

Replicate C1-Feet
Original C1
Replicate C2-Oligo
Original C2
Joint prevalence
100%
0%
Replicate C3-Hand
Original C3
Replicate C4-Polyarthritis
Original C4



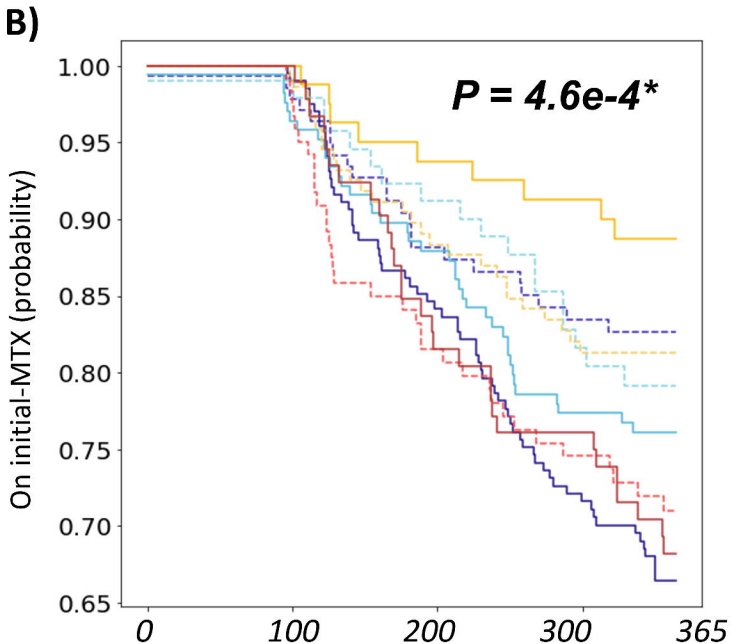


NUMBER OF PATIENTS					
C1-Feet	353	335	284	246	233
C2-Oligo	273	251	222	190	181
C3-Hand	240	228	200	186	177
C4-Poly	218	211	169	153	139
Time	0	100	200	300	365

NUMBER OF PATIENTS					
C1-Feet	208	201	154	131	127
C2-Oligo	124	118	88	77	72
C3-Hand	165	163	102	81	74
C4-Poly	180	176	144	127	118
Time	0	100	200	300	365



NUMBER OF PATIENTS					
ACPA-	522	491	415	372	352
ACPA+	562	534	460	403	374



NUMBER OF PATIENTS					
C1 ACPA+	142	134	115	106	103
C1 ACPA-	211	201	169	140	129
C2 ACPA+	99	91	80	67	63
C2 ACPA-	174	160	142	123	117
C3 ACPA+	156	147	126	114	108
C3 ACPA-	84	81	74	72	69
C4 ACPA+	125	119	94	85	78
C4 ACPA-	93	92	75	68	59