

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

Tjardo Maarseveen

Leiden University Medical Center: Leids Universitair Medisch Centrum <https://orcid.org/0000-0003-1701-1040>

Marc Maurits

Leiden University Medical Center: Leids Universitair Medisch Centrum

Ellis Niemantsverdriet

Leiden University Medical Center: Leids Universitair Medisch Centrum

Annette van der Helm - van Mil

Leiden University Medical Center: Leids Universitair Medisch Centrum

Thomas W.J. Huizinga

Leiden University Medical Center: Leids Universitair Medisch Centrum

Rachel Knevel (✉ R.Knevel@lumc.nl)

Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands
<https://orcid.org/0000-0002-7494-3023>

Research article

Keywords: rheumatoid arthritis, machine learning algorithms, EHR, electronic health records, artificial intelligence, classification criteria, big data, observational research, chart review

Posted Date: April 12th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-394821/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background

Electronic Health Records (EHRs) offer a wealth of observational data. Machine-Learning (ML) methods are efficient at data extraction, capable of processing the information-rich free-text physician notes in EHRs. The clinical diagnosis contained therein represents physician expert opinion and is more consistently recorded than classification criteria components.

Objectives

To investigate the overlap and differences between Rheumatoid Arthritis patients as identified either from EHR free-text through extraction of the rheumatologist diagnosis using machine-learning (ML) or through manual chart-review applying the 1987 and 2010 RA classification criteria.

Methods

Since EHR initiation, 17,662 patients visited the Leiden outpatient clinic. For ML, we used a Support Vector Machine (SVM) model to identify those who were diagnosed with RA by their rheumatologist. We trained & validated the model on a random selection of 2,000 patients, balancing PPV and Sensitivity to define a cutoff, and assessed performance on a separate 1,000 patients. We then deployed the model on our entire patient selection (including the 3,000). Of those, 1,212 patients had both a 1987 and 2010 EULAR/ACR criteria status at one year after inclusion into the local prospective arthritis cohort. In these 1,212 patients we compared the patient characteristics of RA cases identified with ML and those fulfilling the classification criteria.

Results

The ML model performed very well in the independent test set (sensitivity=0.85, specificity=0.99, PPV=0.86, NPV=0.99). In our selection of patients with both EHR and classification information, 406 were recognized as RA by ML and 386 and 457 fulfilled the 1987 or 2010 criteria respectively. Eighty percent of the ML-identified cases fulfilled at least one of the criteria sets.

Both demographic and clinical parameters did not differ between the ML extracted cases and those identified with EULAR/ACR classification criteria.

Conclusions

With ML methods we enable fast patient extraction from the huge EHR resource. Our ML algorithm accurately identifies patients diagnosed with RA by their rheumatologist. This resulting group of RA patients had a strong overlap with patients identified using the 1987 or 2010 classification criteria and the baseline (disease) characteristics were comparable. ML assisted case labelling enables high-throughput creation of inclusive patient selections for research purposes.

Introduction

Electronic Health Records (EHRs) contain a vast amount of observational data. Manual review of these data is time consuming and laborious, hampering the usability of the data. Advancements in Natural Language Processing and Machine Learning (ML) methods have created great potential for processing format-free text data such as present in EHRs [1, 2]. These EHR entries contain the prosaic conclusion of the treating physician, ranging from elaborate discussion of lab results to listed differential diagnoses. This unstructured nature of the records makes them hard to query using simple text matching. Machine learning methods can deduce patterns from a set of training examples, without requiring any domain specific knowledge. The algorithm does not count specific criteria, but rather identifies discriminatory features by learning from an annotated outcome. We have previously developed a pipeline for the identification of rheumatoid arthritis (RA) in format-free text from clinical notes of the rheumatology clinic.[3] The machine learning pipeline generates algorithms (i.e. classifiers) by which we can extract the RA patients from 17,662 patient records in merely 5 seconds with high confidence. These algorithms recruit words and chunks in the text as features to identify the RA-cases with respect to the final diagnosis of the rheumatologist which is treated as the true label.[4] Hereby we ensured capturing even patients diagnosed with RA, for whom the rheumatologist did not register the components of the 1987 or 2010 EULAR/ACR RA criteria.

Now the question arises whether the patient selection of our high-throughput machine learning approach, which is steered by rheumatologist's diagnosis, differs from the traditional manual chart review, which uses classification criteria as golden standards.

In the current descriptive study, we compared the patients selected by our machine learning pipeline to patients selected through (traditional) manual chart review applying the ACR/EULAR 1987 and 2010 criteria.

Method

Patients

We retrieved the records of all patients who visited the Leiden University Medical Center (LUMC) rheumatology clinic between initiation of the EHR system in 2011 and 2019

(n = 17,662). From these dossiers, we used the *Conclusion* section of the physician's notes, which describes the symptoms and differential diagnosis. Using a Support Vector Machine (SVM) model we built a classifier, capable of extracting the rheumatologist's RA diagnosis from these unstructured data. We randomly selected 3,000 patients and asked a rheumatologist to review the medical records and annotate the patients based on the diagnosis of their own rheumatologist after one year of follow-up. To develop our ML model we created two distinct dataset: 2,000 patients were selected to train and validate the model and the remaining 1,000 patients were set apart as an independent test set to evaluate model performance.

The SVM identifies these diagnoses by finding the optimal boundary (hyperplane) separating the different classes (RA and non-RA) using both individual words and chunks of sentences as features. It employs the kernel trick, where it maps the samples into a higher dimensional space in order to find the hyperplane.[5] The output of the classifier is a score from 0–1, where scores represent the likelihood of a patient having RA. The cutoff for binarization (RA yes/no) can be tweaked depending on the need for a particularly precise or sensitive patient selection.

A subset of the patients from the EHR were also included in an observational cohort. Patients presenting at the LUMC rheumatology clinic with arthritis are asked to participate in this cohort, which comprehensively registers a wide variety of medical data.[6] In 1993 this population-based prospective cohort started collecting patient information every three months in the first year of patient follow-up and after that yearly. Inclusion took place when arthritis was confirmed at physical examination and symptom duration was < 2 years. The final diagnosis was obtained at one year follow-up by manual chart review of one rheumatologist who counted the 1987 and 2010 EULAR/ACR criteria in patients that were diagnosed with RA [7, 8].

Ethical consent was obtained from the ethics committee of the LUMC before initiation of the study.

So for the current study, we created a dataset that contained patients who visited the outpatient clinic for the first time between 2011 and 2019 and who were manually checked for fulfilling the 2010 and 1987 classification criteria as part of their enrollment in our early arthritis cohort. To this set of patients we applied three methods to identify RA: the SVM model extracting the rheumatologist's diagnosis, the 1987 classification criteria and the 2010 classification criteria.

Statistical analyses

We describe the model which resulted from training an SVM algorithm on the medical records of 2,000 randomly selected patients in previously published work.[4] The final threshold for ML case identification was set by optimizing the trade-off between positive predictive value (PPV) and sensitivity in the training set. To test whether the model was robust, we evaluated the performance of the SVM-derived classifier (sensitivity, specificity, PPV and negative predictive value (NPV)) using the diagnosis of the treating rheumatologists as gold standard. In addition, we visualized the performance by rendering an ROC and PR curve with Scikit-learn package v0.21.2 in Python v3.5.[9].

Next, we examined the extent of overlap between the patients identified with the machine learning classifier and those identified using the criteria approach. The relationship between the cohorts was visualized in an upset plot with the R UpSetR package v1.4.0 in R v4.0.2 [10]. Finally, we compared the baseline demographics and disease characteristics of the different RA-case selections using Pearson chi-squared and Mann-Whitney U tests ($\alpha = 0.05$).

Figure 1 describes the flow of case selection and tests of our study.

Results

Performance of ML in identifying the RA cases as diagnosed by the treating rheumatologist

The flexible nature of the SVM binarization cutoff (RA yes/no) enables us to choose a very precise, very sensitive or a balanced approach to the performance of the algorithm (Table 1 and Figure S1). To make sure we find the largest number of definite cases, we took a balanced approach between PPV and sensitivity of the SVM, which resulted in a probability cutoff of 0.83 based on our training data. We then applied this cutoff to the independent set of 1,000 annotated patients. In this set, the SVM based ML classifier had an AUC-ROC and AUC-PRC of 0.97 and 0.90 respectively (Fig. 2). The classifier performed very well at identifying patients that were diagnosed with RA by their rheumatologists': sensitivity 0.85, specificity 0.99, PPV, 0.86, NPV 0.99 (Table 1). The most discriminatory features that contribute to the SVM's decision can be found in previously published work [4].

Extent of overlap between machine learning and criteria based selections

A total of 17,662 novel patients visited the Leiden outpatient clinic since the EHR initiation in 2011. In this set, the ML identified 1,508 patients with a diagnosis of RA by their rheumatologist after one year of follow-up. In the same period, the prospective cohort included 1,376 patients with early arthritis. Patients in whom the 2010 and 1987 criteria were not assessed at all were excluded, leaving 1,212 patients for this paper's analyses (Fig. 1).

To visualize the overlap of the ML defined RA cases to the 2010 and 1987 RA criteria selections, we rendered an upset plot (Fig. 3). In our set of 1,212 patients with both EHR data and criteria based annotation 583 unique RA cases were identified. Of these, 406 (69.6%) were identified by our ML as having RA. In the same set, 457 (78.4%) fulfilled the 2010 criteria and 386 (66.2%) the 1987 criteria. The overlap between the different selection methods was substantial: 254 (43.6%) were identified with all three methods, and an additional 94 (16.1%) were identified by both ML and one of the classification criteria (56 (9.6%) and 38 (6.5%) for 2010 and 1987 respectively). The ML identified 58 (9.9%) patients for whom all classification criteria were assessed, but who were negative on both sets, whereas 84 (14.4%) and 31 (5.3%) patients met a single classification criteria set (2010 or 1987 respectively) and were not identified by the ML. A final group of 63 (10.8%) patients met both classification criteria but not the ML cutoff. The ML defined set had slightly more overlapping patients with the 2010 criteria than the 1987 criteria (310 (53.2%) and 292 (50.1%) respectively).

Demographic and baseline differences in machine learning and criteria based selections

In Table 2 we compared the baseline characteristics of the RA cases identified by ML to the patients fulfilling the two sets of criteria. The group of patients that was diagnosed with RA by their rheumatologists had the same median age (57.4), DAS44 at baseline (2.8), prevalence of women (64%), anti-CCP-positivity (52%) and RF-positivity (57%) as patients selected based on fulfilling the 2010 or 1987 classification criteria. We found no statistically significant differences between the three groups.

Description of patients exclusively found by either the ML or criteria

To further elucidate the cases exclusively identified by the ML and those exclusively identified by the criteria, we investigated the baseline characteristics for these subgroups as well (Table 3). The ML identified 58 patients who were not found by the criteria. This group had an abundance of seronegative scoring patients, with a CCP-positivity of 6% and a RF-positivity of 19% respectively. The criteria-based approach identified 178 patients that were not found by the ML. The majority of cases that were only found by the criteria were also anti-CCP2- and RF negative: 16% and 34% respectively. There were no clear differences with regard to other patient characteristics.

Upset & Baseline table for different cutoffs

In addition to the balanced cutoff of the ML probability, we studied the effect of a more stringent and a more lenient cutoff. The ML with the stringent cutoff (0.99) was, as expected, much more precise, but less sensitive (Table 1). With this cutoff the ML identified 303 patients (Table S1), 209 of those overlapped with both 1987 and 2010 criteria selections (Figure S2). This group of patients had a similar age (57.6), prevalence of women (65%) and RF-positivity (62%) as the criteria-based selections. The anti-CCP-positivity prevalence (58%) was substantially higher compared to both the 1987 ($P= 0.014$) and 2010 criteria ($P= 0.013$).

With the lenient cutoff of 0.53 the ML was very sensitive but less precise (Table 1). Here, we identified 466 patients (Table S2) of which 266 patients fulfilled both criteria (Figure S3). The group of ML-identified cases maintained a similar prevalence of women (64%), RF-positivity (53%) and anti-CCP-positivity (47%) as those who fulfilled one or both of the two classification criteria. We did, however, find substantial differences in the disease characteristics. The median number of swollen joints (5) was significantly lower with respect to the 1987 criteria-based selection ($P= 0.022$), whereas the median DAS44 of 2.7 was comparatively low in contrast to the 2010 criteria-based selection ($P= 0.049$).

Discussion

Our study describes the production and validation of a robust machine learning model that extracts high quality patient collections from free written EHR data. By extracting the diagnosis of the treating rheumatologist, we are able to confidently classify patients as RA cases even when information on classification components is missing. Our method is both fast and efficient (3,326 complete medical records per second) and creates a highly similar case selection to criteria based chart review.

Traditionally, researchers use the classification criteria for the creation of their datasets in order to select a homogenous patient cohort. While these criteria are rigorously validated and generally accepted by the community, they are by no means a replacement of the clinical judgment of a rheumatologist [11, 12, 14]. Defining what constitutes a true RA case has always been the prerogative of the expert community of rheumatologists. The aim of both the 1987 and the 2010 criteria is the inclusion of said cases, with a preference for more stringent selection over larger, more heterogeneous cohorts. As a result, the general

population of RA patients extends beyond the scope of the classification criteria. Many research ventures (e.g. drug trials) align with the inclusion of criteria positive patients (high specificity), as their effects should not be diluted by noisy patient populations. However, when we are interested in investigating or redefining the entire disease entity of RA, a broad inclusion is much preferred (high sensitivity). In these latter instances, the rheumatologist diagnosis is the ideal balance between inclusiveness and precision. The EHR format-free text fields filled out by the treating physician are the most extensive collection of these diagnoses, but these are very time and resource consuming to manually peruse. By using high throughput machine learning approaches to crunch outpatient clinic EHRs, it is feasible to classify thousands of patients based on their rheumatologist's diagnosis in mere seconds.

An additional advantage of using EHR data for science is that we solely ask patients for access to their records. No additional visits are required. Thereby we can also include patients who would be unable or who would refuse to enter a research cohort or who missed the inclusion because it was not offered to them. Between 2011 and 2019, 17,662 patients visited the Leiden outpatient clinic. We identified 1,508 that were diagnosed with RA. Of those, 406 were included in the cohort. This higher patient number could increase study power. However, while the ML method increases sensitivity, it slightly reduces precision when one considers fulfilling classification criteria the golden standard. Though manual chart review searching for individual criteria will also have its imprecisions, it might hit a better balance between precision and sensitivity than our probabilistic approach. Furthermore, there is great benefit to data that is registered into the consistently structured prospective cohorts by specialized research personnel. On the other hand, EHRs often contain a larger quantity and higher variety of data whose collection is not constrained by a specific study design. Using these data for research purposes will require stringent data curation. Fortunately, natural language processing and machine learning make the wealth of (noisy) EHR data more accessible than ever. Regardless, depending on one's research question, either structured cohorts or EHRs will be more suited for data collection.

We show that our ML constructed patient selection is highly comparable to the patient groups meeting 1987 and 2010 criteria. Nevertheless, we also identify 2 groups of single positive patients; those who are diagnosed with RA by their rheumatologist, but who do not meet either of the criteria sets, as well as those who meet one or both of the criteria, without officially being diagnosed. When looking at the clinical characteristics of these groups, it becomes apparent that these are both composed of patients with for example a much lower prevalence of anti-CCP and RF. The single positives therefore seem quite divergent from the double positives and those caught extra by one method, seem to balance out well with those missed according to the other.

One might think that ML methods are not truly required to deal with the extensive information in the EHR data; the use of standardized billing codes for the purpose of case identification requires only simple queries. However, previous work has shown a tendency for these codes to identify false positives.[13] The true diagnosis of a treating rheumatologist is undoubtedly more reliable. We show that it is completely feasible to distil this high-quality label from the same EHR data, using a robust ML model.

The expertise required to implement an ML model might seem daunting in comparison to classical approaches such as chart review. The field of bioinformatics is rapidly expanding, and the ever-increasing offer of modelling techniques can come across as overly complicated. However, in previously published work we introduce and validate a user-friendly pipeline for the construction of an ML model tailored to individual healthcare centers.[4] By relieving the need for ML specific knowledge, we enable everyone with access to an annotated set of cases to create classifiers on par with the SVM model presented here.[3]

Conclusion

Using ML methods to extract the physician opinion from free written text, recorded as part of standard clinical care, allows for sensitive collection of cases with clinical manifestations similar to traditional, criteria based, selection of patients. This approach for high throughput identification of disease case selections will be invaluable in research into the larger entity of a disease. Including the widest range of trustworthy cases is crucial when, for example, looking into novel patient subgroups within a disease or when identifying novel risk and susceptibility factors for complex illnesses.

Limitations

In our study, we compare the ML extracted rheumatologist's diagnosis with the criteria based RA identification in one center. It remains unknown whether rheumatologists at other centers would have selected a similar RA population. For future international multi-center research we would need to compare the differences in the patient characteristics of different centers and investigate the effect of EHR language on the model performance. The pipeline employed here for the construction of the ML model is language independent and has been previously shown to perform well in both Dutch and German. Further studies into the consistency of our findings go beyond the scope of the current study.

While we emphasize the very sensitive nature of our approach as a strength, therein lies a seeming limitation as well; case selections made using the initial SVM model (cutoff 0.83) will be diluted by non-cases to a larger extent than the stringent classification selections. However, as we have described in our methodology, we choose to optimize the trade-off between sensitivity and PPV, which results in a particular cutoff point for the binarization of the SVM probabilities. This threshold can be modified to fit one's research purposes, for example optimizing the specificity only or the balance between specificity and sensitivity (Figure S1 and Table S1 and S2). This flexibility opens this methodology up to a wide field of potential applications.

Abbreviations

AUC-PRC: area under the precision recall curve

AUC-ROC: area under the receiver operating characteristic curve

EHR: electronic health record

ML: machine learning

NPV: negative predictive value

PPV: positive predictive value

RA: rheumatoid arthritis

SVM: support vector machine

Declarations

Ethics approval and consent to participate

We obtained ethical consent from the ethics committee of the Leiden University Medical Center before the initiation of our study.

Consent for publication

Not applicable.

Availability of data and materials

In accordance with the FAIR principles, we have made the Machine Learning scripts publicly available at GitHub [3]. The corresponding study data are available from the corresponding author upon reasonable request.

Competing interest

None declared.

Funding

This study would not have been possible without the support of the Dutch Arthritis Association (ReumaNederland) 15-3-301 and the Measurement of Efficacy of Treatment in the 'Era of Outcome' in Rheumatology (project number RP 2014-03)

Authors' contributions

TM and MM contributed equally to this study. The study design was developed by RK together with TM and MM. The ML annotation was provided by RK. AHM and EN annotated the classification criteria. TM developed the model and performed a statistical analysis together with MM. RK, TM, MM and AHM interpreted the results of the comparative analysis. The final manuscript was critically revised and approved by all authors.

Acknowledgements

We would like to thank Rory Monahan and Joy van der Pol for proofreading the paper.

Authors' information

1 Department of Rheumatology, Leiden University Medical Centre (LUMC), Leiden, The Netherlands

References

1. Jamian L, Wheless L, Crofford LJ, Barnado A. Rule-based and machine learning algorithms identify patients with systemic sclerosis accurately in the electronic health record. *Arthritis Res Ther*. 2019;21(1):305. doi:10.1186/s13075-019-2092-7
2. Kirby JC, Speltz P, Rasmussen L V, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc*. 2016;23(6):1046-1052. doi:10.1093/jamia/ocv202
3. Maarseveen T. DiagnosisExtraction_ML. Github. URL: https://github.com/levrex/DiagnosisExtraction_ML [accessed 2021-01-14]
4. 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 Inf*. 2020;8(11):e23930. doi:10.2196/23930
5. Bennett KP, Campbell C. Support Vector Machines: Hype or Hallelujah? *SIGKDD Explor Newsl*. 2000;2(2):1–13. doi:10.1145/380995.380999
6. de Rooy DPC, van der Linden MPM, Knevel R, Huizinga TWJ, van der Helm-van Mil AHM. Predicting arthritis outcomes—what can be learned from the Leiden Early Arthritis Clinic? *Rheumatology*. 2011;50(1):93-100. doi:10.1093/rheumatology/keq230
7. Arnett FC, Edworthy SM, Bloch DA, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum*. 1988;31(3):315-324. doi:10.1002/art.1780310302
8. 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;62(9):2569-2581. doi:10.1002/art.27584
9. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-Learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12(null):2825–2830.
10. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*. 2017;33(18):2938-2940. doi:10.1093/bioinformatics/btx364
11. Landewé RBM, van der Heijde DMFM. Why CAPS criteria are not diagnostic criteria? *Ann Rheum Dis*. 2017;76(4):e7 LP-e7. doi:10.1136/annrheumdis-2016-210656
12. June RR, Aggarwal R. The use and abuse of diagnostic/classification criteria. *Best Pract Res Clin Rheumatol*. 2014;28(6):921-934. doi:10.1016/j.berh.2015.04.004
13. Liao KP, Cai T, Gainer V, et al. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res (Hoboken)*. 2010;62(8):1120-1127. doi:10.1002/acr.20184
14. Aggarwal, R., Ringold, S., Khanna, D., Neogi, T., Johnson, S. R., Miller, A., Brunner, H. I., Ogawa, R., Felson, D., Ogdie, A., Aletaha, D., & Feldman, B. M. (2015). Distinctions between diagnostic and classification criteria?. *Arthritis care & research*, 67(7), 891–897. <https://doi.org/10.1002/acr.22583>

Tables

Table 1 Performance characteristics for different cutoffs of the SVM ML RA identification score in the independent test set.

	0.53	0.83	0.99
Sens	0.93	0.85	0.71
Spec	0.97	0.99	1.00
PPV	0.75	0.86	0.94
NPV	0.99	0.99	1.00

ML = Machine Learning; SVM = Support vector machine; RA= Rheumatoid Arthritis, PPV= Positive predictive value, NPV = Negative predictive value

Table 2 Comparison of baseline characteristics between the Machine Learning defined case selection (cutoff=0.83) and the two criteria based selections.

	Patients from cohort with EHR data and classification data		
	Predicted Case based on machine learning (cutoff=0.83)	1987 Criteria Based Cases	2010 Criteria Based Cases
N	405	386	457
Proportion Women	0.64	0.63	0.66
Proportion anti-CCP2 Positive	0.52	0.49	0.49
Proportion RF positive	0.57	0.58	0.59
Median DAS44 at Baseline	2.8	2.9	2.9
Median BMI	26.0	25.6	25.6
Median BSE	25	28	26
Median CRP	10.0	10.5	9.0
Median Age at Inclusion	57.3	58.7	57.2
Median Symptom Duration at Diagnosis (days)	91.0	89.5	91.0
Median Number of Swollen Joints	5	6	5

*P-values were calculated with the Pearson Chi-Squared for proportions, Mann-Whitney U for medians: * $p < 0.05$; ** $p < 0.01$, *** $p < 0.001$; ☒ Not statistically tested;*

Table 3 Baseline characteristics of the cases exclusively identified by the ML and those exclusively identified by either the 1987 or 2010 criteria.

	Only-ML not criteria	Only-criteria not ML
N	58	178
Proportion Women	0.59	0.64
Proportion anti-CCP2 Positive	0.06	0.16
Proportion RF positive	0.19	0.34
Median DAS44 at Baseline	2.51	2.63
Median BMI	26.6	25.8
Median BSE	19.0	25.0
Median CRP	10.4	6.8
Median Age at Inclusion	56.5	58.7
Median Symptom Duration at Diagnosis (days)	62	79.0
Median Number of Swollen Joints	4.0	4.0

Figures

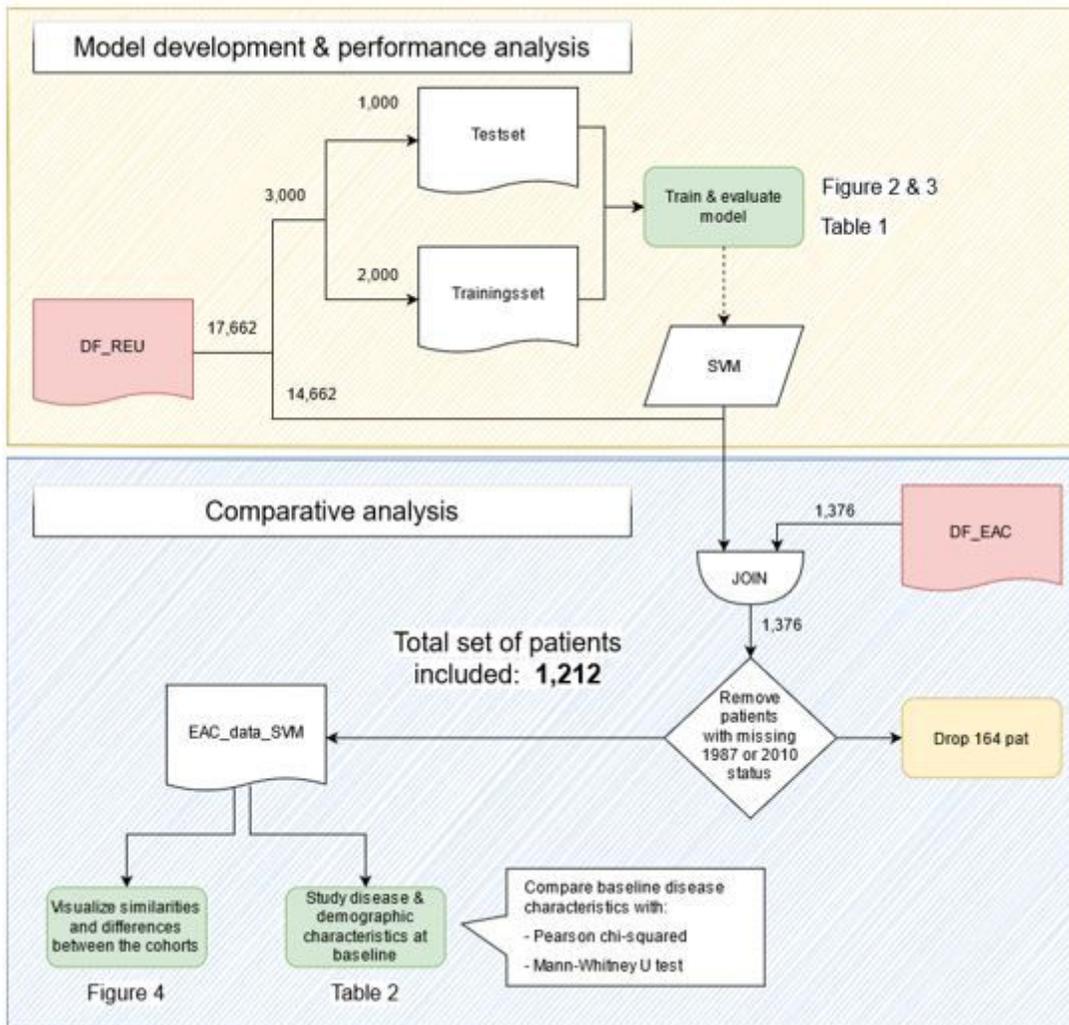


Figure 1

Study workflow depicting the model development & evaluation procedure and the criteria comparison analysis, whereby the important analysis-steps are highlighted in green. The patient collections are indicated by a wave line box, whereby the initial two patient collections are colored red (DF_REU = patients from the Leiden outpatient clinic with first consult after 2011; DF_EAC = Early arthritis cohort patients with first consult after 2011).

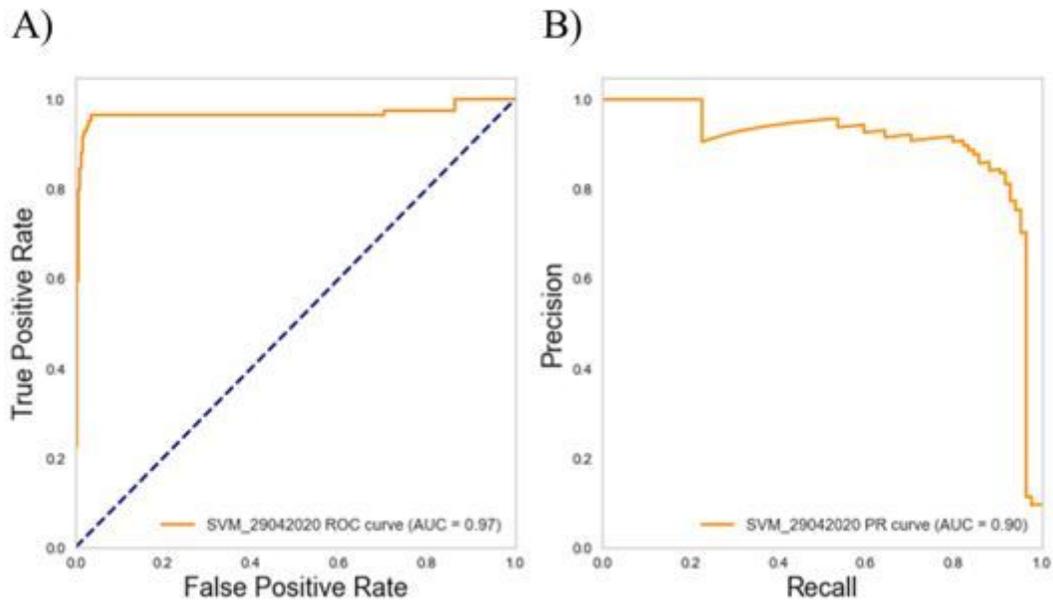


Figure 2

(A) Receiver operating characteristics plotting the sensitivity against the specificity and (B) precision-recall curve plotting the positive predictive value (precision) against the sensitivity (recall) for the support vector machine classifier in the independent test set. The precise features (top 20) that constitute the support vector machine model can be found in the original study by Maarseveen et al (2020).

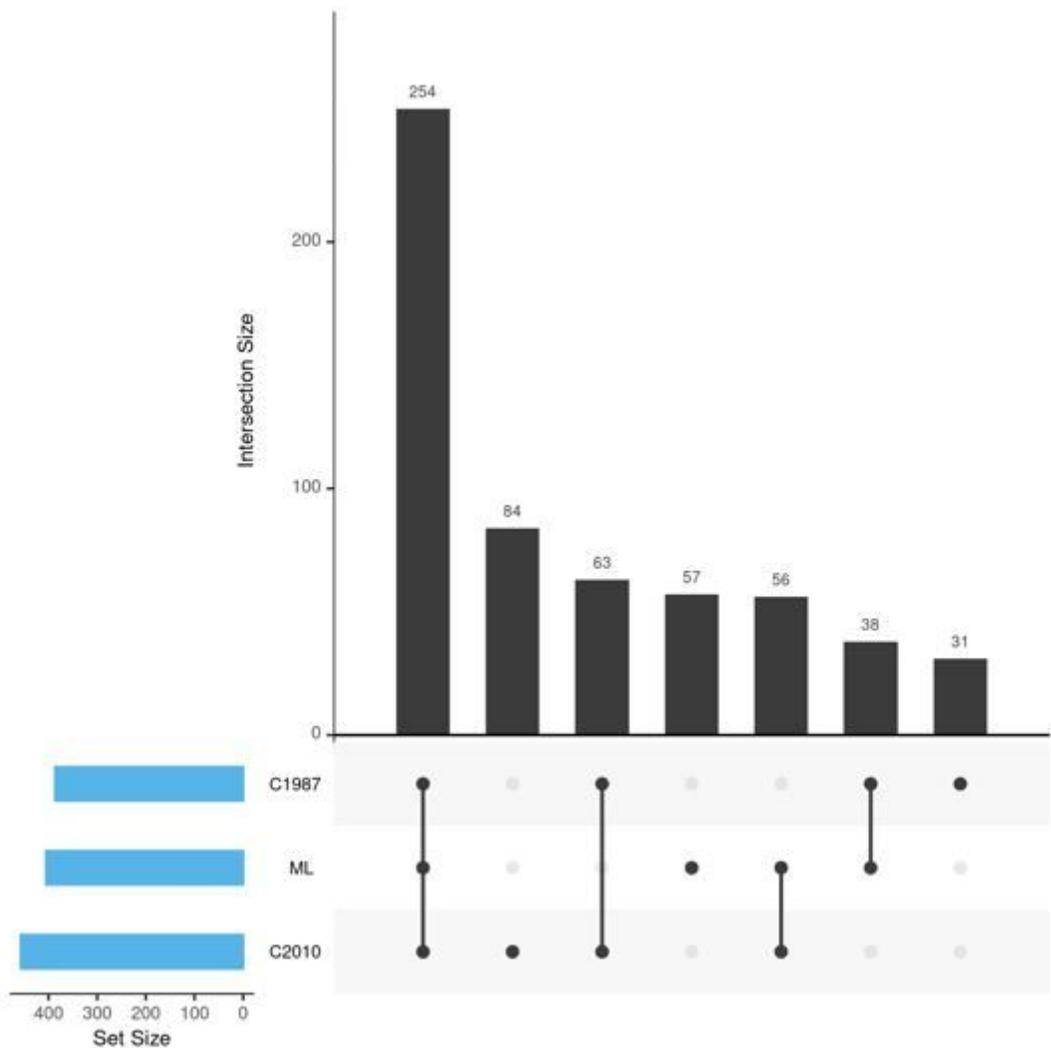


Figure 3

Upset plot visualizing the intersections of the ML defined cohort and the 2 criteria based selections, with a bar chart depicting the total cohort size in the bottom-left. Where C1987 = 1987 criteria based cases; ML = Machine learning based cases; C2010 = 2010 criteria based cases. N = 583 unique cases out of 1,212 records.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SUPPLEMENTARY.docx](#)