Title

Evaluation of an automated phenotyping algorithm for rheumatoid arthritis.

Authors
Zheng, Henry
Ranganath, Veena
Perry, Lucas
et al.

Publication Date
2022-11-01

DOI
10.1016/j.jbi.2022.104214

Peer reviewed

Published in final edited form as:
J Biomed Inform. 2022 November ; 135: 104214. doi:10.1016/j.jbi.2022.104214.

Evaluation of an automated phenotyping algorithm for rheumatoid arthritis

Henry W. Zheng[a,1,*], **Veena K. Ranganath**[b,1], **Lucas C. Perry**[b], **David A. Chetrit**[b], **Karla M. Criner**[b], **Angela Q. Pham**[b], **Richard Seto**[b], **Sitaram Vangala**[c], **David A. Elashoff**[c], **Alex A.T. Bui**[a]

[a]Medical & Imaging Informatics, University of California Los Angeles, 924 Westwood Blvd Suite 420, Los Angeles, CA 90024, USA

[b]Department of Medicine, Division of Rheumatology, David Geffen School of Medicine, 10833 Le Conte Ave, Los Angeles, CA 90095, USA

[c]David Geffen School of Medicine, 10833 Le Conte Ave, Los Angeles, CA 90095, USA

Abstract

To better understand the challenges of generally implementing and adapting computational phenotyping approaches, the performance of a Phenotype KnowledgeBase (PheKB) algorithm for rheumatoid arthritis (RA) was evaluated on a University of California, Los Angeles (UCLA) patient population, focusing on examining its performance on ambiguous cases. The algorithm was evaluated on a cohort of 4,766 patients, along with a chart review of 300 patients by rheumatologists against accepted diagnostic guidelines. The performance revealed low sensitivity towards specific subtypes of positive RA cases, which suggests revisions in features used for phenotyping. A close examination of select cases also indicated a significant portion of patients with missing data, drawing attention to the need to consider data integrity as an integral part of phenotyping pipelines, as well as issues around the usability of various codes for distinguishing cases. We use patterns in the PheKB algorithm's errors to further demonstrate important considerations when designing a phenotyping algorithm.

Keywords

Phenotyping algorithm; Computational phenotyping; Rheumatoid arthritis; PheKB

[*] Corresponding author. henryzheng@g.ucla.edu (H.W. Zheng).
[1] Equally contributed to the work presented.

Appendix A. Supplementary material
Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2022.104214.

Declaration of Competing Interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

1. Introduction and background

There have been broad advancements to identify patients with given medical conditions through the electronic health record (EHR) and promote personalized care delivery, research (e.g., for clinical trial recruitment), registry development, and other applications [1,2]. The challenges of this type of cohort discovery are well-known [3], particularly when clinical case definitions are not easily established and/or the diagnosis entails subjective physician interpretation. As such, an area of active research has been the development of automated phenotyping algorithms that implement a computerized "chart review" to assess features of a patient and judge their likelihood of having a given condition. Concerted efforts have been made to develop "automated" phenotyping algorithms. Phenotyping algorithms have been developed for a wide variety of applications, from cohort discovery to clinical screening and outcome prediction. One example is the electronic Medical Records and GEnomics (eMERGE) network [4], which established phenotyping algorithms built to discover patients and link phenotype to genotype (e.g., for genome wide association studies, GWAS). The Phenotype KowledgeBase (PheKB) is a repository for the phenotypes developed by eMERGE and others [5]. PheKB's curatorial efforts aim to standardize the information needed for phenotyping, including harmonized data elements, (deidentified) patient databases that the algorithms were evaluated on, their performance on those samples, tools used to implement these algorithms, as well as documentation describing the work for others. These data elements most commonly deployed in PheKB phenotyping algorithms include diagnostic codes (e.g., International Classification of Disease, ICD; Current Procedural Terminology, CPT; etc.), lab measurements, medications, and vital signs as well as ontologies for natural language processing (NLP) methods. To date, 70 eMERGE disease phenotypes have been identified, ranging from colorectal cancer to rheumatoid arthritis (RA).

The diagnostic process typically involves holistic clinical reasoning, where the physician weighed a patient's presentation, history, labs, and other medical data in aggregate, known as the "clinical presentation." However, this diagnostic approach may at times be difficult to replicate due to each physician's conceptualization of the clinical presentation [6]. As a result, there has been a push towards clearer and more consistent definitions of diseases and conditions across different clinical disciplines. What form that definition takes depends on the medical condition. In some conditions, it involves a clear gold standard test or validated measurement (e.g., polymerase chain reaction [PCR] assays or imaging) [7]. In other conditions, a set of features and presentations agreed upon by domain experts are codified into a standard or rubric, as diagnostic or classification criteria. A prominent example of this second method is in psychiatry, where the Diagnostic and Statistical Manual of Mental Disorders (DSM) series of diagnostic criteria define psychiatric phenotypes and form diagnostic guidelines [8]. Regardless of how the criteria are built, the goal is to produce an internally – and ultimately externally– valid set of criteria that can differentiate between patients with and without the disease at clinically useful accuracy. Phenotyping algorithms can be understood as the next evolution of this standardization, where clinical features are encoded, and the disease defined algorithmically.

Inherent to any classification algorithm, of which phenotyping is a kind, is the issue of balancing sensitivity and specificity. Increasing specificity entails using more restrictive criteria, causing some positive edge cases to be rejected (and hence, lower sensitivity). Conversely, ensuring higher capture of positive cases raises sensitivity, but at the expense of greater false positives. For example, one RA phenotyping study [9] using only ICD-9 codes yielded a sensitivity of 89 % but specificity of 57 %. A second study [10] that used ICD-9 codes, rheumatoid factor (RF) labs, and a prescription for a disease-modifying antirheumatic drug (DMARD) yielded a specificity of 97 % but a sensitivity of 76.5 %. What features are used and how they are defined are examples of the impact that design choices have on performance and how different parts of the EHR serve to cover different but overlapping information about a condition. Markedly, these studies markedly do not analyze in depth what are the error types and whether that affects the validity of the algorithm design.

Such validation studies have not yet clearly characterized how variations in performance relate to design or implementation differences, and what lessons can be drawn for developing phenotyping algorithms. In this paper, we evaluated the performance of PheKB's phenotyping algorithm for rheumatoid arthritis in an adult patient population at the University of California, Los Angeles (UCLA) so that we can closely study its usability across institutions and to study the methodology of evaluating phenotyping algorithms more generally. We sought to explore the challenges related to implementing the algorithm and required features, as well as to appreciate differences in expected performance. Notably, we had anticipated that given our institution's different demographic admixture (e.g., higher proportion of non-White patients) compared to the original populations on which the PheKB algorithm was developed, notable deviations in performance and biases should arise. Particular attention was also paid to the PheKB algorithm's handling of ambiguous cases, such as patients with other autoimmune diseases (e.g., psoriatic arthritis, systemic lupus erythematosus [SLE], etc.) that may be misdiagnosed as RA. Lastly, an evaluation looking into differences in the diagnosis of RA by clinicians and by the PheKB algorithm was conducted to forge a deeper understanding of practical considerations in using this and other phenotyping algorithms. The result of this evaluation is that we identified several important lessons in thinking about the design of phenotyping algorithms more generally and their systematic assessment.

2. Methods

2.1. Defining rheumatoid arthritis

The diagnosis of RA has evolved in step with evolving understanding of the disease. Creating criteria for this condition has been difficult as RA patients show considerable variation in clinical presentations and may not always present with a consistent set of features (chronic joint swelling and serological/inflammatory markers). In 1987 the American College of Rheumatology (ACR) established RA classification criteria to create a standardized definition for enrollment of a homogenous group of patients for clinical trials [11]. This original definition examined seven clinical features: morning stiffness, arthritis of three or more joints, arthritis of hand joints, symmetric arthritis, rheumatoid nodules, serum rheumatoid factor, and radiological changes. The 1987 ACR RA Classification

Criteria was later revamped to identify RA patients earlier and to incorporate the anticyclic citrullinated peptide antibody (ACPA, ~95 % specific for RA). The current prevailing standard for diagnosing RA is the 2010 American College of Rheumatology/European League Against Rheumatism (ACR/EULAR) classification criteria (Table 1) [12]. Two clinical/laboratory criteria are numerically weighed heavily: joint involvement (swelling and pain) and serological evidence (ACPA and/or RF presence).

Computer-based implementations of the 2010 ACR/EULAR criteria have been frustrating due to the inconsistent ways the relevant data are encoded and collected within the EHR. For example, values such as specific joint inflammation are typically not encoded in structured data but are instead described in free-text clinical notes. PheKB's RA algorithm [13] is constructed using Automated Feature Extraction (AFEP) algorithm [14] that uses a penalized logistic regression (LASSO) approach to select features. As deployed onto their dataset, the algorithm selected a number of ICD-9/10 codes relating to RA and lab results for rheumatoid factor (RF). It also considers the presence of ICD-9/10 codes for psoriatic arthritis and SLE as well as the total number of diagnoses received by the patient. Notably, some of the features selected by AFEP are features also used by the 2010 ACR/EULAR criteria while others were not included, particularly lab results for ACPA or text-based indications of joint involvement. PheKB reports a sensitivity of 87 % and a specificity of 97 % for its algorithm based on a pre-screened cohort of patients with and without RA (n = 100). However, the distribution of symptoms within that cohort is unclear, which may negatively impact algorithm portability.

2.2. Cohort creation

2.2.1. Cases—To evaluate the performance of the PheKB algorithm, we performed a retrospective case-control study on a sample of the UCLA Health System adult (≥18 years) patient population. Candidate cases were screened using the inclusion criteria used by the PheKB algorithm, which were patients with at least one ICD-9 code (714, 714.0, 714.1, 714.2) or ICD-10 code (M05.* and M06.*) that reflected an RA diagnosis. 22,266 patients from 2009 to 2019 met these criteria. Of these, 2,385 patients had been reviewed and identified as having RA through clinician chart evaluation for purposes of RA clinical trial recruitment.

2.2.2. Controls—To examine the ambiguity of RA diagnosis, a control cohort was built from two populations: 1) patients without the above ICD-9/10 codes, and 2) patients with exclusion criteria diagnoses (listed in Table S1). The proportion of these two groups were selected to reflect the prevalence of the exclusion criteria, which in aggregate have a 10 % total prevalence. Patients with psoriatic arthritis and SLE were not excluded as per the PheKB algorithm's exclusion criteria to be able to study the algorithm's performance on these cases. The final breakdown of the control group was:

1. 90 % with no RA diagnosis, of whom some patients have a psoriasis diagnosis, some patients have an SLE diagnosis, and some with neither psoriasis diagnosis nor SLE diagnosis; and
2. 10 % has an exclusion criteria diagnosis, with or without an RA diagnosis.

2,381 patients were selected based on these criteria to form the control cohort. Table S2 details the demographic breakdown of the cohort.

2.3. PheKB algorithm implementation

The 4,766 patients were scored using the PheKB algorithm, which uses a logistic regression [13]:

$$s = \frac{e^{-1.017 + \beta}}{1 + e^{-1.017 + \beta}}$$

where:

$$\beta = 1.937$$

$$* \log(1 + RAICD) + 1.639$$

$$* \; LabCount - 0.529$$

$$* \log(1 + SLE) - 0.122$$

$$* \log(1 + Psoriatic) - 0.954$$

$$* \log(1 + Diagnoses)$$

The formula's variables are defined as: *RAICD*, the total number of RA ICD-9 and ICD-10 diagnoses (714, 714.0, 714.1, 714.2, $M05.*, M06.*$); *LabCount*, whether the highest RF lab value seen in the patient's charts is either abnormal ($LabCount = 1$) or normal/unavailable ($LabCount = 0$); *SLE*, the total number of SLE ICD-9/10 diagnoses, (Table S1); *Psoriatic*, the total number of psoriatic arthritis ICD-9/10 diagnoses, (Table S1); and *Diagnoses*, the total number of ICD-9/10 diagnoses in the patient's charts. Patients who scored greater than 0.632 were predicted as positive for RA. Comparison of the PheKB algorithm's prediction to the patient's case/control assignment was then calculated accordingly. The ICD-9/10 codes used to define diagnoses of SLE and psoriasis are also as defined in the PheKB documentation.

2.4. ACR/EULAR criteria subset

To further study the algorithm's performance as well as identify any patterns in errors, 150 patients from among the cases and 150 from among the controls (for a total of 300) from those sampled above were additionally manually evaluated under the 2010 ACR/EULAR criteria. Under the guidance of a clinical RA expert (VKR), specific instructions for EHR chart review were determined and individuals trained. Patient charts were reviewed by a rheumatologist faculty RA expert (VKR), three rheumatology fellows (KC, DC, RS), one medical student (AP), and one research associate (LP). Reviewers were pre-trained on a known set of 10 positive RA patients and 5 negative RA patients. The six extractors determined the RA status of this subset, with each patient categorized in one of four ways: 1) presence of RA (positive); 2) absence of RA (negative); 3) unsure; and 4) incomplete data to make an evaluation. Of the 300, six cases were used as inter-annotator comparison and reviewed by all five clinicians, leaving 49 unique cases per clinician for four clinicians (DC, KC, AP, RS) and 98 unique cases for VKR. All cases were also reviewed a second time by the research associate (LC).

3. Results

Table 2 summarizes the results of the PheKB algorithm as applied on the full cohort. The sensitivity of the algorithm was 71.9 %, its specificity was 95.2 %, and its F-score was 0.814. Fig. 1 shows the distribution of raw scores calculated by the algorithm, separated by cases and controls. To understand the contribution of each variable in the logistic function to the ultimate score, the distribution of values of each variable as calculated by the algorithm are described in Table S3 and Fig. S1. These results broadly reflect the findings of a generalizability study performed on the algorithm by the PheKB group, which observed sensitivities ranging from 65 % to 82 % across three institutions; however, these studies did not analyze the causes or types of errors nor examine for noteworthy bias [15].

Table 3 summarizes the performance of the PheKB algorithm on the ACR/EULAR criteria subset. On this subset, the PheKB algorithm had a sensitivity of 74.5 % and a specificity of 99.5 %. Among the six cases used to measure inter-annotator agreement, 5 of the 6 (83 %) cases showed perfect agreement across all five clinicians, and one case showed disagreement (the patient had gout and inconsistent notes both diagnosing and rejecting an RA diagnosis), yielding a mean Cohen’s kappa of 0.71. The one false positive was from a seronegative patient who was seen several times for RA but did not respond to methotrexate and is suspected to have osteoarthritis instead of seronegative RA. Seven of the fourteen patients predicted as negative by the PheKB algorithm are considered to have RA based on clinical assessment using clinical notes as well as autoimmune medication usage (methotrexate, prednisone, adalimumab, etanercept).

Fig. 2 illustrates the distribution of raw scores calculated by the PheKB algorithm on the ACR/EULAR criteria subset, and Table S5 breaks down how many patients within the cases and controls were assigned different RA statuses. Markedly, several patients considered negative by the PheKB algorithm but had a diagnosis of RA were observed (i.e., false negatives). The algorithm struggles to identify seronegative RA patients, as 5 of the 14 (36 %) false negative patients either were missing or had normal serology.

There were also differences observed because of different lab usage. Among the false negatives, 3 out of 7 patients (43 %) with positive acute-phase reactant lab tests had a negative RF lab value. In comparison, 9 out of 38 (24 %) correctly identified RA patients had positive acute-phase reactant labs but negative RF. The PheKB algorithm only considers the RF lab, but the 2010 ACR/EULAR criteria also consider erythrocyte sedimentation rate (ESR), ACPA, and C-reactive protein (CRP) as relevant labs for diagnosis.

From among those with a diagnosis of psoriasis and SLE, the PheKB algorithm correctly identified all patients. Within the sampled population, there was one patient with psoriasis and seventeen patients with SLE. None of these eighteen patients had RA.

16.3 % (49/300) of the patients reviewed had incomplete data for proper evaluation using the 2010 ACR/EULAR criteria. Among those, 10.2 % (5/49) were classified as positive by the PheKB algorithm and 89.8 % (44/49) of those were classified as negative.

4. Discussion

We evaluated the PheKB RA algorithm to elucidate the reproducibility of its computational phenotyping methodology. We tested the algorithm on a patient population different from the one PheKB used, studied its performance on both a general set of patients with and without RA, and performed chart review on a subset of 300 patients against the 2010 ACR/EULAR criteria to pinpoint sources of error. When utilized in a UCLA population, the algorithm demonstrated high specificity but modest sensitivity, findings that directly reflect its design choices. These findings also elaborate on and expand upon a similar evaluation of this algorithm performed by Carroll et al. [15], which yielded performance that roughly match this study's but which do not examine in depth the specifics of how, why, or under what circumstances the phenotyping algorithm should be used or the sources of the errors that were observed. Here, we reveal the effects that implementation and design choices make on the usability of a phenotyping algorithm and draw important lessons for consideration when performing such evaluations. Taken together, this evaluation demonstrates the necessity for more comprehensive thinking in developing phenotyping algorithms specific to disease, task, and patient population and the need to incorporate them into the algorithm's design.

4.1. Implementational considerations

We consider some important observations that arose during implementation of PheKB's algorithm. Care was taken to replicate the algorithm and experimental setup as closely as could be surmised from documentation, but nevertheless important ambiguities and issues arose that point to the need for greater rigor in defining or disseminating these algorithms. These implementational considerations were not discussed in the PheKB algorithm's methodology nor in the validation study done by Carroll et al. [15], and so are discussed in detail here as part of this evaluation study.

4.1.1. Disease definition—RA's complex and potentially ambiguous phenotypes were a major complication in constructing a proper evaluation. RA is manifested through several symptom clusters, including joint inflammation, autoimmune serology, and inflammatory markers. None of these symptoms individually are specific to RA but, when taken as a whole, increases its likelihood and contributes to a positive assessment. Thus, a phenotyping algorithm for RA must reflect its clinical complexity and its overlap or confusion with many other conditions such as psoriatic arthritis, SLE, and others. The intricacies of an RA diagnosis became abundantly clear during the chart review of the ACR/EULAR subset, where an RA diagnosis relies on both positive indications such as joint involvement as well as negative indications such as the likelihood of a similarly-presenting disease such as SLE. In clinical practice, differential diagnosis between RA and other rheumatic conditions is accomplished by considering the extent of non-arthritic symptoms like rash, vasculitis, or nephritis; the patient's disease history; physical exam; and their response to targeted RA medications. A primary care physician may give an initial diagnosis of RA that is followed up by a rheumatologist, who confirms or revises the diagnosis. Such differences in diagnostic accuracy due to workflow has been noted in other studies [16]. As a result, the same features are used in different manners at different stages of a RA diagnosis, and a phenotyping algorithm's design should these workflow differences into consideration, contextualizing how features are used based on where in the workflow the algorithm is intended to be deployed. If used for screening, such an algorithm should acknowledge the uncertainty of a single diagnosis and not use features seen only during confirmatory diagnosis. If used for downstream genomic studies, the algorithm should reflect the logic of RA's diagnosis, target particular subgroups or reject patients with diagnoses that would muddy the selected patients' phenotypic profile more thoroughly. The PheKB algorithm neglects to reflect this complexity, resulting in the inclusion of patients with osteoarthritis, Sjögren's syndrome, or gout, conditions that are neither excluded on nor penalized for appropriately.

4.1.2. Patient admixture—

It is notable that the evaluation of the algorithm was performed on a UCLA patient population because the populations that the PheKB algorithm was calibrated and evaluated on previously likely have different demographic admixtures to UCLA's patient population. The exact demographic breakdowns, including race and gender breakdowns, were not reported in their respective publications, but given that the patient populations were drawn from the Partners, Northwestern, and Vanderbilt hospital systems, it is likely that their patient admixtures have a higher proportion of Caucasian patients relative to our Los Angeles population. As described in Table S2, nearly 40 % of the sampled patients were not Caucasian. In this study, the patients were drawn from a wide number of clinics and hospitals within the UCLA hospital system. The original PheKB paper and subsequent validation study state that their respective patient samples were drawn from data marts intended for research use – the preprocessing that went into selecting for or filtering patients that went into the data marts are not described. The patients sampled in this study were chosen because they were seen by clinicians and reflect the feature presentation, data missingness, and other characteristics observed in real-world clinical data. The nature of our UCLA population, along with its larger size (4,766 patients compared to the 100 used in the original calibration paper and the ~400–500 patients used in the evaluation study) arguably captures a more realistic portrayal of how the PheKB algorithm performs on a more diverse (both in terms of demographics and data quality) patient admixture.

4.1.3. Cohort construction—

Identifying candidate cases and control RA cases from the UCLA population had significant ramifications on what the performance of the phenotyping algorithm means. Table S5 illustrates how the control patients align closely with an absence of RA, but among cases there shows significant variation in actual presence or absence of RA. Based on how cases and controls were defined, the large proportion of cases without RA shows that having at least one ICD code diagnosing RA is insufficient in identifying patients with RA. Thirteen out of the 148 clinically reviewed patients deemed cases did not have RA and instead were better diagnosed as osteoarthritis, Sjögren's syndrome, SLE, or gout. It is important to point out that the definitions for cases and controls were pulled from the PheKB algorithm development's own definitions and were mirrored in this study. The case definition of the PheKB algorithm only checks that a patient has at least one RA diagnosis, regardless of its source (was it a rheumatologist or primary care physician who gave the diagnosis?), time (how long ago was the diagnosis provided?), or other complicating factors that go into an evaluation for RA. Thus, the result of only using ICD codes to identify candidate RA cases is a high false positive rate that includes patients initially suspected of having RA but were subsequently diagnosed with something else. This finding correlates with clinical intuition, as the need for a phenotyping algorithm is to refine a disease definition more complexly than merely ICD codes or medications. However, it is important to note that there may exist a population of RA patients without any RA-indicating ICD codes, medications, or other datapoints that were not identified based on case definition. How to capture these pre-diagnosis patients is beyond the scope of this paper.

4.1.4. Feature ambiguity—Several patients were categorized as "uncertain" because of RA diagnoses that were not followed up. These patients were categorized as not having RA by PheKB's algorithm, some with highly confident scores. The 2010 ACR/EULAR criteria's score point for a duration of greater than 6 weeks is assessed by a clinician's history review showing continuous presence of RA symptoms over that time. This nuance is not captured in the PheKB algorithm, which looks at a patient's entire medical history. A large time gap between RA diagnoses without any evidence of follow-up or treatment is inadequately discounted by merely using the total number of diagnoses if the patient had not been to a hospital in a long time or if the patient has had only one visit noting RA. In fact, based on the distribution of total number of diagnoses as seen in Fig. S2, these differences were not as large between cases and controls as other factors, particularly number of RA diagnoses. Thus, it did not serve as a driving factor in the PheKB algorithm's ability to differentiate cases from controls. Ultimately, greater sophistication in modeling the temporality of diagnoses is required to capture significant diagnoses in an overall patient history.

4.1.5. Missing data—One significant class of patients reported by both our chart reviewers and other clinicians [17] are those who are missing too much data to be practically scored on the 2010 ACR/EULAR criteria. The 2010 ACR/EULAR criteria assume that all required information is available, and its weighting scheme is normalized on this assumption. However, the PheKB algorithm relies on counts, automatically considering missing data as a positive mention of no significant result. By way of illustration, the lack of RF lab value in EHR (missing value) is considered by the PheKB algorithm as equal to having a normal RF lab value. The unavailability of data forms the largest proportion of disagreement between a predicted RA diagnosis and the chart-reviewed patients' RA status.

Any implementation of the PheKB algorithm must therefore first validate if the relevant features are missing or whether there is a positive mention of zero or normal for that feature. If automated phenotyping algorithms are to be used in a broader clinical setting, it may be necessary for the EHR to capture in finer detail the necessary data. Alternately, a solution may be to define a more flexible criteria that factors in the possibility of missing data within the phenotyping algorithm. For example, if RF lab tests were not available, the point cutoff for a diagnosis of rheumatoid arthritis might be normalized to five points rather than six. Thus, an ability to accommodate missing or incomplete data is necessary to work with the practical reality of data that arises from clinical care. [18] Whether resolved by improved clinical workflows, EHR design, imputation, or modeling, an implementation of a phenotyping algorithm must consider this fact. Furthermore, how this is resolved will affect downstream data integrity, meaning that these choices must be contextualized within the phenotyping goal.

4.2. Lessons learned for designing phenotyping algorithms

Error analysis of the PheKB algorithm on the UCLA population has shown several important design considerations that impact an algorithm’s transportability onto new populations or tasks. This study shows that the scope that a phenotyping algorithm can have depends on the design of the algorithm’s components, including the features used and the patient cohort chosen for testing and validation. In the PheKB algorithm’s case, it was originally designed for use in downstream genomic studies such as GWAS or PheWAS studies, which aim to have high specificity to ensure the desired phenotype (i.e., no spurious false positive cases). The high specificity of PheKB’s algorithm designs is recapitulated in the UCLA population. By calibrating their model to a specificity of 97 %, the developers chose to minimize false positives; however, the tradeoff is a decrease in sensitivity. Beyond calibration, however, the innate design of the PheKB algorithm does not lend itself to prospective uses such as screening – the choice of features, when contextualized within clinical practice, is at odds with what features are available prospectively. As such, for some applications such as screening or recruitment, the PheKB algorithm’s design would need reconsideration.

The phenotyping algorithm comes at the end of a long sequence of scientific investigation and definitions during which choices made by clinicians, studies, and/or guidelines affect the scope and methodology for its construction. As demonstrated here, such choices impact performance across institutions and affect transportability. Unfortunately, these choices are frequently opaque to the phenotyping algorithm designers – and consequently, downstream implementers – making the task of designing and, by extension, evaluating phenotyping algorithms a tricky-one. Depending on whether a phenotype was defined clinically based on case/control studies, clinical consensus, or other, the resulting definitions of a disease’s phenotype will carry the effects of these methodologies and designing a phenotyping algorithm using methods that go at odds with those embedded within the phenotype itself yields poor-performing or non-transportable algorithms. The error types observed in this study reflect this issue, as a close study of the original design of the 2010 ACR/EULAR criteria suggests that its use case does not reflect the PheKB algorithm’s use cases –

while the criteria are intended to screen for suspected patients, the PheKB algorithm was developed to pick out obvious cases and reject ambiguous ones.

4.2.1. Feature selection—There is misalignment between the PheKB algorithm and the 2010 ACR/EULAR criteria with regards to which laboratory features are used for a diagnosis. The 2010 ACR/EULAR criteria involve several markers of inflammation, including ESR and CRP, and the highly specific ACPA. In contrast, PheKB only looks at one lab, RF, and thus misses out on a class of patients who have not tested for, or potentially have subclinical levels of, RF but who exhibit a long history of radiological evidence for RA such as marginal erosions and periarticular osteopenia and who have other inflammatory markers sufficient for a diagnosis of the disease. This is significant as approximately 20 % of RF-negative patients are ACPA positive. Furthermore, ESR and CRP are more commonly performed tests and are thus more commonly available than an RF lab value. As a result, the PheKB algorithm rejects patients who do not have RA ICD codes yet have abnormal acute-phase reactant labs and abnormal radiological symptoms. This type of patient would be the clinically relevant one for screening, and a set of them were indeed observed among the ACR/EULAR criteria subset. It is important to note, however, that ESR and CRP are sensitive but non-specific measures of inflammation (e.g., elevated due to infection, malignancy, other autoimmune conditions etc.) and that a revised algorithm that uses these labs as features may erroneously capture patients with inflammatory diseases other than RA.

It is likely that only ICD codes and RF factor were used in the PheKB algorithm because an automated feature selection algorithm was used and these features were statistically the most predictive of RA. This evaluation shows that more consideration is needed when selecting features. As an automated approach is blind to prior clinical knowledge, any unintentional biases within the patient data could be picked up by the algorithm as a spurious clinical feature. In the case of RA, automated feature selection method was suboptimal as not all RA patients had all RA symptoms (most commonly seen in patients in remission). Automated feature selection may make sense in diseases where all patients exhibit some degree of pathology in a specific set of symptoms and varied only by their degree – that way, statistically significant deviations from the mean among diagnostic features would arise and allow an automated feature selection method to identify them. But as seronegative RA is a known and potentially growing subtype of RA, the automated feature selection did not identify several serological markers as relevant features. Prior clinical knowledge about existing RA subtypes should be incorporated into the case selection process to more deliberately generate sufficiently sized samples of different RA subtypes. Furthermore, using an automated feature selection method in conjunction with a population-representative case definition results in ignoring clinically important but low-frequency phenotype subtypes and the features therein to identify them. The 2010 ACR/EULAR guidelines evaluate RA on three phenotypic axes – joint pain, antibody serology, and inflammatory markers – that are used together over time to identify an RA patient. A patient with all three and which would not be better explained by another condition would strongly be suspected of having RA. These patients are also the most likely to go to a specialist to be evaluated for RA, creating a patient population that is biased towards showing all three phenotypes, while younger or seronegative RA patients with a less clear history of RA phenotypes are under-sampled.

An automated feature selection method such as AFEP thus would preferentially select features that confirm the diagnoses of patients with clear, already-known cases of RA and reject features seen in under-sampled RA subtypes such as seronegative RA patients with short histories. Indeed, AFEP, which uses a penalized logistic regression model to select features, identified within its given patient admixture the strongly predictive features of numerous RA ICD diagnoses and a positive RF test, which is performed as a confirmatory diagnostic test and are typically only done for patients already being evaluated for RA, while rejecting other features such as joint count or inflammatory markers. In other words, the statistically significant features identified by automated feature selection algorithms arise from the cohort used for development, which was insufficiently enriched for certain patient subtypes such as seronegative RA patients. The end result of pairing an automated feature selection algorithm with an insufficiently enriched or balanced patient cohort is that clinically meaningful but low-frequency subtypes are ignored, limiting the algorithm's clinical validity.

4.2.2. ICD codes and clinical relevance—The PheKB algorithm relies heavily on ICD codes as part of its algorithm as a positive indication of RA presence, as a negative indicator of similar diseases, and as a denominator indicating size of the clinical record. This evaluation demonstrates that a simple use of ICD codes overlooks several complexities that limit their usability. One misalignment between the PheKB algorithm and the 2010 ACR/EULAR criteria is the approach towards capturing duration of diagnosis – an issue inherent to and complicating the longitudinal (retrospective) analysis of EHRs. It is recognized that patients who have received an RA diagnosis but otherwise show no evidence of RA in subsequent visits or during follow-up may have received an erroneous initial RA diagnosis. The PheKB algorithm attempts to capture this by discounting a low "density" of RA diagnoses via a negative weight on the total number of diagnoses as represented by total number of ICD codes in the patient's EHR. This approach is predicated on the idea that if a patient has RA, repeated healthcare visits would reflect repeated diagnoses of RA (e.g., as part of their ongoing medical problem list). But this assumption would not hold for patients who do not have extensive histories, have large gaps in their medical history, who see providers outside the EHR system, or who do not regularly receive care (such as those in remission) [19]. In general, the use of ICD codes as part of an EHR-based algorithm implicitly assumes correct coding and completeness, which may not necessarily be the case, and should be carefully contextualized.

The PheKB algorithm also penalizes via negative weights the presence of a diagnosis for psoriatic arthritis and systemic lupus erythematosus (SLE). This design choice reflects the fact that psoriatic arthritis and SLE are overlapping conditions that share many similar symptoms, and that psoriatic arthritis and/or SLE can even appear along with RA in some patients. Thus, patients with psoriatic arthritis or SLE may be diagnostically confused for RA, and the algorithm aims to reduce these cases of false positives by penalizing their presence, especially in light of the original algorithm's intent for identifying strong cases of RA for downstream biomedical research. From the review of the patient sample, the PheKB algorithm successfully differentiated patients with psoriatic arthritis and SLE from among those with rheumatoid arthritis. It must be noted, however, that the algorithm does not penalize other diseases like such as systemic sclerosis, ankylosing spondylitis, mixed connective tissue disease, vasculitis, and others which, like psoriatic arthritis or SLE, are symptomatically related to and can sometimes be confused with RA.

4.2.3. Temporal relationships—

How features are selected or used should also take into consideration clinical workflow to contextualize when they are usable. In clinical practice, an initial RA/inflammatory arthritis ICD code may be added to a patient’s chart when it is being considered as a diagnosis, which is followed up by a rheumatologist with confirmatory labs such as an RF/ACPA lab. Because the lab was ordered only after an initial suspicion of RA, its diagnostic power is limited to those with suspicion of RA. On the other hand, if a patient with elevated ESR or CRP labs from a different evaluation started exhibiting other RA symptoms, there would be a stronger suspicion of RA. Thus, the sequence that these events occur have different diagnostic power. The EHR may reflect this, showing how an ICD code leads to the ordering or availability of specific other labs, which a physician reviewing a chart would implicitly understand this relationship. Considering the clinical workflow in diagnosing RA reveals that there may be a useful set of temporal features that the PheKB algorithm insufficiently captured and that an explicit temporal model relating labs and diagnostic codes may be warranted.

4.3. Future work

Future work should evaluate the specific performance of the algorithm in an expanded cohort enriched with potentially ambiguous cases of different subtypes, paying attention to areas of success and failure in differentiating patients with the target disease (e.g., rheumatoid arthritis) that overlap with related conditions (e.g., psoriasis or SLE, patients with only psoriasis and SLE, and patients with neither but which have as-yet undiagnosed inflammatory or arthritic symptoms). By characterizing the correlative distributions of these conditions within a patient population and the symptoms and evidence that characterize them, a more detailed understanding of which clinical features to use and how to quantify their relationships can emerge. It is likely that the choices of features were driven in part by the availability of data, as ICD codes and lab values are among the most commonly available data in most HER systems. Ease of implementation, however, should not be the primary consideration when considering the applicability of a phenotyping algorithm for a diagnostic task. Rather, the design of the phenotyping algorithm, the features it uses, and its performance on the intended patient population should be foremost in considering the applicability of a phenotyping algorithm to the end application. Developing NLP methods accurate enough to provide the remaining data in a useful way is an area of active research.

4.4. Limitations

A limitation of this study is the relatively small size of the subset evaluated on the 2010 ACR/EULAR criteria. While several interesting findings emerged from the analysis of patterns in the erroneously categorized patients, evaluating the entire cohort using the 2010 ACR/EU-LAR criteria would allow greater power and statistical certainty in confirming these observations. A larger cohort size could also improve the inter-annotator accuracy measurement, as currently only six cases were common across all reviewers. Second, this study merely reviews the performance of an existing algorithm and provides recommendations but does not test or validate an alternative algorithm. Recalibrating the PheKB algorithm on the existing cohort and comparing the existing PheKB algorithm to a revised algorithm are part of an extended study of RA phenotyping and will be future work.

5. Conclusion

This study looked at the performance of an automated phenotyping algorithm for rheumatoid arthritis, demonstrating a broader set of considerations necessary to designing and applying phenotyping algorithms onto different populations and tasks. Special attention was paid to ambiguous cases, such as those with commonly seen conditions or those included in the original study's exclusion criteria, to reflect the breadth of circumstances in which a phenotyping algorithm might be deployed for. Performing an evaluation on a large cohort of RA patients drawn from the UCLA hospital system yielded a more realistic and comprehensive study of the PheKB algorithm's performance. A major finding of this study was that the cohort of patients were missing a significant amount of data needed for effective or valid use of the phenotyping algorithm, suggesting that an implementation would require adaptation for handling missing data. Furthermore, specific to the condition of RA, the features selected by the algorithm are valid but insufficient in cases where no RF lab was available but who otherwise had ESR or CRP labs sufficient for a diagnosis of RA per the 2010 ACR/EULAR criteria, an insufficiency that arose because of incompatible choices made during design and testing. In the context of understanding the PheKB algorithm's performance in applications outside downstream biomedical research, this study points out areas of improvement. It also provides insight on the impact that phenotyping algorithm design choices can have on their ability to capture the desired patient population and how design choices made at different stages of a phenotyping algorithm development process ultimately affect its applicability and generalizability. It is important that users appreciate the diversity of applications that a phenotyping algorithm can be used for, which risks an algorithm being used at cross purposes with its intended design.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to acknowledge Dr. Anders Garlid and Robert Follett for their advice in the design of the study and for their help in clinical data extraction and management, and the UCLA Institute for Precision Health (IPH) for their support of this project.

Funding

Funding support for this project came in part from NIH/NIBIB T32 EB016640 and NIH/NCATS UL1 TR001881.

References

[1]. Anderson AJM, Click B, Ramos-Rivers C, Babichenko D, Koutroubakis IE, Hartman DJ, Hashash JG, Schwartz M, Swoger J, Barrie AM, Dunn MA, Regueiro M, Binion DG, Development of an inflammatory bowel disease research registry derived from observational electronic health record data for comprehensive clinical phenotyping, Dig. Dis. Sci 61 (11) (2016) 3236–3245, 10.1007/s10620-016-4278-z. [PubMed: 27619390]

[2]. Arzt M, Oldenburg O, Graml A, Erdmann E, Teschler H, Wegscheider K, Suling A, Woehrle H, the Schla HFI, Phenotyping of sleep-disordered breathing in patients with chronic heart failure with reduced ejection fraction—the SchlaHF registry, JAHA 6 (12) (2017), 10.1161/jaha.116.005899.

[3]. Torous J, Staples P, Barnett I, Sandoval LR, Keshavan M, Onnela J-P, Characterizing the clinical relevance of digital phenotyping data quality with applications to a cohort with schizophrenia, npj Dig. Med 1 (1) (2018) 15, 10.1038/s41746-018-0022-8.

[4]. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, Masys DR, Ritchie MD, Roden DM, Struewing JP, Wolf WA, The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies, BMC Med. Genom 4 (1) (2011), 10.1186/1755-8794-4-13.

[5]. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, Pacheco JA, Tromp G, Pathak J, Carrell DS, Ellis SB, Lingren T, Thompson WK, Savova G, Haines J, Roden DM, Harris PA, Denny JC, PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability, J. Am. Med. Inf. Assoc.: JAMIA 23 (6) (2016) 1046–1052, 10.1093/jamia/ocv202.

[6]. Zhou S-M, Fernandez-Gutierrez F, Kennedy J, Cooksey R, Atkinson M, Denaxas S, Siebert S, Dixon WG, O'Neill TW, Choy E, Sudlow C, Follow-up UKB, Outcomes G, Brophy S, Defining disease phenotypes in primary care electronic health records by a machine learning approach: a case study in identifying rheumatoid arthritis, PLoS ONE 11 (5) (2016) e0154515, 10.1371/journal.pone.0154515.

[7]. Ahmed HU, El-Shater Bosaily A, Brown LC, Gabe R, Kaplan R, Parmar MK, Collaco-Moraes Y, Ward K, Hindley RG, Freeman A, Kirkham AP, Oldroyd R, Parker C, Emberton M, Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study, The Lancet 389 (10071) (2017) 815–822, 10.1016/s0140-6736(16)32401-1.

[8]. Jacobson NC, Weingarden H, Wilhelm S, Using digital phenotyping to accurately detect depression severity, J. Nerv. Ment. Dis 207 (10) (2019) 893–896, 10.1097/nmd.0000000000001042. [PubMed: 31596769]

[9]. Gabriel SE, The sensitivity and specificity of computerized databases for the diagnosis of rheumatoid arthritis, Arthritis Rheum. 37 (6) (1994) 821–823, 10.1002/art.1780370607. [PubMed: 8003054]

[10]. Singh JA, Holmgren AR, Noorbaloochi S, Accuracy of Veterans Administration databases for a diagnosis of rheumatoid arthritis, Arthritis Rheum. 51 (6) (2004) 952–957, 10.1002/art.20827. [PubMed: 15593102]

[11]. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, Healey LA, Kaplan SR, Liang MH, Luthra HS, Medsger TA, Mitchell DM, Neustadt DH, Pinals RS, Schaller JG, Sharp JT, Wilder RL, Hunder GG, The american rheumatism association 1987 revised criteria for the classification of rheumatoid arthritis, Arthritis Rheum. 31 (3) (1988) 315–324, 10.1002/art.1780310302. [PubMed: 3358796]

[12]. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO, Birnbaum NS, Burmester GR, Bykerk VP, Cohen MD, Combe B, Costenbader KH, Dougados M, Emery P, Ferraccioli G, Hazes JMW, Hobbs K, Huizinga TWJ, Kavanaugh A, Kay J, Kvien TK, Laing T, Mease P, Ménard HA, Moreland LW, Naden RL, Pincus T, Smolen JS, Stanislawska-Biernat E, Symmons D, Tak PP, Upchurch KS, Vencovský J, Wolfe F, Hawker G, 2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative, Arthritis Rheum. 62 (9) (2010) 2569–2581, 10.1002/art.27584. [PubMed: 20872595]

[13]. Partners Phenotyping G 2016, 2016, Available from: https://phekb.org/phenotype/585.

[14]. Yu S, Liao KP, Shaw SY, Gainer VS, Churchill SE, Szolovits P, Murphy SN, Kohane IS, Cai T, Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources, J. Am. Med. Inform. Assoc 22 (5) (2015) 993–1000, 10.1093/jamia/ocv034. [PubMed: 25929596]

[15]. Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, Pacheco JA, Boomershine CS, Lasko TA, Xu H, Karlson EW, Perez RG, Gainer VS, Murphy SN, Ruderman EM, Pope

RM, Plenge RM, Kho AN, Liao KP, Denny JC, Portability of an algorithm to identify rheumatoid arthritis in electronic health records, J. Am. Med. Inform. Assoc 19 (e1) (2012) e162–e169, 10.1136/amiajnl-2011-000583. [PubMed: 22374935]

[16]. Diaz-Garelli F, Strowd R, Lawson VL, Mayorga ME, Wells BJ, Lycan TW, Topaloglu U, Workflow differences affect data accuracy in oncologic EHRs: a first step toward detangling the diagnosis data babel, JCO Clin. Cancer Inf 4 (2020) 529–538, 10.1200/CCI.19.00114.

[17]. Che Z, Liu Y (Eds.), Deep learning solutions to computational phenotyping in health care, in: 2017 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, New Orleans, LA, 2017 2017/11.

[18]. Wells BJ, Nowacki AS, Chagin K, Kattan MW, Strategies for handling missing data in electronic health record derived data, eGEMs 1 (3) (2013) 7, 10.13063/2327-9214.1035.

[19]. Rusanov A, Weiskopf NG, Wang S, Weng C, Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research, BMC Med. Inf. Decis. Making 14 (1) (2014) 51, 10.1186/1472-6947-14-51.

6.

Statement of significance

Problem or Issue	Phenotyping algorithms are used inconsistently or inappropriately to their design, leading to impaired performance.
What is Already Known	Algorithms for clinical phenotyping of patients have been developed for a variety of purposes. However, the relationship between an algorithm’s design and its usability is unclear.
What this Paper Adds	Through an evaluation of a phenotyping algorithm for rheumatoid arthritis on a novel population, this paper details important considerations when building and using a phenotyping algorithm by showing how design choices made during a phenotyping algorithm’ s development lifecycle can affect applicability and generalizability. It studies holistically how feature choice, cohort selection, and clinical workflow affect an algorithm’s scope.

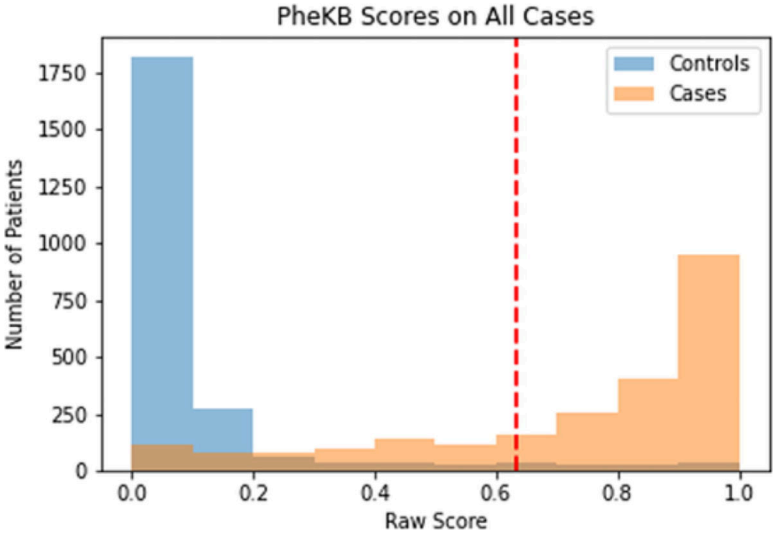

Fig. 1.

Histogram of raw PheKB algorithm scores of cases and controls. The red dotted line represents the PheKB positive cutoff score of 0.632. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

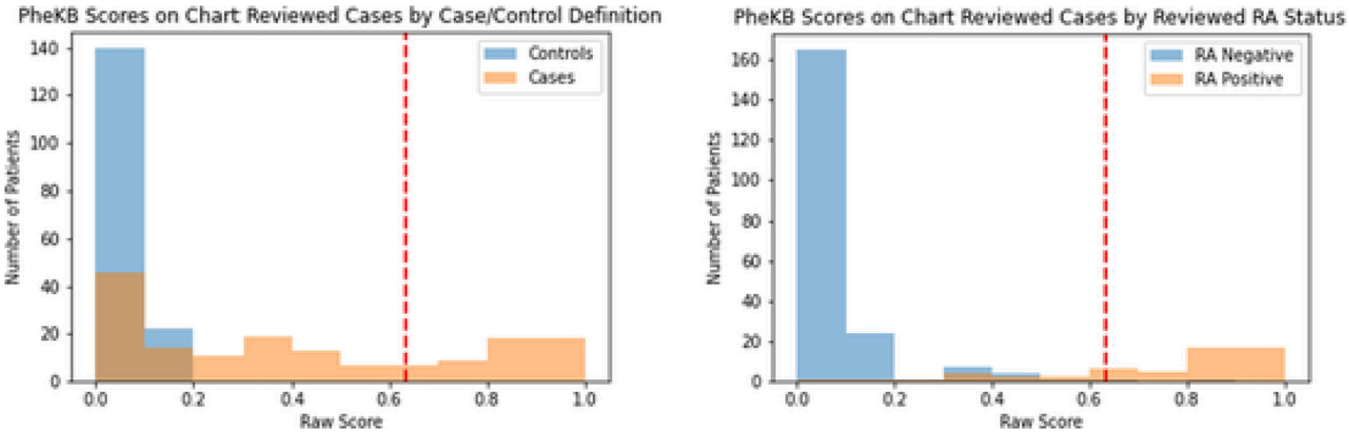

Fig. 2.

Histogram of PheKB algorithm scores of ACR criteria subset, separated by case/control (left) and by RA status based on the 2010 ACR/EULAR criteria (right). The red dotted line represents the PheKB positive cutoff score of 0.632. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1

2010 ACR/EULAR criteria for rheumatoid arthritis. An individual must score a total of 6 points across the different categories to be diagnosed with RA using this framework.

Category	Criteria	Points
Joint involvement	1 large joint	0
	2–10 large joints	1
	1–3 small joints	2
	4–10 small joints	3
	>10 small joints	5
Serology	Negative rheumatoid factor (RF) and negative anti-CCP antibodies (ACPA)	0
	Low-positive RF or low-positive ACPA	2
	High-positive RF or high-positive ACPA	3
Acute-phase reactants	Normal C-reactive protein (CRP) & normal erythrocyte sedimentation rate (ESR)	0
	Abnormal CRP or abnormal ESR	1
Duration	<6 weeks	0
	6 weeks or more	1

Table 2

Performance of PheKB algorithm on entire cohort.

	Cases N = 2385	Controls N = 2381	All N = 4766
PheKB Positive	1714 (36.0 %)	113 (2.4 %)	1827
PheKB Negative	671 (14.1 %)	2268 (47.6 %)	2939
All	2385	2381	4766

Table 3

Performance of PheKB algorithm against 2010 ACR/EULAR criteria.

	PheKB Positive	PheKB Negative	Total
ACR/EULAR reviewed RA Positive	38 (12.9 %)	13(4.4 %)	51 (17.3 %)
ACR/EULAR reviewed RA Negative	1 (0.3 %)	186 (63.3 %)	187 (63.6 %)
ACR/EULAR reviewed, Unclear	0	7 (2.4 %)	7 (2.4 %)
ACR/EULAR reviewed, Missing Data	5(1.7 %)	44 (15.0 %)	49 (16.7 %)
Total	44 (15.0 %)	250 (85.0 %)	294